The exogenous variables are the input prices, *wage rate* (w), and the *rental price of capital* (r). The wage rate, or wage for short, is measured in \$/hour and the rental price of capital is \$/machine. We assume that the firm is a price taker in the markets for labor and capital so it can rent as much L and K as it wants at the given w and r. The amount to produce, q, is also an exogenous variable in this problem. We are not considering how much should be produced, but what is the best way to produce any given amount of output. Finally, the firm's technology, the production function, $f(L, K)$, is also given.

Because the firm has to produce a given amount of output, we know this is a constrained optimization problem. Our work in the Theory of Consumer Behavior has made us expert at solving this kind of problem. As you will see, the analysis is similar, but there are some striking differences.

One thing that does not change is our framework. We first explore the constraint to determine our options, then focus on the goal (to minimize TC), and, finally, we will combine the two to find the initial optimal solution.

The Constraint

The menu of options available to the firm is given by the isoquant. It serves as the constraint because the firm is free to choose L and K on the condition that it must produce the assigned level of output. Mathematically, the equation for the constraint is simply the production function, $q = f(L, K)$.

STEP Open the Excel workbook *InputCostMin.xls*, read the *Intro* sheet, then go to the *Isoquant* sheet to see the isoquant displayed in Figure 11.1.

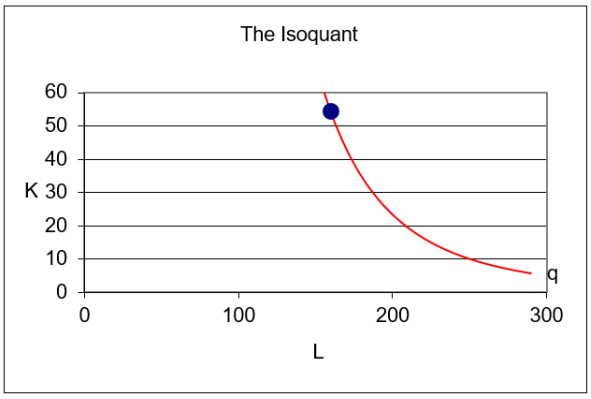


Figure 11.1: An isoquant from a Cobb-Douglas production function. *Source: InputCostMin.xls!Isoquant.*

Like the budget constraint in the Theory of Consumer Behavior gives us consumption possibilities, the isoquant gives the firm its feasible input options. All combinations below and left of the isoquant are ruled out. For example, there is no way to produce 100 units of output, holding quality and everything else constant, with the L,K combination of 100,20. The technology is simply not advanced or powerful enough to make 100 units of output with 100 hours of work and 20 machines.

The points above and to the right of the isoquant are feasible, but they are clearly wasteful. In other words, the firm could produce 100 units of output with an L,K combination of 250,50, but the isoquant tells the firm it does not need that much labor and capital to make 100 units. At 250,50, it could travel straight down to $K = 10$ and still produce $q = 100$ or straight left (on the horizontal line at $K = 50$) until it hit the isoquant and use a lot less labor. The firm could also travel in a diagonal, southwest direction until it hit the isoquant to economize on both inputs.

Points off the isoquant to the northeast (such as 250,50) are said to be *technically inefficient*. The *inefficient* part tells us that the firm is not minimizing its total cost at that point; *technical* describes the fact that the firm is not organizing its inputs so as to maximize output. In other words, the firm is not correctly solving the engineering optimization problem represented by the production function. Making 100 units of output with 250 hours of labor and 50 machines means that you are not getting the most out of your labor and capital. Economists call this situation technically inefficient.

Since the firm cannot choose a combination below the isoquant and it is wasteful to choose a combination above the isoquant, we know the answer has to lie on the isoquant.

STEP Use the scroll bar next to cell B11 to see the input mixes the firm might choose. As you change cell B11, the cell below changes also. It has a formula that computes the amount of K needed to produce the required output when you choose a value for L.

The idea is quite clear: The firm will roll around the isoquant in search of the best combination. Rolling is a good word choice and image to remember—the firm is free to choose a point high up or roll down to the bottom right. Because we do not have the input prices, we cannot find the optimal solution with the isoquant alone.

STEP Change the exogenous variables to see how the isoquant is affected. Increases in A, c, and d pull the isoquant down. That makes sense given that these shocks are all productivity enhancing and the firm will need less L and K to make the given $q = 100$.

Lowering q has the same effect, but this is not a productivity shock. You are simply telling the firm it does not have to produce as much as before so it makes sense that it can use less labor and capital.

Notice how the constraint for this input cost minimization problem is a curve, not a line like it was for the utility maximization problem. Mathematically, that does not matter much, but it will impact the graph we draw to show the initial solution.

Goal

With the constraint in hand, we are ready to model the goal. In this problem, the goal is represented by a series of *isocost* (equal cost) lines.

Total cost is $TC = wL + rK$. If we solve this equation for K (in order to graph it in L-K space), we get the equation of a line:

$$TC = wL + rK$$

$$rK = TC - wL$$

$$K = \frac{TC}{r} - \frac{w}{r}L$$

The K (or y axis) intercept is $\frac{TC}{r}$ and the slope is $-\frac{w}{r}$.

Isocosts are a little tricky at first because you are used to seeing a linear constraint and a set of indifference curves. Input cost minimization has a curved constraint and a set of linear isocosts. In the equation of the line above, TC can take on any value. Thus, there is an isocost for $TC = \$500$ and another for $TC = \$500.01$ and an isocost for every single dollar amount. Every L,K point is on an isocost and the L,K points that have the same TC are on the same isocost.

STEP Proceed to the *Isocost* sheet to see how the isocost lines are used to find the optimal solution.

Each point on a particular isocost line has the exact same total cost. So, the point on Figure 11.2 (and on your screen) has a cost of \$500 (since 2 x 190 + 3 x 40 = 500).

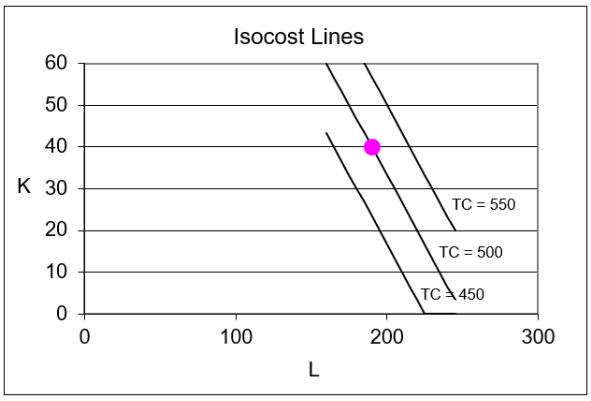


Figure 11.2: Three representative isocost lines.
Source: InputCostMin.xls!Isocost.

STEP Click the Show Goal to see how the firm's cost minimization goal is represented on this graph.

The firm can move to a new point by choosing a different combination of L and K. If the new point has the same TC of \$500 as the initial point, then it will be on the same \$500 isocost.

STEP Increase L by 30 and decrease K by 20 so you will be at another point on the same isocost line of \$500.

Now you know that all points on the TC = \$500 isocost line share the same total cost of \$500. It is also obvious that the slope of each isocost line is $-\frac{2}{3}$ since $w = 2$ and $r = 3$.

Because the firm can choose the input mix, it can choose any combination of L and K, provided that the chosen combination can produce the given amount of output. The firm wants to hire as few inputs as it can (to save on costs), but it has to meet the production target. How can it solve this problem?

The Initial Optimal Solution

We have the constraint (the isoquant) and the goal (get to the lowest isocost possible), so now we combine the two to find the optimal solution.

STEP Proceed to the *OptimalChoice* sheet.

The starting position shows an L,K combination that costs \$482.81. You can confirm this number both in cell B7 and on the chart (the middle label for the middle line).

The idea is to be on the lowest isocost line (i.e., the one with the smallest intercept) that is just touching the isoquant because that means the firm will be minimizing the total cost of producing the given level of output.

Clearly, the starting position is not optimal. You can see that the isocost is intersecting the isoquant. This information is also revealed by the slope and TRS information below the chart. The TRS, which is the slope of the isoquant at a point, is greater (in absolute value) than the slope of the isocost line at that point.

At the opening position, the firm is said to suffer from *allocative inefficiency* because it is on the isoquant, but it fails to choose the cost minimizing input mix. Because it is on the isoquant, we know it is not technically inefficient—it is using the opening combination of L and K to get the maximum output. The problem is that it is using the wrong combination of inputs in the sense that there is a cheaper way to produce the given output.

We know there are two ways to solve optimization problems: analytically and numerically. Because we have Excel and the problem implemented on the sheet, we begin with the numerical approach.

STEP Run Solver. The optimal solution is depicted by the canonical graph displayed in Figure 11.3.

Solver's answer, which is correct, has the firm choose an L,K combination whose isocost just touches the isoquant. There is no cheaper combination that can produce 100 units with the existing technology (given by the production function). If the firm went to an isocost that was one cent lower, it could not rent enough L and K to make 100 units of output.

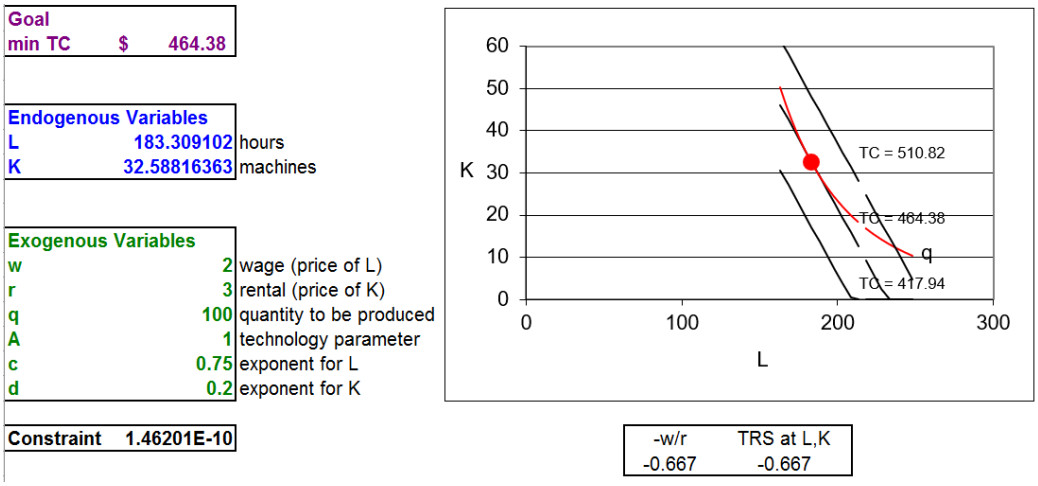


Figure 11.3: The initial optimal solution.
Source: InputCostMin.xls!OptimalChoice, after running Solver.

We can confirm Solver's result by applying the Lagrangean method to solve this constrained optimization problem.

We start by writing down the problem, using the parameter values from the *OptimalChoice* sheet.

$$\min_{L,K} TC = 2L + 3K$$
$$\text{s.t. } 100 = L^{0.75}K^{0.2}$$

The first step is to rewrite the constraint so that it is equal to zero.

$$100 - L^{0.75}K^{0.2} = 0$$

The second step is to form the Lagrangean by adding lambda, λ, times the rewritten constraint to the original objective function. We use an extra-large L for the Lagrangean function that is not at all related to the L for labor.

$$\min_{L,K,\lambda} L = 2L + 3K + \lambda(100 - L^{0.75}K^{0.2})$$

The third step to finding the optimal solution is to take the derivative of the Lagrangean with respect to each endogenous variable and set each derivative to zero (giving us the first-order conditions).

$$\frac{\partial L}{\partial L} = 2 - 0.75\lambda L^{-0.25} K^{0.2} = 0$$
$$\frac{\partial L}{\partial K} = 3 - 0.2\lambda L^{0.75} K^{-0.8} = 0$$
$$\frac{\partial L}{\partial \lambda} = 100 - L^{0.75} K^{0.2} = 0$$

The fourth, and last, step is to solve this system of equations for L^*, K^*, and λ^*. The system of three equations contains the answer, that is, the values of L and K that minimize TC. Our task is to use the equations to find these values that satisfy the three equations.

There are many ways to solve the system, but we will use the same approach that we used in the Theory of Consumer Behavior. We will reduce the system from 3 to 2 to 1 equation and unknown.

We move the terms with lambda in the first two equations to the right-hand side and then divide the first equation by the second. The Cobb-Douglas production function is easy to work with because the exponents of L and K sum to -1 and 1, respectively, when you apply the $\frac{x^a}{x^b} = x^{a-b}$ rule.

$$\frac{2}{3} = \frac{0.75\lambda L^{-0.25} K^{0.2}}{0.2\lambda L^{0.75} K^{-0.8}}$$
$$\frac{2}{3} = \frac{0.75 L^{-0.25} K^{0.2}}{0.2 L^{0.75} K^{-0.8}}$$
$$\frac{2}{3} = \frac{3.75K}{L}$$
$$L = 5.625K$$

As you can see above, this strategy cancels the lambdas and gives an expression for $L = f(K)$, which, in conjunction with the third first-order condition, reduces the system to two equations with two unknowns.

$$L = 5.625K$$
$$100 - L^{0.75} K^{0.2}$$

We substitute the expression for L into the constraint (the third first-order condition) and solve for K^*.

$$
\begin{aligned}
100 - [5.625K]^{0.75} K^{0.2} &= 0 \\
100 &= 3.6525 K^{0.75} K^{0.2} \\
27.3784^{\frac{1}{0.95}} &= (K^{0.95})^{\frac{1}{0.95}} \\
27.3784^{\frac{1}{0.95}} &= \left(K^{0.95}\right)^{\frac{1}{0.95}} \\
K^* &= 32.588
\end{aligned}
$$

Then, substituting K^* back into the expression for $L = f(K)$, we get L^*.

$$
\begin{aligned}
L &= 5.625K \\
L &= 5.625[32.588] \\
L^* &= 183.31
\end{aligned}
$$

Substituting L^* and K^* into the original objective function, we can compute the minimum cost of producing 100 units.

$$
\begin{aligned}
TC &= 2L + 3K \\
TC &= 2[183.1] + 3[32.588] \\
TC^* &= \$464.38
\end{aligned}
$$

The analytical solution agrees with Solver's answer.

The work we did in dividing the first equation by the second yields an equimarginal condition that is similar to the MRS $= \frac{p_1}{p_2}$ rule from constrained utility maximization. At the optimal solution, we have

$$\frac{2}{3} = \frac{3.75K}{L}$$

The left-hand side is the input price ratio and the right-hand side is the TRS. Thus, at the optimal solution we know that input price ratio must equal the TRS. This is a mathematical statement of the tangency we see in Figure 11.3.

If this equimarginal condition is not met, but the firm is on the isoquant (i.e., it is technically efficient), then we have allocative inefficiency. If $|TRS| > \frac{w}{r}$, then the isocost is cutting the isoquant and the firm can lower total costs by rolling down the isoquant. The reverse, of course, applies if $|TRS| < \frac{w}{r}$.

STEP If you have not done so already, double-click inside the box around cell J25 and use the scroll bar to show how the isocost and isoquant graph matches up with the TRS $= \frac{w}{r}$ equimarginal condition.

Comparing Consumer and Firm

Figure 11.3 bears a striking resemblance to the canonical graph used in the Theory of Consumer Behavior and the analytical work also contains strong similarities, but there are some critical differences between the consumer and firm optimization problems. Figure 11.4 presents a side-by-side comparison to highlight the contrasts between them.

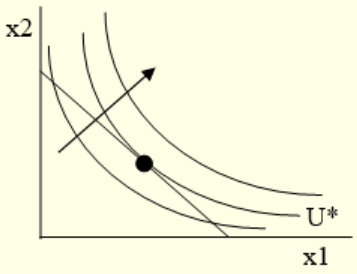
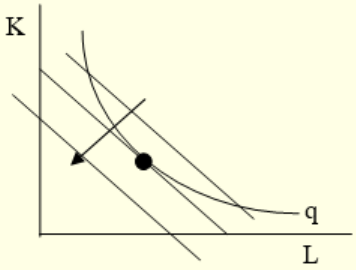
Characteristic	Theory of Consumer Behavior	Isoquant Side of the Theory of the Firm
Goal	maximize utility ($U = f(x_1, x_2)$)	minimize total cost ($TC = wL + rK$)
Canonical Graph of Initial Solution	 The line is the constraint. The curves are the goal. One line and several (representative) curves.	 The curve is the constraint. The lines are the goal. One curve and several (representative) lines.
Function Properties	Utility is a fiction that represents preferences. The actual value of the utility function has no meaning.	The production function gives q as a measurable, cardinal quantity.
Maximum Value Function	The numerical value of maximum utility (U^*) is not important.	The numerical value of minimum total cost (TC^*), measured in dollars, is the *highest* priority.
Key Comparative Statics Exercise	Demand Curve $x_1^* = f(p_1)$, ceteris paribus	Cost Function $TC^* = f(q)$, ceteris paribus
Interpreting λ^*	No real economic meaning because utility is merely ordinal.	λ^* is marginal cost, the additional cost of producing more output.

Figure 11.4: Comparing consumer and firm optimization problems.

It makes sense to use the knowledge and skills learned from the Theory of Consumer Behavior, but do not fall into a false sense of security. The input cost minimization problem has its own characteristics and terminology.

Cost Minimization is One of Three Problems

The Theory of the Firm is actually a set of three interrelated optimization problems. The initial solution to the firm's cost minimization problem focuses attention on the cheapest combination of inputs to produce a given level of output.

We can apply the same techniques we used to solve the consumer's utility maximization problem. The canonical graph is similar to the standard graph from the Theory of Consumer Behavior, but as Figure 11.4 shows, there are substantial differences between utility maximization and cost minimization.

One important similarity is the continued use of the comparison of a price ratio to the slope of a curve to determine whether the optimal solution has been found. In the case of the constrained cost minimization problem, the firm will choose that combination of inputs where TRS $= \frac{w}{r}$. If this condition is not met, the direction of the inequality (¿ or ¡) tells us which way the firm should move to find the minimum total cost.

Now that we understand the firm's cost minimization problem and have found the initial solution, we are prepared to take the next step—comparative statics analysis. The economic approach is unrelenting and monotonous. We apply the same framework to every problem. Through practice and repetition, you will learn to think like an economist.

Exercises

1. The *Q&A* sheet asks you to change r to 30 and use Solver to find the initial solution. Find the initial solution to this same problem via analytical methods and compare the two results. Are they the same? Show your work.

2. The fixed proportions production function, $q = min\{\alpha L, \beta K\}$ is analogous to the perfect complements utility functional form. Suppose $\alpha = \beta = 1$, $w = 10$, $r = 50$, and $q = 100$. Find L^*, K^*, and TC^*. Show your work. Use Word's Drawing Tools to draw a graph of the optimal solution.

3. Given the quasilinear production function, $q = \sqrt{L} + K$, and input prices $r = 2$, and $w = 5$, find the cheapest way to produce 1000 units of output. Use analytical methods and show your work.

4. Set up the problem in question 3 in Excel and use Solver to find the optimal solution. Take a screen shot of the solution on your spreadsheet and paste it into a Word document.

5. Can isoquants intersect? Explain why or why not.

References

The epigraph is from page 1044 of Joseph Schumpeter's *History of Economic Analysis* (published in 1954, shortly after his death). This classic traces the intellectual history of economics from Aristotle to the 20th century. Schumpeter reviews the theories and visions of giants like Adam Smith and Karl Marx, but also an incredible number of philosophers and economists that will not be familiar to you.

Ragnar Frisch, credited by Schumpeter with inventing the term *isoquant*, had a knack for inventing words, e.g., macroeconomics and econometrics. Luckily, "substitumal cost flexibility" did not catch on. A Norwegian, Frisch was part of an exceptionally strong quantitative and empirical tradition in Scandinavian economics that remains alive to this day.

> Several hundred years ago, an unknown inventor combined charcoal, sulfur and saltpeter and lit it afire. When the dust settled the world was changed forever.
>
> *The Story of the Gun*

11.2 The Enfield Arsenal

This chapter departs from the usual presentation style employed in this book. There is no Excel workbook associated with this application. Instead, you will be given the opportunity to answer questions and the answers are provided at the end of the chapter. Each question is highlighted by the usual *Step* marker. Try to work out each question on your own before looking at the answers.

There are four goals:

1. To understand cost minimization with isoquants and isocosts.

2. To provide an example of how theory can be applied to real-world problems.

3. To illustrate how economics can help us understand what we observe.

4. To see that economics has wide and varied application.

The inspiration and source of this application of cost minimization is from Edward Ames and Nathan Rosenberg, "The Enfield Arsenal in Theory and History," *The Economic Journal* (Vol. 78, No. 312, December, 1968), pp. 827–842, www.jstor.org/stable/2229180.

Ames and Rosenberg were economic historians and that immediately leads to a puzzler: how are economic historians different from regular historians? The answer has to do with the economic approach. Once trained as an economist, the methods and way of thinking can be applied to events and outcomes from the past. This is what Ames and Rosenberg did with the Enfield Arsenal. But before we get to that, we need to understand what rifling is all about.

Rifling

Rifles are a relatively recent innovation in firearms. Figure 11.5 shows an early version of the famous Enfield rifle with labels for the three main parts: the lock, stock, and barrel.

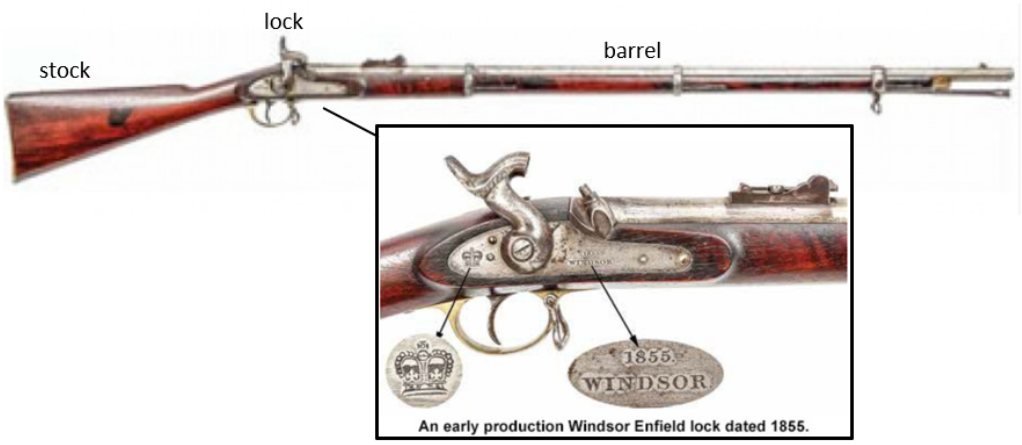


Figure 11.5: The Enfield rifle that was stored in the Enfield Arsenal. *Source:* collegehillarsenal.com/index.php?route=information/blogger&blogger_id=9

It is the barrel that distinguishes rifles from smooth-bore muskets. The barrel of a rifle has a striated pattern that spins the bullet, increasing velocity and accuracy compared with a ball from a musket.

STEP Watch this short video on rifling from *The Story of the Gun*: vimeo.com/25200729.

But the Enfield rifle was important not because it rifled, but because of how it was made.

The American System of Manufacturing

Ames and Rosenberg (p. 827) explain what the Enfield Arsenal was in the introduction to their paper:

> This paper analyses a particular historical event, the establishment of the Enfield Arsenal, in the context of the literature cited. The British Government committed itself to the construction of the Enfield Arsenal in 1854 because it wished to be able to make

> large numbers of rifles for an impending war with Russia (now known as the Crimean War). The event is important because it marked the beginning of the movement of mass-production techniques from the United States to Europe. Technical changes in gunmaking in the nineteenth century were a major source of new machine techniques; and industrialisation in the nineteenth century is overwhelmingly the history of the spread of machine making and machine using.

So an arsenal is an armory, a warehouse of guns and ammunition. Enfield is a place in England and the Enfield Arsenal is literally a building constructed by the British government in 1854 that would be used to store rifles made with mass-production techniques.

The Enfield Arsenal was special because it was the first time the British would use mass-production techniques to make weapons. Up to this point, the British had made guns the old-fashioned way—by hand in the small shops of thousands of skilled artisans in the area around Birmingham. The stock was carefully carved by an experienced craftsman who fitted the stock with the lock and barrel. It was like a tailor making a bespoke suit—each rifle was one of a kind. A work of art.

Ames and Rosenberg (p. 832, footnotes omitted) point out that making the stock by hand was especially slow and expensive to do:

> The gunstock was one of the most serious bottlenecks in firearms production. In England, at the time of the Parliamentary hearings, out of about 7,300 workmen in the Birmingham gun trade, the number of men employed in making gunstocks totalled perhaps as many as 2,000. Its highly irregular shape for long seemed to defy mechanical assistance, and the hand-shaping of the stock was a very tedious operation. Furthermore, the fitting and recessing of the stock so that it would properly accommodate the lock and barrel were extremely time-consuming processes, the proper performance of which required considerable experience. With Birmingham methods, it required 75 men to produce 100 stocks per day. Using the early (1818) version of the Blanchard lathe, 17 men could produce 100 stocks per day.

This quotation requires some explanation. First, the reason for the Parliamentary hearings was that British politicians were angry with the Birmingham gunsmiths for not adopting fast, efficient mass-production techniques.

There was an investigation and testimony was given. How could upstart Americans have better technology than the British, a nation that dominated the entire globe? It was a national embarrassment!

Second, the quotation mentions the *Blanchard lathe*. This is a machine that cuts and shapes wood (and other materials like metal), but it is easier to understand if you see it.

STEP Watch this video, vimeo.com/25200825, to understand how a lathe works and how the production of precision parts makes Diderot's dream come true.

The video explained that the new country of the United States of America needed weapons so the Springfield Armory was built in 1794 in Springfield, Massachusetts. At first, stocks were made by hand, just like in Birmingham. They were then individually fitted to each rifle.

But in 1818, the Blanchard lathe burst on the scene. The narrator, echoing the British Parliamentary hearings, says, "Prior to the Blanchard Lathe, it took one to two days to make a rifle stock by hand. Now, a twelve-year-old boy could turn out a dozen stocks in a single day."

The Blanchard lathe enabled a reorganization of the production process. In factories in the northeast, the United States began to use mass-production techniques to make rifles and pistols (and then bicycles, sewing machines, typewriters, and so on). This is the American system of manufacturing. A key element is that a machine can make a precision part so many almost identical parts can be made and then the product is assembled.

The video points out that the history of gun-making is closely tied to the rise of mass-production techniques and precision manufacturing. In the video, William Ruger cites an idea from French philosopher Denis Diderot (1713–1784). Ruger says Diderot's theory at that time was that "It would be possible to make all of the individual parts alike and then at the last minute assemble them, rather than fitting them together as you went, which was the customary thing up to that time."

Adam Smith (1723 - 1790) was a contemporary of Diderot. For Smith, *the division of labor* explained the explosion in productivity that he saw all around him as the Industrial Revolution began.

Breaking production into a series of steps and then assembling the parts enables many more units of output to be produced. This is called the division of labor. Smith emphasized several reasons for the greater productivity enabled by the specialization of labor:

1. Practice makes perfect: focusing on a single task makes you very good at it.

2. Saves time: no need to set things up when you move to a new task.

3. Innovation: adjustments are made by workers who are expert in a particular task.

Machines such as the Blanchard lathe feed into the division of labor by enabling much finer specialization. For rifles, production with a lathe meant that they were no longer one of a kind. They were all alike and could be easily connected to the lock and barrel to make a rifle.

By applying Diderot's theory of assembling perfectly fitted parts and Smith's division of labor, the Springfield Armory was able to enjoy a huge increase in productivity compared with Birmingham methods.

So now you know exactly what a lathe is and how mass production played a key role in the exponential increases in productivity during the Industrial Revolution, but there is one more important advantage to mass production. Let's see if you can figure it out.

STEP What are the tremendous advantages of interchangeable parts in a rifle (or anything else for that matter) for the end user?

The answer is at the end of this section, but take a few minutes to think about the question. What advantage would soldiers using rifles that were all alike have over enemies using individually made rifles?

Two Big Questions

The key date in this story is 1854. Until this time, the British used Birmingham methods, which means an experienced craftsman made each entire gun by hand. They shaped the stock, then attached it to the lock and the barrel. Each part was slightly different and could not be easily replaced if damaged.

Beginning in 1854, rifles produced for the Enfield Arsenal, however, were made with interchangeable parts (including stocks made on lathes) that could be put together in an assembly line. Once in use, broken parts could be removed and new ones snapped on.

Ames and Rosenberg (pp. 839–840, footnotes omitted) sum up the situation:

> As of 1785, neither the British nor the Americans could make guns with interchangeable parts. As of 1815, Americans could make guns with interchangeable metal parts, but could not make interchangeable gunstocks. As of 1820, they could make interchangeable gunstocks. At any date, presumably, they could use not only current methods but earlier methods which these had displaced.

The United States had been mass-producing guns with interchangeable parts since 1815. The British waited until 1854 to use the superior, mass-production techniques. This gives rise to two big questions:

1. Why did the British wait so long to use mass-production techniques to make rifles with interchangeable parts?
2. Why did the British switch to mass-production techniques in 1854?

1. Why Did the British Wait so Long?

A possible answer to the puzzle of why British gunsmiths did not adopt the new technology is that the British did not know about the Blanchard lathe so that is why they did not use it?

STEP Is lack of knowledge about American technology a good answer? Why or why not?

Another possible answer is poor management. Maybe British rifle manufacturers were lazy, stupid, and careless? The right answer—adopt mass-production techniques—was staring them in the face and they ignored it.

STEP Is managerial failure a good answer? Why or why not?

Economic historians give a third answer to why the British did not adopt the Blanchard lathe. They use the economic way of thinking. They look for differences in the environment that would lead to different optimal solutions.

In other words, Ames and Rosenberg stop searching for why the British made a mistake and accept the fact that their refusal to adopt mass-production techniques was actually smart and correct. They look for reasons that justify the British decision to reject the Blanchard lathe.

This is crazy, right? It is obvious that mass production is better. Well, it turns out that there are two critical differences between the United States and Britain in the first half of the 19th century that play an important role in deciding how to make rifles.

First, the two countries had quite different labor forces. The British had a cohort of skilled rifle craftsmen and the United States did not. As the Parliamentary hearings noted, there were several thousand skilled craftsmen in Birmingham making stocks and rifles. The United States was a young country with mostly unskilled, male workers. Few skilled craftsmen would emigrate to the United States since they had good, high paying jobs at home.

These supply and demand differences meant that, in the United States, wages for skilled craftsmen were much higher than in Britain, and wages for unskilled labor were lower.

Second, wood was plentiful and cheap in the United States, but it was much more expensive in Britain. Ames and Rosenberg offer the following footnote (p. 831) to help explain why wood plays a critical role:

> Report of the Small Arms Committee, op. cit., Q. 7273-81 and Q. 7520-7521; G. L. Molesworth, "On the Conversion of Wood by Machinery," Proceedings of the Institution of Civil Engineers, Vol. XVII, pp. 22, 45-6. In the discussion which followed Mr. Moleworth's paper Mr. Worssam, a prominent English dealer in woodworking machinery, made some interesting comparative observations which were summarised as follows: "He had seen American machines in operation, and he found that, although they might be adapted for the description of work required in that country, they were not so suitable for English work, in which latter high finish and economy of material were of greatest importance. In America the saws were much thicker than those used in the English saw-mills, so that they consumed more power, wasted more material, and did not cut so clean, or so true, though there was less care required in working them" (ibid., pp. 45-6).

A key point in this long quotation is that American saws (and, of course, lathes) "wasted more material." A British skilled craftsman making stocks from lumber would be careful to "economize" on the material. In America, a 12-year old boy working with a lathe (a dangerous job) would not care at all about wasting wood.

The different endowments of wood in the two countries meant that the Blanchard lathe was much more expensive to operate in Britain than in America.

Now that we know how the United States and Britain differed with respect to (1) wages for skilled and unskilled labor and (2) operating costs for the Blanchard lathe, we are ready to make the case for the economic explanation for why the British waited so long to adopt mass-production techniques.

As is typical in economics, the exposition will rely on graphs. But instead of just reading the explanation, you will try to do it yourself first. The idea is to apply the input cost minimization problem to this scenario. You can, of course, simply jump to the end of the section to see the answers, but you will learn much more if you try to do it yourself first. Follow the instructions and hints offered below and see how close you get. Make sure you understand where you made a mistake or in what ways you were confused.

STEP Draw graphs that show how the different resource endowments and input prices affected the optimal input mix. Use the detailed instructions that follow as a guide. How do the graphs explain why the British waited so long to adopt mass-production techniques?

We will use two sets of two graphs. The first set of two graphs will be for the labor force difference between the United States and Britain. The second set shows the effect of the different endowments of wood.

Begin by drawing a graph representing the British situation in 1820 with respect to using skilled and unskilled labor to make, for example, an order of 1,000 rifles. It should have skilled labor on the y axis and unskilled labor on the x axis. Draw in an isoquant (representing the combinations of skilled and unskilled labor that would make the requested 1,000 rifles).

Draw another graph, next to the first one, that is exactly the same. Your second graph represents the United States' options for making 1,000 rifles in 1820. The fact that both isoquants are the same means that the two countries had access to the same technology and are making the same product.

Next, you need to draw the isocost lines. This is where the difference in labor force comes into play. We know the British have skilled labor and the United States does not—immigrants to the United States were not typically experienced, educated workers, but young, unskilled males. That means the price of skilled labor is much higher in the United States. How is that reflected in the isocosts for your two graphs?

The second set of graphs uses L and K as the inputs. As before, draw a pair of graphs side by side, one for the British and the other for the United States, with machinery on the y axis and labor on the x axis. Include the isoquants. Once again, the isoquants are the same, meaning that the British were aware of and could have used American methods.

The key to the economic explanation for why the British did not do what the Americans were doing lies in the isocosts. Remember that early versions of the Blanchard lathe used a lot of wood and this increases the price of machinery. If r is much higher in Britain than in the United States, how does this affect the isocosts?

Take a moment to look at your two sets of graphs. How can they be used to explain why the British rejected mass production before 1854?

Proceed to the end of this section to check your graphs and answers.

2. Why Did the British Switch in 1854?

The second big question revolves around the British decision to switch in 1854 and mass produce the Enfield rifle. Why did they do this? Why did they abandon their decades-old system of production centered in Birmingham, with a network of many small artisans and smiths that made firearms to individual order or in small batches?

Our first possible answer matches up with the lack of knowledge answer to the first big question. Maybe, in 1854, the British heard that mass-production techniques utilizing the Blanchard lathe were available and immediately moved to adopt the new production methods?

STEP Is sudden awareness of new American technology a good answer? Why or why not?

The second possible answer, like before, relies on management. Maybe they wised up? What if British firearms manufacturers recovered from their slumber and moved quickly to modernize their industry?

STEP Is managerial improvement a good answer for the switch? Why or why not?

You probably got the first two right, but the third one is harder. It might be easy in general terms, but getting the details can be complicated.

The third answer is based on economic reasoning. This means that when we see changes in behavior, we look for changes in the environment. We do not search for events or causes that changed a mistake into the right answer. Instead, we accept that the answer to not use mass production was correct for, say, 1830, but the new optimal solution, in 1854, was to switch to the American system.

This is a key aspect of the economic approach, and it can be challenging to grasp. Our instinct when we see something change is to think of correction or improvement. Economists do not think this way. We see optimization everywhere so if something changes, it was optimal before and it has moved to a new optimal solution because of an exogenous shock.

The search is on for shocks that switch the correct answer from "reject" to "accept" interchangeable parts.

There are two ways in which Britain before 1854 differed from Britain after 1854 and these two ways impacted wages and the operating cost of machinery. These changes act as shocks on the input cost minimization problem and produce a new optimal solution. We first have to figure out the shocks, then we can see how they affect the optimal solution.

STEP Answer these two questions:

1. What happened to the British labor force?
2. What happened to the Blanchard lathe?

You may not be an expert on British labor in the 19^{th} century or know anything about the Blanchard lathe, but you can think about what might have happened. Try to come up with a hypothesis. Think of recent changes in the

labor force that you have heard about, especially those driven by technology (e.g., driverless cars and trucks). Think about how machines, computers, and technology in general have changed over time.

After checking your answer at the end of this section, so now you know what happened, you are ready to draw graphs that illustrate the economic historian's explanation for the British switch in production technique.

STEP Draw graphs that show how the changes mentioned affected the optimal input mix. How do the graphs explain why the British switched to mass-production techniques in 1854?

Draw two pairs of graphs just like before (unskilled and skilled labor on one and machinery and labor on the other), but this time we are comparing environments before and after 1854 in Britain (the United States has nothing to do with this). First, compare the optimal mix of unskilled and skilled labor for Britain in 1820 versus 1854. Remember that the skilled craftsmen died and were not replaced so the skilled wage rate rose. How does this make the 1854 graph different from the 1820 graph?

In the second set of two graphs, with machinery and labor on the axes, we know that machinery got better and better (wasting less and less wood) over time so r fell. What will this shock do to the isocost lines?

Check out the suggested answers at the end when you are finished. Take the time to debug any mistakes. Make sure you understand how the isocost lines shift and how the comparison of two graphs yields answers to the questions.

Evaluating this Application

At the beginning, we had four goals:

1. To understand cost minimization with isoquants and isocosts.

2. To provide an example of how theory can be applied to real-world problems.

3. To illustrate how economics can help us understand what we observe.

4. To see that economics has wide and varied application.

You decide to what extent the goals were met. At the very least, you learned a little about American manufacturing in the 19$^{\text{th}}$ century and rifles (including where the phrase "lock, stock, and barrel" comes from).

The application should help you understand the conventional isoquant–isocost graph and the firm's input cost minimization problem. Remember that the higher the price of the input on the x axis or lower the price of the input on the y axis, the steeper the isocosts.

But the real deep learning and big picture idea concerns how economists view the world. This is called economic reasoning or the economic approach. We did "an economic analysis of the Enfield Arsenal."

The idea is that economics is not a discipline organized around content (the stock market or money, for example), but a way of thinking. Economists often interpret observed behaviors as optimal solutions to optimization problems and they see change as driven by a shock that takes us from one optimal solution to another.

Thinking like an economist is difficult and sometimes counter-intuitive, but it can provide an interesting perspective on the world. Certainly, Ames and Rosenberg gave us a novel view of the issues surrounding the Enfield Arsenal.

Exercises

1. Explain why the endowment of wood affects the price of machinery used in producing rifles in the 19th century.

2. What could have caused the British to switch to mass-production techniques before 1854? Give a concrete example.

3. If the British had used the Blanchard lathe in 1820, then that would have been allocatively inefficient. Draw a graph that shows this and explain what it means.

4. Ames and Rosenberg (p. 836) include additional differences between America and Britain, such as the fact that the British consumer liked fancier gunstocks:

   > American machine processes could not produce guns of the kind favoured by English civilians. The Blanchard lathe produced stocks of a standard size, whereas English buyers did

> not want standard gunstocks. The English methods were suited to catering to the idiosyncratic needs of individual users.

How would this information change the comparison of the isoquant–isocost graph in the two countries?

References

The epigraph is on the description of *The Story of the Gun's* DVD set, available online and in many libraries. This entertaining video mixes the history of firearms with military history and technological change.

Ames and Rosenberg's article is an excellent example of economic history. The Cliometric Society is online at www.eh.net/Clio. In Greek mythology, Clio is the muse of history. Cliometricians use economic theory and econometrics to analyze economic history.

Denis Diderot's encyclopedia is available online. The Catholic Church banned it because it was too skeptical about Biblical miracles.

Adam Smith's discussion of the division of labor and his famous pin factory example is in the first book of *The Wealth of Nations*, at www.econlib.org/. Although it seems obvious today, specialization as a way to increase productivity was not so clear and Smith argued it drove the mighty economic engine he saw springing to life.

Appendix: Suggested Answers

STEP What are the tremendous advantages of interchangeable parts in a rifle (or anything else for that matter) for the end user?

Fixing broken rifles! You can quickly repair a mass-produced rifle if one of its pieces (lock, stock, or barrel) breaks. A rifle built by hand is useless once one of its individual parts fails. You would need a skilled craftsman to fix it.

On a battlefield, you could cobble together parts from different broken units to create operating weapons. And anyone could do this—they would not have to be a skilled craftsman.

In general, with precision parts, if the product breaks, you can buy a replacement part to repair the product. With bespoke items, you need an expert to adjust and refit to repair it.

STEP Is lack of knowledge about American technology a good answer? Why or why not?

This is a ridiculous explanation. Granted there is an ocean, but given the common language and communication, this answer makes no sense. In fact, there is lots of evidence that the British knew all about the American methods. They simply chose not to use them.

STEP Is managerial failure a good answer? Why or why not?

Like lack of knowledge, this is not a very satisfying answer. There is no reason to believe these specific people were especially poor managers. Economists are wary of this type of answer. Self-interested agents who respond to incentives are unlikely to make bad decisions, especially with great sums of money and lives at stake.

There is a subtle point to be made here that separates economists and non-economists. The latter are much more likely to accept mistake and stupidity to explain an observed decision or behavior that turned out badly. Economists tend to stick with rational, optimizing agents and explain bad choices as a result of lack of information or differing objectives.

STEP Draw graphs that show how the different resource endowments and input prices affected the optimal input mix. Use the detailed instructions that follow as a guide. How do the graphs answer explain why the British waited so long to adopt mass-production techniques?

The isoquant is exactly the same in each graph in Figure 11.6. US skilled labor wages were very high because there were few experienced craftsmen migrating to the United States, but lots of young, unskilled workers. The slope of each isocost is the input price ratio, $-\frac{w_{Unskilled}}{w_{Skilled}}$. Thus, the US isocost lines are flatter than Britain's. This leads to a different cost-minimizing input mix.

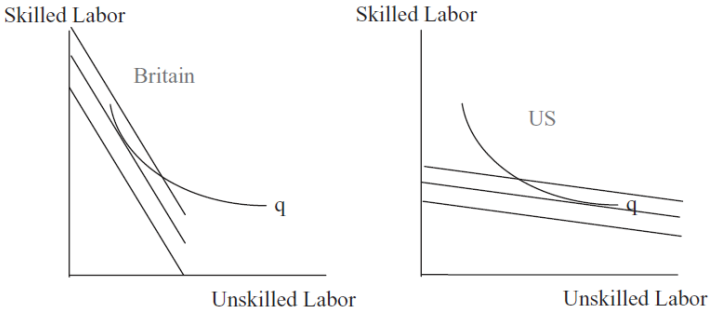


Figure 11.6: The effect of different wages for labor.

The price of machinery includes the cost of wood use just like a car's operating cost includes the cost of gasoline. The early versions of the Blanchard Lathe were quite wasteful, but this did not matter in the heavily forested United States. In Britain, however, wood was expensive. The British Isles were mostly deforested by then. This makes the isocost lines steeper in Figure 11.7 for Britain. Once again, factor prices help determine the input mix.

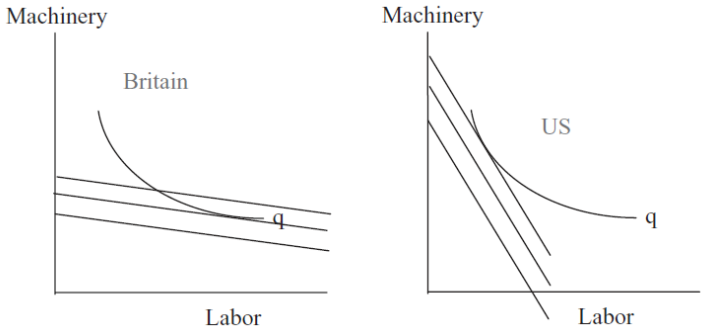


Figure 11.7: The effect of different prices for machinery.

So how do these graphs explain how economists view this historical episode? Varying resource endowments mean that each country faces its own set of input prices, which in turn lead to different cost-minimizing solutions. For the United States, unskilled labor with the Blanchard lathe was the cheapest way to make rifles. Not so for the British. At that time and place, with the skilled craftsmen and lack of cheap wood, rejecting mass production was the optimal decision.

In fact, the economic approach says something even more outlandish. Had the British used mass production for rifles before 1854, that would have been a mistake! Take the US tangency point and transfer it to the British graph in Figures 11.6 and 11.7. Producing with the US input mix is allocatively inefficient for Britain—that is, the British would not be minimizing cost.

Economists have no problem with agents making different choices. This does not mean that one is right and the other is wrong. All it means is that they face different prices. They are both optimizing. That is a difficult idea to wrap your head around. Ponder it.

STEP Is sudden awareness of new American technology a good answer? Why or why not?

This answer makes little sense. American and British people and entrepreneurs moved freely across the Atlantic and were well aware of production methods in each country. The claim that a new technique was suddenly made known to the British is absurd.

STEP Is managerial improvement a good answer for the switch? Why or why not?

This answer is pretty silly. To be credible, it requires an explanation for the sudden change from stupid, lazy, and careless producers of firearms to smart, energetic, and focused ones. There is no evidence of an explosion in managerial aptitude or a burst in managerial education. For this argument to be convincing, we will need a lot more evidence on British management prowess and how it changed over time.

STEP Answer these two questions:

1. What happened to the British labor force?

2. What happened to the Blanchard lathe?

The British labor force underwent a profound structural adjustment. The skilled craftsmen in the Birmingham gun trade died off and were not replaced. No skilled gunstock maker would suggest that his son follow him into the trade. They could see the writing on the wall—the machines were taking over. As the supply of these workers dwindled, the wages of skilled rifle artisans in Birmingham rose.

Perhaps more important is the second shock. The Blanchard lathe was continually improved over time; more modern versions of the lathe wasted a lot less wood. Today, a lathe uses a laser sight to precisely cut the wood. No human could possibly compete with it.

As the lathe wasted less wood, the operating cost of machinery fell. This is a nice example of how the price of an input can represent more than simply the out-of-pocket cost paid for the input. In this example, the price of a lathe is not simply the price paid for the machine itself; it includes the price of the wood used.

So, the shocks to the input cost minimization problem are that the skilled labor wage rose relative to the unskilled and r fell relative to w.

Notice how we first figure out what happened and then we model it. That is, we incorporate the story into one of the variables. In this case, the changing labor force increases the wage of skilled labor and the improving Blanchard lathe decreases r.

STEP Draw graphs that show how the changes mentioned affected the optimal input mix. How do the graphs explain why the British switched to mass-production techniques in 1854?

A high price of skilled labor makes the isocost lines flat (the slope falls in absolute value because the denominator increases). This leads to a more unskilled-labor intensive optimal input mix. As skilled craftsmen disap-

peared and their wages rose, there was greater incentive to use unskilled labor. Notice how the comparison in Figure 11.8 is across time periods.

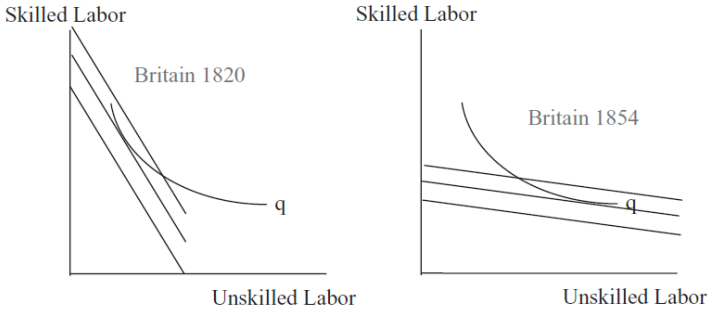


Figure 11.8: The effect of changes in the British labor force.

The price of machinery fell and fell as machines got better and better, making the isocost lines steeper and steeper (r is in the denominator) as shown in Figure 11.9, and leading to the adoption of mass-production techniques in Britain—the Enfield Arsenal was born.

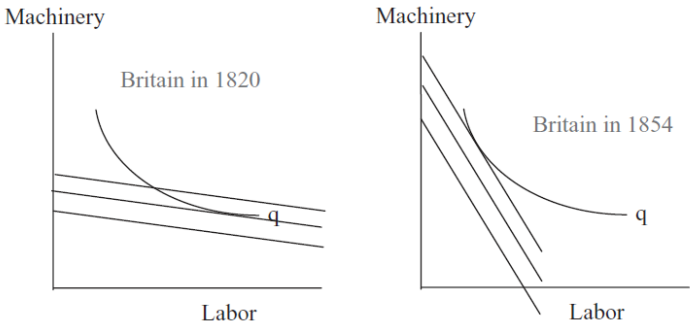


Figure 11.9: The effect of improvement in the Blanchard lathe.

Notice how the Britain in 1854 graphs in Figures 11.8 and 11.9 are the same as the US graphs in Figures 11.6 and 11.7. This shows that when Britain faced the same input prices as the United States, they made the same, optimal decisions.

There are reasons to hope that another type of production function, more diversified than Douglas's, may soon be available, and from these it would be possible to derive cost functions typical for particular industries.

Hans Staehle

11.3 Deriving the Cost Function

We have solved the input cost minimization problem so the next task is comparative statics analysis. We will focus on shocking q (the quantity the firm must produce) and track minimum total cost. The relationship between TC^* and q is called the *cost function*.

The novelty here is that we are not interested in how the optimal values of the endogenous variables, L and K vary as we shock q. Instead, we focus on the objective function, minimum total cost, and how it changes as q changes.

Another important aspect of comparative statics analysis for the input cost minimization problem is that, unlike utility in the Theory of Consumer Behavior, total cost can be cardinally measured. We can compare the total costs of different firms and perform arithmetic on total cost. If the minimum TC for $q = 10$ is \$40 and it rises to \$45 when $q = 11$, we can say TC increased by \$5. Because TC is cardinal, we will be able to interpret and use the Lagrangean multiplier.

As usual, we will explore both ways to do comparative statics:

- Numerical methods using a computer: Excel's Solver and the Comparative Statics Wizard.
- Analytical methods using algebra and calculus: conventional paper and pencil.

Numerical Methods to Derive the Cost Function

STEP Open the Excel workbook *DerivingCostFunction.xls*, read the *Intro* sheet, and proceed to the *OptimalChoice* sheet.

The organization is the same as in the *InputCostMin.xls* workbook. The cost-minimizing way of producing 100 units of output is to use about 183.3 hours of labor with 32.6 machines, which costs $464.38. There is no other combination of L and K that makes 100 units at a lower cost.

What happens if the firm needs to produce more, say, 110 units of output?

STEP Change cell B18 to 110.

The chart updates, showing a new (red) isoquant. The initial combination is not a viable option because it cannot produce 110 units. The firm has to re-optimize.

STEP Run Solver to find the new optimal solution.

The cost-minimizing amounts of labor and capital increase to produce the higher output required and the minimum total cost is now $513.39. We are looking for the minimum total cost. We want to know the cheapest way of producing any given output. This is called the *cost function*.

We can show the comparative statics analysis on the isoquant-isocost graph or on a presentation graph where we plot $TC^* = f(q)$, ceteris paribus. If we connected the points of tangency of isoquants and isocosts, we would get the *least cost expansion path* (LCEP).

Our work thus far has revealed two points on the LCEP and cost function: when $q = 100$, $TC = \$464.38$ and when $q = 110$, $TC = \$513.39$. Let's use the Comparative Statics Wizard to get more data so we can draw the LCEP and cost functions and understand how they are related.

STEP Return cell B18 to 100, then run the Comparative Statics Wizard, applying 10 q shocks in increments of 10.

The *CS1* sheet shows what your results should look like. The *CS1* sheet includes two graphs, the isoquant-isocost graph with the least cost expansion path and the cost function, as shown in Figure 11.10.

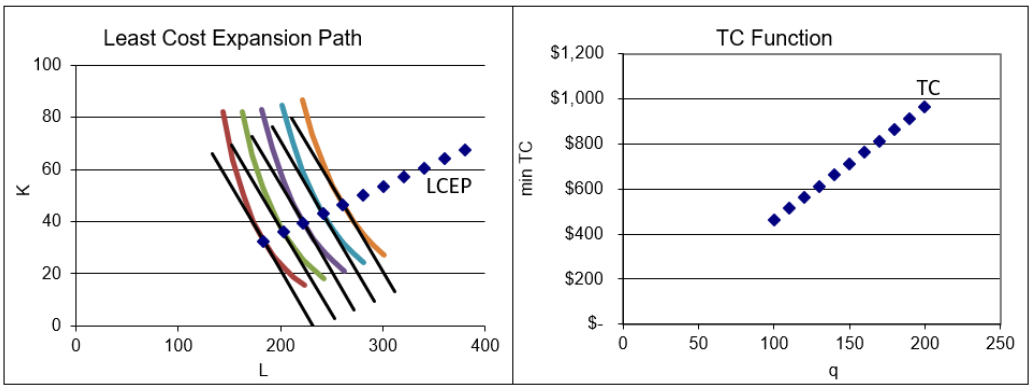


Figure 11.10: Deriving the cost function.
Source: DerivingCostFunction.xls!CS1

Figure 11.10 should remind of you of other graphs we have drawn, such as Engel and demand curves. On the left, using the display of the optimal solution to the input cost minimization problem, we show how different q produce a set of tangency points that comprise the LCEP.

On the right in Figure 11.10, we show only the minimum cost of producing each level of q, and hide everything else. This allows us to highlight the relationship between TC and q.

The two graphs in Figure 11.10 make clear that the source of the cost function is the optimal solution of the cost minimization problem as q varies. Just like demand curves do not come out of thin air, but are derived from utility maximization, cost functions are derived from input cost minimization.

We are interested in the shape of the cost function. It looks like a line, but is it really linear? To find out, we can see if it has a constant slope. If the slope is changing, we know the function is not linear.

STEP In your *CS* sheet, find the slope at different points on the function by computing the change in TC divided by the change in q.

Click the Unhide Column E button (near cell C9 in the *CS1* sheet) if you are stuck or to check your work. It is clear that the slope changes as output changes. This means that the cost function is nonlinear.

Analytical Methods to Derive the Cost Function

We can use the Lagrangean method to find $TC^* = f(q)$. We will leave q as a letter instead of a number so that the reduced-form solution will include q. Then we can plug in any value of q to find minimum cost for that q and easily draw a graph of the cost function.

The solution closely follows the work we did at the beginning of this chapter, but we proceed step-by-step to practice and reinforce the Lagrangean method.

The problem is

$$\begin{aligned} \min_{L,K} TC &= 2L + 3K \\ \text{s.t. } q &= L^{0.75}K^{0.2} \end{aligned}$$

The first step is to rewrite the constraint so that it is equal to zero.

$$q - L^{0.75}K^{0.2} = 0$$

The second step is to form the Lagrangean by adding lambda, λ, times the rewritten constraint to the original objective function. We use an extra-large $\mathcal{L}$ for the Lagrangean function that is not at all related to the L for labor.

$$\min_{L,K,\lambda} \mathcal{L} = 2L + 3K + \lambda(q - L^{0.75}K^{0.2})$$

The third step to finding the optimal solution is to take the derivative of the Lagrangean with respect to each endogenous variable and set each derivative to zero (giving us the first-order conditions).

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial L} &= 2 - 0.75\lambda L^{-0.25}K^{0.2} = 0 \\ \frac{\partial \mathcal{L}}{\partial K} &= 3 - 0.2\lambda L^{0.75}K^{-0.8} = 0 \\ \frac{\partial \mathcal{L}}{\partial \lambda} &= q - L^{0.75}K^{0.2} = 0 \end{aligned}$$

The fourth, and last, step is to solve this system of equations for L^*, K^*, and λ^*. We move the terms with lambda in the first two equations to the right-hand side and then divide the first equation by the second. The exponents

cancel nicely (see section 11.1) and we get $L = 5.625K$. This is not a reduced-form solution because L is not a function of exogenous variables alone. We substitute this expression for L into the third first-order condition to get optimal K and then optimal L as shown below.

$$
\begin{aligned}
&q - [5.625K]^{0.75} K^{0.2} = 0 \\
&q = 3.6525 K^{0.75} K^{0.2} \\
&\frac{q}{3.6525} = K^{0.95} \\
&\left[\frac{q}{3.6525}\right]^{\frac{1}{0.95}} = (K^{0.95})^{\frac{1}{0.95}} \\
&K^* = 0.25574 q^{\frac{1}{0.95}} \Rightarrow L^* = 1.43854 q^{\frac{1}{0.95}}
\end{aligned}
$$

Finally, we substitute the optimal solutions for L^* and K^* into the original objective function.

$$
\begin{aligned}
&TC = wL + rK \\
&TC^* = 2\left[1.43854 q^{\frac{1}{0.95}}\right] + 3\left[0.25574 q^{\frac{1}{0.95}}\right] \\
&TC^* = 2.877 q^{\frac{1}{0.95}} + 0.767 q^{\frac{1}{0.95}} \\
&TC^* = 3.644 q^{\frac{1}{0.95}}
\end{aligned}
$$

This expression is the total cost function. It gives the cheapest cost of producing any given amount of output. If $q = 100$, $TC = \$464.38$. Not surprisingly, this agrees with our results using numerical methods.

Notice also that the cost function is clearly nonlinear. It is increasing at an increasing rate because the exponent on q is greater than one ($\frac{1}{0.95} \approx 1.05$). The derivative of TC with respect to q, the slope, is not constant because it depends on q. If the exponent was exactly 1, then the slope would be constant and the TC would be a line. The fact that this exponent is only slightly greater than one explains why TC looks almost linear in Figure 11.10.

Interpreting Points Off the Cost Function

When we derived the demand curve from the "maximize utility subject to a budget constraint" optimization problem, we explored what it meant to be off the demand curve (see Figure 4.12). We learned that points to the left or right of the inverse demand curve (with price on the y axis) mean that

the consumer is not optimizing, i.e., the consumer is not choosing a point of tangency between the indifference curve and budget constraint.

We can conduct the same kind of inquiry here, asking this question: What does it mean to be off the cost function?

Unlike the inverse demand curve, where the exogenous variable is on the y axis, the cost function is graphed according to usual mathematical convention, with the exogenous variable, output, on the x axis. Thus, points off the curve are interpreted vertically above or below the cost function.

What does it mean if a point is above the cost curve? Figure 11.11 helps us answer this question. On the left is the familiar isoquant/isocost graph. The cheapest way to produce *q0* units of output is with the L and K combination at the point labeled TC^*. The graph on the right of Figure 11.11 shows that TC^* is the point on the cost function at an output of *q0*.

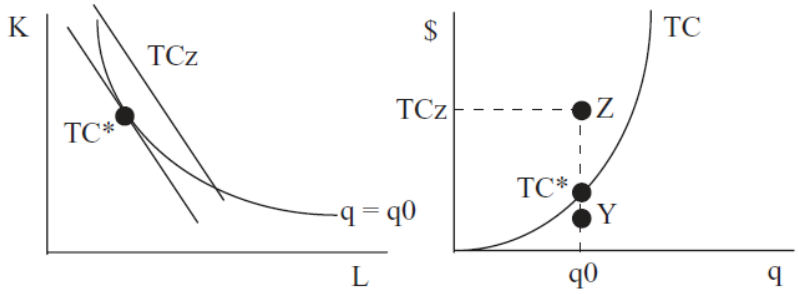


Figure 11.11: Understanding points off the cost function.

Point Z, a point above the cost function, reveals that the firm is producing the level of output *q0* at a total cost *above* the minimum total cost. This means that the firm is choosing an input mix that is not cost minimizing. Point Z on the graph on the left of Figure 11.11 must lie on an isocost above the tangent isocost. We do not know exactly where point Z is on the graph on the left (so we do not know if there is technical or allocative inefficiency), but we do know it has to be somewhere on the isocost labeled *TCz* that has a total cost the same as the cost of producing point Z (on the graph on the right).

Point Y on the right side of Figure 11.11 is below the cost function. How can this point be generated by the graph on the left? It cannot. There is an isocost with a total cost equal to that at point Y, but it is below the iso-

quant and, therefore, unattainable. In other words, point Y does not actually exist. The firm cannot produce *q0* units of output at any cost less than TC^*.

Another way of thinking about TC geometrically, is that there are points above TC, but only empty space below it. Sure, on a printed page, chalkboard, or computer screen, there is white space above and below TC and you can write on it (just like point Y in Figure 11.11), but this is misleading. In fact, below TC there is nothing, total void. If you tried to put a point there, your hand would go through the paper!

This has implications beyond pure theory. The fact that there are no points below the cost function means that we should never fit a line through a cloud of points to estimate a cost function. Instead of a least squares approach to estimating a cost function, estimation techniques in the stochastic frontier literature are based on fitting a curve around the observed points, as in Figure 11.12.

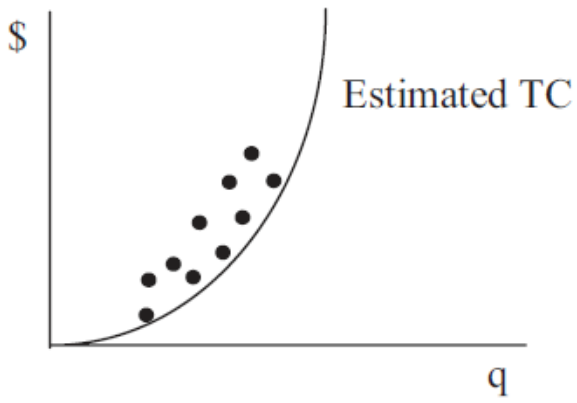


Figure 11.12: Estimating a cost function.

Shifts in the Cost Function

You learned in Introductory Economics that price causes a movement along a demand curve, but other shocks (like increasing income) change demand causing the entire curve to shift. The same thing happens with the cost function. Changing q leads to moving along the TC function, but other exogenous variables cause shifts in the cost function.

STEP Proceed to the *CostFn* sheet.

The sheet displays a cost function charted from the data above it. The data in columns L and M are actually formulas for the reduced-form expressions for L^* and K^*. Column N has the minimum total cost for the benchmark problem and will not change because the cells are merely numbers (so it is labeled "Dead (Initial)"). Column O, however, has the reduced-form expression for TC^* and will update if any of the underlying parameters are changed (hence the "Live" label).

STEP Click on a few cells in columns L, M, N, and O to see the formulas and values.

The general versions of the reduced forms for the Cobb-Douglas production functions are provided and entered in cells. The expressions look daunting (and they are tedious to derive), but the derivation is straightforward: leave every exogenous variable as a letter and find the optimal solution for L,K, λ, and total cost.

Initially, N and O are the same because the exogenous variable values have not been changed yet. Let's do that now.

STEP Change cell B20, the exponent on L, to 0.8.

Your screen looks like Figure 11.13. The increase in labor productivity has shifted down the total cost curve. This makes sense. The increase in c has made it cheaper to produce any given output.

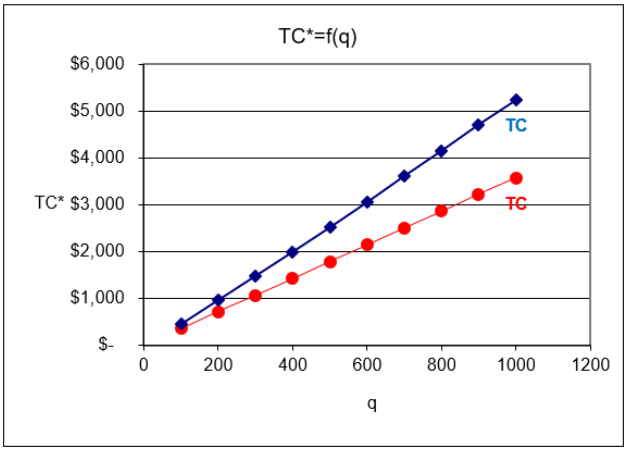


Figure 11.13: Total cost shifts down when labor productivity rises.
Source: DerivingCostFunction.xls!CostFn

You can experiment with other shocks to the cost function. Change input prices, input exponents, or A to see how the cost function shifts. Click the Reset button or ctrl-z (undo) after every trial. Connect what you see on the screen with the shock you applied. Changes in q have no visible effect because you simply move along the cost function.

Interpreting λ*

We end this chapter by showing that the Lagragean multiplier, λ^* has a useful interpretation in the input cost minimization problem. We will see that λ^* gives an easier way to derive a cost function than solving the constrained cost minimization problem with q as a letter and finding $TC^* = f(q)$.

The cost function shortcut uses the fact that λ^* gives the instantaneous rate of change in the optimum value of the objective function as the constraint varies. Thus, λ^* signals how relaxing the constraint would impact the goal.

For utility maximization, we could relax the constraint by increasing income. The budget constraint in the Lagrangean is $m - p_1x_1 - p_2x_2 = 0$ so as m rises, the consumer will be able to reach greater maximum utility. The Lagrangean multiplier tells us how much more utility is gained as income increases. Unfortunately, utility is ordinal so λ^* does not have a useful interpretation in the Theory of Consumer Behavior.

Things are different in the constrained input cost minimization problem. The objective function in this case is minimum total cost and is measured on a cardinal scale. We can directly observe minimum total cost and meaningfully compare how it changes within a firm and across firms. This means we can apply the interpretation of λ^* to input cost minimization.

The constraint in the Lagrangean is $q - f(L, K)$. If we vary the constraint by having the firm produce one more unit of output, we know total cost would rise as we moved to a higher isoquant. The value of λ^* tells us by how much minimum total cost would rise.

For example, at $q = 100$ in *DerivingCostFunction.xls*, λ^* is about \$4.89. You can confirm this by numerical methods (using Excel's Solver and getting the Sensitivity Report) or by analytical methods, solving for λ^* from the three first-order conditions. Either way, you will get (almost exactly) the same answer.

But what does this tell us? The \$4.89 value means that if we increase output by an infinitesimally small amount, minimum total cost will go up by \$4.89-fold. Let's use Excel to work on this.

STEP Click the Reset button in the *CostFn* sheet and take a look at the highlighted cell with a yellow background (P8). Click on it and read the formula.

The value of P8 is \$4.99. That is close to the value of λ^* of \$4.89, but not quite exactly the same. What is going on?

STEP Go to the *CS1* sheet and take a look at the highlighted, yellow-backgrounded cell (E15) (click the Unhide Column E button if needed). Its value is \$4.90.

This is much closer to λ^*'s value of \$4.89. Why? Because the change in q is much smaller in the *CS1* sheet than in the *CostFn* sheet. As the change in q approaches zero, the change in TC^* divided by the change in q will approach λ^*.

STEP Return to the *CostFn* sheet and change cell K8 from 200 to 110. This replicates the *CS1* sheet value for λ^*. Next, set K8 to 101. What do you see?

With K8 set to 101 so that $\Delta q = 1$, $\frac{\Delta TC}{\Delta q} = \$4.89$, the value of λ^*. Well, actually, not exactly \$4.89. If we displayed more decimal places in P8 and computed the value of λ^* to more decimal places, the two would not agree. But they would get closer the smaller we made Δq.

Of course, this is nothing more than a demonstration of the idea of the derivative. If you are puzzled as to how $\frac{\Delta TC}{\Delta q}$ can be that close to λ^* in the *CS1* sheet (a one cent difference seems pretty small), given that the change in q is 10 units (which is hardly infinitesimally small), the answer lies in the total cost function: It simply is not very curvy. Because TC^* follows almost (but not quite) a straight line, computing the slope from $q = 100$ to $q = 110$ is close to the slope of the tangent line at $q = 100$.

The purpose of the work above was to convince you that $\lambda^* = \frac{dTC}{dq}$. The Lagrangean multiplier gives the instantaneous rate of change in minimum total cost with respect to output.

STEP You can confirm the claim that $\lambda^* = \frac{dTC}{dq}$ by changing the parameters in the *CostFn* sheet and keeping your eye on the rose-backgrounded cell H31. It computes the difference between λ^* in H13 and $\frac{dTC}{dq}$ in H30. The difference is always zero because these two things, λ^* and $\frac{dTC}{dq}$ are equivalent.

You might ask, "So what?" In other words, what can we do with the knowledge that $\lambda^* = \frac{dTC}{dq}$? A lot. For one thing, we can easily derive the cost function. After all, the rate of change in total cost as output changes is *marginal cost* (MC). Thus, $\lambda^* = \frac{dTC}{dq} = MC(q)$. This means we can easily get the total cost function by simply integrating λ^* with respect to q.

Furthermore, as we will see when we solve the output profit maximization problem, we usually want marginal revenue and marginal cost, so knowing that $\lambda^* = \frac{dTC}{dq}$ can be a real shortcut. If we have λ^*, then we do not have to derive $TC^* = f(q)$ and then take the derivative to get MC.

The Cost Function has Parents

This section included some complicated ideas, but we end by prioritizing things. There is no doubt that the most important idea is that the cost function has a source and does not appear from nowhere. This is captured by Figure 11.10—the cost function is derived by doing comparative statics analysis on the input cost minimization problem.

Although we are often interested in the response of an endogenous variable to a shock, comparative statics in the input cost minimization problem is focused on how the objective function, minimum total cost, is affected by shocking q. Minimum total cost as a function of q is the cost function.

By explaining what it means to be above or below the cost function in terms of the isoquant–isocost graph, we emphasized the idea that the cost function shows the cheapest way to produce any given output. A good way to remember this is to ponder the striking fact that there is no space below the cost function, meaning that it is impossible to produce the given output any cheaper than the cheapest way possible.

Changes in other parameters besides output cause the entire cost function to shift because minimum total cost depends on all of the exogenous variables. If q changes, we move along the cost function; other shocks shift TC.

Finally, we explained a mathematically sophisticated idea: λ^* provides information on the rate of change of the optimum value of the objective function as the constraint is relaxed. This interpretation of the Lagrangean multiplier holds for every constrained optimization problem.

We did not apply this interpretation in the Theory of Consumer Behavior because utility (the objective function) cannot be cardinally measured. In the old days, when utility was believed to be cardinally measured in utils, λ^* was the marginal utility of money. λ^* would tell you the rate of change in maximum utility if you gave the consumer an infinitesimal increase in income.

Since total cost is directly observable and countable, λ^* can be correctly interpreted as marginal cost, $\frac{dTC}{dq}$. This gives a shortcut to the cost function and MC.

Exercises

1. With the production function, $q = L^{0.75}K^{0.5}$, and exogenous variables $w = 2$, $r = 3$, use Excel to create a graph of the cost function for the same q values as the one in the *CS1* sheet. Copy and paste your graph in a Word document.

2. How is the cost function you just derived different from the one in the *CS1* sheet? Which variable is responsible for generating this difference?

3. From the cost functions in the *CS1* sheet and question 1, what can you deduce about cost functions derived from Cobb-Douglas production functions?

4. If someone solves an input cost minimization problem and finds that $\lambda^* = 50$, what does this mean?

References

The epigraph is from page 333 of Hans Staehle, "The Measurement of Statistical Cost Functions: An Appraisal of Some Recent Contributions," *The American Economic Review*, Vol. 32, No. 2, Part 1 (June, 1942), pp. 321–333, www.jstor.org/stable/1803513. Staehle was optimistic in 1942 that

advances in statistics and data collection would enable economists to estimate cost functions for particular industries. Unfortunately, it is fair to say that Staehle's dream of the discovery of flexible functional forms remains unfulfilled. Empirical work on cost functions usually finds that firms face linear (or nearly linear) total costs (yielding horizontal average and marginal costs) over large ranges of output.

Only 11 percent of firms report that their MC curves are rising. By contrast, about 40 percent claim that their MC curves are falling.

Alan Blinder, Elie Canetti, David Lebow, and Jeremy Rudd

11.4 Cost Curves

In the next chapter, we will work on the firm's second optimization problem: maximize profits by choosing the amount of output to produce. Because profits are revenues minus costs, the cost function plays an important role in the firm's profit maximization problem.

This section is devoted to the terminology of cost curves and an exploration of their geometric properties. Derived from the cost function, a variety of cost curves are used to solve and display the firm's profit-maximization problem. This section defines and derives them.

A basic idea that is easy to forget is that there are many shapes of cost functions. Our work on deriving the cost function used a Cobb-Douglas production function and that gives rise to a particularly shaped cost function. A different production function would give a different cost function. A key idea is that $q = f(L, K)$ determines the shape of $TC^* = f(q)$.

Names and Acronyms

You know that if we track TC^*, minimum total cost, as a function of q, we derive the cost function. Since we will be using other measures of costs, to avoid confusion, we refer to the cost function as the *total cost* (TC) function. The total cost function has units of dollars (\$) on the y axis. We can divide total costs into two parts, *total variable costs*, TVC, and *total fixed costs*, TFC.

$$TC(q) = TVC(q) + TFC$$

If the firm is in the short run, it has at least one fixed factor of production (usually K) and the total fixed costs are the dollar value spent on the fixed inputs (rK). Notice that the total fixed costs do not vary with output. TFC is a constant and does not change as output changes so there is no "(q)" in the TFC function like there is on TVC and TC.

The total variable costs are the costs of the factors that the firm is free to adjust or vary (hence the name "variable costs"), usually L. As output rises, firms need more inputs to produce the increased output so total variable costs rise.

In the long run, defined as a planning horizon in which there are no fixed factors, there are no fixed costs ($TFC = 0$) and, therefore, $TC(q) = TVC(q)$. In other words, the total cost and total variable cost functions are identical.

In addition to total costs, the firm has average, or per unit, costs associated with each level of output. *Average total cost*, ATC (also known as AC), is the total cost divided by the output level.

$$ATC(q) = \frac{TC(q)}{q}$$

Average variable cost, AVC, is total variable cost divided by output.

$$AVC(q) = \frac{TVC(q)}{q}$$

Average fixed cost, AFC, is total fixed cost divided by output.

$$AFC(q) = \frac{TFC(q)}{q}$$

Notice that $AFC(q)$ is a function of q even though TFC is not because AFC is TFC divided by q. Since the numerator is a constant, $AFC(q)$ is a rectangular hyperbola ($y = c/x$) and is guaranteed to fall as q rises. This can be confirmed by a simple example. Say TFC = \$100. For very small q, such as 0.0001, AFC is extremely large. But AFC falls really fast as q rises from zero (and AFC is undefined at $q = 0$). At $q = 1$, AFC is \$100, at $q = 2$, AFC is \$50, and so forth. The larger the value of q, the closer AFC gets to zero (i.e., it approaches the x axis).

It is easy to show that the average total cost must equal the sum of the average variable and average fixed costs:

$$TC(q) = TVC(q) + TFC$$

$$\frac{TC(q)}{q} = \frac{TVC(q)}{q} + \frac{TFC}{q}$$

$$ATC(q) = AVC(q) + AFC(q)$$

We often omit *AFC(q)* from the graphical display of the firm's cost structure (see Figure 11.14) because we know that $AFC(q) = ATC(q) - AVC(q)$. Thus, average fixed cost can be easily determined by simply measuring the vertical distance between ATC and AVC at a given q.

The facts that $AFC(q) = ATC(q) - AVC(q)$ and AFC goes to zero as q rises means that AVC must approach ATC as q rises. Always draw AVC getting closer to ATC as q increases past minimum AVC. Figure 11.14 obeys this condition.

Unlike the total curves, which share the same y axis units of dollars, the average costs are a rate, dollars per unit of output. You cannot plot total and average cost curves on the same graph because the y axes are different.

Another cost concept that we get from the total cost function is *marginal cost (MC)*. Like average costs, MC is a rate and it comes in \$/unit. Marginal cost is often graphed together with the average curves (as shown in Figure 11.14).

Marginal means *additional* in economics. Marginal cost tells you the additional cost of producing more output. If the change in output is discrete, then we are measuring marginal cost from one point to another on the cost curve and the equation looks like this:

$$MC(q) = \frac{\Delta TC(q)}{\Delta q}$$

If, on the other hand, we treat the change in output as infinitesimally small, then we use the derivative and we have:

$$MC(q) = \frac{dTC(q)}{dq}$$

Because TFC does not vary with q, marginal cost also can be found by taking the derivative of *TVC(q)* with respect to q.

Average cost and marginal cost are used to refer to entire functions (see Figure 11.14), but also to specific values. For example, if $ATC = \$10$/unit and $MC = \$3$/unit at $q = 5$, this means that it costs \$10 per unit to make the five units and, thus, the firm had \$50 of total costs to make five units. The MC tells us that the 5^{th} unit costs an additional \$3 so the total cost went from \$47 for 4 units to \$50 for 5 units.

The Geometry of Cost Curves

The average and marginal curves are connected to each other and must be drawn according to strict requirements. Whenever a marginal curve is above an average curve, the average curve must be rising. Conversely, whenever a marginal is below an average, the average must be falling.

For example, consider the average score on an exam. After the first 10 students are graded, there is an average score. The 11th student is now graded. Suppose she gets a score above average. Hers is the marginal score and we know it is above the average so it has to pull the average up. Suppose the next student did poorly. His marginal score is below the average and it pulls the average down. So, we know that whenever a marginal score is below the average, the average must be falling and whenever a marginal score is above the average, the average must be rising. The only time the average stays the same is when the marginal score is exactly equal to the average score.

This relationship between the average and marginal means that the marginal cost curve must intersect the average variable and average total cost curves at their respective minimums, as shown in Figure 11.14. From $q = 0$ to the intersection of MC with ATC, MC is below the ATC and the ATC falls. To the right of the intersection of MC with ATC, MC is above the ATC so the ATC is pulled up. MC and AVC curves share the same relationship.

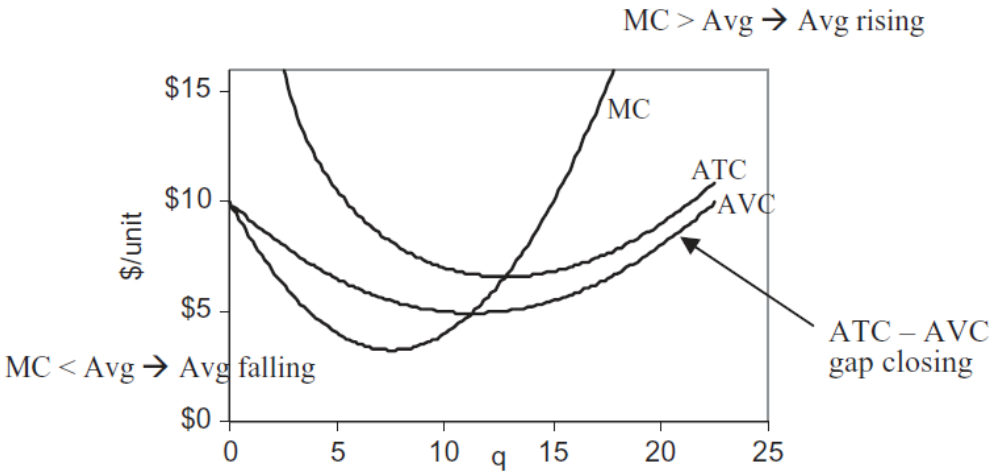


Figure 11.14: Marginal and average relationships.

Figure 11.14 also shows a property that was highlighted earlier: The gap between ATC and AVC must fall as q rises.

You will understand these abstract ideas better by exploring concrete examples. Three cost functional forms will be examined:

1. Cobb-Douglas Cost Curves
2. Canonical Cost Curves
3. Quadratic Cost Curves

Instead of memorizing specific facts or points, look for the pattern and repeated connections. Focus on the relationship between the total and average and marginal curves.

STEP Open the Excel workbook *CostCurves.xls* and read the *Intro* sheet, then go to the *CobbDouglas* sheet to see the first example.

1. Cobb-Douglas Cost Curves

The *CobbDouglas* sheet is the *CostFn* sheet from the *DerivingCostFunction.xls* workbook with the ATC and MC curves plotted below the TC curve. Column I has a formula for the TC curve using L^* and K^*, from which we can compute ATC and MC in columns J and K. Click on an MC cell, for example, cell K4, to see that the cell formula is actually for λ^*. We are using the shortcut that $\lambda^* = MC$.

With L and K both endogenous, there are no fixed factors of production. This means we are in the long run and there are no fixed costs. Thus, $TC = TVC$ and $ATC = AVC$.

It is immediately obvious that the marginal and average curves do not look at all like the conventional family of cost curves as shown in Figure 11.14. In fact, a Cobb-Douglas production function cannot give U-shaped average and marginal cost curves as in Figure 11.14.

Remember that there are many functional forms for cost curves (total, average, and marginal) and the shape depends on the production function. In other words, the production function is expressed in the cost structure of a firm.

STEP Set the exponent on capital, d, to 2 to replicate Figure 11.15.

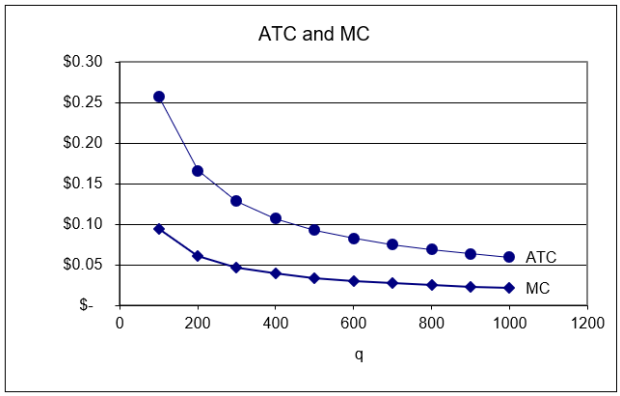


Figure 11.15: Total cost shifts down when labor productivity rises. *Source: CostCurves.xls!CostFn*, after setting $d = 2$.

Because average cost is falling as q rises in Figure 11.15 (and your computer screen), it means that total cost is increasing less than linearly as output rises. The total cost graph on your screen confirms that this is the case. It costs \$33 to make 200 units, but only \$43 to make 400 units. Double output again to 800. How much does it cost? Cell I9 tells you, \$55. This is puzzling. If input prices remain constant, how can we double output and not at least double costs?

The answer lies in the production function. You changed the exponent on capital, d, from 0.2 to 2. Now the sum of the exponents, $c + d$, is greater than 1. For the Cobb-Douglas production function, this means that we are operating under increasing returns to scale. This means that if we double the inputs, we get more than double the output. Or, put another way, we can double the output by using less than double the inputs.

This firm can make 400 units cheaper *per unit* than 200 units. It can make 800 units even cheaper *per unit* because it is taking advantage of the increasing returns to scale.

Increasing returns are a big problem in the eyes of some economists because they lead to a paradox: One firm should make all of the output. There are situations in which increasing returns seem to be justified, such as the case of *natural monopolies*, in which a single firm provides the output for an entire industry because the production function exhibits increasing returns to scale.

The classic examples are utility companies, e.g., electric, water, and natural gas companies. Often, these firms are nationalized or heavily regulated.

We can emphasize the crucial connection between the production function and the cost function via the isoquant map.

STEP Scroll down to row 100 or so in the *CobbDouglas* sheet.

The three isoquants are based on a Cobb-Douglas production function with parameter values from the top of the sheet, except for d, which can be manipulated from the *Set d* radio buttons (above the chart). The three red points are the cost-minimizing input combinations for three different output levels: 100, 120, and 140.

Above the graph, the value of the sum of the exponents, initially 0.95, is displayed. A description of the shape of the total cost function, which depends on the value of $c + d$, and a small picture of that shape is shown. Figure 11.16 has the initial display.

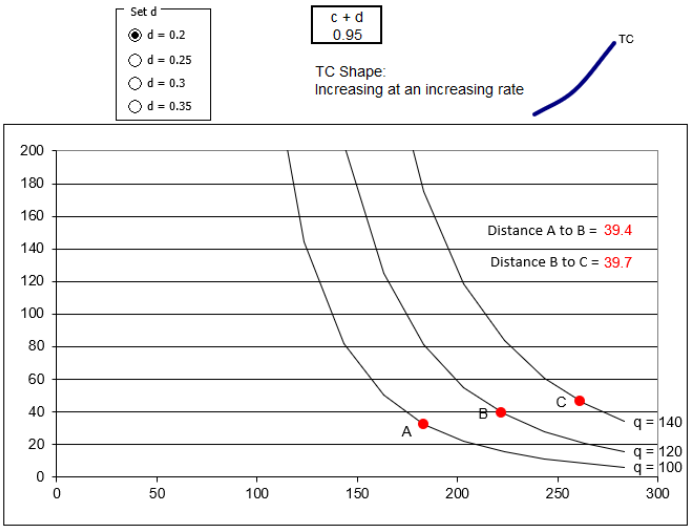


Figure 11.16: Isoquants determine the shape of the cost function.
Source: CostCurves.xls!CobbDouglas.

The spacing between the points is critical. The distance from A to B is a little less than that from B to C. This means that as output is increased from 120 to 140, the firm needs a bigger increase in inputs than when q rose from 100 to 120.

As output continues rising by 20 units, the next isoquant we have to reach is getting farther and farther away, requiring progressively more inputs, and progressively higher costs. This is why TC is increasing at an increasing rate.

STEP Click on the $d = 0.25$ option.

The isoquants shift in because it takes fewer inputs to make the three levels of output depicted. The distance between the isoquants has decreased and TC is linear. Most importantly, the distance between the points is identical.

With $c + d = 1$, the spacing of the isoquants is constant. As q increases by 20, the next isoquant is the same distance away and the firm increases its input use and costs by a constant amount. This is why the TC function is a line, increasing at a constant rate.

STEP Click on the $d = 0.3$ option.

Once again, the chart refreshes and isoquants shift in. Now the distance between the isoquants is decreasing. As q rises, the isoquants get closer together and the total cost function is increasing at a decreasing rate.

STEP Click on the $d = 0.35$ option.

This produces even stronger increasing returns and a TC function that bends faster than $d = 0.3$.

The fundamental point is that the distance between the isoquants reflects the production function. There are three cases:

1. If the distance is increasing as constant increases in quantity are applied, the total cost function will increase at an increasing rate.

2. If the distance remains constant, the cost function will be linear.

3. If the distance get smaller as output rises, the firm has costs that rise at a decreasing rate.

This holds for all production functions and, in the case of Cobb-Douglas, it is easy to see what is going on because the value of $c + d$ immediately reveals the returns to scale and spacing between the isoquants.

But the advantage of Cobb-Douglas in easily displaying the three cases (depending on the value of $C + d$) means it cannot do all three cases at once. A Cobb-Douglas production function can generate a TC functon that is increasing at an increasing or constant or decreasing rate, but not all three.

The shape of the cost function is dependent on the production technology. Repeatedly cycle through the radio buttons, keeping your eye on the isoquants, the distance between the points, and the resulting total cost function. Your task is to understand and cement the relationship between the production and cost functions.

An accordion is a good metaphor for what is going on. When scrunched up, the isoquants are being squeezed together, which gives increasing returns to scale and TC increasing at a decreasing rate. When the accordion is expanded and the isoquants are far apart, we have decreasing returns to scale and TC rising at an increasing rate.

Do not be confused. The reason why increasing (decreasing) returns to scale leads to TC rising at a decreasing (increasing) rate (they are opposite) is that productivity (returns to scale) and costs are opposites. Increased productivity enables slower increases in costs of production. Production increasing at an increasing rate and costs increasing at a decreasing rate are two sides of the same coin.

2. Canonical Cost Curves

STEP Proceed to the *Cubic* sheet.

This sheet displays the canonical cost structure, in other words, the most commonly used cost function. It produces the familiar U-shaped family of average and marginal costs (which Cobb-Douglas cannot).

The canonical cost curves graph can be generated by a cost function with a cubic polynomial functional form.

$$TC(q) = aq^3 + bq^2 + cq + d$$

The d coefficient (not to be confused with the d exponent in the Cobb-Douglas production function) represents the fixed cost. If $d > 0$, then there are fixed costs and we know the firm is in the short run.

Once we have the cost function, the top curve on the top graph in the *Cubic* sheet, we can apply the cost definitions (from the beginning of this section) to get all of the other cost curves. The other total curves are:

$$TVC(q) = aq^3 + bq^2 + cq$$

$$TFC = d$$

STEP Click on each of the three curves in the top graph of the *Cubic* sheet to see the data that are being plotted.

Now turn your attention to the bottom graph. The curves in the bottom graph are all derived from the top graph. Notice that the y axis label is different, the totals in the top have units of \$, while the average and marginal curves have a y scale of \$/unit (of output).

STEP Click on each of the three curves in the bottom graph to see the data that are being plotted.

Custom formatting has been applied to the numbers in the average and marginal cost cells to display "\$/unit" in each cell. It is easy to forget that "\$" is not the units of average and marginal cost curves.

The average total and average variable costs are easy to compute: simply divide the total by q. You can confirm that column E's formula does exactly this. There is no ATC value for $q = 0$ because dividing by zero is undefined.

We can also divide the equation itself by q to get an average. This is done for *AVC*. The formula in cell F2 is "= a_*(A6^2) + b_*A6 + c_" because dividing $TVC(q) = aq^3 + bq^2 + cq$ by q yields $= aq^2 + bq + c$. Notice that *AVC* for $q = 0$ does exist.

Marginal cost is more difficult to understand than average cost. Marginal cost is defined as the additional cost of producing more output. "More" can be an arbitrary, finite amount (such as 1 unit or 10 units) or an infinitesimally small change in the number of units.

If we use an arbitrary, finite amount of increase in q, then we compute MC as $\frac{\Delta TC}{\Delta q}$. We can also compute MC for an infinitesimally small change, using the derivative, $\frac{dTC}{dq}$. These two computations will be exactly the same only if *MC* is a line.

The two approaches are applied in columns G and H. The derivative of TC with respect to q is:

$$TC(q) = aq^3 + bq^2 + cq + d$$
$$\frac{dTC}{dq} = 3aq^2 + 2bq + c$$

Notice how we apply the usual derivative rule, bringing the exponent down and subtracting one from the exponent for each term. The d coefficient, TFC, disappears because it does not have q in it (or, if you prefer, think of d as dq^0). The expression for MC is entered in column G.

Column H has MC for a discrete-size change. You can vary the size of the change in q by adjusting the step size in cell B3.

STEP Make the step size smaller and smaller. Try 0.1, 0.01, and 0.001.

As you make the step size smaller, the values in column H get closer to those in column G. This, once again, demonstrates the concept of the derivative.

Another way to get the cost function is to use the neat result from Lagrangean method. We can simply use $\lambda^* = MC$ and we have the MC curve. No delta-size change or derivative required. If what we really wanted was the total cost function, then we would have to integrate the λ^* function with respect to q. The constant of integration is the fixed cost, which would be zero in the long run.

The family of cost curves in the *Intro* and *Cubic* sheets (and in Figure 11.14) are the canonical cost curves displayed in countless economics textbooks. You might wonder, if not Cobb-Douglas, then what production function could produce such a cost function? That is not an easy question to answer. In fact, the functional form for technology that would give rise to the canonical cost curves is quite complicated and it is not worth the effort to painstakingly derive the usual U-shaped average and marginal cost curves from first principles.

It is sufficient to know that a production function underlies the polynomial TC function and its resulting U-shaped average and marginal cost curves. We also want to keep in mind that if input prices rise, the cost curves shift up and, if technology improves, they shift down.

3. Quadratic Cost Curves

STEP Proceed to the *Quadratic* sheet to see a final example of cost curves.

It is immediately clear that the quadratic functional form is a special case of the cubic cost function, with coefficients a and c equal to zero.

Look at the top chart and connect the shapes of the TC, TVC, and TFC functions to the functional form $TC(q) = bq^2 + d$. Given the coefficient values in the sheet, this gives $TC(q) = q^2 + 1$, $TVC(q) = q^2$, and $TFC = 1$.

The bottom chart does not look familiar, but it obeys the definitions of average and marginal cost explained earlier in this section. ATC is $TC(q)$ divided by q: $ATC(q) = q + \frac{1}{q}$. Similarly, AVC is $TVC(q)/q$, which is q (a ray out of the origin). MC is the derivative of TC with respect to q, which is $2q$.

Although not the usual U-shaped curves, the MC curve (actually, MC is linear) intersects AVC and ATC at their minimums. When MC is below ATC, ATC is falling, but beyond the point at which MC intersects ATC (at the minimum ATC), MC is above ATC and ATC is rising. As q increases, AVC converges to ATC, which implies that AFC goes to zero.

The shapes of the cost curves are not the usual U-shaped average and marginal curves, but this is another of the many possible cost structures that could be derived from a firm's input cost minimization problem.

The Role of Cost Curves in the Theory of the Firm

Cost curves are not particularly exciting, but they are an important geometric tool. When combined with a firm's revenue structure, the family of cost curves is used to find the profit-maximizing level of output and maximum profits.

Cost curves can come in many forms and shapes, but they all share the basic idea that they are derived by minimizing the total cost of producing output, where output is generated by the firm's production function. Different production functions give rise to different cost functions.

The shape of the cost function, rising at an increasing, constant, or decreasing rate, is determined by the production function. With increasing returns to scale, for example, a firm can more than double output when it doubles its input use. That means, on the cost side, that doubling output will less than double total cost. Returns to scale can be spotted by the spacing between the isoquants. With increasing returns to scale, for example, the gaps between the isoquants get smaller as output rises.

No matter the production function, it is always true that for output levels at which marginal cost is below an average cost, the average must be falling and MC above AVC or ATC means AVC or ATC is rising. It is also true that, in the short run (when there are fixed costs), AVC approaches ATC as output rises.

Lastly, consider the message conveyed by Figure 11.17. The arrows show the progression—average and marginal curves come from the total cost function, which comes from the input cost minimization problem (with the production function expressed in the isoquants).

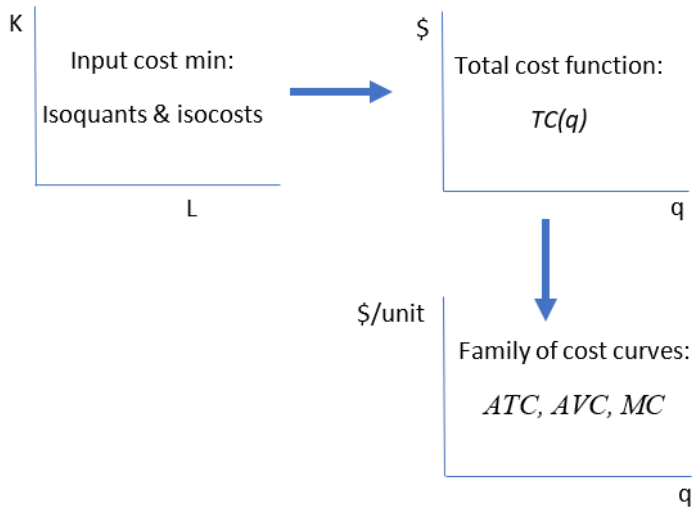


Figure 11.17: Connecting cost graphs.

Economists use graphs to communicate. It may seem like graphs are conjured out of thin air, but this is false. All graphs have a genealogy and a story to tell. When you know where graphs come from, that helps in reading them correctly.

Exercises

1. A Cobb-Douglas production function with increasing returns to scale yields a total cost function that increases at a decreasing rate. Use Word's Drawing Tools to draw the underlying isoquant map for such a production function.

A commonly used specification for production functions in empirical work is the *translog* functional form. There are several versions. When applied to the cost function, you get a result like this:

$$\ln TC = \alpha_0 + \alpha_1 \ln Q + \alpha_2 \ln w + \alpha_3 \ln r + \alpha_4 \ln Q \ln w + \alpha_5 \ln Q \ln r + \alpha_6 \ln w \ln r$$

Notice that the function is a modification of the log version of a Cobb-Douglas function. In addition to the individual log terms there are combinations of the three variables, called interaction terms.

Click the Exercise Questions button at the bottom of the *Q&A* sheet in the *CostCurves.xls* workbook to reveal a sheet with translog cost function parameters. Use this sheet to answer the following questions.

2. Enter a formula in cell B18 for the TC of producing 100 units of output, given the alpha coefficient and input price values in cells B5:B13. Fill your formula down and then create a chart of the total cost function (with appropriate axes labels and a title). Copy and paste your chart in a Word document.

   Hints: $TC = e^{\ln TC}$ and the exponentiation operator in Excel is EXP(). "=EXP(number)" in Excel returns e raised to the power of that number.

3. Compute MC via the change in output from 100 to 110 in cell C19. Report your result.

4. Compute MC via the derivative at $Q = 100$ in cell D18. Report your result.

   Hint: $\frac{d}{dx}(e^{f(x)}) = e^{f(x)} \frac{d}{dx}(f(x))$

5. Compare your results for MC in questions 3 and 4—are your answers the same or different? Explain.

References

The epigraph is from page 218 of Alan Blinder, Elie Canetti, David Lebow, and Jeremy Rudd, *Asking About Prices: A New Approach to Understanding Price Stickiness* (Russell Sage Foundation, 1998). This book reports the results of interviews with more than 200 business executives. The authors explain that asking about a firm's marginal cost "turned out to be quite tricky because the term 'marginal cost' is not in the lexicon of most business people; the concept itself may not even be a natural one" (p. 216). The question was, therefore, phrased in terms of "variable costs of producing additional units."

The results confirmed what many who have attempted to estimate cost curves know: The canonical, U-shaped family of cost curves makes for nice theory, but it is not common in the real world. In fact, many business leaders have no idea what marginal cost is or how to measure it. Do not lose sight, however, of the purpose of the Theory of the Firm. It is not designed to realistically describe a living firm. The Theory of the Firm is a severe abstraction with a primary goal of deriving a supply curve. The next chapter does exactly that.

Chapter 12

Output Profit Maximization

Initial Solution

Deriving the Supply Curve

Diffusion and Technical Change

> There are many occasions, therefore, when several explorers are surprised, and somewhat pained, on meeting each other at the Pole. Of such an occasion the history of the "marginal revenue curve" presents a striking example. This piece of apparatus plays a great part in my work, and my book arose out of the attempt to apply it to various problems, but I was not myself one of the many explorers who arrived in rapid succession at this particular Pole.
>
> Joan Robinson

12.1 Initial Solution

With a total cost function, $TC(q)$, and its associated average and marginal cost curves, we are ready to solve the the firm's output profit maximization problem. The firm chooses the amount of output that maximizes profit, defined as total revenue minus total cost. This is the second of three optimization problems that make up the Theory of the Firm.

All firms face this profit maximization problem, but this chapter works with a perfectly competitive (PC) firm in the short run (SR). There are, of course, many other market structures and types of firms, but perfect competition is the first step from which more sophisticated scenarios arise.

The firm's market structure tells us the environment in which it operates. Its market structure determines the firm's revenue function. A PC firm is the simplest case because it takes price as given. Thus, revenues are simply price times quantity and the revenue function is linear.

Remember that we are not trying to describe the actual operation of a business. In fact, a truly perfectly competitive firm does not exist in the real world. The concept is an abstraction that enables derivation of the supply curve. This is our goal.

Remember also that the short run is defined by the fact that at least one input (usually K) is fixed. In the long run, the firm is free to choose how much to use of every factor. K is fixed not because it is immovable (like a pizza oven or a building), but because the firm has contracted to rent a certain amount. It cannot increase or decrease the amount of K in the short run.

Profit maximization and its graphs may be familiar from introductory economics. This experience will help you, but do not be complacent. Keep your eye on how the economic way of thinking is being applied in this case and make connections with other optimization problems we have explored.

Perfectly Competitive Market Structure

A perfectly competitive firm sells a product provided by countless other firms selling that homogeneous (which means identical) product to perfectly informed consumers. Because the product is homogeneous, there are no quality differences or other reasons for consumers to care about who they buy from. Because consumers are perfectly informed, they know the price of every seller.

Thus, the PC firm's market structure is one of intense price competition. Every firm sells the product at the exact same price because if anyone tried to sell at even a tiny bit higher than the market price, no one would buy from them.

The shorthand term for this environment is *price taking*. The PC firm must take the price and cannot choose its price—price is exogenous to the firm.

In addition to price taking, the market structure of the PC firm is characterized by an assumption about the movement of other firms into and out of the industry: *free entry and exit*. Firms can enter or leave the market, selling the same good as everyone else, at any time.

These two ideas, price taking and free entry, distinguish the PC firm from its polar opposite, monopoly. A monopolist chooses price and has a barrier to entry. Between these two extremes are many other market structures in which real-world firms actually exist.

The PC firm's market structure means that an individual PC firm does not worry about what other firms are doing. Each firm simply chooses its own output to maximize profit and does not watch the other firms to gain a strategic advantage. In this sense, there is no rivalry in perfect competition.

Setting Up the Problem

As usual, we organize the optimization problem into three parts:

1. Goal: maximize profits (π, Greek letter pi), which equal total revenues (TR) minus total costs (TC).
2. Endogenous variable: output (q).
3. Exogenous variables: price of the product (P), input prices (the wage rate (w) and the rental rate of capital (r)), and technology (parameters in the production function).

Unlike the consumer's utility maximization and the firm's input cost minimization problems, this profit maximization problem is unconstrained. The firm does not have a restriction, like a budget constraint or isoquant, that limits its choice of output to a particular range. It can choose any non-negative level of output.

This greatly simplifies the optimization problem. For the analytical method, it means we do not need the Lagrangean method. All we need to do is take a single derivative and set it equal to zero.

Finding the Initial Solution

Suppose the cost function is:

$$TC(q) = aq^3 + bq^2 + cq + d$$

Then we can form the PC firm's profit function and optimization problem like this:

$$\max_q \pi = TR - TC$$

$$\max_q \pi = Pq - (aq^3 + bq^2 + cq + d)$$

As usual, we have two ways to solve this optimization problem: numerically and analytically.

STEP Open the Excel workbook *OutputProfitMaxPCSR.xls* and look over the *Intro* sheet.

The *Intro* sheet is not meant to be immediately understood. It offers highlights of material that will be explained and prints as one landscaped page. It provides a compact summary of the optimal solution of the output profit maximization problem for a perfectly competitive firm in the short run.

STEP Proceed to the *OptimalChoice* sheet to find the initial solution.

The sheet is organized into the components of an optimization problem, with goal, endogenous, and exogenous variable cells.

Initially, the firm is producing nine units of output and making $11.74 of profit. Is this the highest profit it can possibly make?

No. The sheet reveals the information needed to give this answer. By comparing marginal revenue (MR) and marginal cost (MC), we immediately know that the firm would make a mistake (we would say it is inefficient) if it produced just nine units.

The MC of the ninth unit is \$3.52 as shown in cell B22, but what about MR? Perhaps you remember from introductory economics that $P = MR$ for perfectly competitive firms? We can see that the additional revenue produced by the last unit, \$7 (the price), is greater than the additional cost, \$3.52 (cell B22). Thus, the firm should produce more. How much exactly should the firm produce?

STEP Run Solver to find out.

Look carefully at B22. At the optimal solution, $q^* \approx 13.09$, $MC = \$7$ per unit. $P = MC$, a special case of $MR = MC$ for a PC firm, is the equimarginal condition in this problem, analogous to $MRS = \frac{p_1}{p_2}$ and $TRS = \frac{w}{r}$. When the equimarginal condition is met, the firm is guaranteed to be maximizing profits.

To find the optimal solution via the analytical method, we take the derivative of the profit function with respect to q, set it equal to zero, and solve for q^*. Our cubic cost function introduces the complication that the solution has two roots so we have to use the quadratic formula.

STEP Click the Show the Analytical Solution button to see how to solve this problem with calculus.

Cell AC17's formula has the root that maximizes profits (the other root minimizes profits—more on this in the next section). As usual, Solver and calculus agree (not exactly, but they give effectively the same answer).

Representing the Optimal Solution with Graphs

Since this is an unconstrained optimization problem (unlike utility maximization and input cost minimization), the graphical display of the optimal solution is different.

The firm's output profit maximization problem is usually represented by a graph that depicts the family of cost curves along with marginal and av-

erage revenue. Figure 12.1 and the *Intro* sheet shows this canonical graph for a perfectly competitive firm (signaled by the fact that firm demand is horizontal, so marginal revenue equals demand).

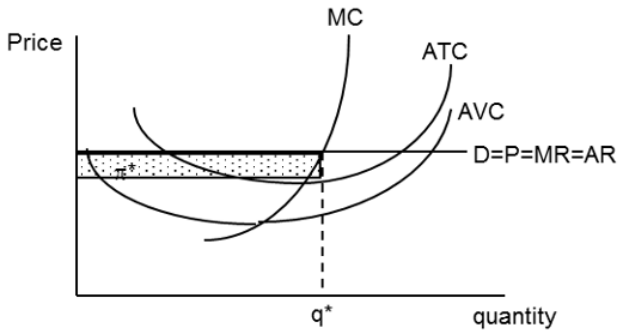


Figure 12.1: The canonical output profit maximization graph.
Source: OutputProfitMaxPCSR.xls!Intro.

Figure 12.1 is the usual display of the optimal solution, but it is actually part of a much larger graphical display.

STEP Proceed to the *Graphs* sheet to see how Figure 12.1 fits into the bigger picture, also shown in Figure 12.2. Zoom out to see all four graphs.

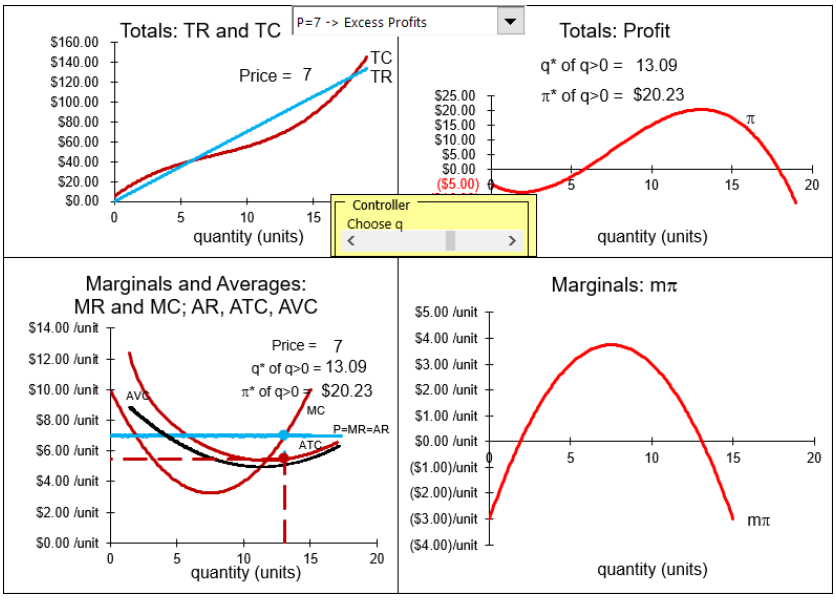


Figure 12.2: Four graphs of output profit maximization.
Source: OutputProfitMaxPCSR.xls!Graphs.

Each of the four graphs in Figure 12.2 and on your screen can be used to show the firm's optimization problem and its solution. We will walk through each one.

1. The top left graph plots total revenue and total cost. TR is linear because the firm's market structure is perfect competition, hence, it is a price taker. The cubic total function produces the shape of TC. The firm wants to choose q to maximize the difference between revenues and costs.

2. The top right graph shows the profit function, which is $TR - TC$. The firm wants to choose q so that it is at the highest point on the profit hill.

3. The bottom right graph displays marginal profit, which can be expressed as the derivative of the profit function with respect to q. The firm can find the maximum profit by choosing q so that marginal profit is zero. This is the first-order condition from the analytical solution.

4. Finally, the bottom left graph is the usual display. The firm chooses q where MR (which equals P given that the firm is a price taker) equals MC. Profits can be calculated as the area of the rectangle $(AR - ATC)q$.

To be clear, all four graphs in Figure 12.2 show the same optimal q and maximum profits, but the graph that is most often used is the bottom left. It highlights the comparison of MR and MC and the family of cost curves provides information about the firm's cost structure. We can also find profits as the area of the rectangle (with blue top and dashed line bottom).

STEP Move the output with the slider control (in the middle of the four charts) to the left and right of q^* to see how the profit rectangle changes.

Only when q is such that $MR = MC$ do you get the maximum area of the profit rectangle. Moving left from optimal q, you can make the rectangle taller, but you must make it shorter to do this and you end up with less area. You can make the rectangle longer by moving right from optimal q, but ATC rises and the rectangle gets thinner, so once again the area falls.

The intersection of MR and MC immediately reveals the optimal q. Profits at any q are also easily seen as the area of a rectangle, length times width, with units in dollars. Because the y axis is a rate, \$/unit, and the x axis is in units of the product, multiplying the two leaves dollars. In other words,

say the product is milk in gallons. Then price, average total, and average variable cost are all in \$/gallon. Suppose that at a price of \$2/gallon, $MR = MC$ at an output of 7,000 gallons and ATC = \$1.50/gallon at this output. Clearly, profits are (\$2/gallon - \$1.50/gallon) x 7,000 gallons, which equals \$3,500.

We can compute profits from the profit rectangle at any level of output. The height of the rectangle is always average revenue (which equals price) minus average total cost. This vertical distance is average profit. When multiplied by the level of output, we get profits, in dollars, at that level of output.

The bottom left graph has another advantage over the other graphs. It can be used to explain a curious and puzzling feature of a firm's short run profit maximization problem. The story revolves around a firm with negative profits and what it should do in this situation.

The Shutdown Rule

The firm has an option when maximum profits are negative: it can simply shut down, close its doors, hire no workers, and produce nothing. The *Shutdown Rule* says the the firm will maximize profits by producing nothing ($q^* = 0$) when $P < AVC$.

The key to whether the firm shuts down or continues production in the face of negative profits lies in its fixed costs. If the firm can do better by shutting down and paying its fixed costs instead of producing and choosing the level of output where $MR = MC$, then it should produce nothing.

Continuing production in the face of negative profits versus shutting down are actually the last two of four possible profit positions for the firm.

1. Excess Profits: $\pi^* > 0$ and $P > ATC$
2. Normal Profits: $\pi^* = 0$ and $P = ATC$
3. Negative Profits, Continuing Production: $\pi^* < 0$ and $P \geq AVC$
4. Shutdown: $\pi^* < 0$ and $P < AVC$

Case 1, excess profits, occurs whenever maximum profits are positive. The example we have been working on is this case. With $P = 7$, we know that $q^* = 13.09$ and $\pi^* = \$20.23$.

STEP In the *Graphs* sheet, click on the pull down menu (over cell R5) and select the *Zero Profits* option.

Your screen now looks like Figure 12.3.

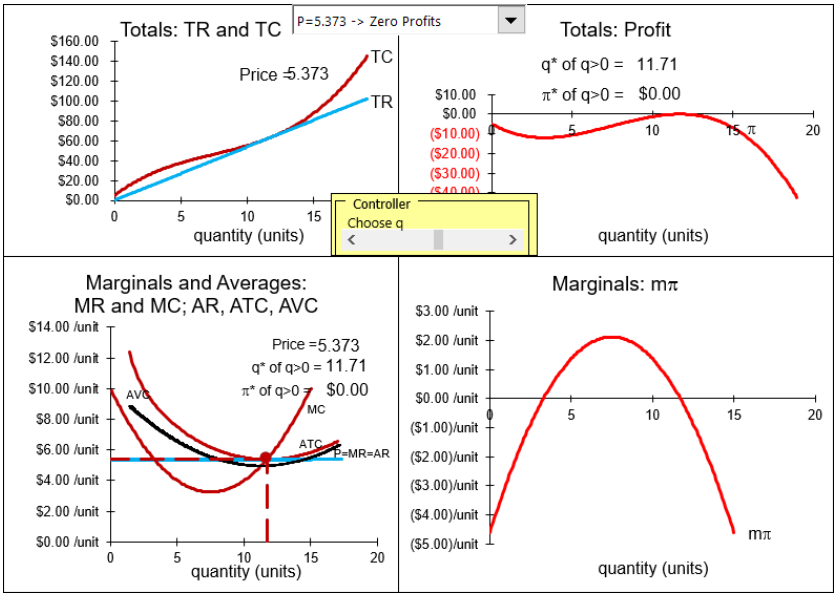


Figure 12.3: Case 2: Normal (zero) profits.
Source: OutputProfitMaxPCSR.xls!Graphs.

Notice that the price ($5.373) in the bottom left chart just touches the minimum of the average total cost curve. The profit rectangle has zero area because it has zero height. The best the firm can do is zero profits—all other choices of q lead to lower (negative) profits.

In the top left graph, you can see that *TR* just touches *TC*. In the top right graph, the top of the profit hill just touches the x axis. These charts confirm what the bottom left chart tells us—with $P = \$5.373$, q^* yields $\pi^* = 0$.

The third and fourth profit cases are the flip side of the first two in the sense that price is so low that profits are now negative. This means firms will leave in the long run, but another question arises: should the firm shut down immediately or continue production?

STEP Click on the pull down menu (over cell R5) and select the *Neg Profits, Cont Prod* option.

With the *Neg Profits, Cont Prod* option selected, $P = 5.10$. The firm produces $q^* = 11.43$ and suffers negative maximum profits of $-\$3.16$. Notice that price is below ATC in the bottom left graph, so that the profit rectangle, $(AR - ATC)q$, will be a negative number. (The area is not negative, but it is interpreted as a negative amount since revenues are below costs.) In the top left graph, the TR line is below the TC curve. In the top right graph, the profit function is below the x axis. There is a maximum, or top of the hill, but it is negative, like a mountain under water.

Keep your eye on the top right graph, reproduced as Figure 12.4. Notice that the top of the profit function is higher than the intercept (where $q = 0$). It is better for the firm to continue production, even though it is earning negative profits of $-\$3.16$ at the optimal output level, because it would make an even lower negative profit of $-\$5$ (the fixed cost) if it shut down.

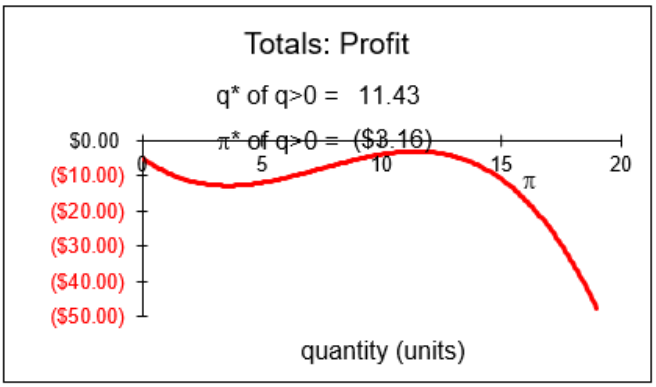


Figure 12.4: Case 3: Negative profits, continuing production.
Source: OutputProfitMaxPCSR.xls!Graphs.

The canonical graph of profit maximization can be used to determine whether the firm should produce or shut down by comparing price to average variable cost. The Shutdown Rule is easy: hire no labor and produce nothing if $P < AVC$.

STEP Look at the bottom left graph on your screen. It confirms that the Shutdown Rule works. Profits are negative because price is below average total cost, but the firm will continue production because $P > AVC$. When the relationship between P and AVC is such that price is greater than average variable cost, it means that the top of the profit function is higher than the y intercept, as in Figure 12.4.

STEP Click on the pull down menu (over cell R5) and select the *Neg Profits, Shutdown* option. Figure 12.5 displays the top right graph.

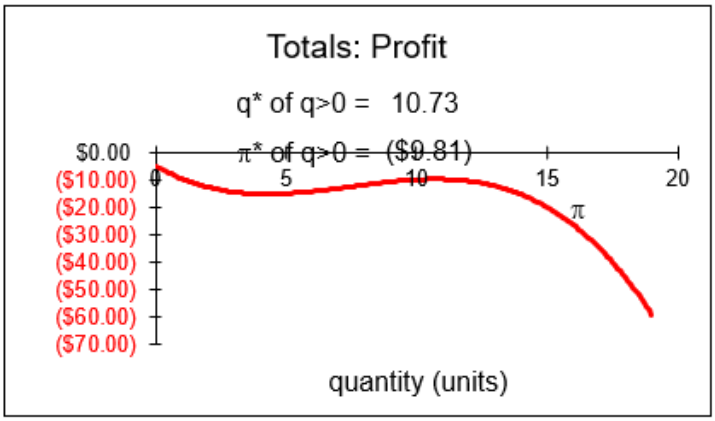


Figure 12.5: Case 4: Negative profits, shut down.
Source: OutputProfitMaxPCSR.xls!Graphs.

In this case, the top of the profit function is below the y intercept. In other words, the maximum profit if the firm produces, $-\$9.81$, is worse than the negative profit incurred if the firm shuts down, $-\$5$. The firm optimizes by choosing $q^* = 0$, that is, shutting down.

STEP Look at the bottom left graph on your screen. Once again, we have confirmation of the Shutdown Rule. With $P = 4.5$, $P < AVC$ and the firm should shut down.

STEP Carefully watch the canonical (bottom left) and profit function (top right) graphs as you change the price (with the pull down menu over cell R5).

As long as $P > AVC$, the top of the profit hill is above the y intercept. If $P = AVC$, the two are exactly equal and the firm is indifferent between producing and shutting down.

$P < AVC$ is the magic cutoff point. When this happens, the top of the hill is below the y intercept (which is the negative profit suffered if the firm produces nothing). Thus, the firm's best choice is to produce nothing.

Here is why the rule works. Multiply the Shutdown Rule by q to get:

$$(P < AVC)q$$
$$Pq < AVCq$$
$$TR < TVC$$

$TR < TVC$ is a restatement of the Shutdown Rule—produce nothing if total revenue cannot cover total variable costs. This makes sense. Why produce if you can't even pay for the variable expenses? You are better off not producing at all.

If total revenue is less than average total cost, then profits are negative. However, the firm can be in a situation where $TR < TC$, but $TR > TVC$. If so, then production makes sense because you will be able to reduce some of the fixed costs you have to pay no matter what you do. Profits are negative, but it is better to produce than not produce because variable costs are covered and fixed costs are at least partially reduced.

STEP For a summary of the four cases and what the Shutdown Rule is doing, click the Show Cases button (over cell AC5).

What's Normal about Zero Profits?

In economics, zero profits are called *normal profits*. This is confusing. Zero sounds bad, not normal. There is a logical explanation, but it requires a clear separation of accounting versus economic profits. They differ because economists include opportunity costs when calculating economic profits.

- Accounting profits = revenues - explicit costs
- Economic profits = revenues - explicit costs - opportunity costs

In economics, without an adjective, "profits" means *economic profits*. So, when profits are zero that means economic profits are zero. Economic profits have had an extra item subtracted, the opportunity costs of using firm resources to make this particular product.

An accountant would subtract explicit (out-of-pocket) costs (wages, rent, etc.) from revenues and if this number is positive, announce that the firm is making money. The economist would then subtract the cost of the profits that could be made by the next best alternative industry that the firm could

be in. If economic profits are zero, it means the opportunity costs are exactly equal to the accounting profit and the firm cannot do better by switching to its next best alternative.

Although this may seem needlessly contorted at first, there is a nice interpretation of economic profits: If positive, the firm will stay in the industry and new firms will enter in the long run; if negative, the firm will exit in the long run; and if zero, there will be neither exit nor entry in the long run. It is in this sense of equilibrium that we say zero profits are normal. With $\pi = 0$, there is stability and no tendency to change in the movement of firms.

The distinction between economic and accounting profits also explains why positive profits are *excess* profits. It is not meant as a pejorative term, but to indicate that the firm is earning greater profits than are needed to keep producing that product in the long run. Excess profits also mean that others are attracted and will enter that industry.

Economists are not concerned with how much money the firm made, but with profits as a signal to entry and exit. Defining economic profits as accounting profits minus opportunity costs gives us a profit measure that tells us whether the firm will stay or leave in the long run.

Shutdown Rule and Corner Solution

The Shutdown Rule is usually covered in introductory economics. Memorization is often all that is achieved. We can do better by properly situating the Shutdown Rule in the landscape of mathematical and economic concepts—it is a corner solution.

Recall that, in the Theory of Consumer Behavior, there are situations in which the MRS does not equal the price ratio, yet the solution is optimal. This is a corner solution.

Food stamps are an example. The fact that food stamps can only be used to buy food creates a horizontal segment on the budget constraint so that a consumer might not be able to make MRS $= \frac{p_1}{p_2}$. At the kink in the constraint, the consumer is optimizing even though the equimarginal condition is not met.

Corner solutions are a general phenomenon. They can be seen whenever a restriction or border blocks further improvement in the objective function. Consider Figure 12.6 which sketches a maximization problem to highlight the difference between an interior and a corner solution. In panel B, the agent cannot choose negative values of the x variable and, therefore, the function is cut off by the y axis.

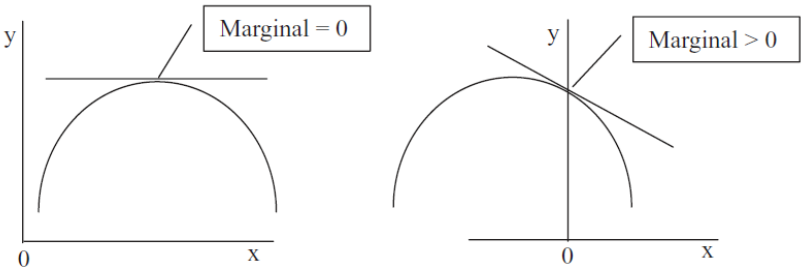


Figure 12.6: Understanding the corner solution.

In panel B, although the marginal condition is not met, we have an optimal solution, defined as doing the best we can without violating any constraints.

Shutting down is another example of a corner solution because, once again, the equimarginal condition is not met at $q = 0$, yet producing nothing is the optimal solution. Shutting down is an unusual example of a corner solution because there *is* a place where the marginal condition is met (there is an output where $MR = MC$), but it is not optimal. The profit function twists in such a way (see Figure 12.5) that profit is decreasing as output rises from zero. This means that profits would go up if we were able to produce negative output. Since we are not allowed to choose $q < 0$, we have a corner solution.

How can we know if we should choose q at $MR = MC$, the interior solution, or shut down, the corner solution? The only way is to compare the profit positions at the two quantities. The good news is that no checking is required for cases 1 and 2. As long as profits are non-negative, there is no way that a profit of minus total fixed cost can be better than the interior solution of q where $MR = MC$. But, whenever, $MR = MC$ yields negative maximum profits, comparing those negative profits to TFC is necessary. Or, you could just use the Shutdown Rule and see if P ¡ AVC, which will give the same, correct answer.

The complexity of the firm's profit maximization problem in the short run, with its shutdown possibility, should increase your sensitivity to lurking problems with analytical and numerical methods. We know neither is perfect so there may be glitches in applying these methods to the firm's profit maximization problem. The *Q&A* sheet provides an example. Be sure to look carefully at questions 2 and 3.

Finding and Displaying the Initial Solution

The output profit maximization problem for a PC firm in the short run is a single-variable (q) unconstrained problem. It can be solved with numerical and analytical methods. The equimarginal rule applied is that $MR = MC$ and since price taking behavior means that $P = MR$ for a PC firm, the equimarginal rule is often shown as $P = MC$.

The firm's profit maximization problem contains a complication in the short run. If maximum profits are negative, it is possible that the firm is better off not producing anything. A shortcut to determine whether or not to produce when $\pi^* < 0$ is the Shutdown Rule, $P < AVC$.

The initial optimal solution is displayed by a canonical graph that superimposes the firm's revenue side (average and marginal revenue) over its cost structure (average and marginal costs). Optimal output is easily found where MR intersects MC (as long as $P > AVC$) and maximum profit is displayed as the area of the appropriate rectangle. The ability to instantly show the optimal solution, maximum profits, and whether or not to shut down explains the popularity of this graph.

You can think of the firm as walking through a series of three steps when solving its profit maximization problem:

1. Choose q where $MR = MC$ in the canonical graph.
2. Compute profits at q^* via $(AR - ATC)q$ (the profit rectangle).
3. If profits are negative, shut down if $P < AVC$.

The PC firm's profit maximization is simpler in the long run. If $\pi < 0$, firms exit the industry; $\pi > 0$ (also known as excess profits) lead to entry. Thus, in long run equilibrium (a state never actually attained), $P = ATC$ and $\pi = 0$ for all firms. This is why zero economic profits are called normal profits.

Exercises

1. Use Excel's Solver to find the optimal output and profit for a firm with cost function $TC = 2q^2 + 10q + 50$ and $P = 40$. Take a screen shot of your optimal solution (including output and profits) and paste it in a Word document.

2. Use analytical methods to solve the problem in the previous question.

3. For what price range will the firm in question 1 shut down? Explain.

4. If fixed costs are higher, will this influence the firm's shutdown decision? Explain.

References

The epigraph is from the foreword (p. vi) of Joan Robinson, *The Economics of Imperfect Competition* (first edition, 1933, followed by many reprints). In a male-dominated profession, Joan Robinson established herself as a well-known, important economist. She helped create the Theory of the Firm, including the canonical graph with average and marginal revenue and cost that is used to this day.

Ironically, however, much of her work was critical of mainstream economics. Her famous Richard T. Ely lecture at the 1971 American Economics Association conference pulled no punches:

> For once the president of the AEA was a dissident. This was the veteran institutionalist and Keynesian John Kenneth Galbraith, a longtime friend of Robinson's and celebrated critic of US capitalism and its apologists in academic economics. Galbraith now offered her the most important platform she had ever occupied. Robinson took full advantage of it, delivering an abrasive, challenging, deliberately provocative indictment of neoclassical economics that was designed to polarize her audience between the old and conservative and the young and progressive. (John Edward King, *A History of Post Keynesian Economics Since 1936* (2002), p. 123.)

As all the functions $\phi_k(D_k)$ are supposed to increase with D_k, the expression for D_k derived from the equation $p = \phi_k(D_k)$ is itself a function of p, increasing with p. [Translation: The supply curve, $q^* = f(P)$ is derived from $P = MC$ and it is upward sloping because MC is upward sloping.]

Augustin Cournot

12.2 Deriving the Supply Curve

The most important comparative statics analysis of the firm's output profit maximization problem is based on tracking q^* (quantity supplied) as price changes, ceteris paribus. This gives us the firm's supply curve.

An important thing to remember is that the supply curve has two parts:

1. MC when $P >$ min AVC
2. Zero otherwise (Shutdown Rule)

As usual, we have numerical and analytical methods at our disposal for the comparative statics analysis that generates the supply curve. Before we begin, we show how Solver can be modified to deal with the shut down possibility and revisit the fact that it is not a silver bullet.

Solver Issues

STEP Open the Excel workbook *DerivingSupply.xls*, read the *Intro* sheet, then go to the *OptimalChoice* sheet to see an implementation of a PC firm's profit maximization problem in the short run.

The sheet looks like the *OptimalChoice* sheet in the *OutputProfitMaxPCSR.xls* workbook (from the previous section), but it has a few additional cells.

The IF statements in cells C4 and C8 of the *OptimalChoice* sheet are a convenient way to incorporate the firm's shutdown option.

STEP Click on C8 to reveal its formula: = IF(max profit >= − d, q, 0). We will use this cell as the correct optimal solution in all cases, including the shutdown case.

It is easy to see that Solver has been run because at $q \approx 10$ in cell B8, $MR = MC$ since $P = 4$ and cell B18 reports $MC = 4$. This q, however, is not the optimal solution because cell B4 shows that $\pi = -15$ (using the common convention that "()" denote negative numbers). This firm would be better off not producing at all and suffering a loss of $TFC = -5$. The Shutdown Rule says the same thing since $P < AVC$ (cell B15 is \$5).

While Solver's answer is wrong (because it found the top of the profit hill, which is lower than the y intercept at $-TFC$), we can add a step to Solver where we check for exactly this situation. This is what cells C8 and C4 do.

The expression $max\_profit \geq -d$ is used to test if Solver's answer (the interior solution) has higher profits than negative total fixed costs (the corner solution). If true, it keeps Solver's solution; if false, the optimal solution is zero (shut down).

Solver will find the best of the positive levels of output in cell B8 and the IF statement in cell C8 checks to make sure that the best solution (of the $q > 0$) is better than shutting down and producing nothing ($q = 0$).

With $P = 4$, the best of all of the positive levels of output, $q = 10$, provides a profit of minus \$15. Cells C4 and C8 show that producing nothing yields a higher profit (and smaller loss) of minus \$5 and is the correct optimal solution.

While this is an improvement over manually checking Solver's answer, there is another potential problem with Solver in this application.

STEP To see the problem, set P (cell B12) to 7 and run Solver.

The optimal q is approximately 13.09 and the firm is enjoying excess profits. Cells B4 = C4 and B8 = C8 because Solver's answer gives profits greater than minus *TFC*. All is well.

STEP Now set cell B8 = 1. Run Solver from this initial value.

Solver's result is disastrous! What happened?

STEP Click the Solver Explained button to see why starting from $q = 1$ leads Solver astray.

The explanation on the sheet makes it clear that the initial or starting value can play a critical role when numerical methods are utilized. This profit maximization problem has a sufficiently complicated surface that a numerical algorithm, such as Solver, cannot easily distinguish between local and global optimal solutions. There is no simple fix. The lesson is that you have to know the optimization problem you are dealing with and be careful interpreting the answers provided by a numerical algorithm.

The explanation of Solver's failure involves the minimum point of the profit function and this provides an opportunity to explain the two roots in the quadratic formula. A picture, in this case, really is worth a thousand words.

STEP Click the Show Analytical Solution button.

Cell Z17 has the other root from the quadratic formula (computed by adding instead of subtracting the square root term). Both roots are places where the profit function is flat (in the top right graph on the sheet). Notice how the dashed lines from the max and min profit points lead to points where marginal profit ($m\pi$) is zero. These are the two roots in the quadratic formula.

The two roots can also be seen in the canonical, bottom left graph as the two points where MR and MC intersect. Of course, we only care about the root that maximizes profits. One way to ensure that $MR = MC$ yields a profit max is to make sure that $MC < MR$ to the left of the intersection. In other words, MC cuts MR from below.

Numerical Methods to Derive the Supply Curve

STEP Set cell B8 back to 10 and $P = 4$ so Solver will converge to the local max at $q = -15$.

STEP Run the Comparative Statics Wizard from $P = 4$ with 0.05 sized shocks 100 times. Track the C4 and C8 cells as endogenous variables. You can safely ignore the warning—you are using the CSWiz to keep track of these cells, but will not include them as changing cells in the Solver dialog box.

Your results will look like those in the *CS1* sheet. Notice that at low prices, the firm is producing nothing. This is the part of the supply curve where the firm shuts down to maximize profits.

The supply curve and inverse supply curves can be graphed with the CSWiz data, as shown in Figure 12.7 and the *CS1* sheet. Of course, the tail runs along the quantity axis all the way to zero. Just as with the demand curve, $q = f(P)$ is the supply curve and flipping the axes, $P = f^{-1}(q)$, gives the inverse supply curve.

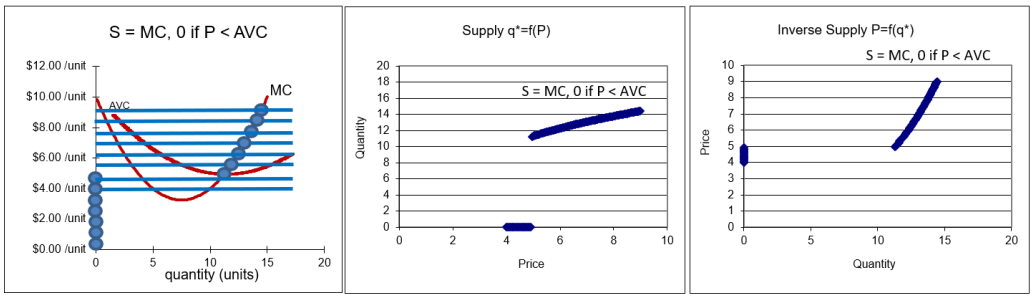


Figure 12.7: Deriving supply and inverse supply curves.
Source: DerivingSupply.xls!CS1.

Figure 12.7 applies our usual graphical exposition. The leftmost chart is the underlying graph from which the other charts are produced. We shock P and track q^*. This gives the supply curve.

Unlike the demand curve, however, notice that the supply curve follows *MC* as long as *P* is not below *AVC*. The discontinuity is at the minimum *AVC*. Row 32 of the *CS1* sheet shows the break occurs for this cost function between \$4.90 and \$4.95. Prices below this minimum *AVC* value result in no quantity supplied since the firm shuts down.

Analytical methods can be used to find the discontinuity. First, we obtain an expression for *AVC*.

$$TC = 0.04q^3 - 0.9q^2 + 10q + 5$$
$$TVC = 0.04q^3 - 0.9q^2 + 10q$$
$$AVC = \frac{TVC}{q} = 0.04q^2 - 0.9q + 10$$

Then we take the derivative of AVC with respect to q and set it equal to zero to find its minimum point.

$$\min_q AVC = 0.04q^2 - 0.9q + 10$$
$$\frac{dAVC}{dq} = 0.08q - 0.9 = 0$$
$$q^* = \frac{0.9}{0.08} = 11.25$$

By plugging this minimum value of output into the AVC function, we know the price at which the discontinuity kicks in.

$$AVC[q = 11.25] = 0.04[11.25]^2 - 0.9[11.25] + 10 = 4.9375$$

In the *CS1* sheet, the discontinuity occurs when price rises from \$4.90 to \$4.95. Our analytical work tells us that the discontinuity is exactly at \$4.9375. Any price below this yields optimal q of zero.

Notice how we used the derivative to find the value of q at which the rate of change for the AVC curve was zero. This is the bottom of the U-shaped AVC curve and prices below this AVC result in shutting down. The lesson is that derivative is a tool that has a variety of uses.

The *CS1* sheet also computes the price elasticity of supply in column E.

STEP Scroll down to see a comparison of slope and elasticities via the Δ and derivative approaches.

In this case, the two approaches are not exactly the same because q^* is non-linear in P. The sheet has all of the details in case you want to refresh your understanding of this concept.

Analytical Methods to Derive the Supply Curve

For the analytical approach, we use a different cost function to give us more practice.

$$TC(q) = q^2 + 20$$

With this quadratic cost function, we can set up and solve the PC firm's profit maximization problem. Because it is a perfectly competitive firm, we know price is given and, thus, $TR = Pq$. Therefore, the optimization problem is:

$$\max_q \pi = Pq - (q^2 + 20)$$

We proceed by taking the derivative with respect to q and setting it to zero, then solving this first-order condition for optimal q.

$$\frac{d\pi}{dq} = P - 2q = 0$$

$$q^* = \frac{1}{2}P$$

This is the supply function. It gives the quantity supplied by a firm at every given price. For example, with $P = 20$, $q^* = 10$.

The inverse supply curve is found by expressing the equation as $P = f(q)$.

$$P = 2q^*$$

The supply function tells us that q^* increases by one-half fold for every increase in P. The size of the change in P does not matter since $\frac{dq}{dP}$ is constant.

The price elasticity of supply is $+1$.

$$\frac{dq}{dP} = \frac{1}{2}$$
$$\frac{dq}{dP}\frac{q}{P} = \frac{1}{2}\frac{P}{\frac{1}{2}P} = 1$$

We can compute the price elasticity of supply from one point to another. We know that at $P = 20$, $q^* = 10$. If $P = 30$, $q^* = 15$. A 50% rise in price led to a 50% increase in quantity supplied so the price elasticity of supply is $+1$. The result is the same as the derivative approach because q^* is linear in P.

A PC firm with a quadratic cost function will not shut down with any price greater than zero. By constructing its family of cost curves and graph of the optimal solution, we can see why. We begin with the cost curves. We know $TVC = 2q$ and $TFC = 20$. Then we can find the average and marginal curves.

$$ATC(q) = \frac{TC}{q} = \frac{q^2 + 20}{q} = q + \frac{20}{q}$$
$$AVC(q) = \frac{TVC}{q} = \frac{q^2}{q} = q$$
$$MC(q) = \frac{dTC}{dq} = \frac{d(q^2 + 20)}{dq} = 2q$$

STEP Proceed to the *Graphs* sheet to see the four graph display of the optimal solution for this problem.

If $P = 20$, then $q^* = 10$ and $\pi^* = \$80$. It is also obvious that there is no positive price at which this firm will shut down because AVC is simply a ray with slope +1 out of the origin. Thus, price can never fall below AVC.

Notice also how there is only one point where $MR = MC$, unlike the two intersections we saw with the cubic cost function. The quadratic cost function cannot produce the S-shape TC needed for the profit function to have a minimum profit at the bottom of a U-shape. The profit function in the top right graph has a single top of the hill (where $m\pi = 0$).

Points Off the Supply Curve

As we did with the demand curve (see Figure 4.12), we can explore the meaning of being off the supply curve. The interpretation is quite similar.

STEP Return to the *CS1* sheet and manipulate the point off the supply and inverse supply curves with the scroll bar in column E.

Figure 12.8 shows what is on your screen, but in Excel you can move the red dot. As you do, the chosen q and profit for that quantity is displayed.

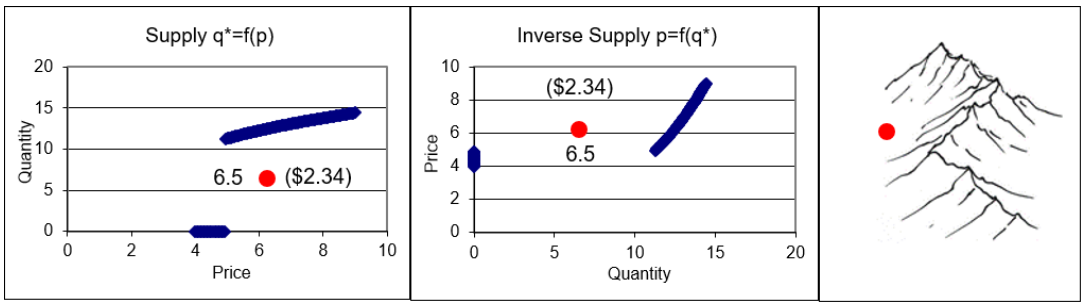


Figure 12.8: Points off the supply curve.
Source: DerivingSupply.xls!CS1.

Profits are maximized when you are on the supply curve. It is clear that the supply curve, like the demand curve, has a hidden third dimension—profit for supply and utility for demand. The right most panel shows the mountain

and how you approach the top at the optimal solution. The ridgeline connecting the mountain tops is the supply curve. Like the demand curve, points off the supply curve are associated with lower values of the objective function.

Notice how the point off the curve moves in a vertical fashion in the supply curve graph and horizontally on the inverse supply curve graph. This happens because price is constant (at $P = 6.25$). With the price on the x axis, points can be above or below the supply curve. Points off the inverse supply curve are to the right or left because P is on the y axis.

Finally, on the inverse supply curve, the inefficiency of being off the curve is obvious because output levels off the inverse supply curve means the firm is not choosing a point where $MR(= P) = MC$.

The Supply Curve has Parents

Like demand and cost curves, supply is derived from an optimization problem. Knowing where key relationships come from separates introductory from more advanced economics and is an important aspect of mastering the economic way of thinking.

The supply curve is a comparative statics analysis of the effects on optimal quantity as price changes, ceteris paribus.

Unlike the demand curve, the supply curve has a discontinuity because the firm will shut down if price falls below AVC. The supply curve depends critically on the firm's cost function. The inverse supply curve is simply MC above AVC and zero otherwise. The firm will choose that level of output where $MR(= P) = MC$ as long as $P > AVC$.

Like the demand curve, points off the supply curve are interpreted as inefficient solutions to the optimization problem. Although possible, no optimizing agent would choose a point off the supply (or demand) curve.

Exercises

1. What happens to the short run supply curve if wages rise? Explain. Use Word's Drawing Tools to create a graph depicting your answer.

2. What happens to the inverse short run supply curve if wages rise?

Explain. Use Word's Drawing Tools to create a graph depicting your answer.

3. What happens to the short run supply curve if the rental rate of capital increases? Explain.

4. What happens to the short run supply curve if the price (P) increases? Explain.

5. Suppose a firm is off its short run supply curve, but at a point where $MR = MC$. Use Word's Drawing Tools to a draw the profit function for this situation and label a point Z that meets the supposed conditions.

References

The epigraph comes from page 92 of the 1897 English translation of Augustin Cournot's *Researches into the Mathematical Principles of the Theory of Wealth*. This book was originally published in French in 1838. It is a remarkable work—truly far ahead of its time.

Cournot (pronounced coor–no) solves profit maximization problems for a variety of market structures, including monopoly, unlimited (today called perfect) competition, and intermediate cases of small numbers of firms. He uses derivatives and integrals with numerous supporting figures, including supply and demand with price on the x axis. Cournot was not bound by Marshall's convention of P on the y axis since Marshall's famous graphs of supply and demand would not appear until 1890.

The mathematical exposition was simply beyond the grasp of many readers in 1838 and the book languished in obscurity until the rise of mathematics in economics. You will hear Cournot's name again in the chapter on Game Theory.

Why was the spread of crops from the Fertile Crescent so rapid? The answer depends partly on that east-west axis of Eurasia.

Jared Diamond

12.3 Diffusion and Technical Change

The Theory of the Firm is a highly abstracted model of a real-world firm, yet there are fundamental ideas that can be applied to observed firm behavior. This section does exactly that, applying the Shutdown Rule to explain differing rates of diffusion of new technology.

The Shutdown Rule, $P < AVC$, says that firms will not produce when price is below average variable cost because profits are maximized (and losses minimized) by shutting down instead of producing at the best of the positive output choices (at $MR = MC$).

Diffusion of new technology is the process by which new methods of production are adopted by firms. The speed of diffusion is critical—the faster firms upgrade and modernize, the richer the society. We will see that some industries have fast and others slow diffusion with the Shutdown Rule playing a key role.

Setting the Table

Consider two thoughts that are both wrong:

1. Always upgrade to have the best equipment or to use "best practice" techniques.

2. Never throw working machinery away or abandon a process that can produce output.

The first statement is wrong because firms would always be replacing almost new machinery, tools, and plant to have the very latest equipment. The second statement is the polar opposite of the first: Now you keep using ancient machinery that was long ago superseded by better technology just because it is still functioning.

There has to be a middle ground between these two extremes and a logical way to determine when to replace equipment.

Consider these two words that are accepted as synonymous in common usage, but are different in the language of the specialized literature of diffusion:

1. *Outmoded*: machinery that is not the best at the time, but is still used.
2. *Obsolete*: machinery that is scrapped (thrown away) yet still functions.

Your phone is *outmoded* if it is not the latest, greatest available version. When you replace your phone with a new one, the old one becomes *obsolete*. At any point in time, a few people have the newest, fanciest model; the rest have versions of outmoded models still in use; and there are many obsolete models that are no longer being used. As time goes by, the newest model becomes outmoded and, eventually, obsolete.

The distinction between outmoded and obsolete sharpens our focus on this question: When does machinery go from being outmoded to obsolete?

Another important idea is *labor productivity*: the ability of labor to make output. This is measured in two ways, output per hour or labor required to produce one unit of output.

The output per hour version is simply the average product of labor, $\frac{q}{L}$. The bigger this ratio, the more productive is labor. You can take the reciprocal and ask, "How much labor is needed to make one unit of output?" This measure, called the *unit labor requirement*, gets smaller as labor productivity improves.

There are two ways of increasing labor productivity:

1. Better labor: increasing education.
2. Better machinery: technical (or technological) change.

Most people only think of the first way. More educated and skilled labor obviously will be more effective in translating labor input into output. But holding labor quality constant, if workers have better technology, such as computers or power tools, then labor productivity rises.

So, if you want to increase ditch digging productivity, you can improve the worker (think ditch digging classes) or you can improve the technology. A

worker with a shovel digs a ditch a lot faster than one without. But the explosion in productivity and output really occurs when you give the worker a backhoe.

But here's the curious thing, after backhoes are invented and brought online, if you look at the entire industry of ditch digging, you will see many different methods being used. Not everyone will instantly adopt the backhoe.

The question we are interested in boils down to explaining the *rate of diffusion*: how rapidly do the latest, best machinery and methods spread?

The mere existence of a new machine (e.g., a backhoe) is not enough to spur economy-wide increases in labor productivity. If the machine is not adopted rapidly, it will have little effect on the economy. We want fast diffusion so new methods spread quickly. This will boost productivity and economic growth.

The rate of diffusion is like adding a drop of red dye in a bucket of water. How rapidly will the water turn red? What factors affect the rate of diffusion? If we stir, the rate of diffusion rockets—how can we "stir" the economy to speed up diffusion?

It turns out that the rate of diffusion of technical change in an economy varies across industries and depends on specific characteristics. We are not searching for an unknown constant, but for the factors that explain wide variation in rates of diffusion—sometimes backhoes are rapidly adopted and other times not.

The rate of diffusion depends on whether machinery is determined to be outmoded versus obsolete. If machines are scrapped and replaced with the latest technology fairly quickly, then the rate of diffusion of technical change will be fast. If old technology is kept online and in production for a long time, then the rate of diffusion of technical change will be slow.

Before we see how the Shutdown Rule plays a critical role in deciding whether machinery is outmoded or obsolete, we review data used by W. E. G. Salter (1960) to support the claim that the rate of diffusion varies across industries. We also introduce a new graph that captures the idea of a distribution of methods or vintages of machinery.

On the Variation of Methods Used Across Industries

Salter presents data on a variety of goods. He focuses on the methods of production used at any point in time. It is quite obvious that there is always a mix of technologies being used. As new plants come online and new machinery is installed, older plants with older machinery remain in operation.

For example, Salter's Table 5, reproduced as Figure 12.9, shows a mix of technologies used in pig-iron production. Notice that the labor productivity of the best-practice plants (the latest technology) rises from 1911 to 1926. The industry average, however, lags behind because the latest technology is not immediately adopted by every manufacturer. The machine charged and cast method (the right most column) is the best technology, but even by 1926, 30.6% of the firms are not using it. These firms remain in operation with older technology. This slow diffusion hampers industry-wide labor productivity.

Table 5. Methods in use in the U.S. blast-furnace industry, selected years, 1911–26

	Gross tons of pig-iron produced per man-hour		Percentage of plants using the following methods		
Year	Best-practice plants	Industry average	Hand-charged and sand-cast %	Mixed types %	Machine charged and cast %
1911	0.313	0.14	50.0	22.7	27.3
1917	0.326	0.15	41.9	34.9	23.2
1919	0.328	0.14	42.0	28.0	30.0
1921	0.428	0.178	22.2	44.3	33.5
1923	0.462	0.213	20.7	39.7	39.6
1925	0.512	0.285	7.2	25.5	67.3
1926	0.573	0.296	6.1	24.5	69.4

Figure 12.9: Slow diffusion in pig-iron production.
Source: DiffusionTechChange.xls!Data.

Figure 12.10 (Salter's Table 6) focuses on the production of five-cent cigars. Salter keeps constant the quality and type of cigar, the five-cent variety, to focus on an apples-to-apples comparison of production methods. Because the measure of productivity is the labor required to make 1,000 five-cent cigars, the *lower* the hours required, the *greater* the labor productivity. The two-operator machine is the best practice, but three other methods are also used. Once again, the point is that a mix of methods are used and all of them combined determines industry-wide productivity.

Table 6. Approximate labour requirements per thousand five-cent cigars for different manufacturing methods, Unites States, 1936

Manufacturing methods in use in 1936	Man-hours per thousand cigars
Hand made	33.38
Machine bunched, hand rolled	27.38
Four-operator machine	15.96
Two-operator machine	11.94

Figure 12.10: Various methods of producing five-cent cigars.
Source: DiffusionTechChange.xls!Data.

Figure 12.11 offers a final example of Salter's point that an economy's labor productivity depends on the technology actually being utilized to make output. The *Range of all plants* column shows substantial variation in output from the best-practice firms to the least productive methods still being used. Notice that lower numbers are higher productivity because, as the title says, we are measuring "labour content per unit of output."

Table 8. Variation in labour content per unit of output in selected industries

			Man-hours per unit of output			Ratio of range to mean	
Industry, time and place	No. of plants	Unit of Output	Mean	Range of all plants	Range of middle 50% of plants	All plants	Middle 50%
Bricks, UK, 1947	17	1000 bricks	1.36	2.12–0.64	1.75–0.93	1.16	0.61
Houses, UK, 1948	160	Standard house	3080	4300–2150	3520–2630	0.66	0.29
Men's shoes, UK, 1949	12	Dozen pairs	9.70	12.34–7.30	11.02–8.53	0.53	0.26
Cement, US, 1935	60	100 barrels	46.7	86.0–25.3	57.9–39.3	1.30	0.40
Beet sugar, US, 1935	59	Ton of beet sliced	1.46	2.81–0.88	1.98–1.20	1.32	0.53
Sole leather, US, 1949	8	1000 lb.	48	–	61–39	–	0.47

Figure 12.11: Variation in labor productivity across six industries.
Source: DiffusionTechChange.xls!Data.

For bricks, with 17 plants in operation, the middle 50% range is from a best 0.93 hours to make 1,000 bricks to 1.75 hours. That is a huge difference and it is just the middle 50%. Take a moment to look at the ranges of the other products in Figure 12.11.

The *Ratio of range to mean* columns measure the rate of diffusion. If somehow every plant adopted the best-practice method, this ratio would be zero. Thus, houses and men's shoes are industries with much faster diffusion than the others.

Pig-iron, five-cent cigars, and products in Figure 12.11 are examples of a widespread phenomenon that was of great interest to Salter. The rate of diffusion of new technology is neither constant nor instantaneously fast. Salter wanted to know what diffusion depends on in the hope of manipulating it. After all, if there is a policy or lever we can pull to speed up diffusion, we would improve productivity and increase output.

A Graph is Born

Salter used an uncommon graph, an ordered histogram, to show how an industry incorporated various technologies in production.

Figure 12.12 (Salter's original Fig. 5) uses rectangles to indicate each method or vintage of machinery. We call this a *Salter graph*.

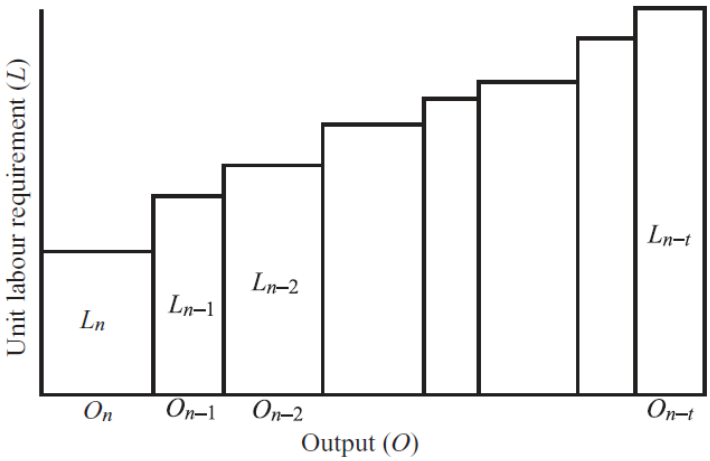


Figure 12.12: Salter graph of the mix of technologies.

The greater the base of each rectangle in Figure 12.12, the greater the share of the industry's output for that particular technology. So, in the middle of the graph, the wider rectangle has a bigger share of the output than the narrower one right next to it. The sum of lengths of the bases have to add up to 100% of the industry output.

The height of each rectangle tells you how much labor is needed to make one unit with that technology. The lower the height (because the y axis shows the labor required to make one unit of output), the greater the labor productivity for that technology.

The Salter graph has to have a stair-step structure because the rectangles are ordered according to when they came online. The oldest technology is to the right and the newest is to the left. The left-most rectangle is the best-practice technology at that time and all of the other rectangles are at different stages of outmodedness.

The Salter graph in Figure 12.12 is actually a single frame of a motion picture. As time goes by, and new techniques are invented and brought online, some of the right most rectangles will "fall over" and be replaced by a new shorter rectangle coming in from the left. Figure 12.13 shows a possibility for the next frame in the movie.

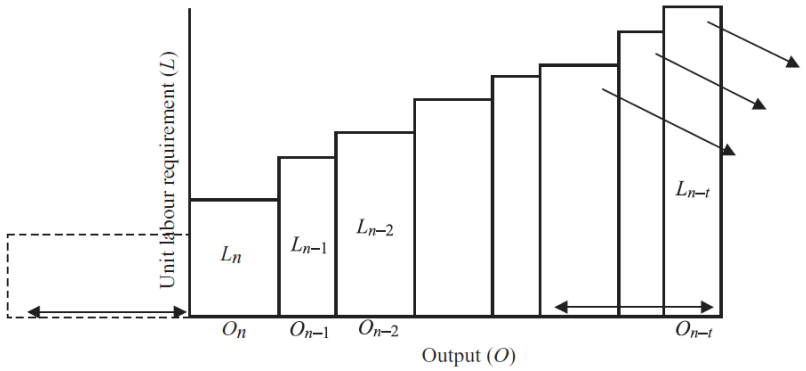


Figure 12.13: Salter graph as time goes by.

The base of the rectangle of the newest technology in Figure 12.13 equals the sum of the widths of the three rectangles representing obsolete technologies, which fall off the graph because they are no longer used.

The wider the base of the newest technology, the better in terms of fast diffusion of technological change and rapid increases in industry-wide productivity. If a new technology swept through an industry like wildfire, the Salter graph would show it as having a very long base, indicating it was producing a large share of industry output.

Another, less favorable possibility is that the newest technology has a small width. This would mean that few firms have adopted the best-practice method and industry-wide productivity will not improve by much. The industry will remain dominated by outmoded methods.

Consider the two Salter graphs in Figure 12.14 (Salter's original Fig. 12). They are enhanced by a strip in the middle, the height of which represents the industry average productivity.

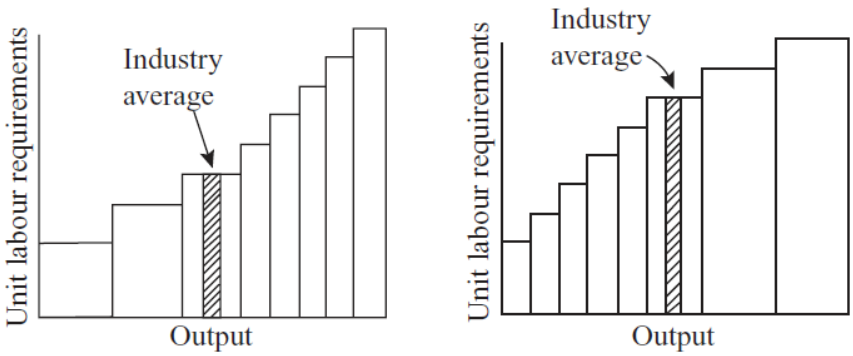


Figure 12.14: A comparison of two industries.

We would much prefer the industry on the left in Figure 12.14 because it has a lower industry average unit labor requirement, which means it has higher productivity. This is a result of much more rapid diffusion of newer, higher productivity technology.

The industry average shaded bar is a *weighted average* of all of the technologies in existence at any point in time. This statistic is the correct way to add up the rectangles with differing widths into a single measure of industry productivity. To understand how to do this, we turn to a concrete example in Excel.

STEP Open the Excel workbook *DiffusionTechChange.xls*, read the *Intro* sheet, then go to the *IndustryAverage* sheet to see how a weighted average is computed and how the Salter graph works.

Cells C9 and C10 show how two technologies contribute to the industry output. Initially, Methods A and B produce 50% of the total output. Because A (the superior, best-practice technology) requires only 1 hour of labor to make a unit of output, whereas B (an outmoded technology) requires 2 hours, the industry average productivity is 1.5 hours per unit of output.

STEP Click on the scroll bar a few times to increase A's share of total output to 90%. Notice how the Salter graph changes as you manipulate the scroll bar.

The Salter graph now shows A's share as a much wider rectangle (indicating much faster diffusion) and the red, industry (weighted) average rectangle is much shorter. Although the simple average does not change, the weighted average falls because more of the output is being generated by the more productive A technology. The weighted average computation (implemented in the formula for cell M10) is:

$$WeightedAverage = \frac{Output_A}{TotalOutput} UnitLReq_A + \frac{Output_B}{TotalOutput} UnitLReq_B$$

STEP Click on the scroll bar to decrease A's share of total output to 10%.

This time, the industry (weighted) average is 1.9 because only 10% of the output is produced with the best-practice technology. This would be an example of slow diffusion.

The contributions of each technology to industry output, weighted by the share of total output, is a good way to show how the rate of diffusion affects industry-wide productivity.

Having seen data that there is substantial variation in the rate of diffusion and that a Salter graph displays this variation, we are ready to explain why we see industries with mixes of technologies. We answer two questions:

1. Why is a machine that works sometimes kept (so it is outmoded) and other times scrapped (so it is obsolete)?

2. What determines the rate of diffusion of technical change?

1. Outmoded versus Obsolete?

We assume that new technologies are being constantly generated in all industries, but some are adopted more quickly. Why is that? Why are some factories and technologies quickly replaced while others remain online? Salter's work pointed to an easily overlooked element: the *cost structure* of the firms in an industry.

STEP Proceed to the *Output* sheet. The opening situation is depicted in Figure 12.15.

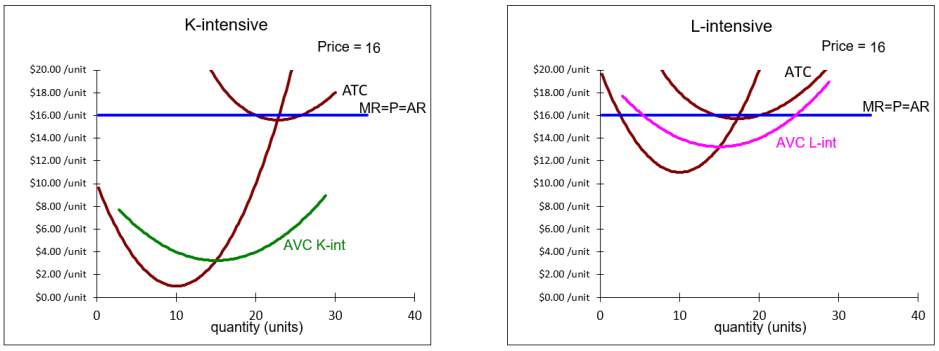


Figure 12.15: Explaining why machinery is outmoded versus obsolete. *Source: DiffusionTechChange.xls!Output.*

The graph shows two firms, one that is labor intensive and the other capital intensive. The capital intensive firm has a larger gap between ATC and AVC because it has higher fixed (capital) costs. The much lower AVC curve will prove to be critical.

Both firms in Figure 12.15 are earning small, but positive economic profits. As time goes by, however, new technologies are introduced and incorporated in newly built factories with shiny, modern equipment. The products from firms with the newest factories with their best-practice methods (the left-most rectangle in a Salter graph) can be made more cheaply so competitive pressure drives the price down.

STEP Click on the scroll bar to lower the price.

Since you know the Shutdown Rule, it is easy to see that the L-intensive firm will shut down first. As soon as you make $P < AVC$, the factory is obsolete and taken offline. The factory on the left will survive as an outmoded technology that is still in operation for much longer. You will have to keep driving the price down for much longer to see it shut its doors.

All firms use the same Shutdown Rule, but differing cost structures is what makes some factories stay in production while others close down.

So, to directly answer the question, Why is a machine that works sometimes kept (so it is outmoded) and other times scrapped (so it is obsolete)? Because the Shutdown Rule, $P < AVC$, determines the difference between outmoded and obsolete technology. Old plants that are kept online, using outmoded machines, operate in an environment in which profits may be negative, but

$P > AVC$. These plants will remain in operation as long as revenues cover variable costs. Once $P < AVC$, we know the machines will be scrapped and become obsolete as the factory is closed down.

2. What Does the Rate of Diffusion Depend On?

Figure 12.15 shows that the firm's cost structure is one of the factors which determine the rate of diffusion of technical change. Industries with capital intensive production and low variable costs will have slow rates of diffusion because plants and technologies will remain online until $P < AVC$.

Steel is a good example of such an industry. Old factories remain in production alongside modern mini-mills. The Salter graph looks like the right panel in Figure 12.14 and the cost structure is given by the left panel in Figure 12.15.

On the other hand, industries who produce in a way that labor is dominant and fixed costs are low will see rapid rates of diffusion of new methods. Legal services are a good example. Cost curves look like the right panel in Figure 12.15 so when new computers and information systems (such as LexisNexis) are developed, they are rapidly adopted and old ways are discarded. Thus, the Salter graph looks like the left panel in Figure 12.14.

Another factor affecting the rate of diffusion is the speed at which price falls. Competition among firms can be intense or muted. If, for example, the government protects an industry from foreign competition with trade barriers, preventing price from falling, the rate of diffusion of new technology and growth of labor productivity are retarded. This has certainly played a role in the rate of diffusion in the steel industry.

So, what determines the rate of diffusion of technical change? There are three factors:

1. New ideas and inventions from research and development (R&D): This is the creativity of the society. Curiosity and willingness to experiment produce a stream of better methods. The faster the flow, the better.

2. The cost structure of the firm: Capital intensive industry with high fixed and low variable costs retards diffusion of new technology. The new ideas are there, but the old ways stay online.

3. The speed at which price falls: If it is slow, we get slow diffusion. We want to encourage competition so price puts pressure on outmoded methods and drives them to be obsolete.

The first factor is the obvious one that everyone thinks of when explaining why technology affects labor productivity and economic growth. Innovation is the implementation of invention—new ideas are the raw material which expand the production function.

But Salter identified another crucial factor: Even if new technology exists, it will be mixed with existing technology and the rate at which it is adopted will depend on the Shutdown Rule. Highly capital intensive industries with low AVC will feel the drag of old technology for a long time because the gap between ATC and AVC will be great. Old methods will stay outmoded as long as $P > AVC$.

The Shutdown Rule compares average variable cost to price. Both matter. Low AVC will keep old methods around, but so will slow decline in P. Although economists usually defend free trade policies on the basis of comparative advantage, this analysis points to another reason for allowing foreign competition in domestic markets. As price is pushed down, firms are forced to modernize, taking old methods offline and investing in the newest technology. Steel tariffs are an example.

You might be confused about the claim that competition makes price fall as time goes by. It seems like inflation, prices rising, is the usual state of affairs. The explanation lies in the difference between real and nominal price.

In nominal terms, also known as current prices, the price of a light bulb is definitely higher today than 10 years ago and much higher than 100 years ago.

But in this application, the correct price to consider is the real price, in terms of actual input use. In real terms, the price of lighting is incredibly lower today. Figure 12.16, created by Nobel Prize winner William Nordhaus, tells an amazing story. In terms of the number of hours of work needed to buy 1,000 lumen hours, the price of light went from incredibly expensive for thousands of years to a free fall since the 1800s. In terms of input use, as technology improves, costs and, therefore, price of the output fall over time.

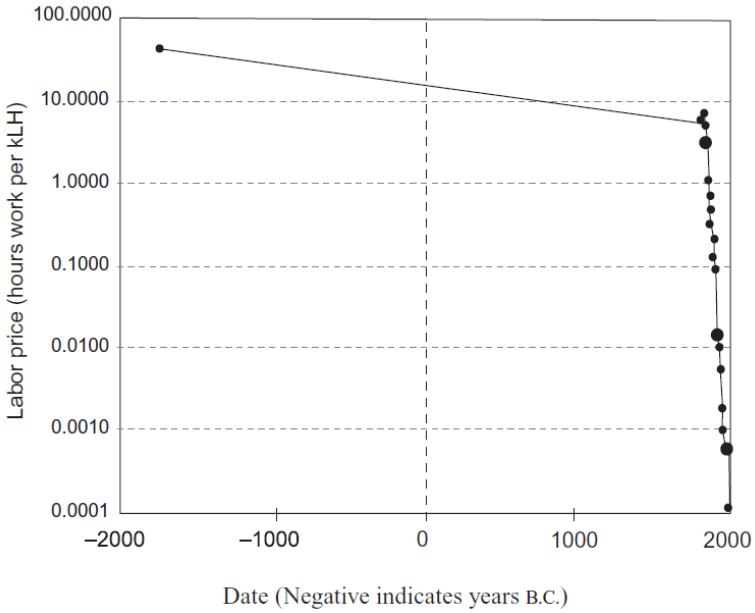


Figure 12.16: Labor price of light: 1750 b.c. to present.
Source: Bresnahan and Gordon (eds.), 1997, p. 54.

Nordhaus argues that "price indexes can capture the small, run-of-the-mill changes in economic activity, but revolutionary jumps in technology are simply ignored by the indexes" (Bresnahan and Gordon, eds., 1997, p. 55). Thus, the real price of lighting, in terms of the labor used, keeps falling and falling as time goes by.

It Is Diffusion, not Discovery, that Really Matters

Wilfred Edward Graham Salter was an Australian economist born in 1929. His promising career was tragically cut short when he died in 1963 after battling heart disease. His dissertation, finished in 1960, was published by Cambridge University Press as *Productivity and Technical Change* and was met with wide acclaim.

Salter was amazed by the ability of markets to incorporate new technology to increase output per person. He realized that scientific knowledge, technology "on the shelf," is not the only or even the most important driver of rapid growth. The new technology has to be implemented, actually used in production, and the faster it is adopted, the faster the economy grows.

Salter's primary contribution was in showing that the rate of diffusion varies tremendously and depends on the cost structures of firms. Industries with high fixed and low variable costs have large $ATC - AVC$ gaps that imply long time spans for outmoded technology.

We want nimble, adaptive firms and startups that challenge established titans. Replacing old with new machinery is necessary for rising productivity. Economies with ossified, rigid institutions are stagnant. There was a silver lining after Germany and Japan's factories were destroyed during World War II. The latest, greatest technology could be used to make all of an industry's output and productivity increased rapidly.

Exercises

1. Sometimes a best practice investment is quickly leapfrogged by newer technology. Google "fiber optic overinvestment" to see an example. Briefly describe what happened and cite at least one web source.

2. Automobile emissions requirements are stricter in Japan than in the United States (where many areas have no vehicle inspection at all). In both countries, newer cars pass inspection (if required) easily, but older cars are more likely to fail inspection and be removed from the operating car fleet. Draw hypothetical Salter graphs, with emissions on the y axis, for the car fleets of Japan and the United States that reflect the stricter emissions standards in Japan.

3. What happens to a late model year Toyota or Honda that has failed an emissions inspection in Japan and, therefore, cannot be used there? Google "Japan used engines" to find out. What effect does this have on the United States Salter graph that you drew above?

4. The National Highway and Traffic Safety Administration maintains a data base of car characteristics by model year. For miles per gallon (MPG) performance, they show the following:

2017	2016	2015	2014	2013	2012	2011	2010	2009	2008	2007	2006	2005	2004
39.2	37.3	37.2	36.3	36.1	34.8	32.7	33.1	32.1	31.2	30.6	30.3	30.5	29.9

Figure 12.17: MPG by year for US domestic passenger cars.
Source: one.nhtsa.gov/cafe_pic/CAFE_PIC_fleet_LIVE.html.

These data cannot be used to show a Salter graph (with MPG on the y axis) of the US car fleet. Why not? What additional information is needed?

References

The epigraph is from page 183 of Jared Diamond, *Guns, Germs, and Steel: The Fates of Human Societies* (W. W. Norton & Company, originally published in 1997). Diamond argues that geography determines historical development. It is not the people, but fortunate geographical circumstances that guaranteed that western Eurasian societies would become disproportionately powerful. It is geography that enabled the rapid diffusion of technology and knowledge. Diamond, like Salter, is concerned with a point that is easily missed—diffusion is more important than discovery. Visit www.pbs.org/gunsgermssteel for the documentary.

The primary source for the application in this chapter is W. E. G. Salter, *Productivity and Technical Change* (Cambridge University Press; 1st edition, 1960; 2nd edition, 1966; 1st paperback edition, 1969).

For more on technological change and the spread of new ideas, see Timothy F. Bresnahan and Robert J. Gordon, *The Economics of New Goods* (The University of Chicago Press, 1997) and David Warsh, *Knowledge and the Wealth of Nations: A Story of Economic Discovery* (W. W. Norton & Company, 2006).

Richard Preston's *American Steel* (Simon and Schuster, 1991) tells the story of a mini-mill in rural Indiana that uses German cold-casting technology. It is an entertaining tale of entrepreneurship—a billion dollar gamble—and an introduction to the exciting world of business management.

Chapter 13

Input Profit Maximization

Initial Solution

Deriving Demand for Labor

What is the Marginal Revenue Product* of Barry Bonds?

John Palmer in The Sports Economist Blog

13.1 Initial Solution

Recall that the firm's backbone is the production function. Inputs, or factors of production, typically labor (L) and capital (K) are used to make output, or product (q).

In previous chapters, we explored the firm's input cost minimization and output profit maximization problems. This chapter returns to the input side and works on the firm's third optimization problem: input profit maximization.

We continue working with a perfectly competitive (PC) firm, but we extend the assumption of perfect competition to input markets. Thus, not only is the firm one of many sellers of a perfectly homogeneous product with free entry and exit, it is also one of many buyers of labor and capital. Our firm is an output and input price taker.

This means that our PC firm only chooses the amount of input to hire, not how much to pay for it. If it has market power, then the firm not only determines how much to hire, but also gets to choose the input price. In this case, we say the firm has *monopsony power*.

While you have surely heard of monopoly, monopsony may be new to you. They are similar in that one is selling (monopoly) and the other buying (monopsony) and that means price (output or input) is no longer exogenous. A classic example is the only hospital in a small town hiring nurses. Another example is a big box retailer. Walmart is such a big buyer that they have monopsony power. They can negotiate with suppliers and extract cheaper prices from them. Notice that a firm can have both monopoly and monopsony power.

In a Labor Economics course, you study how firms can take advantage of the ability to set input prices to make greater profits. We assume this possibility

away and stay with a PC firm that takes the wage rate (w) and rental rate of capital (r) as given. Our PC firm is such a small buyer that it can hire as much L and K as it wants at the going w and r.

Setting Up the Problem

There are three parts to every optimization problem. Here is the framework for a PC firm.

1. *Goal*: Maximize profits (π), which equal total revenues minus total costs. To distinguish the input from the output side, we use the terms total revenue product (TRP) and total factor cost ($TFacC$). The idea is that labor and capital are used to make product that is sold so price times the number of units produced is the TRP.

2. *Endogenous variables*: labor and capital, in the long run; only L in the short run.

3. *Exogenous variables*: price (of the product, P), input prices (the wage rate and the rental rate of capital), and technology (parameters in the production function).

As usual, we will work with a Cobb-Douglas production function, with $\alpha > 0$, $\beta > 0$, and $\alpha + \beta < 1$.

$$q = AK^{\alpha}L^{\beta}$$

Revenues are the output price multiplied by the output produced, $TR = Pq$. We substitute the production function for q in TR to get total revenue product:

$$TRP = PAK^{\alpha}L^{\beta}$$

The units of TRP are dollars (just like total revenue). The "revenue product" language indicates that we are considering the amount of revenue ($) produced by the inputs.

The costs are simply the amounts spent on labor and capital, $wL + rK$. These are called total factor costs.

The firm chooses L and K to max profits.

$$\max_{L,K} \pi = PAK^{\alpha}L^{\beta} - (wL + rK)$$

Finding the Initial Solution

First the problem is solved using numerical methods, and then the analytical approach is used.

STEP Open the Excel workbook *InputProfitMax.xls* and read the *Intro* sheet,then go to the *TwoVar* sheet to see the problem implemented in Excel.

The sheet is named *TwoVar* because both inputs are choice variables, which means this is a long run profit maximization problem. As usual, the sheet is organized into the color-coded components of an optimization problem, with goal, endogenous, and exogenous cells.

STEP Read the description of the firm, a bakery, and scroll down to the endogenous variables.

On opening, the sheet has 500 hours of labor hired and 100 units of capital rented, yielding a profit of $936. Is this the best this firm can do? Cells B48 and B49 show the marginal revenue product of labor and marginal factor cost. By hiring one more hour of labor, revenues would rise by more than costs, so profits would increase. Clearly, therefore, this bakery is not optimizing.

STEP Run Solver to find the initial solution. Your screen should look like Figure 13.1.

Exogenous variables			
Price (P)	$	2.00	$/loaf of bread
Wage (w)	$	20.00	$/hr
Rental (r)	$	50.00	$/machine
alpha		0.20	
beta		0.75	
technology (A)		30	
Prod Fn (q)		19,086	loaves of bread

Endogenous Variables		
Labor (L)	1,431	hours
Capital (K)	153	machines

Goal			
Profit (π)	$	1,908.55	dollars
Revenue	$	38,171	dollars
Cost	$	36,262	dollars

Figure 13.1: The initial optimal solution.
Source: InputProfitMax.xls!TwoVar.

The firm hires roughly 1,431 hours of labor and rents 153 machines (but click on cells B34 and B35 to see more decimal places). This yields a maximum possible profit of just over $1,900.

Notice that the marginal revenue product and marginal factor cost cells are now exactly equal at $20/hour. This is no coincidence. The equimarginal condition for input profit maximization is that $MRP = MFC$. Since the firm is an input price taker, $MFC = w$ (just like $P = MR$ for a PC firm) so it is also true that $MRP = w$ at the optimal solution.

Finally, notice the breakdown of the firms revenues in rows 44 to 46. Labor's share (wL), capital's share (rK), and profits (whatever is left) add up to 100%. K and L's shares, 75% and 20% equal α and β. Is that a coincidence? No, that's a property of the Cobb-Douglas functional form. The exponent tells you the share of revenues that factor will receive.

We can also solve this problem via the analytical approach. We know the objective function and can substitute in each of the parameter values.

$$\max_{L,K} \pi = PAK^{\alpha}L^{\beta} - (wL + rK)$$

$$\max_{L,K} \pi = 2 * 30 * K^{0.2}L^{0.75} - (2L + 3K)$$

Next, we take derivatives with respect to L and K, set them equal to zero, and use algebra to solve the two equation system of first-order conditions.

$$\frac{\partial \pi}{\partial L} = 0.75 \cdot 2 \cdot 30\, K^{0.2} L^{-0.25} - 20 = 0$$

$$\frac{\partial \pi}{\partial K} = 0.2 \cdot 2 \cdot 30\, K^{-0.8} L^{0.75} - 50 = 0$$

We can move the 20 and 50 to the right hand side and this immediately reveals the equimarginal conditions: $MRP_L = w$ and $MRP_K = r$.

We solve the first equation for L and substitute it into the second equation to solve for optimal K. We use the rule that $(x^a)^b = x^{ab}$ to solve for L.

$$45K^{0.2}L^{-0.25} = 20$$

$$2.25K^{0.2} = L^{0.25}$$

$$\left[2.25K^{0.2} = L^{0.25}\right]^4$$

$$L = 2.25^4 K^{0.8}$$

Substitute the expression for L into the second first-order condition.

$$\begin{aligned}
&0.2 \cdot 2 \cdot 30\, K^{-0.8}\left[2.25^4 K^{0.8}\right]^{.0.75} = 50 \\
&12K^{-0.8}2.25^3 K^{0.6} = 50 \\
&K^{-0.2} = 0.365798 \\
&K^* = 152.6842
\end{aligned}$$

Compute optimal L from the expression for L.

$$L^* = 2.25^4 K^{0.8} = 2.25^4[152.6842]^{0.8} = 1431.414$$

Compute maximum profits.

$$\pi^* = 2{*}30{*}[152.6842]^{0.2}{*}[1431.414]^{0.75}{-}2{*}[1431.414]{-}3{*}[152.6842] = \$1908.55$$

This analytical solution is extremely close to Excel's solution. Practically speaking, as we would expect, the two solutions are the same.

The Short Run

A slightly different version of the firm's input profit maximization problem involves the short run when capital is not variable. By putting a bar over K, we highlight that capital is fixed.

$$\max_L \pi = PA\bar{K}^{\alpha}L^{\beta} - wL - r\bar{K})$$

We do the analytical solution first this time and in general form. There is only one derivative (since there is only one choice variable) and one first-order condition.

$$\begin{aligned}
&\frac{\partial \pi}{\partial L} = \beta PA\bar{K}^{\alpha}L^{\beta-1} - w = 0 \\
&\beta PA\bar{K}^{\alpha}L^{\beta-1} = w \\
&L^{\beta-1} = \frac{w}{\beta PA\bar{K}^{\alpha}} \\
&L^* = \left[\frac{w}{\beta PA\bar{K}^{\alpha}}\right]^{\frac{1}{\beta-1}}
\end{aligned}$$

STEP To see the numerical version of this problem, proceed to the *OneVar* sheet.

Notice that there is only one endogenous variable, L. Capital has been moved to the exogenous list because we are in the short run.

Notice also that there are two graphs. Each one can be used to represent the initial solution.

Below the graphs, you can see that the marginal revenue product of labor does not equal the wage. As you know, this means you need to run Solver because the firm is not optimizing.

STEP Run Solver to find the initial solution. Your screen should look like Figure 13.2.

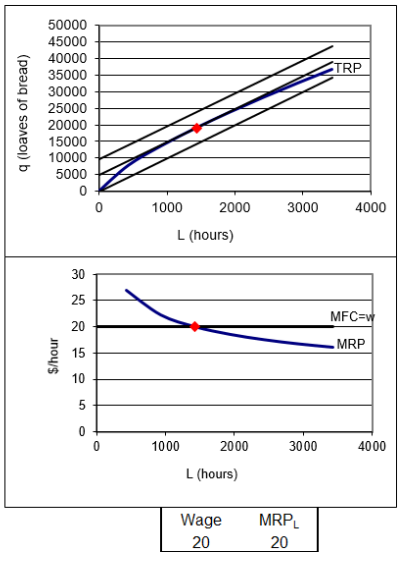
Exogenous variables			
Price (P)	$	2.00	$/loaf of bread
Wage (w)	$	20.00	$/hr
Rental (r)	$	50.00	$/machine
alpha		0.20	
beta		0.75	
technology (A)		30	
Capital (K)		153	machines
Prod Fn (Y)		19,085	loaves of bread

Endogenous Variables			
Labor (L)		1,431	hours

Goal			
Profit (π)	$	1,909	dollars
Revenue	$	38,171	dollars
Cost	$	36,262	dollars

Calculating Distribution of Revenues

		Dollar Payments	Percentage of Total
Labor Share	$	28,628	75%
Capital Share	$	7,634	20%
Profit Share	$	1,909	5%



Wage	MRP_L
20	20

Figure 13.2: The initial optimal solution in the short run.
Source: InputProfitMax.xls!OneVar.

The bottom graph shows that the optimal labor use can be found where the marginal revenue product of labor (the curve) equals the wage (at \$20/hr). This is the canonical graph for the input side profit maximization problem. Like $MR = MC$ on the output side, the intersection of the two marginal relationships instantly reveals the optimal solution.

The top graph is a different way of viewing the exact same problem. It is using the production function as a constraint (the *TRP* curve) and three representative *isoprofit* lines are displayed. Each isoprofit line shows the combination of L and q that gives the same profit. The firm is trying to get on the highest isoprofit (to the northwest) while meeting the constraint. It can roll on the *TRP* curve (like it rolled on the isoquant) until it hits an isoprofit line that is tangent to the *TRP*.

The constrained optimization problem can be written like this:

$$\max_{L,q} \pi = Pq - wL - r\bar{K}$$
$$\text{s.t. } q = A\bar{K}^{\alpha}L^{\beta}$$

The Lagrangean method could be applied to solve this problem. Naturally, the exact same solution is obtained if we use the Lagrangean or the more common approach of directly substituting the constraint (the production function) into the revenue function.

Suppose we wanted to check if the analytical and numerical results are the same. We need to evaluate the expression for optimal L at the parameter values in the *OneVar* sheet.

The expression is complicated enough that entering it in a cell as you would write it is a bad idea. The parentheses are likely to cause confusion. It is better to create *houses* for each part then fill them in. Here's how.

STEP Watch this short video on how to enter a complicated formula in Excel: vimeo.com/415967747.

Entering parentheses as pairs, is a good habit to develop when working in a spreadsheet. It is easy to make an order of operations mistake or get mismatching parentheses if you try to enter the formula like you would on a piece of paper.

STEP Enter the formula in cell M28 (just like in the video) to practice building houses in formulas in Excel.

In so doing, you confirm that the analytical and numerical methods yield substantially the same answer.

Another Short Run Production Function

A Cobb-Douglas production function has many advantages, including that the sum of exponents reveals whether returns to scale are increasing, constant, or decreasing if they are greater, equal, or less than one. However, once the exponents are set, the function can only exhibit those returns to scale.

Likewise, in the short run, with K fixed, our Cobb-Douglas functional form showed the Law of Diminishing Returns because $\beta = 0.75$. A more flexible functional form would enable production to have increasing and diminishing returns as more labor is added.

Like the cubic polynomial we used for the total cost function, a cubic functional form can give us an S-shaped TRP curve.

$$TRP = aL^3 + bL^2 + cL$$

STEP Proceed to the *Graphs* sheet to see this functional form implemented in a set of four graphs that can be used to represent the firm's input profit maximization problem (Figure 13.3).

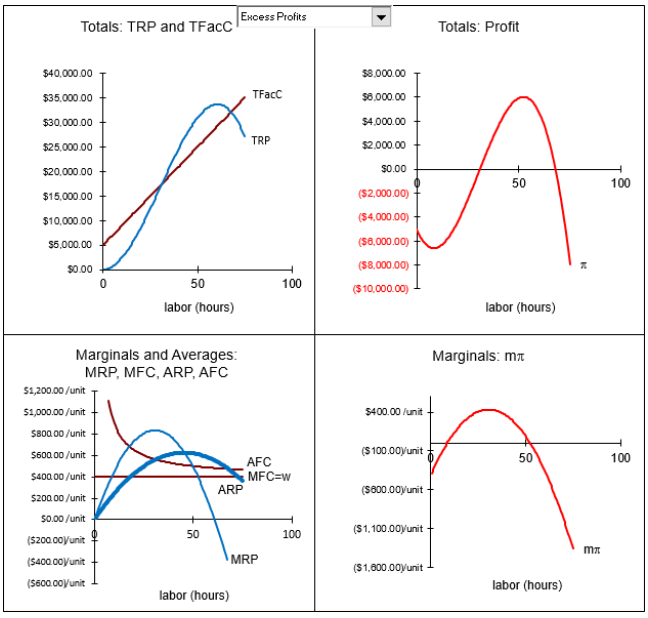


Figure 13.3: Four graphs for input profit maximization.
Source: InputProfitMax.xls!Graphs.

It is striking that these graphs mirror the four graphs we used to describe the firm's output side profit maximization problem. The two top graphs show total revenue and total cost on the top left, along with total profits on the top right. The bottom graphs display a series of marginal and average curves on the bottom left and marginal profit on the bottom right.

If you look carefully, you will notice that things are switched around a bit. Instead of total cost being a curve (as it is on the output side), it is a straight line because total factor cost on the input side in the short run is $wL + r\bar{K}$. On the other hand, total revenue product (so named to distinguish it from total revenue on the output side) is a curve (instead of a straight line).

Unlike the canonical output side profit maximization graph with U-shaped MC, ATC, and AVC curves and a horizontal $P = MR$ line, the bottom left graph has a horizontal MFC line and the MRP and ARP functions are curves and they are upside down.

But there are also key similarities. The equimarginal rule is in play: $MFC = MRP$ reveals the labor use that maximizes profits. Also, a rectangle of $(ARP - AFC)L$ gives an area that is equal to profits. The length of the profit rectangle ranges from zero to the chosen amount of labor hired. The height is the difference between average revenue product, ARP, and average factor cost, AFC. The area of this rectangle is profit because $ARP - AFC$ is profit per hour so multiplying by L, measured in hours, yields profits. Another way to think about this is that multiplying L by ARP yields total revenues (since $L * TRP/L = TRP$) and multiplying L by AFC gives total costs (since $L * TFacC/L = TFacC$). Subtracting the total cost rectangle from the total revenue rectangle leaves the profit rectangle.

Another similarity between output and input profit maximization is that the firm has the same four profit positions.

STEP In the *Graphs* sheet, click on the pull down menu (near cell P4) and cycle through all of the profit positions.

As with the output side, the shock is output price. As it falls, so do maximum profits.

The *Neg Profits, Cont Prod* and *Neg Profits, Shutdown* options show that the firm will shut down when the $w > ARP$. This is analogous to the $P < AVC$ Shutdown Rule. Keep your eye on the total profits in the top right graph

to see that the story is the same—the firm is deciding whether the negative profit at best of the positive levels of L is better than hiring no L at all.

The connection between input and output is simple. The firm shuts down when $w > ARP$ which we can multiply by L to give $wL > TRP$. But wL and TRP are TVC and TR on the output side. Divide both by q and we get $AVC > P$, which is the same as $P < AVC$, the usual output side Shutdown Rule. In addition, the $wL > TRP$ version of the Shutdown Rule supports the claim that revenues must cover variable costs for a firm to produce.

Input Profit Maximization Highlights

At this point, you might be suffering from repetitive stress syndrome—we seem to be going over and over the same ideas. That is an important level to attain in mastering the economic way of thinking. The body of knowledge in economics is grounded in a core methodology of optimization and comparative statics. The framework is used over and over and over again.

Like every optimization problem, the input side profit maximization problem can be organized into a goal, endogenous, and exogenous variables. This problem has a canonical graph (with MFC and MRP as the key elements) and an equimarginal rule $MFC = MRP$.

Because the firm is an input price taker, $MFC = w$. This means that every additional hour of labor adds w to total cost. If the firm was a monopsony, this would not be true and the optimization problem would be more complicated.

Finally, because the input profit maximization problem is the flip side of the output side profit maximization problem, it should not be surprising that we can represent the initial solution with a set of four graphs. The parallelism carries through all the way to the Shutdown Rule, where $w > ARP$ is equivalent to $P < AVC$. We will stress the connections between input and output side again in the next chapter.

Exercises

1. Use the *TwoVar* sheet to compute the long run beta elasticity of L^* from beta = 0.75 to beta = 0.74. Show your work.

2. In the *Q&A* sheet, question 4 asks you to find short run beta elasticity of L^* from beta = 0.75 to beta 0.74. The *InputProfitMaxA.doc* file in the *Answers* folder shows that the answer is about 28. Explain why the short run elasticity (which is admittedly quite large) is much smaller than the long run elasticity that you computed in the previous question.

3. Use Excel to set up and solve (with Solver, of course) the constrained version of the input profit maximization problem in the *OneVar* sheet. Take a screenshot of your solution (including the constraint cell) and paste it in your Word document.

4. In the *Graphs* sheet, select the *Neg Profits, Shutdown* case. Does the top, right graph support the $w > ARP$ Shutdown Rule? Explain.

References

The epigraph, from John Palmer at thesportseconomist.com/what-is-the-marginal-revenue-product-of-barry-bonds, points to two avenues for further reading: sports economics and blogs.

The worlds of economics and sports are increasingly intertwined. There are courses, conferences, and journals dedicated to the economics of sports. For a classic paper on baseball, see Simon Rottenberg's "The Baseball Players' Labor Market," *The Journal of Political Economy*, Vol. 64, No. 3 (June, 1956), pp. 242–258, www.jstor.org/stable/1825886.

There are, of course, many blogs dedicated to economics. The marginalrevolution.com and cafehayek.com are often informative and entertaining. For macroeconomics, see Greg Mankiw at gregmankiw.blogspot.com and Brad DeLong at delong.typepad.com. John Cochrane will give you a free market perspective at johnhcochrane.blogspot.com—plus, *The Grumpy Economist* is a great name for a blog.

> To be sure, we are living in a dessert age. We want things to be sweet; too many of us work to live and live to be happy. Nothing wrong with that; it just does not promote high productivity. You want high productivity? Then you should live to work and get happiness as a by-product.
>
> David Landes

13.2 Deriving Demand for Labor

A profit-maximizing firm with Cobb-Douglas technology and given prices in all markets (P, w, and r) in the short run can be modeled as solving the following optimization problem:

$$\max_L \pi = PA\bar{K}^{\alpha}L^{\beta} - wL - r\bar{K})$$

The previous section found the initial solution for this problem. This section is devoted to comparative statics analysis. How will this firm respond to a change in one of its exogenous variables, ceteris paribus?

Although there are several exogenous variables from which to choose, the responsiveness of optimal L to a change in the wage is of utmost importance. This comparative statics analysis will give us the short run demand for labor.

After deriving the demand for labor in the short run, we will examine the long run demand for labor. A comparison of short and long run wage elasticities of labor reveals that labor demand is more responsive in the long run. We then explore how changes in P affect L^*.

Demand for Labor in the Short Run

We begin with numerical methods for a comparative statics analysis of a change in the wage (also called the wage rate is measured in \$/hr).

STEP Open the Excel workbook *DerivingDemandL.xls* and read the *Intro* sheet, then go to the *OneVar* sheet.

The layout is the same as the *InputProfitMax.xls* workbook in the previous section. It is clear from the graphs and the equivalence of wage and *MRP* below the graphs that the firm is at its optimal solution. The yellow-backgrounded cell, the wage rate, is the shock variable on which we will focus.

STEP Change the wage in the *OneVar* sheet to \$19/hr from the initial value of \$20/hr.

It is difficult to see anything in the top graph, however, the isoprofit line is no longer tangent to the *TRP*. The bottom graph clearly shows that the red diamond (at $L = 1431$ hours) has a marginal revenue product greater than the marginal factor cost (equal to the wage). Cells H40 and I40 show that the wage is less than *MRP*.

STEP Since the firm is no longer optimizing, run Solver to find the new optimal solution.

You will find that, to maximize profits, the firm will hire 1757 hours when the wage falls to \$19/hr, ceteris paribus. At this level of labor use, the marginal revenue product once again equals the marginal factor cost.

Although we have only two data points, it should be clear that the firm will hire that amount of labor where the marginal revenue product equals the wage, in the short run. This means that *the marginal revenue product curve is the firm's (inverse) demand for labor curve.* Quote the firm a wage and it will look to its *MRP* curve to decide how much labor to hire.

We have two points on the demand for labor curve; at $w = \$20$/hr, $L^* = 1431$ hours and at $w = \$19$/hr, $L^* = 1757$ hours. Can we pick more points off of the demand for labor curve?

STEP Set the initial wage back to \$20/hr and use the Comparative Statics Wizard to apply five \$1/hr decreases in the wage. Create charts of the demand for labor and the inverse demand for labor.

Your results should look like those in the *CS1* sheet. The CSWiz output makes common sense. As the wage drops, the firm hires more labor. Look also at the objective function—as wage falls, maximum profits are rising. The key idea here is that firm hiring decisions are driven by profit maximization. The reason why L increases as w falls is that this response is profit maximizing.

Like demand curves in the Theory of Consumer Behavior, the price—the wage in this case—can be placed on the x or y axis. The two displays use the same information and convey the same message.

We can also derive the short run demand for labor via analytical methods. This problem was presented in the previous section. For your convenience, it is repeated below.

We need to leave w as a variable, but for maximum generality we solve for L^* as a function of all parameters.

$$\max_L \pi = PA\bar{K}^\alpha L^\beta - wL - r\bar{K})$$

We take the derivative with respect to L, set it equal to zero, and solve for L^*.

$$\begin{aligned}
\frac{\partial \pi}{\partial L} &= \beta PA\bar{K}^\alpha L^{\beta-1} - w = 0 \\
\beta PA\bar{K}^\alpha L^{\beta-1} &= w \\
L^{\beta-1} &= \frac{w}{\beta PA\bar{K}^\alpha} \\
L^* &= \left[\frac{w}{\beta PA\bar{K}^\alpha}\right]^{\frac{1}{\beta-1}}
\end{aligned}$$

This expression is the demand curve for labor. If we substitute in values for all exogenous variables except w, we can plot L^* as a function of w, ceteris paribus.

Do the numerical methods based on the CSWiz add-in agree with the analytical derivation of the demand for labor?

STEP In the *CS1* sheet, click on cell C16. This is Solver's answer for L^* when the wage is \$20/hr.

Do not be misled by all of the decimal places. That is false precision.

STEP Click on cell E26. It displays L^* when the wage is \$20/hr based on the reduced-form solution.

Do not be misled by the number displayed in cell E26. This is Excel's display for the formula entered into that cell. Excel's memory has a different number.

STEP Widen column E to see more decimal places.

We proceed slowly because things can get confusing here. Consider this hierarchy of truth:

1. Solver is giving a number close to the exact right answer in cell C16.
2. Excel is representing the exact right answer as a decimal in cell E26.
3. The exact right answer is $\frac{w}{\beta PA\bar{K}^{\alpha}}^{\frac{1}{\beta-1}}$ evaluated at $w = \$20/\text{hr}$, along with the other parameter values.

STEP To see that E26 is not the exact answer, make column E very wide, then select cell E26 and click Excel's *Increase Decimal* button repeatedly.

You will see that, eventually, Excel will start reporting zeroes. Excel has finite memory and, therefore, it cannot compute an infinite number of decimal places for the exact answer. The decimal representation of the exact answer stored in Excel's memory is not the exact answer.

To be clear, Excel can display the exact answer if it is an integer or fraction that can be represented with finite memory. For example, $\frac{x}{7}$, evaluated at $x = 14$ is 2 so, no problem for Excel. If 2 is the answer, Excel has it exactly right. Evaluating at $x = 1$ means there is no decimal representation with a finite number of digits. Excel cannot display the exact answer in this case. Enter $= 1/7$ in a cell, widen the column, and click the *Increase Decimal* button repeatedly to see that Excel eventually starts showing zeroes.

Thus, neither E26 nor C16 is the exact answer. They are both so close to the answer, however, that we can say they "substantially agree" and are correct.

We can also use the analytical approach to reinforce the idea that the short-run (inverse) demand for labor is the marginal revenue product of labor.

The first-order condition gives the equimarginal rule.

$$\frac{d\pi}{dL} = \beta PA\bar{K}^{\alpha}L^{\beta-1} = w$$

The term on the left is the *MRP*. Evaluating the $\beta PA\bar{K}^{\alpha}$ portion at their initial values gives 123.0187 (as shown in cell K26 of the *CS1* sheet). Thus, $MRP = 123.0187L^{\beta-1}$ and at $\beta = 0.75$, $MRP = 123.0187L^{0.25}$.

The *CS1* sheet has an inverse demand for labor chart. Is the relationship in this chart the same as the *MRP* function that we just found? Let's find out. By finding the function that fits the data in the inverse demand for labor chart, we can compare this relationship to the *MRP* function.

STEP Right-click on the series in the inverse demand for labor chart and select the *Add Trendline* option. Select the *Power* fit, scroll down and check the *Display equation on chart* option. Click OK. Move the equation (if needed) and increase the font size to see it better. Scroll right to see what your chart should look like.

The answer is clear: The fitted curve that reveals the function for the inverse demand curve for labor is the marginal revenue product of labor curve. The fitted curve's coefficient and exponent are almost exactly that of the *MRP*.

Next, we turn our attention to the wage elasticity of labor demand. We can compute the elasticity at a point or from one point to another. We do the former below and leave the latter as an exercise question.

Elasticity at a point begins by finding the derivative of the reduced-form expression. We substitute in the known value for $\beta P A \bar{K}^{\alpha} = 123.0187$ in the denominator and $\beta = 0.75$ in the exponent.

$$L^* = (\frac{w}{\beta P A \bar{K}^{\alpha}})^{\frac{1}{\beta - 1}} = (\frac{w}{123.0187})^{\frac{1}{0.75-1}} = (\frac{w}{123.0187})^{-4}$$

To take the derivative with respect to w, we isolate w.

$$L^* = (\frac{w}{123.0187})^{-4} = \frac{w^{-4}}{123.0187^{-4}} = (\frac{1}{123.0187^{-4}})w^{-4}$$

Now we can apply our usual derivative rule, moving the exponent to the front and subtracting one from it.

$$\frac{dL^*}{dw} = -4(\frac{1}{123.0187^{-4}})w^{-5}$$

This expression is merely the slope or instantaneous rate of change of optimal labor hired as a function of the wage. To find the elasticity, we must multiply the derivative by the ratio w/L.

$$\frac{dL^*}{dw}\frac{w}{L} = -4(\frac{1}{123.0187^{-4}})w^{-5}\frac{w}{L}$$

But we have an expression for L, so we substitute it in.

$$\frac{dL^*}{dw}\frac{w}{L} = -4(\frac{1}{123.0187^{-4}})w^{-5}\frac{w}{(\frac{1}{123.0187^{-4}})w^{-4}}$$

The 123.0187^{-4} terms cancel. And w^{-5} times w in the numerator is w^{-4} so that cancels with w^{-4} in the denominator. We are left with this.

$$\frac{dL^*}{dw}\frac{w}{L} = -4$$

As has happened before (remember the price and income and cross price elasticity of demand?), the Cobb-Douglas functional form produces a constant wage elasticity of short run labor demand.

This elasticity value says that labor demand is extremely responsive to changes in the wage. We would not expect to find such a large wage elasticity of short-run labor demand in the real world. For a Cobb-Douglas production function, the elasticity is driven by the value of beta. If we had left β in the expression for optimal L instead of using 0.75 (see the first two exercise questions), we would get this expression for the wage elasticity of labor demand:

$$\frac{dL^*}{dw}\frac{w}{L} = \frac{1}{\beta - 1}$$

If we compute the elasticity from one point to another, say from a wage of \$20/hr to \$19/hour (see exercise question 3), we will get a different answer than -4. That makes sense since we know that L^* is non linear in w. As the change in the wage approaches zero, the elasticity computed from one point to another approaches -4.

13.2.1 Demand for Labor in the Long Run

If we relax the assumption that capital is fixed, we change the firm's planning horizon from short to long run. The *TwoVar* sheet implements the firm's long run input profit maximization problem. There are two endogenous variables, labor and capital, and no fixed factors of production.

STEP To derive the firm's long run demand for labor, use the Comparative Statics Wizard from the *TwoVar* sheet. As you did in the short run analysis, apply \$1 decreases in the wage.

Your results should show labor use rising as wage falls, just as in the short run. But what about the elasticity—is it the same in the short and long run?

STEP Use your CSWiz results to compute the wage elasticity of labor demand from a wage of \$20/hr to \$19/hr. Is it close to -4, the point elasticity at w=\$20/hr?

The *CSCompared* sheet is similar, but not the same as your results. It shocks wage by \$1/hr increments in the short and long run.

The difference in the elasticity is dramatic—labor demand is incredibly responsive in the long compared to the short run. The elasticity almost triples, from -3.5 to almost -11. You should find the same result with your CSWiz data for a wage decrease—the long run elasticity is much higher (in absolute value) than in the short run. What is going on?

Figure 13.4 provides an answer to this question. The movement from point A to B is the short run response for a \$1/hr wage increase. As the short run results in the *CSCompared* sheet show, when the wage rises from \$20/hr to \$21/hr, L^* falls from roughly 1,431 hours to 1,178 hours.

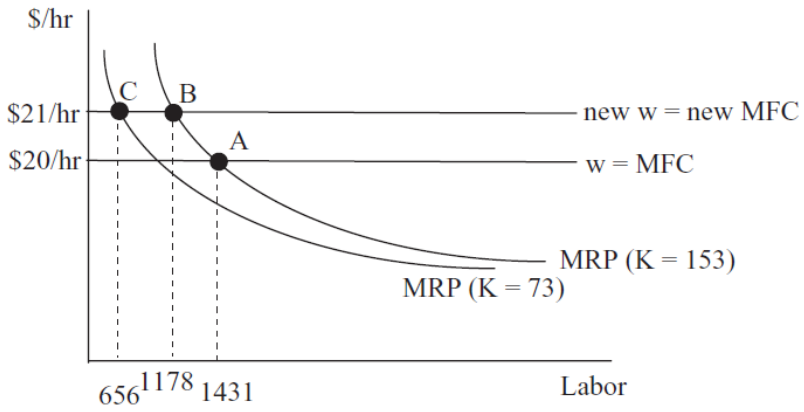


Figure 13.4: Why L^* is more responsive to Δw in the long than short run.

In the short run, capital stays fixed and the firm moves along its marginal revenue product curve (which as we already know is the firm's short run demand for labor) as the wage changes. The $K = 153$ in the parentheses signals that this is the value of K for this MRP schedule.

In the long run, however, the adjustment is different. The data in the *CSCompared* sheet show clearly that the firm will change both labor and capital as the wage rises. Notice that capital falls from 153 machines to 73 machines as the wage rises from \$20/hr to \$21/hr.

This change in capital shifts labor's marginal revenue product curve. As shown in Figure 13.4, the firm's long run response to the change in the wage is from A to C, not simply A to B. It decreases labor use as it moves along the initial *MRP* and then again when *MRP* shifts as K falls. This is the reason why the wage elasticity of labor demand is more responsive in the long run.

Figure 13.5 shows the firm's long run demand for labor and that it is no longer the *MRP* curve. Because capital falls as wage rises, leading to a further decrease in labor hired, the firm is much more responsive to changes in the wage.

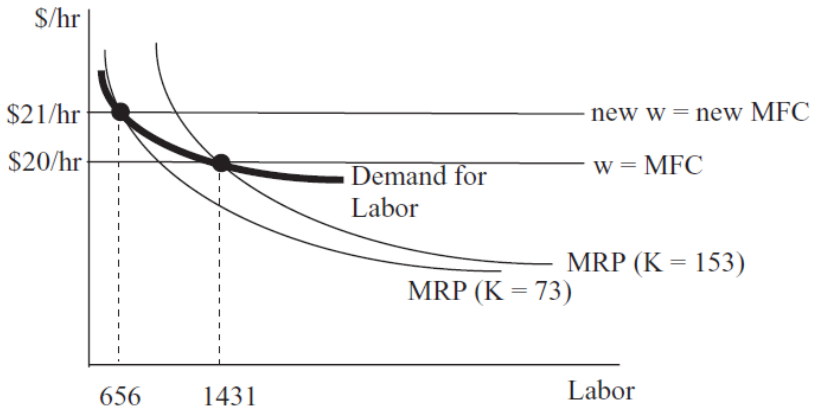


Figure 13.5: The long run demand for labor.

It is clear that the inverse labor demand curve shown in Figure 13.4 is flatter in the long run than the *MRP* curve (which is the short run inverse demand for labor). A wage decrease would stimulate more labor hired in the long than short run because K would rise in the long run.

The Shutdown Rule and the Demand Curve for Labor

Recall that, on the output side, the supply curve is the MC curve when $P > AVC$.If $P < AVC$ where $MR = MC$, then the firm ignores this marginal signal (which is the top of a local profit hill) and shuts down ($q = 0$).

The supply curve has a tail where the quantity supplied is zero when the price falls below average variable cost.

There is a similar tail, with $L = 0$, on the demand curve for labor. The previous section showed that if $w > ARP$, the firm will shut down, hiring no labor and producing no output.

STEP Proceed to the *Graphs* sheet to quickly review this concept. Use the pull down menu to change the firm's output price and place the firm in any of the four profit positions. Select *Neg Profits, Shutdown* to see that the firm will shut down when P is so low that it shifts ARP down so much that $w > ARP$. This is analogous to the $P < AVC$ Shutdown Rule.

The Shutdown Rule means that we have to change our definition of the demand curve for labor to get it exactly right. In the short run, the inverse demand curve is the MRP curve, as long as $w > ARP$; otherwise it is zero, as shown in Figure 13.6.

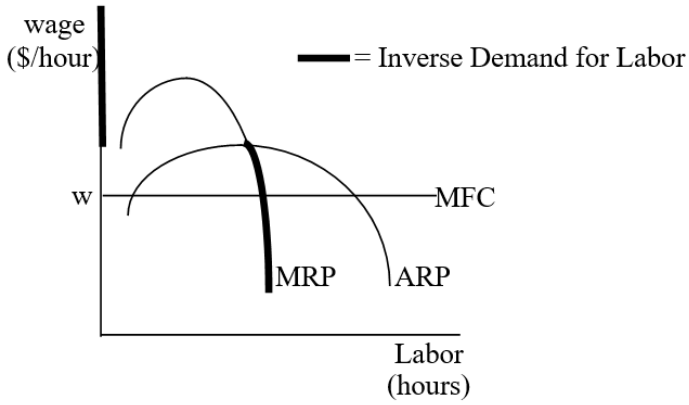


Figure 13.6: The short run inverse demand for labor.

The Shutdown Rule is usually presented from the output side as $P < AVC$. This version of the rule is perfectly compatible with the input side version of the shutdown rule, $w > ARPL$. Either wage increases or output price decreases can trigger a shutdown.

In Figure 13.6, it is easy to see what is happening when wage increases—the horizontal MFC line shifts up and it rises above ARP, the firm shuts down. What is happening on the output side? Remember that as wage rises, cost curves on the output side shift up. At the precise point at which a higher

wage triggers the decision to not hire any labor, the AVC curve will have shifted above P and the firm will decide to not produce any output.

The same story is at work when P falls. On the output side, it easy to see that when the horizontal $P = MR$ line falls below AVC, the firm shuts down. What is happening on the input side? As P falls, the MRP and ARP curves in Figure 13.6 shift down. At the precise moment when P falls below AVC and the firm decides to produce no output, the ARP shift below the horizontal wage line in Figure 13.6 and the firm will decide to hire no labor.

Demand for Labor Depends on *P*

Another comparative statics analysis for input profit maximization revolves around the effect that P has on L^*. This shows how the demand for labor is a derived demand from the desirability of the product. In other words, the stronger the demand for the product, the greater the demand for labor.

Suppose demand for bread rises in our Excel workbook. This increases P, ceteris paribus. What happens to L? We explain the short run response here and leave the long run for exercise questions 4 and 5.

STEP Return to the *OneVar* sheet. Return the wage to \$20/hr. Run Solver.

Instead of simply changing P and running Solver again, we want to see what effect P has on the graphs that show the initial solution.

STEP Change P to \$2.10 and look carefully at the charts.

It is difficult to see that the TRP curve has changed so that it is no longer tangent to the isoprofit line, but the bottom chart clearly shows that the initial solution is no longer optimal. What happened?

From our analytical work, we know that $MRP = \beta P A \bar{K}^{\alpha} L^{\beta-1}$ so it is clear that an increase in P will shift the MRP curve up. That is what you are seeing in the bottom graph on the *OneVar* sheet. Return P to \$2/unit to see that MFC stays constant (w remains unchanged), but MRP is moving.

STEP With P=\$2.10, run Solver. What happens to L^*?

Not surprisingly, the firm wants to hire more labor. The reason is that the MRP curve shifts and a new solution is found where the new $MRP = w$. Labor cost and productivity are unchanged, but the demand for labor is affected by consumer's desire for the product (expressed through the P). We say that that demand for labor is a *derived demand*—the firm's need for labor (and other inputs) comes from the fact that it has customers who want its product.

Figure 13.7 shows what happens as you increase the product price. If the demand for a firm's output is high, the price will be high, and this will induce an increased demand (shift) for labor.

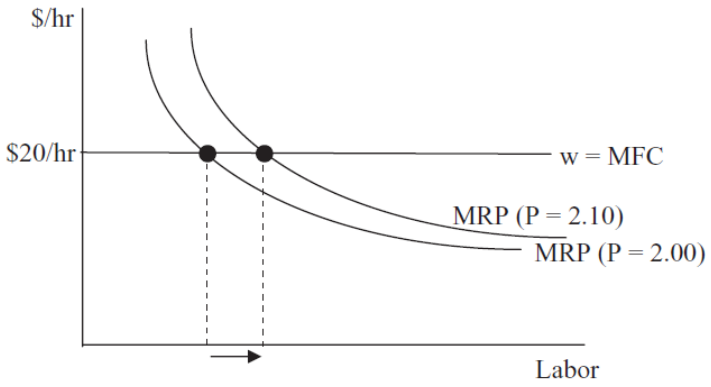


Figure 13.7: Demand for L is a derived demand.

It is easy to see that labor is a derived demand by considering professional sports. Pro athletes in major sports make a lot of money because they are in high demand. Sports teams know that the price of the good they produce (including broadcast and streaming revenue) is high. The output side is most definitely reflected in the input side via the product price.

Marginal Productivity Theory of Distribution

The input side profit maximization problem can be used to examine the distribution of firm revenues. The basic idea is that shares are a function of an input's productivity: The more productive the input, the greater its share.

STEP From the *TwoVar* sheet, run a comparative statics experiment that changes the exponent on labor from 0.75 to 0.755 (5 shocks of 0.001). In the endogenous variables input box, be sure to track not only L and K, but also the shares received in cells C44:C46.

Check your results with the *CS3* sheet. The *CS2* sheet has the outcome of a change in alpha, the exponent on capital. It explains how "large" shocks of, say, 0.1 will cause catastrophic failure as $\alpha + \beta$ approaches +1. This is why the change in beta so small—to stay away from the singularity.

By increasing the exponent on labor in the Cobb-Douglas production function, labor's productivity rises. In other words, labor can make more output, ceteris paribus, as the exponent on labor increases. The firm maximizes profit by using more labor and labor's share of firm revenues rises.

The CSWiz data show that we can immediately determine the percentage share of revenues gained by each input by the input's exponent in the production function. Although a different production function may not have this simple short-cut to determine the percentage share of revenues accruing to each input, it remains true that an input's share will depend on its marginal productivity.

Whereas algebraic convenience and simplicity are often invoked as a rationale for utilizing the Cobb-Douglas functional form, in the case of factor shares, a strong empirical regularity supports the use $AK^{\alpha}L^{\beta}$. About 2/3 of national income has gone to labor and 1/3 to capital. "In fact, the long-term stability of factor shares has become enshrined as one of the "stylized facts" of growth" (Gollin, 2002, pp. 458–459). More recent measurements of factor shares shows that capital is gaining a greater share and this is an active, exciting area of research.

Labor Demand Highlights

The most important comparative statics exercise on the input side is to derive the demand for inputs. This chapter focused on labor demand and showed that the short run demand for labor is the marginal revenue product of labor curve.

In the long run, however, the demand for labor is not the MRP curve because K^* changes as w changes. For this same reason, labor demand is more responsive to changes in the wage in the long run.

Whether in the long or short run, the demand curve for labor is subject to the same Shutdown Rule qualification as the supply curve for output. If the wage is higher than the ARP at the point at which $MRP = MFC$, the firm

will hire no labor. This coincides perfectly with the firm's decision to shut down on the output side, producing no output.

In addition to changes in the wage, this chapter explored the effects of a change in product price. As P increases, L^* rises. In terms of the canonical graph, an increase in P shifts the MRP and leads to a new optimal solution. This leads economists to think of and say that labor demand is a derived demand because the price of the product influences how much labor the firm wants.

This section ended by pointing out that an input's productivity determines its share of firm revenues. As productivity rises, so does the percentage share accruing to that input. Productivity is a key variable in determining input use and distribution of revenues.

Exercises

1. Derive the wage elasticity of short run labor demand for the general case where $L^* = (\frac{w}{\beta P A \bar{K}^{\alpha}})^{\frac{1}{\beta-1}}$. Show your work, using Word's Equation Editor.

2. Does your result from the previous question agree with the -4 value obtained in the text?

3. Compute the wage elasticity of short run labor demand (using the parameter values in the *OneVar* sheet) from w=\$20/hr to \$19/hr. Show your work.

4. Use the Comparative Statics Wizard to analyze the effect of an increase in the product price in the long run. Compute the P elasticity of L^* from $P = 2.00$ to $P2.10$. Copy and paste your results in a Word document.

5. Is L^* more responsive to changes in P in the short run or long run? Explain why.

References

The epigraph is from page 523 of David S. Landes, *The Wealth and Poverty of Nations: Why Some are So Rich and Some So Poor* (paperback edition, 1999; originally published, 1998). Landes was an economic historian interested in economic development. He asked really difficult, fascinating questions: "How

and why did we get where we are? How did the rich countries get so rich? Why are the poor countries so poor? Why did Europe ('the West') take the lead in changing the world?" (p. xxi). His answers are opinionated and clear.

The idea that a profit-maximizing firm will use and reward factors according to productivity has a normative or ethical dimension. John Bates Clark, one of the first well-known American economists, argued in *The Distribution of Wealth* (1899) that the equimarginal principle was not only efficient, but also fair. Paying factors according to productivity showed that capitalism was just. For more modern reading on morality or ethics in economics, from one end of the spectrum to the other, see Robert Nozick, *Anarchy, State, and Utopia* (1974) and John Rawls, *A Theory of Justice* (1971).

You are undoubtedly familiar with the Nobel Prize in Economic Sciences, but the John Bates Clark Medal is given every two years "to that American economist under the age of forty who is adjudged to have made the most significant contribution to economic thought and knowledge." See www.aeaweb.org/about-aea/honors-awards/bates-clark for a complete list of winners—it is peppered with Nobel Prize winners.

In his paper reconciling time series and cross section data, Douglas Gollin, "Getting Income Shares Right," *The Journal of Political Economy*, Vol. 110, No. 2 (April, 2002), pp. 458–474, www.jstor.org/stable/10.1086/338747, says that Cobb and Douglas "were among the earliest authors to point out that, for the United States, the labor share of income appeared to be roughly constant over time, regardless of changes in factor prices" (pp. 460–461). As mentioned in this section, this remarkable constancy of labor shares has crumbled recently, as labor's share has fallen. For a more recent review of labor's share, see conversableeconomist.blogspot.com/2018/02/behind-declining-labor-share-of-income.html.

[That long run responses are more elastic than short run responses] is commonly believed to be empirically true, simply as a matter of assertion. It is interesting and noteworthy that this type of behavior is in fact mathematically implied by a maximization hypothesis.

Eugene Silberberg

Chapter 14

Consistency

We have considered three separate optimization problems in our study of the perfectly competitive (PC) firm. Figures 14.1, 14.2, and 14.3 provide a snapshot of the initial solution and the key comparative statics analysis from each of the three optimization problems.

This chapter ties things together with the fundamental point that these three problems are tightly integrated and are actually different views of the same firm and same optimal solution. Change an exogenous variable and all three optimization problems are affected. The new optimal solutions and comparative statics results are *consistent*—i.e., they tell you the same thing and are never contradictory.

Figure 14.1 shows the input side cost minimization problem. Quantity is exogenous in this problem and the firm looks for the input mix that minimizes the total cost of producing the given q.

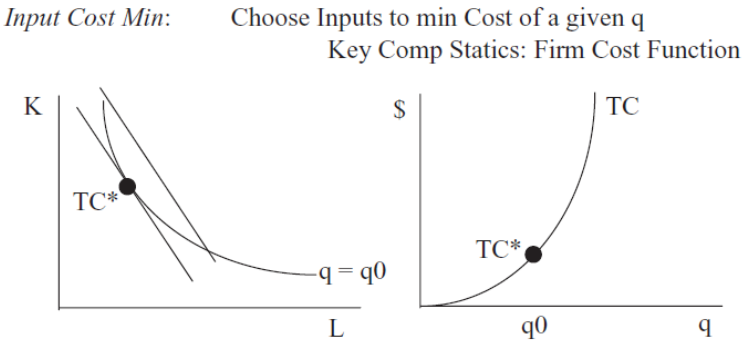


Figure 14.1: Initial cost minimization and cost function.

The right panel in Figure 14.1 shows the cost function that comes from tracking minimum total cost as q varies, ceteris paribus.

Figure 14.2 shows output side profit maximization. The PC firm (since $P = MR$ is a horizontal line) gets average and marginal cost curves from the cost function and finds the quantity that maximizes profit.

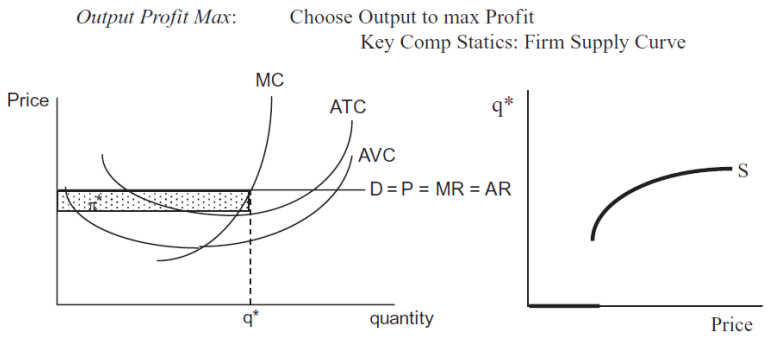


Figure 14.2: Initial profit maximization and the supply curve.

The right panel in Figure 14.2 shows where supply curves come from: shock P, ceteris paribus, and track optimal q.

Figure 14.3 returns to the input side, but this time the firm solves a profit maximization problem, choosing how much labor to hire.

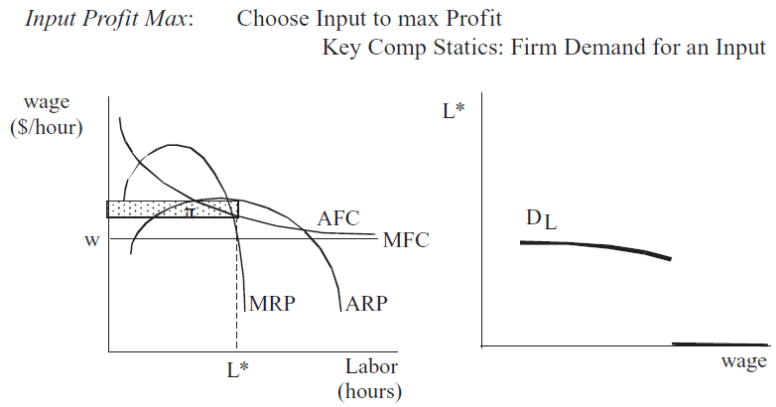


Figure 14.3: Initial profit maximization and the demand for labor.

The right panel in 14.3 shows how changing w, ceteris paribus, produces the demand curve for labor.

These three optimization problems share a common methodology. In each case, we set up and solve the problem, then do comparative statics analysis. There are other shocks that can be explored, but the one shown here is the most important.

But there is one last crucial concept that is the focus of this chapter: these three problems do not exist in isolation, instead, they are woven together to comprise the Theory of the Firm. The relationships among the three exhibit a consistency that can be demonstrated with Excel.

Perfect Competition in the Long Run

STEP Open the Excel workbook *Consistency.xls* and read the *Intro* sheet; then proceed to the *TheoryoftheFirmLongRun* sheet. Use the Zoom In button to fit the graphs on your screen so that all of them can be seen simultaneously.

The first and most important point is that all three optimization problems, in unison, comprise the Theory of the Firm. Perhaps because they see it in introductory economics, many students think of the output profit maximization graph as the firm. The display in *Consistency.xls* gives a strong visual presentation and constant reminder that the firm has three facets.

Gray-backgrounded cells are dead (click on one to see that it has a number, not a formula). They serve as benchmarks for comparisons when we do comparative statics.

The output and input profit maximization graphs do not have the usual U-shaped curves because the production function is Cobb-Douglas. This functional form cannot generate conventional U-shaped MC and AC curves (or upside down U-shaped MRP and ARP). There is no separate AVC curve because we are in the long run, so $AC = AVC$.

STEP Compare the initial solutions for each of the three problems.

There are several ways in which they agree.

1. L^* and K^* are the same in the Input Profit Max (left) and Input Cost Min (middle) graphs.

2. If you use these amounts of L and K, you will produce 636 units of output, as shown in the Output Profit Max (right) graph.

3. π^* is the same in the Input and Output Profit Max graphs. There is no profit in the Input Cost Min graph because there is no output price (P) and, therefore, no revenue in that optimization problem.

4. Total cost from each side is exactly the same. You can find TC from the Input Profit Max by creating a cell that computes wL^*+rK^*. This will equal \$36,262. From the Output Profit Max side, calculate TC by subtracting revenue, Pq, from profit. Again, you get \$36,262.

We can also see consistency in the ways in which the three optimization problems respond to shocks. As you would expect, the comparative statics results are identical.

STEP Wage increase of 1%. Change cell B2 to 20.2. Use the Zoom In button if needed to see more clearly how the graphs have changed.

Figure 14.4 shows the results.

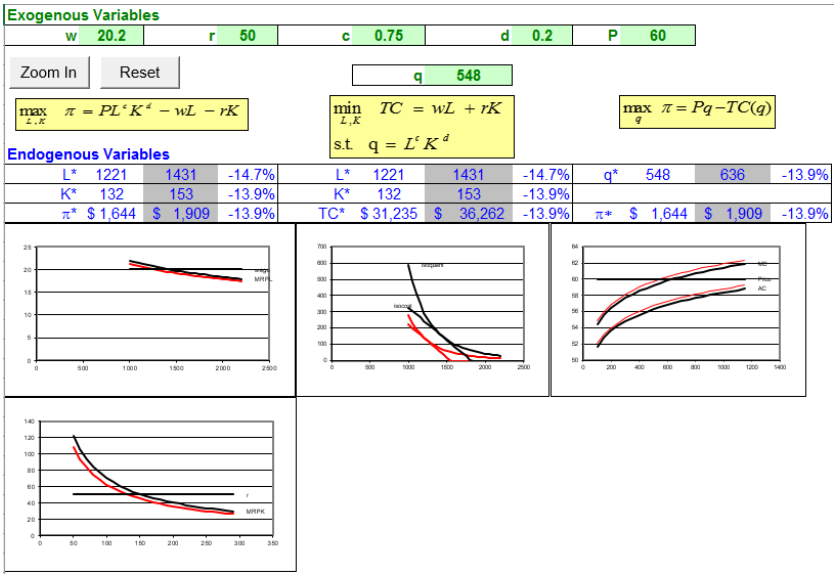


Figure 14.4: Wage shock in the long run.
Source: Consistency.xls!TheoryoftheFirmLongRun

On the Input Profit Max graph, we see that optimal labor use has fallen by 14.7% as wage rose by 1% (so the wage elasticity of labor from wage = \$20/hr

to \$20.20/hr is −14.7). Labor demand collapsed because the horizontal wage line shifted up and because the *MRP* schedule shifted left. The latter effect is due to the fact that optimal K fell.

On the Input Cost Min graph, we see that the firm is minimizing the cost of producing a lower level of output. In other words, we are on a new isoquant. Notice that the changes in L^* and K^* are consistent with the decreases reported from the Input Profit Max results.

The wage increase in the Output Profit Max graph is felt via the shifting up of the cost curves. The firm decreases q^* because *MC* shifted up and therefore the intersection of *MR* and *MC* occurs to the left of the initial solution.

Figure 14.4 and your screen shows how the Theory of the Firm reacts in a consistent manner to a wage shock. Is this true of other shocks? Yes. Here is another example.

STEP Click the Reset button, and then implement a labor productivity increase to 0.751 by changing cell F2.

Figure 14.5 shows the dramatic results of this shock. Input use and output produced have increased by about 18% in response to this tiny change in c.

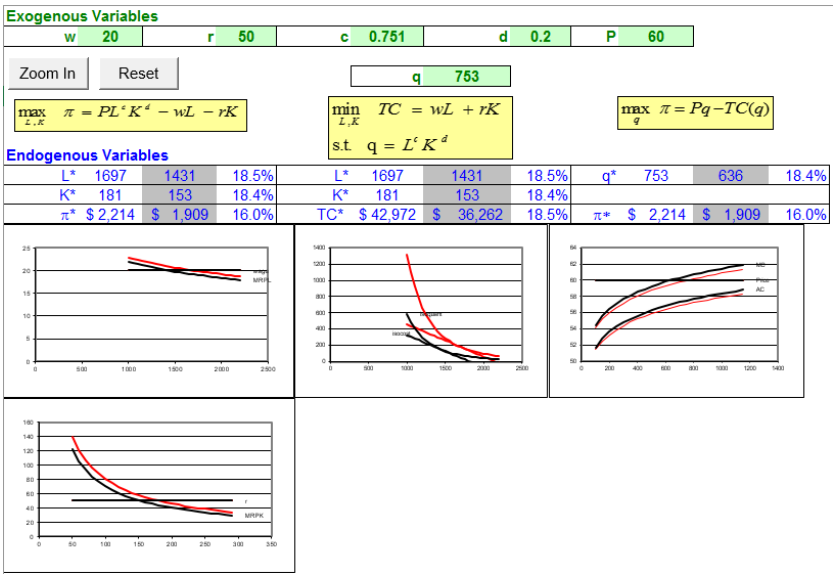


Figure 14.5: Labor productivity shock in the long run.
Source: Consistency.xls!TheoryoftheFirmLongRun

As with the wage shock, comparison of the effects of the change in c on the three optimization problems shows consistency. The two input side problems show that input use is the same and the inputs used will make the desired output on the output side. Profits on the input and output sides are the same. The productivity increase has shifted MRP up and cost curves down.

Other shocks are explored in Q&A and exercise questions. In every case, changing an exogenous variable, ceteris paribus, produces effects felt throughout the three optimization problems and the results are always consistent.

Perfect Competition in the Short Run

STEP Go to the *TheoryoftheFirmShortRun* sheet to explore the comparative statics properties of the three optimization problems in the short run.

This sheet has several differences compared to the previous overall view of the firm in the long run.

- There is an additional exogenous variable, K, because we are in the short run. Its value is set to the long run optimal solution for the initial values of the other parameters.
- There is a missing graph in the input profit max problem. With K fixed, we no longer need to depict its optimal solution.
- There is a straight, horizontal line in the isoquant side graph. With K fixed, the firm will not be able to roll around the isoquant to find the cost-minimizing input mix. It must use the given amount of K.
- There is an extra cost curve in the output profit max graph. Having K fixed means there is a fixed cost so we now have separate average total and average variable costs.

STEP Compare the initial solutions for each of the three problems. As expected, they agree in input use, output produced, and profits generated.

As before, we can change the light-green-backgrounded exogenous variable cells in row 2 and follow the results in the graphs.

STEP Apply a wage increase of 1%. Change cell B2 to 20.2. Use the Zoom In button if needed to see more clearly how the graphs have changed.

Figure 14.6 shows the results of this shock.

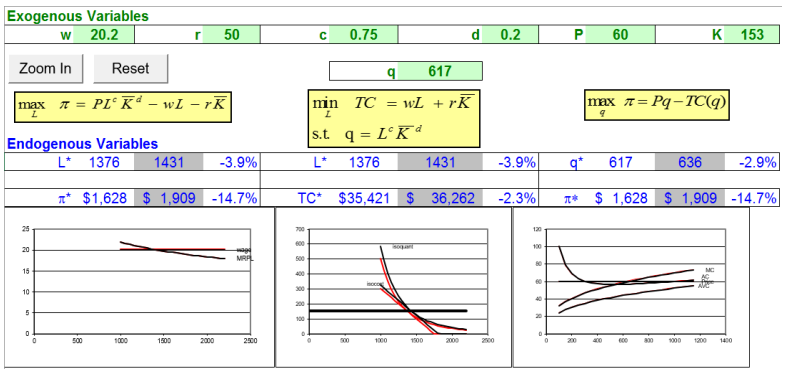


Figure 14.6: Wage shock in the short run.
Source: Consistency.xls!TheoryoftheFirmShortRun

The usual consistency properties are readily apparent. We observe the same change in L^*, q^*, and π^* across the board. Notice that the input profit max problem does not show a shift in MRP because K is fixed.

If we compare the short (Figure 14.4) to the long run (Figure 14.6), we see that the responsiveness of the changes in endogenous variables is greater in the long run. Labor and output fall by more in the long run. Profits, however, fall by less in the long run.

STEP Click the Reset button, then implement a labor productivity increase to 0.751 by changing cell F2.

Figure 14.7 displays the results.

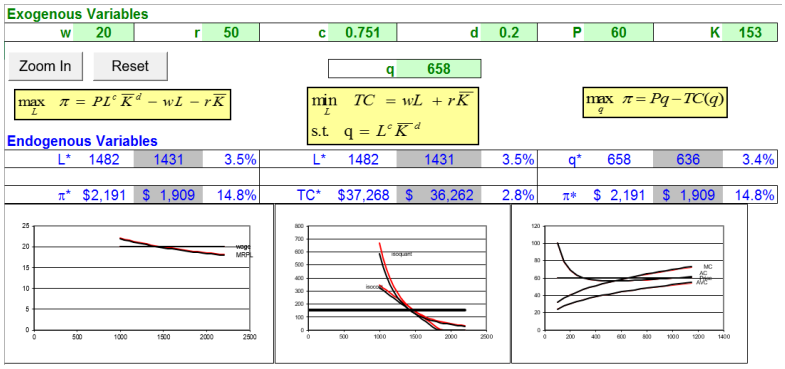


Figure 14.7: Labor productivity shock in the short run.
Source: Consistency.xls!TheoryoftheFirmShortRun

Figure 14.7 shows consistency in the results and, once again, the long run changes are more responsive than in the short run. L and K fall by more and the increase in profits is higher in the long run.

Long versus Short Run

When we compared the short and long run results for shocks in w and c, the long run exhibited greater responsiveness in labor and output. Is there a general principle at work?

Yes. The general law is that long run responses are always at least as or more elastic than in the short run. This is known as the *Le Chatelier Principle*.

Le Chatelier's idea, which he originally applied to the concept of equilibrium in chemical reactions, was introduced to economics by Nobel laureate Paul Samuelson in 1947.

> The Le Chatelier principle explains how a system that is in equilibrium will react to a perturbation. It predicts that the system will respond in a manner that will counteract the perturbation. Samuelson, following the methods of the hard sciences, has transported this principle of chemist Henri-Louis Le Chatelier to economics, to study the response of agents to price changes given some additional constraints. In his extension of this principle, Samuelson uses the metaphor of squeezing a balloon to further explain the concept. If you squeeze a balloon, its volume will decrease more if you keep its temperature constant than it will if you let the squeezing warm it up. This principle is now considered as a standard tool for comparative static analysis in economic theory. (Szenberg, et al., 2005, p. 51, footnote omitted)

In the context of the short and long run responses to shocks by a firm, the Le Chatelier Principle says that long run effects are greater because there are fewer constraints.

When the wage rises, a firm in the short run is stuck with its given quantity of K. In the long run, however, it will be able to adjust both L and K and it is this additional freedom of movement that guarantees at least as great or a greater response in input use and output produced.

For increasing c, the Le Chatelier Principle is reflected in the fact that labor demand is much more responsive in the long run than the short run. In the long run, the firm is able to take greater advantage of the labor productivity shock by renting more machines and hiring even more labor. This is, of course, reflected in the greater profits obtained in the long run in response to the increased c.

A Holistic View of the Firm

Figures 14.1, 14.2, and 14.3 are fundamental graphs for the Theory of the Firm. They represent the three optimization problems that, in unison, comprise the theory. The firm is not merely its output side representation, but includes all three optimization problem, as shown in the *Consistency.xls* workbook.

The input cost min (isoquants and isocosts that can be used to derive the cost function), output profit max (horizontal P with the family of cost curves that yield a supply curve), and input profit max graphs (horizontal w with MRP generating a demand curve for an input) are all intertwined. Not only do they all yield consistent answers for the initial solution, they all provide consistent comparative statics responses.

If we compare short and long run effects of shocks, we see that the firm responds more energetically in the long run. The wage elasticity of labor is greater (in absolute value) in the long run and, via consistency, so is the wage elasticity of output. Similarly, the c elasticities of labor and output are also greater in the long run.

Both of these results are examples of the Le Chatelier Principle: With fewer constraints, responsiveness increases. Since the short run prevents K from varying, the firm is less able to adjust to a shock. It can only vary L and, thus, its adjustment is more restricted and inelastic.

Exercises

1. What happens in the long run when price increases by 1%? Implement the shock and take a picture of the results, then paste it in a Word document. Comment on the changes in optimal labor, capital, output, and profits.

2. Compute the long run output price elasticity of labor demand. Show your work.

3. Apply the same 1% price increase in the short run. Take a picture of the results, then paste it in your Word document. Comment on the changes in optimal labor, capital, output, and profits.

4. Compute the short run output price elasticity of labor demand. Show your work.

5. Compare the price elasticities of labor demand in the long (question 2) and short run (question 4). Is the Le Chatelier Principle at work here? Explain why or why not.

6. With output price 1% higher, increase the wage by 1% in the long and short run. Do these two shocks cancel each other out in either case? Explain.

References

The epigraph is from page 116 of Eugene Silberberg, *The Structure of Economics: A Mathematical Analysis* (1978). This is a classic Math Econ book that was a popular graduate-level text.

Michael Szenberg, Aron Gottesman, and Lall Ramrattan, *Paul Samuelson On Being an Economist* (2005), explore the life and contributions of one of the most important economists of the 20^{th} century.

Instead of using marginal conditions as Cournot had done, Marshall used total ones. Perhaps for that reason, Cournot's marginal-revenue concept was forgotten and had to be rediscovered in the 1930s.

Hans Brems

Chapter 15

Monopoly

Like the perfectly competitive firm, a monopolist has three interrelated optimization problems. Attention is focused on the output profit max problem because that is where the essential difference lies between a perfectly competitive (PC) firm and a monopoly. We know that via consistency, monopoly power manifests itself on the input side also. A monopoly will produce less than a PC firm and, in turn, hire less labor and capital.

Unlike a PC firm, a monopoly chooses output and the price at which to sell the product. This makes the monopoly problem harder to solve. Fortunately, your experience with optimization, comparative statics, and graphical displays give you the background needed to understand and master monopoly.

Definition and Issues

A *monopoly* is defined as a firm that is the sole seller of a product with no close substitutes. The definition is inherently vague because there is no clear demarcation for what constitutes a close substitute.

Consider this example: In the old days, a local cable provider might have an exclusive agreement to provide cable TV in a community. One could argue that the cable provider was a monopoly because it was the sole seller of cable TV. But what are the substitutes for cable TV?

Years ago, cable TV was the only way to access subscription channels such as ESPN and HBO. Commercial broadcasts (with national broadcasters such as ABC, NBC, and CBS and local channels) were a poor substitute for cable TV. In this environment, cable TV would be a good example of a monopoly.

Today, however, cable TV has strong competition from satellite services and streaming services from the web. Even if a firm had an exclusive franchise to deliver cable TV in a community, there are many ways to get essentially the same package of channels. Today, cable TV is not a monopoly.

Of course, cable TV is not a good example of perfect competition either. The cable company does not accept price as a given variable. It is in the middle, somewhere between perfect competition and monopoly. Markets served by a few firms are called *oligopolies*. Add more firms and you eventually get monopolistic competition. The study of how firms behave under a variety of market structures is part of the subdiscipline of economics called Industrial Organization. Figure 15.1 sums things up.

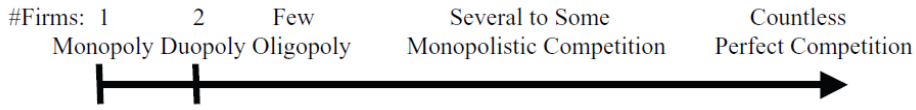


Figure 15.1: A continuum of market structures.

Barrier to Entry

To remain a monopoly, the firm must have a *barrier to entry* to prevent other firms from selling its product. In the cable TV example, the barrier to entry was provided by the exclusive agreement with the community. Such governmental restriction is a common form of a barrier to entry.

Another way to erect a barrier to entry is control over a needed input. ALCOA (the Aluminum Corporation of America) had a monopoly in aluminum in the early 20^{th} century because it owned virtually all bauxite reserves.

If a product requires entry on a large scale, like automobile manufacturing, this is considered a barrier to entry. To compete against established car companies, a firm must not only produce cars, but also many spare parts and figure out how to sell the product.

Like the concept of a close substitute, a barrier to entry is not a simple yes or no issue. Barriers can be weak or strong and they can change over time. Cable TV's barrier was eroded not by changes in legal rules, but by technological change—the advent of satellite TV and the web.

Monopoly's Revenue Function

We know that the firm's market structure impacts its revenue function. The simplest case is a perfectly (or purely) competitive firm. It takes price as given and, therefore, revenues are simply price times quantity. For a perfect competitor, even though market demand is downward sloping, the firm's own individual demand curve is perfectly elastic at the given, market price.

Because the PC firm can sell as much as it wants at the given price, selling one more unit of output makes total revenue (TR) increase by the price of the product. Marginal revenue (MR) is defined as the change in TR when one more unit is sold. Thus, for a PC firm, $MR = P$.

This is not true for a monopoly. A critical implication of monopoly power is that MR diverges from the demand curve. But this is too abstract. We can use Excel to make these concepts clearer.

STEP Open the Excel workbook *Monopoly.xls* and read the *Intro* sheet, then go to the *Revenue* sheet to see how monopoly power affects the firm's revenue function.

The sheet opens with a perfectly competitive revenue structure. Total revenue is a linear function of output and, therefore, $P = MR$ with a horizontal line in the bottom graph. A graph with a linear TR and corresponding horizontal MR means it is a PC firm.

Unlike a PC firm, a monopoly faces the market's downward sloping demand curve. We can model a linear inverse demand curve simply as $P = p_0 - p_1 q$. Because the slope parameter, p_1, in cell T2 is initially zero, TR is linear and MR is horizontal.

STEP To show how monopoly power affects the firm's revenue function, click on the *Price Slope* scroll bar.

Notice that as you increase the slope parameter, MR diverges more from D.

The smaller (in absolute value) the price elasticity of demand, the greater the divergence of MR from D and the stronger the monopoly power.

We will see that the monopolist uses the divergence of MR from D to extract higher profits than would be possible if there were other sellers of the product.

When drawing MR and D in the case of a linear inverse demand curve, keep in mind these two basic rules:

1. MR and D have the same intercept.
2. MR bisects the y axis and D.

We can derive these properties easily. With our inverse D curve, $P = p_0 - p_1 q$, we can do the following:

$$TR = Pq$$
$$TR = (p_0 - p_1 q)q$$
$$TR = p_0 q - p_1 q^2$$
$$MR = \frac{dTR}{dq} = p_0 - 2p_1 q$$

Clearly, both D and MR share the same intercept, p_0. Because the slope of MR is $-2p_1$, it is twice the slope of D, which is simply $-p_1$.

Thus, when you draw a linear inverse demand curve and then prepare to draw the corresponding MR curve, remember the two rules: (1) the intercept is the same and (2) MR has twice the slope so at every y axis value, MR is halfway between the y axis and the D curve.

Figure 15.2, with an inverse demand curve slope of -1, shows the monopoly's revenue function. Unlike the PC firm, TR is a curve and MR diverges from D. MR bisects the y axis and D. The dashed line at \$20/unit, for example, shows the distance from the y axis to MR is 10, the same as MR to D.

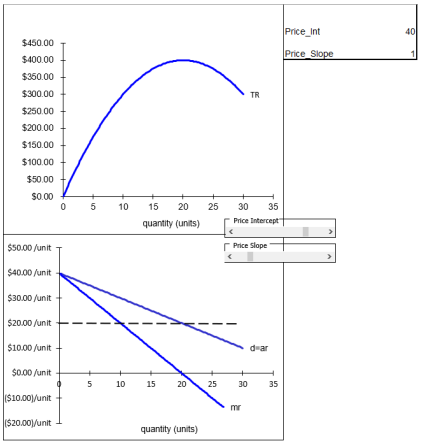


Figure 15.2: TR, D, and MR functions for a monopolist.
Source: Monopoly.xls!Revenues

Notice that where $MR = 0$ at $q = 20$, TR is at its maximum. At this quantity, the price elasticity of demand is exactly -1.

Figure 15.2 shows that MR can be negative. This can happen because there are two opposing forces at work. Increasing quantity increases TR, since $TR = Pq$. However, the only way to sell that extra product is to lower the price (by traveling down the demand curve) so TR falls. When the increase to TR by selling additional output outweighs the effect of the drop in the price, MR is positive. Eventually, however, with a linear demand curve, the monopolist will reach a point at which the increase in revenue for selling one more unit is negative. In the range of output ($q > 20$ in Figure 15.2) where $MR < 0$, the effect of the decreased price outweighs the positive effect of selling more output.

When $MR > 0$, the price elasticity of demand is greater than 1 (in absolute value). When MR is negative, demand is inelastic. The monopolist will never produce on the negative part of MR, which is the same as the inelastic portion of the demand curve.

There is a neat formula that expresses the relationship between MR and P. With an inverse demand curve, $P(Q)$, we know that $TR = P(Q)Q$. From the TR function we can take the derivative with respect to output to find the MR function. We use the Product Rule:

$$MR = \frac{dTR}{dQ} = P + \frac{dP}{dQ}Q$$

If we factor out P from this expression, then MR can be rewritten as:

$$MR = P + \frac{dP}{dQ}Q = P(1 + \frac{dP}{dQ}\frac{Q}{P}) = P(1 + \frac{1}{\epsilon})$$

The Greek letter epsilon (ϵ) is the price elasticity of demand ($\frac{dQ}{dP}\frac{P}{Q}$). The expression shows that $MR = P$ under perfect competition because an individual firm faces a perfectly elastic demand curve. This means epsilon is infinite and its reciprocal is zero.

It also shows that the more inelastic the demand curve (the closer ϵ is to 0), the greater the separation between MR and the demand curve (P). If $\epsilon = 0$, then MR is undefined. With $\epsilon = 0$, inverse demand is a vertical line. The monopoly would charge an infinite price.

Setting Up the Problem

There are three parts to every optimization problem. Here is the framework for a monopolist's output side profit maximization problem.

1. Goal: maximize profits (π), which equal total revenues (TR) minus total costs (TC).

2. Endogenous variables: output (q) and price (P)

3. Exogenous variables: input prices (the wage rate and the rental rate of capital), demand function coefficients, and technology (parameters in the production function).

A monopoly differs from a PC firm only on the revenue side—price is now endogenous. The cost structure is the same. The monopoly has an input cost min problem and it is used to derive a cost function. Increases in input prices shift cost curves up and improvements in technology shift cost curves down. The monopolist has a long and short run, just like a PC firm, and in the short run there is a gap between ATC and AVC that represents fixed costs.

Finding the Initial Solution

We will show the conventional approach to solving the monopoly problem first, then turn to an alternative formulation based on constrained optimization.

The conventional approach is to find optimal q where $MR = MC$, then get optimal P from the demand curve, and then compute optimal π as a rectangle. This is the standard approach and there is a canonical graph that goes along with this approach. Its primary virtue is that it can be easily compared to the perfectly competitive case.

The conventional approach can be demonstrated with a concrete problem. Suppose the cost function is $TC = aq^3 + bq^2 + cq + d$. Suppose the market (inverse) demand curve is $P = p_0 - p_1 q$. Thus, $TR = Pq = (p_0 - p_1)q$.

With this information, we can form the firm's profit function and optimiza-

tion problem, like this:

$$\max_q \pi = TR - TC$$
$$\max_q \pi = (p_0 - p_1)q - (aq^3 + bq^2 + cq + d)$$

We first solve this problem with numerical methods, then analytically.

STEP Proceed to the *OptimalChoice* sheet and look it over.
The profit function has been entered into cell B4. Quantity and price are displayed as endogenous variables, but q is bolded to indicate that it is the primary endogenous variable. In other words, Solver will search for the profit-maximizing output and, having found it, will compute the highest price that can be obtained from the demand curve.

The firm is making \$245 in profits by producing 10 units of output and charging \$34.50 per unit, but this is not the profit-maximizing solution. We know this because the marginal revenue of the 10^{th} unit is \$29/unit, whereas the marginal cost of that last unit is only \$4/unit. Clearly, the firm should produce more because it is making more in additional revenues from the last unit produced than the additional cost of producing that unit.

STEP Run Solver to find the optimal solution.

At the optimal solution, the equimarginal condition, $MR = MC$, is met. With positive profits, this is a clear signal that we have found the answer.

Before you click the Analytical Solution button, try doing the problem on your own. This is a single variable unconstrained maximization because $P = p_0 - p_1 q$ has been substituted into the profit function. Take the derivative with respect to q, set it equal to zero, and solve for optimal q. Substituting in the parameter values to make it a concrete problem makes it easier to do the math:

$$\max_q \pi = (40 - 0.55)q - (0.04q^3 - 0.9q^2 + 10q + 50)$$

You can check your work by clicking the Analytical Solution button. You can also confirm that the two approaches, Solver and calculus, agree.

STEP Proceed to the *OutputSide* sheet to see a familiar set of four graphs.

As usual, the totals are on the top and the average and marginal curves on the bottom. The cost curves are the quite similar to the PC firm's output profit maximization graphs, but the revenue curves are quite different.

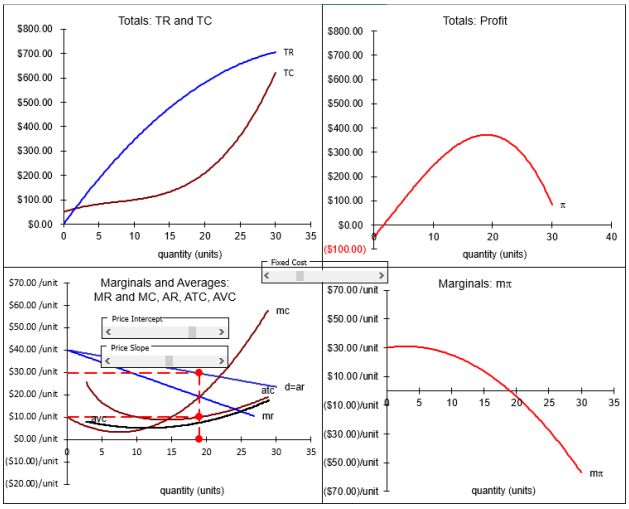


Figure 15.3: Monopoly output profit maximization graphs.
Source: Monopoly.xls!OutputSide

The bottom left-hand corner graph in Figure 15.3 is the canonical graph for a monopolist. It can be used to quickly find q^*, P^*, and π^*. Here's how to read and use the conventional monopoly graph:

1. Finding q^*: Choose q where $MR = MC$. This gives the biggest the difference between TR and TC and puts you on top of the profit hill (in the top right graph).

2. At q^*, travel straight up until you hit the demand curve to get P^*. This is the highest price that the monopolist can get for the chosen level of output.

3. Create the usual profit rectangle as $(AR–ATC)q^*$. It has length q^* and height $AR - ATC$ (where $AR = P$). The area of this rectangle equals the distance of the line segment between TR and TC, which is the height of the profit hill.

Play with the slider controls to improve your understanding of the graphs and relationships.

STEP Click the Fixed Cost slider to manipulate total fixed costs (d in the cubic cost function).

Changes in fixed costs do not affect the monopolist's optimal quantity and price solution. This is just like the perfectly competitive case.

STEP Click the Reset button; explore changes in the price intercept to see how the firm responds. At a low enough price intercept, profits become negative and, just like a PC firm, if $P < AVC$, the firm will shut down.

You can also control the firm's monopoly power by manipulating the inverse demand curve's slope.

STEP Set the Price Slope slider to zero. What happens?

You stripped the monopoly of its price power and it is a PC firm.

No Supply Curve for Monopoly

Monopolists do not have a supply curve. This seems like a strange statement since monopolies produce output and so "supply" whatever good or service of which they are the sole seller. But the key lies in the definition of a supply curve: given price, the supply curve gives the quantity that will be produced.

Because a PC firm is a price taker, it is possible to shock P and see how the optimal output changes. We can derive $q^* = f(P,$ ceteris paribus) and this is called a supply curve.

Unlike a perfectly competitive firm, for which price is exogenous, a monopoly chooses the price. Thus, we cannot ask, "Given this price, what is the optimal quantity supplied?" With price as an endogenous variable, it cannot serve as a shock variable in a comparative statics analysis.

We can (and you just did) shock a monopolist's demand curve parameters such as the intercept and slope, but this is not an exogenous change in the price of the product. The experiment of changing the price cannot be applied to a monopolist and, therefore, the monopolist has no supply curve.

Measuring Monopoly Power

Another common misconception is that monopoly is either zero or one. In fact, it is a continuum and you can have more or less monopoly power. There are several ways to measure it.

STEP Proceed to the *Lerner* sheet.

This sheet demonstrates the point that the more inelastic the demand faced by a monopolist, the greater the monopoly power. In other words, from a profit-maximizing point of view, it is better to have a monopoly over a product that everyone desperately needs (i.e., very inelastic) than to be the sole seller of a product that has a highly elastic market demand curve.

Abba Lerner formalized this idea in a mathematical expression that bears his name, the *Lerner Index*. "If P = price and MC = marginal cost, then the index of the degree of monopoly power is $\frac{P-MC}{P}$." (Lerner, 1934, p. 169). This measure of monopoly power uses the gap between P and MC as a percentage of P.

The Lerner Index takes advantage of the fact that a monopolist will choose that quantity where $MR = MC$, then charge the highest price possible for that quantity. The higher the price that can be charged, the more inelastic is demand and the greater the monopoly power.

The *Lerner* sheet compares two monopolies with the exact same cost structure (assumed for simplicity to have a constant $MC = AC$). They both produce the same profit-maximizing quantity, but Firm 2 faces a more inelastic demand curve than Firm 1 and, therefore, it has a bigger gap between price and marginal cost.

STEP Click on cells B16 and I16 to see the simple formulas for the Lerner Index.

The idea is that the bigger the divergence between price and marginal cost, the greater the monopoly power. Firm 2 has more monopoly power than Firm 1 and more monopoly profits. The Lerner Index for each firm reflects this.

Notice that a perfectly competitive firm that sets $MC = P$ will have a Lerner Index of zero. As the index approaches one, monopoly power rises.

STEP Change Firm 2's demand parameters to 130 for the intercept and 20 for the slope. The y axis is locked down so the entire D and MR functions are not displayed.

The optimal quantity is still 3, but P and profits are higher, as is the Lerner Index.

STEP Make the demand curve more inelastic at $Q = 3$ by setting the demand parameters to 190 and 30.

Optimal P has increased again, along with profits. The Lerner Index reflects the greater monopoly power.

STEP One last time, change the demand parameters to 6010 and 1000. The graph is hard to read because only MR is shown; D is literally off the chart.

Firm 2 continues to produce the same output as Firm 1, but has a much, much higher optimal price and maximum profits. Its Lerner Index is close to one. It cannot rise above one, but the closer it gets, the greater the divergence of P and MC so the greater the monopoly power.

The *Lerner* sheet also shows that the Lerner Index can be expressed as the reciprocal of the price elasticity of demand at the profit-maximizing price. The few algebra steps needed to connect the Lerner Index to the price elasticity start in row 25.

STEP Set Firm 2's demand parameters back to 70 and 10, and then click the Show Elasticity button.

The price elasticity of demand for the two firms is displayed. If you click in the cells, you can see the formula. Notice that the reciprocal of the inverse demand curve's slope is used to compute the price elasticity of demand correctly.

Firm 2's price elasticity of demand at the profit-maximizing price is lower than Firm 1's. The lower the price elasticity and the higher the Lerner Index, the greater the firm's monopoly power.

STEP Proceed to the *Herfindahl* sheet for a quick look at another way to measure monopoly power.

Instead of measuring the markup of price over marginal cost, we can see how big the firms are in an industry. Strictly speaking, a monopoly is one firm so it would have a 100% market share, but in practice, firms have monopoly power even though they are not technically monopolies. Any firm that faces a downward sloping demand curve and has the ability to set its price is said to have monopoly power.

If a market has many firms, each with the same share of total sales, we have a competitive market structure. If, on the other hand, only a few firms exist, the market is monopolized. The question is how to measure the degree of monopolization?

We can sort the firms in an industry from highest to lowest share and then add the shares of the four biggest firms. This gives the *four firm concentration ratio* in cell D5. It turns out this is not a very good way to distinguish between concentrated and unconcentrated industries.

The problem is that the four firm concentration ratio tells you nothing about the sizes of the top four firms or the rest of the market. The four firm concentration ratio is 70%, which seems pretty highly concentrated. The biggest firm's share, 30%, is almost one-third of the entire industry.

STEP Click on the Distribution A button.

The four firm concentration ratio is the same as before (70%), but this industry is clearly much more concentrated. Firm A is even bigger and the others are tiny.

STEP Click on the Distribution B button.

The four firm concentration ratio is the same as before (70%), but this industry is clearly less concentrated. The four top firms are equal so no one firm really dominates.

The primary virtue of the four firm concentration ratio is that it is easy to compute and understand. However, because we have three scenarios with wildly different shares for the top four firms yielding the same four firm con-

centration ratio, we can conclude that this ratio is a poor way to determine whether firms in a market are in a competitive or monopolistic environment. The four firm concentration ratio might be easy to compute and understand, but it is incapable of picking up differences in the distribution of shares.

A better way to judge concentration is via the *Herfindahl Index*. Unlike the Lerner Index, there is confusion about who invented it. Hirschman concludes, "The net result is that my index is named either after Gini who did not invent it at all or after Herfindahl who reinvented it. Well, it's a cruel world" (Hirschman, 1964, p. 761). It is sometimes called the Herfindahl-Hirschman Index (HHI).

Fortunately, its computation is simpler than its paternity. The idea is to square each share and sum, like this:

$$H = \sum_{i=1}^{n} S_i^2$$

The index ranges from $1/n$ to 1 (when using decimal values of shares). The higher the index, the greater the concentration. By squaring the shares, it gives more weight to bigger firms: for example, $0.1^2 = 0.01$, while $0.3^2 = 0.09$.

The *Herfindahl* sheet shows the computation. Notice how each value in column B is squared in column G. The sum of the squares is in cell G15 and it is the value of the Herfindahl Index.

STEP Click on the three buttons one after the other to cycle through them. Notice how the Herfindahl Index changes (but the four firm concentration ratio does not).

For Distribution A, the H value is 0.325. This is quite high. The 0.1375 value with Distribution B means there is more competition in this scenario than the other two.

The Herfindahl Index is not perfect because no single number can completely describe an entire distribution. It is, however, better than the four firm concentration ratio and often used to measure the degree of market competition.

The United States Department of Justice is charged with regulating the conduct and organization of businesses. The mission of the Antitrust Division is to promote economic competition. They use the Herfindahl Index as part of

their Horizontal Merger Guidelines (www.justice.gov/atr/horizontal-merger-guidelines-08192010). Markets with a Herfindahl Index less than 0.15 are "unconcentrated," values between 0.15 and 0.25 are "moderately concentrated," and anything over 0.25 is "highly concentrated."

The Department of Justice deems any proposed merger that increases the Herfindahl Index by more than 0.01 (100 points in the scale they use) in concentrated markets as warranting scrutiny. They can go to court to block mergers to prevent too much concentration. They can also break up companies that have too much monopoly power. This is known as *antitrust law* and is part of the Industrial Organization field of economics.

An Unconventional Approach

The monopolist's profit maximization problem can also be solved by choosing P and q simultaneously subject to the constraint of the demand curve. While this is not the usual way of framing the monopoly's optimization problem, it enables practice with the Lagrangean method of solving constrained optimization problems and reading isoprofit curves.

The analytical solution is based on rewriting the constraint so it is equal to zero ($P - (p_0 - p_1 q) = 0$), forming the Lagrangean, setting derivatives equal to zero, and solving the system of equations for the optimal solution.

$$\max_{p,q} L = Pq - (aq^3 + bq^2 + cq + d) + \lambda(P - (p_0 - p_1 q))$$
$$\frac{dL}{dP} = q + \lambda$$
$$\frac{dL}{dq} = P - 3aq^2 - 2bq - c + \lambda p_1$$
$$\frac{dL}{d\lambda} = P - (p_0 - p_1 q)$$

Set each derivative equal to zero and solve the three first-order conditions for q^*, P^*, and λ^*. From the first equation, $\lambda = -q$, substitute into the second equation:

$$P - 3aq^2 - 2bq - c + [-q]p_1 = 0$$

From the third first-order condition, $P = p_0 - p_1 q$, so

$$(p_0 - p_1 q) - 3aq^2 - 2bq - c - qp_1 = 0$$

Rearrange the terms to prepare for using the quadratic formula.

$$-3aq^2 - 2(b + p_1)q + (p_0 - c) = 0$$

$$\frac{-b + \sqrt{b^2 - 4ac}}{2a}$$

$$\frac{2(b + p_1) \pm \sqrt{4(b + p_1)^2 - 4(-3a)(p_0 - c)}}{2(-3a)}$$

STEP Proceed to the *ConOpt* sheet to see formulas based on the Lagrangean solution starting in cell F24.

Naturally, we get the same, correct answer as the unconstrained version.

The *ConOpt* sheet shows that monopoly as a constrained optimization problem can be depicted with a graph. The pink curves are isoprofit curves and the black line is inverse demand. The *MR* curve is not drawn because it is not used. The firm is trying to get to highest isoprofit without violating the demand curve constraint. Clearly, the opening values are not optimal.

STEP Run Solver and get a Sensitivity Report to confirm the value of lambda star is minus optimal quantity. Notice how the Solver dialog box is set up so Solver chooses cells B8 and B9 subject to the constraint.

After running Solver, the graph, reproduced in Figure 15.4, shows the usual tangency result.

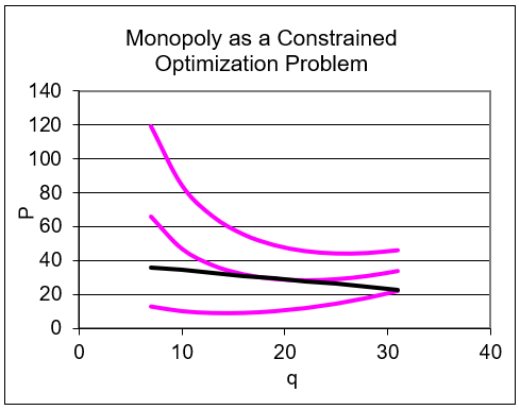


Figure 15.4: The constrained optimization version of the monopoly problem. *Source: Monopoly.xls!ConOpt*

The point of tangency provides the optimal q and P solution, while the value of the isoprofit curve at that point is the level of profits.

Do not be confused. The constrained version is rarely used. The conventional approach is the canonical output profit maximization graph (bottom left in Figure 15,2). This graph shows the optimal q where $MR = MC$ and easily displays P^* from the demand curve and π^* as a rectangle.

Figure 15.4 gives the same optimal solution, but presents the problem in a different way. Understanding that the demand curve serves as a constraint on monopoly is helpful. Monopoly power is not infinite. A monopolist cannot choose a ridiculously high price *and* a high quantity. As price rises, quantity sold must fall.

Monopoly Basics

A monopoly differs from a perfectly competitive firm in that a monopolist can choose the quantity and price, whereas a perfect competitor is a price taker. In addition, a monopolist has a barrier to entry that enables it to maintain positive economic profits even in the long run.

The two are the same, however, in the cost structure (like a perfect competitor, the monopolist derives its cost function from the input cost minimization problem) and the fact that it seeks to maximize profits (where $MR = MC$ as long as $P > AVC$).

We depict the monopolist's optimal solution with a graph that superimposes D and MR over the family of cost curves (MC, ATC, and AVC). Like a PC firm, a monopolist can suffer negative profits in the short run and it will shut down when $P < AVC$.

Monopoly's canonical graph (the bottom left chart in Figure 15.2) belongs in the pantheon of fundamental graphs in economics. Like the indifference curves with a budget constraint or supply and demand, a linear inverse demand with its associated marginal revenue showing optimal q (at the intersection of MR and MC, of course) and optimal P is a truly classic graph.

One way to measure monopoly power is by the Lerner Index. The greater the gap between price and marginal cost, the greater the monopoly power. The greater the price elasticity of demand, the lower the Lerner Index and

the weaker the monopoly power.

The Herfindahl Index is another way to measure the strength of monopolization in a market. It measures industry concentration. Unlike the four firm concentration ratio, it uses market shares of every firm to create a single number that reflects the concentration of an industry. Mergers that boost the Herfindahl Index by more than 0.01 (100 points) in concentrated markets are carefully scrutinized by the Department of Justice because it is presumed that the market will not be competitive.

We concluded this chapter with an unconventional analysis. The monopoly's profit maximization problem can be cast as a constrained optimization problem. In addition to providing practice with the Lagrangean method, this way of looking at monopoly makes quite clear that the monopolist must obey the demand curve.

Exercises

1. De Beers is an internationally famous company that had a monopoly over diamonds. Google "synthetic diamonds" to learn more. Include web citations with supporting evidence in your answers to these two questions.

   (a) What was their barrier to entry when they had a monopoly?

   (b) What happened to their monopoly?

2. Use Word's Drawing Tools to depict a monopoly shutting down in the short run. Explain the graph.

3. In the *ConOpt* sheet, set the demand intercept (cell B13) to 9 and the fixed cost (B18) to 180. Run Solver. Why is Solver generating a miserable result? What is the correct answer?

4. Use Word's Drawing Tools to depict the effect of monopoly from the input side profit maximization perspective. Explain the graph.

   Hint: With perfect competition, L^* is found where $w = MRP$ (where MRP is based on the given, constant price, $PxMP$). With monopoly, however, P and MR diverge.

5. Is the effect of monopoly on the input side consistent with the effect of monopoly on the output side? Explain.

References

The epigraph is from page 149 of Hans Brems, *Pioneering Economic Theory, 1630–1980: A Mathematical Restatement* (1986). This book recasts ideas in the history of economics in mathematical terms. Seeing the thoughts of Smith, Ricardo, Marx, and others presented as mathematical models provides an uncommon perspective.

On the Lerner Index, see Abba P. Lerner, "The Concept of Monopoly and the Measurement of Monopoly Power," *The Review of Economic Studies*, Vol. 1, No. 3 (June, 1934), pp. 157–175, www.jstor.org/stable/2967480.

On the Herfindahl Index, see Albert O. Hirschman, "The Paternity of an Index," *The American Economic Review*, Vol. 54, No. 5 (September, 1964), p. 761, www.jstor.org/stable/1818582.

> Von Neumann hovered for a moment by two rather sloppily dressed graduate students who hunched over a peculiar-looking piece of cardboard. It was a rhombus covered with hexagons. It looked like a bathroom floor. The two young men were taking turns putting down black and white go stones and had very nearly covered the entire board.
>
> Later that evening, at a faculty dinner, he buttonholed Tucker and asked, with studied casualness, "Oh, by the way, what was it they were playing?" "Nash," answered Tucker, allowing the corners of mouth to turn upwards ever so slightly, "Nash."
>
> ---
>
> Sylvia Nasar

Chapter 16

Game Theory

In perfect competition, firms are price takers with no power to affect the market price. Each firm optimizes by choosing q to equalize MC and P.

In monopoly, the sole seller of a product with no close substitutes optimizes by choosing q to equalize MC and MR and then charges the highest price that clears the market (given by the demand curve).

In both market structures, the profits of the individual firm are not affected by what anyone else does. In perfect competition, there are so many other firms that Firm i does not care about what Firm j is doing. In monopoly, there is no other firm to worry about.

What about market structures between the extremes of perfect competition and monopoly? *Oligopoly* is a market dominated by a few firms. Their decisions are interdependent. In other words, what each individual firm chooses does affect the sales and profits of the other firm. To optimize, each firm must anticipate what their rivals will do and then choose its best options. This is clearly a more realistic model than that of perfect competition and monopoly, which rely on idealized, abstract descriptions of firms that have no real-world counterparts.

How do oligopolies behave? We know that, like other firms, they optimize given the economic environment, but because of interdependence, it is much more difficult to analyze.

This chapter opens the door to the analysis of strategic behavior. It presents a few basic ideas from the fields of Game Theory and Industrial Organization.

Interdependence and Nash Equilibrium

It seems obvious when we say that firms are interdependent, but exactly what does this mean? Consider two power companies that generate and sell electricity. This is a good example of a homogeneous product. We assume consumers do not care at all which of the two firms provides electricity to their homes.

To keep it simple, suppose that each power company can choose either a high level of output or a low level of output. Market price is a function of the output decisions of the two firms. Each power company's profits are functions of their own decision to produce and the market price.

Figure 16.1 displays a *payoff matrix*, which shows the possible choices and outcomes. You read the entries in the payoff matrix like coordinate pairs on a graph, the first part is for Firm 1 and the second for Firm 2. The \$300, \$300 pair in the top left of the four entries says that Firm 1 chose high output and Firm 2 chose high output. Each firm ends up with low profits.

		Firm 2			
		High Output		Low Output	
Firm 1	High Output	\$300 profits,	\$300 profits	\$1000 profits,	\$200 profits
	Low Output	\$200 profits,	\$1000 profits	\$800 profits,	\$800 profits

Figure 16.1: The payoff matrix.

If Firm 2 had chosen low output (top right), Firm 1 profits would be much higher, \$1,000, because it made a lot of output and price rose when Firm 2 decided to cut back.

This particular game is a one-shot, simultaneous-move game known as the *Prisoner's Dilemma*. You have probably seen it before. Two criminals are arrested and questioned separately. If both stay silent, they get 1 year in jail. If both confess, they get 3 years. But if one confesses and the other does not, the one who talks gets no jail time and the silent one gets 10 years.

You can match those outcomes to the payoff matrix in Figure 16.1. The outcome that is best for both firms together is \$1600 total, with \$800 for each company. But, like the criminals version of the game, that is going to be an unlikely outcome. Suppose that both agree beforehand that they

are going to collude and both choose low output. Unless they can write a binding agreement that is enforceable (so a cheater can be punished), there is an incentive for each firm to change its decision and choose high output if it thinks that the other firm will stick with low output. As a result, both firms end up with low profits (and both criminals confess).

If you think the other firm is going to cheat, your best move is to also cheat. If you think the other firm is going to honor the agreement, your best move, in the sense of profit maximization, is to cheat and produce a high output (assuming this is a one-time game and you never have to see your opponent again). It looks like cheating, producing high output (or confessing), is the best move no matter what the other firm does. We say that this game has a dominant strategy—produce high output (confess).

This result illustrates the reason why cartels—groups of firms that get together to charge the monopoly price and split the monopoly profits—are unstable. It is difficult for oligopolistic firms to get together and act like a monopoly because there is an incentive for individual firms to cheat on the agreement and produce more to take advantage of high prices.

Because of the interdependence of firms' decision making, competition among firms in an oligopoly may resemble military operations involving tactics, strategies, moves, and countermoves. Economists model these sophisticated decision making processes using game theory, a branch of mathematics and economics that was developed by John von Neumann (pronounced noy-man) and Oskar Morgenstern in the 1930s. One of the most important contributors to game theory is John Nash, a mathematician who shared the Nobel Prize in Economics.

A game-theoretic analysis of oligopoly is based on the assumption that each firm assumes that its rivals are optimizing agents. That is, managers act as though their opponents or rivals will always adopt the most profitable countermove to any move they make. The manager's job is to find the optimal response.

Nash's most important and enduring contribution is the concept named after him, the Nash equilibrium. Once we are in a world where firms are interdependent and one firm's profits depends on what other firms do, we are out of the world of exogenously given price that we used for perfect competition and out of the isolated world of the monopolist. John Nash invented an equilibrium concept that describes a state of rest in this new world of

interdependence.

A *Nash equilibrium* exists when each player, observing what her rivals have chosen, would not choose to alter the move she herself chose. In other words, this is a *no regrets equilibrium*: After observing the outcome, the player does not wish she would have done something else instead.

We will explore in detail a concrete example of a *duopoly* (a market with two firms) with a single Nash equilibrium. Remember, however, that this is simply one example. Some games have one Nash equilibrium, some have many, and some have none. There are many, many games and scenarios in game theory and we will look at just one simple example.

The Cournot Model

Augustin Cournot (pronounced coor-no) was a remarkably creative 19^{th}-century French economist (see the References in section 12.2). Cournot originally set up a model of duopolists who produce the same good and optimize by choosing their own output levels based on assumptions about what the rival will do.

Here is the set up:

- Two firms.
- Each produces the exact same product.
- Constant unit cost.
- Firms choose output levels at the same time.
- Both know the market demand for the product.

The profit of each firm depends on how much it produces and how much its rival produces. If the rival produces a lot, the the market price falls. The interdependence is that one firm's decision about how much to produce affects the price and, thus, the rival's profit.

What strategy should each firm use to choose its output level? The answer depends on its beliefs regarding its rival's behavior.

STEP Open the Excel workbook *GameTheory.xls* and read the *Intro* sheet, then go to the *Parameters* sheet.

Market demand is given by the linear inverse demand curve and, for simplicity, we assume a linear total cost function. This means that $MC = AC$ is a horizontal line.

STEP Proceed to the *PerfectCompetition* sheet.

With many small PC firms, the industry as a whole will produce where demand intersects supply (which is the sum of the individual firm's MCs). The graph shows that a perfectly competitive market will produce 15,000 kwh at a price of 5¢/kwh.

What happens if a single firm takes over the entire market?

STEP Proceed to the *Monopoly* sheet. Use the *Choose Q* slider control to determine the profit-maximizing quantity. Keep your eye on cell B18 as you adjust output. The optimal output is found where $MR = MC$.

The monopolist will produce 7500 kwh and charge a price of 12.5¢/kwh. This solution nets a maximum profit of 56,250 cents.

Not surprisingly, compared to the perfectly competitive results, monopoly results in lower output and higher prices.

Cournot was the first to ask the question, "What happens if the industry is shared by two firms?"

To understand the answer, the concept of residual demand is crucial because it enables us to solve the firm's optimization problem. *Residual demand* is the demand curve facing the firm after the sales from the other firm are subtracted. From there, the reaction function for each firm is derived from a comparative statics analysis. The two reaction functions are then combined to yield the Nash equilibrium, which is the answer to Cournot's question. That is confusing. We turn to Excel to see each step and how it all works.

Residual Demand

To figure out the quantity and price combination with two competing firms, we need to understand how the firms will behave.

STEP Proceed to the *ResidualDemand* sheet.

This sheet shows how Firm 1 decides what to do, given Firm 2's output decision. Think of the chart as belonging to Firm 1. It will use this chart to decide what to do, given different scenarios.

Conjectured Q2, in cell B14, is the key variable. A conjecture is an educated guess. It is based on incomplete information. Firm 1 does not know and cannot control what Firm 2 is going to do. Firm 1 must act, however, so it treats Firm 2's output decision as a conjecture and proceeds based on that projected value.

Conjectured Q2 is an exogenous variable for Firm 1. It does not know what Firm 2 will do and cannot control it. The conjectured output of Firm 2 may be different from Firm 2's actual output. Firm 1 can, however, examine how it would react to different possible values of Firm 2 output.

The *ResidualDemand* sheet opens with *Conjectured Q2* $= 0$. In this scenario, Firm 2 produces nothing and Firm 1 behaves as a monopolist, producing 7,500 kwh and charging a price of 12.5¢/kwh.

STEP Click five times on the scroll bar in cell C14. With each click, *Conjectured Q2* rises by 1,000 units and the red lines in the graph shift left.

The red lines are the critical factor for Firm 1. They represent residual demand and residual marginal revenue. The idea behind residual demand is that Firm 2's output will be sold first, leaving Firm 1 with the rest of the market.

The residual in the name refers to the fact that Firm 2 will supply a given amount of the market and then Firm 1 is free to decide what to do with the demand that is left over.

With each click, Firm 2 was producing more and so the demand left over for Firm 1 was falling. This is why the residual demand shifts left when Firm 2 produces more.

As the *Parameters* sheet shows, the inverse demand curve for the entire market is given by the function $P = 20 - 0.001Q$. If *Conjectured Q2* = 5,000, then the residual inverse demand curve is $P = 20 - 0.001Q - 0.001(5000)$. In other words, we subtract the amount supplied by Firm 2. Thus, the residual inverse demand curve is $P = 15 - 0.001Q$.

Figure 16.2 shows how the residual demand is shifted left by 5,000 kwh when *Conjectured Q2* is 5,000. The key idea is that Firm 2's output is subtracted from the demand curve and what is left over, the residual, is the demand faced by Firm 1.

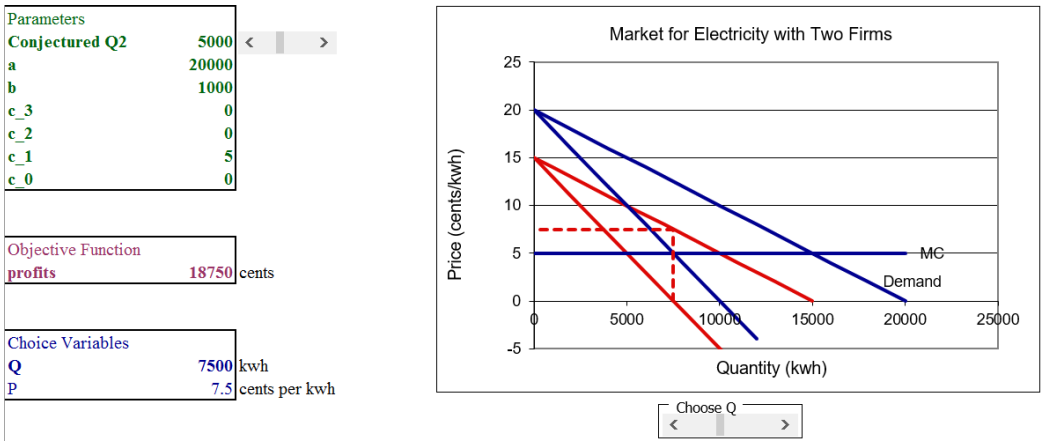


Figure 16.2: Residual demand.
Source: GameTheory.xls!ResidualDemand

Once we have residual demand for Firm 1, we can find the profit-maximizing solution. Firm 1 derives residual MR from its residual demand curve and uses this to maximize profits by setting residual $MR = MC$. In Figure 16.2, Firm 1 is not maximizing profits by producing 7,500 units and charging 7.5¢/kwh. Notice that the price is read from the residual demand curve, not the full market demand curve.

STEP Use the scroll bar (below the chart) to find Firm 1s optimal solution when *Conjectured Q2* is 5,000.

You should have found that optimal Q is 5,000 kwh, optimal $P = 10$¢/kwh and maximum π are 25,000 cents.

The Reaction Function

Now that we know how the duopolist uses residual demand to choose the quantity (and price) that maximizes profits, we can proceed to the next step in answering Cournot's question: "What happens if the industry is shared by two firms?"

We track each duopolist's optimal output as a function of *Conjectured Q2*. This gives the *reaction (or best response) function*. The reaction function is a comparative statics analysis based on shocking *Conjectured Q2*.

STEP Fill in the table in the *Residual Demand* sheet. You are picking points off of Firm 1's reaction function.

You already have two of the rows. In addition to the optimal solution at *Conjectured Q2* = 5,000 which we just found, when *Conjectured Q2* = 0, optimal output is 7,500 and optimal price is 12.5¢/kwh. Fill in the rest of the table.

STEP Check your work by clicking the Check Table button.

The filled in table is giving us Firm 1's reaction function. It is similar to the output of the CSWiz—the leftmost column is the exogenous variable and the other columns are endogenous responses.

Deriving Firm 1's reaction function is an important step in figuring out how two firms will interact. The reaction function gives us Firm 1's optimal response to Firm 2's output decision. We do not know, however, what Firm 2 will actually do. It has a reaction function just like Firm 1. The two firms must interact to determine what will happen in the market.

Finding the Nash Equilibrium

Residual demand enabled us to understand the reaction function. We are now ready for the third and final step so we can answer Cournot's question concerning the results of a duopoly. Remember, perfect competition gives 15,000 kwh of output and monopoly gives only 7,500 (and at a higher price). Presumably, duopoly is between them, but where?

STEP Proceed to the *Duopoly* sheet.

The display is new, but easy to understand. Instead of working with just Firm 1, both are shown. They have the same costs.

The sheet has buttons that make it a snap to see what each firm will do. The analytical solution is used so you do not have to run Solver every time *Conjectured Q2* changes.

STEP Notice that *Conjectured Q2* (in cell B13) is zero. To find the optimal solution, click the Choose q_1* button.

Not surprisingly (given our earlier work with the residual demand graph) since *Conjectured Q2* is zero, Firm 1 chooses to produce 7500 kwh.

But look at cell G13—Firm 1 has optimized, but now we need to ask what Firm 2 would do if Firm 1 made 7,500 kwh? Firm 2 wants to maximize profits just like Firm 1.

STEP Click the Choose q_2* button.

Firm 2's solution makes sense. If Firm 1 makes 7,500 kwh, then Firm 2 maximizes profits by taking the residual demand and producing 3,750 kwh. Their combined output means $P = 8.75$.

This is not, however, an equilibrium solution because Firm 1 is not going to produce 7,500 kwh. Why not?

STEP Look at cell B13. Click on cell B13.

B13's formula, =G20, makes clear how Firm 1's decision is connected to its rival. If Firm 2 says it wants to produce 3,750, then Firm 1 regrets and will change its previous choice. We need to find the optimal output for Firm 1 given Firm 2's new level of output.

STEP Click the Choose q_1* button.

Firm 1 chooses to make 5,625 kwh (based on Firm 2's output of 3,750 kwh), but now we return to Firm 2. Will it produce 3,750 kwh? No. When Firm 1 changed its output, cell G13 updated. Like B13, G13 connects Firm 2's optimal decision to Firm 1's output choice.

It is Firm 2's turn to regret its previous decision. Firm 2 can make higher profits by changing its output when Firm 1 makes 5,625 kwh. How much will Firm 2 want to produce? Let's find out.

STEP Click the Choose q_2* button.

Firm 2 is set, but what about Firm 1? Does it regret making 5,625? Yes, it does because it can make higher profits by changing its decision.

We will not be in equilibrium until both firms are happy with their output choice and do not wish to change it. Since Firm 2 changed its output, Firm 1 will want to change its output.

STEP Click the Choose q_1* button.

You might be thinking that this will never end. That is incorrect. It will end. You can actually see it end.

STEP Repeatedly move back and forth, clicking the Choose q_2* and Choose q_1* buttons, one after the other. What happens?

After repeatedly clicking, you are looking at convergence. Clearly, the two optimal output levels closed in on 5,000—this is the Nash equilibrium solution to this problem and the answer to Cournot's question. The duopoly will produce a combined total of 10,000 kwh with a price of 10¢/kwh. This makes sense since it is in between the perfectly competitive (15,000 kwh) and monopoly outcomes (7,500 kwh).

Manually optimizing for each firm in turn, back and forth, until the equilibrium solutions comes into focus is a great way to understand the concept of a *Nash equilibrium*. It is a position of rest where neither firm regrets its previous decision. In fact, a Nash equilibrium is often referred to as a "no regrets" point. There is, however, a faster way to find the position of rest.

STEP click the Nash Equilibrium button.

This button does all of the hard work for you. It alternately solves one firm's problem given the other firm's output many times. It continues to maximize firm profits until there is less than a 0.001 difference between a firm's optimal output and its optimal output based on the conjectured output of its rival.

STEP To see this, click on cells B20 and G20. They are close to 5,000, but not exactly 5,000.

The Nash Equilibrium button also displays the individual firm's reaction functions (scroll down if needed). In this case, the two reaction functions are identical.

Finally, the Nash Equilibrium button shows the two reaction functions on the same chart and the intersection instantly reveals the Nash equilibrium. Figure 16.3 shows the Nash equilibrium chart with additional elements to help explain it.

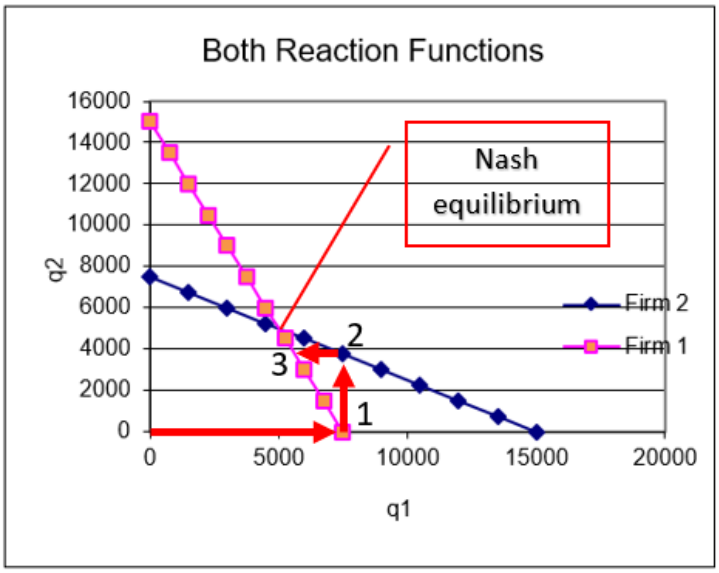


Figure 16.3: Nash equilibrium.
Source: GameTheory.xls!Duopoly

Point 1 in Figure 16.3 represents the first time Firm 1 maximized profits, with *Conjectured Q2* of zero. Point 2 shows Firm 2's optimization based on Firm 1 making 7,500 kwh. You can see, by following the arrows, how this would lead to the intersection as the Nash equilibrium.

You might wonder why the reaction functions are not the same in Figure 16.3 since they are identical when graphed by themselves (as shown below the buttons in the *Duopoly* sheet). The answer lies in the axes—to plot them

both on the same graph, we use the reaction function for Firm 2 and the *inverse* reaction function for Firm 1. Scroll down to see the inverse reaction function starting in row 63.

Remember: A Nash equilibrium exists when each player, observing what her rivals have chosen, would not choose to alter the move she herself chose. Nash equilibrium is a no regrets point for all players.

Figure 16.3 shows that the Nash equilibrium is at the intersection of the two reaction functions. Only there will both firms decline the offer to change their optimal decisions. This is a position of rest.

Evaluating Duopoly's Nash Equilibrium

We know the answer to Cournot's question. Duopoly, at its Nash equilibrium, leaves us in between perfect competition and monopoly. But we can say more about the duopoly outcome. We focus on profits.

STEP In cell D16 in the *Duopoly* sheet, enter a formula that adds the profits of the two firms at the Nash equilibrium. What are industry profits?

You might recall monopoly had maximum profits of 56,250 cents. That is better than the 50,000 cents you just computed with your formula in cell D16 of = B16 + G16.

Can duopolists increase their profits to 56,250 like a monopolist? Yes, they can, but they will not be able to honor their commitments.

STEP Set quantities for both firms (in cells B20 and G20) to 3,750. What happens to profits?

Amazingly, they go up. If the two rivals can agree to simply split the monopoly output of 7,500 kwh, each will make 28,125 cents and match the monopoly outcome.

But this will not last. Why not? Why don't the two firms get together and produce 3,750 units each and make greater joint profits than the Nash equilibrium solution? A single click reveals the answer.

STEP Click the Choose q_1* button or the Choose q_2* button.

If the rival makes 3,750 kwh, the firm maximizes profits at 5,625 kwh. In other words, they have an incentive to cheat—just like in the Prisoner's Dilemma game.

As soon as one takes advantage, the other fires back and they spin back to the Nash equilibrium.

You might suggest writing a contract, but that is illegal and unenforceable in the United States. There are other options and strategies, but they would take us too far from Intermediate Microeconomics. One strong attraction that is easy to see is merger. If the two firms combine into a single entity, they will be a monopoly and enjoy monopoly profits. Presumably, the Department of Justice would object.

Interdependence

Game theory is an exciting, growing area of economics. Its primary appeal lies in the realistic modeling of agents as strategic decision makers playing against each other, moving and countering. This is obviously what a real-world firm does.

The Cournot model is a simple game matching two firms against each other. It illustrates nicely the notion of interdependence and how one firm moves, and then the other responds, and so on. Whereas some games do not have a Nash equilibrium, the Cournot duopolists do settle down to a position of rest.

The *Summary* sheet has the outcomes from perfect competition, duopoly, and monopoly. It is clear that monopoly maximizes firm profits, but perfect competition offers the consumer the lowest price and most output. We will return to this comparison in the third and final part of this book.

We have just scratched the surface of game theory. There are many, many more games. The workbook *RockPaperScissors.xls* lets you play this child's game in Excel. Section 17.7 on Cartels and Deadweight Loss has another application of game theory.

For an entertaining version of the Prisoner's Dilemma in a game show, see this *Golden Balls* episode finale: tiny.cc/splitsteal. And for a really clever twist, watch this one: tiny.cc/ibrahim. Nick's strategy has been outlawed from the show. The Cornell game theory blog has an entry explaining it: tiny.cc/splitstealanalysis.

Exercises

These exercises are based on $c_1 = 5$. If you did the *Q&A* questions and changed this parameter, change it back to its original value.

1. If *Conjectured Q2* is 15,000, why does Firm 1 decide to produce nothing? Use the *ResidualDemand sheet* to support your explanation.

2. Suppose Firm 1 produces 4,500 kwh and Firm 2 produces 6,000 kwh. Does Firm 1 have any regrets? Does Firm 2 have any regrets? Enter these two values in the *Duopoly* sheet and click the Choose q buttons. Which firm changed its mind? Why?

3. Click the Reset All button in the *Duopoly* sheet. Explore the effect of changing Firm 1's cost function so that c_2 (cell B10) is 0.001 (with B11 = 5). How does this affect the Nash equilibrium?

References

The epigraph is from page 75 of Sylvia Nasar, *A Beautiful Mind* (1998). This biography of Nash has won countless awards and was made into an Academy Award-winning motion picture, with Russell Crowe starring as John Nash. Although much of the book is devoted to Nash's personal struggle with schizophrenia, Nasar's book gives a clear and engaging review of game-theoretic concepts before Nash and of the Nash equilibrium.

On the game Nash invented that is mentioned in the epigraph, Nasar writes,

> That spring, Nash astounded everyone by inventing an extremely clever game that quickly took over the common room. Piet Hein, a Dane, had invented the game a few years before Nash, and it would be marketed by Parker Brothers in the mid-1950s as Hex. But Nash's invention of the game appears to have been entirely independent. (p. 76)

A Brilliant Madness, www.pbs.org/wgbh/americanexperience/films/nash/, is an excellent 2002 documentary on Nash's life, struggles, and contributions.

In 1994, Nash, John C. Harsanyi, and Richard Selten shared the Nobel Prize "for their pioneering analysis of equilibria in the theory of non-cooperative games." (www.nobelprize.org/prizes/economic-sciences/1994/summary/)

The first edition of *The Theory of Games and Economic Behavior* by John von Neumann and Oskar Morgenstern was published in 1944.

Thinking Strategically: The Competitive Edge in Business, Politics, and Everyday Life by Avinash K. Dixit and Barry J. Nalebuff (originally published in 1991) explains and applies game theory to a variety of interesting examples and situations.

Part III

The Market System

The Butterfly Effect
acquired a technical name:
sensitive dependence on
initial con-ditions.

James Gleick

Overview

The first part of this book was the Theory of Consumer Behavior. It modeled a consumer's utility maximization problem and emphasized deriving a Demand Curve as the key result.

The Theory of the Firm comprised the second part. Firm decisions about inputs and outputs were modeled as optimization problems. The key result was deriving a Supply Curve from the perfectly competitive firm's output profit maximization problem.

This third part will put together consumers' demand and firms' supply in an equilibrium model. This will show how individual markets solve society's resource allocation problem. In addition, we will introduce an equilibrium model that incorporates all markets simultaneously.

Before we begin, we review these three key ideas:

1. Optimization versus equilibrium.
2. Partial and general equilibrium.
3. Society's resource allocation problem.

1. Optimization Versus Equilibrium

The stress thus far has been on optimization. Consumers maximize utility, firms minimize costs and maximize profits. We have used numerical and analytical methods, including the Lagrangean, to solve these problems.

The market system, however, is an equilibrium model. There are similarities between optimization problems and equilibrium models. They both rely heavily on comparative statics and we will continue to use numerical and analytical methods, but there are critical differences.

In an optimization problem, an agent explicitly chooses, setting the values of endogenous variables. For example, a consumer picks from available options to maximize utility and a firm manipulates variables to maximize profit. An optimal solution means the best choice is made (from the decision maker's point of view).

Unlike optimization problems, equilibrium models do not have an agent directly controlling or setting values of a variable. Instead, forces within the model drive variables to positions of rest. No agent actually picks the solution in an equilibrium model. Instead, the equilibrium solution means that there is no tendency to change in the endogenous variables (those determined within the model).

The notation we will use is common in economics, but often goes unremarked and unnoticed. A star, or asterisk, means optimal. We have found $x_1{}^* = 25$ and $L^* = 1431$. The star means this value is the best value the agent can choose.

In equilibrium models, the solution is denoted by a subscript "e." We might find that $Q_e = 100$. This means that the system settles down and is at rest at this value.

Unlike optimization problems, an equilibrium solution says nothing about the desirability of the solution. In other words, we cannot conclude that an equilibrium solution is a good one simply because it is the equilibrium solution. We could be at rest at a bad place.

Finally, unlike optimization problems, economists are often interested in the equilibration process, that is, the path followed to the final resting place. If it exists, the type of convergence, direct or oscillatory, can be studied. The equilibration process is beyond the scope of this book, but it helps show the difference between optimization and equilibrium. There is no process in optimization—the agent chooses the best solution and if there is a shock, the agent instantly re-optimizes. Not so with equilibrium. A shock will put forces into play that move the system.

Confusing equilibrium with optimal is common, but bad practice. They are different in the fundamental fact that optimization has an agent choosing and equilibration does not. Never automatically assume that an equilibrium solution is optimal.

2. Partial and General Equilibrium

While all equilibrium models rely on the concept of rest or stability as a key marker of the equilibrium solution, the market system was analyzed in two fundamentally different ways:

1. Partial equilibrium: Focus on a single good or service, in isolation.
2. General equilibrium: Consider all of the goods or services together.

Partial equilibrium was made famous by Alfred Marshall. He not only popularized putting price on the y axis, he made the graphical display of supply and demand curves for individual goods and services popular, especially in the English-speaking world. It is easy to see the equilibrium solution at the intersection of the supply and demand curves and, we will see, the graph can be used to evaluate the equilibrium outcome.

In the rest of Europe, a different tradition arose. Spearheaded by increasingly sophisticated mathematical economists, such as Leon Walras and Vilfredo Pareto, a more holistic approach to the market system was developed. Instead of looking at a single product or industry, all goods and services are simultaneously analyzed.

You are already familiar with partial equilibrium because supply and demand graphs are a staple of high school and introductory economics courses. General equilibrium theory, however, will be new and challenging.

Make no mistake, they are not equal. General equilibrium is superior to partial equilibrium analysis, but it is also more complicated and difficult. In our study of the market system, we will first analyze individual markets using conventional supply and demand graphs, then we turn to general equilibrium analysis.

In both partial and general equilibrium analyses, we first determine the equilibrium solution and then judge it by comparing it to an optimal solution. We avoid the fundamental error of conflating equilibrium with optimal. We may find that an equilibrium solution is, in fact, optimal, but we will also see situations where this is not so and the market fails.

3. Society's Resource Allocation Problem

The partial and general equilibrium models explain how markets function in solving a particularly fundamental optimization problem. It is so important that is it often referred to as *The Economic Problem*.

Figure III.1 depicts the problem. Given scarce resources of labor and capital (representing all inputs), society must decide what to produce, how much of each product to make, and how to distribute the output.

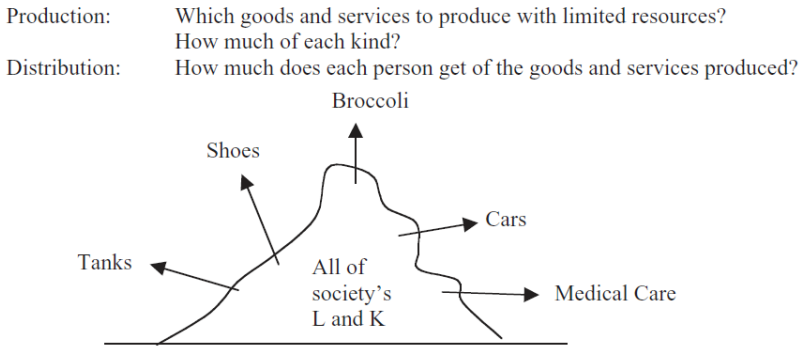


Figure III.1: Society's resource allocation problem.

This problem can be solved by tradition, authority, or the market system. Most people do not realize that the last way is a brand new approach. Of the roughly 200,000 years that humans have been on this planet, traditional and authoritarian arrangements are by far the usual ways to solve society's resource allocation problem. The market system emerged only in the last couple of hundred years.

This may seem incorrect given that money and prices have been around for a long, long time. A moment's reflection should convince you that trading is not a sufficient condition to determine whether a market system is being used to solve society's resource allocation problem. After all, societies in Biblical times had bazaars where people bought and sold goods and the former Soviet Union had stores where people paid rubles for groceries, but neither of these societies had market economies.

Cuba has had not one, but two currencies for decades (tourists must use the Cuban convertible peso or CUC, while Cubans use pesos), but no one would say it has a market economy. No, the presence of money is not a litmus test for a market system.

Societies based on the market system do not use supply and demand to allocate resources for every good and service. It is obvious that military equipment, such as tanks, in Figure III.1, are not produced according to perfectly competitive conditions via supply and demand. There is only one buyer, the government, and a few sellers (manufacturers of military vehicles). Likewise, no modern society uses the market system for medical care.

One could argue that all goods and services are regulated or controlled to some degree and, while there is some truth to this, it is also mostly true that many individual farmers decide what to grow based on market prices and this is a hallmark of the market system.

Unlike other ways of allocating resources, the market system allows each agent to decide how to use their labor and other privately owned resources. In a market system, individual resource owners respond to incentives. Unlike traditional and authoritarian systems, which rely on custom and command to get work done and products made, markets use the lure of gain to attract effort and capital.

The market system takes advantage of individual self-interest, using prices as incentives and signals. Whether self-interest is innate or learned is a deep philosophical question, but there is no doubt that players in a market system are driven to succeed and they calculate (and maximize, as they see it) before deciding what to do.

Although the market system, or simply markets, is the usual terminology today, other names have been used, such as capitalism, private property, free enterprise, price system, and *laissez faire*. Adam Smith's *An Inquiry into the Nature and Causes of the Wealth of Nations* (1776) is the first attempt at a comprehensive explanation of how a decentralized system that allows individual resource owners to decide where and how to use society's inputs can give a reasonable solution to society's economic problem. Notice the date—1776—before then, no one had to explain the market system because it did not exist.

This is not a history book, but you should be aware that the market system first emerged in western Europe around the 1700s, give or take a hundred years. It is difficult to pinpoint exactly where and when because there is no single event or marker. From close up, focusing on the 15^{th} to the 20^{th} centuries, it was a long, gradual transformation of society that took a few

hundred years. From far away, on a scale of centuries stretching back thousands of years, it was a sudden, explosive societal change.

One way to convey the stunning explosion in economic output before and after the emergence and spread of the market system is by examining the historical performance of different countries. We know the world was poor for millennia and then things changed fast, but economic historians have painstakingly compiled estimates of output per person to help us understand the evolution. Angus Maddison, for example, devoted his career to measuring long run economic growth around the world. The data are here: www.ggdc.net/maddison/Maddison.htm. There are output and population measures for countries all the way back to 1 AD. Figure III.2 plots real GDP per capita for 12 western European countries.

Figure III.2: Western Europe's historical economic performance.

The hockey stick depicted in Figure III.2 tells a remarkable story. Before the market system, although individual people (kings or other elites) could be rich, almost everyone knew only grinding poverty. Then, suddenly, something happened in western Europe that changed everything. Economies literally took off and the modern world was born. For an excellent, brief review of the rise of the market system, see the second chapter, "The Economic Revolution," in Robert Heilbroner's classic best-seller, *The Worldly Philosophers*.

The intellectual history of research on capitalism and markets is also quite fascinating. A great deal of work revolves around the idea of patterns emerg-

ing without direct, top-down control. Smith invoked the image of an "invisible hand" and Nobel Prize winning economist Friedrich Hayek coined the oxymoron "spontaneous order." In mathematics today, nonlinear dynamics and chaos theory focus on "self-organizing behavior." This idea, a pattern out of nothing, is critical to understanding the market system and the role of supply and demand.

Many have noticed that birds fly in a V, ants can form long chains and never seem to get stuck in traffic, and many animals (bees, locusts, and fish) swarm—they seem to act as if they had a collective mind. How do they do it? They do not rely on a single command center or leader to tell each one what to do. There are no orders given. There is no central direction. Instead, each individual follows simple rules that, taken together, produce a pattern or coherent order.

In computer science, the Game of Life is an artificial world that produces patterns from trivially simple rules. There are many examples on the web, such as this recent one, in honor of John Conway who recently passed away: b3s23life.blogspot.com/2020/01/a-gentleman-and-scholar.html. Search "game of life excel" for spreadsheet versions. LifeWiki has history, explanations, and many examples: www.conwaylife.com/wiki.

The point is that complicated movements of gliders and other objects that would seem to require central control can be generated in a decentralized way. Thus, the Game of Life is just another application—like supply and demand—of the general principle that patterns can be formed not only by top-down direction (like a marching band), but by decentralized systems with no controller at all.

The difficult idea to grasp is that supply and demand analysis is more than two intersecting lines. We are actually studying a pattern-generating system. Supply and demand is the model used by economists to explain how multitudes of interacting agents in markets can solve society's incredibly complicated resource allocation problem.

For the purposes of understanding how the market system works, an individual market will be defined by the commodity bought and sold. Thus, there is a market for broccoli and a market for engineers and a market for tutors. Every good and service allocated by the market system has a supply and demand.

By having a market for each product, we can use each individual market's equilibrium output as the market system's answer to the resource allocation problem. There is no central planner or controller who decides how much gas to produce. Buyers and sellers interact and establish an equilibrium price and quantity that determines how much of society's scarce resources are devoted to gas. If you think the price of gas is too high or we are allocating too much of society's scarce resources to producing gas, there is no one to call. The price and output are determined by the decentralized market system based on the operation of supply and demand.

The idea that a pattern emerges from the interaction of agents is fundamental to the market system. Watch the *The Invisible Hand and the Market System*, freely available at vimeo.com/econexcel/invisiblehand, to get a deeper understanding of these issues.

Organization

The organization of material in this part is straightforward, perhaps deceptively so. Figure III.3 shows the overall view of the book and the two chapters in this part.

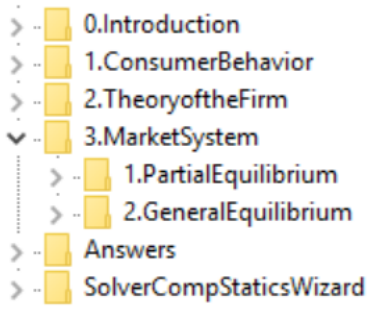


Figure III.3: Content map with focus on the market system.

The partial equilibrium chapter has sections devoted to both theory and applications (including government interventions) of supply and demand analysis. General equilibrium presents only exchange to give you a glimpse of how the model works.

It seems odd to say, but we will ignore a critical, fascinating part of the market system. Even a casual observer would notice that the market system exhibits high rates of innovation and technological change (which is what

produced the striking Figure III.2), but we will limit our analysis to exploring how the market system functions in a static environment in which the only issue is resource allocation (given constant technology).

In the conclusion, after you have mastered partial and general equilibrium analysis, we will return to the question of the dynamic analysis of the market system.

References

The epigraph is from page 23 of James Gleick, *Chaos: The Making of a New Science* (New York: Penguin Books, 1987). This serves as an excellent, friendly introduction to nonlinear dynamics and chaotic systems.

As mentioned in Chapter 3, the online version of *The Wealth of Nations* by Adam Smith, is freely avaiable at www.econlib.org/.

Robert Heilbroner, *The Worldly Philosophers: The Lives, Times, and Ideas of the Great Economic Thinkers* (New York: Touchstone, 1999, 7th edition, originally published 1953), remains one of the best summaries of the history of economics and the market system.

The Game of Life is associated with John Conway, who passed away in April of 2020. See www.nytimes.com/2020/05/16/science/john-conway-math.html for a tribute to this "mathemagician," with many links for further exploration.

Chapter 17

Partial Equilibrium

Supply and Demand

Consumers' and Producers' Surplus

Tax Incidence and Deadweight Loss

Inefficiency of Monopoly

Sugar Quota

Externality

Cartels and Deadweight Loss

Signaling Theory

> Credit for the ubiquitous demand and supply diagrams in principles texts is usually given to Fleeming Jenkin [1870]. ...For the first time, a real visual sense of the market is located. Pride of place goes to the equilibrium price.
>
> Judy Klein

17.1 Supply and Demand

We begin our analysis of the market system by making an obvious, but necessary point: A market demand (or supply) curve is the sum of individual demand (or supply) curves.

STEP Open the Excel workbook *SupplyDemand.xls*, read the *Intro* sheet, then go to the *SummingD* sheet.

The sheet has three consumers, with three different utility functions and different incomes. We assume the consumers face the same prices for goods 1 and 2. We set $p_2 = 10$, but leave p_1 as a variable to derive the individual demand curve for each consumer.

STEP Confirm, by clicking on a few cells in the range B18:D22, that the formulas in these cells represent the individual demand curves for each consumer. Notice that the graphs below the data represent the individual demand ($x_1{}^* = f(p_1)$) and inverse demand ($p_1 = f(x_1{}^*)$) curves.

Given individual demands, market demand can be found by simply summing the optimal quantity demanded at each price.

STEP Confirm, by examining the formula in cell E18, that market demand has been computed by adding the individual demands at $p_1 = 1$. The same, of course, holds true for the other points on the market demand curve.

Because we often display demand schedules as inverse demand curves, with price on the y axis, the red arrow (see your screen and Figure 17.1) shows that market demand is the result of a horizontal summation. At $p_1 = 5$, we read off each of the individual quantities demanded and add them together to obtain the market quantity demanded of 24.3 units.

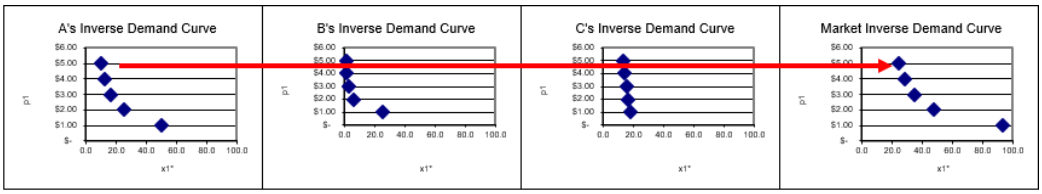


Figure 17.1: Horizontal summation to get market demand.
Source: SupplyDemand.xls!SummingD.

Supply works just like demand. We add individual supply curves (horizontally if we are working with inverse supply curves) to get the market supply curve. Because individual supply curves are P above AVC, we know that the market supply curve is simply the sum of the marginal costs above minimum AVC of all the firms producing the particular good or service sold in this market.

So the way it works is that each of the individual buyers and sellers optimizes to decide how much to buy or sell at any given price. The Theory of Consumer Behavior and the Theory of the Firm are the sources of individual demand and supply.

Once we have the many individual demand and supply curves, we add them up. So market demand and supply are composed of the sum of many individual pieces. Some consumers want a lot of the product at a given price, while others want less (or maybe none at all), but they all get added together to form market demand. The same is true for supply.

Initial Solution

The next step is obvious: market supply and demand are combined to generate an equilibrium solution that determines the quantity produced and consumed. This equilibrium solution is the market's answer to society's resource allocation problem.

The simple story is that price adjusts, responding to surpluses and shortages, until it settles down at its equilibrium level, where quantity demanded equals quantity supplied. This is the intersection of the two curves.

It is confusing, but true that in the supply and demand model, price and quantity are endogenous variables. How can price be endogenous—don't

consumers and PC firms take the price as given? Yes, they do and for individual buyers and sellers, price is exogenous, but, for the system as a whole, price is endogenous.

At the individual agent level, price is given and cannot be controlled by the agent so it is exogenous. But we are now at a different level. We are allowing forces of supply and demand to move the price until it settles down. Thus, at the level of the market, we say price is endogenous because it is determined by forces within the system.

It is worth repeating that equilibrium means no tendency to change. When applied to the model of supply and demand, equilibrium means that price (and therefore quantity demanded and supplied) has no tendency to change. A price that does have a tendency to change (because there is a surplus or shortage) is a disequilibrium price.

We can put these ideas in the same framework that we used to solve optimization problems. There are two ways to find the equilibrium solution and they yield the same answer:

1. Analytical methods using algebra: conventional paper and pencil.

2. Numerical methods using a computer: for example, Excel's Solver.

STEP Proceed to the *EquilibriumSolution* sheet to see how the supply and demand model has been implemented in Excel.

The information has been organized into three main areas: endogenous variables, exogenous variables, and an equilibrium condition. Excel's Solver will be used to find the values of the endogenous variables that meet the equilibrium condition.

As usual, green represents exogenous variables, the coefficients on the demand and supply curves.

Although price and quantity are both endogenous variables, price is bolded to indicate that the model will be solved by finding the equilibrium price and then the equilibrium quantity (demanded and supplied) is determined. This is similar to the approach we took with monopoly where we maximized profits by choosing q, then found P from the demand curve.

Finally, the equilibrium condition is represented by the difference between quantity demanded and supplied.

On opening, the price is too high. At $P = 125$, quantity demanded (Q_d) is 112.5 and Q_s is about 173. Thus, have a surplus ($Q_d < Q_s$) and, therefore, price is pushed down (as firms seek to unload unsold inventory).

STEP Use the scroll bar next to the price cell to set the price below the intersection of supply and demand. The dashed line (representing the current price) responds to changes in the price cell (B12).

Notice how the quantity demanded and supplied cells also change as you manipulate the price, which makes the equilibrium condition cell (B17) change.

With P below the intersection, the market experiences a shortage ($Q_d > Q_s$) and price is pushed up. The force in the market model is the pressure generated by surpluses (excess supply) or shortages (excess demand).

Obviously, the equilibrium price is found where supply and demand intersect. At this price, there is no tendency to change. The forces of supply and demand are balanced. We can find this price by adjusting the price manually and keeping our eye on the chart or by using Excel's Solver.

STEP Open Solver.

The Solver dialog box appears, as shown in Figure 17.2. Notice that the objective is not to Max or Min, but to set an equilibrium condition equal to zero. Notice also that P, price, is being used to drive the market to equilibrium and there are no constraints.

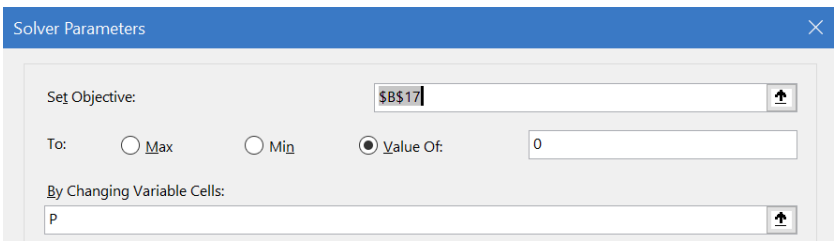


Figure 17.2: Solver dialog box.
Source: SupplyDemand.xls!EquilibriumSolution.

STEP Click Solve to find the equilibrium solution.

The chart makes it easy to see that Solver is correct. At $P = 100$, $Q_d = Q_s = 125$. Without a surplus or shortage, there is no tendency for the price to change and we have found the equilibrium resting point.

The equilibrium quantity, 125 units, is the market's answer to society's resource allocation problem. It says that we should send enough resources from the scarce, finite amount of inputs available to produce 125 units of this product.

We envision a supply and demand diagram for every product and the equilibrium quantity, in each market, is the market's answer to how much we should have of each commodity.

The analytical approach is easier than the math we applied for optimization problems because there is no derivative or Lagrangean. All we need to do is find the intersection of supply and demand.

Given either market supply and demand curves $Q = f(P)$ or inverse supply and demand functions, $P = f(Q)$, we find the equilibrium solution by setting supply and demand equal to each other.

The inverse functions in the Excel workbook are:

$$P = 350 - 2Q_d$$

$$P = 35 + 0.52Q_s$$

Setting the inverse functions equal to each other, we replace the Q_d and Q_s with Q_e because we are finding the value that lies on both of the curves:

$$350 - 2Q_e = 35 + 0.52Q_e$$

$$385 = 2.52Q_e$$

$$Q_e = \frac{315}{2.52} = 125$$

Substituting this solution into either inverse function yields $P_e = 100$.

We can also easily flip the inverse functions, solving for Q in terms of P, to obtain the demand and supply functions:

$$P = 350 - 2Q_d \rightarrow 2Q_d = 350 - P \rightarrow Q_d = 175 - \frac{1}{2}P$$

$$P = 35 + 0.52Q_s \rightarrow 0.52Q_s = P - 35 \rightarrow Q_s = \frac{1}{0.52}P - \frac{35}{0.52}$$

If we set demand equal to supply, using P_e to denote the common value we seek, we find the equilibrium price:

$$175 - \frac{1}{2}P_e = \frac{1}{0.52}P_e - \frac{35}{0.52}$$

$$175 + \frac{35}{0.52} = \frac{1.26}{0.52}P_e$$

$$P_e = \frac{175 + \frac{35}{0.52}}{\frac{1.26}{0.52}} = 100$$

Plugging this equilibrium price into either function gives $Q_e = 125$.

This work shows something obvious, but worth making clear: we can use $P = f(Q)$ functions to find Q_e, then P_e or we can use $Q = f(P)$ functions to find P_e, then Q_e. We get the same result either way since we are merely flipping the axes.

If you think using supply and demand functions ($Q = f(P)$) to get P_e and then Q_e is more faithful to what is going on in the market, you are a Marshallian for that is exactly how he saw markets functioning. And that is why P is on the y axis—so the reader sees it fluctuate up and down until it settles down to its equilibrium value.

We finish our work on the initial solution by pointing out that it is not surprising that numerical methods, using Solver, agree with the analytical approach. Given supply and demand for this product, we know that the market equilibrium solution would call for producing 125 units. The market system would, therefore, allocate the labor and capital needed to make this amount.

Elasticity

We can compute the price elasticity of demand and supply at the equilibrium price (the point elasticity) by applying our usual formula, $\frac{dQ}{dP}\frac{P}{Q}$. This time, we must use the demand and supply curves, $Q = f(P)$.

STEP Click the Show Point Elasticity button to see the calculation.

Although it has text wrapped around it, the number displayed for the price elasticity of demand is based on this part of the formula: $(-1/d1) * (P/Qd)$. With $Q_d = \frac{d_0}{d_1} - \frac{1}{d_1}P$, it is easy to see that $\frac{dQ}{dP} = -\frac{1}{d_1}$ and then we multiply by $\frac{P}{Q}$. Likewise, the price elasticity of supply is the slope of the supply function times the $\frac{P}{Q}$ ratio.

At the equilibrium price and quantity, demand is much more price inelastic than supply. This does not matter right now, but it will in future work.

STEP With $P = 100$, click on the price scroll bar and watch the price elasticities. Keep clicking until you set $P = 125$.

As you increase price, the elasticities change. Even though the slopes are constant, the supply and demand elasticities change because the $\frac{P}{Q}$ ratio is changing. Multiplying the slope by a price-quantity coordinate produces a percentage change measure of responsiveness.

The price elasticity of demand at $P = 125$ is -0.56 means that a 1% increase in the price leads only to a 0.56% decrease in the quantity demanded. This means demand is not very responsive since the percentage change in quantity is less than the percentage change in the price. Notice, however, the demand is more responsive at $P = 125$ than it was at $P_e = 100$.

We will see in future applications of the supply and demand model that the price elasticities play crucial roles. For now, remember that slope and elasticity are not the same and that the price elasticity tells us how responsive quantity demanded or supplied is to a change in the price.

Economists can be sloppy and say things like "demand is elastic" or "inelastic supply." This, of course, is nonsense. All downward-sloping, linear, inverse demand curves that cut both axes have elasticities that range from negative infinity at y-intercept to zero at the x-intercept. Statements like "demand is elastic" typically refer to a specific, usually equilibrium, price.

Long Run Equilibrium

Another concept at play in the model of supply and demand is that of long run equilibrium.

In the long run (when there are no fixed factors of production), a competitive market has another adjustment to make. In addition to responding to pressure from surpluses and shortages, the market will respond to the presence of non-zero profits.

The story is simple. Excess profits (economic profits greater than zero) will lead to the entry of more firms. This will shift the inverse supply curve right, lowering the price until all excess profits are competed away.

If the long run price is too low, firms suffering negative profits will exit, shifting the inverse supply curve left and raising prices. Thus, a long run competitive equilibrium has to look like Figure 17.3.

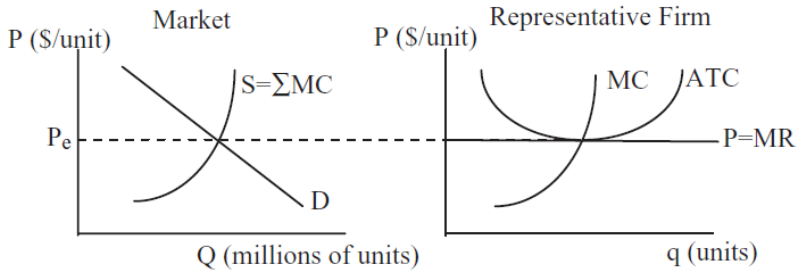


Figure 17.3: Long run equilibrium.

The left panel in Figure 17.3 shows supply and demand in the market as a whole, while the right panel depicts a single firm that is just one of the many firms in this perfectly competitive industry. The two graphs have the same y axis, but the scale of the x axis is different. A single firm can only produce a few units (q), but "millions" (an arbitrary number chosen just as an example) are bought and sold in the market (uppercase Q for emphasis). The idea is that there are many firms, each producing small amounts of the same output. In the aggregate, they make "millions" of units, but one individual firm produces only a tiny amount of the total.

Notice how the market demand curve is downward sloping, but the firm's demand curve is horizontal. This is the classic price taking environment in which a PC firm operates. Notice also that the market supply curve is the sum of the individual firm MC curves because individual firm supply is MC where $P > AVC$. We could chop off the bottom of the market supply curve (below P_e), but that would be confusing.

The long run adjustment process endogenizes the number of firms. This means that forces within the model determine how many firms there will be. This is not true in the short run, where the number of firms is assumed fixed (although they can shutdown if $P < AVC$) and the only adjustment is that market surpluses and shortages are eliminated by price movements.

Notice that the long run equilibrium price meets two equilibrium conditions:

1. Quantity demanded equals quantity supplied so there is no surplus or shortage in the market.

2. Economic profits are zero so there is no incentive for entry or desire to exit.

Long run equilibrium is even more fanciful and unrealistic than our abstract models of the consumer and firm. There has never been and never will be a market in long run equilibrium. Its primary purpose is as an indicator of where a market is heading.

The long run equilibrium model tells us that even though we are at an equilibrium with no surplus or shortage (such as with $P_e = 100$ in the Excel workbook), further adjustments will be made depending on the profit position of the firms. If profits are positive, entry will increase supply and lower price; while negative profits will lead to exit, decreased supply and higher prices.

In the Excel workbook, we do not know if the market is in long run equilibrium when $P_e = 100$ because we do not have a representative firm with its cost curves so we can determine its profits.

A key takeaway is that, like price, the number of firms is endogenous in the long run because there are forces in the model that determine its value. No one sets the number of firms. The interaction of buyers and sellers is generating the number of firms as an equilibrium outcome.

Comparative Statics

Comparative statics analysis with the supply and demand equilibrium model is familiar. Most introductory economics courses emphasize shifts in supply and demand. Here is a quick review, with special emphasis on equilibrium as an answer to society's resource allocation problem.

A change in any variable that affects supply or demand, other than price, causes a *shift* in the inverse supply or demand curve. A change in price causes a *movement along* stationary supply and demand curves. An increase in demand or supply means a rightward shift in inverse demand or supply.

For demand, the shift factors are income, prices of other goods related in consumption (i.e., complements and substitutes), tastes, consumers' expectations about future prices, and the number of buyers. The usual shift factors for supply include input prices, technology, firms' expectations, and the number of sellers.

As usual, comparative statics analysis consists of finding the initial solution, applying the shock, determining the new solution, and comparing the initial to the new solution. In the case of supply and demand, we want to make statements about the changes in equilibrium price and quantity. P_e and Q_e are the endogenous variables in the equilibrium model and we track how they respond to shocks.

For example, new technology lowered costs, What would that do to equilibrium price and quantity? We can use the *EquilibriumModel* sheet to see what happens.

STEP Make sure $P = 100$ so the market is in equilibrium, then click on the *s0* slider to lower the inverse supply curve intercept to 15.

The graph updates as you change the *s0* and a new, red inverse supply curve appears. The original, black line remains as a benchmark, but there is only one demand and supply at any point in time.

At $P = 100$, there is a surplus. We need to find the new equilibrium solution.

STEP Run Solver to find the new P_e and Q_e.

Figure 17.4 shows the result. The equilibrium price falls (from \$100/unit

to roughly \$84/unit) and the equilibrium quantity rises from (from 125 to about 133 units).

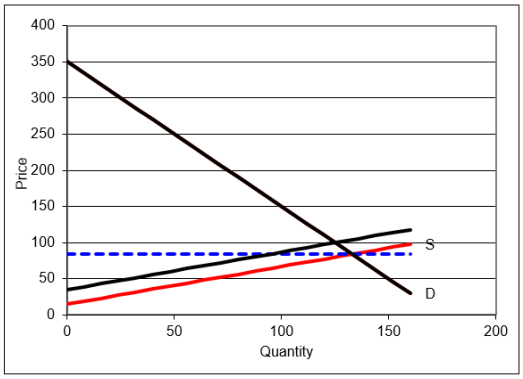


Figure 17.4: Comparative statics with the supply and demand model. *Source: SupplyDemand.xls!EquilibriumSolution.*

The decentralized market system has generated a new answer to society's resource allocation problem. Ceteris paribus, if a product enjoys a productivity increase from a new technology, making it cheaper to produce the product, the system will produce more of it.

This response makes common sense, but it is absolutely critical to understand that the increase in output is not decreed from on high. It is bubbling up from below—output rises because supply shifts and market forces lower prices and raise output.

We do not examine the equilibration process from the initial to the new solution when doing comparative statics analysis. We might directly converge to the new equilibrium, with price falling gradually until $Q_d = Q_s$. Or, price might collapse, falling below the equilibrium price, then rising above it, and so on. This would be oscillatory convergence.

With comparative *statics*, however, the focus is entirely on comparing the new to the initial solution. We may, in fact, be interested in the path to the new equilibrium, but that would take us into comparative *dynamics*—a topic for advanced microeconomics.

Applying Supply and Demand

To escape the usual trap of thinking of supply and demand in purely graphical terms, we apply the model to a real world example. We avoid graphs completely and focus on the mechanics and logic of supply and demand.

The market system uses supply and demand for outputs *and* inputs. This example focuses on labor, but there are many applications of supply and demand for capital—perhaps the stock market is the most prominent.

Consider that most fans of American football would not know the second highest paid position in the NFL. Everyone knows quarterbacks are the highest paid, but what position is second? Are star running backs, wide receivers, or maybe linebackers then next highest paid? No, the answer is left tackles—www.spotrac.com/nfl/positional/.

In *The Blind Side: Evolution of a Game* (2006), Michael Lewis explains that free agency, allowing players to sell their services to the highest bidder, radically altered the pay structure of the NFL. How did this happen? Supply and demand.

First, Lewis (p. 33) explains, there is little supply for the left tackle position.

> The ideal left tackle was big, but a lot of people were big. What set him apart were his more subtle specifications. He was wide in the ass and massive in the thighs: the girth of his lower body lessened the likelihood that Lawrence Taylor, or his successors, would run right over him. He had long arms: pass rushers tried to get in tight to the blocker's body, then spin off it, and long arms helped to keep them at bay. He had giant hands, so that when he grabbed ahold of you, it meant something.
>
> But size alone couldn't cope with the threat to the quarterback's blind side, because that threat was also fast. The ideal left tackle also had great feet. Incredibly nimble and quick feet. Quick enough feet, ideally, that the idea of racing him in a five-yard dash made the team's running backs uneasy. He had the body control of a ballerina and the agility of a basketball player. The combination was just incredibly rare. And so, ultimately, very expensive.

In addition to low supply, there is high demand. The left tackle is charged with protecting the quarterback's blind side, the direction from which defensive ends and blitzing linebackers come shooting in, causing sacks, fumbles, and worst of all, injuries. Because the quarterback is the team's most prized asset, the left tackle position is a highly sought-after bodyguard.

But even more surprising than the fact that blind side tackles are the second highest paid players in the NFL is that this was not always the case. Lewis reports that for many years, linemen were low paid, as shown in Figure 17.5.

Linemen	$398,000
Wide receiver	$504,000
Defensive end	$551,000
Running back	$620,000
Quarterback	$1,250,000

Figure 17.5: NFL salaries in 1990, before free agency.
Source: Lewis, p. 227.

So, why do blind side tackles make so much money today? NFL players did not enjoy free agency until the 1993 season. Up to that time, players were drafted or signed by teams and could move only by being traded.

Then the players' union and team owners signed a contract that enabled free agency for players so they could move wherever they wanted. In return, the players agreed to a salary cap that was a percentage of league-wide team revenue. Free agency meant that a player could sell himself to the highest bidder—in other words, the market would operate to establish player salaries.

At first, everyone was shocked. Teams spent outlandish sums on unknown linemen. Players that most fans never heard of made millions. Then a starting left tackle for the Bills, Will Wolford, announced his deal: $7.65 million over three years to play for the Colts. No one had ever paid so much money for a mere lineman. Not only that, his contract stipulated that Wolford was guaranteed to be the highest paid player on offense for as long as he was on the team.

The NFL threatened to invalidate this outrageous contract. In the end, the deal was allowed, but the commissioner decreed that such terms in a contract could never be used again.

Lewis, pp. 227–228 (emphasis added), explains what had happened:

> The curious thing about this market revaluation is that nothing had changed in the game to make the left tackle position more valuable. Lawrence Taylor had been around since 1981. Bill Walsh's passing game had long since swept across the league. Passing attempts per game reached a new peak and remained there. There had been no meaningful change in strategy, or rules, or the threat posed by the defense to quarterbacks' health in ten years. There was no new data to enable NFL front offices to value left tackles—or any offensive linemen—more precisely. *The only thing that happened is that the market was allowed to function.* And the market assigned a radically higher value to the left tackle than had the old pre-market football culture.

Economics students around the world study supply and demand, but they think it is a graph. It is so much more than an X. It is a model that explains how pressures from buyers and sellers are balanced.

This example shows that markets value commodities by reflecting the underlying demand and supply conditions. Blind side tackles are worth a lot of money in the NFL. Before markets were used, they were grossly underpaid. There were no statistics for linemen like yards rushing or field goal percentage so they could not differentiate themselves. The market system, however, expressing the desires of general managers and reflecting the true importance of the blind side tackle, correctly values the position.

Markets are neither moral nor caring. They are a way to consolidate information from disparate sources. Prices are high when everyone wants something or there is very little of it available. For blind side tackles, with both forces at work, the market system was a bonanza.

Supply and Demand and Resource Allocation

This introduction to the market system via partial equilibrium showed how an individual market settles down to its equilibrium solution. Much of this material is familiar because most introductory economics courses emphasize supply and demand analysis.

There are two fundamental concepts, however, that are critical in gaining a deep understanding of supply and demand.

1. Supply and demand curves do not materialize out of thin air. They are the result of comparative statics analyses on consumer and firm optimization problems. In other words, supply and demand must be interpreted as the reduced-form solutions from utility- and profit-maximizing agents. Figure 17.6 drives this point home by adding representative consumer and firm graphs to supply and demand.

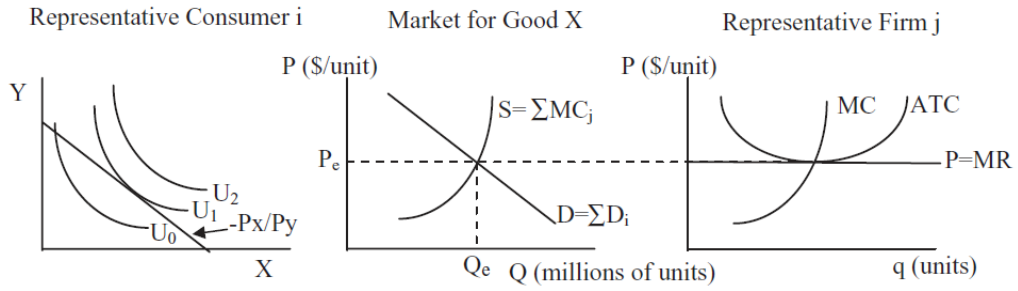


Figure 17.6: An overall view of supply and demand.

The notation in Figure 17.6 is awkward because we are combining consumer and firm theories which have their own individual histories. Thus, X in the left panel is the number of units of the same good that is produced by the firm in the right panel with label "q (units)." Likewise, P in the middle and right panels equals P_x in the left panel. Notational awkwardness notwithstanding, it is true that consumers generate demand for every good and service and the sum of individual demands is market demand. The same holds for supply and firms. Figure 17.6 is a great way to put it all together.

2 Supply and demand is a resource allocation mechanism. It is the equilibrium quantity that is of greatest importance in the supply and demand model because this is the market's answer to society's resource allocation problem. The price is the variable that drives a market to equilibrium, but it is Q_e that represents how much of society's scarce resources are to be allocated to the production of each commodity, according to the market system.

   A picture of this is in the *Intro* sheet. Now that you have finished this section, take another look at it and walk through it carefully.

Introductory economics students are taught supply and demand, but they do not understand that the market demand and supply curves are reduced forms from individual optimization problems. Deriving demand and supply is a bright line separating introductory from intermediate courses.

In addition, introductory courses stress price and equilibration (surpluses and shortages) as students learn the basics of supply and demand. Unfortunately, this means students miss the fundamental point: the equilibrium quantity is the decentralized, market system's answer to how much of society's scarce resources should be devoted to this particular commodity. There are graphs like Figure 17.6 for every good or service allocated by the market.

While the graphics in the *Intro* sheet emphasize the importance of Q_e, Figure 17.7 offers another way to explain what supply and demand is really all about. Filling in the mountain of society's finite resources with a checkerboard pattern conveys that the factors of production are individually owned and controlled. Each square represents the resources controlled by each person. Every person owns a tiny piece of the mountain and decides what do with that labor and capital.

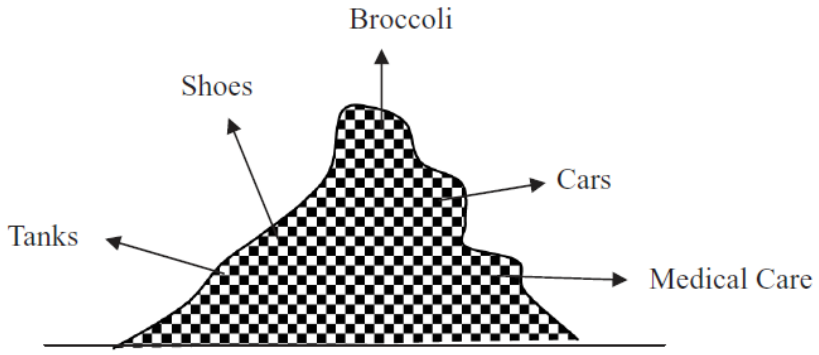


Figure 17.7: Individual ownership of resources.

Every product allocated by the market system has a supply and demand that attracts individual resources owners. Out of this cacophony of interactions, an equilibrium is found and resources flow to the production of an amazing variety of goods and services. This is the truly fascinating aspect of supply and demand. Each agent is self-interested and thinking only of their own gain, but the outcome of the market system establishes a pattern that answers the question of how to use scarce resources.

Of course, the checkerboard pattern in Figure 17.7 makes it seem like everyone controls equal shares, yet there is no question that some people own more resources than others. Inequality in the distribution of resources can be a serious obstacle facing the market system. It will not work well if resources are grossly unequally distributed.

This leads to another common misconception regarding equilibrium and desirability. Can we conclude, by virtue of the fact that the market is in equilibrium, that the market system has correctly solved society's optimization problem? Absolutely not. Equilibrium does not automatically equal optimal. The next section tackles this issue.

Exercises

STEP Click the Reset button in the *EquilibriumSolution* sheet to set the coefficients to their initial values.

1. Use the scroll bar in cell C7 of the *EquilibriumSolution* sheet to set the intercept of the inverse demand curve to 375. Use Excel's Solver to find the equilibrium solution. Take a picture of the answer and paste it in your Word document.

2. Solve the equilibrium model with $d_0 = 375$ via analytical methods. Show your work, using Word's Equation Editor as needed.

3. Because the intercept increased compared with the initial values of the parameters, we know there has been an increase in demand. How has the market responded to this shock? Is the market's response reasonable?

References

The epigraph is from page 111 of Judy Klein, "The Method of Diagrams and the Black Arts of Inductive Economics," published in Ingrid Hahne Rima, *Measurement, Quantification and Economic Analysis: Numeracy in Economics* (1995). As mentioned in section 4.3's References, credit for drawing supply and demand curves is usually given to Jenkin in 1870 and then Marshall in 1890 made the diagrams popular. Klein reviews precursors and how graphs evolved and came to be so important in economics.

The basic questions about resource use that must be answered by society can be traced to Paul Samuelson's *Introductory Economics* textbook (first published in 1948) and Frank Knight's *The Economic Organization* (1933). "Samuelson boiled Knight's five functions down to three: i) What commodities shall be produced and in what quantities?, ii) How shall they be produced?, and iii) For whom are they to be produced? 'These three questions,' Samuelson adds, paraphrasing Knight, 'are fundamental and common to all economies.'" See Ross B. Emmett, "Frank H. Knight and The Economic Organization," Michigan State University Working Paper No. 0405–01, p. 16, papers.ssrn.com/sol3/papers.cfm?abstract_id=922531.

Michael Lewis' book, The Blind Side: The Evolution of a Game was a huge hit in 2006. It was made into a movie in 2009, winning a Best Picture nomination and the Academy Award for Best Actress for Sandra Bullock.

It follows that consumer's surplus is not a concept which can be attributed to Marshall as something rather peculiarly his own. All that belongs exclusively to him is the name.

R. W. Houghton

17.2 Consumers' and Producers' Surplus

Society's resource allocation problem is an especially important optimization problem. It is an easy problem to envision. Pile up of all of society's factors of production and then ask, "How should we use these resources? What should we make? How much of each product should be produced? How should we distribute the output?" These are questions about resource allocation.

An important idea is that of a constraint. Needs and wants by consumers far outstrip available resources. More of one means less of other goods and services.

The previous section showed how supply and demand establishes an equilibrium price and output. The latter is the market system's answer to the resource allocation questions.

Although we are not studying alternative resource allocation methods, it is worth pointing out that if supply and demand is not used, that does not make difficult choices go away. Scarcity means there is not enough to go around. We may decide we do not want to use markets to allocate scarce organs, but we will still need a mechanism to decide whose lives are saved.

This section changes the focus from how supply and demand works to an evaluation of the market system's solution. The approach is clear: We first consider what an optimal allocation would look like, and then check to see whether the market's allocation conforms to the optimal solution.

Finding the Optimal Quantity in a Single Market

To find the optimal solution, we conduct a fanciful analysis. Like the imaginary budget line we used to find income and substitution effects, we work out a thought experiment that actually can never be carried out.

Suppose you had special powers and could allocate resources any way you wanted? Your official title might be *Omniscient, Omnipotent Social Planner*, or OOSP, for short. You are omniscient, or all knowing, so you know every consumer's desires and every firm's costs of production. Because you are omnipotent, or all powerful, you can decide how much to produce of each good and service and how it is produced and distributed.

Because this is partial equilibrium analysis, we focus on just one good or service. The question for you, OOSP, is, "How much should be produced of this particular commodity?"

One way for you to answer this question is to measure the total gain obtained by the consumers and producers of the good. When we compute the gain, we subtract the cost of acquiring the product for consumers and, for firms, the costs of production. The plan is to compute the total net gain for different quantities and pick that quantity at which the total gain is maximized.

The notion of net gain, something above the cost that is captured by consumers and firms, is the fundamental idea behind consumers' and producers' surplus. *Consumers' surplus* is the gain from consumption after accounting for the costs of purchasing the product. *Producer's surplus* is the difference between total revenues and total variable costs. In the long run, it is profit.

We begin with producers' surplus because it is uncontroversial. We will see that consumers' surplus is problematic.

Producers' Surplus

At any given price, if sellers get that price for all of the units sold, they get a surplus from the sale of each unit except the last one. The sum of these surpluses is the producer's surplus. The sum of all of the producer's surpluses in the market is the producers' surplus, *PS*.

The location of the apostrophe matters. Producer's surplus is the surplus obtained by one firm. If the focus is on all of the firms, we use producers' surplus.

STEP Open the Excel workbook *CSPS.xls*, read the *Intro* sheet, then go to the *PS* sheet.

The sheet displays an inverse supply curve given by $P = 35 + 0.52Q_s$. The area of the green triangle is *PS*. To see why, consider the situation when output is 75 units and the price is \$74/unit.

The very last unit sold added \$74 to total cost (given that we know that the supply curve is the marginal cost curve). Thus, the 75^{th} unit sold yielded no surplus. In general, the marginal unit yields no surplus.

But what about the other units? All of the other units are *inframarginal* units. In other words, these are units below the marginal (last) unit and, in general, the inframarginal units generate surplus. The firm is receiving a price in excess of marginal cost for these units, from 1 to 74, and, therefore, it is reaping a surplus each of those units. We can add them up to get producer's surplus.

Consider the 50^{th} unit. The marginal cost of the 50^{th} unit is given by $35 + 0.52 * 50 = \$61$. The firm would have been willing to sell the 50^{th} unit for \$61, but it was paid \$74 for that 50^{th} unit. So, it made \$13 on the 50^{th} unit.

STEP Look at cell Q68. It reports the surplus generated by the 50^{th} unit, \$13, as we computed above. Look at cell Q28. It reports the surplus generated by the 50^{th} unit, \$33.80.

Cell R19 adds the surpluses from all of the inframarginal units. Notice how PS steadily falls from the first to the last unit. The key to PS is that all quantities are sold at the same price, but marginal cost starts low and rises. The firm makes a surplus above *MC* on all output except the last one.

Cell R19 differs from cells B19 and B21 because cell R19 is based on an integer interpretation of output. If output is continuous, then we can compute the *PS* as the area of the triangle created by the horizontal price and the supply curve.

Notice that cell B19 offers another way to understand *PS*. If supply is marginal cost, then the area *under* the marginal cost curve is total variable cost. Because marginal cost is linear, the computation is easy. If *MC* was a curve, we would have to integrate. Total revenue is simply price times quantity. Cell B19 computes $TR - TVC$, the excess over variable cost, which is the producers' surplus.

STEP If $Q_s = 95$, what is *PS*? Use the scroll bar in cell C12 to set quantity equal to 95.

At 95 units of units of output, *MC* is \$84.40. At that price, the 95$^{\text{th}}$ unit has no surplus. But all of the other, inframarginal units generate surplus, adding up to \$2,346.50.

STEP Explore other quantities and confirm that as output rises, so does producers' surplus.

Consumers' Surplus

The idea is the same. At any given price, if a buyer pays that price for all of the units bought, she gets a surplus from the purchase of each unit except the last one. The sum of these surpluses is the consumer's surplus. The sum of all of the consumer's surpluses is the consumers' surplus, *CS*.

STEP Proceed to the *CS* sheet.

Given the inverse demand curve, $P = 350 - 0.2Q_d$, we can easily compute *CS* for a given quantity as the area of the pink triangle.

At $Q_d = 95$, the price so consumers will buy 95 units is \$160/unit. The last unit purchased provides no surplus, but the inframarginal units generate *CS*. The area under the demand curve, but above the price, is a measure of the net satisfaction enjoyed by consumers.

Consumers' surplus comes from the fact that consumers would have paid more for each inframarginal unit than the price they actually paid so they get a surplus for each marginal unit.

STEP Use the quantity scroll bar to confirm that as output rises, so does consumers' surplus.

As mentioned earlier, there is a problem with consumers' surplus. We will finish how OOSP could use *CS* and *PS* before explaining the problem.

Maximizing *CS* and *PS*

Producers' surplus is the amount by which the total revenue exceeds variable costs and measures gain for the firm. Consumers' surplus also measures gain because it is the amount by which the total satisfaction provided by the commodity exceeds the total costs of purchasing the commodity.

Both parties, consumers and producers, gain from trade. This is why a trade is made—both buyer and seller are better off. When you buy something, you part with some money in exchange for the good or service. If the purchase is voluntary, you must value what you are getting more than what you paid for it or else you would not have bought it. Similarly, the seller values the money you pay more than the good or service or else she would refuse to sell at that price. The gains from voluntary trade are captured in the terms consumers' and producers' surplus.

Casting the problem in terms of surplus received by buyers and sellers leads naturally to this question: What is the level of output that maximizes the total surplus? After all, it is clear that as quantity changes the *CS* and *PS* also change.

Thus, OOSP is faced with the following optimization problem:

$$\max_q CS(q) + PS(q)$$

The idea is to maximize the gains from trade for all buyers and sellers. This problem can be solved analytically and numerically. We focus on the latter.

STEP Proceed to the *CSandPS* sheet.

This sheet combines the surpluses enjoyed by producers and consumers into a single chart, shown in Figure 17.8.

Understanding this chart is fundamental. We proceed slowly. The vertical dashed line represents the quantity, which OOSP controls and will choose so that $CS + PS$ is maximized.

There are two prices on the chart, one for the firm and the other for the consumer. The idea is that OOSP uses the quantity to determine the prices needed for firms to be willing to produce the output level and for consumers to want to buy that amount of output.

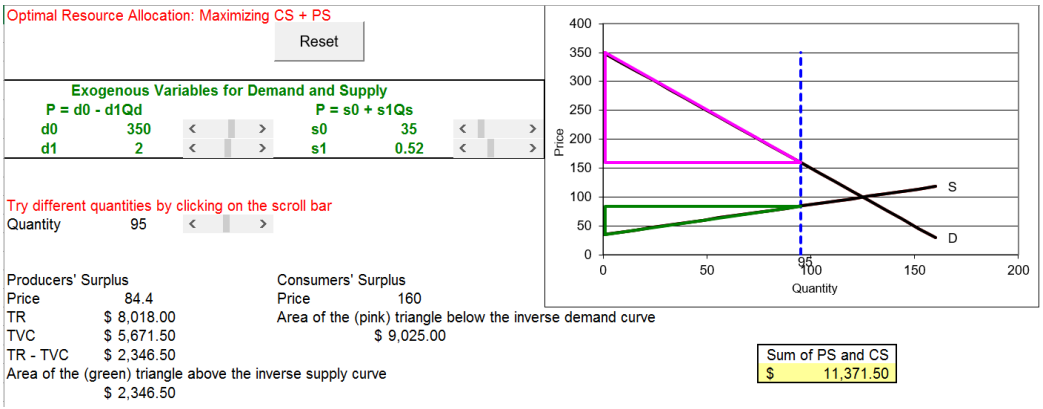


Figure 17.8: *CS* and *PS* at Q=95.
Source: CSPS.xls!CSandPS.

This is *not* an equilibrium model of supply and demand. OOSP cares only about choosing the optimal output. Price for consumers and firms is used only to compute surplus.

In Figure 17.8 (and on your computer screen), producers receive a price of $84.40 for each of the 95 units, yet consumers pay $160.00 per unit. Remember that OOSP, our benevolent dictator, has magical powers so she can charge one price to consumers and give a different price to producers. By adding the values in cells E18 and B21, we get the value in cell J20. It is highlighted in yellow and maximizing it is the goal.

STEP Click on the slider control (over cell C12), to increase output in increments of five units.

As output increases, *CS* and *PS* both rise.

STEP Continue clicking on the slider control so that output rises above 125 units.

Now the sum of *CS* and *PS* is falling. That is confusing because the two triangles are getting bigger. But once the price to consumers falls below the price to the firms, we have to pay the difference. This is explained below in more detail. For now, let's work finding optimal Q.

STEP Launch Solver and use it to find Q^*.

With an empty Solver dialog box, you have to provide the objective (J20) and changing cell (B12). We find that $CS + PS$ is maximized at $Q^* = 125$ units.

In other words, OOSP should order the manufacture of 125 units of this product, allocating the inputs needed from society's scarce resource endowment. This level of output maximizes the sum of *CS* and *PS*.

We have seen this number before. In the previous section, we found that the equilibrium solution was $Q_e = 125$ units. This means that the market's solution is the optimal solution. This is a remarkable result.

No one intended this. No one chose this. No one directed this. Supply and demand established an equilibrium output which answered the question of how much to produce and we now see that it is the same solution we would have chosen if our goal was to maximize consumers' and producer's surplus. This is truly amazing.

Deadweight Loss

If OOSP chooses an output level below 125 and charges a price to consumers based on the inverse demand curve and pays producers a price based on the inverse supply curve, it will generate a smaller value of $CS + PS$.

How much smaller? The amount of surplus not captured is given by the trapezoid between the consumers' and producers' surpluses. This area is called *deadweight loss*, *DWL*. It is a fundamental concept in economics and merits careful attention.

STEP Enter 95 in cell B12, then click the Show DWL button.

Not only do data appear below the button, but the chart has been modified to include a red trapezoid. The area of the trapezoid is displayed in cell D30.

STEP Click on cell D26.

The formula is simply the solution of the intersection of the supply and demand curves. We know this quantity is the solution to the problem of maximizing *CS* and *PS*.

STEP Click on cell D28.

This seemingly complicated formula is not really that hard. It displays the maximum possible total surplus. Two things are being added, *CS* and *PS*. The first part of the formula is *PS*: 0.5*(((s0_ + s1_ *D26) − s0_)*D26). It is half the height of the *PS* triangle times the length (or quantity produced). The second part of the formula uses the same area of the triangle formula to compute the *CS*: 0.5*((d0_ - (d0_ − d1_*D26))*D26).

STEP Click on cell D30.

The formula, = D28 − J20, makes crystal clear that deadweight loss is maximum total surplus minus the sum of *CS* and *PS* at any value of output. In other words, deadweight loss is a measure of the inefficiency of producing the wrong level of output in a particular market. Deadweight loss vaporizes surplus so that it disappears into thin air. Deadweight loss is pure waste.

STEP Click on the slider control (over cell C12) to increase output in increments of five units.

As you increase output, note that the deadweight loss falls as the output approaches the optimal quantity. There is no deadweight loss when the output is at 125 because this is the optimal level of output.

Another way to expressing the efficiency in the allocation of resources of the equilibrium solution is to say it has no deadweight loss. That is, no inefficiency in allocating resources.

As Q approaches Q^* we reach the maximum possible $CS + PS$ and DWL goes to zero. As Q keeps rising, past $Q > Q^*$, we get less total $CS + PS$ and deadweight loss rises. We get deadweight loss on either side of Q^*. The explanation for deadweight loss when $Q > Q^*$ is more complicated. Let's look at some concrete numbers.

STEP Set output above the optimal level, for example, $Q = 150$.

Your screen should look like Figure 17.9. It is true that *CS* and *PS* triangles are large, but with a higher price to firms than consumers, society has to pay for the difference. Once we account for this, the total gain is less than that at $Q = 125$ and we suffer deadweight loss, as shown by the red triangle.

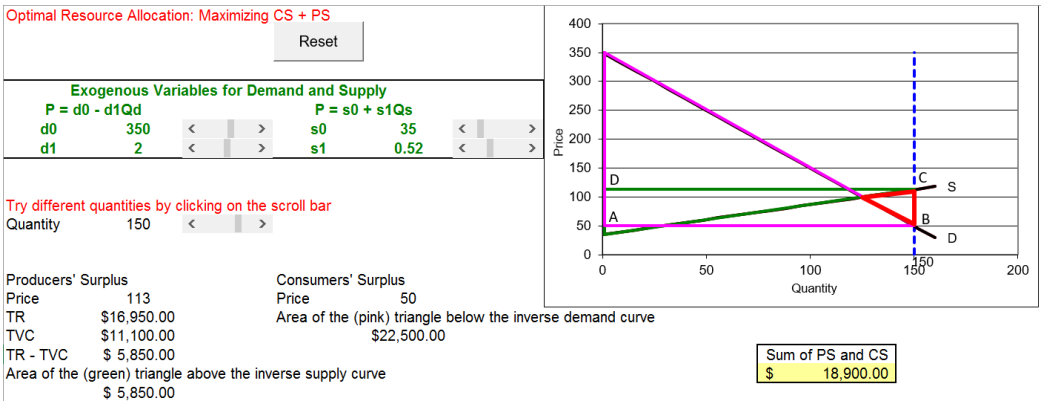


Figure 17.9: *DWL* at Q=150.
Source: CSPS.xls!CSandPS.

Figure 17.9 shows that it is possible to have sellers receive \$113 per unit sold yet have buyers pay only \$50 per unit sold, but someone is going to have to make up that \$63 per unit difference. The total value of the subsidy, \$63/unit times 150 units is \$9,340. This amount (rectangle ABCD in Figure 17.9) must be subtracted from the sum of *CS* and *PS*.

When we add everything up, we get a total surplus of \$18,900 at $Q = 150$, which is lower than the maximum total surplus. Cell J20 uses an IF statement to get the calculation right. The deadweight loss from producing 150 units is \$787.50 (cell D30).

The deadweight loss at $Q = 150$ is given by the area of the red triangle in Figure 17.9. The geometry is easy. We must subtract a rectangle with height 63 and length 150 from the sum of the pink *CS* and green *PS* triangles. This leaves the red triangle as the *DWL* caused by producing too much output.

There is one optimal output and at that value, deadweight loss is zero. Outputs above and below Q^* produce inefficiency in the allocation of resources because we fail to maximize $CS + PS$. This is called deadweight loss.

Price Controls

Price controls are legally mandated limits on prices. A price ceiling sets the highest price at which the good can be legally sold. A price floor does the opposite: The good cannot be sold any lower than the given amount.

To be effective, a price ceiling has to be set below and a price floor has to be set above the equilibrium price.

Most introductory economics students are taught that price ceilings generate shortages and price floors lead to surpluses. For most students, the take-home message is that market forces cannot push the price above the ceiling or below the floor so the market cannot clear and this is why price controls are undesirable.

It turns out that this is not exactly right. Although it is true that ceilings lead to persistent excess demand and floors prevent the market from eliminating excess supply, the real reason behind the unpopularity (among economists) of price controls is the fact that they cause a misallocation of resources.

STEP Proceed to the *PriceCeiling* sheet.

Suppose there is a price ceiling on this good at \$84.40. At this price, there is a shortage of the good because quantity demanded at \$84.40 is 132.8 units (cell B13) while quantity supplied is only 95 (cell B12).

The price cannot be bid up because \$84.40 is the highest price at which the good can be legally sold. Thus, with this price ceiling, the output level is 95. We know this is an inefficient result because we know $Q^* = 125$. This is the real reason why this price ceiling is a poor policy, not because it causes a shortage. The price ceiling fails to maximize total surplus.

To be clear, with this price ceiling, too few resources are allocated to the production of this good or service. There will be only 95 units of it produced, not the optimal 125 units. The fact that there is a shortage is true, but it is the misallocation of resources that is the problem.

While the misallocation of resources is easy to see since the quantity is wrong, deadweight loss is more complicated. It depends on the story about the price control and how agents react.

Suppose, for example, that market players are all honest so there is no illegal selling of the good above the maximum price. In other words, producers do not violate the law. Suppose further that the good is allocated via lottery so there are no lines of buyers or resources spent waiting. This means that consumers' surplus is now a trapezoid instead of a triangle.

STEP Click the Show DWL button.

As shown in Figure 17.10 (and on your screen), a rectangle has been removed from deadweight loss so it is now just the red triangle.

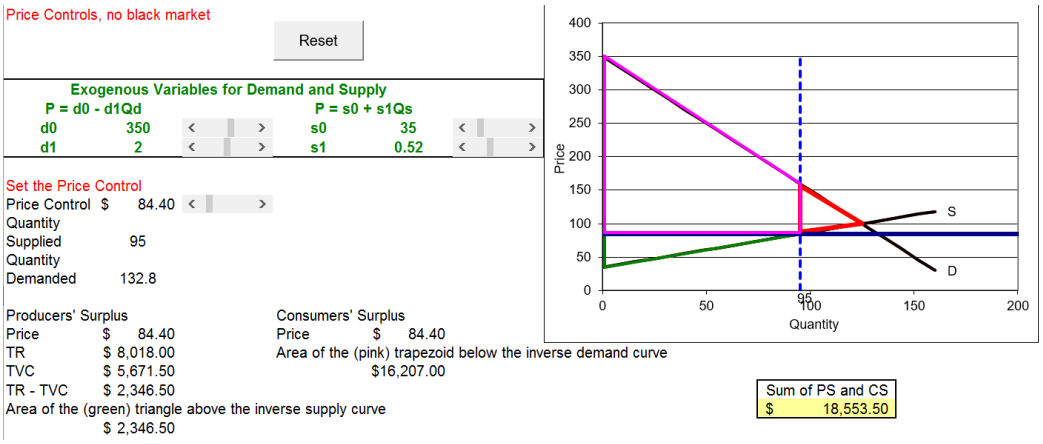


Figure 17.10: *DWL* with no illegal market.
Source: CSPS.xls!PriceCeiling.

In addition to the usual *CS* triangle in Figure 17.10, consumers enjoy the area of the rectangle computed by multiplying a price of \$160 (which is the price consumers are willing to pay for 95 units of the good) minus \$84.40 (the price consumers actually pay) times 95 units.

The good news behind this price ceiling with no cheating story is that the deadweight loss is much smaller than in the *CSandPS* sheet with $Q = 95$ because the lucky consumers who can purchase the good do not have to pay \$160/unit. The bad news is that there is still a deadweight loss of \$1,134. This is a measure of the inefficiency of the price ceiling with no illegal market.

Suppose instead that there are unlawful sales of the product at the illegal market price, \$160/unit (this is the most buyers are willing to pay for 95 units). Suppose, in addition, that somehow there are no wasted resources associated with this illegal market. No police investigations, court cases, or any other resources are spent on stopping criminal sales. Then the producers get the rectangle. With this idealized illegal market, the rectangle is transferred from consumers to producers, but the deadweight loss stays the same. The Q&A sheet asks you to demonstrate this.

If, as is almost surely true, illegal selling results in more resources being spent, then the deadweight loss is larger than the red triangle. Illegal activity often leads to violence (think of illegal drugs, which are a market with a price ceiling of zero) and we would subtract that from $CS + PC$ and thereby increase DWL.

Consider two other stories about the price ceiling. A limited set of buyers are given coupons to buy the product. To buy the good (at the legal price), you must have a coupon. If a rationing coupon scheme is used, the sellers of the coupons get the rectangle. The deadweight loss remains the same.

Suppose, finally, that a price ceiling is set and the good is allocated on a first-come-first-serve basis. In other words, buyers have to wait in line. With this story, the time, effort, and other resources buyers waste standing in line (or paying others to stand in line for them) must be subtracted from the total surplus. The deadweight loss rises. If the entire rectangle is lost, then the deadweight loss is the same as that in the *CSandPS* sheet when 95 units of output are produced.

Price controls are a popular way to modify market results. Unfortunately, from a resource allocation standpoint, price controls suffer from the fact that they fail to maximize total surplus. It is this property and not that they produce shortages that earn price ceilings criticism. We want the allocation mechanism to give optimal Q.

It is confusing that correctly measuring deadweight loss depends on the story, but do not be distracted by the many ways buyers and sellers can respond to price controls. The take-home message is that any deviation from Q^* means that the allocation scheme has failed. Deadweight loss, which gives a measure of the inefficiency in monetary units, depends on the specific implementation of the price control, but the fact that it is not zero is evidence that it has failed.

Caveat Emptor

"Let the buyer beware" is the meaning of the Latin phrase, *caveat emptor*. This idea from contract law is a warning to the buyer that they are responsible for what they are buying. The consumer needs to be careful so they aren't tricked or end up with a poor quality, unsuitable product.

Caveat emptor applies to deadweight loss. On the one hand, deadweight loss is a common way that economists measure inefficiency. It is based on the idea that the maximum total surplus is not attained from a particular output level. But users need to know what they are getting themselves into—deadweight loss has two glaring weaknesses.

The first has to do with our calculation of consumers' surplus. For technical reasons, restrictive assumptions about the utility function must be imposed. For example, a Cobb-Douglas utility function for individual consumers will not work because it has an income effect. A quasilinear utility function will work (no income effect), but it is unlikely that all consumers have quasilinear utility.

Consumers' surplus violates the rule that we should not make interpersonal utility comparisons. We are using the demand curve to add up dollar measures of the extra satisfaction that different people get from consuming a product. That is unsound and breaks a basic tenet of modern utility theory.

The second weakness stems from the use of partial equilibrium analysis. We are calculating deadweight loss based on the impact in a single market of a deviation in output from its optimal level. The focus on one market is too limited. If we apply too many or too few resources to the production of one good, we will cause deviations from optimal output for other goods and services. So, the deadweight loss computation based on one market is a lower bound. To get it exactly right, we would have to analyze effects on other markets and do a general equilibrium analysis.

Regarding deadweight loss, it is *caveat emptor*. Remember that deadweight loss measures inefficiency and it is commonly used in applied work, but it is not exactly right. The best way to think of deadweight loss is as an approximation.

Some economists are appalled at the thought of using deadweight loss. These most strident critics are usually more theoretically oriented. Economists who do empirical work are more likely to argue that deadweight loss is imperfect, but practically speaking, it is a useful way of measuring inefficiency.

Optimal Allocation of Resources

This is an important section. It introduced producers' and consumers' surpluses, which are key elements in the omnipotent, omniscient social planner's objective function.

The idea that there is an optimal level of output for each good and service is fundamental. From this idea we get the procedure for evaluating any allocation scheme or government policy: We compare an observed result to the optimal answer.

It is obvious that quantities below the intersection of supply and demand cannot be optimal because both CS and PS rise as Q increases. The situation with quantity above the intersection of supply and demand is more subtle. To get the calculation right, whenever quantity is above the intersection point, we must subtract from the sum of CS and PS a rectangle that is the difference between prices multiplied by quantity.

The most important and remarkable result from this section is that $Qe = Q^*$. This says that in a properly functioning market, the equilibrium quantity (which is the market system's answer to society's resource allocation problem) yields the socially optimal level of output.

Price controls lead to inefficient allocation of resources. The output generated does not match the optimal output. The deadweight loss associated with a price control depends on the story of how the particular implementation of the price control is enforced and responded to by buyers and sellers.

There is no question that deadweight loss is a linchpin of policy analysis. Countless estimates of deadweight loss and cost–benefit studies have been conducted. It is, however, flawed. Measuring consumers' surplus in value of money terms from a market demand curve in a partial equilibrium setting leaves us on very thin ice. Applications and estimates of deadweight loss should be seen as an approximation to the exact measure of the loss from the misallocation of resources (if such a measure exists).

While deadweight loss is flawed, the notion of a misallocation of resources is not. The idea that there is an optimal solution to society's resource allocation problem is perfectly valid. So is defining an allocation that deviates from optimal as a misallocation of resources. These are bedrock ideas in microeconomic theory.

This should mark the end of this section, but because there is so much confusion about equilibrium and optimal resource allocation, what follows is an attempt to provide some clarity.

Take a moment, before you begin reading the next section, to think about what supply and demand is really all about. What is the point? What are we trying to explain? Read the next section, maybe even repeatedly, to make sure you can answer these fundamental questions.

Equilibrium and Optimal Resource Allocation

The material below is being repeated for emphasis. The Theory of Consumer Behavior and Theory of the Firm are stepping stones to the $Q_e = Q^*$ result. Let's put things in perspective and explain why this is so fundamental.

In the 1700s, it is absolutely true that philosophers and deep thinkers of the day were baffled by the market system. There was active debate about how and why Europe and, especially, England was getting so rich. How could the unplanned, individual decisions of many buyers and sellers produce a pattern, much less a good result? It seemed obvious that a leaderless, fragmented system would produce chaos.

In the previous section, we saw that the equilibrium quantity, Q_e, generated by a properly functioning market is located at the intersection of supply and demand. The market uses a good's price to send signals to buyers and sellers. Prices above equilibrium are pushed down, whereas prices below equilibrium are pushed up. At the equilibrium solution, the price has no tendency to change and output is also at rest. The equilibrium level of output is the market's answer to how much of society's resources will be devoted to producing this particular good.

Our work in this section on consumers' and producers' surplus takes a much different perspective on the resource allocation problem. Instead of examining how the market works, we have created a thought experiment, giving an imaginary social planner incredible powers. Given the goal of maximizing total surplus, OOSP would choose an optimal quantity, Q^*, that should be produced. If we produce less or more than this socially optimal amount, society would forego surpluses that would make producers and consumers better off. Producing the wrong Q yields deadweight loss.

If we compare the market's equilibrium quantity to the socially optimal quantity, we are struck by an amazing result: $Q_e = Q^*$. This critical equivalence means that we do not need a dictator, benevolent or otherwise, to optimally allocate resources. The market, using prices, can settle down to a position of rest where all gains from trade are completely exploited and the sum of producers' and consumers' surplus is maximized.

There is no guarantee, however, that $Q_e = Q^*$—there are conditions under which the invisible hand does not lead the market to optimality. We will see examples where the equality does not hold and the market is said to fail.

As you work on this section and this part of the book, do not lose sight of the main point: The market's ability to generate an equilibrium quantity that is socially optimal is nothing short of amazing and unbelievable. It is equivalent to geese flying in a V. A pattern is generated by the interactions of individuals with no awareness or intent to make the pattern.

Consider this hypothetical: we learn that broccoli cures cancer. Would we need a president, prime minister, or king to tell farmers to grow more broccoli? Of course not. Broccoli would fly of the shelves, its price would rocket, and farmers would *automatically* plant and produce more broccoli. They would not try to figure out what would be best for society, but simply respond to the market signal. That is what the supply and demand model is really all about.

Analogies from biology are many, but this one might be so shocking and different from anything you have seen before that it will convey why supply and demand is so fascinating to economists.

STEP Visit http://tiny.cc/siphonophore to learn about this creature and see it in action.

Exercises

1. From the *CSandPS* sheet, click the Reset button, then set $d_0 = 375$ and use Solver to find the optimal quantity. Take a picture of the cells that contain your answer and paste it in a Word doc.

2. Click the Show DWL button. Suppose there was a price ceiling of \$84.40. What is the story about price ceilings assumed by the chart and *DWL* computations on the sheet?

3. Suppose the government implemented a price support scheme (this is a type of price floor that is used frequently for agricultural products) where they only allowed 95 units to be produced. Cell E16 shows that the market price would be \$185. Compute the deadweight loss and explain it.

References

The epigraph is from the first page of R. W. Houghton, "A Note on the Early History of Consumer's Surplus," *Economica*, New Series, Vol. 25, No. 97 (February, 1958), pp. 49–57, www.jstor.org/stable/2550693 A French engineer, Jules Dupuit (pronounced doo-pwee) presented the idea of *utilite relative* in 1844, but Alfred Marshall independently rediscovered and popularized the notion of consumer's surplus.

Almost immediately after Marshall introduced consumers' surplus, the concept came under attack. It has survived the move from a cardinal to an ordinal perspective on utility and a variety of other criticisms. Economists know that *CS* is built on shaky foundations, but they often use it in practical, policy-oriented, real-world discussions. In a review of the state of *CS*, Abram Bergson concludes, "Despite theoretic criticism, practitioners have continued to apply consumer's surplus analysis through the years. As some have argued, that must already say something about the usefulness (as well as the use) of such analysis, but just what it says has remained more or less in doubt." See "A Note on Consumer's Surplus," *Journal of Economic Literature*, Vol. 13, No. 1 (March, 1975), pp. 38–44, www.jstor.org/stable/2722212

> Harberger triangles, now common fare, were once rare delicacies. ...While the theory of deadweight loss measurement was well-established by the 1950s, economists very rarely estimated deadweight losses prior to the appearance of Harberger's work.
>
> James R. Hines, Jr.

17.3 Tax Incidence and Deadweight Loss

Many goods and services are taxed. Sales taxes (also called value added or *ad valorem* taxes) are a percentage of the monetary amount spent; quantity taxes are levied per unit bought. Quantity taxes are applied, for example, to gasoline, alcohol, and cigarettes.

In chapter 3.4, we examined cigarette taxes. It was shown that, for a particular consumer, lump sum (fixed amount) taxes are better than quantity taxes. In this section, we turn from an analysis of taxes on the individual to their effect on society and the resource allocation problem.

We will use supply and demand in a partial equilibrium setting to evaluate the effects of taxes on goods and services allocated by the market. We work with quantity taxes because our linear supply and demand curves will shift vertically as the tax is applied. Sales taxes are harder to analyze, but the qualitative results we derive for quantity taxes carry over to sales taxes.

There are two basic issues:

1. *Tax incidence*: determining the tax split between buyer and seller.
2. *Deadweight loss*: evaluating the inefficiency generated by the tax.

Our work will show a counterintuitive proposition: It does not matter whether consumers or producers pay the tax. In the end, neither the tax burden nor the deadweight loss depends on who sends tax payments to the government.

Our approach to the second—and more important—issue relies on comparing the output after the tax is imposed to the socially optimal output (based on maximizing consumers' and producers' surplus). Deviations from optimality are said to be inefficient solutions to society's resource allocation problem. We will use deadweight loss to measure the inefficiency. This is known as *welfare analysis*, where welfare means the well-being of a person or group.

It Does Not Matter Who Sends the Tax Payment

Suppose you are renting an apartment for $700 a month. Suppose further that property taxes rise $100. If your landlord raises the rent to $800 a month and you agree, it is easy to see that you are paying for the entire tax increase. The landlord pays the property tax to the government, but you are bearing the burden of the tax.

But what if you refuse to pay the $100 increase and move out. The landlord cannot find anyone to rent the apartment for $800 and, eventually, agrees to rent the apartment for $725 a month to a new tenant. The computation of the tax burden is easy. The new tenant is bearing the burden of $25 or 25% of the tax increase, while the landlord's burden is $75 or 75%.

No matter what the rent ends up being, the landlord sends the tax payment to the government, but that does not answer the question of who is really responsible for the tax. The landlord may be able to shift some of the tax onto the renter.

It turns out that the elasticities of demand and supply determine who bears the burden. The more inelastic, or price insensitive, the higher the burden.

Tax incidence is the analysis of who bears the burden of a tax. In a moment, we will be working with complicated supply and demand graphs, but the analysis is basically the same as the story of the tenant and the landlord.

Supplier Pays

For most products, the supplier or firm is responsible for collecting the tax when the good is purchased and for sending in the tax payments to the government. This is what is meant by "supplier pays." Of course, we know that who collects and pays the tax is different from the tax incidence because anywhere from 0 to 100% of the tax may be shifted to the consumer.

The elasticities of supply and demand determine how the tax is split between consumer and firm.

STEP Open the Excel workbook *Taxes.xls*, read the *Intro* sheet, then go to the *SupplierPays* sheet.

The sheet has parameters for linear demand and supply curves. Initially, there is no tax so the equilibrium price is \$100/unit and the equilibrium quantity is 125 units. Cell B17 shows that the government collects no revenue and cell E17 shows that there is no deadweight loss (because the market's equilibrium quantity equals the socially optimal quantity).

The price elasticities at the *initial* equilibrium solution are $\epsilon_D = -0.4$ and $\epsilon_S = 1.54$, for demand and supply. The sum of the absolute values is 1.94.

STEP Click on the scroll bar next to cell B14 five times to impose a tax.

A red line appears on the chart and it shifts with each click. Five clicks will set the tax at \$50 and the spreadsheet will look like Figure 17.11.

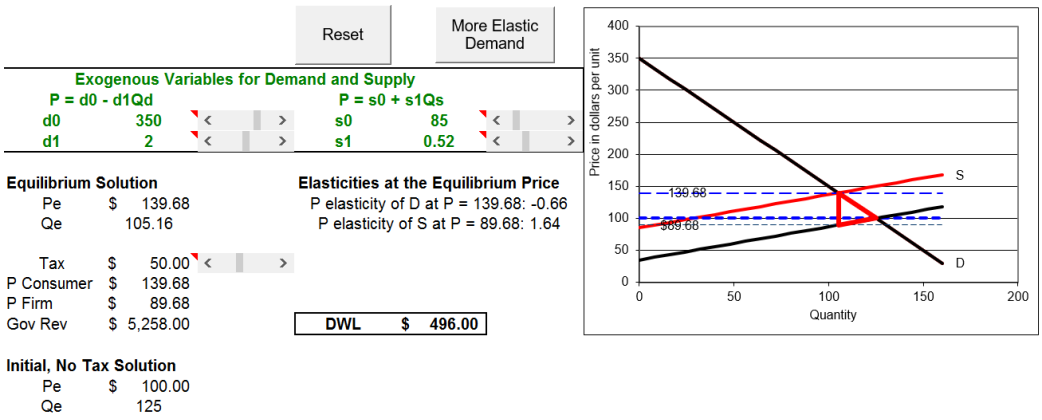


Figure 17.11: Supplier pays a \$50 quantity tax.
Source: Taxes.xls!SupplierPays.

The inverse supply curve has shifted up by \$50/unit because in order for the suppliers to offer a given quantity, they have to receive \$50/unit more than the original supply curve (without the tax). They will not get to keep the extra \$50 per unit—they have to send it to the government.

For example, to offer 125 units at the initial equilibrium solution, firms needed a price of \$100, but now they will need \$150/unit. The value of P is \$150 for $Q = 125$ with the red line in Figure 17.11. Every quantity has the same \$50 increase in price on the red line.

The spreadsheet displays the information we need to compute the tax incidence. We can see that the consumer is bearing the majority of the tax by

looking at the new equilibrium price. The dashed line (and cell B15) shows the new $P_e = 139.68$. We can compute the fraction of the tax borne by the consumer: $\frac{39.68}{50} \approx 79.4\%$. The supplier has managed to pass along all but about one-fifth of the tax to the consumer.

We can also use the absolute values of the *pre-tax* (initial) price elasticities to get the relative burdens for consumer and firm:

$$1 - \frac{0.4}{1.94} \approx 79.4\% \text{ and } 1 - \frac{1.54}{1.94} \approx 20.6\%$$

The *Tax Incidence Formula* to determine the share of the tax burden using demand and supply price elasticities is:

$$1 - \frac{\epsilon_i}{\epsilon_D + \epsilon_S} \text{ for } i = D, S$$

Notice that the formula drops the minus sign for the price elasticity of demand and for the rest of this section, we will mean the absolute value when we refer to the price elasticity of demand.

The elasticity values from the spreadsheet and the *Tax Incidence Formula* make clear that the lower the price elasticity, the higher the tax incidence. As $\epsilon_i \to 0$ (for either D or S), the burden (for D or S) goes to 100%. The consumer is paying four-fifths of tax in Figure 17.11 because demand is much more inelastic than supply at the initial equilibrium price.

We will discuss tax incidence in more detail below, but we turn now to the second, more important issue, the welfare implications of per unit taxes.

With a \$50 quantity tax, the *SupplierPays* sheet shows a deadweight loss of \$496 in cell E17. The deadweight loss can be calculated by finding the difference of the maximum possible surplus minus the surpluses enjoyed by the consumers, producers, and government. This is equivalent to the (red) triangle on the chart, which is also known as a *Harberger triangle*.

We proceed carefully. Consumers' surplus (CS) and producers' surplus (PS) after the tax is imposed have both been reduced by the trapezoidal shapes in Figure 17.12. Clearly, CS has fallen by much more than PS. More importantly, however, is the fact that the deadweight loss (DWL) is not the sum of lost CS and PS because we have introduced a third player—the government. They will get most of CS and PS lost in the form of tax revenue.

The total tax payments of \$5,258 is the area of the rectangle with height \$139.68 − \$89.68 = \$50 and length 105.16 units of output.

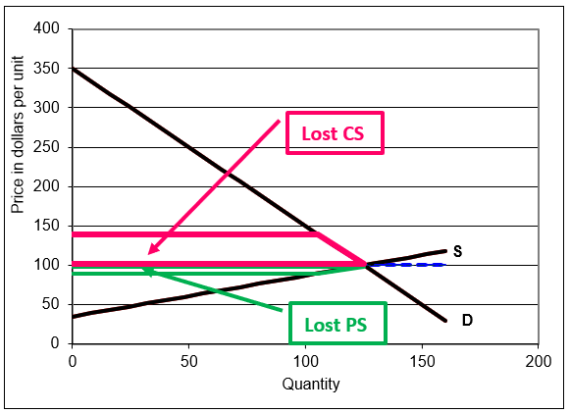


Figure 17.12: Lost *CS* and *PS* from the \$50/unit tax.
Source: Taxes.xls!SupplierPays!AN.

Once we recognize that the tax has lowered *CS* and *PS*, but that part of the surplus is captured by the government, we can see that the deadweight loss is the Harberger (red) triangle in Figure 17.13, with area displayed in cell E17. The surplus in the Harberger triangle vaporizes into thin air, captured by no one.

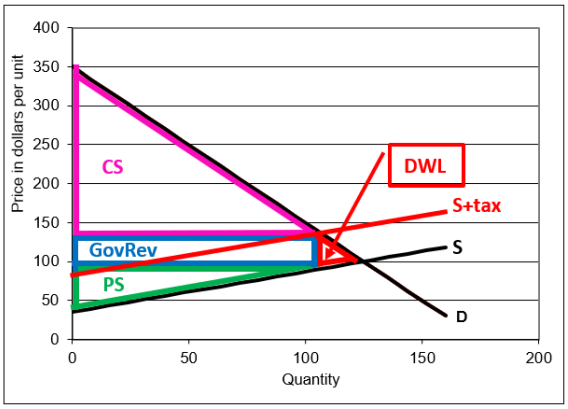


Figure 17.13: *CS*, *PS*, and *GovRev* from the \$50/unit tax.
Source: Taxes.xls!SupplierPays!AN.

The height of the Harberger triangle is the price the consumer pays minus the price received by the firm, which is called the *tax wedge*. This distance is the amount of the tax. When you clicked five times to impose the tax, you

could see the wedge expanding, creating a space between what the consumer pays and the firm receives.

The tax wedge takes surplus from consumers and producers, but this is not a problem. Presumably, the government is building schools, roads, and providing services. As long as someone gets the surplus, partial equilibrium surplus analysis counts it as a successful outcome.

Figure 17.13 shows, however, that the Harberger triangle goes to no one. This is a problem. Deadweight loss is surplus that simply vanishes. It is a loss of surplus that is not recouped by anyone.

The width of the *DWL* triangle is the distance from the new equilibrium quantity after the tax to the original equilibrium quantity. The bigger this distance, the greater is the distortion of the tax in terms of resource allocation.

STEP Click on cell E17 to see its formula. It simply computes the area of the red Harberger triangle.

We summarize and repeat a few key ideas. Deadweight loss is a dollar measure of the distortion caused by the tax—the "market with a tax" scheme is no longer producing the optimal quantity. This is a misallocation of resources. Deadweight loss represents gains from trades that are not being exploited. There is $496 in value that no one is getting. It is simply vaporized and disappears into thin air.

The rectangle formed by the tax times the equilibrium quantity (after the tax is imposed) is a transfer from consumers and producers to the government. This does not count as deadweight loss because someone (the government) is getting it. The key to understanding deadweight loss is that it accrues to no one—it is unclaimed surplus and, therefore, pure waste.

Demander Pays

Suppose that instead of the firm it is the consumer who is responsible for collecting the quantity tax when the good is purchased and for sending in the tax payments to the government. This may seem a little strange at first, but there are cases where this occurs.

For example, if you buy online and the seller does not charge you state and local taxes, you should pay those taxes. At the dawn of the internet, this gave online retailers a big advantage over brick and mortar stores that added sales and other taxes to the total. Few people pay taxes when they are not collected by the seller. Today, almost all online retailers include taxes.

For the purposes of comparing what happens when the buyer or seller pays the tax, forget about administrative costs or the fact that firms are much better tax collectors than consumers. We assume that consumers and firms will both comply and send the correct tax payment to the government even though that is obviously not true.

STEP Go to the *DemanderPays* sheet and impose a $50 tax. Pay attention to the screen as you click. You can watch the tax wedge emerge.

Figure 17.14 shows the result, with the *DWL* triangle displayed.

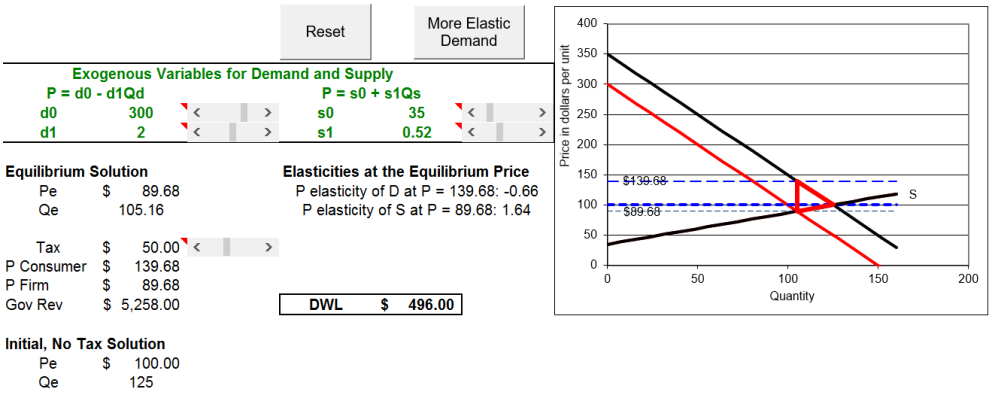


Figure 17.14: Demander pays a $50 quantity tax.
Source: Taxes.xls!DemanderPays.

This time, it is the demand curve that is shifting. Instead of the firm, it is the buyer who must compute the tax and send in the payments. A $50/unit tax will shift the inverse demand curve *down* (not up) by $50 because each consumer is willing to buy any given quantity for $50 less than before since she will have to pay an additional $50 to the government per unit purchased.

As before, a deadweight loss triangle appears when you impose the $50 tax. The tax drives a wedge between the total price the consumer pays and the amount the firm receives. This is the height of the triangle.

The deadweight loss triangle's width is the difference between the initial and new Q_e. The equilibrium quantity is driven down by the tax and, therefore, it no longer equals the socially optimal quantity. The tax causes an inefficient allocation of resources. The deadweight loss of \$496 is a measure of the inefficiency caused by the tax.

The tax incidence can be found by computing the share of the tax paid by the consumer versus the firm. The sellers receive a price of \$89.68 so they bear roughly \$10 of the \$50 tax. The consumer pays the firm \$89.68 and the government \$50 for each unit for a total price of \$139.68. The buyer's share of the tax is about 80%.

The government's revenue is the \$50 tax on each unit sold times the new equilibrium quantity, 105.16. This yields \$5,258 and can be represented as a rectangle in the supply and demand graph.

It is obvious that these numbers are the same as the suppliers pays scenario, but a fun and memorable way to show that it does not matter who pays the government is to toggle back and forth between the two sheets.

STEP Click the *SupplierPays* sheet tab, then click the *DemanderPays* sheet tab. Repeat this several times while keeping your eye on the screen. What do you notice?

The chart is different, of course, and the *d0* and *s0* parameters are different because the demand and supply intercepts do change based on who collects the tax for the government. But the price paid by the consumer, the price received by the firm, government revenue, and, most importantly, equilibrium quantity and deadweight loss are all exactly the same.

There is no doubt about it—tax incidence and deadweight loss do not depend at all on who physically collects and sends tax payments to the government (compliance being equal). If it does not matter if the buyer or seller pays the tax, then what do tax incidence and deadweight loss depend on?

Elasticities Drive Tax Incidence and Deadweight Loss

The relative price elasticities of demand and supply determine both the tax incidence (the distribution of the tax burden) and the deadweight loss (the measure of inefficiency in the allocation of society's resources).

Nothing else matters—certainly not who collects the tax, but also nothing else either. Price elasticities of demand and supply are all you need.

The more inelastic is demand at the initial equilibrium price, given supply, the more the consumer will bear the burden of the tax and the lower the deadweight loss. Likewise, the more inelastic is supply at the initial equilibrium price, given demand, the more the supplier bears the burden of the tax and the lower the deadweight loss.

We return to the apartment rent example to see how supply and demand analysis would work in an extreme case. If you agree to a \$100 increase in rent, your demand for apartments is perfectly inelastic in this price range. The price increase from \$700 to \$800 has no effect on the quantity demanded. In this case, you bear the entire burden of the tax and there is no deadweight loss. The situation is depicted in Figure 17.15.

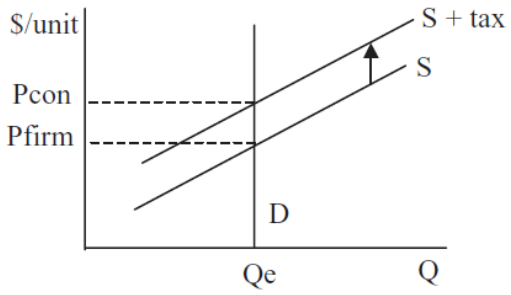


Figure 17.15: Tax effects with perfectly inelastic demand.

If you had to pay the property tax, you would be unable to shift it onto the landlord. In Figure 17.15, D would shift down, but it is a vertical line so it would shift on top of itself. The landlord would get \$700 from you (the initial equilibrium price) and you would pay an additional \$100 to the government.

Our Tax Incidence Formula yields the same result. With perfectly inelastic demand, $\epsilon_D = 0$. Thus, we have:

$$1 - \frac{\epsilon_D}{\epsilon_D + \epsilon_S} = 1 - \frac{0}{0 + \epsilon_S} = 100\%$$

This says the buyer bears the burden of the entire tax. Notice the formula does not have an input for who is writing the check to the government—that does not affect the outcome at all.

The formula also tells us that ϵ_S does not matter at all in the extreme case of perfectly inelastic demand. Any price elasticity of supply greater than zero leaves the buyer bearing the full burden of the tax.

The situation is reversed, of course, for the tax incidence if supply is perfectly inelastic. We would have a vertical S line that shifts up onto itself when the supplier pays the government. This leaves equilibrium price and quantity unchanged so the consumer pays the same amount as before and bears none of the tax burden. Once again, deadweight loss is zero.

Once again, the Tax Incidence Formula gives the same result. With $\epsilon_S = 0$, the ϵ_D in the numerator and denominator cancel and we get zero. This means the consumer bears no burden from a tax on a perfectly inelastically supplied good.

Of course, the main result that relative price elasticities determine tax incidence and deadweight loss applies in general and not just to these extreme cases. We can demonstrate this with the Excel workbook.

STEP To enable comparison, copy the *SupplierPays* sheet by right-clicking the sheet tab and selecting *Move or Copy*.) Select *SupplierPays* so the sheet is inserted before the *SupplierPays* sheet and check the *Create a Copy* box.

Excel inserts a new sheet in the workbook, named *SupplierPays (2)*. We will apply the same \$50 tax with a more elastic demand curve at the initial equilibrium price to see the effect on tax incidence and deadweight loss.

STEP Click the Reset button, then click the More Elastic Demand button in your new sheet.

A new, red inverse demand curve appears that is flatter, yet it goes through the initial equilibrium solution. The button simply sets the intercept and slope to 225 and 1, respectively. The price elasticity of demand at the initial equilibrium solution has risen (in absolute value) to -0.8 (as shown in cell E11).

It is important to not confuse slope and elasticity. The new, red inverse demand curve is more price elastic at $P = 100$ because it is flatter at that point. It is incorrect to say, however, that flatter lines are more elastic as a whole than steeper lines—both the initial and new inverse demand curves have varying elasticity all along the line. Thus, it does not make sense to say

that flatter lines are more elastic. Elasticity refers to a percentage change response at a point. Only at $P = 100$ do we know that elasticity is higher for the flatter, red inverse demand curve.

STEP Click the tax scroll bar five times to impose a \$50 per unit tax.

Figure 17.16 shows that the consumer bears less of the tax burden than before (but still more than the seller) and deadweight loss has risen.

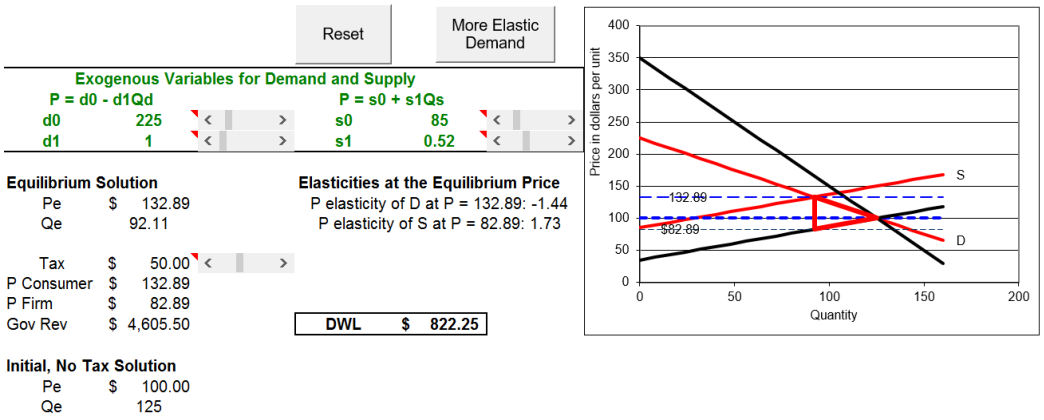


Figure 17.16: Tax effects with a more elastic D.
Source: Taxes.xls!SupplierPays.

With $\epsilon_D = 0.8$ instead of 0.4, ceteris paribus, the tax incidence on the consumer has fallen because the price has risen only to \$132.89 as opposed to \$139.68 on the *SupplierPays* sheet. So, the consumer bears \$32.89 of the \$50 tax or $\frac{32.89}{50} \approx 65.8\%$ of the tax. Notice that firms will now only net \$82.89 per unit instead of \$89.68 when $\epsilon_D = 0.4$. Suppliers tax burden rises to 34.2%.

The Tax Incidence Formula corroborates this result.

$$1 - \frac{\epsilon_D}{\epsilon_D + \epsilon_S} \text{ at } \epsilon_D = 0.4 = 1 - \frac{0.4}{0.4 + 1.54} \approx 79.4\%$$

$$1 - \frac{\epsilon_D}{\epsilon_D + \epsilon_S} \text{ at } \epsilon_D = 0.8 = 1 - \frac{0.8}{0.8 + 1.54} \approx 65.8\%$$

More importantly, deadweight loss has risen after the increase in the price of elasticity of demand from 0.4 to 0.8. Toggle back and forth from the original and new *SupplierPays* sheets to see that deadweight loss increases from \$496 to \$822.25.

While the height of the Harberger triangle has remained the same (the \$50/unit tax), the length has increased because the new equilibrium quantity is farther from the initial $Q_e = 125$.

If you toggle back and forth a few times, you can see how the more elastic demand curve is creating a *DWL* triangle that is longer, but with the same \$50 height. If you keep flattening the inverse demand curve (making sure that it passes through the initial equilibrium solution), the triangle keeps lengthening, but the height stays the same. A perfectly elastic (horizontal) D curve would produce the greatest deadweight loss possible.

STEP After thinking about it a bit, you can verify the claim above by using the control just to the right of the chart. Try all five scenarios.

With *Equal Burden* selected the demand and supply elasticities at $P = 100$ are the same so the \$50 tax is split evenly. The consumer pays \$125/unit and the firm receives \$75/unit.

The bigger drop in equilibrium output with more elastic demand is also responsible for the fall in government revenues. Instead of collecting \$5,258 in tax revenues, the government only gets \$4,605.50. It gets \$50/unit in both scenarios, but equilibrium quantity has fallen to 92.11 units with $\epsilon_D = 0.8$.

But neither the incidence of a tax nor the effect on government revenues is the highest priority issue. The top concern is the misallocation of society's scarce resources caused by taxation. It is this that leads to a theory of optimal taxation.

Optimal Taxation

Figure 17.15 shows why it makes sense to tax inelastically demanded goods. If we could find perfectly inelastically demanded or supplied goods, we would tax them because then we would not distort the allocation of resources.

Our goal is to raise government revenue for needed projects by causing the smallest misallocation of resources. Thus, the optimal tax is the one that has the least deviation of equilibrium output from optimal output, which is equivalent to minimizing deadweight loss.

Clearly, it is better, ceteris paribus, to tax goods with low price elasticities of demand or supply. In the introduction to this section, gasoline, cigarettes, and alcohol were mentioned as goods that carry quantity taxes. It is no surprise that these goods are quite price inelastic at their usual sales prices.

Granted, there may be other reasons to tax these products (and we will see one of them in the section on externalities), but to the extent that government seeks revenue from taxing individual products, it should tax those that will not lead to large deadweight losses.

There is no quantity tax on *Milky Ways*, a scrumptious chocolate candy. Obviously, the government could never generate the same tax revenue from *Milky Ways* as gasoline, but even if it could, with so many substitutes, *Milky Ways* must be very price elastic. A tax on *Milky Ways* would lead to a great fall in equilibrium output. Government revenue would be quite low and deadweight loss very high.

Elasticity Rules

Public Finance (also known as Public Economics) is a subdiscipline of economics that includes the study of government tax policy. The theory of optimal taxation focuses on the best way to tax. The analysis in this section says that quantity taxes should not be applied to goods that are relatively price elastic because the deadweight loss will be high. Instead, by taxing goods with inelastic demand or supply curves, government can raise needed revenue with a minimum of distortion in the allocation of society's resources.

This section also focused on the issue of tax incidence, who really bears the burden of a tax. This is a secondary issue compared to that of the optimal allocation of resources, but there is a surprising key result: It does not matter who collects the tax for the government (ignoring administrative costs and assuming equal compliance) because that party may be able to shift the tax onto someone else. Like deadweight loss, the tax incidence depends only on the elasticities of demand and supply. The more inelastic one of the curves is versus the other, the more that party will bear the burden of the tax. The Tax Incidence Formula sums this up conveniently:

$$1 - \frac{\epsilon_i}{\epsilon_D + \epsilon_S} \text{ for } i = D, S$$

Unfortunately, it is easy to confuse elasticity and slope. Do not fall into the trap of thinking that flat demand and supply curves are elastic and steep ones inelastic. If linear, slope is constant, but elasticity varies—for a linear, inverse demand curve, it rises as you go up and get closer to the vertical axis. Although you should not describe an entire curve as elastic or inelastic, you can correctly infer that where two lines cross, the flatter one is more elastic.

The French economist Frederic Bastiat (1801 - 1850) had a clever way of explaining what economists do. In his final essay, titled "What is Seen and Unseen," Bastiat argues we need to be aware of invisible costs and effects.

Taxes are a good example. It is easy to think that property taxes are paid by property owners, but this is simply not necessarily true. What is seen, a tax payment, is not the whole story. It is amazing, but true, that who pays the tax bill is irrelevant. It is also amazing that price elasticities, which are unseen, completely determine tax incidence and deadweight loss.

Exercises

1. Do we get the same result if we have consumers or firms pay the tax to the government with a perfectly inelastic supply curve? To support your answer, use Word's Drawing Tools to draw graphs. Explain the graphs and the result.

2. Use Word's Drawing Tools to draw a graph where supply is more inelastic than demand at the initial equilibrium price. Apply a quantity tax. Comment on the tax incidence and deadweight loss.

3. In 1937, when Congress set up the Social Security system, it was decided that firms and workers each pay half of the total tax so the tax burden is equally shared. Today, workers and employers each pay 6.2% of wages up to maximum that changes each year. Do you think that by each party paying the same tax the burden is equally shared? Why or why not?

4. Suppose the demand for labor is more elastic than the supply of labor at the equilibrium wage. Use Use Word's Drawing Tools to draw a graph that shows the tax incidence of the Social Security tax.

   Hint: You have to shift both demand and supply by the same amount, and then find the new equilibrium point.

References

The epigraph comes from page 168 of James R. Hines, Jr., "Three Sides of Harberger Triangles," *The Journal of Economic Perspectives*, Vol. 13, No. 2 (Spring, 1999), pp. 167–188, www.jstor.org/stable/2647124. Hines explains that the theory of deadweight loss dates back to Dupuit, Jenkin, and Marshall, but Harberger's papers in the 1950s and 1960s "illustrated the techniques, the usefulness, and the realistic possibility of performing such calculations, and in so doing, ushered in a new generation of applied normative work" (p. 168). For this reason, argues Hines, "Welfare loss triangles are 'Harberger triangles' because Harberger's papers measured them, did so in a consistent manner, and assisted and encouraged a host of others to do likewise" (p. 185).

Arnold Harberger published a number of papers, but perhaps his key contribution was "The Measurement of Waste," *American Economic Review*, Vol. 54, No. 3 (May, 1964), pp. 58–76, www.jstor.org/stable/1818490.

> Marginal cost pricing as a policy is largely without merit. How then can one explain the widespread support that it has enjoyed in the economics profession? I believe it is the result of economists using an approach which I have termed "blackboard economics."
>
> Ronald Coase

17.4 Inefficiency of Monopoly

Partial equilibrium analysis is based on the idea that each good and service with resources allocated via the market system has supply and demand curves. Prices signal quantities demanded and supplied and are pushed toward equilibrium by market forces. The equilibrium quantity is the market's answer to society's resource allocation problem.

If an omniscient, omnipotent social planner, OOSP, were to maximize the consumers' and producers' surplus of an individual good or service, she would explicitly order the production of the socially optimal amount of each good and service.

A critical result from this analysis is that a properly functioning market's equilibrium quantity equals the socially optimal quantity. This is what we mean when we say that a properly functioning market correctly solves society's resource allocation problem. There is no deadweight loss because the correct output is produced.

This section focuses on the following question: What happens if one of the goods is produced by a single seller (instead of the many individual firms that define perfect competition)?

In other words, we explore the welfare effects of monopoly. Our analysis is based on partial equilibrium and uses the tools of consumers' and producers' surplus. We evaluate monopoly by figuring out what a monopolist would produce, and then compare the monopoly output to the socially optimal output.

STEP Open the Excel workbook *MonopolyDWL.xls*, read the *Intro* sheet; then go to the *PC* sheet.

The linear demand and supply curves have the same parameter values used in previous examples. The equilibrium price is $100, which yields an equilib-

rium output of 125 units. Because the socially optimal level of production is also 125 units, the market yields an efficient allocation of resources.

Notice that at the socially optimal and competitive market solution, since supply is the sum of firm's marginal costs, we know that aggregate marginal cost equals demand. This is called *marginal cost pricing* and is indicative of a socially optimal solution. We will see in a moment that monopoly does not share this property.

The Monopoly Solution

Suppose all of the firms that produce a product in a perfectly competitive market were to merge into a giant, single firm. We assume that the cost structure stays exactly the same. In other words, the supply curve, which was the sum of the individual marginal cost curves, now becomes the monopolist's marginal cost curve.

Assuming that the costs of many firms would be the same costs faced by a single firm is a stretch. After all, the monopolist needs only one CEO and one customer service hotline. In other words, there are likely to be economies of scale in administration, distribution, and other areas. We assume this away in our comparison of perfect competition and monopoly. The idea is that the only difference is in the impact on the observed output when we have many firms in competition versus a single firm.

The monopolist will behave differently than the many firms did because there is no competition. Unlike the competitive result, where price is determined by the interaction of many buyers and sellers, the monopolist will choose the profit-maximizing price and quantity.

Chapter 15 explained monopoly profit maximization. What is different in this section is that, after determining the output chosen by the monopolist, we want to evaluate it using the tools of partial equilibrium analysis.

Our path is straightforward: we will solve the monopolist's problem with analytical and numerical methods, then we judge the monopoly outcome.

We know the monopolist will maximize profit by finding that quantity where $MR = MC$. The former is given by the demand curve, but what about MC?

The MC function is given by the supply curve parameters in the PC sheet. Once a monopoly takes over, it does not have a supply curve, but it does have a marginal cost function, which is the same as the supply curve (because of our assumption that there is no difference in costs between a competitive industry and a monopoly).

Thus $MC = 35 + 0.52Q$ and we can derive demand from the demand curve, as we have done before:

$$TR = P(Q)Q = (350 - 2Q)Q = 350Q - 2Q^2$$

$$MR = \frac{dTR}{dQ} = 350 - 4Q$$

As expected, we see that MR has twice the slope of the demand curve.

To find the monopolist's optimal Q, we set $MR = MC$ and solve for Q^*:

$$350 - 4Q^* = 35 + 0.52Q^*$$

$$4.52Q^* = 315$$

$$Q^* \approx 69.69$$

To find P^*, we use the demand curve to compute the highest price obtainable for that quantity.

$$P = 350 - 2Q = 350 - 2[69.69] \approx \$210.62$$

STEP Proceed to the *Monopoly* sheet to use numerical methods.

The graph has been augmented with the MR curve and the supply curve is now labeled MC. The MR curve was always there, but perfectly competitive firms cannot exploit it.

The sheet shows the monopoly price and output in cells B15 and B16 based on the analytical solution. Before we examine the deadweight loss and surplus information, we confirm that numerical methods agree.

When you run Solver, notice that the Solver dialog box is set up to choose that quantity that sets cell B20 to zero. The initial output of 50 units is too low. The fact that $MR - MC$ is \$89 means that the 50th unit of output adds \$89 more in profits and, therefore, more should be produced.

STEP Run Solver to find the Q that sets $MR - MC$ equal to zero.

After running Solver, you should see that cell B20 equals zero and that the Solver solution agrees (not exactly, but practically speaking) with the analytical method. This is not a surprise.

We now arrive at the key moment. How to judge the monopoly solution?

Evaluating Monopoly

We know the monopolized market will have an optimal output of 69.69 units and a price of \$210.62/unit. The evaluation of this outcome is based on computing the consumers' surplus, *CS*, and producers' surplus, *PS*, generated by the monopoly, and then comparing it to the socially optimal result.

The socially optimal result, at $Q = 125$ units, yields \$19,688 of total surplus.

STEP Cell F19 displays \$15,625 of consumers' surplus. Click on the cell to see its formula: = 0.5*(d0_ − P)*Q. *P* and *Q* are named cells for the perfectly competitive solution of 100 and 125, respectively.

Cell F20 has producers' surplus at $Q = 125$. Cell F21 adds *CS* and *PS*. The total surplus of \$19,688 is the maximum surplus possible and it is obtained when 125 units are produced.

Now, consider what happens under monopoly.

STEP Cell I19 shows a dramatic drop in *CS*. Click on the cell to see its formula: =0.5*(d0_-Pm)*Qm. *Pm* and *Qm* are named cells for the monopoly price and output.

The monopolist has lowered output and raised the price, relative to the competitive solution. This has greatly reduced consumer's surplus.

Cell I20 shows producers' surplus. It has more than doubled from what it was when the market was competitive. Its formula is: =(Pm-I18)*Qm + 0.5*(I18-s0_)*Qm. The first part of the formula is a rectangle. The height is the monopoly price minus the *MR* (or *MC* given that they are equal). The width is the monopoly output. A large part of this rectangle—from the monopoly price to the perfectly competitive equilibrium price—used to

belong to the consumers. It has been taken by the monopolist and helps explain why *CS* and *PS* have changed so dramatically.

So, CS has fallen and PS has risen, what is the overall outcome? Cell I21 adds *CS* and *PS* under monopoly. The total surplus of \$15,833 is lower than the maximum possible surplus of \$19,688. The difference, \$3,855 (in cell I23), is the lost surplus due to monopoly. This is also known as the deadweight or welfare loss.

STEP Click the Show DWL in Chart button to see a visual presentation in the graph of the deadweight loss of monopoly. It is a Harberger triangle.

Figure 17.17 is a canonical graph in microeconomics. It shows that the monopoly output is too low (so too few resources are allocated to this market) and the deadweight loss or Harberger triangle is used to indicate the inefficiency generated by monopoly.

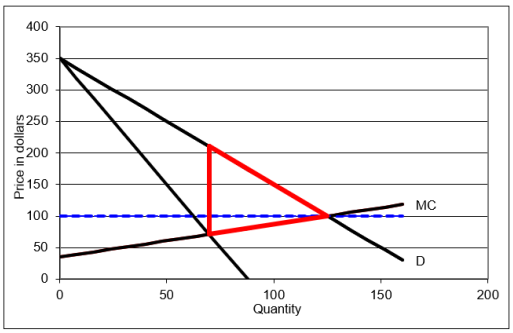


Figure 17.17: The deadweight loss from monopoly.
Source: MonopolyDWL.xls!Monopoly.

Because the monopoly solution does not equal the socially optimal output, we say there is a *market failure*. It is a failure in the sense that resources are not optimally allocated from society's point of view.

Inframarginal thinking can be applied to Figure 17.17. The basic idea is that all of the output in the range from the monopoly solution, roughly 70 units, up to the socially optimal output level of 125 units, exhibits unrealized gains from trade. For example, the marginal cost of producing the 100^{th} unit is $35 + 0.52 \times 100$, which equals \$87. The demand curve tells us that consumers are willing to pay up to \$150 for the 100^{th} unit. Clearly, the 100^{th} unit should

be produced because the additional satisfaction (as measured by willingness to pay) is greater than the additional costs of production.

The monopolist refuses to produce and sell the 100^{th} unit, however, because of an implicit restriction. Monopoly power allows the firm to set the price, but all units must be sold at the same price. Selling the 100^{th} unit at a price of \$150/unit means that all units must be sold at this price. Doing this would lower monopoly profit.

But the partial equilibrium welfare analysis critique of monopoly does not ride on the fact that monopoly forces consumers to pay higher prices than under a competitive market. The real problem with monopoly is that it produces too little output—it produces less than the socially optimal level. This causes too few resources to be allocated to the production of the monopolized good or service. We measure the amount of this inefficiency in resource allocation by the deadweight loss.

Yet another way to frame the inefficiency of monopoly is to focus on the fact that the monopolist produces where $MR = MC$ and this differs from $P = MC$ because MR diverges from the demand curve. A competitive market yields a socially optimal output because output is produced up to the point at which marginal cost equals the price (i.e., marginal cost pricing).

Figure 17.17 makes clear that the monopolist does not conform to marginal cost pricing. $MR = MC$ yields the output that maximizes profits, but $P = MC$ (where demand intersects supply or the aggregate marginal cost curve) is the socially optimal output. The monopolist is not interested in social optimality and, therefore, does not obey marginal cost pricing.

Elasticity Rules Again

In the previous section, we saw that the deadweight loss from a quantity tax depended on the price elasticities of supply and demand. The same holds true for monopoly.

STEP In the *Monopoly* sheet, display the red *DWL* triangle (if needed), and click the D More Elastic button .

Demand is flatter, while going through the same competitive equilibrium

point, $Q = 125$, $P = 100$. Thus, demand is more elastic at this point.

The button is actually a toggle. By clicking it repeatedly, you can switch back and forth from the original, more inelastic demand (price elasticity of -0.4 at $P = 100$) to the more elastic demand (price elasticity of -0.8 at $P = 100$).

STEP Click the D more and less elastic button a few times to convince yourself that the deadweight loss from monopoly is in fact larger when demand is more inelastic at $P = 100$.

While cell E17 shows that DWL is higher when demand is more inelastic at $P = 100$, we can make a graph that clearly shows this.

STEP Copy the *Monopoly* sheet and make the elasticity on the new sheet different than on the original sheet. Copy the chart in one sheet and paste it on top of the chart in the other sheet.

There is no fill in the chart so it is transparent. Your chart should look like Figure 17.18.

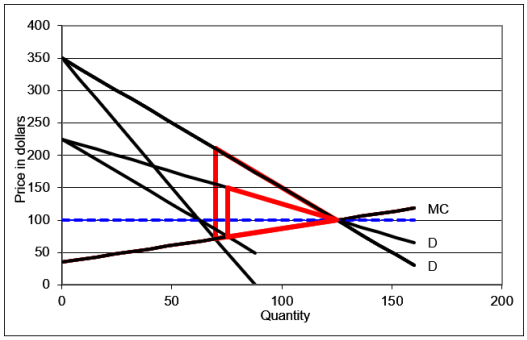


Figure 17.18: Comparing deadweight loss with different price elasticities of demand.

In Figure 17.18, the larger red triangle is the deadweight loss of \$3,855 in the initial case, with a price elasticity of demand of -0.4 at $P = 100$. The smaller red triangle is DWL with more elastic demand of -0.8. The DWL is lower, falling to \$1,870, when demand is more elastic.

Deadweight loss falls when demand is more elastic because the output does not deviate as much from the socially optimal result and the monopoly price is much lower. Hence, the Harberger triangle is both shorter and thinner.

Intuitively, the more inelastic is demand at given price, the greater is the monopoly power. A monopolist who enjoys demand that responds little to price is able to charge very high prices and the gap from marginal cost to demand for the inframarginal units will be large. This increases the deadweight loss from monopoly.

This example shows why economists use deadweight loss to measure inefficiency instead of simply the deviation in output from its optimal value. The monopolist does not change output by much when demand is more inelastic (-0.4), but the fact that consumers are willing to pay a lot more for the inframarginal units drives the large increase in deadweight loss.

Notice that the effect of elasticity on DWL is different than what we obtained for quantity taxes. In that case, more inelastic demand led to lower deadweight loss. The effect is reversed with monopoly, but the principle that elasticity rules remains true.

Monopoly and Price Discrimination

Although we usually assume a monopolist must charge the same price for all units sold, sometimes a seller can charge different prices for the same product. This is known as *price discrimination* and it enables profits to be even greater than when a single price is charged to all customers.

Charging different prices to see a movie in the afternoon versus the evening, different prices for coach versus first-class on a plane or train, and different net tuition to students (in the form of differing amounts of financial aid) are all examples of price discrimination. In each case, the firm is able to increase its profits by separating consumers into different groups and charging them different prices for the same good or service.

Sometimes firms try to slightly change the product so it isn't so obvious that the exact same thing is being sold at different prices. Offering first-class passengers pre-boarding and free drinks on a plane is an example of this. As is the bigger portions of a dinner versus lunch version of a dish at a restaurant. The difference in prices for the first-class and dinner versions of these products is not grounded in higher costs.

What is really going on here is coming entirely from the demand side. Some consumers are willing to pay more and firms are taking advantage of this.

People can get really upset at price discrimination. Dry cleaners can get in hot water when they charge different prices for cleaning men's versus women's clothing that is almost identical. It can be a fun game to spot examples of price discrimination.

There are three requirements for price discrimination to work:

1. Some degree of monopoly power (facing a downward sloping demand curve).

2. The firm must be able to segregate customers into groups (splitting the overall demand curve into subgroup demands).

3. There must be a way to seal the markets to prevent resale from the low-price to the high-price market, which is called *arbitrage*.

Assuming these requirements are met, we can construct a simple example that illustrates the essential logic of price discrimination. To increase profits, the idea is to separate price sensitive from insensitive consumers and then charge insensitive ones more.

STEP From the *Monopoly* sheet, click the Reset button and change cell E8 to 0 (zero).

This makes *MC* constant at \$35/unit and makes it easy to find the optimal solution and deadweight loss.

STEP With *MC* constant at \$35/unit, run Solver to find the monopolist's optimal solution.

Your screen shows that the monopolist will produce 78.75 units of output and charge a price of \$192.50. *CS* under monopoly is \$6,202 and *PS* is \$12,403.

STEP Click the Show DWL in chart button to display the Harberger triangle. Its area of \$6,202 is the *DWL*.

The fact that *CS* equals the *DWL* is not a coincidence. This is a property of linear demand and constant *MC*.

Now, suppose that this monopolist can separate the overall market demand, given by the inverse demand function of $P = 350 - 2Q$, into two separate

markets with two subdemands. For example, the two subdemands could be given by

$$\text{Market 1: } P = 450 - 6Q$$

$$\text{Market 2: } P = 300 - 3Q$$

The coefficients in the two separate markets must be consistent with the coefficients in the overall market inverse demand curve. The intercept and slope are not randomly drawn. If the price is zero, quantity demanded in Market 1 is 75 (= 450/6), while Market 2's quantity demanded would be 100 (= 300/3). The sum of the two is 175, which equals the quantity demanded at $P = 0$ using the overall inverse demand curve. At $P = 300$, Market 1's quantity demanded is 25 and Market 2's is zero, and this sum equals the quantity demanded using the overall demand curve.

How can a monopolist take advantage of the ability to separate the overall market into two sealed, separate subdemands?

The intuitive answer is simple: Instead of charging the same price, $192.50, to all customers, increase the price in the market with demand less sensitive to price and reduce it in the other market. The customers in Market 1 can be charged a higher price than those in Market 2. This will lead to greater profits.

We can see a concrete demonstration of this and figure out exactly what prices we should charge in our example.

STEP Proceed to the *TwoPriceDisc* sheet to see this plan in action.

Unlike the *Monopoly* sheet, there is no need to run Solver. The analytical solution has been entered and all cells and charts will instantly respond to changes in parameter values.

The top of the sheet shows how a perfectly competitive market would behave if there were two separate markets. Marginal cost pricing would result from competition so both markets would have the same price of $35/unit (cells B11 and B15). Market 2 would produce slightly more (B16) than Market 1 (B12), but the sum of the two (B20) would equal the perfectly competitive output of perfect competition for a single market. Thus, the ability to price discriminate, separating a single market demand into two separate, sealed subdemands, has no effect under perfect competition.

The outcome is different for monopoly. We begin by pointing out that the price elasticity of demand, while quite inelastic in both submarkets, is higher in Market 2 at the perfectly competitive price of \$35/unit.

The chart in the sheet is reproduced in Figure 17.19 and helps explain what is going on. This clever display shows the conventional monopoly graph for Market 1 on the right and uses the left side as a mirror for Market 2. Although the x axis shows output as negative on the left side, that is just a consequence of using Excel to draw the chart. Read the output as a positive number.

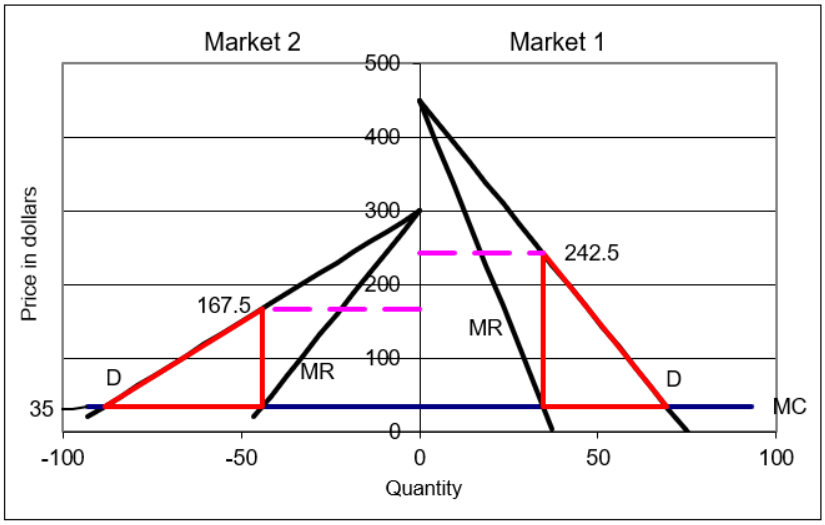


Figure 17.19: Price discrimination.
Source: MonopolyDWL.xls!TwoPriceDisc.

Figure 17.19 shows that the price discriminating monopolist will choose output where $MR = MC$ in each market, then charge the highest price obtainable for that output in each market. The price in each market is indicated by the dashed line and it is clear that price is higher in Market 1. This makes sense because consumers in Market 1 are less price sensitive so the monopolist takes advantage of this to generate higher profits.

STEP To easily compare the results of the single-price monopolist in the *Monopoly* sheet to the price discriminator in the *TwoPriceDisc* sheet, click the Show Single Price Monopoly button and notice that Market 1 is more price inelastic at the single-market monopoly price.

The price discriminator has the same total output, but it splits the single price into two prices. Cell B34 in the *TwoPriceDisc* sheet computes a weighted average of the two prices and it is higher than the single price of $192.50 charged by the conventional monopolist. This enables the two-price monopolist to make greater profits, as shown by the increase in PS from $12,403 to $13,028.

The monopolist will always be able to increase profits if it can split a market and keep the submarkets sealed off from each other, as long as the submarkets have different price elasticities. If so, a monopoly will charge the relatively more inelastic market a higher price and this is the source of the increase in profits.

Profits will continue to rise as markets are ever more finely subdivided. Amazon and other online retailers use your previous buying history, click behavior, and other information to serve up a personal price, just for you. Search for amazon+pricing+algorithm to learn more.

If you never heard about this before and think this is eye-opening or maybe even unfair, think about what colleges and universities do. They require their customers to provide detailed financial information about their ability to pay. They will, naturally, explain this as a benign effort to help the disadvantaged, but you should be glad your grocery store does not do this to you when you walk in the door.

The welfare consequences of price discrimination are not as clear. Comparing cells L38 and H34 shows that *DWL* has increased from $6,202 to $6,514 when the monopolist separated the markets and charged different prices. Of course, the monopolist does not care about deadweight loss; she is focused on maximizing profits. We, however, use *DWL* to evaluate outcomes and we would rather have the single market than the two submarkets exactly because deadweight loss is higher with price discrimination.

Unfortunately, these results do not generalize so we cannot say this will always happen. Higher *DWL* with price discrimination is guaranteed only for linear demand functions. In general, with nonlinear demands, we cannot state with certainty the effects on output and welfare. In other words, it is possible for output to rise and *DWL* to fall with a two-price discriminating monopolist. The effect on output and *DWL* depends on the shapes of the individual market demand curves.

For a concrete scenario of price discrimination improving welfare, consider the following from Scherer, (1970, p. 259):

> It is possible, for instance, that no physician would be attracted to a small town if he were required to charge the same fee to rich patients as to poor. Since profits can be increased by discriminating, the added revenue attainable through discrimination may be sufficient to make the difference between having a service provided and not having it.

Returning to the idea of subdividing the market more finely, there is a special case of price discriminating monopoly power that is a bit mystifying, but does yield a definitive result. The *perfectly price discriminating monopolist* has the ability to charge different prices for different output levels down to each individual consumer. This remarkable power enables the monopolist to sell every unit of output at the highest price the market will bear.

In the *Monopoly* sheet, the first unit goes for \$348, the 100^{th} for \$150, and the 125^{th} is priced at \$100. The perfectly price discriminating monopolist takes every bit of consumers' surplus, making the greatest profit possible, but does produce the socially optimal level of output. Thus, she has no deadweight loss!

Pondering the idea of perfect price discrimination and the fact that we would judge it as a socially optimal outcome cements the idea of surplus and deadweight loss. As long as someone, anyone, even if it is a single monopolist, gets the surplus, we count it as a successful outcome. Deadweight loss is tragic precisely because no one gets it. Deadweight loss vaporizes surplus and it disappears into thin air.

Monopoly Results in Market Failure

Monopoly leads to market failure because, to maximize profits, it restricts output and, therefore, this produces a misallocation of resources. The canonical monopoly graph (see Figure 17.17) has MR splitting off of D so that $MR = MC$ is less than the optimal output where $P = MC$.

While most people do not like monopoly because it charges higher prices than a competitive market, this is not why economists dislike the monopoly outcome. Partial equilibrium supply and demand analysis is based on maximizing consumers' and producer's surplus. The logic of deadweight loss

rides on the idea of waste. The monopolist does not take advantage of infra-marginal sales that would lower its profit, but increase society's total surplus. Any mechanism that generates deadweight loss is said to fail in the sense of not generating an optimal solution.

Another difference in outlook is that economists do not believe monopolists are inherently bad folks. The monopolist, like the perfectly competitive firm and consumer, is optimizing. Monopolies are in a position to improve their individual outcome and they take advantage. According to the economists, put anyone of us in the same position and we do the same thing. Do not blame the monopolist; blame the market structure for the deadweight loss.

We conclude with some advanced and heretical thinking. There is another, radically different view of monopoly that is based on the work of Joseph Schumpeter (1883 - 1950). He argued monopoly was actually a good thing because he had an evolutionary, dynamic view of capitalism. Striving for monopoly drives capitalism and monopolies are toppled by new firms in a process he named *creative destruction*. This oxymoron conveys Schumpeter's vision of capitalism, with entrepreneurs engaged in an epic battle of rising firms slaying established leaders.

Schumpeter's perspective is not that of solving society's resource allocation problem. He considered this static optimization problem to be uninteresting because it did not apply to the real world and it had been solved already. He did not believe that price competition was the real driver of capitalism's success. For Schumpeter, the serious open problem was how and why markets generated so much innovation and growth.

One important difference between mainstream economics and Schumpeter revolves around the government's role. Partial equilibrium analysis says monopolies should be broken up because they generate a misallocation of resources. Schumpeterians reject the need for government to intervene, arguing that dynamic competition will erode monopoly positions through entrepreneurial innovation.

Take an *Industrial Organization* course, an upper-level elective taught in most economics departments around the world, to learn more about monopoly, price discrimination, and Schumpeter's ideas.

Exercises

1. To punish a monopolist, your friend suggests applying a quantity tax on the monopoly's commodity. Is this a good idea? Explain why or why not, using the initial values of the parameters for supply and demand in the *Monopoly* sheet for a concrete example.

2. Another friend suggests a quantity subsidy to eliminate the deadweight loss caused by monopoly. The idea would be to shift down MC via the subsidy until output equaled the socially optimal output. Does this make sense?

3. Consider a monopoly that sells its output in two completely separated and sealed markets. Marginal cost is constant at \$35 per unit.

   Inverse demand in the two markets is given by

   $$P_1 = 200 - 2Q_1$$

   $$P_2 = 300 - 3Q_2$$

   (a) Solve this problem via analytical methods. Report optimal quantity and price in each market. Use Word's Equation Editor as needed.

   (b) Solve this problem with the *TwoPriceDisc* sheet. Enter the appropriate coefficients on the sheet. Take a picture of the results and paste it in a Word doc.

   (c) Which market has a higher price?

   (d) How does the price elasticity of demand in each market affect the price in each market?

   (e) Which market has greater deadweight loss?

   (f) How does the price elasticity of demand affect the deadweight loss?

   (g) The overall market demand is given by $P = 240 - 1.2Q$. How does price discrimination affect welfare loss in this case?

4. Suppose that, in the long run, average cost is decreasing throughout and marginal cost is below average cost, as shown in Figure 17.20. This is called a *natural monopoly*. The profit-maximizing level of output for the monopolist is where $MR = MC$. The socially optimal result is where $P = MC$.

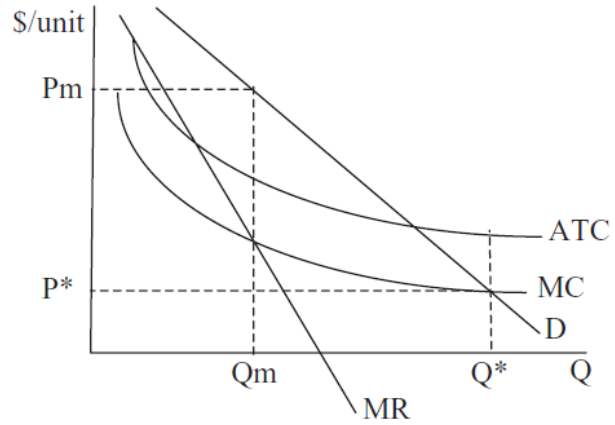


Figure 17.20: Natural monopoly.

(a) What is the problem with using competitive markets to achieve the socially optimal result in this situation?

(b) What government policy could be used to help the market reach the social optimum?

References

The epigraph is from page 19 of Ronald Coase, *The Firm, the Market and the Law* (paperback edition, 1990; originally published in 1988). Coase, a Nobel Prize winner for his work on transactions costs and property rights, criticized economics for its simplified, mathematical models that were completely stripped of real-world nuance and complexity. In "The Marginal Cost Controversy" (originally published in *Economica* in 1946 and reprinted in *The Firm, the Market and the Law*), Coase rejected the standard view that everywhere decreasing average cost (as in Figure 17.20) implied government intervention was the only solution.

Like Schumpeter, Coase has a broader, less mathematical view of economic analysis. Coase says on page 20 of *The Firm, the Market and the Law*,

> Blackboard economics is undoubtedly an exercise requiring great intellectual ability, and it may have a role in developing the skills of an economist, but it misdirects our attention when thinking about economic policy. For this we need to consider the way in which the economic system would work with alternative institutional structures. And this requires a different approach from that used by most modern economists

Visit www.hetwebsite.net/het/profiles/schumpeter.htm for more on Schumpeter. His most accessible work is *Capitalism, Socialism and Democracy*.

F. M. Scherer's *Industrial Market Structure and Economic Performance* came out in 1970 and was an instant classic Industrial Organization textbook. It was by far the most popular text of its day.

Today, almost all IO courses and books begin with a review of perfect competition and its polar opposite monopoly, but most of the course is about the complicated, fascinating, and vast area in between. Once you get past perfect competition and monopoly, the ways in which firms interact, compete, and strategize, is truly amazing.

We apply nonparametric regression models to estimation of demand curves of the type most often used in applied research. From the demand curve estimators we derive estimates of exact consumers' surplus and deadweight loss, which are the most widely used welfare and economic efficiency measures in areas of economics such as public finance.

Jerry A. Hausman and Whitney K. Newey

17.5 Sugar Quota

This section applies the tools of partial equilibrium analysis and deadweight loss to analyze import restrictions on sugar in the United States. Supply and demand analysis is shown to be a flexible, powerful tool.

Before analyzing the US sugar quota through the lens of surplus and deadweight loss, we take a crash course on sugar—production, pricing, and how import quotas on sugar are implemented.

Facts about Sugar

Everyone knows you can buy sugar in any grocery store and pour it into your coffee or use it to bake cookies. But there are many other kinds of white granulated sugars (like confectioners' sugar) and also brown and liquid sugars.

No matter the final form, "All sugar is made by first extracting sugar juice from sugar beet or sugar cane plants" (www.sugar.org/sugar/types/). Cane sugar is grown in warmer areas, whereas beets come from cooler climates. Once refined, you cannot easily tell the difference. Unless you are an expert, sugars from beet versus cane are perfect substitutes.

Some sugars are used only by industrial food manufacturers and not available in the grocery store. Home and commercial users can choose from many other sweeteners, such as high fructose corn syrup, and a long list of artificial sweetener options.

In addition to eating it, sugar can be made into ethanol and used to power a car. Most cars in Brazil are flex-fuel and growing huge quantities of cane has enabled Brazil to greatly reduce oil imports.

Many countries produce sugar. The United States grows both beet and cane sugar, but domestic production does not meet total demand so the United States imports sugar. Figure 17.21 is a subsection of a bigger table that shows sources of US sugar.

Table 1: U.S. sugar: Supply and use, by fiscal year (Oct./Sept.), May 2020

Items	2018/19	2019/20 (estimate)	2020/21 (forecast)
	1,000 Short tons, raw value		
Beginning stocks	2,008	1,783	1,273
Total production	8,999	8,024	9,005
Beet sugar	4,939	4,285	4,965
Cane sugar	4,060	3,740	4,040
Florida	2,005	2,100	2,105
Louisiana	1,907	1,513	1,800
Texas	147	127	135
Hawaii	0	0	0
Total imports	3,070	3,731	3,456
Tariff-rate quota imports	1,541	2,180	1,395
Other program imports	438	350	350
Non-program imports	1,092	1,200	1,710
Mexico	1,000	1,050	1,660
High-duty	91	150	50
Total supply	14,076.75	13,538	13,733
Total exports	35	35	35

Figure 17.21: US sugar sources.
Source: www.ers.usda.gov/webdocs/outlooks/98469/sss-m-381.pdf.

The numbers in the table come in units of short tons, raw value, STRV. A short ton is 2,000 pounds. Raw value means the dry weight of raw sugar. You get 1 ton of refined sugar (the white crystals you buy in the store) from 1.07 tons of raw sugar.

Beets are grown in many states so they are not all listed, but half of US beet production comes from the Red River Valley in Minnesota and North Dakota. The table shows the US domestic sugar industry is split roughly evenly between beet and cane, producing about 4,000 thousand STRVs (or 4,000,000 STRVs) from each crop.

Figure 17.21 makes clear that the United States imports a great deal of sugar, 3,070 thousand STRVs in 2018/19 and approaching 4,000 in 2019/20 (although this estimate was made before the covid 19 pandemic). So, roughly, the United States grows 2/3 of its own sugar and imports the rest.

Figure 17.21 shows that sugar is imported under several categories, the most important of which is the *tariff-rate quota*, TRQ. This is a complicated scheme for controlling the amount of sugar imported from different countries. The details are available at www.ers.usda.gov/topics/crops/sugar-sweeteners/policy.aspx.

A TRQ is a type of import restriction where a split tariff (or tax on imported goods) is employed. There is an extremely low tariff (zero or a nominal charge) applied to imports under a given amount (called the in-quota tariff) and a really high tariff applied to quantities imported beyond the given amount (so little is imported after the in-quota tariff is exhausted).

The TRQ was created in 1990 after multilateral trade agreements forced elimination of traditional quotas. In Europe, the EU Sugar Protocol is similar to the US TRQ system. The U.S. Department of Agriculture (USDA) runs the TRQ. The overall allotment is established by multilateral trade agreements and the USDA decides on the country allocations.

We can look at reports issued by the USDA as Excel spreadsheets to understand the TRQ.

STEP Open the Excel workbook *SugarQuota.xls*, read the *Intro* sheet, then go to the *TRQ* sheet and scroll around.

The data are constantly updated, so the specific numbers are not our chief concern. What matters is that column A has a list of countries and column O has *FY 2020 TRQ Original Allocations*.

As an example, consider the Dominican Republic. As of May 18, 2020, it had used 114,516 STRVs of its 185,335 TRQ allocation. The USDA has given every country in column A an amount that they can import. Beyond the TRQ amount, a hefty tax is applied so imports stop.

Outside of sugar producers and commercial food manufacturers that buy sugar, very few people in the US know or care much about this. In many countries, like the Dominican Republic, however, the US TRQ is a big news. When it is announced, there is intense media coverage and discussion.

If you scroll up and down, you will see that the Dominican Republic has the highest TRQ allocation, even bigger than Brazil, which is obviously a much larger country. What is going on here? In addition to protecting domestic US

sugar producers, the United States uses the TRQ as a major foreign policy lever, using allocations as punishments and rewards for foreign governments.

Now that we know a little about quantities of domestically produced sugar and imports, we turn to the price of sugar.

STEP Proceed to the *Price* sheet to see US and world raw sugar prices.

Figure 17.22 shows that prices have fluctuated over time, but US prices are always higher than world prices. The 1970s produced sharp spikes, followed by a period of calm until another spike during the Great Recession.

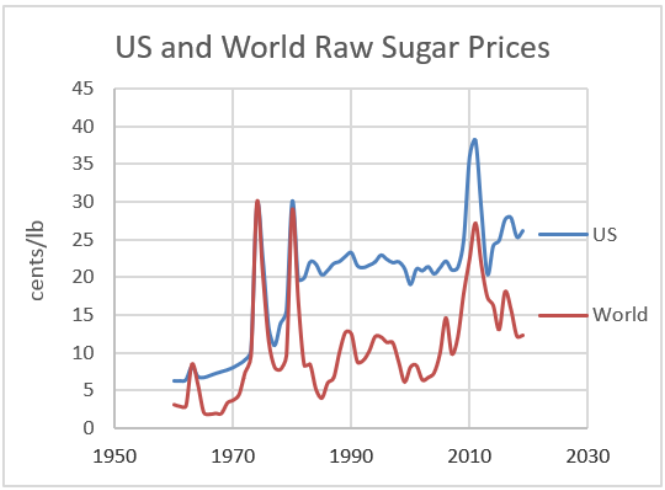


Figure 17.22: Nominal raw sugar prices.
Source: SugarQuota.xls!Price.

Since the TRQ was implemented in 1990, US sugar prices are consistently about 10 cents per pound higher than world prices. That might not sound like much of a difference, but think of it this way: US sugar prices are roughly double what others pay for sugar. If you make ice cream or candy or soft drinks or any one of the many products that uses sugar, doubling costs for this input is a really big deal.

STEP Review the price adjusted for inflation chart in the *Price* sheet.

The real price of sugar had been falling steadily, but it seems to have leveled off more recently. We can expect technological change (especially genetic engineering of cane and beet plants) to lower prices in the future.

We have ended our whirlwind tour of sugar production, the US TRQ system, and prices. Obviously, the sugar quota is causing higher prices for US consumers (including commercial buyers of sugar) and it benefits US producers. But we can say more and evaluate the US sugar quota by applying partial equilibrium analysis.

Supply and Demand for US Sugar

To analyze the effects of the sugar quota, we need estimates of demand and supply curves for sugar in the United States. Because we will work with linear functions, we need intercept and slope parameters for the demand and supply of sugar.

The USDA reports roughly 12,000 thousand STRVs of sugar are bought and sold in the United States for about 25 cents per pound for raw sugar. We assume the market is in equilibrium so we interpret these values as the equilibrium quantity and price.

There is a vast literature on sugar with countless estimates of demand, supply, and elasticities. Since this is an exercise in showing how partial equilibrium analysis works, we will use hypothetical demand and supply functions that are calibrated to the observed values in the US sugar market.

Our linear demand and supply curves are

$$Q_D = 15000 - 120P$$

$$Q_S = 400P$$

At $P = 25$ cents per pound, quantity demanded is 12,000 thousand STRVs (our equilibrium P and Q) and the price elasticity of demand is $\frac{\Delta Q_D}{\Delta P}\frac{P}{Q_D} = \frac{1}{-120}\frac{25}{12000} = -0.25$. That is quite inelastic and conforms with estimates of sugar price elasticities of demand. Although there are substitutes, in many recipes (especially for commercial products), precise amounts of sugar are absolutely required. The price elasticity of supply in our simple model is +1.

The inverse demand and supply curves are

$$P = 125 - \frac{1}{120}Q_D$$

$$P = \frac{1}{400}Q_S$$

You probably did not do this, but computing the quantity supplied from the supply curve with $P = 25$ gives $Q = 10000$. Something is wrong because the quantity demanded does not equal the quantity supplied. For sugar, we need to include imports.

Free Trade

We begin our partial equilibrium analysis of the US sugar quota in Fantasyland—we assume that there is no restriction of any kind on the importation of sugar.

STEP Proceed to the *FreeTrade* sheet to see how the market would work under a regime of no restrictions on imports.

Figure 17.23 reproduces the graph. The demand curve is straightforward, but the supply curve merits special attention.

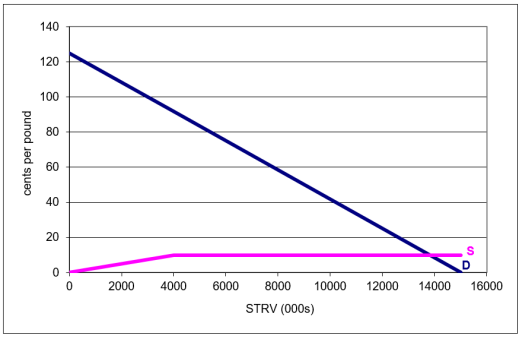


Figure 17.23: Supply and demand with hypothetical free trade.
Source: SugarQuota.xls!FreeTrade.

The first part of the supply curve (from the origin to the kink at $Q = 4000$) is domestically produced US sugar. As long as the price is below the world price of 10 cents per pound, the best, lowest cost US producers will supply the market.

Beyond 4,000 units (measured in thousands of STRVs for consistency with USDA TRQ units), world suppliers take over. It is assumed that the United States has access to as much sugar as it wants at the world raw sugar price of 10 cents per pound. Thus, the market would not continue to use US produced sugar beyond 4,000 units. Instead, supply would come from the perfectly elastic world supply curve.

US consumers (home and industrial buyers) would enjoy a 10 cent per pound price for raw sugar and the equilibrium quantity would be 13,800 thousands STRVs. Over 2/3 of sugar consumed would be imported.

The sum of US consumers' and producers' surplus would be more than $16 billion. In this properly functioning market, this is the maximum possible total surplus.

STEP Click on cells G33 and G34 to see the formulas used to compute *CS* and *PS*.

Notice that many US producers would be driven out of the market because they cannot make sugar at the low world price. Those US producers that remain (selling the first 4,000 units) would earn $400 million in producers' surplus under a free trade regime. As will be clear in a moment, this is an important number to keep in mind.

Incorporating an Import Quota

The TRQ system is too complicated to exactly implement in Excel so we model a simple quota that is easier to understand and acts similarly to the TRQ scheme.

STEP Proceed to the *ImportQuota* sheet to see what happens with an import quota on sugar.

As before, we focus on the supply curve. It is crucial to the analysis.

The *ImportQuota* sheet shows that the supply curve has an upward sloping part, then a flat part, and then it starts sloping up again. The first part is the US domestic supply curve. The lowest cost US firms will supply the market when the price is below the 10 cents per pound world price.

The flat part is the amount of imported sugar allowed. Cell H6 shows this amount is 2,000 units, so the flat segment is 2,000 units long.

The last, rising part of the supply curve is, once again, the domestic US supply curve. Once the quota is filled and no more foreign sugar is allowed into the United States, domestic producers that could not survive in a free market supply sugar.

Notice how the supply curve is pink, indicating it is domestic US sugar, at low and high levels of output. Imports snap the US supply curve, inserting a flat portion of length equal to amount of imports.

Cells I6 and J6 report equilibrium price and quantity (where S and D intersect). Compared to the *FreeTrade* sheet, P_e has risen from 10 to 25 cents per pound, while quantity has fallen from 13,800 to 12,000 thousand STRVs (2,000 of which are imported). Remember, we chose parameter values for the supply and demand curves to match real-world data from the US sugar market.

STEP Move the import slider control left and right to see how the import allotment affects the supply curve.

As you increase the amount of imports, you lengthen the flat segment and push the pink part of the S curve to the right. Decreasing the import allotment does the opposite. The beginning of the supply curve, below 4,000 units remains unchanged.

It is also easy to see how tightening the import allotment increases the equilibrium price and lowers the equilibrium quantity. Relaxing the imports allowed does the reverse.

STEP Enter 9800 in cell H6. This is the same as moving the import slider control all the way to the right.

This mimics the *FreeTrade* sheet. The import allotment is set so high that foreign sugar producers supply all of the US market after the first 4,000 units. Equilibrium price falls back to its free trade level of 10 cents per pound and quantity rises to 13,800 units.

Evaluating an Import Quota

We know import quotas raise prices and lower output, but this is just the outcome of the mechanism. To evaluate import quotas, we use the concepts of surplus and deadweight loss.

STEP Return the import quota to 2000 in cell H6 and then click the *Show CS* checkbox (cell C7).

Cells are displayed in columns A and B that are the source data for the blue consumers' surplus triangle that has been added to the chart. Under the sugar quota, the *CS* no longer extends to the world price of 10 cents per pound and quantity is smaller than the optimal quantity. Consumers lose $3.87 billion in surplus compared to the optimal solution.

STEP Click the *Show US PS* checkbox.

This adds the producers' surplus gained by sugar manufacturers in the United States. Their total *PS* is composed of two separate parts. On the left is a trapezoid and on the right is a triangle. What is in the middle?

STEP Click the *Show Foreign PS* checkbox.

The orange rectangle added to the chart is *PS* that goes to foreign producers. Notice that this is not deadweight loss because someone gets it.

Clearly, producers' surplus is much higher with the quota, rising from a mere $400 million with free trade (which equals the optimal solution) to $2.5 billion for US producers and $3.1 billion for all producers.

The transfer of *CS* to *PS* under the quota system, however, is not without waste. Consumers lost $3.87 billion of surplus and producers gained $2.7 billion. What happened to the rest?

STEP Click the *Show DWL* checkbox.

The red triangle has an area of $1.17 billion. This is the amount of *CS* that was lost during the transfer of surplus from consumers to firms. Figure 17.24 shows the chart with all checkboxes checked.

A leaky bucket is an apt metaphor. While siphoning off billions of dollars from consumers and delivering them to producers, $1.17 billion leaked and was wasted, captured by no one. We can express the leakage as a percentage, $\frac{1.17}{3.8} \approx 30\%$. That is a pretty big hole in the bucket.

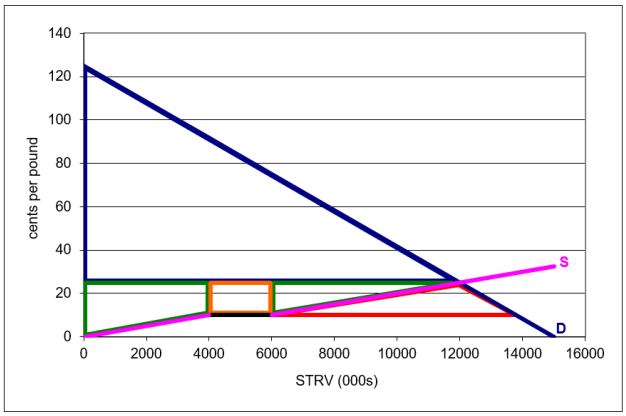


Figure 17.24: Partial equilibrium analysis of a sugar quota.
Source: SugarQuota.xls!ImportQuota.

Notice the geometry in this example. The *DWL* triangle in Figure 17.24 is not the usual bow tie shape (as in the price ceiling, tax, and monopoly applications). In this case, the *DWL* is a triangle under supply and demand. But the interpretation is the same—we are measuring surplus foregone and using this as an indicator of the damage done by the misallocation of resources.

You might wonder why consumers are not up in arms. In fact, commercial sugar buyers do lobby Congress and when prices spike, the TRQ allotments are relaxed. The vast majority of buyers in the supermarket, however, simply have no idea that this is happening. A five pound bag of refined sugar that costs $2 is just another item in the shopping cart.

This is a common problem surrounding import quotas: costs are diffused widely while the benefits are concentrated on a few key players. Thus, although the costs add up to a large number, $3.7 billion in this example, no one individual is impacted enough to object. The handful of US sugar producers, however, have strong incentives to maintain the system to keep their profits. You will see what this means when you answer the last exercise question.

The transfer of surplus, no matter how unfair it may seem, is not the real problem in the eyes of partial equilibrium analysis. The fact that surplus is vaporized and vanishes into thin air so no one gets it—this is the real problem.

It is easy to be confused by the shapes on the graph and concerns that prices are higher and producers are stealing surplus from consumers. None of that really matters. Here is the takeaway: the import quota is causing

a misallocation of resources. The United States is using land, labor, and capital to make sugar when it would be better off buying foreign sugar and using these inputs to make other goods and services.

Comparative Statics

We can explore the effects of changing demand and supply coefficients on the equilibrium price and quantity of sugar, but the natural question to ask is, what is the effect of the import allotment? We are chasing the *import elasticity of price* and the *import elasticity of quantity*. We want to know how responsive price and quantity are to shocking the import allotment.

We can also explore how the surplus and deadweight loss changes. These variables are also endogenous in this model because they are generated by the forces of supply and demand.

We have the initial position. With H6 = 2000, $P_e = 25$ and $Q_e = 12000$.

STEP Set H6 to 3000.

The length of the orange rectangle expands and the rising part of the US supply curve is pushed right. Equilibrium price falls to just over 23 cents per pound and output rises to about 12,231 thousand STRVs. *CS* and foreign *PS* rise. Deadweight loss falls. This is better for US consumers and foreign sugar producers than the initial quota of 2,000 units.

United States *PS*, however, falls. Domestic sugar producers are not happy with this. They prefer a lower import quota.

Elasticities give us more information than the qualitative statements (up or down) made above. We can compute the percentage change in price, quantity, surplus, and deadweight loss for the 50% increase in import (from 2,000 to 3,000 units).

The import elasticity of price $\approx \frac{\frac{23-25}{25}}{\frac{3000-2000}{2000}} = \frac{-0.8}{0.5} = -0.16$. This tells us that equilibrium price is quite unresponsive to the import allotment.

The import elasticity of quantity $\approx \frac{\frac{12231-12000}{12000}}{\frac{3000-2000}{2000}}$ $approx \frac{-0.02}{0.5} = -0.04$ is even smaller. Equilibrium quantity is extremely unresponsive to the import allotment.

These elasticity estimates are for illustration. Our model relies on rudimentary, linear demand and supply curves. The framework, however, is exactly how an economist would model the sugar market and interpret the effects of a sugar quota.

Do as I Say, not as I Do

Rich, developed countries talk a lot about free trade, especially to lesser developed countries, but it is clear that powerful special interests can and do dominate individual markets in the rich countries of the world. The tools of partial equilibrium analysis can be used to (approximately) evaluate the results of protectionist policies.

In the case of the US sugar TRQ program, data provided by the USDA can be used to estimate the size of the deadweight loss. With a total import level of 2,000 thousand STRVs, assuming price elasticities of demand and supply of -0.25 and $+1.0$, the deadweight loss is around one billion dollars. United States consumers bear the brunt of the costs of the TRQ system, while US and foreign producers enjoy much higher profits.

But remember caveat emptor. Partial equilibrium deadweight loss analysis is a rough, back-of-the-envelope calculation. Although progress has been made in estimating deadweight loss (see the references to this chapter), consumers' surplus using demand curves makes interpersonal utility comparisons, violating one of the principles of modern utility theory.

Even more importantly, by focusing on a single market, we ignore the ramifications of the sugar quota on other goods and services. We are not counting lost output of other goods by devoting resources to producing sugar in the United States. We are also not counting health effects of sugar.

Now that you know about the US sugar quota, you can take a break and watch comedian Stephen Colbert's brief segment from 2009: tiny.cc/TRQ. Recall from Figure 17.22 that sugar prices spiked to an all-time high back then.

Exercises

1. Use the *ImportQuota* sheet to figure out what happens if all imports are banned. Explain your procedure and take screenshots as needed. Would you support a ban of all imports? Explain.

2. The deadweight loss estimates in the text are sensitive to the demand and supply curve parameters. Suppose that the inverse supply curve had a slope of 1/100 instead of 1/400. Be sure to change this parameter in both the *FreeTrade* and *ImportQuota* sheets to 1/100. What effect would this have on the TRQ system? Explain your procedure and take screenshots as needed.

3. Search the web for information about how much money US sugar producers contribute to the political campaigns of members of the US Congress. Copy and paste one sentence from a web site that you think shows the influence US sugar producers have on the US Congress. Please document your sentence with a URL and date visited citation.

References

The epigraph is from the abstract of Jerry A. Hausman and Whitney K. Newey, "Nonparametric Estimation of Exact Consumers Surplus and Deadweight Loss," *Econometrica*, Vol. 63, No. 6 (November, 1995), pp. 1445–1476. www.jstor.org/stable/2171777. This paper has a nice explanation of developments in estimating deadweight loss and an example application to gasoline demand.

Economists know that using ordinary demand curves to measure *CS* and deadweight loss (the Marshallian approach) is a mistake, but some argue the error is small enough to ignore. This paper says the error matters. Pascal Lavergne, Vincent Requillart, and Michel Simioni, "Welfare Losses Due to Market Power: Hicksian Versus Marshallian Measurement," *American Journal of Agricultural Economics*, Vol. 83, No. 1 (February, 2001), pp. 157–165, www.jstor.org/stable/1244307.

When the beekeeper's bees fly into the adjoining apple orchard and pollinate the apple-grower's apple blossoms, they are conferring a positive benefit on the apple-grower that the beekeeper cannot take advantage of directly (i.e., a positive externality).

Eric S. Maskin

17.6 Externality

This section is devoted to explaining the concept of externality, why it causes a market failure, and how the inefficiency in the allocation of resources can be corrected.

The core idea is that externalities cause markets to fail—too much or too little is produced. Society's resources are inefficiently allocated. The reason why markets fail in the presence of externalities is that decision makers (consumers or firms) fail to incorporate the full costs or benefits of an action so they make a poor decision (from society's point of view).

There are three questions to answer:

1. What is an externality?

2. Why do externalities break the market?

3. How can we fix the market?

1. What is an Externality?

An *externality* is a cost or benefit not taken into account by the decision maker. An agent takes an action that impacts others, but she does not incorporate this "external impact" (hence the name externality) into her optimization problem. The decision maker considers only personal or private cost and benefit, not the full or social cost and benefit.

Externalities can arise on the cost or benefit side of an optimization problem. The private costs or benefits are the ones included in the agent's calculations. The external costs or benefits are ignored. The full or total costs or benefits are called social costs or benefits.

We can better understand externalities by looking at examples. The key is always that the optimizing agent is not considering all of the costs and benefits. Costs are imposed, but not felt by the agent or benefits are conferred on others, but not captured by the agent. This leads to a privately optimal solution that diverges from the socially optimal solution and produces a misallocation of resources.

A classic example of an externality is industrial pollution. When the cost of pollution is not taken into account by the firm, this is called a *negative production externality*. A steel firm deciding how much steel to produce factors into its choice of output level the revenue from making steel and a whole series of costs: labor, raw materials, and equipment. The costs that are counted are private costs.

If the firm pollutes the air through a smokestack, but does not have to pay for polluting the air, this is an external cost. Social costs include private costs and external costs. It is a negative externality because costs are imposed on others that are not taken into account by the decision maker. It is a production externality because the decision is made by a firm deciding how much to produce.

A college education is another classic example of a situation where the decision maker fails to consider the total picture. It is often used to explain a *positive consumption externality* because there are benefits to education that are not taken into account by the student.

The choice variable is how many years of schooling to acquire beyond high school. The costs are huge—out-of-pocket costs of a 4-year college degree include tuition and books, but opportunity costs are even greater. The benefits are also quite large, including access to better jobs, higher pay, and greater quality of life. These private benefits are considered when high school students decide whether or not to go to college so they are not part of the externality.

But society benefits from education also. College-educated people have lower unemployment rates, smoke less, and are more likely to vote. These social benefits are ignored by individuals making a decision about whether or not to acquire a college education. It is a positive externality because benefits flow to others that are not taken into account by the decision maker. It is a consumption externality because the decision is made by a consumer deciding how much to purchase.

Many studies attempt to estimate the gap between the social rate of return and private rate of return to a college degree. Social rates of return to education are several percentage points higher than the private return. This gap is an estimate of the external value generated by education.

Externalities are everywhere. Some are easy to spot, like the loud music your next door neighbor plays (a negative consumption externality). To the extent that you ignore the impact on others, your decision about which shirt to wear contains an externality.

But externalities can be subtle also. Consider an army with soldiers that were drafted into service. The externality is that the government does not take into account the full cost of acquiring its soldiers. This externality disappears with a volunteer army because the military has to pay enough to entice people to join.

Externalities are all about impacts on others so it is easy to see why they are also known as *spillover effects*. Remember, the private costs and benefits are counted by the decision maker, but the external effects are not.

2. Why Do Externalities Break the Market?

Recall Figure 17.6, reproduced below as Figure 17.25 for your convenience.

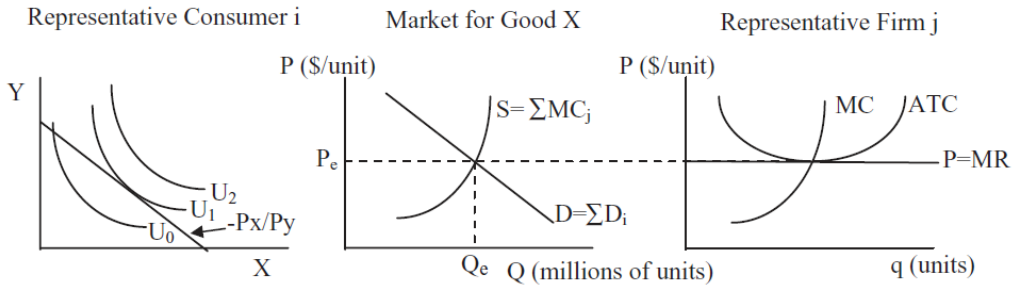


Figure 17.25: An overall view of supply and demand.

This figure has three canonical graphs: the Theory of Consumer Behavior on the left, the Theory of the Firm on the right, and supply and demand in the middle. It says that the equilibrium solution is found at the intersection of supply and demand, which come from the firm and consumer graphs.

We can show that the equilibrium quantity equals the quantity that would maximize consumers' and producers' surplus. Price controls (such as ceilings or floors), taxes, and monopoly all generate market failures, defined as quantities that do not maximize *CS* and *PS*.

We can add externalities to this list. Negative externalities are costs not taken into account and they produce too much output, while positive externalities do the reverse.

Look carefully at Figure 17.25. For the market system to yield a socially desirable outcome, supply and demand must reflect the full costs and benefits of the product. But this is precisely what is not happening if an externality is present. There are positive or negative spillover effects that result in a market equilibrium that is sub-optimal.

Suppose we have a situation where producers do not take into account the costs of pollution created as a by-product of manufacturing. Then the *MC* curve in Figure 17.25 is not incorporating the full costs of production and the supply (which is the sum of individual *MC* curves) is also too low.

There is a *marginal social cost*, *MSC*, curve that does include all costs and it does yield the socially optimal solution. Figure 17.26 shows the canonical graph of a negative externality in production. It is easy to see that the *marginal private cost*, *MPC*, which firms use to decide how much to produce to maximize profits, is too low. This produces an equilibrium output that is too high.

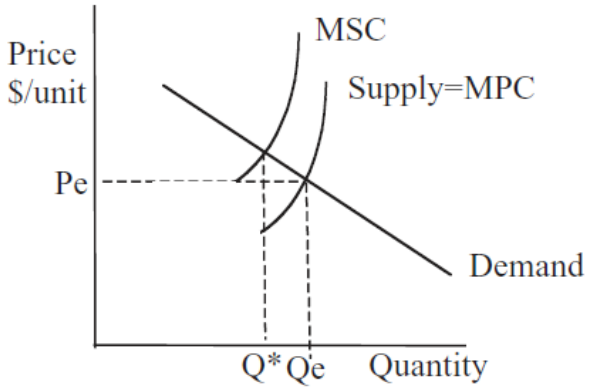


Figure 17.26: A broken market with a negative production externality.

Q^* in Figure 17.26 shows the optimal output for society. The socially optimal level of output is based on the full, social cost of production. Q_e shows the (broken) market's output. The market's equilibrium output is based only on the private cost of production so it is too high.

To sum up, a negative production externality means that firms fail to include all costs and, therefore, $MPC < MSC$, and, therefore, $Q_e < Q^*$. This is why externalities cause market failure.

We can use Excel to create a simple spreadsheet that demonstrates the concepts of externality and market failure.

STEP Open the *Externality.xls* workbook, read the *Intro* sheet, then proceed to the *Externalities* sheet.

Let's take a quick tour of the screen. On the left are the total and marginal graphs for a single firm. We ignore the average cost curves (ATC and AVC) because we are not interested in this firm's profit position. All we care about is how much it will produce. The cost function is a simple quadratic and the market price is \$40/unit so the revenue function is $40q$.

On the right is the conventional supply and demand graph. Notice that the y axes of the individual and market graphs are the same. The x axes, however, are different. There are 1,000 firms and, combined, they produce tens of thousands of units of output.

Initially, this firm is producing 10 units of output. What would you advise this firm to do? Why?

STEP Use the firm's scroll bar control to adjust its output level.

To maximize profits, this firm will choose output where $MR = MC$. This output level will generate the maximum difference between the total revenue and total cost curves in the top graph.

The problem is easily solved via analytical methods.

$$\max_q \pi = 40q - (200 + q^2)$$

$$\frac{d\pi}{dq} = 40 - 2q = 0 \rightarrow q^* = 20$$

Both analytically and with Excel, we can see that the firm will produce 20 units when equilibrium price in the market is \$40/unit. When all 1,000 firms do this we get a market equilibrium output of 20,000 units. This is the socially optimal allocation of resources to this product.

STEP To implement the externality, slide the *Set Externality* control all the way to the right (so the red lines and curve are above the black ones in the three graphs).

The red objects are not labeled. What do they represent?

STEP Insert text boxes to label the red curve in the top graph, the red line in the bottom graph, and the red line above the supply curve.

The correct labels must include the word *social*. The red line in the top graph is the *total social cost*, *TSC*, and its marginal counterpart is the *marginal social cost*, *MSC*. The divergence between the red social cost and the black private cost signals the presence of an externality. The distance between the curves are costs not taken into account by the firm.

In the market graph, the red line is *MSC*, by which we mean the sum of the indivdual marginal social costs. Like in the individual graph, divergence between supply and MSC is a clear marker of the presence of an externality.

Note that neither the firm's profit-maximizing output level nor the market's equilibrium solution changes in the presence of the externality. We have imposed an added cost, yet the firms and market do not respond because the cost is ignored.

The dashed line from the intersection of *MSC* and demand is the socially optimal level of output. An omniscient, omnipotent social planner, OOSP, would incorporate the full costs of production in determining the optimal solution to society's resource allocation problem. OOSP would choose output at the intersection of *D* and *MSC*.

We could measure the inefficiency caused by the externality by the deadweight loss. This would be the area of the triangle shown in Figure 17.27.

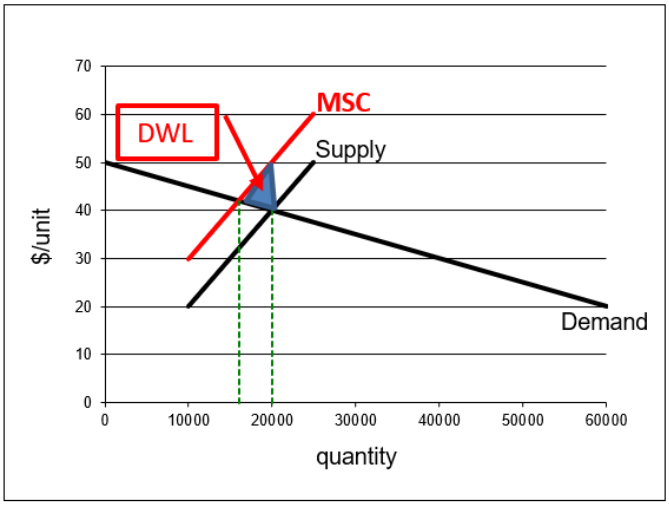


Figure 17.27: Deadweight loss from a negative externality.
Source: Externality.xls!Externalities.

The market in the presence of a negative externality has produced too much output. Units beyond 16,000 have greater marginal social cost than marginal benefit (as given by the demand curve) and should not be produced. The market produces an extra 4,000 units because it ignores the external costs of production.

3. How Can We Fix the Market?

Externalities break the market because costs and benefits are not fully incorporated into the agent's optimization problem. There are two possible solutions: government regulation and more property rights.

There are several regulatory approaches the government can take to fix the market failure caused by externality. They are united by the use of authority to correct the equilibrium output level so that it equals the socially optimal output.

Perhaps the most obvious regulatory fix is a strict limit on production, for example, a quota on pollution. If firms are allowed to pollute only a certain amount, they cannot produce as much as they want.

This is known as *command and control*, a term borrowed from the military, where top down decision making is the norm.

But this approach suffers from a serious drawback. It requires massive amounts of information to set the total amount of pollution and output.

Furthermore, if everyone is forced to reduce pollution by, say 20%, this does not take advantage of the fact that some firms can reduce pollution more cheaply than others. In other words, the government not only has to determine the total amount of pollution and output, it has to tell each individual firm exactly what and how to produce.

Command and control has long been used in environmental regulation. In the case of pollution, the Environmental Protection Agency (EPA) still uses effluent restrictions, but the EPA has moved toward other regulatory strategies.

Another government focused approach to fixing a market failure brought on by externality allows firms to decide how much to produce, but uses taxes and subsidies to incentivize decision makers to choose the socially optimal outcome.

This is based on the work of Arthur C. Pigou (rhymes with zoo, 1877 - 1959). He was a student of Marshall's and in 1908 he was appointed to Marshall's chair in economics at the University of Cambridge. Pigou argued that whenever private and social costs or benefits diverged, the government could offer incentives to align individual optimal solutions with socially optimal levels of output. Thus, today we call this solution a *Pigovian tax or subsidy.*

By imposing a Pigovian tax on polluting firms, producers are forced to consider the full costs of production in a roundabout way—the tax takes the place of the external cost.

The Pigovian tax shifts the supply curve up so that, if properly calibrated, the amount of the tax reflects the external cost not taken into account. Figure 21.28 shows how a Pigovian tax fixes the market failure caused by externality. Notice that the *Supply + Tax* curve equals the *MSC*. This enables the market equilibrium solution to equal the socially optimal solution.

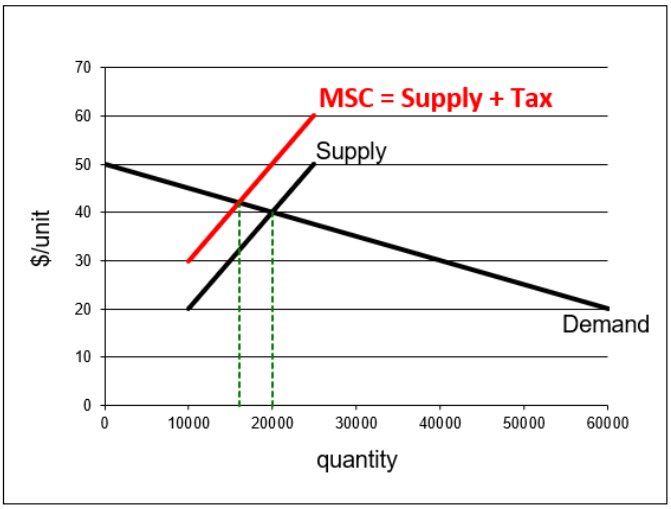


Figure 17.28: Pigovian tax correcting the inefficiency from an externality. *Source: Externality.xls!Externalities.*

The Excel workbook *Externality.xls* enables you to correct the externality with a Pigovian tax.

STEP With an externality in place, click the Set Pigovian Tax/Subsidy scroll bar to fix the inefficiency.

With every click, the market supply curve shifts up because you are imposing additional tax. A Pigovian tax works like a regular tax—it shifts the supply curve up. Obviously, you want to set the tax so that the black supply curve is coincident with and covers the red *MSC* curve.

The Pigovian tax fixes the inefficiency caused by the negative externality when the amount of the tax takes the place of the divergence between marginal social and private cost. You know you have the right tax when the market's equilibrium output equals the socially optimal level of output at 16,000 units.

Unlike regular taxes, which are applied to generate revenue for the government and cause the equilibrium quantity to be less than the optimal quantity, Pigovian taxes are actually applied to correct a market failure. They do generate revenue, but the primary purpose of a Pigovian tax is to change the market's equilibrium output to allocate resources optimally.

Pigou's approach dominated economics for many years. Then, in 1960, Ronald Coase (1910 - 2013), who spent most of his long career at the University of Chicago, offered an ingenious alternative: Define property rights over all resources (such as clean air) to internalize the externality. It took some time, but Coase's approach caught on and would win him the 1991 Nobel award in economics.

In essence, Coase cures the "market failure" by creating more markets. Market failure is in quotation marks because the argument is that it is *not* a market failure since we do not have complete property rights over all resources. A little intellectual history will help clear this up.

Frank Knight (at the University of Chicago) disagreed with Pigou in an article way back in 1924. Pigou used too much traffic as an example of a market failure in his influential book, *The Economics of Welfare*, in 1920. On page 194, Pigou explained that individual drivers would fail to take into account the additional congestion they caused when deciding whether to take one road versus another. Thus, the drivers would distribute themselves inefficiently. He pointed out that the government could impose a toll, a tax to use the road, to fix this market failure (Pigou used the phrase *laissez-faire* and it would not be until the 1950s that "market failure" was coined).

In his 1924 paper, Knight replied that, far from this being a market failure, the problem created by the externality was that there was a missing market! He said Pigou's logic was error free. It is true that drivers following their own self-interest would produce too much congestion. It is true that this decentralized system failed and a corrective tax would fix it. But, said Knight, while decentralized, this is not a market system because nobody owns the roads. Not all decentralized systems are automatically market systems.

Knight maintained that you cannot blame the market system for a lack of property rights. In Knight's view, a properly functioning market system would force firms to pay for all of the resources used. A negative externality meant that firms would treat some resources as free and it is no surprise that they would overuse those resources.

Pigou removed the traffic congestion externality example from the next edition of his book. He left, however, the overall framework of corrective taxes and subsidies intact and it became part of the paradigm of economics. For decades, students learned that corrective taxes and subsidies could and should be used to fix inefficient levels of equilibrium output.

In 1960, Coase wrote his most famous article (and perhaps the most often cited article in the history of economics), "The Problem of Social Cost." He explained how more property rights would enable markets to cure externalities. For a negative spillover like pollution, instead of command and control or a government tax, Coase advocated establishing property rights to clean air and letting the market work its magic. Firms would no longer treat the air as free if they had to pay to use it.

There is no Excel implementation of Coase's solution. The idea is simply that unpriced resources be priced. This happens when unowned resources are assigned owners. This creates a market, buyers and sellers, for the resource. This directly internalizes the externality.

Coase has said that the property rights solution was influenced by Knight. They were colleagues at the University of Chicago for many years. Knight is known as the father of Chicago School economics and an impact on the work of many social scientists at Chicago and around the world.

A theorem bears Coase's name and a brief explanation of its content is in order. The *Coase Theorem* arises out of the idea that more finely delineated property rights enable the market to solve the problem of externality. The word *theorem* is loosely used here and Coase never claimed to have found or in any sense proved the Coase Theorem.

Coase showed that by settling property rights disputes, courts played a key role in enabling markets to work. Before the court ruled, trade would be impossible because there was disagreement over ownership. These high transactions costs would prevent negotiation.

Once the court ruled, there would be a clear potential buyer and seller. Coase argued that it was not important who won the case because the resource would end up with whoever valued it more. By giving one party the property right, the court established ownership and enabled the resource to be traded. If the winner valued the resource more, the loser would be unwilling to buy it. If the winner valued it less, the loser would buy the resource. Either way, said Coase, once the judge ruled, the resource would end up at its most highly valued use. This idea is now known as the Coase Theorem.

So, in the case of pollution, perhaps homeowners would sue the polluting firm. The court would rule and, either way, once the property right was es-

tablished, the market would begin to function. Assuming the polluting firm values the property more, it will buy the right to pollute if it loses and will not sell the right if it wins the court case. Either way, it incorporates the cost of pollution because it has to purchase the right if it loses or it recognizes the opportunity cost of having the asset if it wins.

Coase criticized Pigovian taxes and subsidies as a way to fix inefficiency in the allocation of resources by a market system. Coase saw Pigou's approach as hopelessly idealistic and impossible to implement in the real world. It is easy to draw Figure 17.27 and a snap to show that the correct tax or subsidy enables the market to hit the socially optimal output as in Figure 17.28.

Unfortunately, this blackboard economics (as Coase derisively called it) is easy to draw and teach, but almost impossible to implement. The government regulator will know neither the demand nor the supply functions, and changes over time imply constant tweaking of optimal taxes or subsidies.

Economists think of Coase and Pigou as locking horns and often cast the issue as free market versus regulation. It is clear, however, that Coase and Pigou share some common ground. They both seek to maximize the value of output; they want to optimally allocate resources.

Both offer solutions that work well in theory, but can prove difficult to implement. Once we recognize that neither approach is perfect, we can begin the difficult task of deciding which approach is better in a particular situation.

The EPA and Acid Rain

Although Pigovian taxes remain a staple of economics, in recent years, market-based strategies relying on Coase's logic have gained popularity.

For example, *cap and trade* works by creating a total amount of allowable pollution and creating a market where firms can buy and sell rights to pollute. This forces firms to take into account the full costs of their production decisions. They must buy a permit in order to pollute and this forces them to internalize the externality.

The EPA's sulfur dioxide (SO_2) cap and trade program is aimed at decreasing pollutants that cause acid rain, www.epa.gov/airmarkets/allowance-markets. Instead of command and control or taxes, the EPA sets a total emissions con-

straint, or bubble, then allows firms in the bubble to buy and sell pollution permits. This scheme is equivalent to setting up a market for pollution.

There are many details to be worked out when setting up a market. For example, the government can give each firm an initial allocation of permits or they can auction off the permits.

Some environmentalists remain strongly opposed to market-based solutions to pollution abatement. They see such programs as "licenses to pollute." But the market's ability to price resources correctly and enable socially optimal resource allocation is a powerful factor in favor of the market.

Other countries (including such different places as Europe, Costa Rica, and China) have started emissions trading programs. The idea of creating a market for pollution to correct the market failure caused by externality is most definitely a real, practical solution that continues to grow in popularity.

Externalities, Market Failure, and Corrective Action

Externalities are costs or benefits not taken into account by the decision maker. Externalities cause inefficiency because the equilibrium level of output does not equal the socially optimal level of output. As usual, we can measure the inefficiency in the allocation of resources caused by an externality by computing the deadweight loss.

The inefficiency caused by externality can be corrected by command and control, but this approach requires micromanagement by government regulators. Pigovian taxes and subsidies are a type of government regulation that allows individual agents to decide what to do. A firm, for example, would decide how much to pollute and produce, but they have to pay tax. The Pigovian tax is optimized to push the market to the socially optimal output.

The Pigovian approach is definitely at play in the area of education. Truancy laws and other absolute requirements concerning schooling are an example of command and control. Government support of higher education through student grant and loan programs are Pigovian subsidies. The idea is to help students capture the full benefit of a college education and ensure that private decision making is socially optimal.

Another repair relies on market-based solutions to the inefficiency created by externality. Instead of taxing or subsidizing buyers or sellers, property rights

for unowned and unpriced resources are established and then the market is left to work its magic. Cap and trade is an example of this approach.

Coase is credited with the idea of fixing inefficient market outcomes with property rights, but Knight definitely had an influence. Knight's criticism of Pigou's toll road example is long forgotten, but it contained the seed of the logical argument that a Pigovian market failure is no such thing because not all decentralized systems are market systems.

Mechanism design is a new subfield in economics where we consciously design a game and then let agents play to reach a desired result. This is totally different than the evolution of the market system. Adam Smith did not draw up a blueprint for a market-based society. It happened organically. But now that we know how it works, we are trying to design institutions that give desirable results.

Exercises

1. Give an example of a positive externality in consumption.
2. Analyze the welfare effects of a positive externality in consumption. Use Word's Drawing Tools to support your answer with a demand and supply graph.
3. In each case that follows, describe the regulatory strategy to correct the market failure caused by a positive externality in consumption.
   (a) Command and control
   (b) Pigou
   (c) Coase

References

The epigraph is from the first page of Eric S. Maskin, "The Invisible Hand and Externalities," *The American Economic Review*, Vol. 84, No. 2, Papers and Proceedings of the Hundred and Sixth Annual Meeting of the American Economic Association (May, 1994), pp. 333–337, www.jstor.org/stable/2117854.

"The Problem of Social Cost" was published by Ronald Coase in the fledgling journal of a new field, *Journal of Law and Economics*, Vol. 3 (October, 1960), pp. 1–44, www.jstor.org/stable/724810.

The transcript of a conversation about Coase's article and the Chicago School in general is in Edmund W. Kitch, "The Fire of Truth: A Remembrance of Law and Economics at Chicago, 1932–1970," *Journal of Law and Economics*, Vol. 26, No. 1 (April, 1983), pp. 163–234, www.jstor.org/stable/725189. George Stigler describes the presentation by Coase to a "collection of superb theorists" as "one of the most exciting intellectual experiences of my life:"

> My recollection is that Ronald didn't persuade us. But he refused to yield to all our erroneous arguments. Milton would hit him from one side, then from another, then from another. Then to our horror, Milton missed him and hit us. At the end of that evening the vote had changed. There were twenty-one votes for Ronald and no votes for Pigou. (p. 221)

That was no typo on Coase's lifespan in the text, 1910 - 2013. And he worked to the end. In 2009, at the age of 99, he published a book, with Ning Wang, How China Became Capitalist.

Arthur Cecil Pigou's *The Economics of Welfare* was first published in 1920 and is freely available online at the Library of Economics and Liberty, www.econlib.org/library/NPDBooks/Pigou/pgEW.html.

Frank H. Knight's criticism of Pigou's traffic congestion example is in "Some Fallacies in the Interpretation of Social Cost," *The Quarterly Journal of Economics*, Vol. 38, No. 4 (August, 1924), pp. 582–606, www.jstor.org/stable/1884592.

> Lack of legal sanctions means that loyal members of the cartel must exact penalties against deviants in the market place. Unless such disciplinary actions (mainly price cuts) can be localized, every member of the cartel, loyalist and defector alike, suffers. That is a very severe (if little remarked) limitation on the efficiency of cartels.
>
> Oliver E. Williamson

17.7 Cartels and Deadweight Loss

We know that the equilibrium output of a competitive market equals the output that maximizes consumers' and producers' surplus. We also know that monopoly produces too little output and the resulting deadweight loss is a measure of the inefficiency of monopoly. But competition and monopoly mark opposite ends of a spectrum that includes a wide range of other market structures.

A cartel is a type of market structure in which a group of firms cooperate to control output and price. Perhaps the most famous international cartel is the *Organization of the Petroleum Exporting Countries*, OPEC. Cartels are not monopolies because there are several independent firms in the syndicate or trust, but they hope to act like a monopolist, restricting output and raising price, to earn monopoly profits. Cartels are inherently unstable because it is in the interest of each member to cheat and sell more than the agreed amount.

This section explores the welfare properties of a specific type of cartel. The application is based on the workings of the Norwegian cement cartel as explained by Röller and Steen (2006). Analyzing the cartel involves solving a two-stage game and the cartel result is compared to monopoly and non-cooperative, Cournot competition. This material is advanced and it is recommended that the chapter on Game Theory be completed before proceeding.

A Brief History of Norwegian Cement

Cement output in Norway (and in other countries that use the metric system) is measured in tonnes (pronounced tons). This is not simply a foreign spelling for a ton. A ton is 2,000 pounds. A tonne, sometimes called a metric ton, is 1,000 kilograms. Given there are roughly 2.2 kilos in a pound, a tonne is about 2200 pounds. Thus, a tonne is bigger than a ton.

Figure 17.29 shows that production rose dramatically during the second half of the 1960s, greatly outpacing demand. This excess output was sold at a loss in other countries. A balance between production and consumption was restored by the early 1980s.

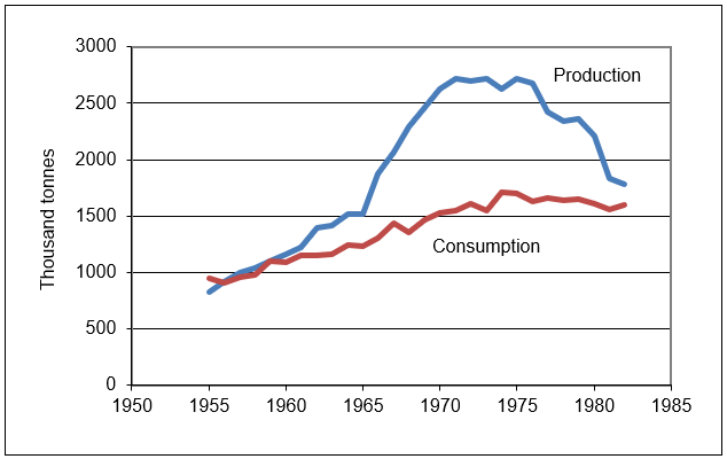


Figure 17.29: Norwegian cement production and consumption, 1955-1982. *Source: CartelDWL.xls!Data.*

Production rocketed because of the sharing rule adopted by the Norwegian cement industry. A *sharing rule* determines how the monopoly rent is to be distributed among the firms in the cartel. Each firm's share of the domestic market was based on its fraction of total industry capacity. We will see that this gives each firm an incentive to expand plant capacity and led to the explosion in output shown in Figure 17.29.

In 1968, the three producers in the cement industry abandoned the cartel market structure and merged to form a monopoly. By then, however, plant capacity had been expanded and it took years to reduce output.

Röller and Steen explain that there are few empirical studies of cartels because they are illegal in many places (including the United States) so obtaining data is difficult. Such is not the case in Norway. "Given the legality of the Norwegian cement cartel, we have a large amount of primary data allowing us to do a complete welfare analysis." (Röoller and Steen, 2006, p. 321.

Monopoly Review

STEP Open the Excel workbook *CartelDWL.xls*, read the *Intro* sheet, then go to the *Monopoly* sheet.

Given the linear inverse demand curve and constant marginal cost, finding the monopolist's profit-maximizing solution is easy.

STEP Use the scroll bar under the chart to find Q^*. As you change the quantity, you can see the corresponding price in the chart and in cell B11. You can also see the producers' surplus (also known as profits) change in cell B19 as you set Q.

You can choose Q^* by watching cell B19, but you could also find Q^* by choosing the intersection of MR and MC.

Excel's Solver offers yet another alternative to finding the profit-maximizing level of output.

STEP Run Solver and configure the Solver dialog box to solve the monopolist's profit maximization problem.

Finally, click on cells B18, B19, and B21 to show the consumers' surplus (CS), producers' surplus (PS), and deadweight loss (DWL) from the monopoly solution in the chart.

Having found the monopoly solution, we turn to output (and price) under a noncooperative, Cournot environment.

Cournot Review

STEP Proceed to the *CournotFirm* sheet.

Chapter 16 on game theory presented the material reviewed here, which assumes a basic understanding of the Cournot model and Nash equilibrium.

Instead of a single firm, there are three firms making a homogeneous product. They do not collude or combine forces. Instead, they compete. Unlike

perfect competition, however, there are so few producers that they impact each other's decision making. If one firm decides to produce a lot, this will lower the price for all three firms.

How will an individual firm decide how much output to make? The core idea is that each firm will make profit-maximizing output decisions based on conjectures about what the other firms will do. The output level at which each firm's decision is consistent with the output chosen by the other firms is the solution, called a Nash equilibrium.

The *CournotFirm* sheet opens with cell B10 set equal to zero. This means that Firm 1 is exploring what its best option is if the other firms produce nothing.

STEP Use the scroll bar under the chart to find the profit-maximizing output for the conjecture that the other firms produce nothing.

If the other firms decide to produce zero output, Firm 1 will produce 2.3 million units of output. But this is not an equilibrium solution because the other firms would not choose to produce zero units of output when this firm produced 2.3 million tonnes. How much would the other firms produce?

STEP Click the Set Exo Q button to copy Firm 1's optimal solution (in cell B15) to the conjectured output in cell B10.

Notice how the chart shows new red D and MR curves. These are the residual demand and residual marginal revenues curves for Firm 2, given that Firm 1 produces 2.3 million and Firm 3 produces nothing.

STEP Use Excel's Solver to find the profit-maximizing output for the conjecture that the other firms produce 2.3 million units.

You should find that Firm 2 will produce 1,150,000 units when the other two firms produce 2.3 million. We have stumbled upon the Nash equilibrium solution! If each firm produces 1.15 million units, then none of them will regret its output decision. In other words, each firm's optimizing decision (1.15 million) is consistent with the conjectured output (2.3 million).

Notice that the Nash equilibrium is not Firm 1 = 2.3 million, Firm 2 = 1.15 million, and Firm 3 = 0. Both Firms 1 and 2 would regret their decisions

and would opt for different output choices. It should be clear, however, that if each one makes 1.15 million, then none of the firms would regret or wish to change its chosen output level.

The Cournot solution can be found via iteration (which was easy in this example) or by analytical methods (see work starting in cell A28). The reduced form for the industry's Nash equilibrium output in this Cournot model (linear demand and cost function and n firms) is:

$$Q_e = \frac{n}{n+1}\frac{d_0 - MC}{d_1}$$

Price, of course, is simply read from the inverse demand curve.

STEP Proceed to the *Cournot* sheet to see the welfare implications of the Cournot solution. Click on cell B14 to see that the formula for the Nash equilibrium has been entered.

Notice that the Cournot output level is between the perfectly competitive ($D = MC$) and monopoly ($MR = MC$) output levels.

STEP Click on cells B18, B19, and B21 to highlight *CS*, *PS*, and *DWL* in the chart.

Once again, notice that the *DWL* for the Cournot solution is between the monopoly (highest *DWL*) and perfect competition with many firms (no *DWL*) extremes.

STEP Increase the number of firms in cell B10 to 5, 10, and 20.

As n rises, *DWL* falls because as n rises, we are approaching the ideal solution of competition with many firms. Thus, perfect competition is simply an n-firm Cournot model with an infinite number of firms. You can confirm that at $n = 1$, the monopoly solution is found.

Having covered the monopoly and competitive Cournot models, you are ready to tackle yet another market structure: the cartel.

Cartel Behavior

Suppose an industry, made up of several firms, organized into a *cartel*. In other words, the firms would join forces and cooperate in making decisions.

The cartel would decide the total domestic output and price for the product. In addition, the cartel would have to determine how much each firm would produce. This is called the sharing rule.

Different sharing rules yield different results. Suppose that the sharing rule applied is that each firm's output reflects its share of total industry capacity. There are no limits on each firm's capacity and any output not sold domestically could be exported at the world price.

Although each firm chooses capacity first and then the cartel chooses total output (and price), we solve the two-step optimization problem recursively. This means we start at the second stage, then work backwards to the first stage.

Stage 2: Choosing Total Domestic Output (and Price)

STEP Proceed to the *CartelStage2* sheet.

The information is laid out as in the *Monopoly* sheet, but there are additional variables. The world price (below marginal cost) has been added in cell F8 and to the chart. Individual firm parameters start in row 26. The three firms have chosen their capacities (cells B30:B32), determining total capacity (B28) and shares of domestic output (C30:C32).

STEP Use the scroll bar under the chart to explore different quantities of domestic output. This is the cartel's key choice variable. It can choose anywhere from no output to the vertical, total capacity, line (which is determined by the firm's capacity decisions in stage 1 and is now an exogenous variable to the cartel).

STEP Click on cell B19, which is the *PS* and also the profit generated by a given output level, to highlight the *PS* in the chart. The formula and the chart reveal that *PS* has two parts: = (P − s0_)*Q − (s0_ − R_)*(B28 − Q).

The first part is a rectangle with height from *MC* to price and width from zero to the chosen output. This would be *PS* under monopoly.

But the cartel has a second component to *PS*. This is the smaller rectangle on the chart and it is subtracted from the bigger rectangle. This second part is the excess output that is exported and sold at the world price. It

is subtracted from profits because the world price is below MC. Thus, these units are sold at a loss.

STEP Use the scroll bar to find the cartel's Q^*. Notice that you can find the optimal output by keeping an eye on PS (in cell B19) or by setting $MR = R$. You can also use Excel's Solver to find the optimal output.

Cell B13 shows the optimal output and your cell B12 should equal this solution. The cartel will produce 3,150,000 units and charge $1,725 per unit. This is a higher output (and lower price) than the monopoly solution.

R is a key variable. It plays the role of MC in the cartel's optimization problem. What effect does changing R have on Q^* and P^*? What welfare effect does changing R have? We can answer these questions with Excel.

STEP Change R to 500 in cell F8. Solve the cartel's optimization problem again.

You should see that optimal domestic quantity is lower and price is higher.

STEP With the new optimal solution for $R = 500$ in B12 ($Q^* = 2.8$ million), click the Compare Surplus button. It displays the initial CS, PS, and DWL values (for $R = 150$) and computes the difference between the new and initial values.

As R rises, CS falls and PS rises. Total DWL is bigger by $136 million, with both parts of DWL (the traditional triangle that represents domestic DWL and the export loss) rising.

STEP Click the Reset button (or reset R to 150).

We conclude our analysis of the cartel's first stage of the optimization problem by examining the effect on the individual firms. Cells D30:G32 show how the sharing rule is applied to determine how much each firm produces, given the cartel's total domestic output decision. The blue text color means these variables are endogenous—they are determined by the cartel's domestic output decision.

STEP Adjust Q via the scroll bar under the chart and keep your eye on cells D30:G32. As Q changes, so do the individual firm variables in blue.

Because the firms have equal capacities, each sells a third of the domestic output and exports the rest. Domestic and export sales for each firm are displayed.

STEP Enter 3,150,000 in cell B12 (the value of Q^* at the initial values of the exogenous variables) to see the PS earned by each firm at the cartel's optimal output.

From the cartel's point of view, the individual firm capacities are given. But would profit-maximizing firms choose these particular capacities? This question is at the heart of the first stage of the cartel's two-stage optimization problem.

Stage 1: Choosing Capacity

Now that we know how the cartel is going to decide how much domestic output to produce and the sharing rule, we can tackle the question facing each firm: How much capacity?

At any point in time, firms have a given maximum total production, or capacity, determined by factory size. To increase capacity, firms must expand factory size and this takes time.

Notice that the marginal cost of cement production is different from the marginal cost of capacity. The former is assumed to be low and it does not play a role in this analysis. In fact, it is assumed that firms always produce up to capacity.

The capacities of each firm and hence total capacity are given to the cartel but are chosen by each firm. Each firm would pick that capacity that would maximize its profits.

The profit function has revenue from two sources: domestically sold output at price P (chosen by the cartel) and the excess output that is exported and sold at the world price, R. The cost of capacity function is linear, with constant marginal cost.

STEP Proceed to the *CartelStage1* sheet and click on cell B11 to see that the formula reflects the firm's profit function.

Cells B19:B23 have the exogenous variables. Each firm chooses capacity (q_i) to maximize profits.

The sheet opens with the firm having a capacity level of 1,200,000 units, the same as the other two firms, so the total industry capacity is 3,600,000 units.

STEP Click on the scroll bar (next to B27) to increase capacity.

Notice that the larger the chosen capacity, the greater the share of the domestic sales (Q), which is chosen by the cartel, and thus domestic revenues (B13) rise. As capacity increases, exports also rise (because only a share of the firm's output is sold domestically) and this hurts because the world price is below marginal cost. Of course, increasing output is going to increase costs because the firm has to build a bigger plant.

Given these trade-offs, what level of capacity should this firm select?

STEP Keep your eye on cells F27:H27 as you adjust the scroll bar to select the profit-maximizing output.

As usual, you can equate *MR* to *MC* to find the optimal solution.

STEP Check your work by using Excel's Solver.

The optimal capacity, 1,342,758 units, differs from the original 1.2 million units. This means that the optimizing firm would choose to make 1,342,758 units when the other two firms make a total of 2.4 million.

STEP Copy the optimal capacity in cell B27 and paste it in cell K9 (or enter 1,342,758 units in cell K9).

We are not done yet because if this firm wants to make 1,342,758 units, it stands to reason that the other firms (with identical cost structures) will also want to do this.

STEP Return to the *CartelStage2* sheet, select cell B30, and paste (or type in) 1,342,758.

Notice that cells B31 and B32 change to the value of cell B30. Cell B28, *Total Capacity*, is now higher and, thus, the vertical line in the chart has shifted right.

We do not need to run Solver again because the cartel's optimal output and price combination in the domestic market is unaffected by the total industry capacity. The extra output is simply exported and sold at the world price.

STEP Return to the *CartelStage1* sheet and notice that *MR* no longer equals *MC*. Click on cell B20 to see that it has a formula.

Cell B20, *Other Capacity*, has changed because the other two firms have selected different capacities.

STEP Copy cell B20, select cell J10, and *Paste Values*, then run Solver. Copy the new optimal Q (cell B27) and paste it in cell K10.

Notice that we still do not have an internally consistent solution between the two optimization problems. The firm capacity optimal solution is different from the total capacity used by the cartel. We must iterate.

STEP Do these three steps:

1. In the *CartelStage2* sheet, select cell B30, and paste the value of optimal capacity.

2. Return to the *CartelStage1* sheet and copy cell B20, select cell J11, and *Paste Values*.

3. Run Solver to find the new optimal solution. Copy the optimal Q (cell B27) and paste it in cell K11.

We still do not have a situation in which the optimal capacity decision of Firm 1 agrees with the total capacity parameter used by the cartel.

STEP Fill in the Stage 1 and Stage 2 Consistency table. You will need to iterate, repeating the process of solving for Firm 1's optimal capacity, pasting that result in the CartelStage1 sheet, then returning to the CartelStage2 sheet to see if the two solutions coincide (the three steps above).

STEP When you have finished completing the table, click the Show Data button.

This reveals results in columns L, M, and N that are based on your iterations. It also shows the Nash equilibrium solution for q_i*. As with our work in the

Cournot model earlier, there is an analytical solution to each firm's optimal and consistent capacity and we entered it in cell K19.

Figure 17.30 shows what your screen should look like. The total capacity, the vertical line in the *CartelStage2* chart, is driven to an equilibrium value of 3,891,176 units. The total capacity line bounces right and left until settling down at a value that is consistent with the optimal solution to the individual firm's profit maximization problem. In equilibrium, each firm will have a capacity of 1,297,059 units. This is consistent in the sense that each firm would choose this capacity if it knew the sharing rule adopted by the cartel.

Stage 1 and Stage 2 Consistency

Iteration	Other Capacity	qi*	Starting Total Capacity	Ending Total Capacity	Difference
1	2,400,000	1,342,758	3,600,000	3,742,758	(142,758)
2	2,685,515	1,273,616	4,028,273	3,959,131	69,142
3	2,547,232	1,308,620	3,820,848	3,855,852	(35,004)
4	2,617,240	1,291,240	3,925,860	3,908,480	17,380
5	2,582,480	1,299,959	3,873,719	3,882,438	(8,719)
6	2,599,917	1,295,607	3,899,876	3,895,524	4,352
7	2,591,213	1,297,784	3,886,820	3,888,997	(2,178)
8	2,595,569	1,296,696	3,893,353	3,892,265	1,088
9	2,593,392	1,297,240	3,890,088	3,890,632	(544)

Nash eq q_i	1,297,059
Nash eq Total Q	3,891,176

Figure 17.30: Nash Equilibrium capacity.
Source: CartelDWL.xls!CartelStage1.

Given the demand curve parameters, marginal cost, and the world price, we know the cartel's profit-maximizing domestic output and price. Because we know the equilibrium solution to each firm's capacity decision, we can compute the total output produced and export loss. Thus, we can compute *CS*, *PS*, and *DWL*.

STEP Copy cell K19 from the *CartelStage1* sheet and *Paste Values* in cell B30 of the *CartelStage2* sheet.

STEP Click on cells B18, B19, and B21 to display the *CS*, *PS*,and *DWL* generated by the cartel solution.

Cartel Model Summary

Determining the cartel's output is not easy. One has to solve a two-stage game. The cartel's sharing rule means that each profit-maximizing firm is willing to trade off export losses in order to get a share of high-priced domestic output.

The vertical total capacity line in the *CartelStage2* chart is actually an equilibrium solution to the first stage of the game. There is only one value of total capacity that is internally consistent with individual firm capacity decisions.

The cartel game-theoretic model also can be solved via analytical methods. The mathematics is not easy, but if you are interested in seeing the solution, click the Show Analytical button near cell M5 of the *CartelStage1* sheet.

Having determined the output and price solutions to each of the three market structures, we are ready for the welfare analysis.

Comparing Monopoly, Cournot, and Cartel Solutions

STEP Proceed to the *Compare* sheet.

Given the parameter values (in the shaded cells), the table displays the output, price, *CS*, *PS*, and *DWL* associated with perfect competition, monopoly, cartel (with the sharing rule), and Cournot market structures.

Cells B18:B21 are connected to the market structure currently displayed on the chart. Initially, the perfectly competitive result is displayed. *DWL* will be computed against this standard.

STEP Click the *Monopoly* option.

Cell range B18:B21 is updated and the chart displays the monopoly result. Notice that the monopolist ignores the world price and does not export cement. She maximizes profits by choosing output where $MR = MC$.

Compared to perfect competition (in cells B10:B14), monopoly generates much lower *CS*, higher *PS*, and a substantial *DWL*.

STEP Click the *Cartel* option.

The chart displays the total capacity vertical line and the exports are highlighted. We can compare the cartel to the monopoly and PC results by looking at the cells in columns B, C, and D, in rows 10 to 14.

Note that for the cartel option, cell D13 shows the value of profits for the cartel. This is domestic *PS* less export loss. Cell B19, also labeled *PS*, shows domestic producers' surplus (and leaves out the export loss). This is confusing, but it allows separation of the two sources of total *DWL*, domestic *DWL*, given in B21, and export loss, shown in cell B22, and ensures that domestic *DWL* plus total surplus will sum to total surplus in the perfect competition case. Total *DWL*, the sum of domestic *DWL* and the export loss, is reported in cell D14.

Compared to perfect competition, the cartel generates lower output and higher prices, but it is better than monopoly. Cells G10:G14 show what happens when you move from cartel to monopoly.

STEP Click on cells G10 to G14 to see their formulas.

If the Norwegian cement industry merged to monopoly from a cartel, we would see the following: Output falls, price rises, *CS* falls, *PS* rises, and *DWL* rises.

The increase in *DWL* would enable to us to judge such a move as a failure in terms of resource allocation in the Norwegian economy.

STEP Click the *Cournot* option.

Comparing cartel and monopoly to perfect competition is not particularly useful, because we are not going to get a perfectly competitive cement industry. There are only three firms. If we had competition, it would be Cournot competition. The three firms would not collude, but they would behave strategically.

If the industry went from cartel to Cournot, cells F10:F14 show what would happen. As with cells G10:G14, these cells report the difference from the cartel to the Cournot market structure. Notice that output rises, price falls, *CS* rises, *PS* rises, and *DWL* falls.

Of these effects, *PS* rising is surprising, but remember that under Cournot, the export losses are eliminated.

This completes the theoretical welfare analysis. The results are clear: To maximize surplus, the Norwegians should have moved from a cartel to Cournot competition. Of the three market structures, Cournot has the lowest *DWL*.

There is, however, one important issue left unresolved: These results apply only to the parameter values on the sheet. We do not know the intercept or slope of the Norwegian demand curve for cement, nor do we know *R* or *MC*. We need to get these parameter values, and then do the analysis based on these real-world parameter values.

Welfare Analysis for 1968

STEP In the *Compare* sheet, scroll to the right of the graph and click the Show Actual button (over cell N1).

After clicking the button, a new sheet appears, populated with key parameters for 1968, the last year of the cartel.

Figure 17.31 shows the results for the various market structures for the estimated demand curve for 1968. The conclusion is clear—Cournot is the best of the three feasible market structures. It produces the highest output, lowest price, highest *CS*, and lowest *DWL*.

Exogenous Variables for Demand and Marginal Cost						
P = d0 - d1Qd			MC = s0		World Price	
d0	953.7203057		s0	288.6596342	(R)	
d1	0.000252144				235.0740945	
	Perfect Competition	Monopoly	Cartel	Cournot	Cartel to Cournot	Cartel to Monopoly
Q (domestic)	2,637,622	1,318,811	1,425,071	1,978,217	553,146	(106,260)
P	kr 288.66	kr 621.19	kr 594.40	kr 454.92	(kr 139.47)	kr 26.79
CS (millions)	kr 877	kr 219	kr 256	kr 493	kr 237.332	(kr 36.758)
PS (millions)	kr 0	kr 439	kr 391	kr 329	(kr 61.745)	kr 47.891
DWL (millions)	kr 0	kr 219	kr 230	kr 55	(kr 175.587)	(kr 11.133)

Figure 17.31: Welfare analysis.
Source: CartelDWL.xls!CompareActual.

Figure 17.31 also makes clear why the industry went to monopoly instead of Cournot after the cartel collapsed (under the weight of overproduction

and export losses). PS would rise when moving from Cartel to monopoly (by 47,891,000 kroner), but fall (by 61,745,000 kroner) if the industry had adopted a noncooperative Cournot arrangement. Thus, it is clear that the cement industry chose to maximize its own PS instead of $CS + PS$. This is not surprising.

In fact, Röller and Steen build an even stronger case by exploring the welfare effects over several years. Scroll to column AE and read the text box if you are interested.

STEP Click the *Monopoly* option to display the monopoly solution in the graph.

The monopolist would choose output where $MR = MC$ and charge the highest price possible for that level of output. Monopoly profit in 1968 would have been 439 million kroner. Consumer surplus would be much smaller than under perfect competition and Norway would suffer a deadweight loss from monopoly of 219 million kroner.

But the Norwegians did not have a monopoly before 1968, they had the cement cartel.

STEP Click the *Cartel* option.

The cartel chooses output where $MR = R$, allocates the domestic output to the three firms based on capacity shares, and exports the excess output.

Notice, however, that Röller and Steen do not use the predicted capacity based on the demand curve parameters. Instead, they use actual exports. The story here is that capacity takes time to build. The cartel puts persistent pressure on expansion, but the firms do not actually reach their goal of vast capacity because the cartel collapses.

STEP You can check the theoretical cartel solution for the estimated parameters by simply copying the range A5:F8 from the *CompareActual* sheet and pasting in the same range in the *Compare* sheet. Click Yes if prompted to replace the destination cells. You may need to click the *Cartel* option to refresh the screen.

Figure 17.32 shows the result. Capacity is huge and export losses are staggering. This is the capacity that would have been installed in the long run

under the cartel. Röller and Steen do not use this capacity value. Instead, they use actual exports, based on the actual capacity in 1968.

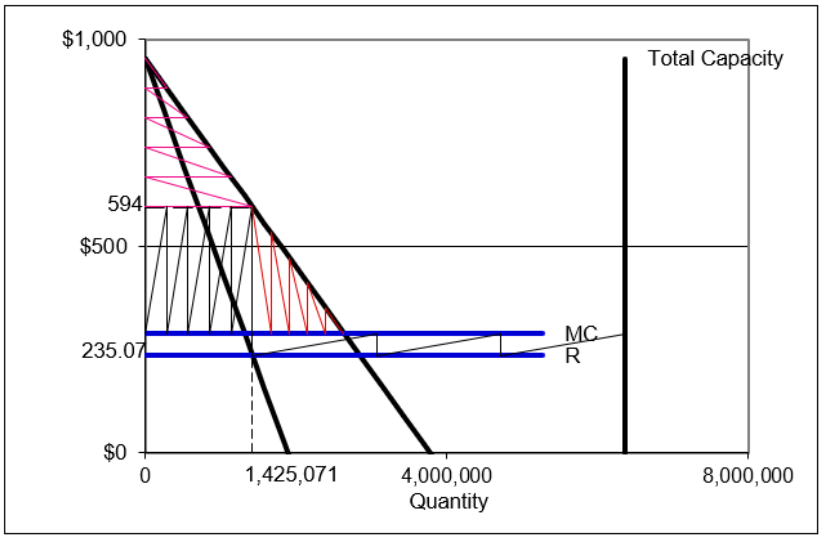


Figure 17.32: Cartel results with capacity determined theoretically. *Source: CartelDWL.xls!Compare* using estimated parameters for 1968.

STEP Return to the *CompareActual* sheet. Focus on columns F and G.

We know the firms merged to monopoly and the cartel to monopoly column (G) shows the welfare implications of this move for just 1968. As expected, output falls and price rises, *CS* falls and *PS* rises. The net welfare effect can be computed as the sum of the changes in *CS* and *PS*, which is an 11 million kroner increase (in cell G15).

Alternatively, the net welfare effect can be determined by looking at the reduction in *DWL* in cell G14. Because *DWL* falls as we move from cartel to monopoly, this number is negative. But notice that the absolute values are the same.

Our standard models tell us that merger to monopoly is the worst possible outcome—monopoly generates the greatest *DWL* of any market structure. However, because of the sharing rule, welfare actually increases when the cartel merges to monopoly because monopoly does not suffer export loss.

STEP Compare the values in Table 3 for the cartel to monopoly in 1968 to the values in column G.

The slight differences are due to rounding and precision differences.

Although monopoly beats the cartel, this is a poor argument for supporting monopolization. After all, the cartel could have dissolved into a noncooperative, Cournot competition. We must examine the welfare effects of this move and compare it to moving to monopoly to find the better option.

STEP Compare the red circled value of the change in *PS* when moving from *Cartel* to *Cournot* in Table 3 to cell F13. These numbers should be the same, but they are not.

Röller and Steen made a mistake in computing the net welfare effect for the move from cartel to Cournot in Table 3. They report the change in domestic *PS* in the table, not the change in total *PS*, which includes the export loss. As a result, the net welfare effect for cartel to Cournot in Table 3 is also incorrect. By failing to include the export loss in the reported *PS*, they underestimated the welfare gain from adopting a Cournot noncooperative market structure.

This error does not change Röller and Steen's conclusion. In fact, if anything, their results are strengthened once the export loss is accounted for. The loss in *PS* that the cement industry undergoes in moving to Cournot competition is not as bad as Table 3 suggests because of the elimination of the export loss. The true net change in welfare is some 45 million kroner higher than Table 3 estimates.

Consequences of Using Actual Versus Theoretical Total Capacity

Now that we understand how net welfare effects for 1968 are computed, we turn to the issue of how the export loss is measured.

Cell D20, the export loss in 1968, is based on actual exports—the difference between actual capacity (total production) and domestic output.

Figure 17.32 and your *Compare* sheet show that at the Nash equilibrium, long run capacity is much higher than the actual capacity (based on actual total production). How does this impact the analysis? This is an important question with a surprising answer.

STEP Compare the formulas in cells G16 and G17. Both display the same number, but the formulas are different.

G16 computes the net welfare gain from going to Cournot instead of monopoly (from the cartel, of course) by taking *DWL* from cartel to Cournot minus the *DWL* from cartel to monopoly. Cournot beats monopoly by about 165 million kroner.

G17 computes the same net welfare gain, but does so by subtracting the net welfare effect from going to monopoly from the net welfare effect from going to Cournot. Once again, the move to Cournot beats the move to monopoly by roughly 165 million kroner.

STEP Copy the two cells, G16:G17, and go to the *Compare* sheet, pasting these cells in the same range.

The result is surprising—the superiority of Cournot over the cartel remains exactly the same, even though the *Compare* sheet is using theoretical, long run total capacity and the export losses are huge.

If you compare the values in columns F and G in both sheets, you will find that for both the move to monopoly and the move to Cournot, the change in PS and the change in net welfare are much higher if the theoretical capacity is used. This makes sense because the export loss is much greater.

However, the relative improvement in Cournot over monopoly remains the same because both Cournot and monopoly avoid export losses. Thus, the size of the export loss does not matter.

Had Röller and Steen used the theoretical, long run total capacity level based on the estimated parameters in 1968, their qualitative and quantitative conclusion regarding the superiority of Cournot over monopoly would remain completely unaffected.

Lessons from the Norwegian Cement Cartel

Röller and Steen (2006) evaluate the effectiveness of the (legal) cement cartel in Norway over the period 1955 to 1968. They solve monopoly, Cournot, and cartel models and compare the results. They find that because of the sharing rule adopted by the cartel, consumers actually did better (in terms of con-

sumer surplus) than they would have if the industry had been monopolized. Producers, on the other hand, lose in the domestic market with the cartel compared to a monopoly. Producers suffer an additional export loss under the cartel and this leads to a key result: The merger to monopoly that occurred in 1968 actually improved net welfare relative to the cartel outcome. This is certainly a surprise, given that we expect monopoly to be the worst market structure. The authors point out, however, that simply breaking up the cartel and allowing Cournot competition would have improved welfare even more.

The fact that Röller and Steen used actual exports instead of estimated exports makes no difference to their final conclusion that Cournot competition would have been the first-best choice. The reason it does not matter is that both monopoly and Cournot competition result in the elimination of the export loss, so in comparing a move to either Cournot competition or monopoly, the actual size of the export loss is irrelevant.

Röller and Steen (2006) give an excellent example of how economists use *CS*, *PS*, and *DWL* in policy analysis. It also enables deeper understanding of game theory by examining the two-stage game played by members of the cartel.

This section is certainly not typical of an Intermediate Micro course, but it offers advanced students a chance to see a sophisticated application of welfare analysis.

Exercises

Suppose the inverse demand curve is $P = 1000 - 0.5Q$, marginal cost is constant at \$100 per unit, and the world price is \$50. Enter these parameter values in the *Compare* sheet and answer the questions below. Enter the demand slope as a positive number, 0.5, and click one of the market structure options to refresh the chart.

The math theory prep section showed two surprising results. First, consumers' and producers' surplus under the cement cartel do not depend on the marginal cost of capacity. Second, as the number of firms in the cartel rises, the likelihood a merger to monopoly will be welfare enhancing rises.

To answer the questions that follow, taking pictures is helpful. You can select cells (e.g., A1:M25) and copy as a picture, then paste.

1. Increase *MC* from 100 to 200 and determine the impact on the cartel's *Q*, *P*, *CS*, *PS*, *DWL*, and export loss. What happens to each of these variables as *MC* rises?

   Be sure to click the *Perfect Competition* and then *Cartel* option button to refresh the data below the buttons.

2. Which changes, if any, in the variables are surprising? Why?

3. At what value of *MC* will there be no exports? Take a picture of this situation and paste it in your Word document.

4. Increase the number of firms from 3 to 5 (with *MC* at the no export loss value). What effect does this have on the cartel's *Q, P, CS, PS, DWL*, and export loss?

5. What can you conclude about the effect of the number of firms on *PS* from a merger to monopoly (from the cartel)?

References

The epigraph is from page 278 of Oliver E. Williamson, *The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting* (1985). Williamson applies the standard tools of economic reasoning (optimization and comparative statics) to transactions and argues that the institutions we observe are the evolutionary product of selection based on optimization.

This application is based entirely on the excellent paper by Lars-Hendrik Röller and Frode Steen, "On the Workings of a Cartel: Evidence from the Norwegian Cement Industry," *American Economic Review*, Vol. 96, No. 1 (January, 2006), pp. 321–338, www.jstor.org/stable/30034368. Special thanks to Frode Steen for making additional data available.

> The presence of people in the market who are willing to offer inferior goods tends to drive the market out of existence—as in the case of our automobile "lemons." It is this possibility that represents the major costs of dishonesty—for dishonest dealings tend to drive honest dealings out of the market.
>
> George A. Akerlof

17.8 Signaling Theory

We all want to live in a world in which every buyer and seller is always completely honest, dependable, and trustworthy. In such a world, no one would lie, cheat, or steal. No one would misrepresent a product or hide a defect to make a sale, and the buyer would always alert the cashier when receiving too much change. Even politicians and children would always tell the truth.

Plainly, we do not live in such a world. Cigarette manufacturers swear under oath that their products are safe and that there is no proof that tobacco causes lung cancer. Management lies to labor about the true profitability of the firm and the size of the wage increase that the firm can really afford. It seems that we live in the midst of lies and deceit. Few can be trusted and few trust us.

This then is the problem: How can we make our world—the one full of distrust and scams—more like the world we all agree is better—the one in which individuals are sincere and open? How can we get people to tell the truth?

Three Ways to Handle Dishonesty

We review utopian and authoritarian solutions to fighting dishonesty, and then focus on a third way that most people rarely consider.

If somehow it were possible to create a perfectly honest person, we could attain our goal of living in an honest world. People could be counted on, with no doubt or reservation whatsoever, to be completely clear and forthright. This is the *utopian solution.*

Karl Marx believed private property, money, and the capitalist system created an all-encompassing greed that generated fraud, deception, and a variety of other reprehensible individual behaviors. For Marx, the solution to the

problem was quite simple: Replace vicious capitalism with its superior evolutionary offspring, communism, and replace the money-hungry *homo economicus* with the noble *new socialist man*.

Although seemingly hopelessly idealistic, in certain cases, reliance on people's good qualities is, in fact, possible. We all have close friends and family whom we can trust to be sincere and truthful. In our daily lives, however, we deal with countless strangers, and we cannot rely on personal relationships to ensure honest behavior. In a modern society that incorporates the actions and decisions of millions of individuals, it is simply impractical to expect trustworthiness from everyone.

To protect against dishonesty, many people think immediately of monitoring. This second approach can be called the *authoritarian solution*.

If a store owner thinks customers are going to steal, valuable merchandise can be put under a glass counter, security cameras installed, and guards can watch the customers. If the government knows that citizens will cheat on their taxes, a sample of tax returns will be audited carefully to check for full compliance and severe penalties will be imposed on those caught cheating.

In general, the authoritarian approach to solving the problem of dishonesty requires a powerful judge who can check the truthfulness of statements and punish those who are caught violating the rules. Monitoring and punishment can work well when it is clear what constitutes a lie, and it is easy to observe the dishonest behavior.

Unfortunately, in many cases, it is quite difficult to determine dishonest behavior because there are shades of deceitfulness, ambiguities in truthfulness, and inherent uncertainty in the world. For example, if I sell you an expensive product, promising that it is of high quality, and then it breaks, am I a liar? It may very well be a high-quality good that just happened to break. Of course, I may have known that it was really shoddy merchandise and I just tricked you. How can you know which case is true?

In addition to that rather large subset of cases in which detecting dishonesty is nearly impossible, every application of the authoritarian approach suffers from a much larger drawback. To be effective, the powerful judge must be able to monitor individuals, including investigating alleged wrongdoing, determining guilt, and meting out punishment accordingly. This raises a serious concern: Who watches the watcher?

The inescapable paradox is that the stronger the authority, the more it will be able to control the individual, but also the more dangerous it becomes to the individual. Secret police, neighbors spying on friends, and severe control of individual behavior via strict rules and regulations seem the destiny of authoritarian schemes to coerce honesty from unwilling individuals.

There is little doubt that the authoritarian approach to the problem of dishonesty is the most common solution contemplated and applied. Faced with severe cheating, our first instinct is to call the referee and demand that force be applied to ensure truthfulness. There is, however, another alternative—one that does not suffer from the dangers inherent in the authoritarian solution.

Transforming humans to remove the driving force of self-interest or imposing authoritarian control to repress behavior driven by greed is like swimming against a powerful tide. The third approach is completely different. It is based on accepting self-interest and greed as immutable forces, but using them to get desired behavior. We can harness the power of self-interest in favor of our desired end. Individuals are free to decide to lie or not, but lying leaves them worse off. If honesty is the best choice from a self-interested point of view, then honesty is what we will get. This is the key idea underlying signaling theory.

An Economic Model of Used Cars

Suppose that there are only two kinds of used cars: high-quality A cars and low-quality B cars (called *lemons* in the United States). To keep things simple, suppose that there are equal numbers of each and that the high-quality A car is worth $10,000 while the low-quality B car is worth only $5,000.

The seller knows whether his or her car is of low or high quality, but the buyer does not. This is called *asymmetric information* because one party has knowledge and the other does not. The general problem of honesty, in this case, is reduced to figuring out a way to get sellers to tell the truth about the quality of the cars they are selling.

It is important to emphasize that, as illustrated in Figure 17.33, the buyer has no easy way to tell the cars apart. The underlying distribution of cars is on the left, and is known to the seller, but what the buyer actually sees is on the right.

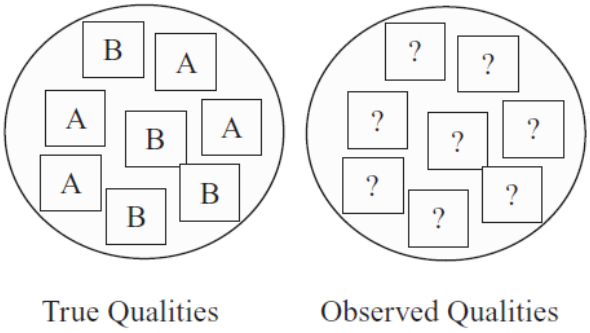


Figure 17.33: The problem of asymmetric information.

In a world where buyers cannot tell the difference between low- and high-quality cars and there are equal numbers of each type, buyers would expect to get a car worth \$7,500 on average. Half of the time they would get a \$10,000 car and the other half a \$5,000 car. Thus, on average, a used car would be worth \$7,500 and this is the amount buyers would be willing to pay for a used car.

Whereas sellers of low-quality cars would be quite happy getting \$7,500 for their low-quality cars, sellers of high-quality cars would be upset. After all, owners of A cars have a product worth \$10,000.

They might try to convince buyers to pay \$10,000 by making claims about the high quality of the car. Declarations about high quality, however, are likely to be ignored because the buyer has no way of knowing if the seller is telling the truth. After all, the seller might actually have a low-quality car worth \$5,000 and is lying to make more money. The buyer would worry that the seller's self-interest would dominate any desire to be honest.

The frustrated sellers of high-quality used cars simply leave the market. This phenomenon is an example of *Gresham's Law*, "bad money drives out good." It was first stated in the 16th century, when monarchs would debase coinage (by adding filler) to get more coins out of a given amount of gold. People would exchange the less valuable coins (bad money) and hoard the pure gold ones (good money). With more bad money in circulation, prices would rise.

Applied to the used car market, the low-quality used cars can be seen as driving out the high-quality cars. Left alone, we would not expect to see high-quality used cars for sale. In fact, that is not what happens—high-quality used cars are sold. How?

Instead of fixing the problem of dishonesty (lying about the quality of the car) by attempting to correct the unethical behavior of the sellers of low-quality used cars (whose dishonesty is causing the trouble here) or imposing authoritarian control over the used car sellers, an alternative scheme has evolved that has certain appealing properties—not the least of which is that car sellers truthfully reveal the qualities of their cars without any central, controlling authority.

Before explaining signaling theory, it is worth pointing out that what is happening here is actually an externality problem. The low-quality sellers fail to take into account the full cost of their lying and, therefore, they lie too much. No individual seller is aware, or would care, that his or her lying is contributing to the elimination of high-quality goods.

Another point that merits attention is that no one designed the system you are about to see. It emerged out of the interaction of buyers and sellers. Probably, some seller of a high-quality car got the idea and, when it worked, it was imitated, but you are about to meet another example, like supply and demand, of a decentralized system.

Signaling Theory

Developed by Spence (1973), the idea behind signaling theory is simple: when we cannot directly observe quality, we use a substitute that is observable (a signal) to enable the market to function. The signal is like a stoplight, green means go and red means stop. The signal will sort the combined low- and high-quality cars into separate markets.

Buyers cannot directly observe the quality of the car, but there are other observable characteristics bundled with the car and seller. *Indices* are attributes that cannot be changed, such as the age of the seller. *Signals*, on the other hand, are observable markers that can be acquired.

The signal, however, must have some special properties to be effective. The signal must be correlated with the underlying, unobservable characteristic. It must be something the A car owner is willing to do, but the B car owner is not, so that it is not immediately copied by unscrupulous sellers of low-quality cars.

In the case of used cars, a common signal is a warranty. Suppose that high-quality cars will have low warranty costs to the seller because they are unlikely to break, but the sellers of low-quality cars would face high warranty costs for their cars that will probably require many repairs.

We have gone as far we can in abstract terms and we are ready to see an Excel implementation of the signaling model.

STEP Open the Excel workbook *SignalingTheory.xls*, read the *Intro* sheet, then go to the *Optimizing* sheet.

The cost of the warranty to the sellers of A and B cars is depicted in Figure 17.34. With no warranty at all (the car is sold "as is"), at a warranty level of zero, a seller has no warranty costs—if something breaks after the car is sold, it is the buyer's problem.

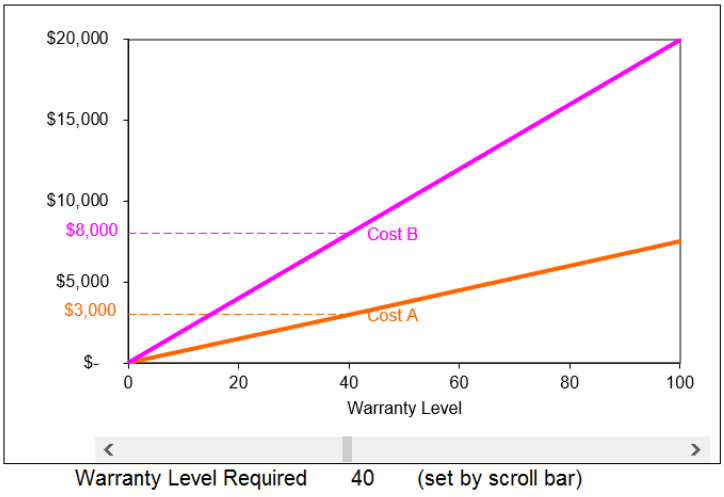


Figure 17.34: Seller's cost of warranty for each type of car.
Source: SignalingTheory.xls!Optimizing.

As the amount of warranty coverage offered by the seller increases, however, costs rise. The seller of the B car's costs rise faster so the gap between the two seller's warranty costs expands.

At a warranty level of 40 (this might be repairs covered by the seller for the first 12 months or 12,000 miles), in Figure 17.34, sellers of high-quality cars expect to incur costs of about $3,000, whereas the sellers of low-quality cars will pay around $8,000 for repairs.

The warranty cost functions are determined by the slopes in cells C6 and C7. It is easy to see that a seller's warranty cost is simply the slope parameter times the warranty level.

Now, suppose there was a warranty level, which is set at 40, initially. Buyers are willing to believe anyone who claims that their cars are high quality and pay the $10,000 price if and only if the car comes with a warranty level of 40.

So the warranty is the signal and any seller who acquires it will sell a car for $10,000. It seems like everyone will offer the warranty, right? Not so fast.

STEP Click the Show Price button.

Excel adds a price function to the chart. It is simply two horizontal lines with a break at a warranty level of 40. The hollow and solid dots mark the discontinuity. The solid dot means the endpoint is included and the hollow dot indicates it is not. Thus, any warranty level from zero up to the signal level (the hollow dot) means the car sells for $5,000. As soon as the signal level is reached, the price jumps to $10,000.

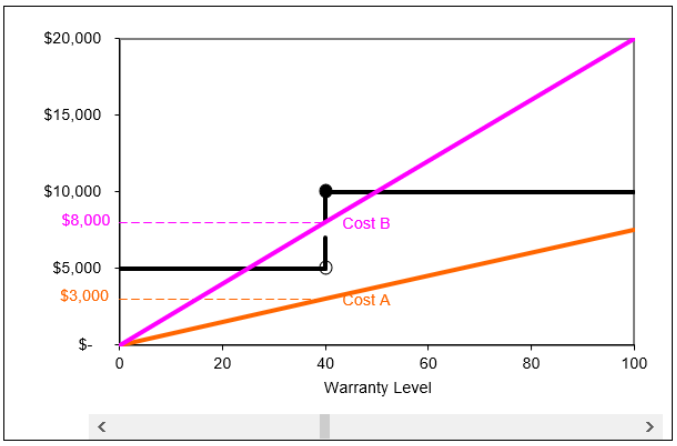
Anyone buying a car with a warranty level below 40 will be willing to pay, at most, $5,000 because it is assumed that the car is of low quality. Even if the car is actually a high-quality car, if it fails to come with the warranty level for high-quality cars, no buyer will pay $10,000 for it because the claim that the car is of high quality is unbelievable without the warranty. On the other hand, a buyer would be willing to pay $10,000 for any car with a warranty level of 40, even if it is actually a low-quality car.

It is now up to the sellers of used cars to make a decision of whether or not to lie. Sellers of low-quality used cars can claim that their cars are high quality and thereby receive the $10,000 high-quality price.

They will not misrepresent the quality of the car, however, because they would end up worse off. Their individual self-interest will drive them to tell the truth.

STEP Click the Show Net Gain button to see why low-quality sellers will not lie.

Figure 17.35 shows what is on your computer screen. We use data from the graph to create a table below that explains how the two sellers will behave.



Warranty Level Required 40 (set by scroll bar)

Decision Making
Net Gain Calculation

	Warranty Offered? No	Yes	Choice
A	$ 5,000	$ 7,000	Yes
B	$ 5,000	$ 2,000	No

Figure 17.35: Understanding why sellers will not lie.
Source: SignalingTheory.xls!Optimizing.

All sellers seek to maximize the net gain, or profit, from the sale of their goods and services. Sellers of used cars would not look simply at the fact that they can make $10,000 by offering a warranty level of 40. This decision-making strategy completely ignores the cost of the warranty. Instead, sellers must compare the net gain, price minus cost of the warranty, to arrive at an optimal decision concerning the warranty level.

The table below the graph contains each type of seller's net gain from selling a car with no warranty versus selling the same car with warranty level of 40. Read the table horizontally—for each type of seller, compare the net gain without and with the warranty, and choose the higher number.

It is clear that sellers of high-quality used cars will offer the warranty level and make $7,000 in profit because that beats the $5,000 net gain if no warranty is chosen. The sellers of low-quality used cars will choose to forgo the warranty and walk away with $5,000 because that is superior to the $2,000 net gain from choosing to lie and offering the warranty.

This is a rather remarkable result. To restate the outcome, the sellers of low-quality used cars will voluntarily and honestly admit that their used cars are of low quality and only worth $5,000. The sellers of low-quality used cars will not lie to the buyers. Is this because they suddenly were overcome by their conscience? No. They are the same fallible, less than perfectly honest people before and after the warranty scheme. Are they telling the truth because an authority figure is watching them, ready to punish liars? No. No one is watching them.

The sellers of low-quality used cars can lie if they so wish. They will not lie, however, because it is not in their self-interest. They end up worse off if they lie in this situation. The warranty scheme has managed to successfully separate or sort the two qualities of cars into their respective groups. This result is called a *separating equilibrium*.

Figure 17.36 shows that the warranty acts as a screen, separating the true car qualities into two distinct groups, Xs and Ys, from which it easy to tell which cars are high quality and which are not.

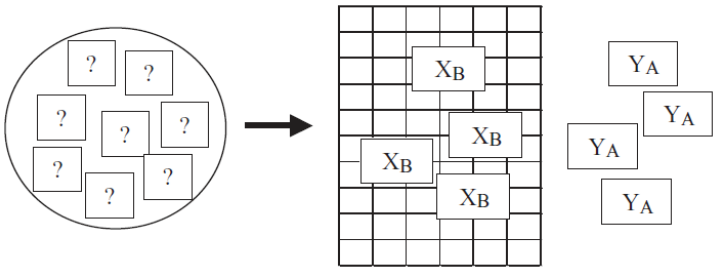


Figure 17.36: Warranty as a screen.

In essence, two markets for cars are created, one for low- and the other for high-quality cars, each with their own prices. Sellers of low-quality cars, although they are able to do so, will not lie and enter the high-quality car market because the price of admission is too high. Lying is not profit maximizing; therefore, sellers will not lie.

Let's repeat a key idea: no individual or organization runs this scheme. No one sets the warranty level and no one sets the price of the cars. The whole system bubbles up from the interaction of the two kinds of sellers and the buyers. Adam Smith would have called it an example of the *invisible hand* of

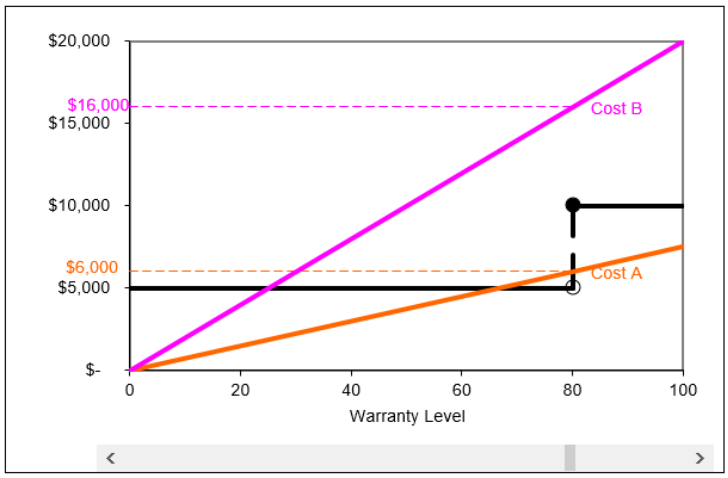
the market; Friedrich Hayek would have described it as a *spontaneous order*; and modern day mathematicians would speak of *self-organizing systems*. It is all the same thing—individual interaction generating a quite agreeable systemwide result. To see how the equilibrating forces operate in this model, we examine how the signaling scheme can break down.

Signaling Failures and Equilibrium

One way that a signal can fail is if it is set too high.

STEP Use the scroll bar to set a high warranty level like 80 or so.

In this case, as shown in Figure 17.37 and your computer screen, not even the sellers of high-quality cars find it in their self-interest to offer the warranty level that brings the $10,000 price. The signal has failed to separate the two qualities of cars.



Warranty Level Required 80 (set by scroll bar)

Decision Making
Net Gain Calculation

	Warranty Offered? No	Yes	Choice
A	$ 5,000	$ 4,000	No
B	$ 5,000	$ (6,000)	No

Figure 17.37: Signaling failure from a warranty level set too high. *Source: SignalingTheory.xls!Optimizing.*

On the other hand, if the signal is set too low, sellers of B cars will find it in their self-interest to lie and claim their cars are actually high quality. They

will choose the warranty level that brings the $10,000 price.

STEP To see this, use the scroll bar to set a low warranty level, 20 or less.

Your screen should show that both sellers opt to acquire the signal. The low-quality seller will lie and claim that the car is of high quality because the net gain from lying (cell H27) is greater than the net gain from telling the truth (cell G27). Once again, this signal has failed.

When the signal is too high, the holes in the screen are too small and no one can get through. If the signal is too low, the holes are too large and everybody passes through. In a separating equilibrium, the level of the signal is such that the two types are sorted and grouped together so they are easily identifiable.

The fact that signals can be observed as failing provides the key to understanding how the system can settle down to a result that effectively solves the problem without central control. If the signal is too low, self-interested sellers of high-quality cars will offer higher warranty levels in order to block their lying brethren from diluting their market. The sellers of high-quality cars want to distance themselves from low-quality sellers.

If the signal is too high, no one will take it and buyers will lose the means by which to identify the two qualities of cars. The market will collapse so pressure will push the level down.

The forces inherent in the system, self-interested behavior by the interacting agents, will conspire to generate an equilibrium signal level that effectively sorts the two qualities of cars. The process works just like supply and demand—pressure in disequilibrium pushes the signal in one direction or another until it equilibrates.

STEP Play around with the warranty level to reveal the range for which it effectively separates the two qualities of cars.

You already know 80 is too high and 20 is too low. Look at the chart to help you see what must be true for the signal to succeed. When you are ready to check your answer, click the Range button.

Other Applications of Signaling Theory

We have barely scratched the surface of signaling theory. There are many situations in which one party to a transaction has available information that the other party lacks and this asymmetric information puts honesty in peril.

Consider the job market (which was Spence's original example). Faced with many job applicants, all claiming to be high-productivity A workers, the firm might insist on a signal, a college degree, to back the claims made by job applicants.

Suppose that low-productivity workers are also likely to be weaker students, and that it is more costly for them to acquire the educational signal. As in the used car case, the successful screen will separate the two worker groups into their respective low- and high-productivity categories. The signal will elicit honest responses from low-productivity workers because lying requires a college degree to be believed and this is not in their best interest.

Additional applications of signaling include insurance, legal bargaining, and firm entry models. In both health and life insurance, asymmetric information is critical. The insurance company does not know the health status of the applicant. If the price of the insurance depends on the applicant's health, just saying they are healthy is not enough for the insurance company to believe it.

In a lawsuit, where the plaintiff seeks damages from the defendant, asymmetric information means neither party knows the other's true intentions and beliefs. They can signal the strength of their case by demanding a high pre-trial settlement.

Firm entry models use signaling to convey the degree of confidence and strength of incumbent firms to potential newcomers. Incumbents can signal or make reliable claims about their low costs and ability to compete by charging low pre-entry prices.

In these cases, an incentive mechanism has developed that accepts self-interest among buyers and sellers as a powerful, immutable, driving force. Instead of fighting self-interest by removing or suppressing it, the incentive mechanism uses self-interest to reach the desired end.

The Economics of Honesty

Dishonesty exacts a large cost on society. For lesser developed countries, corruption is a severe obstacle to economic growth. Getting people to be truthful is a serious, critically important goal.

The primary solutions to the problem of dishonesty have centered on utopian and authoritarian approaches. The former seeks to perfect human behavior; the latter to directly control it. A third, somewhat counterintuitive, alternative exists that relies on self-interest to yield an agreeable systemwide result.

This third alternative is marked by individuals following their self-interest. When geese fly in a V-shaped pattern over thousands of miles, they do so not under the guidance of an authoritarian drill sergeant or master goose who tells each bird where to fly, but because they obey a simple rule that says, "If there are no birds around, fly; if a bird is in front, fly just off its wing because it is easier." This minimizes the effort for each bird and produces a pattern which no bird intended.

Likewise, modern society is composed of millions of individual agents whose interaction establishes a systemwide pattern. Unsatisfactory results can be changed via transmuting the motivating forces of each agent, imposing decisions on each agent, or changing the incentives faced by each agent. The last option is rarely considered, but may be the most effective and best of the three.

Signaling theory says that by making honesty the best policy—for the selfish, greedy individual—we will get honesty. Sellers reveal the truth because lying leaves them worse off than telling the truth. This is the economics of honesty.

To be sure, signaling requires rules and institutional support. If the seller of low-quality used cars knows that he can renege on warranties or other contracts because the court system is nonexistent or corrupt, then signaling will be useless.

There is, however, a world of difference between an authoritarian approach that relies on a central power to coerce honesty and the system that evolves out of the interaction of the buyers and sellers given appropriately supporting institutions. The decentralized system avoids the question of "Who watches the watcher?" because there is no dominant, central power. And in the end, this may be its most significant advantage.

Exercises

1. Suppose a firm is trying to determine whether an applicant is of low or high ability and it believes people with long fingernails have higher ability. Would fingernail length be an effective signal? Draw a graph to support your answer.

2. Draw a graph that shows how education as a signal could be used to separate low- and high-ability job applicants. Explain how education as a signal works.

3. Draw a graph in which education as a signal fails because the signal level is set too high. Explain why the signal fails.

4. College education as a signal clashes with *human capital theory*, which says that educated workers earn more because they were made more productive by their education. What does signaling theory say about the value of education? In other words, according to signaling, why are educated workers paid more?

5. Why has it been difficult to determine with data whether human capital or signaling theory is right about college education?

References

The epigraph is from page 495 of George A. Akerlof, "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism," The *Quarterly Journal of Economics*, Vol. 84, No. 3 (August, 1970), pp. 488–500, www.jstor.org/stable/1879431. This is the paper that Michael Spence described as "quite electrifying" in his 1991 Nobel acceptance lecture (available at www.nobelprize.org). The Nobel Prize was shared that year by Akerlof, Spence, and Joseph E. Stiglitz "for their analyses of markets with asymmetric information."

Akerlof's paper led to an exchange concerning the empirical validity of the claim that lemons drove out high-quality used pickup trucks. Eric W. Bond, "A Direct Test of the 'Lemons' Model: The Market for Used Pickup Trucks," *The American Economic Review*, Vol. 72, No. 4 (September, 1982), pp. 838–840, www.jstor.org/stable/1810022 found no evidence for the claim.

Michael Pratt and George Hoffer, "Test of the Lemons Model: Comment," *The American Economic Review*, Vol. 74, No. 4 (September, 1984), pp.

798–800, www.jstor.org/stable/1805151 conduct a "finer test" and "conclude that the market for used pickup trucks is a lemons market."

In a reply, Eric W. Bond, "Test of the 'Lemons' Model: Reply," The *American Economic Review*, Vol. 74, No. 4 (September, 1984), pp. 801–804, www.jstor.org/stable/1805152 said that "Pratt and Hoffer find used trucks to be of lower quality not because they have a 'finer' test, but because they fail to adjust for observable quality differences and include trucks that are more than 10 years old." Bond believes there is no lemons effect for used pickups because institutions have arisen to counteract the effects of asymmetric information.

This debate and projects involving new data to test signaling would make an excellent thesis topic.

The signaling model is laid out in Michael Spence, "Job Market Signaling," The *Quarterly Journal of Economics*, Vol. 87, No. 3 (August, 1973), pp. 355–374, www.jstor.org/stable/1882010. This article was based on his doctoral dissertation and published as a book titled *Market Signaling* in 1974.

George Selgin, "Gresham's Law," available at eh.net/encyclopedia/greshams-law/, offers a short explanation of the history and application of this concept.

Chapter 18

General Equilibrium

The Edgeworth Box

General Equilibrium Market Allocation

Pareto Optimality

General Equilibrium Monopoly

"[Irma Adelman] was an early proponent of simulation models. In addition to work with input-output and linear-programming models, she was one of the pioneers in developing computable general equilibrium (CGE) models and applying them to developing countries, especially for analysis of income distribution."

Distinguished Fellow Citation for Irma Adelman

18.1 The Edgeworth Box

We have become quite familiar with society's resource allocation problem. We have used partial equilibrium analysis to focus on a single commodity, exploring how supply and demand determine an equilibrium quantity that is the market's answer to the resource allocation question.

We know all about consumers' and producers' surplus, market failure, and deadweight loss. We have repeatedly drawn supply and demand graphs and emphasized comparison of equilibrium to socially optimal output.

But the focus on a single commodity is limiting. In fact, the market system uses supply and demand for each good or service to answer the fundamental production and distribution questions. In other words, there are many interacting markets (one for each commodity) simultaneously in operation.

If we monopolize one commodity, we cause a misallocation of resources in the monopolized market (too little is produced). Partial equilibrium analysis stops there. But the low output and high price in the monopolized market reverberates throughout the economy. After all, resources that would have gone into that market are going to go somewhere else and the high price in the monopolized commodity will shift demand curves for substitutes and complements of that good.

General equilibrium analysis attempts to account for supply and demand in *all* markets at once. As you can imagine, it is much more difficult than partial equilibrium analysis, but it is also superior because the entire resource allocation question is under consideration.

This book focuses on general equilibrium *theory*, but as the epigraph to this chapter explains, computable general equilibrium models are used to estimate the general equilibrium effects of tax policies, monopoly power, and other events. Economists have always been aware of the limitations of par-

tial equilibrium analysis, but it was not until the development of modern computers that these complicated models could be solved and applied.

Before beginning our study of general equilibrium theory, two observations are in order.

1. Society can decide which goods and services are handled by the market. Society may decide that human organs or votes may not be legally bought and sold. Different market-based societies may choose different lists of commodities to be allocated by the market. We call a society *market based* if individual resource owners make decisions about how to allocate the inputs they manage, even if particular commodities are regulated or entire sectors of the economy (such as education or health) are not privately owned.

2. A complete general equilibrium analysis of the market system is beyond our scope. There are three parts, of which this book covers only the first one.

   (a) Pure exchange: Assume each consumer has endowments of already produced goods and allow trade to occur.

   (b) Production: Allow goods to be produced from inputs.

   (c) Combine pure exchange and production in a general equilibrium analysis.

We focus solely on pure exchange and ignore the next two stages. This means we will not complete a true general equilibrium analysis of the market system. Emphasizing only the problem of pure exchange enables you to see the core concepts of general equilibrium, including the Edgeworth Box graph, without overwhelming complexity.

Even limiting ourselves to a situation where all products are already made requires serious investment of intellectual capital. As we will see, the Edgeworth Box is a clever graph, but it takes some practice to read it.

Our work on pure exchange will enable us to come full circle and return to the beginning—consumers decide what to buy and sell based on the optimal solution to an Endowment Model. As you work on the model and recall ideas and terminology, you will further cement truly fundamental knowledge.

Constructing the Edgeworth Box

The canonical graph used to depict a pure exchange economy is called the *Edgeworth Box*. It is also commonly referred to as the Edgeworth-Bowley Box. It turns out that both names are wrong. Blaug (1996, p. 523), discussing something called the *Ricardo Effect*, points out an interesting thing about names:

> Whether it really is in Ricardo is a nice question. The fact that the Ricardo Effect is hard to find in Ricardo exemplifies a general rule. According to R. K. Merton, 'eponymy' is the "the practice of affixing the name of the scientist to all or part of what he has found" but it is a striking fact that the outcome of eponymy is almost always to hang the right label on the wrong person. Thus, Thomas Gresham never stated Gresham's Law. Jean Baptiste Say only stated Say's Law after James Mill had stated it for him. Robert Giffen never stated Giffen's Paradox. Francis Edgeworth never drew the Edgeworth Box. Ernst Engel never drew an Engel's curve. Walras never stated Walras' Law. Irving Fisher did not invent the Ideal Index Number and actually pleaded (in vain) that it should not be named after him. Arthur Bowley did not enunciate Bowley's Law. Arthur Pigou did not state the Pigou Effect—and so on. Indeed S. M. Stigler has advanced "Stigler's Law of Eponymy: No scientific discovery is named after its original discoverer," a law which is confirmed as soon as it is stated (see *Transactions of the New York Academy of Sciences*, Series 11, 39, 1980). Nevertheless, there are also counter-examples in economics to Stigler's Law, such as Pareto-optimality and the Wicksell Effect.

If it was not Edgeworth, then who created the canonical graph of general equilibrium analysis? According to Tarascio (1972), it was Vilfredo Pareto (pronounced pa-ray-toe) who should be credited with inventing the graph that we call the Edgeworth Box. Because no one has ever heard of the Pareto Box, we will continue to call it the Edgeworth Box, but now you know the truth behind the name.

The Edgeworth Box is a graph that is constructed by putting together the consumer choice problem graphs from two consumers. It ends up looking like a box; hence its name. While most books just draw a box, we can use Excel to see exactly how you build an Edgeworth Box.

STEP Open the Excel workbook *EdgeworthBox.xls* and read the *Intro* sheet, then go to the *A* sheet to see consumer A's optimization problem.

Take the time to look over the sheet. The goal is to maximize satisfaction, given by a Cobb-Douglas utility function that faithfully reflects the consumer's preferences. The budget constraint's slope is $-\frac{p_1}{p_2}$ and at the initial endowment (35,10), the MRS is less than the price ratio.

You know you do not need to run Solver because at $25{,}16\frac{2}{3}$ (the actual values on the sheet are Solver's false precision) the equimarginal condition is met and the consumer is reaching the highest attainable indifference curve.

At the given prices, the sheet shows that A will maximize utility, subject to the budget constraint, by selling 10 units of x_1 and buying $6\frac{2}{3}$ units of x_2. These are the *net demands* for x_1 and x_2.

STEP Proceed to the *B* sheet to see consumer B's optimal solution.

Notice that B has a different initial endowment (5,30) than A, but the rest of the optimization problem is the same. Given the same prices faced by consumer A, consumer B optimizes by buying 20 units of x_1 and selling $13\frac{1}{3}$ units of x_2.

Figure 18.1 has Endowment Model graphs for the two consumers. We can see that they make different decisions about what to buy and sell. A moves up the constraint (selling x_1 and buying x_2), while B does the reverse.

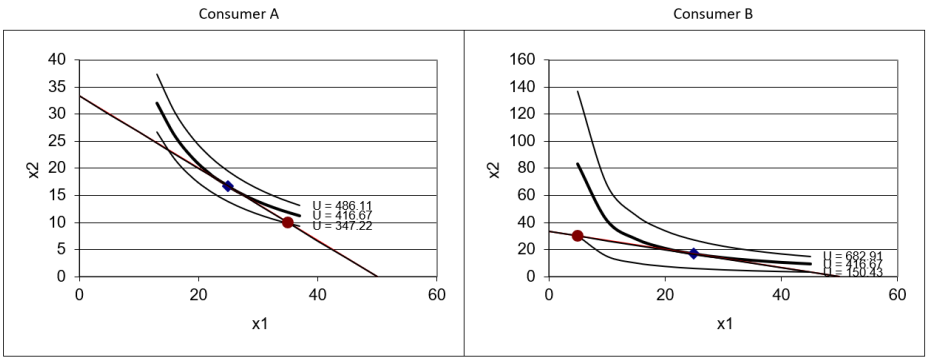


Figure 18.1: Preparing to build the Edgeworth Box.
Source: EdgeworthBox.xls!A and B.

Figure 18.1 shows the two consumers side by side and that helps us see what they are both doing, but it does not show how their plans for buying and selling match up. This is the key to the Edgeworth Box. We want to be able to instantly see if the two consumer's optimal decisions mesh.

The crucial step in understanding the Edgeworth Box is the next one: Flip consumer B's graph, as shown in Figure 18.2. Sheet *B* in Edgeworth-Box.xls shows how to do this.

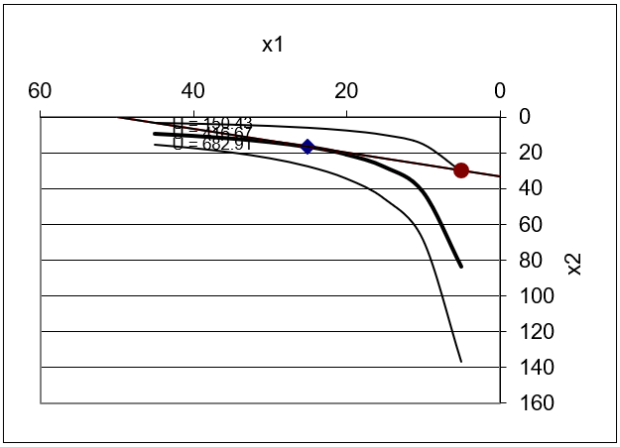


Figure 18.2: Flipping B's graph.
Source: EdgeworthBox.xls!B.

STEP Follow the instructions in column F of sheet *B* to replicate Figure 18.2.

Actually flipping B's graph will help you remember that B's decisions about buying and selling are always read from the perspective of the northeast (top right) corner of the Edgeworth Box.

The last step in constructing the Edgeworth Box is to join A's graph with B's flipped graph. The result of this operation is a graph that looks like a box.

STEP Proceed to the *EdgeworthBox* sheet for your first look at an Edgeworth Box. You may need to scroll down a bit to see it.

How is this chart created? By following the instructions above and taking advantage of Excel's ability to make transparent objects.

STEP Click on the graph to select it, and then drag the graph to the right.

It comes apart! Clearly, the Edgeworth Box is simply two separate graphs superimposed on top of each other. The top graph has no fill, so it is transparent.

STEP Click the Make Box button to put the box back together. The button simply lines up the two graphs precisely to make it easy to create the box.

STEP Scroll back up to see the organization of the sheet.

Let's take a tour of the sheet. The two consumers' optimization problems are represented in columns A and B and columns M and N. In the middle (columns G and H), market information is displayed. Cells H16 and H17 contain the prices of the two goods.

The price of good x_2, called the *numeraire*, has been set equal to 1 and p_1 is expressed as $\frac{p_1}{p_2}$. Instead of $p_1 = 2$ and $p_2 = 3$, we can focus on $\frac{p_1}{p_2} = \frac{2}{3}$ as the relative price. With many goods, a single one is chosen (think gold) as the numeraire and everything is priced relative to that good. In the next chapter, we will see how prices respond to supply and demand.

Properties of the Edgeworth Box

The Edgeworth Box has properties and conventions that will be helpful in our future work. Here are a few of them.

1. The sides of the box give the total amounts of the two goods available. Total $x_1 = 40$ units and total $x_2 = 40$ units so this box is a square.

2. If there is more total x_1 than x_2, then the box is wider than it is tall (if the same axis scale is used for both goods). The first exercise question asks what it means if the box is tall and skinny.

3. Since consumers face the same prices, one budget line is shared for both consumers.

4. The slope of the budget line is the price ratio, $\frac{p_1}{p_2}$, and that is what matters, not the individual prices themselves. By convention, we normalize the problem and set $p_2 = 1$, and call x_2 the numeraire.

5. Net demands for x_1 and x_2 for both A and B can be read from the box. This requires careful attention because it is easy to be tricked. Remember to read B's decisions about buying and selling from the top right corner.

6. The Edgeworth Box has enough information to figure out how prices will change and where the equilibrium solution lies. The next section shows how.

Edgeworth Box Basics

This section introduced the canonical graph of general equilibrium theory. It is unlikely that you have seen this graph before so we are proceeding slowly. Figure 18.3 shows the chart the *EdgeworthBox* sheet.

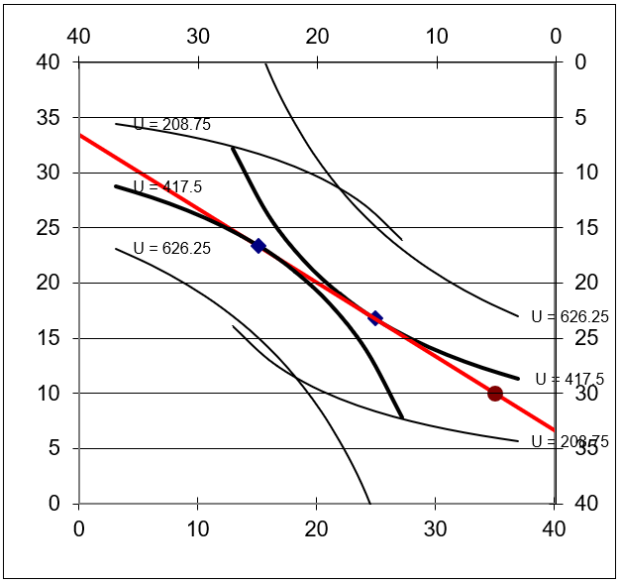


Figure 18.3: An Edgeworth Box in disequilibrium.
Source: EdgeworthBox.xls!EdgeworthBox.

The Edgeworth Box simultaneously displays the optimization problems of two consumers. A's view is the usual x–y axis configuration with the origin in the lower left corner of the graph. B's graph has been flipped so

the origin is at the top right corner. Thus, x_1 rises as you move to the left on the top of the box and x_2 rises as you move down the right side of the box.

If you drew an Edgeworth Box on a piece of paper or are reading this on a laptop or tablet, you could literally rotate the paper or device so that B was the usual configuration at the bottom left and A's axes were at the top and right. This would not change anything substantive.

In the next section, we will use the Edgeworth Box to see how both markets equilibrate simultaneously. This is the hallmark of general equilibrium analysis. Figure 18.3 is not in equilibrium. There are forces that will make the red budget line swing.

The Edgeworth Box will also be used to explain the concept of Pareto optimality and the idea of economic efficiency in a general equilibrium setting. Although it does not have the widespread recognition of supply and demand, the Edgeworth Box is a truly foundational graph in general equilibrium theory. It is important to grasp how it is constructed and read to be able to understand future concepts that rely on the Edgeworth Box.

Exercises

1. Suppose an Edgeworth Box was very tall and very skinny. What would that tell you?

2. Use Word's Drawing Tools to draw an Edgeworth Box that is the same as the *EdgeworthBox* sheet except B's utility function is $U = minx_1, x_2$. Draw three representative indifference curves for B.

   Hint: Return to the Theory of Consumer Behavior to find out what the indifference curves look like for this utility function.

3. Click the Reset button in the *EdgeworthBox* sheet and set cB in cell M21 to 0.1. Click the Take a Picture button and paste the graph in your Word document.

4. Explain B's buy/sell decision for each good.

5. How does B's buy/sell decision make sense given that B has so little of x_1 and so much of x_2?

References

The epigraph is from "Irma Adelman: Distinguished Fellow 2003," *The American Economic Review*, Vol. 94, No. 3 (June, 2004), www.jstor.org/stable/i369727. Advances in computers have enabled real-world, empirical applications of general equilibrium analysis. Computable general equilibrium models (CGEs) are used to find equilibrium solutions with many agents and commodities. The effects of taxes and other shocks are simulated and evaluated.

The history of how computers have been used in ever more sophisticated economic models is a story of determination and grit. See Irma Adelman, "The Research for the Paper on the Dynamics of the Klein-Goldberger Model," *Journal of Economic and Social Measurement*, Vol.32 (2007), pp. 29–33, content.iospress.com/articles/journal-of-economic-and-social-measurement/jem00269, to learn how Adelman and her physicist husband, Frank Adelman, used an IBM 650 mainframe computer in 1958 to produce one of her most famous articles, "The Dynamic Properties of the Klein-Goldberger Model," *Econometrica*, Vol. 27, No. 4 (October, 1959), pp. 596–625, www.jstor.org/stable/1909353. This was the first attempt to solve an econometric model with an electronic computer. Adelman (2007) also says that the work was "I believe, a first application of Monte Carlo techniques in economics." (p. 32)

In an introduction to Adelman's description of how the model was estimated, Renfro describes the IBM 650 and how incredibly impressive it was that Adelman managed to use it to estimate the model. In addition to covering her den with pieces of paper indicating the contents of each memory register at each step in the computation and having to pay more than a month of her salary for 1 hour of computing time, Renfro (p. 24) points out that the work had to be done at night. "Throughout the entire mainframe era, those who needed to get something done quickly worked through the night. Computers in those days had multiple users; this was the time of day that provided the best turnaround, when only the most serious were awake." See Charles G. Renfro, "Introduction," *Journal of Economic and Social Measurement*, Vol. 32 (2007), pp. 23–28, content.iospress.com/articles/journal-of-economic-and-social-measurement/jem00271.

Agent-based computational economics (ACE) is related to CGE. To learn more about "growing economies from the bottom up," visit www2.econ.iastate.edu/tesfatsi/ace.htm.

On the claim that it should be called the *Pareto Box*, see Vincent Tarascio, "A Correction on the Geneology of the So-Called Edgeworth-Bowley Diagram," *Economic Inquiry*, Vol. 10 (1972), pp. 193–197, onlinelibrary.wiley.com/doi/10.1111/j.1465-7295.1972.tb01599.x.

Economic Theory in Retrospect is a classic book on the history of economic thought (the intellectual history of the discipline) is Mark Blaug (1962 originally published, 5$^{\text{th}}$ edition, 1996).

> Without Pareto, the Theory of General Equilibrium, of which Walras was without question the real founder, would never have acquired the fame which it has now, nor indeed would it have been possible to speak of the Lausanne School.
>
> Umberto Ricci

18.2 General Equilibrium Market Allocation

Partial equilibrium analysis relies on supply and demand for a particular commodity to explain how the market establishes an equilibrium output that is society's answer to the resource allocation question. The figure X traced out by supply and demand lines is perhaps the most basic and well known picture in economics.

Compared to the easy, familiar supply and demand graph, general equilibrium analysis labors and struggles with a new graph, the Edgeworth Box, that is confusing when first encountered. It is busy, with many elements, and requires the user to change persepective to read it. As you work on mastering the Edgeworth Box, remember this: the equilibration process in an Edgeworth Box is based on the same logic used in supply and demand analysis.

We will leverage knowledge of supply and demand to explain how general equilibrium works and to learn how to read the Edgeworth Box.

Tatonnement: The Equilibration Process

Introductory economics students know that shortages cause prices to rise and surpluses push prices downward. In a supply and demand graph, the price is displayed as a horizontal line that falls when it is above the intersection and rises when it is below.

In the Edgeworth Box, there are two markets simultaneously equilibrating. The prices of the two goods are displayed by a single line, which is the budget constraint faced by the two consumers. The slope of the price line, also known as the *price vector*, is $-\frac{p_1}{p_2}$.

Just like supply and demand, shortages and surpluses push prices up and down. In the Edgeworth Box, this translates to the price vector swinging.

Remember that we are considering the special case of a pure exchange economy. All products have been produced and individuals are trading from their initial endowments. Prices are determined competitively by the interaction of all buyers and sellers—every consumer takes prices as given.

A two-dimensional Edgeworth Box allows for only two consumers. A third consumer would make it a cube and, beyond that, we run out of dimensions and cannot draw the object (although it exists). Our two-consumer, toy model version implements price-taking behavior by supposing that there is an *auctioneer* who shouts out prices. Our consumers take these prices as given and use them to make buy and sell decisions.

Although each commodity has a price, in general equilibrium analysis, only relative prices matter. We can arbitrarily take one good and set its price to 1. This makes that good the numeraire.

Our two consumers hear the prices and make optimizing decisions based on those prices. If the buy and sell decisions do not match, the prices are adjusted by the auctioneer. No trades are actually made until all markets are in equilibrium.

As prices are called out by the auctioneer, the price vector rotates around the initial endowment, swinging to and fro. It becomes more vertical as $\frac{p_1}{p_2}$ rises and flatter if $\frac{p_1}{p_2}$ falls. We mean, of course, rising and falling in absolute value.

At any moment, the consumers can compute the optimal amounts of each good to buy and sell. If the amounts each wants to buy and sell are not mutually compatible, then the price vector swings toward the equilibrium price vector.

The word *tâtonnement* (pronounced ta-tone-mon) was used by the French economist Leon Walras (1834 - 1910) (pronounced Val-rasse) to describe the equilibration process. Google translates it as groping. Walras visualized the market groping, feeling, working its way through an iterative process that converged to a position of rest. In the technical literature of general equilibrium theory, the word tatonnement (without the circumflex) is accepted without italics.

You may have noticed that the terminology of general equilibrium analysis has a decidedly French-language flavor to it. Walras, the father of general

equilibrium theory (and described by Schumpeter as "the greatest economist ever") was French. His successor at the School of Lausanne was Vilfredo Pareto (1848 - 1923), a native Italian with a background in math and engineering, who invented the concept of Pareto optimality (and is the actual originator of the Edgeworth Box).

In the second half of the 19th century, continental European economists were at the leading edge of general equilibrium theory and mathematical economics. This strong mathematical tradition continues today. French-born Gerard Debreu and Maurice Allais have won Nobel Prizes in Economics for their work in general equilibrium theory.

We will use Excel to implement a concrete problem with actual prices, surpluses, and shortages to see how the Walrasian model works.

STEP Open the Excel workbook *EdgeworthBoxGE.xls*, read the *Intro* sheet, then go to the *EdgeworthBox1* sheet.

We review the display, piece by piece. It is worth going slowly and being careful. There is a lot going on and the details matter.

Consumer A's optimization problem is in columns A and B. No need to run Solver—cells B11 and B12 contain A's optimal reduced-form expression. With a price vector with slope $-\frac{2}{3}$, consumer A would like to sell 10 units of good 1 and buy $6\frac{2}{3}$ units of good 2.

Columns M and N display consumer B's optimization problem. Like A, we have entered the reduced-form formulas for B's optimal consumption of the two goods. At the initial prices, consumer B wants to buy 20 units of good 1 and sell $13\frac{1}{3}$ units of good 2.

This information is all we need to know that the p_1 relative price in cell H16 is not an equilibrium, or market clearing, price. After all, A wants to sell more x_1 than B wants to buy and vice versa for x_2.

Thus, no trades will be made at these prices and the Walrasian auctioneer will call out new prices as the search for equilibrium goes on.

We can also use the Edgeworth Box to reach this same conclusion about the plans not matching at the initial relative price of -0.67.

STEP Scroll down to see the Edgeworth Box.

Figure 18.4 reproduces a portion of what is on your screen, augmented with arrows and dashed lines to help explain what is going on.

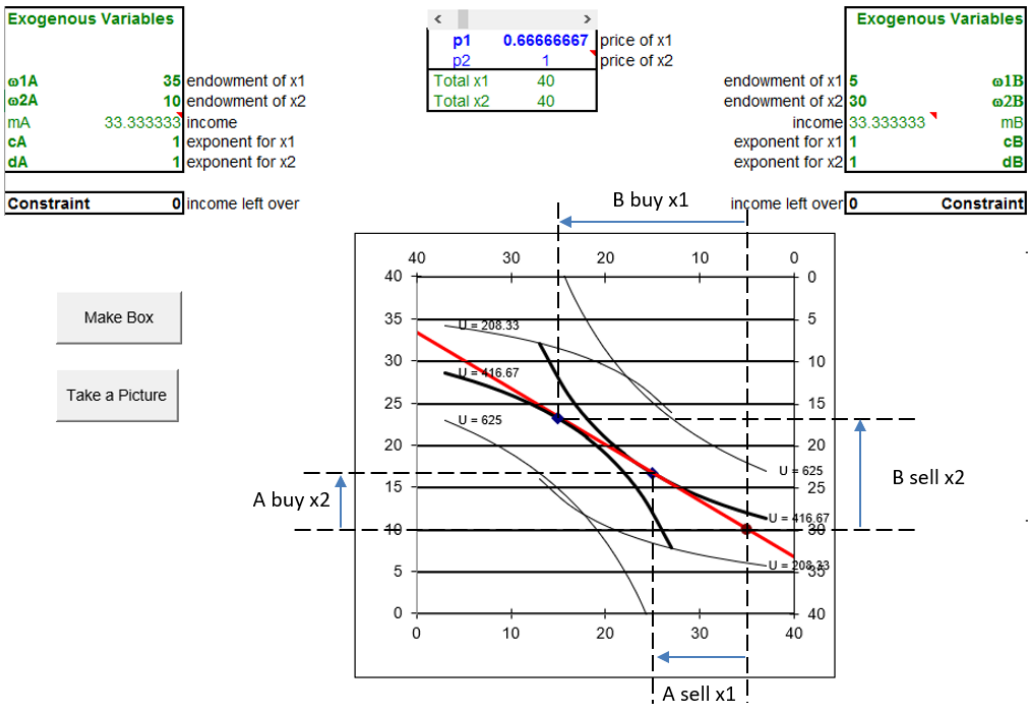


Figure 18.4: An Edgeworth Box in disequilibrium.
Source: EdgeworthBoxGE.xls!EdgeworthBox1.

We begin with A, which is easier than B. In Figure 18.4, arrows along the bottom and left sides of the box indicate what A wants to do: sell x_1 and buy x_2. It is natural to read the dashed lines from A's optimal solution and see that left on the x axis means sell, while up on the y axis means buy.

Reading B is trickier. B also has arrows, but they run the reverse of the usual because we read B's graph from the northeast corner. B wants to buy x_1 and sell x_2.

The direction of the arrow indicates buying or selling. Although one wants to buy and the other sell, the length of the arrows in Figure 18.4 show that the plans do not match. The length of the arrows indicate the *amounts* to be bought and sold. If the lengths are not equal, we are not in equilibrium.

We review the buy and sell decisions of B more carefully, to make sure there is no confusion. B wants to buy 20 units of good 1. From her initial endowment of 5 units, she wants to move *left* along the top axis, which means acquiring *more* x_1, until she ends up with 25 units. On the other hand, she wants to sell $13\frac{1}{3}$ units of good 2, moving *up* the right axis —which means she is *reducing* her desired amount of x_2.

If you get in the habit of drawing dashed lines on an Edgeworth Box, either on a piece of paper or by inserting dashed line shapes in Excel or Word, *from the optimal solution of A and B*, you greatly increase your chances of reading the graph correctly. Those dashed lines are a visual cue that remind you to read A from the bottom left and B from the top right.

STEP Scroll down below the Edgeworth Box to see two supply and demand graphs.

These are the partial equilibrium markets for the two goods. Good 1 shows a shortage, with price below the intersection of supply and demand. Good 2 has demand and supply reversed from the usual display because the price on the y axis is p_1/p_2. There is a surplus of x_2 at $p_1/p_2 = \frac{2}{3}$.

Both markets adjust simultaneously. We know there is upward pressure on p_1 from the shortage and downward pressure on p_2 from the surplus. This will make the price ratio rise and the price vector will become steeper.

STEP Use the scroll bar (over cells G15 and H15) to see how price changes affect the box. Set the price ratio to 1.5.

The spreadsheet does most of the hard work for you. A's and B's optimal solutions are instantly calculated. The market position cells immediately reflect the position of markets for each good at the new prices (where good 1 is one and a half times as expensive as good 2).

The Edgeworth Box is a live graph that reflects the new price vector. It shows that we have overshot the equilibrium price vector because we now have a surplus of good 1 and a shortage of good 2.

STEP Practice reading the Edgeworth Box. With $\frac{p_1}{p_2} = 1.5$, use the graph to read the amounts that A and B want to buy and sell. Compute the surplus and shortage of each good from the box alone.

Verify (using the cells in the Market Position part of the sheet) that your answers are correct. Look at the graphs below the Edgeworth Box to make sure you understand that the Edgeworth Box conveys the same information about the position of each market.

STEP Play with the price vector, adjusting the scroll bar to set different price ratios and interpreting how the consumers will respond to each price ratio by using the Edgeworth Box.

As you rotate the price vector, you are the Walrasian auctioneer. You are calling out prices and the two consumers are reacting to them. The more you practice reading the Edgeworth Box, the more comfortable you will get with it.

As you adjust the price ratio, the price vector swings to and fro. It always rotates around the initial endowment (which would change if and only if any of the four initial endowment parameter values change). The tatonnement process is how the market responds to shortages and surpluses by changing prices in such a way that the surpluses and shortages are reduced, until they are completely eliminated.

There is, of course, no auctioneer in the real world, but price pressure from surpluses and shortages are quite real. Our model captures these pressures by the fiction of the auctioneer changing prices in response to disquilibrium in the two markets.

General Equilibrium

You have seen how shortages and surpluses push the price line to and fro, swinging around the initial endowment point.

We know that equilibrium means *no tendency to change*. We apply this definition of equilibrium to this particular model: when $\frac{p_1}{p_2}$ has no tendency to change, we know we have settled to the equilibrium solution. The equilibrium solution generated by the market tells us how much x_1 and x_2 each consumer will end up with if the market is used and how much each consumer wants to buy and sell of each good.

STEP Use the scroll bar to find the equilibrium price vector.

The equilibrium solution in a General Equilibrium Pure Exchange Model is a canonical economics graph that is reproduced as Figure 18.5. If your screen does not look like this graph, set the price ratio to 1.

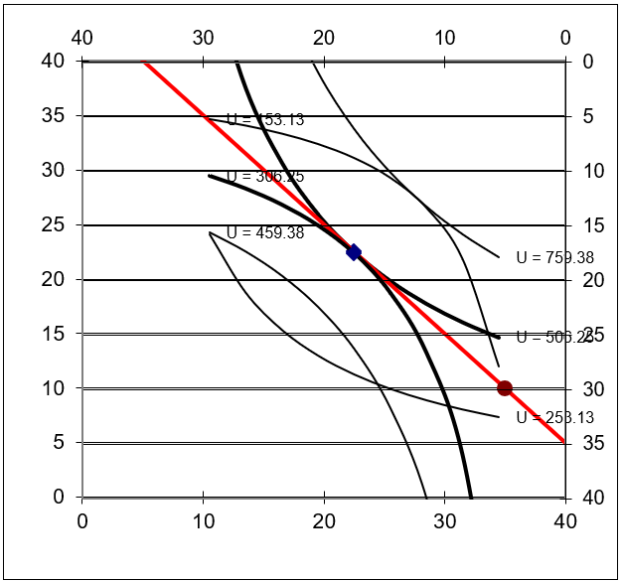


Figure 18.5: The canonical graph of general equilibrium.
Source: EdgeworthBoxGE.xls!EdgeworthBox1 with $\frac{p_1}{p_2} = 1$.

As Figure 18.5 clearly shows, when the equilibrium position is reached, the optimal solution of both consumers lies on the same point. This eliminates all shortages and surpluses (as shown in the supply and demand graphs below the Edgeworth Box) so the price ratio has no tendency to change.

The single point in the Edgeworth Box represents a mutually compatible solution for both consumers and is the hallmark of a general equilibrium solution. The single point is akin to the intersection of supply and demand in a partial equilibrium analysis.

Our general equilibrium model shows how the market is an allocation mechanism. It will redistribute the initial endowments of the two consumers by using prices until it settles down to a position where plans match and forces in the model are in balance.

Notice, however, that the two consumers don't get equal amounts of the two goods. Why does A end up with more? Because A started out richer. At the equilibrium price vector, the market values A's endowment at $45 and

B's at $35. General equilibrium theory does not ask why A is richer. It takes the initial endowment as given.

Walras' Law

Leon Walras is the father of General Equilibrium Theory. The law that bears his name states the following: *The value of aggregate excess demand is identically zero.*

Using Walras' Law, we can deduce the following logical result: If n–1 markets are in equilibrium, then the last market must be in equilibrium.

A concrete demonstration of Walras' Law is the best way to understand what it means.

STEP With $p_1 = 1$ (at the equilibrium solution), change p_2 (cell H17) to 2. Find the equilibrium p_1.

The equilibrium p_1 is now 2. This shows that, no matter the value of p_2, the equilibrium solution will be found when $\frac{p_1}{p_2}$ equals one.

Thus, it looks like there are two endogenous variables here, p_1 and p_2, but there is really only one endogenous variable, $\frac{p_1}{p_2}$. This is the idea behind Walras' Law and why we can find equilibrium in both markets by varying only p_1.

STEP Click the Reset button. Scroll right to cell V5 and click the Show Walras' Law button to reveal calculations that demonstrate Walras' Law in action.

Although the two markets are not in equilibrium, the sum of the value of aggregate net demands in cell Y11 is zero. Look at the cell formulas in row 11 to see how they are computed.

STEP Change p_1 (via the scroll bar) and notice that no matter the price, the sum of the value of aggregate net demand is always zero.
A direct implication of Walras' Law is that in a general equilibrium system with n goods, we do not have to find n prices. If $n - 1$ markets are in equilibrium, the last one automatically has to be in equilibrium.

This is why we actually have only a single endogenous variable, the price ratio, in the two-good case. All that matters is the relative price, not the two individual prices. With n goods, one good would be the numeraire (historically, gold has played that role) and all other goods would be valued in terms of the numeraire.

Comparative Statics with the Edgeworth Box

Having found the initial equilibrium solution, we could pursue a variety of comparative statics experiments, shocking an exogenous variable and tracking how the equilibrium solution (of various endogenous variables) responds.

STEP Click the Reset button and then set cA (cell B21) to 2. What happened to A's indifference curves and optimal solution?

With steeper indifference curves (since A likes good 1 more than before), A's new tangency point is quite close to the initial endowment. This means A wants to sell little x_1. You can scroll down to see how the partial equilibrium graphs have changed—the chart of x_1 confirms we have a big shortage.

STEP Where is the new equilibrium solution? If you decide to use Solver to answer this question, please make the target cell H15 because that is the cell that the scroll bar is affecting. This way you will not destroy the formula in cell H16.

You should find a new equilibrium solution at a relative price ratio of about 1.53. Approximately 7.3 units of good 1 will be traded and 11.8 units of good 2 will be exchanged.

Two Advanced Ideas

In a mathematical sense, General Equilibrium Theory is perhaps the most abstract and sophisticated area of economics. Two questions that have been studied intensively involve existence and uniqueness.

The question of the existence of an equilibrium solution was posed by Walras himself. The issue, loosely stated, is that we cannot be sure that a general equilibrium system with thousands or millions of individual goods has a place where the entire system is at rest. In fact, from an intuitive point of view,

given the huge number of products, consumers, and firms in a real-world economy, we might doubt that an equilibrium solution exists at all.

Walras and other early theorists thought that if the number of endogenous variables (unknowns) equaled the number of equations, then a solution was guaranteed. This is not so. Existence proofs in the 1950s utilized *fixed-point theorems* to prove rigorously the conditions under which an equilibrium solution was guaranteed to exist. Brouwer and Kakutani fixed point theorems are examples of this approach.

Closely tied to existence is the problem of the uniqueness of a general equilibrium solution. Even if an equilibrium solution is proved (in a rigorous mathematical sense) to exist, the worry is that there may be multiple equilibria in a general equilibrium system. Research has focused on what assumptions must be invoked to guarantee a single equilibrium solution.

Existence and uniqueness proofs are well beyond the scope of this book. They rely on topology and advanced mathematical concepts. This is another way of saying that our presentation of the Edgework Box and general equilibrium in a pure exchange economy is introductory and rudimentary. General Equilibrium Theory is a vast ocean and we are paddling near the shore.

Market Allocation in an Edgeworth Box

The canonical supply and demand graph is used in partial equilibrium analysis to find the equilibrium solution. General equilibrium uses the Edgeworth Box to do the same thing.

It appears cumbersome and tedious at first, but, in fact, it is an ingenious graphical device. By representing two consumers simultaneously, while sharing a common budget constraint (given that they face identical prices), the box enables one to quickly see whether the two-good, pure exchange economy is in equilibrium. It also reveals how prices must change as the system finds its way to equilibrium via the tatonnement process.

Whether a pure exchange economy is in a general equilibrium can be determined in an instant by seeing whether the optimal solutions of the two consumers are compatible—that is, if there is a single point where the two consumers want to be, given the existing price ratio.

But what about the final, equilibrium allocation generated by the market—what are its properties? This is a fundamental question that leads to the famous Pareto optimality conditions and the First Fundamental Theorem of Welfare Economics. It is explained in the next section.

Although we have used numerical methods (implementing the problem in Excel) to analyze and find the general equilibrium solution, you should be aware that there are analytical approaches also. We could write down demands for goods by each consumer and impose the equilibrium condition that $Q_D = Q_S$ in each market. This would enable solution of the equilibrium price vector with the aid of algebra (and, as soon as we left the simple world of two or three goods, linear algebra).

Exercises

1. Use Word's Drawing Tools to draw your own Edgeworth Box. Place the initial endowment so that A has more x_2 than x_1.

2. Add a price vector to your box in the previous question that generates a shortage of x_1. Draw arrows along the bottom and top x_1 axes to show the amount of x_1 each consumer wants to buy or sell.

3. Use Word's Drawing Tools to draw a supply and demand graph for x_1. Include a horizontal line in the graph that shows the current price of x_1.

4. Add the equilibrium price vector to your Edgeworth Box graph in question 1. Explain why this price vector is the equilibrium solution.

   Hint: Add indifference curves to your graph to support your explanation.

References

The epigraph is from page 11 of Umberto Ricci, "Pareto and Pure Economics," *The Review of Economic Studies*, Vol. 1, No. 1 (October, 1933), pp. 3–21, www.jstor.org/stable/2967433. You can learn more about Walras, Pareto, and the Lausanne School by visiting the History of Economic Thought web site at www.hetwebsite.net/het/.

Perhaps no area of economics is as mathematically sophisticated and intense as General Equilibrium Theory. There has always been disagreement among

economists regarding the use and necessity of mathematics in economics. Pareto sneered at the literary economists and the use of math as a weapon continues today.

Akerlof says that economists only value "hard" and ignore "soft" questions so the discipline stifles research into issues that cannot be answered with formal tools and models. See George Akerlof (2020),"Sins of Omission and the Practice of Economics," https://doi.org/10.1257/jel.20191573, 58(2), 405–418, doi.org/10.1257/jel.20191573.

Roy Weintraub traces the influence of math in economics in *How Economics Became a Mathematical Science*, published in 2002. For the connection between economics and physics, see Phil Mirowski, *More Heat than Light*, published in 1989.

> Except during short intervals of time, people are always governed by an elite. I use the word elite (It. *aristorocrazia*) in its etymological sense, meaning the strongest, the most energetic, and most capable—for good as well as evil. However, due to an important physiological law, elites do not last. Hence, the history of man is the history of the continuous replacement of certain elites: as one ascends, another declines.
>
> Vilfredo Pareto

18.3 Pareto Optimality

Evaluating the welfare effects with general equilibrium is the same as with partial equilibrium. First we determine the equilibrium solution, then we find the optimal solution, and last we compare the equilibrium to the optimal solution.

The previous section used an Edgeworth Box with a price vector to find the initial equilibrium solution. We know that shortages and surpluses swing the price line to and fro until it settles down where the plans of the two consumers are mutually compatible.

In this chapter, we use the Edgeworth Box to display the optimal solution. The price vector is removed because prices play no role in determining the optimal solution. Just as with partial equilibrium, we logically separate the equilibrium from the optimal solution. If the two agree, then we know we have a good result.

Optimality

STEP Open the Excel workbook *EdgeworthBoxParetoOpt.xls*, read the *Intro* sheet, then go to the *EdgeworthBox* sheet.

The workbook is quite similar to the EdgeworthBox sheet from the previous section, except there is no price or market position information. We are not interested in markets right now. We are focused on determining the optimal solution.

An omniscient, omnipotent social planner, OOSP, is charged with determining the optimal allocation, given the initial endowment.

With OOSP's special powers, we can reallocate the initial endowment as we see fit. Each point in the box is an allocation, distributing the total amounts of the two goods to A and B. We can arbitrarily give and take from one person to the other, choosing any point in the box. What should we do?

At first glance, it might seem that we would want to solve an optimization problem like this:

$$\begin{aligned} &\max U_A(x_{1A}, x_{2A}) + U_B(x_{1B}, x_{2B}) \\ &\text{s.t. } x_{1A} + x_{1B} = \text{Total } x_1 \text{ and } x_{2A} + x_{2B} = \text{Total } x_2 \end{aligned}$$

In other words, we could give consumers A and B the amounts of goods 1 and 2 that maximize the sum of the individual utilities subject to the total goods available.

This strategy suffers from a serious problem: *We cannot make interpersonal utility comparisons.* This brings us full circle to work we did at the very beginning of this book in the Theory of Consumer Behavior. Utility is ordinal, not cardinal. Monotonic transformations (that keep rankings intact) of utility are allowed. Utility has no meaning in terms of its units.

Thus, an optimization problem that aggregates individual utilities is invalid. It makes no sense to say that the utility of A is added to the utility of B to get a total utility. There are no common units with which to measure and add utility. You might as well say that you added three cars and four pencils and got seven carpencils.

There is, however, a way to judge and evaluate different allocations of goods to A and B. This is Pareto's great contribution to welfare economics.

Pareto developed logical rules that enable us to get around the limitations of utility. His basic idea was that you can compare two allocations in terms of better or worse so you can make statements about one allocation compared with another. He invented a new vocabulary for his rules and today we use his name when we work with these rules.

Pareto knew we cannot add utility, but we might be able to compare two allocations and declare which one is better. We proceed by example, using the Excel workbook and Figure 18.6.

From the initial endowment point in Figure 18.6, suppose we consider the point (30,15) for A and (10,25) for B.

Figure 18.6 reproduces what is on your screen. The two thicker indifference curves going through the initial endowment are the starting point. They represent the benchmark satisfaction to which we will compare other allocations.

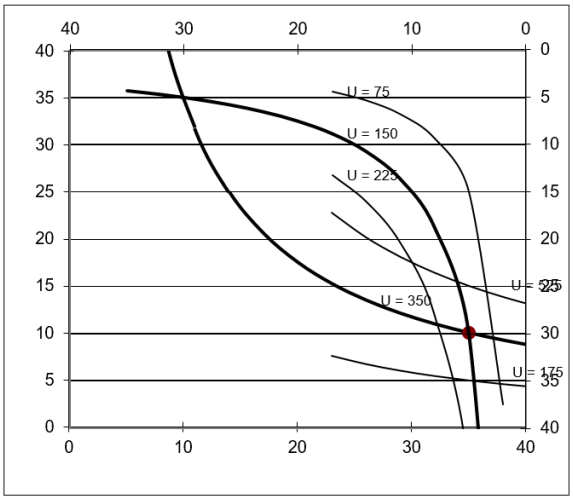


Figure 18.6: Edgeworth Box for Pareto criteria.
Source: EdgeworthBoxParetoOpt.xls!EdgeworthBox1.

From the initial endowment point in Figure 18.6, suppose we consider the combination of 30,15 for A and 10,25 for B.

STEP Click the (30,15) button. A red point appears at that coordinate in the box along with a text box.

Is A better off at the new point compared with the initial endowment? How about B?

As the text box explains, although the indifference curves for A and B are not drawn through the red point, we know they exist because the indifference map is dense—there is an indifference curve through every point in the box. If we draw an indifference curve for A through that point and it lies above the indifference curve that goes through the initial endowment, we know that A prefers 30,15 to the initial endowment.

In fact, indifference curves extend beyond the box in a northeast direction for A and southwest for B. The box just shows the total amounts available for exchange.

The same argument we made for A can be made for B. The only trick for B is to remember that you interpret the box from the top, right corner and B's satisfaction increases as the indifference curves move farther away from the northeast corner in a southwesterly direction.

Because both A and B are better off at 30,15 than the initial endowment, we know that the 30,15 allocation is *Pareto Superior* to the initial endowment. We can also flip the statement to say that the initial endowment is *Pareto Inferior* to point 30,15.

Pareto Superior means that it is possible to make at least one person better off without making anyone else worse off. We make no claims as to how much better off. We do not use the units of utility at all. This is similar to how we first discussed satisfaction in the Theory of Consumer Behavior. We asked consumers to simply choose between one bundle and another. The same logic is being used here.

Consider another point that is 30,10 for A and 10,30 for B.

STEP Click the (30,10) button.

As before, a red dot is placed on the chart and a text box appears. We want to compare the red dot to the initial endowment. Is A better off? How about B?

Because the point 30,10 is better for B, but worse for A, then this allocation is *Pareto Noncomparable* to the initial endowment because at least one person is made worse off. As soon as at least one person is made worse off, it is removed as a candidate for evaluation.

We certainly cannot evaluate these points by saying B's utility goes up by more than A's falls because utility is only ordinal. According to Pareto, we can never trade off a small decrease in satisfaction for one person for a large gain in satisfaction for one or many people because you cannot add up utility.

Now that we understand Pareto Superior and Pareto Noncomparable points, we can shade in all of the points that are Pareto Superior to the initial endowment. This is called the *lens* for reasons that will be obvious in a moment.

STEP Click the Show Lens button.

Every point in the space between and including the two indifference curves going through the initial endowment is shaded red, representing the area of Pareto Superior points. With usually shaped indifference curves, this is a lens-shaped object.

We return to the first point, 30,15. It is, of course, inside the lens so it is Pareto Superior to the initial endowment, but does it have any points that are Pareto Superior to it?

STEP Click the Hide Lens button and then the 30,15 button.

The 30,15 point, like the initial endowment, has a whole set of points that are Pareto Superior to it. These points also form a lens, albeit smaller that the lens formed by the Pareto Superior points to the initial endowment, that stretch from the point 30,15 to where the two indifference curves intersect again.

Clearly, whenever indifference curves from A and B cross at a point, such as the initial endowment or 30,15, we can find Pareto Superior points in a lens from that starting point. What happens when the indifference curves are tangent?

STEP Click the Hide Lens (if needed) and Show B Curve buttons.

A red dot is shown on an indifference curve for B that is tangent to A's highest displayed indifference curve. We will call this point of tangency between the indifference curves point PO1. This point PO1 is obviously Pareto Superior to the initial endowment since it is inside the lens.

But there is something special about PO1. It has a property that contains Pareto's key idea: Does PO1 have any Pareto Superior points to it? No, it does not. Movement in any direction from point PO1 lowers someone's satisfaction. There is no lens from point PO1.

Thus, we say that PO1 is a *Pareto Optimal* point—one that has no Pareto Superior points to it. You cannot make someone better off without hurting someone else. Pareto Optimal points are where we want to be!

It is important to note that there are an infinite number of Pareto Optimal points. Wherever the indifference curves are tangent, we are at a Pareto Optimal point.

The set of all Pareto Optimal points is called the *contract curve*. A minimalist version of a contract curve for an unknown (but well-behaved) pair of utility functions is displayed in Figure 18.7. A few indifference curves are displayed, but you should understand that every point on the contract curve is a point of tangency between two indifference curves. The sides of the box are not labeled, but you know how to read an Edgeworth Box.

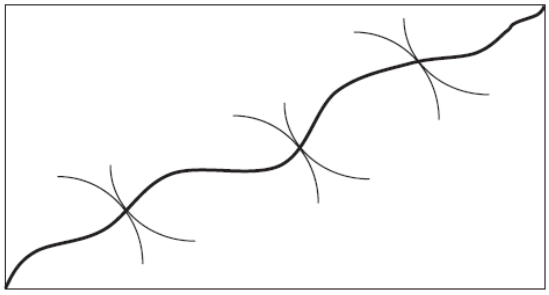


Figure 18.7: The contract curve.

Pareto Optimal points are especially desirable because they ensure that there is no way to improve the allocation without harming someone. In other words, given the limitations of ordinal utility, we can say that we have wrung out as much gain as possible if we are at a Pareto Optimal point. Thus, from any given initial endowment, OOSP would want to reallocate the two goods so that the allocation is on the contract curve.

One drawback of the Paretian framework is that there are many Pareto Optimal points when starting from an arbitrary, non-Pareto Optimal point. There is no way to choose between Pareto Optimal points.

Mathematically, it should be clear that Pareto Optimal points occur only when $MRS_A = MRS_B$. When this condition holds, the two indifference curves are tangent. This means we have a Pareto Optimal point and we are on the contract curve.

Pareto Optimality with Solver

One way to find Pareto Optimal points is to solve an optimization problem. It is not the silly, nonsensical "sum the utilities" objective function, however.

STEP From the *EdgeworthBox* sheet, open Solver.

Your Solver dialog box should look like Figure 18.8. Notice the $UtilityB = Initial\_UtilityB$ constraint. We are going to maximize A's utility without harming B. The constraint requires that B's utility be the same as the initial utility. Thus, B will be indifferent between the new allocation and the initial endowment.

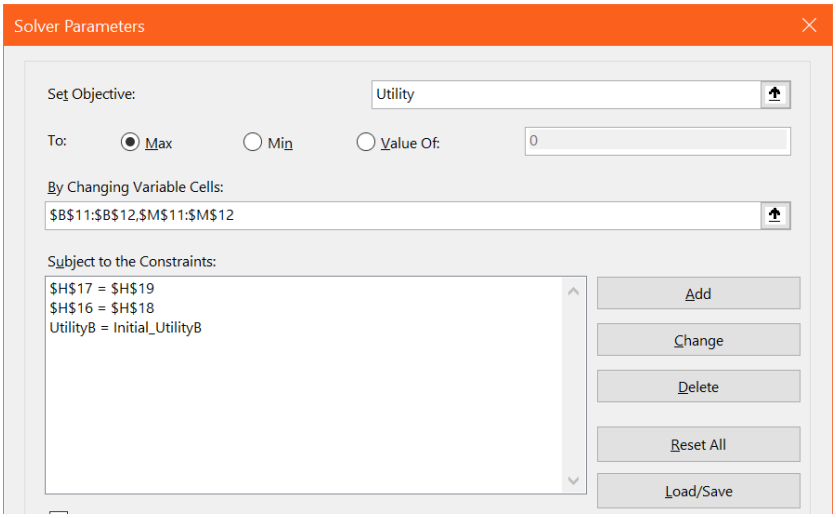


Figure 18.8: Solver parameters dialog box.

STEP Click Solve to find an optimal solution to this problem.

Scroll down (if needed) to see the Edgeworth Box. We are at the top most (from A's point of view) Pareto Optimal point. This point is on the contract curve.

What if we ran the same analysis, but maximized B's utility subject to maintaining A's utility constant? This is yet another Pareto Optimal point.

Some students want to make claims about points in the middle of the contract curve in the lens as being somehow better than the two extreme points, but the Pareto analysis does not allow for such distinctions.

The Contract Curve with Excel

STEP Proceed to the *ContractCurve* sheet.

It is set up just like the *EdgeworthBox* sheet, except A's Initial Endowment cells (B18 and B19) have a formula, =ROUND(randomnv()*38+1,0).

This formula allows you to generate random initial endowments, then you can use Excel's Solver to find a point on the contract curve from that initial endowment. You can use the "max A's utility keeping B's utility constant" or "max B's utility keeping A's utility constant" strategies. In the former case, you are finding the highest indifference curve of A that is tangent to B's indifference curve that goes through the initial endowment. You are doing the reverse when you maximize B's utility subject to A's indifference curve that goes through the initial endowment.

STEP Click the New Endowment button a few times to move the initial endowment point around the box. When you find one you like (it does not matter), find and record a point on the contract curve. Do this several times.

You are sampling points on the contract curve and this helps you learn how Pareto optimality works. Can you discover the shape of the contract curve?

STEP Change A's preferences by setting cA to 0.5. Sample points on the contract curve (using the same method as in the previous step). What effect does this have on the contract curve?

To see the answers to these two questions (but first try to answer them on your own), click the Show CC with cA = 0.5 button.

The First Fundamental Theorem of Welfare Economics

It is no exaggeration to say that we have reached the summit of this book. We are about to see the crowning achievement of economic theory—a demonstration of the welfare effects of the market system in a general equilibrium framework.

With the Pareto criteria in hand, we are ready to judge the market allocation. Recall that the market uses prices to establish an equilibrium solution. Surpluses and shortages push the price vector to and fro until it settles down to its equilibrium solution. What can we say about the market's solution?

We can say that it is Pareto Optimal! In fact, we can say that starting from any initial endowment, a market allocation mechanism yields a Pareto Optimal solution. This is the *First Fundamental Theorem of Welfare Economics*:

> If preferences are well-behaved, a properly functioning market's equilibrium solution is Pareto Optimal.

Figure 18.9 reproduces Figure 18.5 for your convenience. It is the canonical graph of general equilibrium analysis and shows the equilibrium solution from the Edgeworth-BoxGE.xls workbook. We know we have the equilibrium solution because there is a single, common tangency point. Consumer A maximizes by choosing that combination where he reaches the highest indifference curve subject to the constraint. Consumer B does the same.

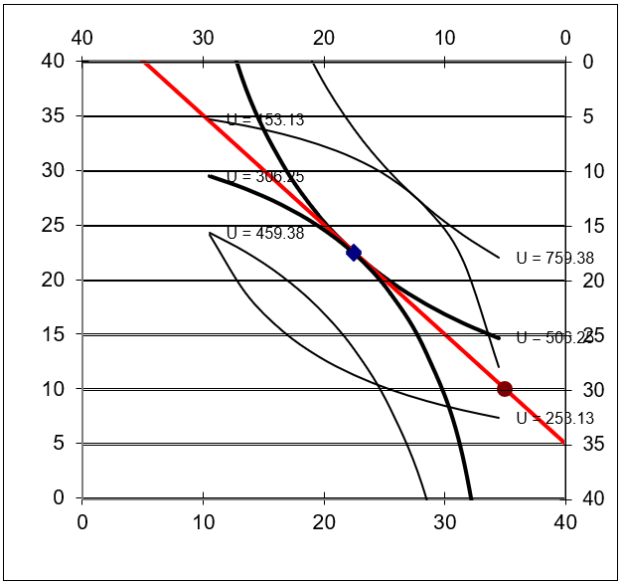


Figure 18.9: Evaluating the market allocation.
Source: EdgeworthBoxGE.xls!EdgeworthBox1 with $\frac{p_1}{p_2} = 1$.

But it is immediately obvious, given our work in this section, that the market allocation is Pareto Optimal. There are no Pareto Superior points to it.

We can use the equimarginal principle to help explain this result. Each consumer is finding a point of tangency that obeys the mathematical condition, $MRS = \frac{p_1}{p_2}$. From A's perspective, we have $MRS_A = \frac{p_1}{p_2}$. Similarly, B chooses that combination where $MRS_B = \frac{p_1}{p_2}$. Unbeknownst to them, they are ending up at a point where $MRS_A = MRS_B$.

In other words, by paying attention to prices and optimizing, the equilibrium generated by exchanging consumers is at the same time generating a Pareto Optimal solution. There is an invisible hand aspect to this in the sense that the consumers do not know and do not care about Pareto Optimality.

Geese fly in a V pattern over thousands of miles by drafting—wind resistance is minimized by aligning one-self at angle to the goose ahead, instead of flying directly behind or next to a fellow goose. The geese are completely unaware that they are generating a V-shaped pattern. Consumers in a market are just like geese—they are completely unaware that they are solving a much bigger optimization problem.

Geese also synchronize their wing beats because they take advantage of updraft. If you watch a flock, it looks like they are coordinating their flapping. This was discovered recently (see Portugal, et al., 2014) and provides an excellent example of how economists see the market system.

With each agent following a simple rule, the system produces a pattern. In the case of the market, it is an incredible result that the market allocation is Pareto Optimal.

What can't we say about the market allocation?

We certainly can't say that it is *fair*. The market will grind to a Pareto Optimal point from any initial endowment. The Pareto logic takes the initial endowment as given. What if A starts out with much more than B? What if the market does not value B's resources? The Pareto criteria have nothing to say about this. Economists have tried to include fairness in welfare analysis, but there is little consensus.

If there's a First Theorem, there must be a Second Theorem, right?

If preferences are well behaved, a properly functioning market can reach any Pareto Optimal point if the appropriate initial endowment is provided.

The Second Fundamental Theorem says that you can use the market to reach any Pareto Optimal allocation—that is, any point on the contract curve. All you have to do is set the initial endowment appropriately, then let the market work its magic.

The last two problems in the Q&A sheet ask you to show that the Second Fundamental Theorem works.

That Markets Generate Pareto Optimal Solutions Is a Truly Fundamental Idea

This section marks the end of a long trek. We began with the Theory of Consumer Behavior and learned that consumers maximize satisfaction subject to a budget constraint. An important extension of this basic model utilizes an initial endowment instead of cash income.

In a Pure Exchange Model, we combine two optimizing consumers in an Edgeworth Box. Their interaction results in an equilibrium solution.

Using the Pareto criteria, we can compare allocations and determine which ones are Pareto Optimal. These are allocations that have no Pareto Superior points. The set of all Pareto Optimal points forms the contract curve.

Students struggle with the term Pareto optimality. Its definition, that there is no way to make someone better off without hurting someone else, can become a jumble of words with little real meaning. Here is the crucial idea: Pareto Optimality means no waste. The allocation at a Pareto optimal point cannot be improved upon (without harming someone). Thus, Pareto optimality means we have an unbeatable allocation.

The First Fundamental Theorem of Welfare Economics makes a powerful statement because it says that a properly functioning market yields a Pareto Optimal allocation. This is a highly desirable result.

It is also shocking because individual consumers have no idea they are participating in solving a resource allocation problem. Each consumer is simply maximizing utility subject to a budget constraint. Like geese that fly in a V, each consumer is responding to a signal (in the consumer's case, prices) and then the interaction is producing the coordination.

Notice that the work here has said nothing about innovation or technological change. In fact, the analysis assumes constant technology and no new products. The analysis is completely static and based solely on the market's ability to reach a Pareto Optimal solution in terms of allocating already produced goods in a pure exchange economy.

You might be wondering if all equilibria in an Edgeworth Box are Pareto Optimal? Absolutely not. The next section shows how the market can fail.

Exercises

1. Why do the Pareto criteria fail to provide a single point that is the best allocation?

2. What must be true about the exponents in the Cobb-Douglas utility functions for consumers A and B to generate a linear contract curve? Describe your procedure and explain your answer.

3. Use Word's Drawing Tools to draw an Edgeworth Box with well-behaved preferences and a point Z, where the $MRS_A > MRS_B$. Explain why point Z is not Pareto Optimal.

4. The contract curve (with $cA = 0.5$) can be transformed into a utility possibilities frontier, as shown in Figure 18.10. Where would point Z (from the previous question) be on this graph? Explain why.

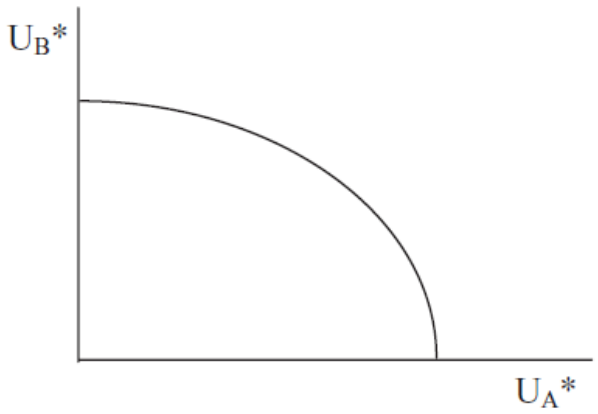


Figure 18.10: A utility possibilities frontier.

References

The epigraph is from page 36 of Vilfredo Pareto's *The Rise and Fall of Elites: An Application of Theoretical Sociology* (originally published in Italian in 1901, translated to English in 1968, and published as a paperback in 1991 and 2000). The back jacket says,

> Here in brief and incisive outline are the major ideas for which Pareto was later to become famous ... This slim volume is more readable and disciplined than most of the later elaborations, and serves well as an introduction to Pareto's political sociology ... Pareto's irony shows in his attack on elites that become humanitarian and tender-hearted rather than tough-minded.

Most economists know Pareto through his work on utility, General Equilibrium Theory, and the idea of Pareto Optimality, but Pareto grew disenchanted with "pure economics" (what we would call today economic theory) and turned to sociology. His most famous sociological work is *Mind and Society* (originally published in 1916 and first translated into English in 1935), in which he explains how the circulation of elites drives history.

See Vincent J. Tarascio, *Pareto's Methodological Approach to Economics* (published in 1968) for a comparison of Pareto's views on the scope and method of economics, especially as contrasted with Alfred Marshall. Whereas Marshall saw mathematics as a language, capable of being translated so nonmathematicians could understand, Pareto believed that "mathematics makes it possible to express relations between facts which are not possible with other facilities or ordinary language" (Tarascio, p. 106, footnote omitted). Pareto saw no need to translate heavily mathematical papers for the "literary economists." Many of Pareto's ideas on optimization and equilibrium were presented in prose form by Philip H. Wicksteed, *Common Sense of Political Economy* (first published in 1910) and available online at www.econlib.org/library/Wicksteed/wkCS.html.

On geese flying in a V and coordinating flapping, see Steven J. Portugal, Tatjana Y. Hubel, Johannes Fritz, Stefanie Heese, Daniela Trobe, Bernhard Voelkl, Stephen Hailes, Alan M. Wilson and James R. Usherwood (2014), "Upwash exploitation and downwash avoidance by flap phasing in ibis formation flight," *Nature*, 505, pp. 399–402, www.nature.com/articles/nature12939. My video, *The Invisible Hand and the Market System*, freely available at vimeo.com/econexcel/invisiblehand, has a clip of the authors explaining how the geese do it.

1) Let E be an economy such that, for every i,
(a) X_i is convex,
(b) if x_i^1 and x_i^2 are two points of X_i and if t
is a real number in $]0,1[$, then
$x_i^2 \succ x_i^1$ implies $tx_i^2 + (1-t)x_i^1 \succ x_i^1$.
An equilibrium $((x_i^*), (x_j^*))$ relative to a price system p, where no x_i^* is a satiation consumption, is an optimum.

Gerard Debreu

18.4 General Equilibrium Monopoly

Partial equilibrium analysis tells us that monopoly causes an inefficient allocation of resources—too little output (compared with the socially optimal level) is produced.

This section explores the welfare implications of monopoly in a general equilibrium setting. The procedure is the same as the one used for judging competitive markets: We determine the monopoly allocation and then test it by comparing it to the set of Pareto Optimal points (i.e., the contract curve).

To reiterate, monopoly results in an inefficient allocation of resources. There is no dispute about that. However, General Equilibrium Theory is the best way to demonstrate this inefficiency.

Monopoly in an Edgeworth Box

Suppose we start with the usual Edgeworth Box. It has an initial endowment that is the point of departure for trade between the two consumers.

Competitive markets are modeled in an Edgeworth Box by supposing that prices are determined by the interaction of many buyers and sellers. To implement price-taking behavior in a two-person Edgeworth Box, we use an auctioneer who calls out prices. Each consumer determines optimal amounts to buy and sell based on the given prices. The Edgeworth Box is used to check whether the amounts that each consumer wants to buy and sell are compatible. If not, prices adjust based on the shortages and surpluses generated by the plans of each consumer.

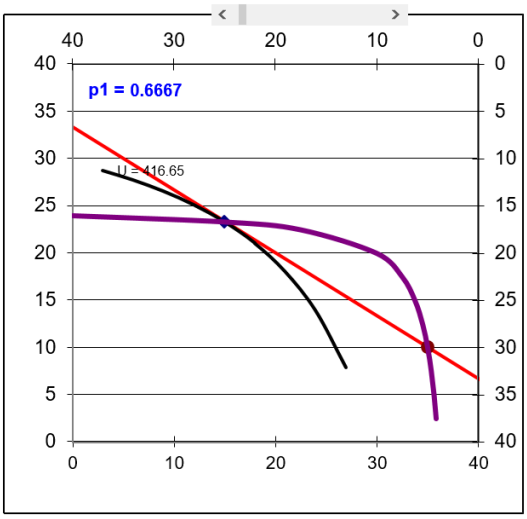
We model monopoly in a pure exchange Edgeworth Box by eliminating the auctioneer. We give one of the consumers monopoly power. They can set the price vector to have any slope.

Suppose that A is a monopolist. What does this mean in the context of the Edgeworth Box? A will quote prices to B and let B decide how much to buy and sell. A will choose a price ratio and this determines the final allocation.

We can think of A as an auctioneer who first shouts out prices to see how B will respond, then picks the best prices—from A's point of view.

STEP Open the Excel workbook *EdgeworthBoxMonopoly.xls*, read the *Intro* sheet, then go to the *PriceOfferCurveB* sheet.

Figure 18.11 (and your screen) shows B's *price offer curve*, which tells A how much x_1 and x_2 B wishes to hold given the price ratio, $\frac{p_1}{p_2}$.



B's Price Offer Curve

p1/p2	x1B	x2B
0.4	40.00	16.00
0.6667	25.00	16.67
1	17.50	17.50
2	10.00	20.00
3	7.50	22.50
4	6.25	25.00
5	5.50	27.50
6	5.00	30.00
7	4.64	32.50
8	4.38	35.00
9	4.17	37.50

Figure 18.11: B's offer curve.
Source: EdgeworthBoxMonopoly.xls!PriceOfferCurveB.

Initially, A has set $p_1 = 0.6667$ (p_2 is the numeraire). B maximizes utility, given that price ratio, by choosing the combination 25,16.67. This is shown by the black indifference curve that is tangent to the red price vector. B will want to buy 20 units of good 1 and offer (hence the name offer curve) 13.33 units of good 2 for sale to A.

A can set any price for good 1 she wishes, but B gets to decide how much to buy and sell at A's chosen price. Also, we assume A will honor the deal and buy the amount B wants to sell.

STEP Click the scroll bar above the graph a few times to change the price of good 1.

With each click, the red budget constraint line rotates about the initial endowment and B chooses a new optimal bundle.

The locus of points that B chooses as p_1 is varied, ceteris paribus, is the *price offer curve*. For any given price, B finds the place at which the highest indifference curve is tangent to the budget constraint—and this point is on the price offer curve.

Having explained B's price offer curve, we bring A into the picture. A knows B's price offer curve and has the monopoly power to set any price for x_1. Given $p_2 = 1$, A has the power to set the slope of the price vector. The key question is: Which price will A choose?

In one sense, the answer is obvious: Choose p_1 that maximizes satisfaction for A. But how can this problem be solved so we find the best price from A's point of view?

STEP Proceed to the *EdgeworthBox* sheet.

The display is the same as on the *PriceOfferCurveB* sheet, except that now we have added A's indifference curves. We also can easily see A's utility in cell C28.

Is the initial price of 0.6667 the best solution for A? No, because by increasing p_1, A gets greater satisfaction.

STEP Confirm that this is true by clicking on the scroll bar to increase p_1 and keeping your eye on A's utility in cell C28.

You can also control the price with the scroll bar over cells A9 and B9. Notice how the price has been moved under the heading of *Endogenous Variables*. Because A chooses the price—this is what monopoly power means—price is endogenous to the monopolist.

In the *Wealth of Nations*, Adam Smith says, "The price of monopoly is upon every occasion the highest which can be got" (Book I, Chapter VII, www.econlib.org/library/Smith/smWN.html?chapter_num=10#book-reader).

But is this true? Would the monopolist literally charge the highest price possible?

STEP Drag the scroll box in the scroll bar all the way to the right.

The chart is hard to read, but we can see from the table next to the chart that with $p_1 = 9$ (the highest price we can set with the scroll bar), B wants to end up with 4.17 units of x_1 and 37.5 units of x_2. This means p_1 is so high that B does not want to buy any of it and, in fact, wants to sell 0.83 units to A!

More importantly, a quick glance at cell C28 reveals that A's utility is under 90. This means that, taken literally, a monopolist will not charge the highest price possible.

Just like a monopoly in a partial equilibrium setting, A is operating under a constraint. A monopolist takes the demand curve as given. Consumer A takes B's offer curve as given and B's offer curve acts as constraint for A.

With this knowledge, can you solve A's problem? What is A's optimal p_1?

STEP Use the scroll bar to manipulate p_1. Keep an eye on A's utility. Can you find the value of p_1 that maximizes A's utility?

You cannot beat $p_1 = 2$. This is the optimal solution. This is what A will charge B for x_1. At this price for good 1, B wants to have 10 and 20 units of goods 1 and 2. B will buy 5 units of x_1 (adding this to the initial endowment of 5 units) financed from the sale 10 units of x_2. A ends up with 30 and 20 units of goods 1 and 2. A sells 5 of her initial endowment of 35 units of x_1 for \$2/unit and buys 10 units of good 2. The plans match and we are at a stable position.

You can also find this answer with Solver.

STEP Click the scroll bar so p_1 is not equal to 2 and run Solver.

Notice that the changing cell is B9, which is the cell connected to the scroll bar. Solver does not need a constraint because the sheet is set up so that B optimizes based on p_1 and then A's x_1 and x_2 are the total units available for each good minus B's optimal decision. Thus, B's offer curve has been included in A's optimization problem.

In addition, you could use analytical methods, using A's utility as the objective function and B's offer curve as the constraint. All of these methods give the same answer—A's utility maximizing p_1 is 2.

The monopoly solution is displayed in Figure 18.12. Notice that A's indifference curve is tangent to B's offer curve. This is how a monopolist maximizes utility.

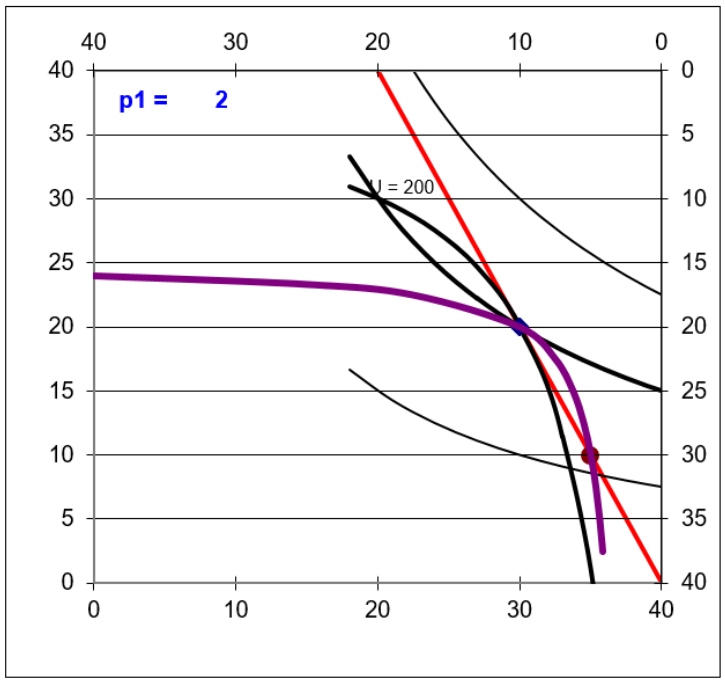


Figure 18.12: Monopoly's optimal solution.
Source: EdgeworthBoxMonopoly.xls!EdgeworthBox with $p_1 = 2$.

Judging Monopoly

What can we say about the monopoly allocation? With Pareto's criteria we can instantly proclaim: Monopoly is not Pareto Optimal.

Figure 18.12 shows that the monopoly allocation is at a point (from A's view it is coordinate 30,20) where the $MRS_A \neq MRS_B$ because the indifference curves intersect. This means that there are Pareto Superior points to the monopoly allocation. It also means that the monopoly allocation is not on the contract curve.

By moving northwest, into the lens created by the two indifference curves at the monopoly solution, an omniscient, omnipotent social planner could make both A and B better off.

Why doesn't A do this? Because all A can do is set the price of good 1 and with this monopoly power, she must charge the same price for all the units sold. This leads to the allocation in Figure 18.12.

If A could perfectly price discriminate, charging different prices for different units, we would get a different result. A could sell the first unit of x_1 at a high price and decrease the price as B purchased more units. As explained in the chapter on monopoly in a partial equilibrium setting, this is called perfect price discrimination. The Q&A sheet asks you to work out the welfare implications of this type of monopoly in a general equilibrium analysis. The welfare results for perfect price discrimination in partial and general equilibrium are the same.

Unlike partial equilibrium, we report no deadweight loss measure in this pure exchange, general equilibrium analysis. We simply note that the monopoly allocation is not Pareto Optimal and this is enough to doom monopoly because we know there are Pareto Superior allocations to the monopoly result.

We cannot say how much damage the inefficiency of monopoly causes because utility can only be measured ordinally. We cannot express, in utils or dollars, the wasted value from monopoly, but we know it is there. Once we say that there are Pareto Superior points, we stamp monopoly as a poor allocation mechanism.

Monopoly is not Pareto Optimal

We found, as we did with partial equilibrium analysis, that monopoly is inefficient. This time, however, we used a general equilibrium analysis that adhered to the strict limitations imposed by ordinal utility. Thus, this analysis is theoretically sound.

In a pure exchange Edgeworth Box, if one agent is granted monopoly power, he or she will choose a price to maximize his or her utility. This does not generate a Pareto Optimal allocation. The monopolist is not interested in Pareto optimality—she simply wants to maximize her own utility.

Recall, however, that this is simply a pure exchange economy. A true general equilibrium model must include production of goods and services and then combine production and exchange. This is beyond the scope of this book. The monopoly result stays the same; however, it still fails to yield a Pareto Optimal allocation.

Exercises

1. Is the monopoly solution better than the initial endowment? Explain.

   Hint: Use Figure 18.12 as a reference.

2. Suppose A really liked x_1, so that cA (cell B21) was 2. How would this change A's utility maximizing price of x_1? What is the monopoly solution? Describe your procedure.

3. In the previous chapter, we used a supply and demand (partial equilibrium) analysis to show that price ceilings in a competitive market cause an inefficient allocation of resources. Use Word's Drawing Tools to create an Edgeworth Box with a price ceiling on x_1. Explain why price ceilings are undesirable in this general equilibrium setting.

References

The epigraph is from page 94 of Gerard Debreu's *Theory of Value: An Axiomatic Analysis of Economic Equilibrium* (originally published in 1959). Debreu won the Nobel Prize in Economics in 1983 "for having incorporated new analytical methods into economic theory and for his rigorous reformulation of the theory of general equilibrium" (www.nobelprize.org/prizes/economic-sciences/1983/summary).

The Nobel Prize web site explains Debreu's contribution in more detail, of course, but for the real scoop, consider this excerpt from E. Roy Weintraub's *How Economics Became a Mathematical Science* (published in 2002):

> While it was the case that most economists would have been unfamiliar at that time with the novel tools of set theory, fixed point theorems, and partial preorderings, there was something else that would have taken them by surprise: a certain take-no-prisoners attitude when it came to specifying the "economic" content of the exercise. Although there had been quantum leaps

> of mathematical sophistication before in the history of economics, there had never been anything like this (p. 114).

Weintraub reports that he had better luck interviewing Debreu than did George Feiwel, who prefaced many of his questions with, "For the benefit of the uneducated." When Feiwel asked why existence of an equilibrium solution is so important, "Debreu shot back, 'Since I have not seen your question discussed in the terms I would like to use, I will not give you a concise answer'" (Weintraub, p. 113). In addition to providing an entire transcript of the interview, Weintraub explains how Debreu led a wave of mathematical formalism into economics in the 1950s.

The general equilibrium ideas you have encountered in this book are a mathematical step below the more formal, axiomatic exposition of General Equilibrium Theory developed in the 1950s and used in graduate economics courses. Pick up Debreu's *Theory of Value* or a modern, PhD-level Micro Theory text (such as David M. Kreps, *A Course in Microeconomic Theory*, 1990) to see exactly what a formal, axiomatic exposition of general equilibrium entails.

Part IV

Conclusion

> But when time and the means for achieving ends are limited and capable of alternative application, and the ends are capable of being distinguished in order of importance, then behaviour necessarily assumes the form of choice. Every act which involves time and scarce means for the achievement of one end involves the relinquishment of their use for the achievement of another. It has an economic aspect . . . Here, then, is the unity of the subject of Economic Science, the forms assumed by human behaviour in disposing of scarce means.
>
> Lionel Robbins

Conclusion

Throughout this book, Excel has been used to solve optimization problems and equilibrium models. Repeated emphasis has been placed on comparative statics and elasticity.

This conclusion has three parts:

1. Excel's Solver: There is a review of basic Solver skills with emphasis on the lesson that Solver is not perfect.
2. Overall view: A quick tour of the topics covered enables a clear statement of the economic way of thinking.
3. An open problem: Markets in a static framework are well understood, but the economic growth generated over time by capitalism is not.

1. Excel's Solver

Consider a perfectly competitive (PC) firm with a total cost function given by $TC = 100q^{\frac{1}{2}}$. Dividing both sides by q gives us the average cost function, $ATC = 100q^{-\frac{1}{2}}$. Taking the derivative of TC with respect to q yields $MC = 50q^{-\frac{1}{2}}$.

If this PC firm faced a market price of \$5/unit, what is the profit-maximizing level of output?

This book has solved optimization problems via numerical and analytical methods. We will apply both methods to this problem. First, we will use Solver.

But we will not use a prepared Excel workbook. Instead, you will create your own implementation of this problem. There are, of course, helpful steps to guide you.

STEP Open a blank Excel workbook. In cell A1, type the word *quantity*. Cell B1 will hold a number that represents the quantity. In cell A2, type the word *profits*. In cell B2, enter the formula for profit.

The price is \$5/unit and $TC = 100q$ so the formula in cell B2 is: = 5*B1 − 100*SQRT(B1).

STEP Run Solver. The target cell is B2, the goal is obviously to maximize profits, and the changing cell is B1. There are no constraints because the PC firm is free to produce as much output as it wants at the given price.

Excel gives a miserable result. Depending on your Solver defaults, it might go negative and, since Excel cannot take the square root of a negative number, it gives up and announces its failure.

If so, make A1 zero and run Solver again, but this time, check the *Make Unconstrained Variables Non-negative* option. Your Solver may be set up so the *Make Unconstrained Variables Non-negative* option was already checked so you might not see the first miserable result.

Starting from zero (or a blank cell) in A1, with the non-negativity constraint, Solver says the answer is zero. This is worrisome. Could the optimal quantity really be zero?

Maybe the issue is that we are starting from blank cell, which is zero. This is poor practice. Excel interprets blanks as a zero and the formula in B1 evaluates to zero. Treating blanks as zero is one of the most dangerous things a spreadsheet does (Google sheets behaves the same way). You should always avoid this.

We can change where Solver starts from to see if that helps.

STEP Change cell B1 to 25. Cell B2 should display −375. Run Solver.

Solver appears convinced that the optimal solution is zero. We turn to analytical methods to see if we can confirm Solver's result.

We know $MC = 50q^{-\frac{1}{2}}$ and since it is a PC firm $MR = P$ so $MR = 5$. We can set $MR = MC$ and solve for optimal q.

$$5 = 50q^{-\frac{1}{2}} \rightarrow q^{\frac{1}{2}} = 10 \rightarrow q^* = 100$$

This is confusing. We now have two answers: $q = 0$ and $q = 100$. Which one is correct?

Maybe a graph will help. We can draw the canonical graph of the firm's output profit maximization problem. Figure IV.1 shows the cost curves and we can clearly see that $MR = MC$ yields a negative profit rectangle.

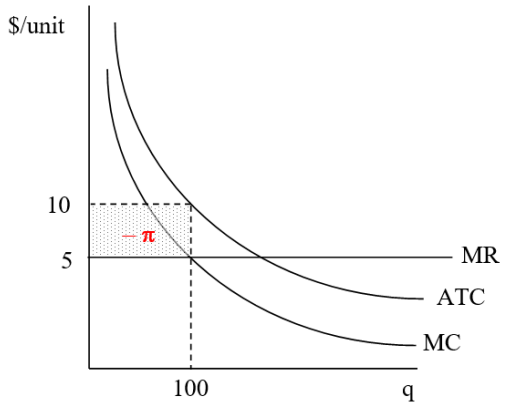


Figure IV.1: The firm at $q = 100$ where $MR = MC$.

This graph helps explain what is going on here, but we need a better visual. This book claimed that looking directly at the profit function made clear the Shutdown Rule so let's try that approach.

STEP Create a column from 0 to 500 by 10. This is the quantity. Use the profit formula to create a column for profit based on the quantity. Create a graph of the two columns.

If you get stuck, this 2-minute video at vimeo.com/425873093 shows how to do it.

Figure IV.2 shows the graph made in the video. It makes clear that the point where $MR = MC$ is actually a point of minimum profit. Although the first-order condition is met (we did find a flat spot on the profit function at $q = 100$), this solution fails the second-order condition for a maximum.

Thus, the correct answer is to produce an infinity of output. Profits rise as more is produced past 100 units of output. Higher output leading to greater profit continues forever so the optimal solution is infinity.

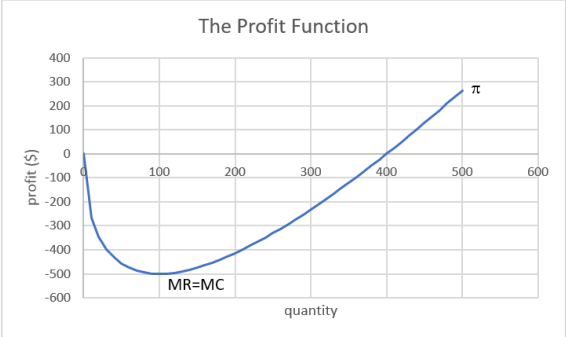


Figure IV.2: The profit function shows that optimal q is infinity.

How can we explain Solver's answer of zero? Why doesn't it give us the correct answer? When Solver starts from below 100 (we started from zero and 25), it goes to zero (or negative output if you do not have a non-negativity constraint). What happens if it starts from a number greater than 100?

STEP Enter 110 in cell A1 and run Solver.

Solver reports that "Objective cells do not converge." Is this a miserable result? No, actually, it is the correct answer! When Solver starts from more than 100, it goes right on the x axis and profits rise and it keeps going and going. As we know, this is the right answer.

It is worth remembering that Solver's algorithm is naive. It evaluates the function at the starting value, then moves left and right. The size of the move depends on the numerical values in the problem. Starting from $q = 25$, for example, Solver moves a little bit right, sees that profits fell, then goes in the opposite direction and lowers output. You can see Solver's steps by checking the *Show Iteration Results* option after clicking the *Options* button in the Solver dialog box.

You might be thinking that since we are in the long run, $ATC = AVC$ and it is clear that $P < AVC$ at $MR = MC$, which means the firm should shut down. That is not bad thinking, except the rule does not work at $MR = MC$ in this case because that is not the profit-maximizing output.

The takeaway of this final example is that you have to know what you are doing with Solver. It is not perfect and you cannot blindly rely on its results. This example shows that numerical methods are to be used with caution. Be careful out there.

2. Overall View

This book covered modern-day, orthodox microeconomic theory at the college undergraduate level. It used Excel to present difficult material and showed how mathematics can be used to solve problems in economics.

The economic approach or economic way of thinking provided the framework for analyzing observed behavior. The basic idea is to set up and solve an optimization problem or equilibrium model. Next, a single variable is changed, ceteris paribus, and the new solution is compared to the initial solution. This procedure is called comparative statics. Elasticity captures the logic of comparative statics in a single number.

When the economic approach is applied to consumers, it is called the Theory of Consumer Behavior. The key comparative statics analysis is deriving the demand curve. Figure IV.3 is a canonical graph of deriving demand.

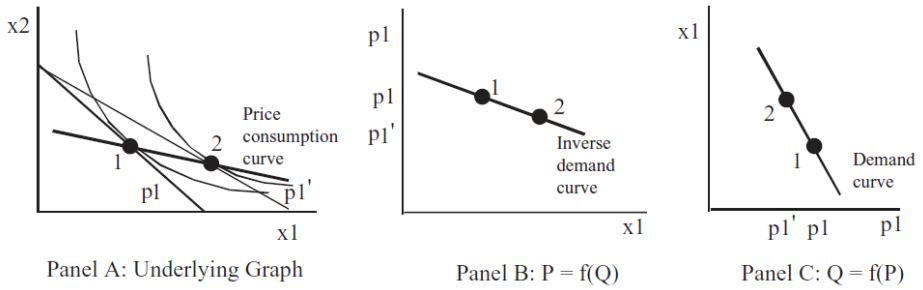


Figure IV.3: Deriving the demand curve.

When the economic approach is applied to producers, it is called the Theory of the Firm. The key comparative statics analysis is deriving the supply curve. Figure IV.4 is a canonical graph of deriving supply.

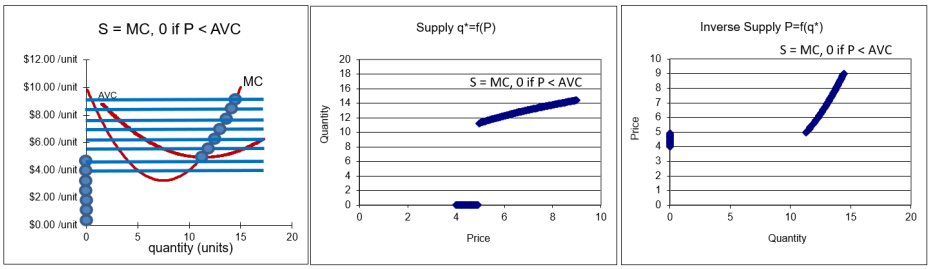


Figure IV.4: Deriving the supply curve.

The firm is more complicated than the consumer because firms hire inputs to produce output. In fact, the firm is really a set of three interrelated optimization problems: input cost minimization, output profit maximization, and input profit maximization.

The individual demand and supply curves derived from the consumer and firm models can be added up to produce market demand and supply curves. This enables a partial equilibrium analysis of how markets solve society's resource allocation question. Figure IV.5 shows supply and demand flanked by its consumer and firm source graphs.

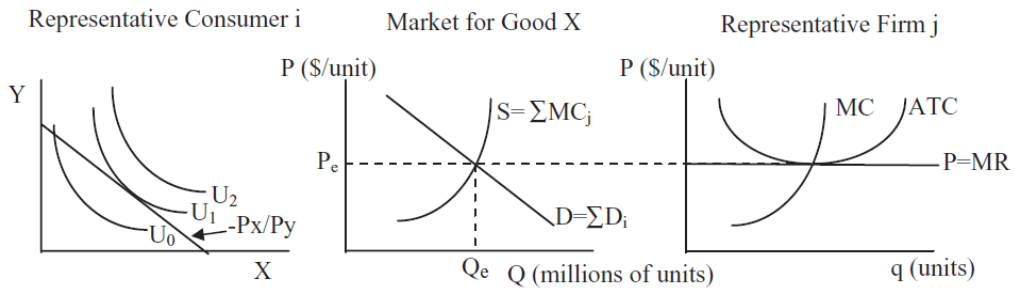


Figure IV.5: The market's resource allocation solution for one good.

Price ceilings, taxes, monopoly, import quotas, and externalities are all examples of situations where we have a misallocation of resources in a single market.

Partial equilibrium enables calculation of a measure of inefficiency called deadweight loss (also known as the Harberger triangle), but this should be interpreted as an approximation because consumers' surplus requires that an adjustment be made to the ordinary demand curve (compensated demand must be used) and the effects on other markets are ignored. Partial equilibrium analysis is commonly used in empirical work. Think of deadweight loss as a rough measure of inefficiency in the allocation of resources.

General equilibrium is a more rigorous and sophisticated analysis because it looks at all markets as a total system. Pareto's criteria show that a properly functioning market yields an optimal allocation and monopoly is not Pareto optimal. Figure IV.6 is the canonical graph of a market's general equilibrium and it makes clear that the market's allocation has no Pareto Superior points.

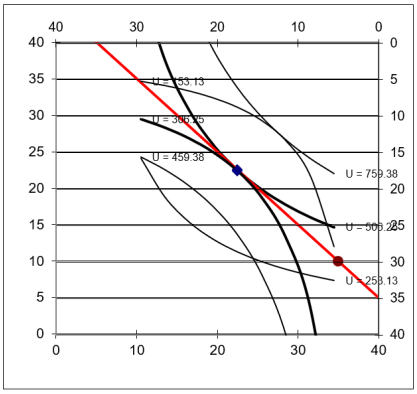


Figure IV.6: The market's allocation in an Edgeworth Box.

General equilibrium does not suffer from the same problems as partial equilibrium, but it is much harder to implement in the real world. In the epigraph to the section introducing the Edgeworth Box, mention was made of computable general equilibrium models. This shows that there is an empirical side to general equilibrium analysis, but it is a relatively modern development.

It is reasonable to view mainstream microeconomics as a theory of the price mechanism. The market system uses prices as signals to allocate resources. Optimizing agents react to price changes and their interactions as buyers and sellers drive the system toward equilibrium. The Theories of Consumer Behavior and the Firm are stepping stones that explain how the market answers society's resource allocation question. Figure IV.5 puts the Theory of Consumer Behavior, Theory of the Firm, and partial equilibrium analysis together. These three graphs and how they fit together are worth remembering.

Another organization of microeconomics splits it into two parts—individual agents (consumers and firms) that optimize and what happens when these optimizing agents interact in a market. The former is about optimization and the latter is about equilibrium. The order that is spontaneously generated by interacting, optimizing agents is a remarkable result. Economists see supply and demand not as the simple intersection of two lines, but as a pattern that is unwittingly generated by the agents themselves—just like geese that fly in a V.

This book was designed to provide you with practice in applying the economic approach. We tackled unconstrained and constrained optimization

problems, computed many different elasticities, and solved several equilibrium models at the partial and general levels.

The many applications of the economic approach demonstrate its remarkable flexibility. The Theory of Consumer Behavior, at first, seems ridiculously unrealistic—a robot consumer chooses between two goods with prices, tastes, and income given! But that is just the basic model. By changing the goods to consumption in the present and the future, it becomes an intertemporal choice model. We analyzed charitable giving, portfolio theory, and the effect of safety features in automobiles with the Theory of Consumer Behavior.

In every application, the economic way of thinking was prominent. We set up and solved an optimization problem, then changed a variable, ceteris paribus, to see how the optimal solution changed. There are countless applications of the economic approach, but they share the same framework and logic.

In fact, the economic approach is what defines economics today. It may be the only discipline that defines itself by a methodology instead of by what it studies. Most people have a content-based definition of economics: They think that the study of interest rates, unemployment, and money is economics. But this is wrong. The proper definition of economics is the application of the economic approach to explain observed behavior. Crime, marriage, and war, if analyzed with the economic approach, fall under the heading of economics.

From now on, when you hear the phrase "an economic analysis of," you will know that the economic approach is about to be applied, you will know what to expect, and you will be comfortable as the speaker talks about constraints, optimality, comparative statics, and elasticity.

3. An Open Problem

Neither this book nor modern, mainstream economics explains the dynamic process of capitalism. A few hundred years of the market system make it obvious that creativity, innovation, and technological change are endogenously generated by market-based societies. No one really knows why.

The question has been with economics since the very beginning. Many people know that Adam Smith wrote a book called the *Wealth of Nations*, but only a few know that the actual title is, *An Inquiry into the Nature of Causes of the Wealth of Nations*. But what was Smith's inquiry, simply put?

He wanted to know why England was so much richer than its neighbors. In 1776, Smith could see British wealth all around him. He could see the economy taking off and he wondered why some places develop and grow, while others cannot seem to do so? This question remains unanswered and, in the language of mathematics, it is the biggest open problem in economics.

Explaining the dynamism of the market system is a much different question than the static optimization and equilibrium models that explain why markets allocate resources efficiently. In the static world, there are no new products, cost-saving innovations, or new firms. The static world is stable and markets are in equilibrium.

This static model clashes violently with reality. Joseph Schumpeter's portrayal of what he called plausible (i.e., real-world) capitalism, captured in the oxymoron "creative destruction," highlights the rise and fall of firms, explosive growth, and dislocation produced by markets. For Schumpeter, the driving force is the entrepreneur, a hero whose desire to dominate the business world results in economic success for society. But Schumpeter's story (best captured in *Capitalism, Socialism and Democracy*, originally published in 1942), thrilling though it may be, is not part of mainstream economics today.

It is plainly clear that markets do generate spectacular economic growth, unparalleled by any other organizational form. Even the harshest critics of capitalism concede this point:

> The bourgeoisie, during its rule of scarce one hundred years, has created more massive and more colossal productive forces than have all preceding generations together. Subjection of Nature's forces to man, machinery, application of chemistry to industry and agriculture, steam-navigation, railways, electric telegraphs, clearing of whole continents for cultivation, canalisation of rivers, whole populations con-jured out of the ground—what earlier century had even a presentiment that such productive forces slumbered in the lap of social labour?

That was written by Karl Marx and Friedrich Engels in *The Communist Manifesto* in 1848, available at www.marxists.org/archive/marx/works/download/pdf/Manifesto.pdf.

Marx and Engels argued capitalism will self-destruct, but not because it failed to make goods and services. They thought it was the most productive system ever devised. They were amazed by capitalism's ability to generate output.

Marx and Engels were not the first nor the last to be awed by the productive power of the market system. Yet, even though we can easily see that productive power, we simply do not know the answer to basic questions about how markets generate growth. Beyond superficial generalities about the institutional environment, such as needing rule of law and established property rights, we have no explanation for how the interaction of multitudes of agents drives the system over time. We cannot even answer the most basic question, posed by Adam Smith—why are some countries rich and others poor?

If we knew how and why markets caused technological change and output per person to grow exponentially, we would know how to help those societies mired in poverty. Nobel Prize winning economist Robert Lucas poses the issue this way:

> Is there some action a government of India could take that would lead India's economy to grow like Indonesia's or Egypt's? If so, what, exactly? If not, what is it about 'the nature of India' that makes it so? The consequences for human welfare involved in questions like these are simply staggering: Once one starts to think about them, it is hard to think about anything else. (Lucas, 1988, p. 5)

The point is this: Markets can be analyzed from static and dynamic perspectives. The former focuses on resource allocation at a single moment in time. It freezes the movie and asks how markets work in this motionless environment. We know how markets work as a resource allocation mechanism.

The latter perspective is about the dynamic nature of markets, we want to know how markets work over time. The movie runs—spurts of rapid growth are followed by recessions, then more growth, but output per person trends upward. Will this continue? We do not know. How do the institutions we rely on (including property rights) emerge from the interaction of optimizing agents? We do not know.

Explaining markets as a dynamic process remains the most important open problem in economics. Perhaps you can work on it.

References

The epigraph is from pages 14 and 15 of Lionel Robbins, *An Essay on the Nature and Significance of Economic Science* (originally published in 1932) and available online at mises.org/library/essay-nature-and-significance-economic-science.

We began with a famous quotation from Robbins, defining economics as "the science which studies human behavior as a relationship between given ends and scarce means which have alternative uses." This book takes this definition seriously and has stressed static optimization, but this last chapter makes clear that we have a great deal to discover and learn about dynamics and technological progress.

Robert Lucas, "On the Mechanics of Economic Development," *Journal of Monetary Economics*, Vol. 22 (1988), pp. 3–42, www.sciencedirect.com/science/article/abs/pii/0304393288901687.

The fact that perfect competition is incompatible with increasing returns (as the Solver example with $TC = 100q^{1/2}$ showed) led to a heated debate in the 1920s. Economics continues to struggle to develop a model that combines the fact that average cost falls as output rises for many products with competitive markets. See David Warsh (2006), *Knowledge and the Wealth of Nations: A Story of Economic Discovery*, for a review of how economics has grappled with the issue of increasing returns.

If you are interested in the trajectory of capitalism and markets, then modern economic theory will not be of much help. For an entertaining review of capitalism and how it has been treated in economics, no one has beaten this classic: Robert Heilbroner, *The Worldly Philosophers: The Lives, Times, and Ideas of the Great Economic Thinkers* (New York: Touchstone, 1999, 7th edition, originally published 1953).

INTERMEDIATE MICROECONOMICS WITH MICROSOFT EXCEL®

This unique text uses Microsoft Excel® workbooks to instruct students. In addition to explaining fundamental concepts in microeconomic theory, readers acquire a great deal of sophisticated Excel skills and gain the practical mathematics needed to succeed in advanced courses. In addition to the innovative pedagogical approach, the book features explicitly repeated use of a single central methodology, the economic approach. Students learn how economists think and how to think like an economist. With concrete, numerical examples and novel, engaging applications, interest for readers remains high as live graphs and data respond to manipulation by the user. Finally, clear writing and active learning are features sure to appeal to modern practitioners and their students. The website accompanying the text is found at www.depauw.edu/learn/microexcel.

Humberto Barreto is Professor of Econmics and Management at DePauw University. He earned his Ph.D. from the University of North Carolina at Chapel Hill. Professor Barreto has lectured around the world on teaching economics with computer-based methods, including Cuba, Brazil, Canada, Scotland, Spain, Poland, India, Burma, Japan, and Taiwan. He was a Fulbright Scholar in the Dominican Republic and has taught National Science Foundation (NSF) Chautauqua short courses using simulation. He has received several research and teaching awards. His book, *The Entrepreneur in Microeconomic Theory*, was translated into Arabic in 1999. He has written numerous articles and books on using Excel to teach economics (including introductory level material, micro, macro and econometrics). He offers an annual workshop for faculty on teaching economics. Visit his website for more information: Teaching Economics with Excel.