

4 Producer Theory

The most basic theory of the *firm* views the firm as a means of transforming things into other, more valuable things, which is known as *production*. Thus, smelting of copper or gold removes impurities and makes the resulting product more valuable. Silicon valley transforms silicon, which is the primary ingredient of sand, along with a thousand other chemicals and metals, into computer chips used in everything from computers to toasters. Cooking transforms raw food, adding flavor and killing bacteria. Moving things to locations where they have higher value is a form of production. Moving stone to the location of a house where the stone can be installed in the exterior, or bringing the King Tut museum exhibit temporarily to Chicago, or a basketball team to the playoffs, are all examples of production. In this simplistic view, a firm is comprised of a technology or set of technologies for transforming things and then chooses the transformation to maximize the net profits. This “firm as a production function” view of the firm is adequate for some purposes, especially when products or services are relatively standardized and technologies widely available, but fares poorly when the internal organization of the firm matters a great deal. Nevertheless, the “firm as a production function” model is a natural starting point in the investigation of competition.

4.1 The Competitive Firm

4.1.1 Types of Firms

There are four major types of firms created in law, although these broad types have many subtypes. At the smallest end is the *proprietorship*, in which a firm is owned by a single individual (the proprietor) or perhaps a family, and operated by a relatively small number of people. The family farm, many restaurants, convenience stores, and laundromats are operated this way. Debts accrued by the proprietorship are the personal responsibility of the proprietor. Professionals like attorneys and accountants are often organized as *partnerships*. Partnerships share profits according to a formula (some equally by partner, some assigning shares or points to partners so that ‘rainmakers’ who generate more of the business obtain a larger share of the profits) and usually all are liable for losses incurred by the partnership. Thus, if a partner in a law firm steals a client’s money and disappears, the other partners are generally responsible for the loss. In contrast, a *corporation* is, by a legal fiction, a person, which means a corporation itself can incur debt and the responsibility for repayment of that debt is with the corporation, not with the officers or owners of the corporation. When the energy trader company Enron collapsed, the shareholders in Enron lost their investment in the stock, but were not responsible for the remaining debts of the corporation. Moreover, executives of the company are also not financially responsible for debts of the corporation, provided the executives act legally and carry out their responsibilities appropriately. If a meteor strikes a manufacturing facility and bankrupts the corporation, the executives are not personally responsible for the debts the corporation fails to pay. On the other hand, breaking the law is not permitted, and executives at Archer Daniels Midland, the large agriculture firm, who colluded in the fixing of the price of lysine went to jail and were personally fined. The corporation was fined as well.

Corporations shield company executives and shareholders from liability, and are said to offer “limited liability.” So why would anyone in their right mind organize a firm as a proprietorship or a partnership? Corporations cost money to organize, about \$1,000 per year at the time of this writing, and are taxed, which is why many small businesses are organized as proprietorships: it is cheaper. Moreover, it may not be possible for a corporation owned by a family to borrow money to open a restaurant: potential lenders fear not being repaid in the event of bankruptcy, so insist on some personal liability on the part of the owners. So why are professional groups organized as partnerships and not corporations? The short answer is that a large variety of hybrid organizational forms exist. The distinctions have been blurred and organizations like “Chapter S Corporations” and “Limited Liability Partnerships” offer the advantages of partnerships (including avoidance of taxation) and corporations. The disadvantages to these forms is primarily larger legal fees, and limitations on the nature of ownership and rules specific to individual states.

It is usually the case that proprietorships are smaller than partnerships, and partnerships smaller than corporations, although there are some very large partnerships (e.g. the big four accounting firms) and some tiny corporations. The fourth kind can be of any size, for its distinction is not how it is organized internally but what it does with the revenue. The *non-profit* firm is prohibited from distributing a profit to its owners. Religious operations, academic associations, environmental groups, most zoos, industry associations, lobbying groups, many hospitals, credit unions (a type of bank), labor unions, private universities and charities are all organized as non-profit corporations. The major advantage of non-profit firms is that the government doesn’t tax them. In exchange for avoiding taxes, non-profits must be engaged in government-approved activities, meaning generally that the non-profit operates for the benefit of some segment of society. So why can’t you establish your own non-profit, that operates for the benefit of you, and avoid taxes? Generally you alone aren’t enough of a socially worthy purpose to meet the requirements to form a non-profit.¹⁹ Moreover, you can’t establish a non-profit for a worthy goal and not serve that goal but just pay yourself all the money the corporation raises, because non-profits are prohibited from overpaying their managers, since overpaying the manager means not serving the worthy corporate goal as well as possible. Finally, commercial activities of non-profits are taxable. Thus, when the non-profit zoo sells stuffed animals in the gift-shop, generally the zoo collects sales tax and is potentially subject to corporate taxes.

The modern corporation is a surprisingly recent invention. Prior to World War I, companies were generally organized in a pyramid structure, with a president at the top, and vice-presidents who reported to him, etc. In a pyramid structure, there is a well-defined chain of command, and no one is ever below two distinct managers of the same level. The problem with a pyramid structure is that two retail stores that want to coordinate have to contact their managers, and possibly their managers’ managers, and so on up the pyramid until a common manager is reached. There are circumstances where such rigid decision-making is unwieldy, and the larger the operation of a corporation, the more unwieldy it gets.

¹⁹ Certainly some of the non-profit religious organizations created by televangelists suggest that the non-profit established for the benefit of a single individual isn’t far-fetched.

Four companies – Sears, DuPont, General Motors and Standard Oil of New Jersey (Exxon) – found that the pyramid structure didn't work well for them. Sears found that its separate businesses of retail stores and mail order required a mix of shared inputs (purchased goods) but distinct marketing and warehousing of these goods. Consequently, retail stores and mail order needed to be separate business units, but purchasing had to answer to both of them. Similarly, DuPont's military business (e.g. explosives) and consumer chemicals were very different operations serving very different kinds of customers, yet often selling the same things, so again the inputs needed to be centrally produced and to coordinate with two separate corporate divisions. General Motors' many car divisions employ 'friendly rivalry,' in which technology and parts are shared across the divisions but the divisions compete in marketing their cars to consumers. Again, technology can't be under just one division, but instead is common to all. Finally, Standard Oil of New Jersey was attempting to create a company that managed oil products from oil exploration all the way through pumping gasoline into automobile gas tanks. With such varied operations all over the globe, Standard Oil of New Jersey required extensive coordination and found that the old business model needed to be replaced. These four companies independently invented the modern corporation, which is organized into separate business units. These business units run as semi-autonomous companies themselves, with one business unit purchasing, at a negotiated price, inputs from another unit, and selling outputs to a third. The study of the internal organization of firms and its ramifications for competitiveness is fascinating, but beyond the scope of this book.²⁰

4.1.2 Production Functions

The firm transforms inputs into outputs. For example, a bakery takes inputs like flour, water, yeast, labor, and heat and makes loaves of bread. An earth-moving company takes capital equipment, ranging from shovels to bulldozers, and labor and digs holes. A computer manufacturer buys parts, generally "off-the-shelf" like disk-drives and memory, along with cases and keyboards and other parts that may be manufactured specially for the computer manufacturer, and uses labor to produce computers. Starbucks takes coffee beans, water, some capital equipment, and labor and produces brewed coffee.

Many if not all firms produce several outputs. However, we can view a firm producing multiple outputs as using several distinct production processes, and thus it is useful to start by looking at a firm that produces only one output. Generally, we can describe this firm as buying an amount x_1 of the first input, x_2 of the second input, and so on (we'll use x_n to denote the last input) and producing an amount y of the output, that is, the production function is

$$y = f(x_1, x_2, \dots, x_n).$$

Mostly we will focus on two inputs in this section, but carrying out the analysis for more than two inputs is straightforward.

²⁰ If you want to know more about organization theory, I happily recommend *Competitive Solutions: The Strategist's Toolkit*, by R. Preston McAfee, Princeton: Princeton University Press, 2002.

Example: The *Cobb-Douglas* production function is the product of the x 's raised to powers, and comes in the form:

$$f(x_1, x_2, \dots, x_n) = a_0 x_1^{a_1} x_2^{a_2} \dots x_n^{a_n}$$

The constants a_1 through a_n are positive numbers, generally individually less than one. For example, with two goods, capital K and labor L , Cobb-Douglas can be expressed as $a_0 K^a L^b$. We will use this example frequently. It is illustrated, for $a_0 = 1$, $a=1/3$ and $b=2/3$, in Figure 4-1.

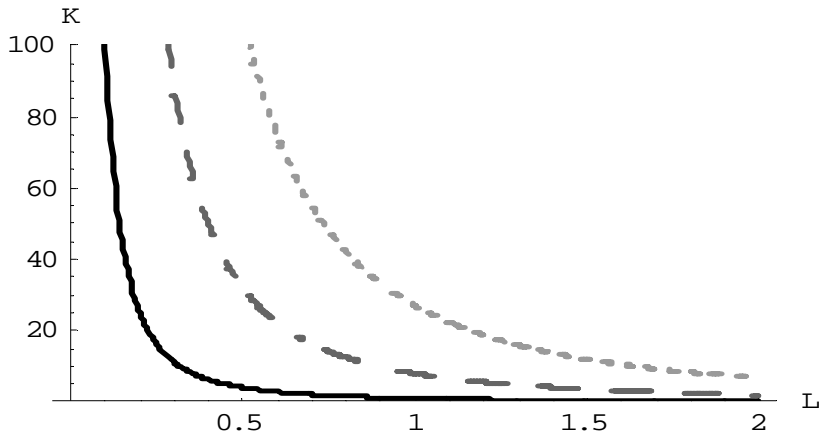


Figure 4-1: Cobb-Douglas Isoquants

Figure 4-1 shows three *isoquants* for the Cobb-Douglas production function. An isoquant, meaning “equal quantity,” illustrates the input mixes that produce a given output level. In this case, given $a=1/3$ and $b=2/3$, we can solve $y = K^a L^b$ for K to obtain $K = y^3 L^{-2}$. Thus, $K = L^{-2}$ gives the combinations of inputs yielding an output of 1, and that is what the dark, solid line represents. The middle, grey dashed line represents an output of 2, and finally the dotted light-grey line represents an output of 3. Isoquants are familiar contour plots used, for example, to show the height of terrain or temperature on a map. Temperature isoquants are, not surprisingly, called isotherms.

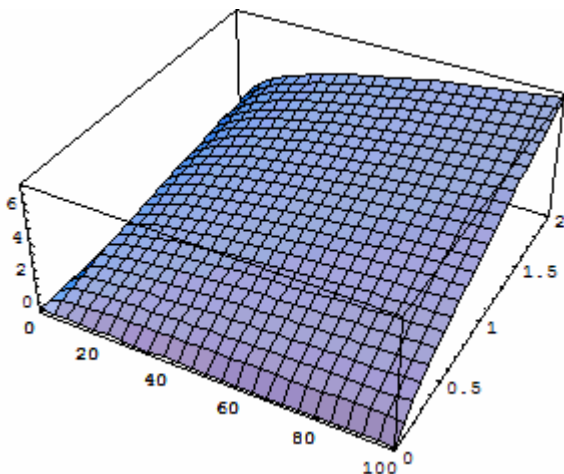


Figure 4-2: The Production Function

Isoquants provide a natural way of looking at production functions and are a bit more useful to examine than 3-D plots like the one provided in Figure 4-2.

The *fixed-proportions production function* comes in the form

$$f(x_1, x_2, \dots, x_n) = \text{Min} \{a_1 x_1, a_2 x_2, \dots, a_n x_n\}$$

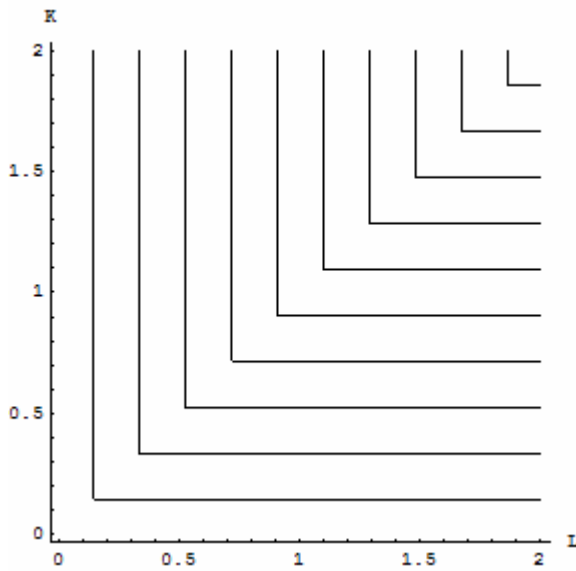


Figure 4-3: Fixed Proportions

The fixed proportions production function has the property that adding an input beyond a necessary level does no good. For example, the productive value of having more than one shovel per worker is pretty low, so that shovels and diggers are reasonably modeled as producing holes using a fixed proportions production function. Moreover, without a shovel or other digging implement like a backhoe, a bare-handed worker produces so little digging as to be nearly useless, so extra workers beyond the number of shovels have little effect. Ultimately, the size of the holes is pretty much determined by $\text{Min} \{\text{number of shovels, number of diggers}\}$. The Figure 4-3 illustrates the isoquants for fixed proportions. As we will see, fixed proportions makes the inputs “perfect complements.”

Two inputs K and L are *perfect substitutes* in a production function f if they enter as a sum, that is, $f(K, L, x_3, \dots, x_n) = g(K + cL, x_3, \dots, x_n)$, for a constant c . With an appropriate scaling of the units of one of the variables, all that matters is the sum of the two variables, not the individual values. In this case, the isoquants are straight lines that are parallel to each other, as illustrated in the Figure 4-4.

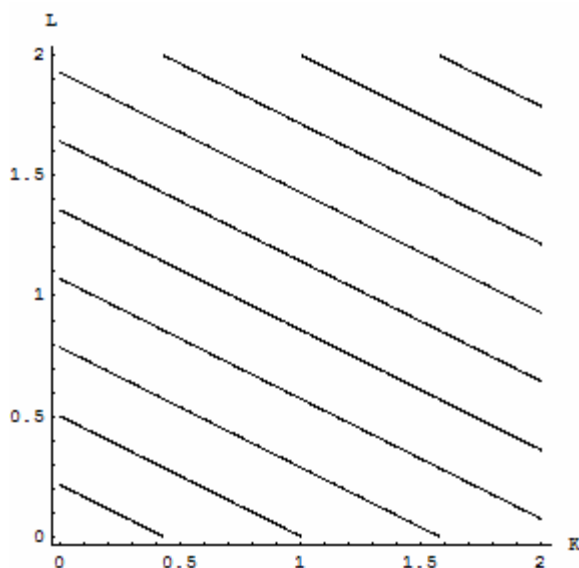


Figure 4-4: Perfect Substitutes

The *marginal product* of an input is just the derivative of the production function with respect to that input.²¹ An important aspect of marginal products is that they are affected by the level of other inputs. For example, in the Cobb-Douglas case with two inputs²² and for constant A :

$$f(K, L) = AK^\alpha L^\beta,$$

the marginal product of capital is

$$\frac{\partial f}{\partial K}(K, L) = \alpha AK^{\alpha-1} L^\beta.$$

If α and β are between zero and one (the usual case), then the marginal product of capital increases in the amount of labor, and decreases in the amount of capital. For example, an extra computer is very productive in a situation with lots of workers and few computers, but not so productive in a situation where there are lots of computers and few people to operate them.

The *value of the marginal product* of an input is just the marginal product times the price of the output. If the value of the marginal product of an input exceeds the cost of that input, it is profitable to use more of the input.

Some inputs are more readily changed than others. It can take five years or more to order and obtain new passenger aircraft, four years to build an electricity generation

²¹ This is a partial derivative, since it holds the other inputs fixed. Partial derivatives are denoted with the symbol ∂ .

²² The symbol α is the Greek letter “alpha.” The symbol β is the Greek letter “beta.” These are the first two letters of the Greek alphabet, and the word alphabet itself originates from these two letters.

facility or a pulp and paper mill. Very skilled labor – experienced engineers, animators, patent attorneys – is often hard to find and challenging to hire. It usually takes three to five years to hire even a small number of academic economists. On the other hand, it is possible to buy shovels, telephones, and computers and to hire a variety of temporary workers quite rapidly, in a matter of a day or so. Moreover, additional hours of work can be obtained by an existing labor force simply by hiring them “overtime,” at least on a temporary basis. The amount of water or electricity a production facility uses can be varied second by second. If you run a restaurant, you can use more water tonight to wash dishes if you need it. If you start in the morning, you can probably get a few additional workers by that evening by paying overtime to those who aren’t scheduled to work. It will probably take a few days or more to hire additional waiters and waitresses, and perhaps more than a few days to find a skilled chef. You can obtain more ingredients, generally the same day, and more plates and silverware pretty quickly. You can lease more space, but it will probably take more than a month to actually occupy a larger space, what with finding the space for rent, renting it, remodeling it and obtaining the necessary permits.

That some inputs or *factors* can be varied quickly, others only slowly, leads to the notions of the long-run and short-run. In the short-run, only some inputs can be adjusted, while in the long-run, all inputs can be adjusted. Traditionally, economists viewed labor as quickly adjustable, and capital equipment as more difficult to adjust. That is certainly right for airlines – obtaining new aircraft is a very slow process – and for large complex factories, and for relatively low-skilled and hence substitutable labor. On the other hand, obtaining workers with unusual skills is a slower process than obtaining warehouse or office space. Generally speaking, the long-run inputs are those that are expensive to adjust quickly, while the short-run factors can be adjusted in a relatively short time frame. What factors belong in which category is dependent on the context or application under consideration.

4.1.2.1 (Exercise) For the Cobb-Douglas production function, suppose there are two inputs K and L , and the sum of the exponents is one. Show that if each input is paid the value of the marginal product per unit of the input, the entire output is just exhausted. That is, for this production function, show

$$K \frac{\partial f}{\partial K} + L \frac{\partial f}{\partial L} = f(K, L).$$

4.1.3 Profit Maximization

Consider an entrepreneur that would like to maximize profit, perhaps by running a delivery service. The entrepreneur uses two inputs, capital K (e.g. trucks) and labor L (e.g. drivers), and rents the capital at cost r per dollar of capital. The wage rate for drivers is w . The production function is $F(K, L)$, that is, given inputs K and L , the output is $F(K, L)$. Suppose p is the price of the output. This gives a profit of:²³

²³ Economists often use the Greek letter π to stand for profit. There is little risk of confusion because economics doesn’t use the ratio of the circumference to the diameter of a circle very often. On the other hand, the other two named constants, Euler’s e and i , the square root of -1, appear fairly frequently in economic analysis.

$$\pi = pF(K, L) - rK - wL.$$

First, consider the case of a fixed level of K . The entrepreneur chooses L to maximize profit. The value L^* of L that maximizes the function π must satisfy:

$$0 = \frac{\partial \pi}{\partial L} = p \frac{\partial F}{\partial L}(K, L^*) - w.$$

This expression is known as a *first order condition*, because it says the first derivative of the function is zero.²⁴ The first order condition shows that we add workers to the production process until reaching a worker who just pays his salary, in that the value of the marginal product for that worker is equal to the cost of the worker.

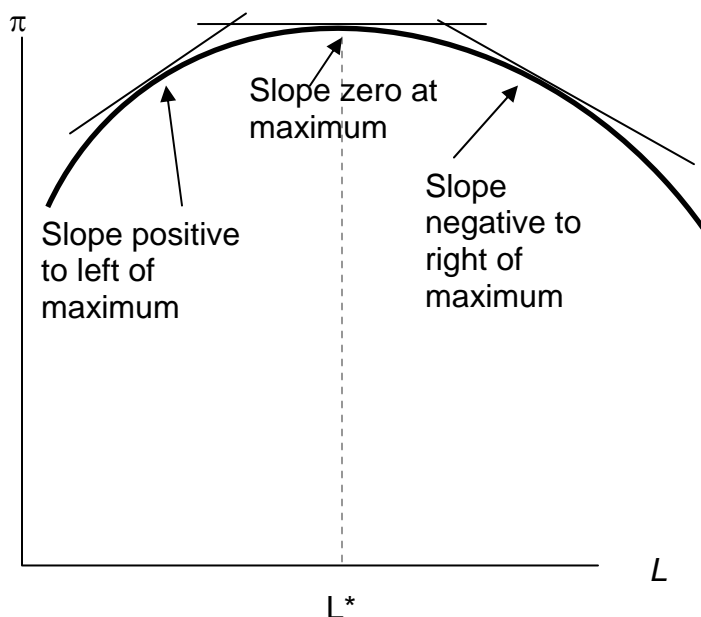


Figure 4-5: Profit-Maximizing Labor Input

In addition, a second characteristic of a maximum is that the second derivative is negative (or non-positive). This arises because, at a maximum, the slope goes from positive (since the function is increasing up to the maximum), to zero (at the maximum), to a negative number (because the function is falling as the variable rises past the maximum). This means that the derivative is falling, that is, the second derivative is negative. This logic is illustrated in the Figure 4-5.

²⁴ It is possible that $L=0$ is the best that entrepreneur can do. In this case, the derivative of profit with respect to L is not necessarily zero. The first order condition instead would be:

Either $0 = \frac{\partial \pi}{\partial L}$, or $L=0$ and $0 \geq \frac{\partial \pi}{\partial L}$. The latter pair of conditions reflects the logic that either the

derivative is zero and we are at a maximum, or $L = 0$, in which case a small increase in L must not cause π to increase.

The second property is known as the *second order condition*, because it is a condition on the second derivative.²⁵ It is expressed as:

$$0 \geq \frac{\partial^2 \pi}{(\partial L)^2} = p \frac{\partial^2 F}{(\partial L)^2} (K, L^*).$$

This is enough of a mathematical treatment to establish comparative statics on the demand for labor. Here, we treat the choice L^* as a function of another parameter – the price p , the wage w , or the level of capital K . For example, to find the effect of the wage on the labor demanded by the entrepreneur, we can write:

$$0 = p \frac{\partial F}{\partial L} (K, L^*(w)) - w.$$

This expression recognizes that the choice L^* that the entrepreneur makes satisfies the first order condition, and results in a value that depends on w . But how does it depend on w ? We can differentiate this expression to obtain:

$$0 = p \frac{\partial^2 F}{(\partial L)^2} (K, L^*(w)) L^{*'}(w) - 1,$$

or

$$L^{*'}(w) = \frac{1}{p \frac{\partial^2 F}{(\partial L)^2} (K, L^*(w))} \leq 0.$$

The second order condition lets the derivative be signed. This form of argument assumes that the choice L^* is differentiable, which is not necessarily true.

Digression: In fact, there is a *revealed preference* form of argument that makes the point without calculus and makes it substantially more generally. Suppose $w_1 < w_2$ are two wage levels, and that the entrepreneur chooses L_1 when the wage is w_1 and L_2 when the wage is w_2 . Then profit maximization requires that these choices are optimal. In particular, when the wage is w_1 , the entrepreneur earns higher profit with L_1 than with L_2 :

$$pf(K, L_1) - rK - w_1 L_1 \geq pf(K, L_2) - rK - w_1 L_2.$$

When the wage is w_2 , the entrepreneur earns higher profit with L_2 than with L_1 .

²⁵ The orders refer to considering small but positive terms Δ , which are sent to zero to reach derivatives. The value Δ^2 , the “second order term” goes to zero faster than Δ , the first order term.

$$pf(K, L_2) - rK - w_2 L_2 \geq pf(K, L_1) - rK - w_2 L_1.$$

The sum of the left hand sides of these two expressions is at least as large as the sum of the right hand side of the two expressions:

$$pf(K, L_1) - rK - w_1 L_1 + pf(K, L_2) - rK - w_2 L_2 \geq pf(K, L_1) - rK - w_2 L_1 + pf(K, L_2) - rK - w_1 L_2$$

A large number of terms cancel, to yield

$$-w_1 L_1 - w_2 L_2 \geq -w_2 L_1 - w_1 L_2.$$

This expression can be re-arranged to yield

$$(w_1 - w_2)(L_2 - L_1) \geq 0.$$

This shows that the higher labor choice must be associated with the lower wage. This kind of argument, sometimes known as a “revealed preference” kind of argument because choices by consumers were the first place the type of argument was applied, can be very powerful and general, because issues of differentiability are avoided. However, we will use the more standard differentiability type argument, because such arguments are usually more readily constructed.

The effect of an increase in the capital level K on the choice by the entrepreneur can be calculated by considering L^* as a function of the capital level K .

$$0 = p \frac{\partial F}{\partial L}(K, L^*(K)) - w.$$

Differentiating this expression with respect to K , we obtain

$$0 = p \frac{\partial^2 F}{\partial K \partial L}(K, L^*(K)) + p \frac{\partial^2 F}{(\partial L)^2}(K, L^*(K)) L^{*'}(K),$$

or,

$$L^{*'}(K) = \frac{-\frac{\partial^2 F}{\partial K \partial L}(K, L^*(K))}{\frac{\partial^2 F}{(\partial L)^2}(K, L^*(K))}.$$

We know the denominator of this expression is not positive, thanks to the second order condition, so the unknown part is the numerator. We then obtain the conclusion that:

An increase in capital increases the labor demanded by the entrepreneur if $\frac{\partial^2 F}{\partial K \partial L}(K, L^*(K)) > 0$, and decreases the labor demanded if $\frac{\partial^2 F}{\partial K \partial L}(K, L^*(K)) < 0$.

This conclusion looks like gobbledygook but is actually quite intuitive. Note that $\frac{\partial^2 F}{\partial K \partial L}(K, L^*(K)) > 0$ means that an increase in capital increases the derivative of output with respect to labor, that is, an increase in capital increases the marginal product of labor. But this is the definition of a complement! That is, $\frac{\partial^2 F}{\partial K \partial L}(K, L^*(K)) > 0$ means that labor and capital are complements in production – an increase in capital increases the marginal productivity of labor. Thus an increase in capital will increase the demand for labor when labor and capital are complements, and will decrease the demand for labor when labor and capital are substitutes.

This is an important conclusion because different kinds of capital may be complements or substitutes for labor. Are computers complements or substitutes for labor? Some economists consider that computers are complements to highly skilled workers, increasing the marginal value of the most skilled, but substitute for lower skilled workers. In academia, the ratio of secretaries to professors has fallen dramatically since the 1970s as more and more professors use machines to perform secretarial functions. Computers are thought to have increased the marginal product of professors and reduced the marginal product of secretaries, so the number of professors rose and the number of secretaries fell.

The revealed preference version of the effect of an increase in capital is to posit two capital levels, K_1 and K_2 , with associated profit-maximizing choices L_1 and L_2 . The choices require, for profit maximization, that

$$pF(K_1, L_1) - rK_1 - wL_1 \geq pF(K_1, L_2) - rK_1 - wL_2$$

and

$$pF(K_2, L_2) - rK_2 - wL_2 \geq pF(K_2, L_1) - rK_2 - wL_1.$$

Again, adding the left-hand-sides together produces a result at least as large as the sum of the right hand sides:

$$pF(K_1, L_1) - rK_1 - wL_1 + pF(K_2, L_2) - rK_2 - wL_2 \geq pF(K_2, L_1) - rK_2 - wL_1 + pF(K_1, L_2) - rK_1 - wL_2.$$

Eliminating redundant terms yields

$$pF(K_1, L_1) + pF(K_2, L_2) \geq pF(K_2, L_1) + pF(K_1, L_2),$$

or,

$$F(K_2, L_2) - F(K_1, L_2) \geq F(K_2, L_1) - F(K_1, L_1)$$

or,

$$\int_{K_1}^{K_2} \frac{\partial F}{\partial K}(x, L_2) dx \geq \int_{K_1}^{K_2} \frac{\partial F}{\partial K}(x, L_1) dx,^{26}$$

or

$$\int_{K_1}^{K_2} \frac{\partial F}{\partial K}(x, L_2) - \frac{\partial F}{\partial K}(x, L_1) dx \geq 0,$$

and finally,

$$\int_{K_1}^{K_2} \int_{L_1}^{L_2} \frac{\partial^2 F}{\partial K \partial L}(x, y) dy dx \geq 0,$$

Thus, if $K_2 > K_1$ and $\frac{\partial^2 F}{\partial K \partial L}(K, L) > 0$ for all K and L , then $L_2 \geq L_1$, that is, with

complementary inputs, an increase in one input increases the optimal choice of the second input. In contrast, with substitutes, an increase in one input decreases the other input. While we still used differentiability of the production function to carry out the revealed preference argument, we did not need to establish that the choice L^* was differentiable to perform the analysis.

Example: Labor Demand with the Cobb-Douglas production function. The Cobb-Douglas production function has the form $F(K, L) = AK^\alpha L^\beta$, for constants A , α and β , all positive. It is necessary for $\beta < 1$ for the solution to be finite and well-defined. The demand for labor satisfies

$$0 = p \frac{\partial F}{\partial L}(K, L^*(K)) - w = p\beta AK^\alpha L^{*\beta-1} - w,$$

or

²⁶ Here we use the standard convention that $\int_a^b \dots dx = -\int_b^a \dots dx$.

$$L^* = \left(\frac{p\beta AK^\alpha}{w} \right)^{1/1-\beta}.$$

When $\alpha+\beta=1$, L is linear in capital. Cobb-Douglas production is necessarily complementary, that is, an increase in capital increases labor demanded by the entrepreneur.

4.1.3.1 (Exercise) For the fixed proportions production function $\text{Min} \{K, L\}$, find labor demand (capital fixed at K).

4.1.4 The Shadow Value

When capital K can't be adjusted in the short-run, it creates a constraint on the profit available on the entrepreneur – the desire to change K reduces the profit available to the entrepreneur. There is no direct value of capital, because capital is fixed. That doesn't mean we can't examine its value, however, and the value of capital is called a *shadow value* because it refers to the value associated with a constraint. Shadow value is well-established jargon.

What is the shadow-value of capital? Let's return to the constrained, short-run optimization problem. The profit of the entrepreneur is:

$$\pi = pF(K, L) - rK - wL.$$

The entrepreneur chooses the value L^* to maximize profit, but is stuck in the short-run with the level of capital inherited from a past decision. The shadow value of capital is the value of capital to profit, given the optimal decision L^* . Because

$$0 = \frac{\partial \pi}{\partial L} = p \frac{\partial F}{\partial L}(K, L^*) - w,$$

the shadow value of capital is

$$\frac{d\pi(K, L^*)}{dK} = \frac{\partial \pi(K, L^*)}{\partial K} = p \frac{\partial F}{\partial K}(K, L^*) - r.$$

Note that this could be negative; the entrepreneur might like to sell some capital but can't, perhaps because it is installed in the factory.

Any constraint has a shadow value. The term refers to the value of relaxing a constraint. The shadow value is zero when the constraint doesn't bind; for example, the shadow value of capital is zero when it is set at the profit-maximizing level. Technology binds the firm; the shadow value of a superior technology is the increase in profit associated with it. For example, parameterize the production technology by a parameter a , so that $aF(K, L)$ is produced. The shadow value of a given level of a is, in the short-run,

$$\frac{d\pi(K, L^*)}{da} = \frac{\partial\pi(K, L^*)}{\partial a} = pF(K, L^*).$$

A term is vanishing in the process of establishing the shadow value. The desired value L^* varies with the other parameters like K and a , but the effect of these parameters on L^* doesn't appear in the expression for the shadow value of the parameter because

$$0 = \frac{\partial\pi}{\partial L} \text{ at } L^*.$$

4.1.5 Input Demand

Over a long period of time, an entrepreneur can adjust both the capital and the labor used at the plant. This lets the entrepreneur maximize profit with respect to both variables K and L . We'll use a double star, **, to denote variables in their long-run solution. The approach to maximizing profit over two variables is to maximize it separately over each variable, thereby obtaining

$$0 = p \frac{\partial F}{\partial L}(K^{**}, L^{**}) - w,$$

and

$$0 = p \frac{\partial F}{\partial K}(K^{**}, L^{**}) - r.$$

We see for both capital and labor, the value of the marginal product is equal to purchase price of the input.

It is more of a challenge to carry out comparative statics exercises with two variables, and the general method won't be developed here.²⁷ However, we can illustrate one example as follows.

Example: The Cobb-Douglas production function implies choices of capital and labor satisfying two first order conditions, one each for labor and capital.²⁸

$$0 = p \frac{\partial F}{\partial L}(K^{**}, L^{**}) - w = p\beta AK^{**\alpha} L^{**\beta-1} - w,$$

$$0 = p \frac{\partial F}{\partial K}(K^{**}, L^{**}) - r = p\alpha AK^{**\alpha-1} L^{**\beta} - r.$$

To solve this expression, first rewrite to obtain

²⁷ If you want to know more, the approach is to arrange the two equations as a vector with $\mathbf{x} = (K, L)$, $\mathbf{z} = (r/p, w/p)$, so that $\mathbf{0} = \mathbf{F}'(\mathbf{x}^{**}) - \mathbf{z}$, and then differentiate to obtain $d\mathbf{x}^{**} = (\mathbf{F}''(\mathbf{x}^{**}))^{-1} d\mathbf{z}$, which can then be solved for each comparative static.

²⁸ It is necessary for $\alpha + \beta < 1$ for the solution to be finite and well-defined.

$w = p\beta AK^{**\alpha} L^{**\beta-1}$ and $r = p\alpha AK^{**\alpha-1} L^{**\beta}$, then divide the first by the second to yield

$\frac{w}{r} = \frac{\beta K^{**}}{\alpha L^{**}}$, or $K^{**} = \frac{\alpha w}{\beta r} L^{**}$. This can be substituted into either equation to obtain

$$L^{**} = \left(\frac{Ap\alpha^{\alpha}\beta^{1-\alpha}}{r^{\alpha}w^{1-\alpha}} \right)^{\frac{1}{1-\alpha-\beta}} \text{ and } K^{**} = \left(\frac{Ap\alpha^{1-\beta}\beta^{\beta}}{r^{1-\beta}w^{\beta}} \right)^{\frac{1}{1-\alpha-\beta}}.$$

While these expressions appear complicated, in fact the dependence on the output price p , and the input prices r and w are quite straightforward.

How do equilibrium values of capital and labor respond to a change in input prices or output price for the Cobb-Douglas production function? It is useful to cast these changes in percentage terms. It is straightforward to demonstrate that both capital and labor respond to a small percentage change in any of these variables with a constant percentage change.

4.1.5.1 (Exercise) For the Cobb-Douglas production function $F(K, L) = AK^{\alpha}L^{\beta}$, show

$$\frac{r}{L^{**}} \frac{\partial L^{**}}{\partial r} = -\frac{\alpha}{1-\alpha-\beta}, \quad \frac{w}{L^{**}} \frac{\partial L^{**}}{\partial w} = -\frac{1-\alpha}{1-\alpha-\beta}, \quad \frac{p}{L^{**}} \frac{\partial L^{**}}{\partial p} = \frac{1}{1-\alpha-\beta},$$

$$\frac{r}{K^{**}} \frac{\partial K^{**}}{\partial r} = -\frac{1-\beta}{1-\alpha-\beta}, \quad \frac{w}{K^{**}} \frac{\partial K^{**}}{\partial w} = -\frac{\beta}{1-\alpha-\beta} \text{ and } \frac{p}{K^{**}} \frac{\partial K^{**}}{\partial p} = \frac{1}{1-\alpha-\beta}.$$

An important insight of profit maximization is that it implies minimization of costs of yielding the chosen output, that is, *profit-maximization entails efficient production*. The logic is straightforward. The profit of an entrepreneur is revenue minus costs, and the revenue is price times output. For the chosen output, then, the entrepreneur earns the revenue associated with the output, which is fixed since we are considering only the chosen output, minus the costs of producing that output. Thus, for the given output, maximizing profits is equivalent to maximizing a constant (revenue) minus costs. Since maximizing $-C$ is equivalent to minimizing C , the profit-maximizing entrepreneur minimizes costs. This is important because profit-maximization implies not being wasteful in this regard: a profit-maximizing entrepreneur produces at least cost.

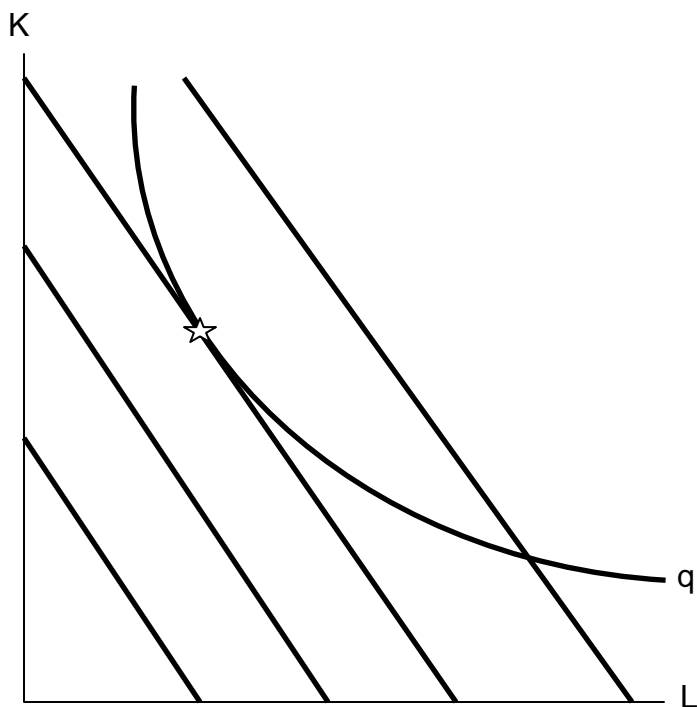


Figure 4-6: Tangency and Isoquants

There are circumstances where the cost-minimization feature of profit maximization can be used, and this is especially true when a graphical approach is taken. The graphical approach to profit-maximization is illustrated in Figure 4-6. The curve represents an isoquant, which holds constant the output. The straight lines represent “isocost” lines, which hold constant the expenditure on inputs. Isocost lines solve the problem

$$rK + wL = \text{constant}$$

and thus have slope $\frac{dK}{dL} = -\frac{w}{r}$. Isocost lines are necessarily parallel – they have the same slope. Moreover, the cost associated with an isocost line rises the further northeast we go in the graph, or the further away from the origin.

What point on an isoquant minimizes total cost? The answer is the point associated with the lowest (most southwest) isocost that intersects the isoquant. This point will be tangent to the isoquant and is denoted by a star. At any lower cost, it isn’t possible to produce the desired quantity. At any higher cost, it is possible to lower cost and still produce the quantity.

That cost minimization requires a tangency between the isoquant and the isocost has a useful interpretation. The slope of the isocost is minus the ratio of input prices. The slope of the isoquant measures the substitutability of the inputs in producing the output. Economists call this slope the *marginal rate of technical substitution*, which is the amount of one input needed to make up for a decrease in another input and hold output constant. Thus, one feature of cost minimization is that the input price ratio equals the marginal rate of technical substitution.

4.1.6 Myriad Costs

How much does it cost to produce a given quantity q ? We already have a detailed answer to this question, but now need to focus less on the details and more on the “big picture.” First, let’s focus on the short-run, and suppose L is adjustable in the short-run, but K is not. Then the short-run total cost of producing q , given the capital level, is

$$SRTC(q|K) = \min_L rK + wL, \text{ over all } L \text{ satisfying } F(K,L) \geq q.$$

In words, this equation says the short-run total cost of the quantity q given the existing level K is the minimum cost, where L gets to vary (which is denoted by “min over L ”), where the L considered is large enough to produce q . The vertical line $|$ is used to indicate a condition or conditional requirement; here $|K$ indicates that K is fixed. The minimum lets L vary but not K . Finally, there is a constraint $F(K,L) \geq q$, which indicates that one has to be able to produce q with the mix of inputs because we are considering the short-run cost of q .

The short-run total cost of q given K has a simple form. First, since we are minimizing cost, the constraint $F(K,L) \geq q$ will be satisfied with equality, $F(K,L) = q$. This equation determines L , since K is fixed, that is, $F(K, L_S(q, K)) = q$ gives the short-run value of L , $L_S(q, K)$. Finally, the cost is then $rK + wL = rK + wL_S(q, K)$.

The *short-run marginal cost* given K is just the derivative of total cost with respect to q . To establish the short-run marginal cost, note that the equation $F(K,L) = q$ gives

$$\frac{\partial F}{\partial L}(K, L_S(q, K)) dL = dq,$$

or

$$\left. \frac{dL}{dq} \right|_{F=q} = \frac{1}{\frac{\partial F}{\partial L}(K, L_S(q, K))}.$$

The tall vertical line, subscripted with $F=q$, is used to denote the constraint $F(K,L) = q$ that is being differentiated. Thus, the short-run marginal cost is

$$SRMC(q|K) = SRTC'(q) = \frac{d}{dq}(rK + wL) = w \left. \frac{dL}{dq} \right|_{F=q} = \frac{w}{\frac{\partial F}{\partial L}(K, L_S(q, K))}.$$

There are two other short-run costs that will be needed to complete the analysis. First, there is the notion of the short-run average cost of production, which we obtain by dividing the total cost by the quantity:

$$SRAC(q|K) = \frac{SRTC(q|K)}{q}.$$

Finally, we need one more short-run cost: the short-run average variable cost. The variable cost eliminates the fixed costs of operation, which in this case are rK . That is,

$$SRAVC(q|K) = \frac{SRTC(q|K) - SRTC(0|K)}{q} = \frac{wL_S(q|K)}{q}.$$

The short-run average variable cost is the average cost ignoring the investment in capital equipment.

The short-run average cost could also be called the short-run average total cost, since it is the average of the total cost per unit of output, but “average total” is a bit of an oxymoron.²⁹ Consequently, when total, fixed or variable is not specified, the convention is to mean total. Note that the marginal variable cost is the same as the marginal total costs, because the difference between variable cost and total cost is a constant – the cost of zero production, also known as the fixed cost of production.

At this point, we have identified four distinct costs, all relevant to the short-run. These are the total cost, the marginal cost, the average cost, and the average variable cost. In addition, all of these can be considered in the long-run as well. There are three differences in the long-run. First, the long-run lets all inputs vary, so the long-run total cost is

$$LRTC(q) = \min_{L,K} rK + wL, \text{ over all } L \text{ and } K \text{ combinations satisfying } F(K,L) \geq q.$$

Second, since all inputs can vary, the long-run cost isn't conditioned on K . Finally, the long-run average variable cost is the same as the long-run average total cost. Because in the long-run a firm could use no inputs and thus incur no costs, the cost of producing zero is zero. Therefore, in the long-run, all costs are variable, and the long-run average variable cost is the long-run average total cost.

4.1.6.1 (Exercise) For the Cobb-Douglas production function $F(K, L) = AK^\alpha L^\beta$, with $\alpha + \beta < 1$, with K fixed in the short-run but not in the long-run, and cost r of capital and w for labor, show

$$SRTC(q|K) = rK + w \left(\frac{q}{AK^\alpha} \right)^{\frac{1}{\beta}},$$

²⁹ An oxymoron is a word or phrase which is self-contradictory, like “jumbo shrimp,” “stationary orbit,” “virtual reality,” “modern tradition,” or “pretty ugly.” Oxymoron comes from the Greek oxy, meaning sharp, and moros, meaning dull. Thus oxymoron is itself an oxymoron, so an oxymoron is self-descriptive. Another word which is self-descriptive is “pentasyllabic.”

$$\text{SRAVC}(q|K) = w \frac{q^{\frac{1-\beta}{\beta}}}{\left(AK^\alpha\right)^{\frac{1}{\beta}}},$$

$$\text{SRMC}(q|K) = w \frac{q^{\frac{1-\beta}{\beta}}}{\beta \left(AK^\alpha\right)^{\frac{1}{\beta}}},$$

$$\text{LRTC}(q|K) = \left(\left(\frac{\alpha}{\beta}\right)^{\frac{\beta}{\alpha+\beta}} + \left(\frac{\beta}{\alpha}\right)^{\frac{\alpha}{\alpha+\beta}} \right) w^{\frac{\beta}{\alpha+\beta}} r^{\frac{\alpha}{\alpha+\beta}} \left(\frac{q}{A}\right)^{\frac{1}{\alpha+\beta}}.$$

Note that the easiest way to find the long-run total cost is to minimize the short-run total cost over K . Since this is a function of one variable, it is straightforward to identify the K that minimizes cost, and then plug that K into the expression for total cost.

One might want to distinguish the very short-run, from the short-run, from the medium run, from the long-run, from the very long-run. But a better approach is to view adjustment as a continuous process, with a gradual easing of the constraints. Faster adjustment costs more. Continuous adjustment is a more advanced topic, requiring an Euler equation approach.

4.1.7 Dynamic Firm Behavior

In this section, we consider a firm or entrepreneur that can't affect the price of output or the prices of inputs, that is, a competitive firm. How does such a competitive firm respond to price changes? When the price of the output rises, the firm earns profits

$$\pi = pq - c(q|K),$$

where $c(q|K)$ is the total cost of producing given that the firm currently has capital K . Assuming the firm produces at all, the firm maximizes profits by choosing the quantity q_s satisfying $0 = p - c'(q_s|K)$, that is, choosing the quantity where price equals marginal cost. However, this is a good strategy only if producing a positive quantity is desirable, that is, if $pq_s - c(q_s|K) \geq -c(0, K)$, which can be rewritten as $p \geq \frac{c(q_s|K) - c(0, K)}{q_s}$.

The right-hand-side of this inequality is the average variable cost of production, and thus the inequality implies that a firm will produce provided price exceeds the average variable cost. Thus, *the profit-maximizing firm produces the quantity q_s where price equals marginal cost, provided price is as large as minimum average variable cost. If price falls below minimum average variable cost, the firm shuts down.*

The behavior of the competitive firm is illustrated in Figure 4-7. The thick line represents the choice of the firm as a function of the price, which is on the vertical axis. Thus, if the price is below the minimum average variable cost (AVC), the firm shuts down. When price is above the minimum average variable cost, the marginal cost gives the quantity supplied by the firm. Thus, the choice of the firm is composed of two distinct segments – the marginal cost, where the firm produces the output where price equals marginal cost, and shutdown, where the firm makes a higher profit, or loses less money, by producing zero.

Figure 4-7 also illustrates the average total cost, which doesn't affect the short term behavior of the firm but does affect the long term behavior, because when price is below average total cost, the firm is not making a profit, but instead would prefer to exit over the long term. That is, when the price is between the minimum average variable cost and the minimum average total cost, it is better to produce than to shut down, but the return on capital was below the cost of capital. With a price in this intermediate area, a firm would produce, but would not replace the capital, and thus would shut down in the long-term if the price is expected to persist. As a consequence, minimum average total cost is the long-run "shut down" point for the competitive firm. (Shutdown may refer to reducing capital rather than literally setting capital to zero.) Similarly, in the long term, the firm produces the quantity where the price equals the long-run marginal cost.

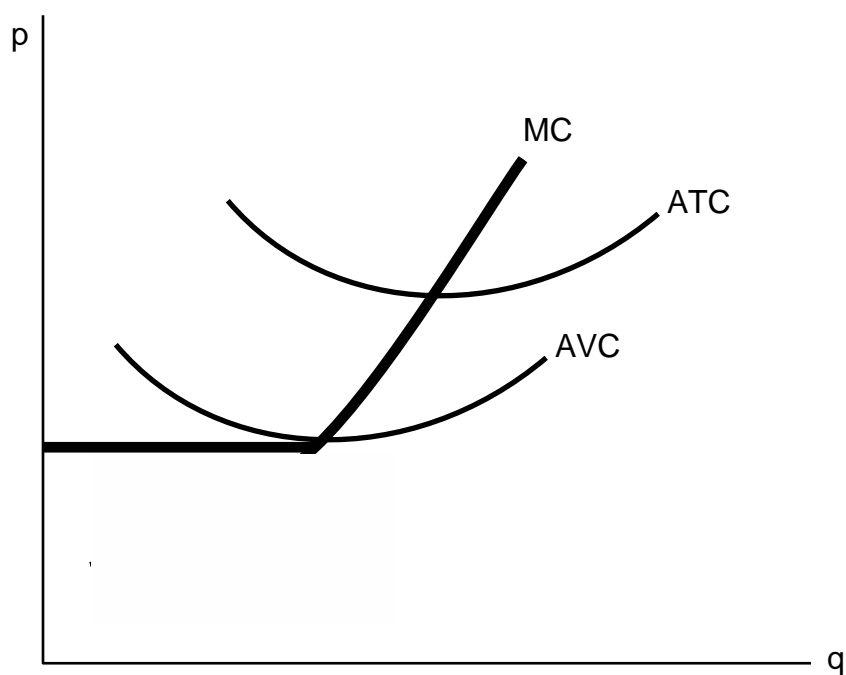


Figure 4-7: Short-Run Supply

Figure 4-7 illustrates one other fact: the minimum of average cost occurs at the point that marginal cost equals average cost. To see this, let $C(q)$ be total cost, so that average cost is $C(q)/q$. Then the minimum of average cost occurs at the point satisfying:

$$0 = \frac{d}{dq} \frac{C(q)}{q} = \frac{C'(q)}{q} - \frac{C(q)}{q^2}.$$

But this can be rearranged to imply $C'(q) = \frac{C(q)}{q}$, that is, marginal cost equals average cost at the minimum of average cost.

The long-run marginal cost has a complicated relationship to short-run marginal cost. The problem in characterizing the relationship between long-run and short-run marginal costs is that some costs are marginal in the long-run that are fixed in the short-run, tending to make long-run marginal costs larger than short-run marginal costs. However, in the long-run, the assets can be configured optimally, while some assets are fixed in the short-run, and this optimal configuration tends to make long-run costs lower.

Instead, it is more useful to compare the long-run average total costs and short-run average total costs. The advantage is that capital costs are included in short-run average total costs. The result is a picture like Figure 4-8.

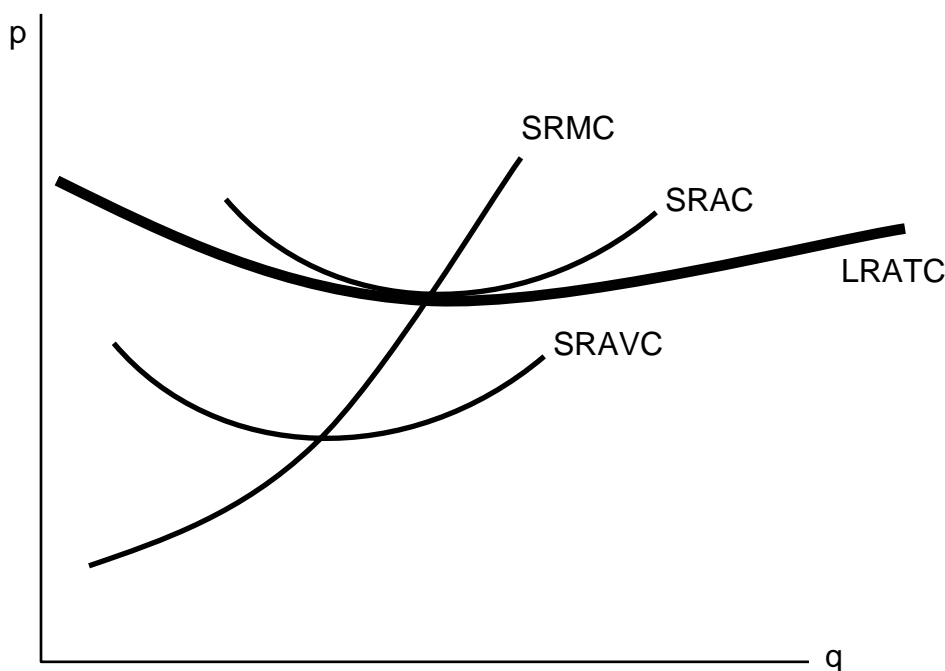


Figure 4-8: Average and Marginal Costs

In Figure 4-8, the short-run is unchanged – there is a short-run average cost, short-run average variable cost, and short-run marginal cost. The long-run average total cost has been added, in such a way that the minimum average total cost occurs at the same point as the minimum short-run average cost, which equals the short-run marginal cost. This is the lowest long-run average cost, and has the nice property that long-run average cost equals short-run average total cost equals short-run marginal cost. However, for a

different output by the firm, there would necessarily be a different plant size, and the three-way equality is broken. Such a point is illustrated in Figure 4-9.

In Figure 4-9, the quantity produced is larger than the quantity that minimizes long-run average total cost. Consequently, as is visible in the picture, the quantity where short-run average cost equals long-run average cost does not minimize short-run average cost. What this means is that a factory designed to minimize the cost of producing a particular quantity won't necessarily minimize short-run average cost. Essentially, because the long-run average total cost is increasing, larger plant sizes are getting increasingly more expensive, and it is cheaper to use a somewhat "too small" plant and more labor than the plant size with the minimum short-run average total cost. However, this situation wouldn't likely persist indefinitely, because, as we shall see, competition tend to force price to the minimum long-run average total cost, and at that point, we have the three-way equality between long-run average total cost, short-run average total cost, and short-run marginal cost.

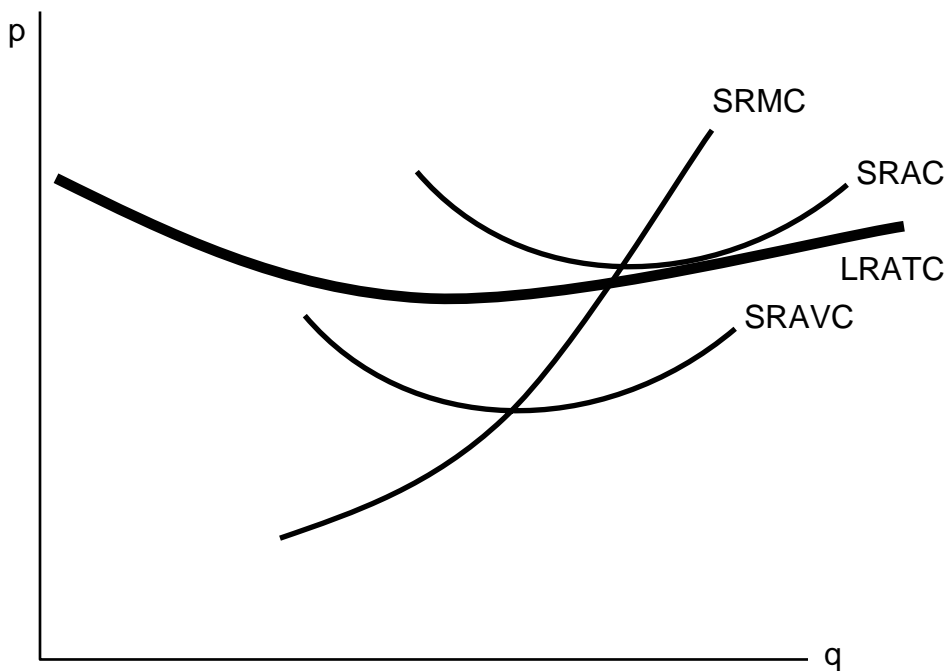


Figure 4-9: Increased Plant Size

4.1.7.1 (Exercise) Suppose a company has total cost given by $rK + \frac{q^2}{2K}$, where capital K is fixed in the short-run. What is short-run average total cost and marginal cost? Plot these curves. For a given quantity q_0 , what level of capital minimizes total cost? What is the minimum average total cost of q_0 ?

4.1.8 Economies of Scale and Scope

An economy of scale – that larger scale lowers cost – arises when an increase in output reduces average costs. We met economies of scale, and their opposite, diseconomies of

scale, in the previous section, with an example where long-run average total cost initially fell, then rose, as quantity was increased.

What makes for an economy of scale? Larger volumes of productions permit the manufacture of more specialized equipment. If I am producing a million identical automotive tail lights, I can spend \$50,000 on an automated plastic stamping machine and only affect my costs by five cents each. In contrast, if I am producing 50,000 units, the stamping machine increases my costs by a dollar each, and is much less economical.

Indeed, it is somewhat more of a puzzle as to what produces a diseconomy of scale. An important source of diseconomies are managerial in nature – organizing a large, complex enterprise is a challenge, and larger organizations tend to devote a larger percentage of their revenues to management of the operation. A bookstore can be run by a couple of individuals who rarely if ever engage in management activities, where a giant chain of bookstores needs finance, human resource, risk management and other “overhead” type expenses just in order to function. Informal operation of small enterprises is replaced by formal procedural rules in large organizations. This idea of managerial diseconomies of scale is reflected in the aphorism that “A platypus is a duck designed by a committee.”

In his influential 1975 book *The Mythical Man-Month*, IBM software manager Fred Brooks describes a particularly severe diseconomy of scale. Adding software engineers to a project increases the number of conversations necessary between pairs of individuals. If there are n engineers, there are $\frac{1}{2} n(n - 1)$ pairs, so that communication costs rise at the square of the project size. This is pithily summarized in *Brooks' Law*: “Adding manpower to a late software project makes it later.”

Another related source of diseconomies of scale involves system slack. In essence, it is easier to hide incompetence and laziness in a large organization than in a small one. There are a lot of familiar examples of this insight, starting with the Peter Principle, which states that people rise in organizations to the point of their own incompetence, which means eventually people cease to do the jobs that they do well.³⁰ That slack grows as an organization grows implies an diseconomy of scale.

Generally, for many types of products, economies of scale from production technology tend to reduce average cost, up to a point where the operation becomes difficult to manage, at which point diseconomies tend to prevent the firm from economically getting larger. Under this view, improvements in information technologies over the past twenty years have permitted firms to get larger and larger. While that seems logical, in fact firms aren't getting that much larger than they used to be, and the share of output produced by the top thousand firms has been relatively steady. That is, the growth in the largest firms just mirrors world output growth.

Related to an economy of scale is an *economy of scope*. An economy of scope is a reduction in cost associated with producing several distinct goods. For example, Boeing, which produces both commercial and military jets, can amortize some of its R&D costs over both types of aircraft, thereby reducing the average costs of each. Scope

³⁰ Laurence Johnston Peter (1919–1990).

economies work like scale economies, except they account for advantages of producing multiple products, where scale economies involve an advantage of multiple units of the same product.

Economies of scale can operate at the level of the individual firm but can also operate at an industry level. Suppose there is an economy of scale in the production of an input. For example, there is an economy of scale in the production of disc drives for personal computers. That means an increase in the production of PCs will tend to lower the price of disc drives, reducing the cost of PCs, which is a scale economy. In this case, it doesn't matter to the scale economy whether one firm or many firms are responsible for the increased production, and this is known as an *external economy of scale* or an *industry economy of scale*, because the scale economy operates at the level of the industry rather than in the individual firm. Thus, the long-run average cost of individual firms may be flat, while the long-run average cost of the industry slopes downward.

Even in the presence of an external economy of scale, there may be diseconomies of scale at the level of the firm. In such a situation, the size of any individual firm is limited by the diseconomy of scale, but nonetheless the average cost of production is decreasing in the total output of the industry, through the entry of additional firms. Generally there is an external diseconomy of scale if a larger industry drives up input prices, for example increasing land costs. Increasing the production of soybeans significantly requires using land that isn't so well suited for them, tending to increase the average cost of production. Such a diseconomy is an external diseconomy rather than operating at the individual farmer level. Second, there is an external economy if an increase in output permits the creation of more specialized techniques and a greater effort in R&D to lower costs. Thus, if an increase in output increases the development of specialized machine tools and other production inputs, an external economy will be present.

An economy of scale arises when total average cost falls as the number of units produced rises. How does this relate to production functions? We let $y=f(x_1, x_2, \dots, x_n)$ be the output when the n inputs x_1, x_2, \dots, x_n are used. A rescaling of the inputs involves increasing the inputs by a fixed percentage, e.g. multiplying them all by the constant λ (the Greek letter lambda), where $\lambda > 1$. What does this do to output? If output goes up by more than λ , we have an economy of scale (also known as *increasing returns to scale*): scaling up production increases output proportionately more. If output goes up by less than λ , we have a diseconomy of scale or *decreasing returns to scale*. And finally, if output rises by exactly λ , we have *constant returns to scale*. How does this relate to average cost? Formally, we have an economy of scale if

$$f(\lambda x_1, \lambda x_2, \dots, \lambda x_n) > \lambda f(x_1, x_2, \dots, x_n) \text{ if } \lambda > 1.$$

This corresponds to decreasing average cost. Let w_1 be the price of input 1, w_2 the price of input 2, and so on. Then the average cost of producing $y=f(x_1, x_2, \dots, x_n)$ is

$$AVC = \frac{w_1 x_1 + w_2 x_2 + \dots + w_n x_n}{f(x_1, x_2, \dots, x_n)}.$$

What happens to average cost as we scale up production by $\lambda > 1$? Call this $AVC(\lambda)$.

$$\begin{aligned} AVC(\lambda) &= \frac{w_1 \lambda x_1 + w_2 \lambda x_2 + \dots + w_n \lambda x_n}{f(\lambda x_1, \lambda x_2, \dots, \lambda x_n)} = \lambda \frac{w_1 x_1 + w_2 x_2 + \dots + w_n x_n}{f(\lambda x_1, \lambda x_2, \dots, \lambda x_n)} \\ &= \frac{\lambda f(x_1, x_2, \dots, x_n)}{f(\lambda x_1, \lambda x_2, \dots, \lambda x_n)} AVC(1) \end{aligned}$$

Thus, average cost falls if there is an economy of scale and rises if there is a diseconomy of scale.

Another insight about the returns to scale concerns the value of the marginal product of inputs. Note that, if there are constant returns to scale:

$$\begin{aligned} x_1 \frac{\partial f}{\partial x_1} + x_2 \frac{\partial f}{\partial x_2} + \dots + x_n \frac{\partial f}{\partial x_n} &= \frac{d}{d\lambda} f(\lambda x_1, \lambda x_2, \dots, \lambda x_n) \Big|_{\lambda \rightarrow 1} = \\ &= \lim_{\lambda \rightarrow 1} \frac{f(\lambda x_1, \lambda x_2, \dots, \lambda x_n) - f(x_1, x_2, \dots, x_n)}{\lambda - 1} = f(x_1, x_2, \dots, x_n) \end{aligned}$$

The value $\frac{\partial f}{\partial x_1}$ is the marginal product of input x_1 , and similarly $\frac{\partial f}{\partial x_2}$ is the marginal product of input 2, and so on. Consequently, if the production function exhibits constant returns to scale, it is possible to divide up output in such a way that each input receives the value of the marginal product. That is, we can give $x_1 \frac{\partial f}{\partial x_1}$ to the suppliers of input 1, $x_2 \frac{\partial f}{\partial x_2}$ to the suppliers of input 2, and so on, and this exactly uses up the all the output. This is known as “paying the marginal product,” because each supplier is paid the marginal product associated with the input.

If there is a diseconomy of scale, then paying the marginal product is feasible, but there is generally something left over, too. If there are increasing returns to scale (an economy of scale), then it is not possible to pay all the inputs their marginal product, that is, $x_1 \frac{\partial f}{\partial x_1} + x_2 \frac{\partial f}{\partial x_2} + \dots + x_n \frac{\partial f}{\partial x_n} > f(x_1, x_2, \dots, x_n)$.

4.1.8.1 (Exercise) Given the Cobb-Douglas production function

$f(x_1, x_2, \dots, x_n) = x_1^{a_1} x_2^{a_2} \dots x_n^{a_n}$, show there is constant returns to scale if $a_1 + a_2 + \dots + a_n = 1$, increasing returns to scale if $a_1 + a_2 + \dots + a_n > 1$, and decreasing returns to scale if $a_1 + a_2 + \dots + a_n < 1$.

4.1.8.2 (Exercise) Suppose a company has total cost given by $rK + \frac{q^2}{2K}$ where capital

K can be adjusted in the long-run. Does this company have an economy of scale, diseconomy of scale, or constant returns to scale in the long-run?

4.1.8.3 (Exercise) A production function f is *homogeneous of degree r* if

$f(\lambda x_1, \lambda x_2, \dots, \lambda x_n) = \lambda^r f(x_1, x_2, \dots, x_n)$. Consider a firm with a production function that is homogeneous of degree r . Suppose further that the firm pays the value of marginal product for all its inputs. Show that the portion of revenue left over is $1 - r$.

4.2 Perfect Competition Dynamics

The previous section developed a detailed analysis of how a competitive firm responds to price and input cost changes. In this section, we consider how a competitive market responds to demand or cost changes.

4.2.1 Long-run Equilibrium

The basic picture of a long-run equilibrium is presented in Figure 4-10. There are three curves, all of which are already familiar. First, there is demand, considered in the first chapter. Here demand is taken to be the “per period” demand. Second, there is the short-run supply, which reflects two components – a shut down point at minimum average variable cost, and quantity such that price equals short-run marginal cost above that level. The short-run supply, however, is the market supply level, which means it sums up the individual firm effects. Finally, there is the long-run average total cost at the industry level, thus reflecting any external diseconomy or economy of scale. As drawn in Figure 4-10, there is no long-run scale effect. The long-run average total cost is also the long-run industry supply.³¹

As drawn, the industry is in equilibrium, with price equal to P_0 , which is the long-run average total cost, and also equating short-run supply and demand. That is, at the price of P_0 , and industry output of Q_0 , no firm wishes to shut down, no firm can make positive profits from entering, there is no excess output, and no consumer is rationed. Thus, no market participant has an incentive to change their behavior, so the market is in both long-run and short-run equilibrium.

³¹ This may seem confusing, because supply is generally the *marginal* cost, not the average cost. However, because a firm will quit producing in the long term if price falls below its minimum average cost, the long-term supply is just the minimum average cost of the individual firms, because this is the marginal cost of the industry.

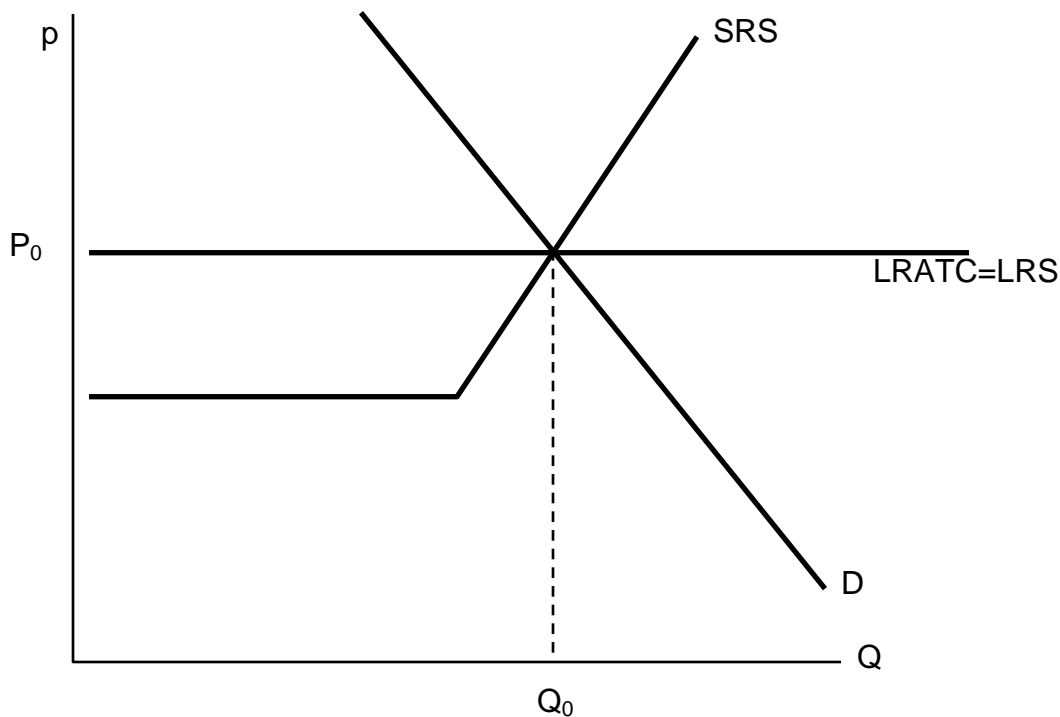


Figure 4-10: Long-Run Equilibrium

4.2.2 Dynamics with Constant Costs

Now consider an increase in demand. Demand might increase because of population growth, or because a new use for an existing product is developed, or because of income growth, or because the product becomes more useful. For example, the widespread adoption of the Atkins diet increased demand for high protein products like beef jerky and eggs. Suppose that the change is expected to be permanent. This is important because the decision of a firm to enter is based more on expectations of future demand than on present demand.

Figure 4-11 reproduces the equilibrium figure, but with the curves “grayed out” to indicate a starting position, and a darker new demand curve, labeled D_1 .

The initial effect of the increased demand is that the price is bid up, because there is excess demand at the old price P_0 . This is reflected by a change in both price and quantity to P_1 and Q_1 , to the intersection of the short-run supply SRS and the new demand curve. This is a short-run equilibrium, and persists temporarily because, in the short-run, the cost of additional supply is higher.

At the new, short-run equilibrium, price exceeds the long-run supply cost. This higher price attracts new investment in the industry. It takes some time for this new investment to increase the quantity supplied, but over time the new investment leads to increased output, and a fall in the price, as illustrated in Figure 4-12.

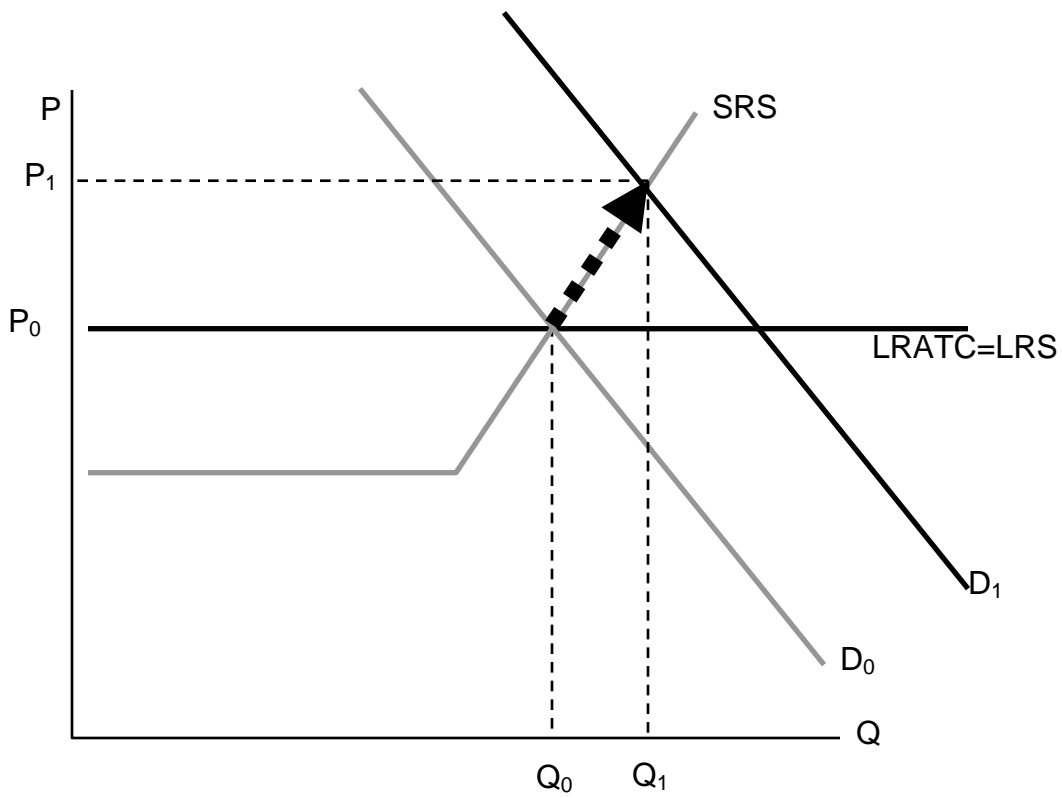


Figure 4-11: A Shift in Demand

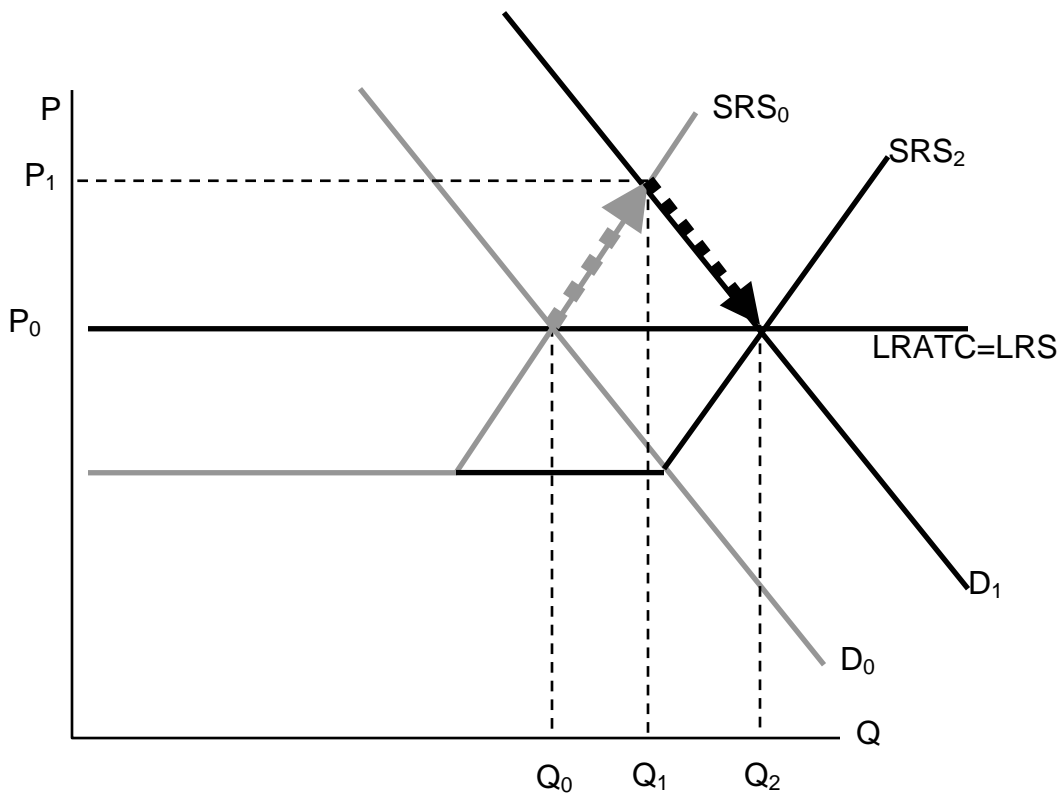


Figure 4-12: Return to Long-Run Equilibrium

As new investment is attracted into the industry, the short-run supply shifts to the right, because with the new investment, more is produced at any given price level. This is illustrated with the darker short-run supply, SRS_2 . The increase in price causes the price to fall back to its initial level, and the quantity to increase still further to Q_2 .

It is tempting to think that the effect of a decrease in demand just retraces the steps of an increase in demand, but that isn't correct. In both cases, the first effect is the intersection of the new demand with the old short-run supply. Only then does the short-run supply adjust to equilibrate the demand with the long-run supply. That is, the initial effect is a short-run equilibrium, followed by adjustment of the short-run supply to bring the system into long-run equilibrium. Moreover, a small decrease in demand can have a qualitatively different effect in the short-run than a large decrease in demand, depending on whether the decrease is large enough to induce immediate exit of firms. This is illustrated in Figure 4-13.

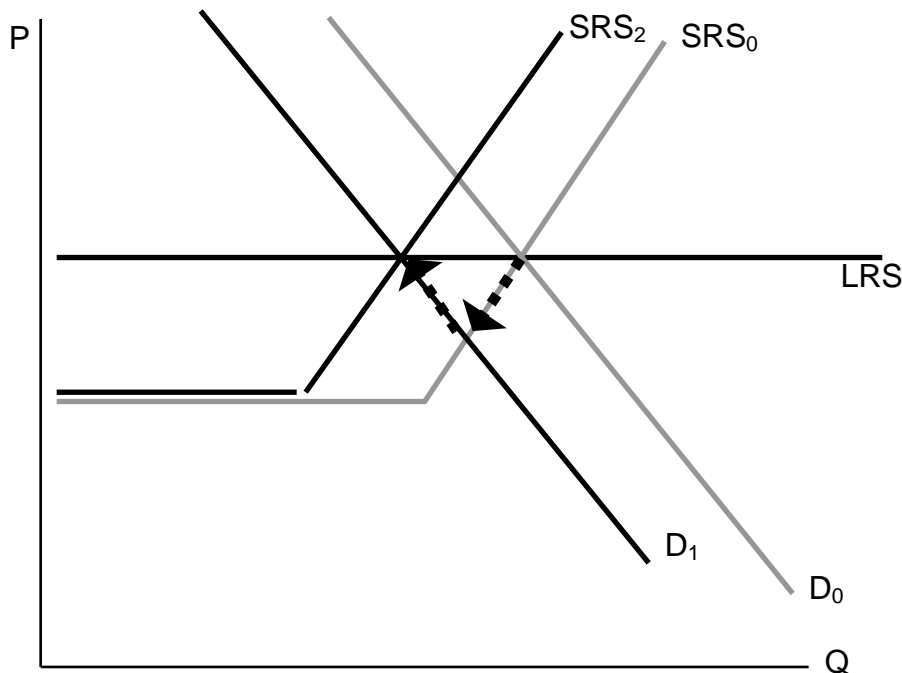


Figure 4-13: A Decrease in Demand

In Figure 4-13, we start at the long-run equilibrium where LRS and D_0 and SRS_0 all intersect. If demand falls to D_1 , the price falls to the intersection of the new demand and the old short-run supply, along SRS_0 . At that point, exit of firms reduces the short-run supply and the price rises, following along the new demand D_1 .

If, however, the decrease in demand is large enough to push the industry to minimum average variable cost, there is immediate exit. In Figure 4-14, the fall in demand from D_0 to D_1 is sufficient to push the price to minimum average variable cost, which is the shutdown point of suppliers. Enough suppliers have to shutdown to keep the price at this level, which induces a shift in of the short-run supply, to SRS_1 . Then there is

additional shutdown, shifting the short-run supply in still further, but driving up the price (along the demand curve) until the long-term equilibrium is reached.

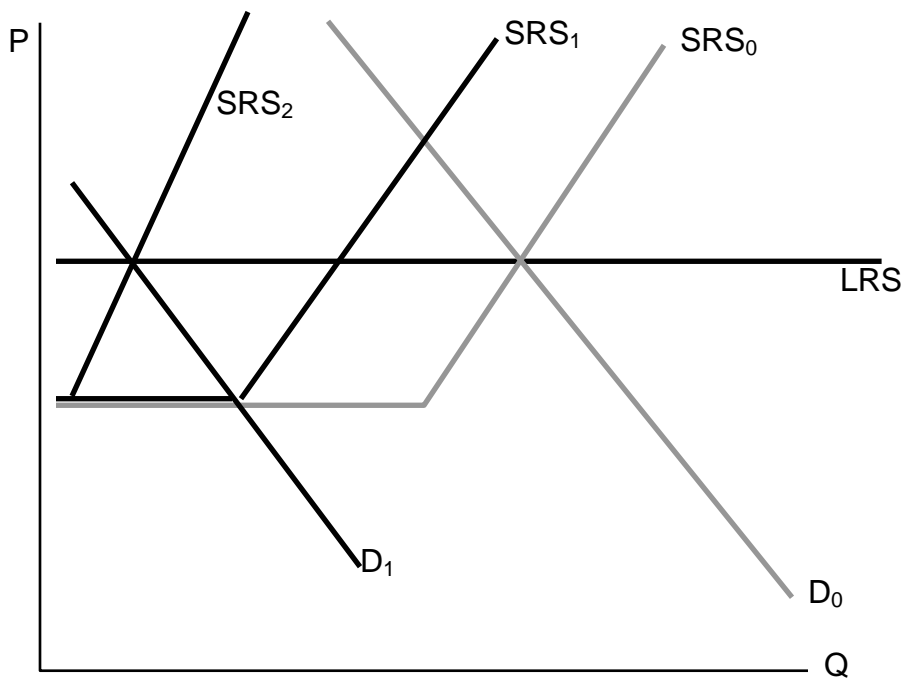


Figure 4-14: A Big Decrease in Demand

Consider an increase in the price of an input into production. For example, an increase in the price of crude oil increases the cost of manufacturing gasoline. This tends to decrease (shift up) both the long-run supply and the short-run supply, by the amount of the cost increase. The effect is illustrated in Figure 4-15. The increased costs reduce both the short-run supply (prices have to be higher in order to produce the same quantity) and the long-run supply. The short-run supply shifts upward to SRS_1 , and the long-run supply to LRS_2 . The short-run effect is to move to the intersection of the short-run supply and demand, which is at the price P_1 and the quantity Q_1 . This price is below the long-run average cost, which is the long-run supply, so over time some firms don't replace their capital and there is *disinvestment* in the industry. This disinvestment causes the short-run supply to be reduced (move left) to SRS_2 .

The case of a change in supply is more challenging because both the long-run supply and the short-run supply are shifted. But the logic – start at a long-run equilibrium, then look for the intersection of current demand and short-run supply, then look for the intersection of current demand and long-run supply – is the same whether demand or supply has shifted.

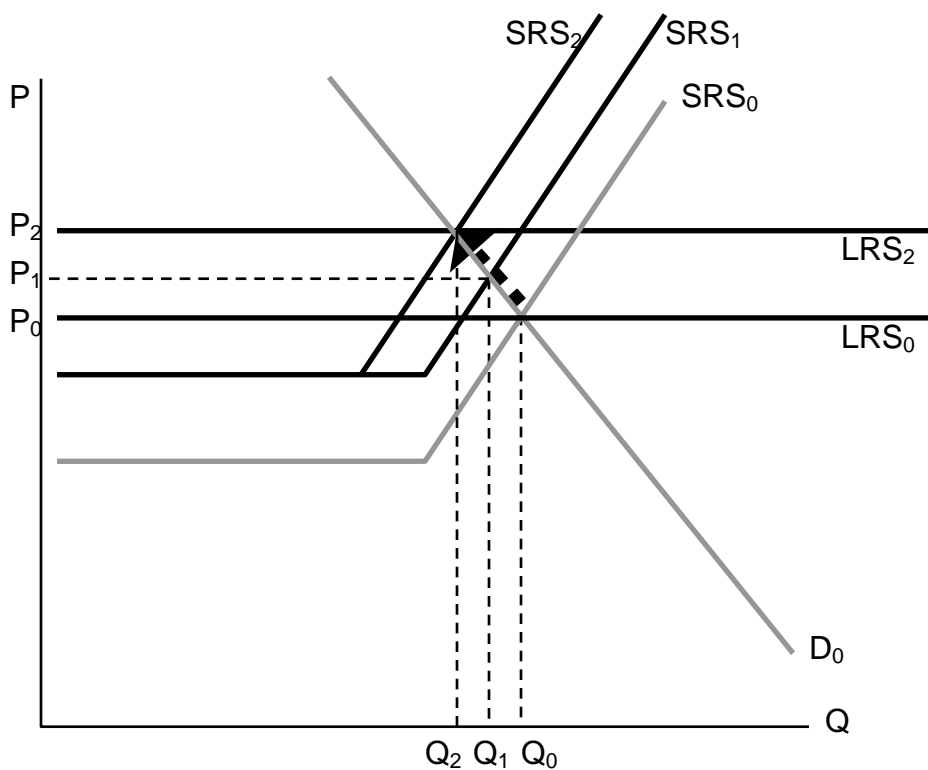


Figure 4-15: A Decrease in Supply

4.2.3 General Long-run Dynamics

The previous section made two simplifying assumptions that won't hold in all applications of the theory. First, it assumed constant returns to scale, so that long-run supply is horizontal. A perfectly elastic long-run supply means that price always eventually returns to the same point. Second, the theory didn't distinguish long-run from short-run demand. But with many products, consumers will adjust more over the long-term than immediately. As energy prices rise, consumers buy more energy-efficient cars and appliances, reducing demand. But this effect takes time to be seen, as we don't immediately scrap our cars in response to a change in the price of gasoline. The short-run effect is to drive less in response to an increase in the price, while the long-run effect is to choose the appropriate car for the price of gasoline.

To illustrate the general analysis, we start with a long-run equilibrium. Figure 4-16 reflects a long-run economy of scale, because the long-run supply slopes downward, so that larger volumes imply lower cost. The system is in long-run equilibrium because the short-run supply and demand intersection occurs at the same price and quantity as the long-run supply and demand intersection. Both short-run supply and short-run demand are less elastic than their long-run counterparts, reflecting greater substitution possibilities in the long-run.

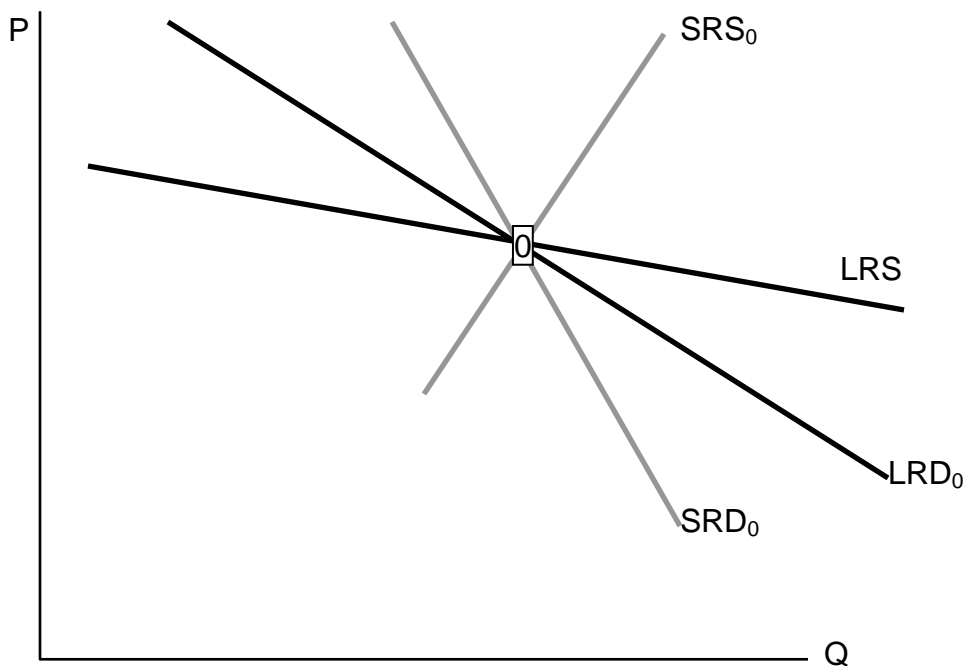


Figure 4-16: Equilibrium with External Scale Economy

Now consider a decrease in demand, decreasing both short-run and long-run demand. This is illustrated in Figure 4-17. To reduce the proliferation of curves, we color the old demand curves very faintly, and mark the initial long-run equilibrium with a zero inside a small rectangle.³² The intersection of short-run supply and short-run demand is marked with the number 1. Both long-run supply and long-run demand are more elastic than their short-run counterparts, which has an interesting effect. The short-run demand tends to shift down over time, because the price associated with the short-run equilibrium is *above* the long-run demand price for the short-run equilibrium quantity. However, the price associated with the short-run equilibrium is below the long-run supply price at that quantity. The effect is that buyers see the price as too high, and are reducing their demand, while sellers see the price as too low, and so are reducing their supply. Both short-run supply and short-run demand fall, until a long-run equilibrium is achieved.

In this case, the long-run equilibrium involves higher prices, at the point labeled 2, because of the economy of scale in supply. This economy of scale means that the reduction in demand causes prices to rise over the long-run. The short-run supply and demand eventually adjust to bring the system into long-run equilibrium, as Figure 4-18 illustrates. The new long-run equilibrium has short-run demand and supply curves associated with it, and the system is in long-run equilibrium because the short-run demand and supply, which determine the current state of the system, intersect at the

³² The short-run demand and long-run demand have been shifted down by the same amount, that is, both reflect an equal reduction in value. This kind of shift might arise if, for instance, a substitute had become cheaper, but the equal reduction is not essential to the theory. In addition, the fact of equal reductions often isn't apparent from the diagram, because of the different slopes – to most observers, it appears that short-run demand fell less than long-run demand. This isn't correct, however, and one can see this because the intersection of the new short-run demand and long-run demand occurs directly below the intersection of the old curves, implying both fell by equal amounts.

same point as the long-run demand and supply, which determine where the system is heading.

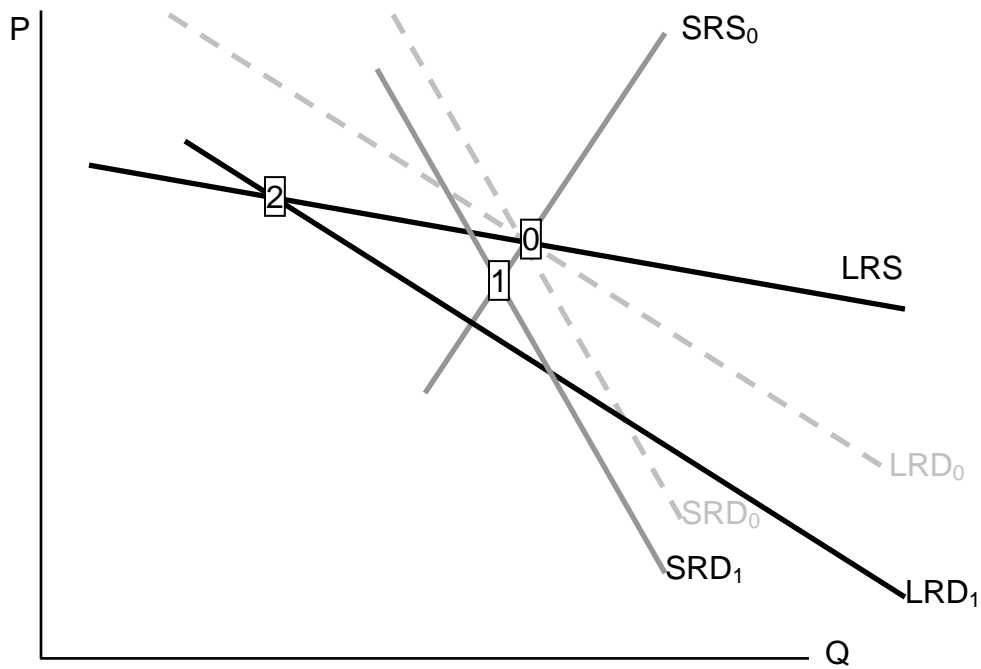


Figure 4-17: Decrease in Demand

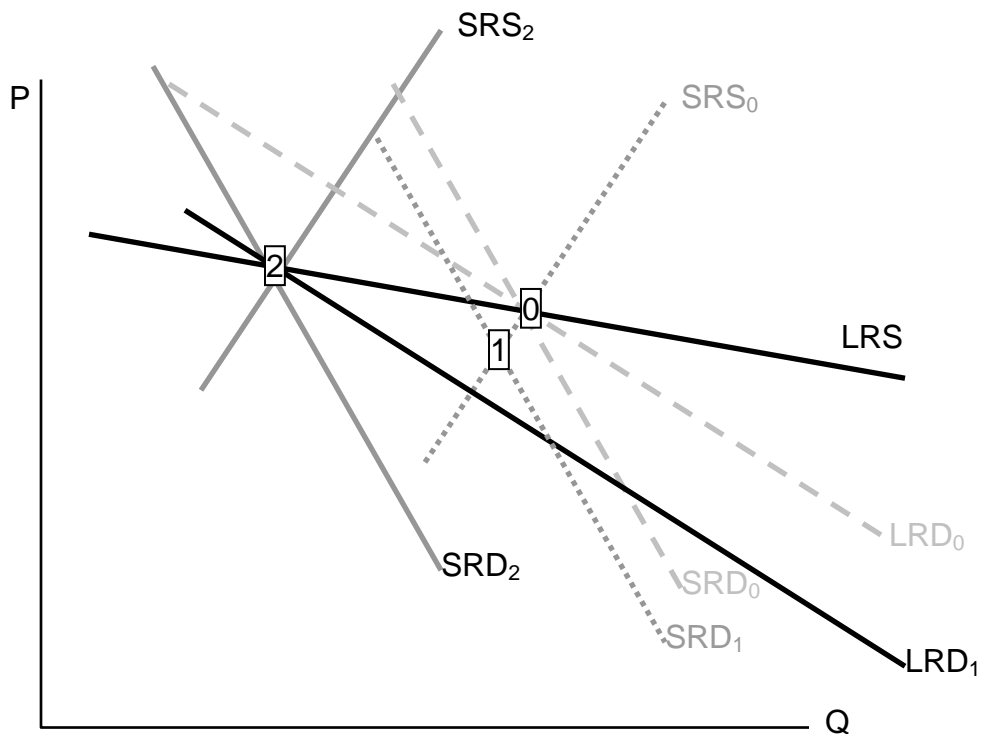


Figure 4-18: Long-run After a Decrease in Demand

There are four basic permutations of the dynamic analysis – demand increase or decrease, and a supply increase or decrease. Generally, it is possible for long-run supply to slope down – this is the case of an economy of scale – and for long-run demand to slope up.³³ This gives sixteen variations of the basic analysis. In all sixteen cases, the procedure is the same. Start with a long-run equilibrium, shift both the short-run and long-run levels of either demand or supply. The first stage is the intersection of the short-run curves. The system will then go to the intersection of the long-run curves.

An interesting example of competitive dynamics concepts is the computer memory market, which was discussed above. Most of the costs of manufacturing computer memory are fixed costs. The modern DRAM plant costs several billion dollars; the cost of other inputs – chemicals, energy, labor, silicon wafers – are modest in comparison. Consequently, the short-run supply is vertical until prices are very, very low; at any realistic price, it is optimal to run these plants 100% of the time.³⁴ The nature of the technology has let manufacturers cut the costs of memory by about 30% per year over the past forty years, demonstrating that there is a strong economy of scale in production. These two features – vertical short-run supply, strong economies of scale – are illustrated in the Figure 4-19. The system is started at the point labeled with the number 0, with a relatively high price, and technology which has made costs lower than this price. Responding to the profitability of DRAM, short-run supply shifts out (new plants are built and die-shrinks permits increasing output from existing plants). The increased output causes prices to fall, relatively dramatically because short-run demand is inelastic, and the system moves to the point labeled 1. The fall in profitability causes DRAM investment to slow, which lets demand catch up, boosting prices to the point labeled 2. (One should probably think of Figure 4-19 as being in a logarithmic scale.)

The point labeled with the number 2 looks qualitatively similar to the point labeled 1. The prices have followed a “saw-tooth” pattern, and the reason is due to the relatively slow adjustment of demand compared to supply, as well as the inelasticity of short-run demand, which creates great price swings as short-run supply shifts out. Supply can be increased quickly, and is increased “in lumps” because a die-shrink (making the chips smaller so that more fit on a given silicon wafer) tends to increase industry production by a large factor. This process can be repeated starting at the point labeled 2. The system is marching inexorably toward a long-run equilibrium in which electronic memory is very, very cheap even by 2004 standards and used in applications that haven’t yet been considered, but the process of getting there is a wild ride, indeed. The saw-tooth pattern is illustrated in Figure 4-20, which shows DRAM industry revenues in billions of dollars from 1992 to 2003, and projections of 2004 and 2005.³⁵

³³ The demand situation analogous to an economy of scale in supply is a *network externality*, in which the addition of more users of a product increases the value of the product. Telephones are a clear example – suppose you were the only person with a phone – but other products like computer operating systems and almost anything involving adoption of a standard represent examples of network externalities. When the slope of long-run demand is greater than the slope of long-run supply, the system will tend to be inefficient, because an increase in production produces higher average value and lower average cost. This usually means there is another equilibrium at a greater level of production.

³⁴ The plants are expensive in part because they are so clean, because a single speck of dust falling on a chip ruins the chip. The Infineon DRAM plant in Virginia stopped operations only when a snow-storm prevented workers and materials from reaching the plant.

³⁵ Two distinct data sources were used, which is why there are two entries for each of 1998 and 1999.

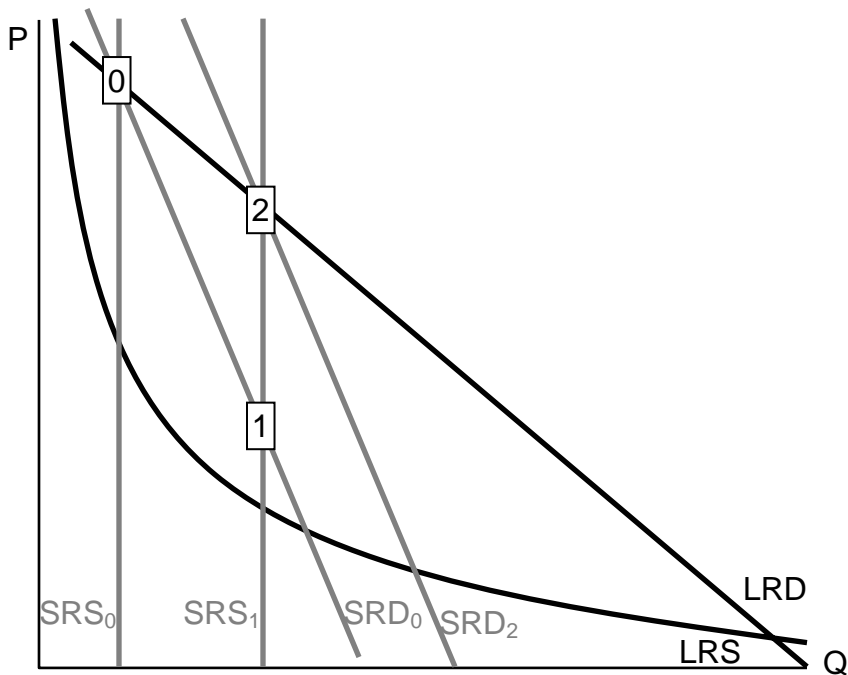


Figure 4-19: DRAM Market

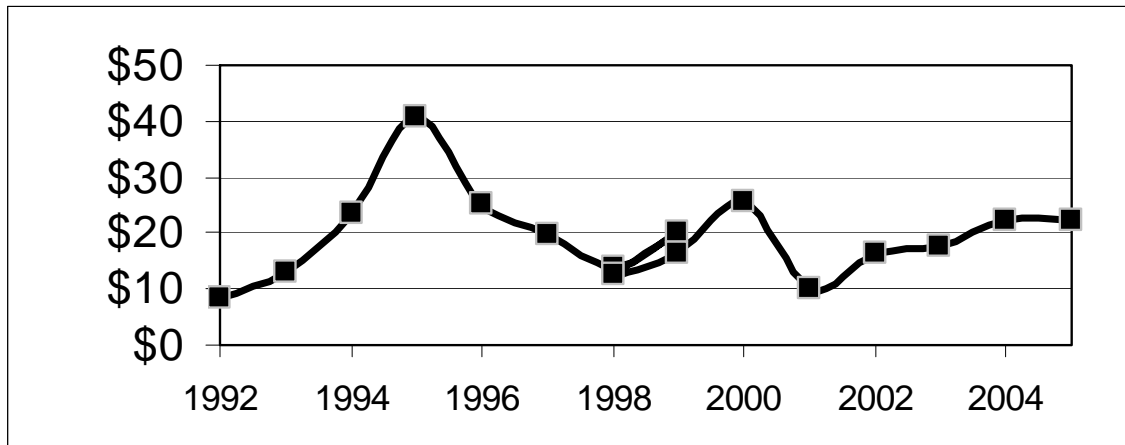


Figure 4-20: DRAM Revenue Cycle

4.2.3.1 (Exercise) Land close to the center of a city is in fixed supply, but it can be used more intensively by using taller buildings. When the population of a city increases, illustrate the long- and short-run effects on the housing markets using a graph.

4.2.3.2 (Exercise) Emus can be raised on a wide variety of ranch land, so that there are constant returns to scale in the production of emus in the long-run. In the short-run, however, the population of emus is limited by the number of breeding pairs of emus and the supply is essentially vertical. Illustrate the long- and short-run effects of an increase in demand for emus. (In the late 1980s,

there was a speculative bubble in emus, with prices reaching \$80,000 per breeding pair, in contrast to \$2,000 or so today.)

4.2.3.3 (Exercise) There are long-run economies of scale in the manufacture of computers and their components. There was a shift in demand away from desktop computers and toward notebook computers around the year 2001. What are the short- and long-run effects? Illustrate your answer with two diagrams, one for the notebook market and one for the desktop market. Account for the fact that the two products are substitutes, so that if the price of notebook computers rises, some consumers shift to desktops. (To answer this question, start with a time 0 and a market in long-run equilibrium. Shift demand for notebooks out and demand for desktops in. What happens in the short-run? What happens in the long-run to the prices of each? What does that price effect do to demand for each?)

4.3 Investment

The distinction between the short-run supply and the long-run supply is governed by the time that investment takes. Some of the difference between the short-run demand and the long-run demand arises because we don't scrap capital goods – cars, fridges, and air conditioners – in response to price changes. In both cases, investment is an important component of the responsiveness of supply and demand. In this section, we take a first look at investment. We will take a second look at investment from a somewhat different perspective later when we consider basic finance tools near the end of the book. Investment goods require expenditures today to produce future value, so we begin the analysis by examining the value of future payments.

4.3.1 Present value

The promise of \$1 in the future is not worth \$1 today. There are a variety of reasons why a promise of future payments is not worth the face value today, some of which involve risk that the money may not be paid. Let's set aside such risk for the moment; we'll consider risk separately later. Even when the future payment is perceived to occur with negligible risk, nevertheless most people prefer \$1 today to \$1 payable a year hence. One way of expressing this is that the *present value* – the value today – of a future payment of a dollar is less than a dollar. From a present value perspective, future payments are *discounted*.

From the individual perspective, one reason that you should value a future payment less than a current payment is due to *arbitrage*.³⁶ Suppose you are going to need \$10,000 one year from now, to put a down-payment on a house. One way of producing \$10,000 is to buy a government bond that pays \$10,000 a year from now. What will that bond cost you? At current interest rates, a secure bond³⁷ will cost around \$9700. This means

³⁶ Arbitrage is the process of buying and selling in such a way to make a profit. For example, if wheat is selling for \$3 per bushel in New York, but \$2.50 per bushel in Chicago, one can buy in Chicago and sell in New York and profit by \$0.50 per bushel, minus any transaction and transportation cost. Such arbitrage tends to force prices to differ by no more than transaction costs. When these transaction costs are small, as with gold, prices will be about the same worldwide.

³⁷ Economists tend to consider US federal government securities secure, because the probability of such a default is very, very low.

that no one should be willing to pay \$10,000 for a future payment of \$10,000, because instead one can have the future \$10,000, by buying the bond, and have \$300 left over to spend on cappuccinos or economics textbooks. In other words, if you will pay \$10,000 for a secure promise to repay the \$10,000 a year hence, then I can make a successful business selling you the secure promise for \$10,000, and pocketing \$300.

This arbitrage consideration also suggests how to value future payments: discount them by the relevant interest rate.

Example (Auto loan): You are buying a \$20,000 car, and you are offered the choice to pay it all today in cash, or to pay \$21,000 in one year. Should you pay cash (assuming you have that much in cash) or take the loan? The loan is at a 5% annual interest rate, because the repayment is 5% higher than the loan amount. This is a good deal for you if your alternative is to borrow money at a higher interest rate, e.g. on (most) credit cards. It is also a good deal if you have savings that pay more than 5% -- if buying the car with cash entails cashing in a certificate of deposit that pays more than 5%, then you would be losing the difference. If, on the other hand, you are currently saving money that pays less than 5% interest, paying off the car is a better deal.

The formula for present value is to discount by the amount of interest. Let's denote the interest rate for the next year as r_1 , the second year's rate as r_2 , and so on. In this notation, a \$1 invested would pay $\$1+r_1$ next year, or $\$(1+r_1)\times(1+r_2)$ after 2 years, or $\$(1+r_1)\times(1+r_2)\times(1+r_3)$ after 3 years. That is, r_i is the interest rate that determines the value, at the end of year i , of \$1 invested at the start of year i . Then, if we obtain a stream of payments A_0 immediately, A_1 at the end of year 1, A_2 at the end of year 2, and so on, the present value of that stream is

$$PV = A_0 + \frac{A_1}{1+r_1} + \frac{A_2}{(1+r_1)(1+r_2)} + \frac{A_3}{(1+r_1)(1+r_2)(1+r_3)} + \dots$$

Example (Consolidated annuities or *Consols*): What is the value of \$1 paid at the end of each year forever, with a fixed interest rate r ? Suppose the value is v . Then³⁸

$$v = \frac{1}{1+r} + \frac{1}{(1+r)^2} + \frac{1}{(1+r)^3} + \dots = \frac{1}{1-\frac{1}{1+r}} - 1 = \frac{1}{r}.$$

At a 5% interest rate, \$1 million per year paid forever is worth \$20 million today. Bonds that pay a fixed amount every year forever are known as consols; no current government issues consols.

Example (Mortgages): Again, fix an interest rate r , but this time let r be the monthly interest rate. A mortgage implies a fixed payment per month for a large number of

³⁸ This development uses the formula, for $-1 < a < 1$, that $\frac{1}{1-a} = 1 + a + a^2 + \dots$ which is readily verified.

Note that this formula involves an infinite series.

months (e.g. 360 for a 30 year mortgage). What is the present value of these payments over n months? A simple way to compute this is to use the consol value, because

$$\begin{aligned}
 M &= \frac{1}{1+r} + \frac{1}{(1+r)^2} + \frac{1}{(1+r)^3} + \dots + \frac{1}{(1+r)^n} = \frac{1}{r} - \frac{1}{(1+r)^{n+1}} - \frac{1}{(1+r)^{n+2}} - \frac{1}{(1+r)^{n+3}} \dots \\
 &= \frac{1}{r} - \frac{1}{(1+r)^n} \left(\frac{1}{(1+r)} + \frac{1}{(1+r)^2} + \frac{1}{(1+r)^3} \dots \right) \\
 &= \frac{1}{r} - \frac{1}{(1+r)^n} \frac{1}{r} = \frac{1}{r} \left(1 - \frac{1}{(1+r)^n} \right).
 \end{aligned}$$

Thus, at a monthly interest rate of $\frac{1}{2}\%$, paying \$1 per month for 360 months produces a present value M of $\frac{1}{.005} \left(1 - \frac{1}{(1.005)^{360}} \right) = 166.79$. Thus, to borrow \$100,000, one would have to pay $\frac{\$100,000}{166.79} = \599.55 per month. It is important to remember that a different loan amount just changes the scale; borrowing \$150,000 requires a payment of $\frac{\$150,000}{166.79} = \899.33 per month, because \$1 per month generates \$166.79 in present value.

Example (Simple and Compound Interest): In the days before calculators, it was a challenge to actually solve interest rate formulas, so certain simplifications were made. One of these was “simple” interest, which means that daily or monthly rates are translated into annual rates by incorrect formulas. For example, with an annual rate of 5%, the simple interest daily rate is $\frac{5\%}{365} = .07692\%$. That this is incorrect can be seen

from the calculation that $\left(1 + \frac{.05}{365} \right)^{365} = 1.051267\%$. Simple interest increases the

annual rate, so it benefits lenders and harms borrowers. (Consequently, banks advertise the accurate annual rate on savings accounts – when consumers like the number to be larger – and not on mortgages, although banks are required by law to disclose – but not to advertise widely – actual annual interest on mortgages.)

Obligatory Lottery Example: You win the lottery, and the paper reports you’ve won \$20 million. You’re going to be paid \$20 million, but is it worth \$20 million? In fact, you get \$1 million per year for 20 years. However, in contrast to our formula, you get the first million right off the bat, so the value is

$$PV = 1 + \frac{1}{1+r} + \frac{1}{(1+r)^2} + \frac{1}{(1+r)^3} + \dots + \frac{1}{(1+r)^{19}} = 1 + \frac{1}{r} \left(1 - \frac{1}{(1+r)^{19}} \right).$$

Table 3.1 computes the present value of our \$20 million dollar lottery, listing the results in thousands of dollars, at various interest rates. At ten percent interest, the value of the lottery is less than half the “number of dollars” paid, and even at 5%, the value of the stream of payments is 65% of the face value.

r	3%	4%	5%	6%	7%	10%
PV (000s)	\$15,324	\$14,134	\$13,085	\$12,158	\$11,336	\$9,365

The lottery example shows that interest rates have a dramatic impact on the value of payments made in the distant future. Present value analysis is the number one tool used in MBA programs, where it is known as Net Present Value or NPV analysis. It is accurate to say that the majority of corporate investment decisions are guided by an NPV analysis.

Example (Bond prices): A standard *treasury bill* has a fixed future value. For example it may pay \$10,000 in one year. It is sold at a discount off the face value, so that a one-year \$10,000 bond might sell for \$9,615.39, producing a 4% interest rate. To compute the effective interest rate r , the formula relating the future value FV , the number of years n , and the price is

$$(1+r)^n = \frac{FV}{\text{Price}}, \text{ or } r = \left(\frac{FV}{\text{Price}} \right)^{1/n} - 1.$$

We can see from either formula that treasury bill prices move inversely to interest rates – an increase in interest rates reduces treasury prices. Bonds are a bit more complicated. Bonds pay a fixed interest rate set at the time of issue during the life of the bond, generally collected semi-annually, and the face value is paid at the end of the term. These bonds were often sold on long terms, as much as 30 years. Thus, a three-year \$10,000 bond at 5% with semi-annual payments would pay \$250 at the end of each half year for three years, and pay \$10,000 at the end of the three years. The net present value, with an annual interest rate r , is

$$NPV = \frac{\$250}{(1+r)^{1/2}} + \frac{\$250}{(1+r)^1} + \frac{\$250}{(1+r)^{3/2}} + \frac{\$250}{(1+r)^2} + \frac{\$250}{(1+r)^{5/2}} + \frac{\$250}{(1+r)^3} + \frac{\$10000}{(1+r)^3}.$$

The net present value will be the price of the bond. Initially, the price of the bond should be the face value, since the interest rate is set as a market rate. The U.S. Treasury quit issuing such bonds in 2001, replacing them with bonds in which the face value is paid and then interest paid semi-annually.

4.3.1.1 (Exercise) At a 7% annual interest rate, what is the present value of \$100 paid at the end of one year, and \$200 paid at the end of the second year?

4.3.1.2 (Exercise) Compute the NPV of the 3 year, \$10,000 bond, with \$250 payments semi-annually, that was described above, at an interest rate of 4%.

4.3.1.3 (Exercise) You can finance your \$20,000 car with a straight 5% loan paid monthly over 5 years, or get one year interest free, but then pay 7% over the following four years. Which is a better deal? (Hint: In both cases, figure out the fixed monthly payments that produce a net present value equal to \$20,000.)

4.3.1.4 (Exercise) You win the lottery. At what interest rate should you accept \$7 million today over twenty annual payments of \$500,000?

4.3.2 Investment

A simple investment project involves spending an investment, I , and then reaping a return over time. If you dig a mine, drill an oil well, build an apartment building or a factory, or buy a share of stock, you spend money now, in the hopes of earning money subsequently. We will set aside the very important risk issue until the next subsection, and ask how to make the decision to invest.

The *NPV* approach involves assigning a rate of return r that is reasonable for, and specific to, the project and then computing the present value of the expected stream of payments. Since the investment is initially expended, it is counted as negative revenue. This gives an expression that looks like:

$$NPV = -I + \frac{R_1}{1+r} + \frac{R_2}{(1+r)^2} + \frac{R_3}{(1+r)^3} + \dots$$

where R_1 represents first year revenues, R_2 represents second year revenues, etc.³⁹ The investment is then made when *NPV* is positive – since this would add to the net value of the firm.

Carrying out an *NPV* analysis essentially requires two things. First, investment and revenues must be estimated. This is a challenge, especially for new products where there is no direct way of estimating demand, or with uncertain outcomes like oil wells or technological research.⁴⁰ Second, an appropriate rate of return must be identified. The rate of return is a problem, mostly because of risk associated with the payoffs to the investment, but also because of the incentives of project managers to inflate the payoffs and minimize the costs to make the project look more attractive to upper management. In addition, most investment undertaken by corporations is financed not with borrowing but with retained earnings, that is, with profits from previous activities. Thus a company that undertakes one investment can't carry out some other investment, and the interest rate has to account for the internal corporate value of funds. As a result of these factors, interest rates of 15%-20% are common for evaluating the *NPV* of projects of major corporations.

³⁹ The most common approach is to treat revenues within a year as if they are received at the midpoint, and then discount appropriately for that mid-year point. The present discussion oversimplifies in this regard.

⁴⁰ The building of the famed Sydney Opera House, which looks like billowing sails over Sydney harbor, was estimated to cost \$7 million and actually cost \$105 million. A portion of the cost overrun was due to the fact that the original design neglected to install air conditioning. When this oversight was discovered, it was too late to install a standard unit, which would interfere with the excellent acoustics, so instead an ice hockey floor was installed as a means of cooling the building.

Example (Silver Mine): A company is considering whether to develop a silver mine in Mexico. The company estimates that developing the mine (building roads and opening a large hole in the ground) would require \$4 million per year for four years and no revenues would accrue during this time. Starting in year 5 the expenses fall to \$2 million per year, and \$6 million in net revenue is earned off the mined silver for each of the subsequent 40 years. If the company values funds at 18%, should it develop the mine?

The earnings from the mine are calculated in the following table. First, the NPV of the investment phase during years 0, 1, 2, and 3 is

$$NPV = -4 + \frac{-4}{1.18} + \frac{-4}{(1.18)^2} + \frac{-4}{(1.18)^3} = -12.697.$$

A dollar earned in each of years 4 through 43 have a present value of

$$\frac{1}{(1+r)^4} + \frac{1}{(1+r)^5} + \frac{1}{(1+r)^6} + \dots + \frac{1}{(1+r)^{43}} = \frac{1}{(1+r)^3} \times \frac{1}{r} \left(1 - \frac{1}{(1+r)^{40}} \right) = 13.377$$

The mine is just profitable at 18%, in spite of the fact that its \$4 million payments are made in four years, after which point \$4 million dollar revenues are earned for forty years. The problem in the economics of mining is that 18% makes those future revenues have quite modest present values.

Year	Earnings (\$M) / yr	PV (\$M)
0-3	-4	-12.697
4-43	4	13.377
Net		0.810

There are other approaches to deciding whether to take an investment. In particular, the *Internal Rate of Return* approach solves the equation $NPV=0$ for the interest rate, and then the project is undertaken if the rate of return is sufficiently high. This approach is flawed because the equation may have more than one solution, or no solutions and it is not transparent what the right thing to do should be in these events. Indeed, the IRR approach gets the profit-maximizing answer only if it agrees with NPV. A second approach is the payback period, which asks how many years a project must be run before profitability is reached. The problem with the payback period is deciding between projects – if I can only do one of two projects, the one with the higher NPV makes the most money for the company. The one with the faster payback may make a quite small amount of money very quickly; it isn't apparent that this is a good choice. When a company is in risk of bankruptcy, a short payback period might be valuable, although this would ordinarily be handled by employing a higher interest rate in an NPV analysis. NPV does a good job when the question is whether to undertake a project or not, and it does better than other approaches to investment decisions. For this reason, NPV has become the most common approach to investment decisions. Indeed, NPV analysis is more common than all other approaches combined. NPV does a poor job,

however, when the question is whether to undertake a project, or delay the project. That is, NPV answers “yes or no” to investment, but when the choice is “yes or wait,” NPV requires amendment.

4.3.2.1 (Exercise) Suppose that, without a university education, you’ll earn \$25,000 per year. A university education costs \$20,000 per year, and you forgo the \$25,000/year you would have earned for four years. However, you earn \$50,000 per year for the following forty years. At 7%, what is the NPV of the university education?

4.3.2.2 (Exercise) Now that you’ve decided to go to university based on the previous answer, suppose that you can attend East State U, paying \$3,000 per year for four years and earning \$40,000 when you graduate, or North Private U, paying \$22,000 per year for the four years and earning \$50,000 when you graduate. Which is the better deal at 7%?

4.3.3 Investment Under Uncertainty

Risk has a cost, and people, and corporations, buy insurance against financial risk.⁴¹ The standard approach to investment under uncertainty is to compute an NPV, with the revenues composed of expected values, and the interest rate used adjusted to compensate for the risk.

For example, consider a project like oil exploration. The risks are enormous. Half of all underwater tracts in the Gulf Coast near Louisiana and Texas that are leased are never drilled, because later information makes them a bad bet. Half of all the tracts that are drilled are dry. So right off the bat, three-quarters of the tracts that are sold produce zero or negative revenue, and positive costs. To see how the economics of such a risky investment might be developed, suppose that the relevant rate of return for such investments is 18%. Suppose further the tract can be leased for \$500,000 and the initial exploration costs \$1 million. If the tract has oil (with a 25% probability), it produces \$1 million per year for twenty years, and then runs dry. This gives an expected revenue of \$250,000 per year. To compute the expected net present value, we first compute the returns:

Table 4-1: Oil Tract Return

	Expected revenue	EPV
0	-\$1.5M	-\$1.5M
1-20	\$0.25M	\$1.338M
Net		-\$0.162

At 18%, the investment is a loss – the risk is too great given the average returns.

A very important consideration for investment under uncertainty is the choice of interest rate. The most important thing to understand is that the interest rate is specific

⁴¹ For example, NBC spent \$6 million in buying an insurance policy against US nonparticipation in the 1980 Moscow summer Olympic games, and the US didn’t participate (because of the Soviet invasion of Afghanistan), and NBC was paid \$94 million from the policy.

to the project, and not to the investor. This is perhaps the most important insight of corporate finance generally: the interest rate should adjust for the risk associated with the project and not the investor. For example, suppose hamburger retailer McDonald's is considering investing in a cattle ranch in Peru. McDonald's is overall a very low-risk firm, but this particular project is quite risky, because of local conditions. McDonald's still needs to adjust for the market value of the risk it is undertaking, and that value is a function of the project risk, not the risk of McDonald's other investments.

This basic insight of corporate finance – the appropriate interest rate is determined by the project, not the investor – is counter-intuitive to most of us because it doesn't apply to our personal circumstances. For individuals, the cost of borrowing money is mostly a function of their own personal circumstances, and thus the decision of whether to pay cash for a car or borrow the money is not so much a function of the car being purchased but of the wealth of the borrower. Even so, personal investors borrow money at distinct interest rates. Mortgage rates on houses are lower than interest rates on automobiles, and interest rates on automobiles lower than on credit cards. This is because the "project" of buying a house has less risk associated for it: the percentage loss to the lender in event of borrower default is lower on a house than on a car. Credit cards carry the highest interest rates because they are unsecured by any asset.

One way of understanding why the interest rate is project-specific but not investor-specific is to think about undertaking the project by creating a separate firm to make the investment. The creation of subsidiary units is a common strategy, in fact. This subsidiary firm created to operate a project has a value equal to the NPV of the project using the interest rate specific to the subsidiary, which is the interest rate for the project, independent of the parent. For the parent company, owning such a firm is a good thing if the firm has positive value, and not otherwise.⁴²

Investments in oil are subject to another kind of uncertainty: price risk. Prices of oil fluctuate and aren't constant. Moreover, oil pumped and sold today is not available for the future. Should you develop and pump the oil you have today, or should you hold out and sell in the future? This question, known as the *option value of investment*, is generally somewhat challenging and arcane, but a simple example provides a useful insight.

To develop this example, let's set aside some extraneous issues first. Consider a very simple investment, in which either C is invested or not.⁴³ If C is invested, a value V is generated. The cost C is a constant; it could correspond to drilling or exploration costs, or in the case of a stock option, the *strike price* of the option, which is the amount one pays to obtain the share of stock. The value V , in contrast, varies from time to time in a random fashion. To simplify the analysis, we assume that V is uniformly distributed on the interval $[0,1]$, so that the probability of V falling in an interval $[a, b]$ is $b-a$ if $0 \leq a \leq b \leq 1$. The option only has value if $C < 1$, which we assume for the rest of this section.

⁴² It may seem that synergies between parent and subsidiary are being neglected here, but synergies should be accounted for at the time they produce value, i.e. as part of the stream of revenues of the subsidiary.

⁴³ This theory is developed in striking generality by Avinash Dixit and Robert Pindyck, *Investment Under Uncertainty*, Princeton University Press, 1994.

The first thing to note is that the optimal rule to make the investment is *cutoff value*, that is, to set a level V_0 and exercise the option if, and only if, $V \geq V_0$. This is because, if you are willing to exercise the option and generate value V , you should be willing to exercise the option and obtain even more value. The NPV rule simply says $V_0 = C$, that is, invest whenever it is profitable. The purpose of the example developed below is to provide some insight into how far wrong the NPV rule will be when option values are potentially significant.

Now consider the value of option to invest, given that the investment rule $V \geq V_0$ is followed. Call this option value $J(V_0)$. If the realized value V exceeds V_0 , one obtains $V - C$. Otherwise, one delays the investment, producing a discounted level of the same value. This logic says

$$J(V_0) = (1 - V_0) \left(\frac{1 + V_0}{2} - C \right) + V_0 \left(\frac{1}{1+r} J(V_0) \right).$$

This expression for $J(V_0)$ arises as follows. First, the hypothesized distribution of V is uniform on $[0,1]$. Consequently, the value of V will exceed V_0 with probability $1 - V_0$. In this event, the expected value of V is the midpoint of the interval $[V_0, 1]$, which is $\frac{1}{2}(V_0+1)$. The value $\frac{1}{2}(V_0+1) - C$ is the average payoff from the strategy of investing whenever $V \geq V_0$, which is obtained with probability $1 - V_0$. Second, with probability V_0 , the value falls below the cutoff level V_0 . In this case, no investment is made, and instead, we wait until the next period. The expected profits of the next period are $J(V_0)$ and these profits are discounted in the standard way.

The expression for J is straightforward to solve:

$$J(V_0) = \frac{(1 - V_0) \left(\frac{1 + V_0}{2} - C \right)}{1 - \frac{V_0}{1+r}}.$$

Rudimentary calculus shows

$$J'(V_0) = \frac{1 + 2rC + V_0^2 - 2(1+r)V_0}{2(1+r) \left(1 - \frac{V_0}{1+r} \right)^2}.$$

First, note that $J'(C) > 0$ and $J'(1) < 0$, which together imply the existence of a maximum at a value V_0 between C and 1, satisfying $J'(V_0) = 0$. Second, the solution occurs at

$$V_0 = (1+r) - \sqrt{(1+r)^2 - (1+2rC)} = (1+r) - \sqrt{r^2 + 2r(1-C)}.$$

The positive root of the quadratic has $V_0 > 1$, which entails never investing, and hence is not a maximum. The profit-maximizing investment strategy is to invest whenever the value exceeds V_0 given by the negative root in the formula. There are a couple of notable features about this solution. First, at $r=0$, $V_0 = 1$. This is because $r=0$ corresponds to no discounting, so there is no loss in holding out for the highest possible value. Second, as $r \rightarrow \infty$, $V_0 \rightarrow C$. As $r \rightarrow \infty$, the future is valueless, so it is worth investing if the return is anything over costs. These are not surprising findings, quite the opposite – they should hold in any reasonable formulation of such an investment strategy. Moreover, they show that the NPV rule, which requires $V_0 = C$, is correct only if the future is valueless.

How does this solution behave? The solution is plotted as a function of r , for $C=0, 0.25$ and 0.5 , in Figure 4-21.

The horizontal axis represents interest rates, so this picture shows *very* high interest rates by current standards, up to 200%. Even so, V_0 remains substantially above C . That is, even when the future has very little value because two-thirds of the value is destroyed by discounting each period, the optimal strategy deviates significantly from the NPV strategy. Figure 4-22 shows a close-up of that picture for a more reasonable range of interest rates, for interest rates of zero to ten percent

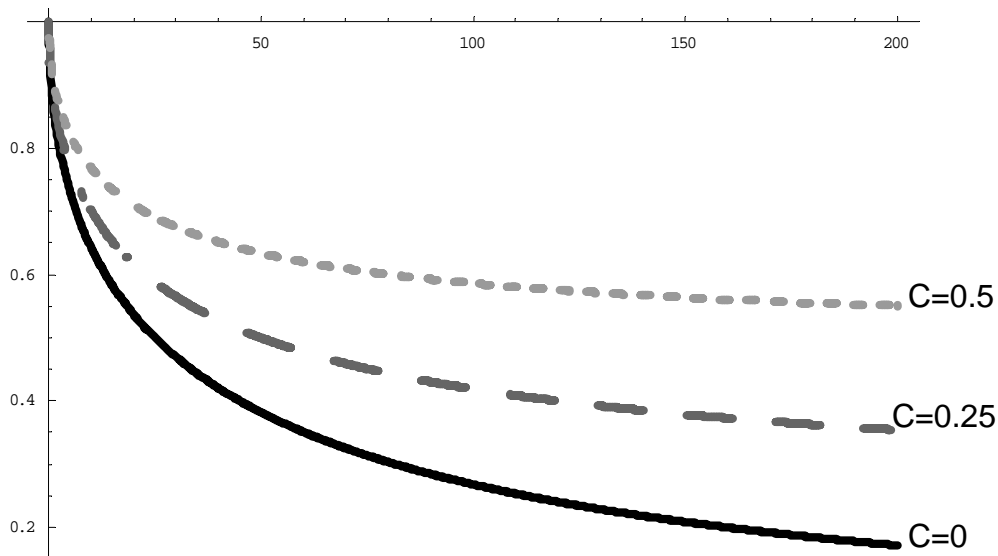


Figure 4-21: Investment Strike Price Given Interest Rate r in Percent

Figure 4-22 shows the cutoff values of investment for three values of C , the cost of the investment. These three values are 0 (lowest curve), 0.25 (the middle dashed curve), and 0.5, the highest, dotted line. Consider the lowest curve, with $C=0$. The NPV of this project is *always* positive – there are no costs and revenues are positive. Nevertheless, because the investment can only be made once, it pays to hold out for a higher level of payoff, indeed, for 65% or more of the maximum payoff. The economics at an interest rate of 10% is as follows. By waiting, there is a 65% chance that ten percent of the potential value of the investment is lost. However, there is a 35% of an even higher value. The optimum value of V_0 trades these considerations off against each other.

For $C = 0.25$, at 10% the cutoff value for taking an investment is 0.7, nearly three times the actual cost of the investment. Indeed, the cutoff value incorporates two separate costs: the actual expenditure on the investment C , and the lost opportunity to invest in the future. The latter cost is much larger than the expenditure on the investment in many circumstances, and in this example, can be quantitatively much larger than the actual expenditure on the investment.

Some investments can be replicated. There are over 13,000 McDonald's restaurants in the United States, and building another doesn't foreclose building even more. For such investments, NPV analysis gets the right answer, provided that appropriate interest rates and expectations are used. Other investments are difficult to replicate or logically impossible to replicate – having pumped and sold the oil from a tract, that tract is now dry. For such investments, NPV is consistently wrong because it neglects the value of the option to delay the investment. A correct analysis adds a lost value for the option to delay the cost of the investment, a value which can be quantitatively large, as we have seen.

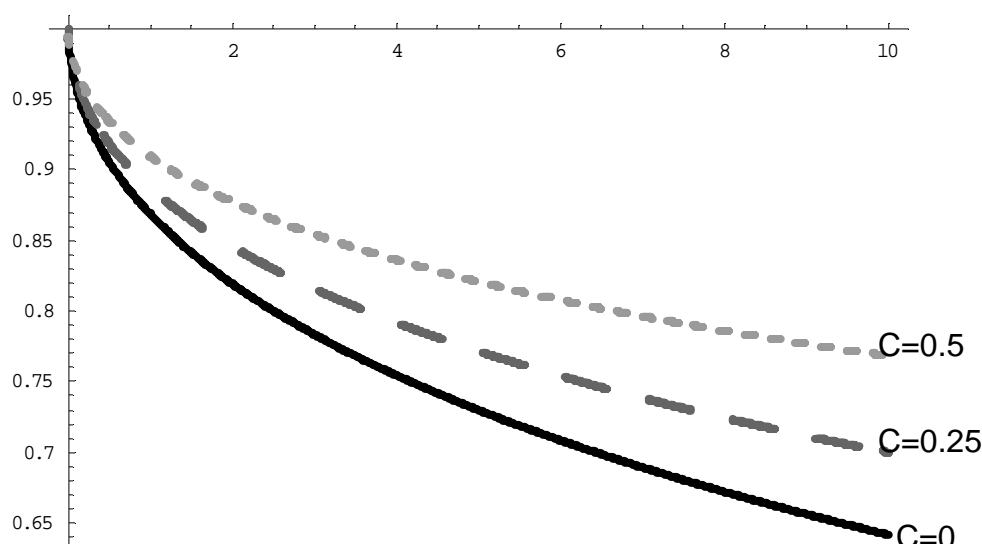


Figure 4-22 Investment Strike Price Given Interest Rate r in Percent

Example: When should you refinance a mortgage? Suppose you are paying 10% on a \$100,000 mortgage, and it costs \$5,000 to refinance, but refinancing permits you to lock in a lower interest rate, and hence pay less. When is it a good idea? To answer this question, we assume that the \$5,000 cost of refinancing is built into the loan, so that in essence you borrow \$105,000 at a lower interest rate when you refinance. This is actually the most common method of refinancing a mortgage.

To simplify the calculations, we will consider a mortgage that is never paid off, that is, one pays the same amount per year forever. If the mortgage isn't refinanced, one pays ten percent of the \$100,000 face value of the mortgage each year, or \$10,000 per year. If one refinances at interest rate r , one pays $r \times \$105,000$ per year, so the NPV of refinancing is

$$\text{NPV} = \$10,000 - r \times \$105,000.$$

Thus NPV is positive whenever $r < \frac{10}{105} = 9.52\%$.

Should you refinance when the interest rate drops to this level? No. At that level, you would exactly break even, but would also be carrying a \$105,000 mortgage rather than a \$100,000 mortgage, making it harder to benefit from any further interest rate decreases. The only circumstance in which refinancing at 9.52% is sensible is if interest rates can't possibly fall further.

When should you refinance? That depends on the nature and magnitude of the randomness governing interest rates, preferences over money today versus money in the future, and attitudes to risk. The model developed in this section is not a good guide to answering this question, primarily because the interest rates are strongly correlated over time. However, an approximate guide to implementing the option theory of investment is to seek an NPV of twice the investment, which would translate into a refinance point of around 8.5%.

4.3.3.1 (Exercise) You are searching for a job. The net value of jobs that arise is uniformly distributed on the interval $[0,1]$. When you accept a job, you must stop looking at subsequent jobs. If you can interview with one employer per week, what jobs should you accept? Use a 7% annual interest rate.

Hint: Relate the job search problem to the investment problem, where accepting a job is equivalent to making the investment. What is c in the job search problem? What is the appropriate interest rate?

4.3.4 Resource Extraction

For the past sixty years, the world has been “running out of oil.” There are news stories about the end of the reserves being only ten, fifteen or twenty years away. The tone of these stories is that, at that time, we will run out of oil completely and prices will be extraordinarily high. Industry studies counter that more oil continues to be found and that the world is in no danger of running out of oil.

If you believe that the world will run out of oil, what should you do? You should *buy and hold*. That is, if the price of oil in twenty years is going to be \$1,000 per barrel, then you can buy oil at \$40 and hold it for twenty years, and sell it at \$1,000. The rate of return from this behavior is the solution to

$$(1 + r)^{20} = \frac{1000}{40}.$$

This equation solves for $r = 17.46\%$, which represents a healthy rate of return on investment. This substitution is part of a general conclusion known as the *Ramsey*⁴⁴

⁴⁴ The solution to this problem is known as *Ramsey pricing*, after the discoverer Frank Ramsey (1903-1930).

rule: for resources in fixed supply, prices rise at the interest rate. With a resource in fixed supply, owners of the resource will sell at the point maximizing the present value of the resource. Even if they do not, others can buy the resource at the low present value of price point and resell at the high present value, and make money.

The Ramsey rule implies that prices of resources in fixed supply rise at the interest rate. An example of the Ramsey rule in action concerns commodities that are temporarily fixed in supply, such as grains, after the harvest. During the period between harvests, these products rise in price on average at the interest rate, where the interest rate includes storage and insurance costs, as well as the cost of funds.

Example: Let time run $t = 0, 1, \dots$ and suppose the demand for a resource in fixed supply has constant elasticity: $p(Q) = aQ^{-1/\varepsilon}$. Suppose there is a total stock R of the resource, and the interest rate is fixed at r . What is the price and consumption of the resource at each time?

Solution: Let Q_t represent the quantity consumed at time t . Then the arbitrage condition requires:

$$aQ_0^{-1/\varepsilon}(1+r)^t = p(Q_0)(1+r)^t = p(Q_t) = aQ_t^{-1/\varepsilon}.$$

Thus, $Q_t = Q_0(1+r)^{-t\varepsilon}$.

Finally, the resource constraint implies

$$R = (Q_0 + Q_1 + Q_2 + \dots) = Q_0 \left(1 + (1+r)^{-\varepsilon} + (1+r)^{-2\varepsilon} + \dots \right) = \frac{Q_0}{1 - (1+r)^{-\varepsilon}}.$$

This solves for the initial consumption Q_0 . Consumption in future periods declines geometrically, thanks to the constant elasticity assumption.

Market arbitrage insures the availability of the resource in the future, and drives the price up to ration the good. The world runs out slowly, and the price of a resource in fixed supply rises on average at the interest rate.

Resources like oil and minerals are ostensibly in fixed supply – there is only so much oil, or gold, or bauxite, or palladium in the earth. Markets, however, behave as if there is an unlimited supply, and with good reason. People are inventive, and find substitutes. England's wood shortage of 1651 didn't result in England being cold permanently, nor was England limited to the wood it could grow as a source of heat. Instead, coal was discovered. The shortage of whale oil in the mid-nineteenth century led to the development of oil resources as a replacement. If markets expect that price increases will lead to substitutes, then we rationally should use more today, trusting that

technological developments will provide substitutes.⁴⁵ Thus, while some believe we are running out of oil, most investors are betting that we are not, and that energy will not be very expensive in the future, either because of continued discovery of oil, or because of the creation of alternative energy sources. If you disagree, why not invest and take the bet? If you bet on future price increases, that will tend to increase the price today, encouraging conservation today, and increase the supply in the future.

4.3.4.1 (Exercise) With an elasticity of demand of 2, compute the percentage of the resource that is used each year if the interest rate is 10%. If the interest rate falls, what happens to the proportion quantity used?

4.3.5 A Time to Harvest

A tree grows slowly, but is renewable, so the analysis of Section 4.3.4 doesn't help us understand when it is most profitable to cut the tree down. Consider harvesting for pulp and paper use. In this use, the amount of wood chips is what matters to the profitability of cutting down the tree, and the biomass of the tree provides a direct indication of this. Suppose the biomass sells for a net price p , which has the costs of harvesting and replanting deducted from it, and the biomass of the tree is $b(t)$ when the tree is t years old. It simplifies the analysis slightly to use continuous time discounting

$$e^{-\rho t} = \left(\frac{1}{1+r} \right)^t,$$

where $\rho = \log(1+r)$.

Consider the policy of cutting down trees when they are T years old. This induces a cutting cycle of length T . A brand new tree will produce a present value of profits of:

$$e^{-\rho T} pb(T) + e^{-2\rho T} pb(T) + e^{-3\rho T} pb(T) + \dots = \frac{e^{-\rho T} pb(T)}{1 - e^{-\rho T}} = \frac{pb(T)}{e^{\rho T} - 1}.$$

This profit arises because the first cut occurs at time T , with discounting $e^{-\rho T}$, and produces a net gain of $pb(T)$. The process then starts over, with a second tree cut down at time $2T$, and so on.

Profit maximization gives a first order condition on the optimal cycle length T of

$$0 = \frac{d}{dT} \frac{pb(T)}{e^{\rho T} - 1} = \frac{pb'(T)}{e^{\rho T} - 1} - \frac{pb(T)\rho e^{\rho T}}{(e^{\rho T} - 1)^2}.$$

This can be rearranged to yield:

⁴⁵ Unlike oil and trees, whales were overfished and there was no mechanism for arbitraging them into the future, that is, no mechanism for capturing and saving whales for later use. This problem, known as the tragedy of the commons, results in too much use and is taken up in Section 6.3.6. Trees have also been over-cut, most notably on Easter Island.

$$\frac{b'(T)}{b(T)} = \frac{\rho}{1 - e^{-\rho T}}.$$

The left hand side of this equation is the growth rate of the tree. The right hand side is approximately the continuous-time discount factor, at least when T is large, as it tends to be for trees, which are usually on a 20 to 80 year cycle, depending on the species. This is the basis for a conclusion: cut down the tree slightly before it is growing at the interest rate. The higher are interest rates, the shorter the cycle on which the trees should be cut down.

The pulp and paper use of trees is special, because the tree is going to be ground up into wood chips. What happens when the object is to get boards from the tree, and larger boards sell for more? In particular, it is more profitable to get a 4×4 than two 2×4 s. Doubling the diameter of the tree, which approximately raises the biomass by a factor of six to eight, more than increases the value of the timber by the increase in the biomass.

It turns out our theory is already capable of handling this case. The only adaptation is a change in the interpretation of the function b . Now, rather than representing the biomass, $b(t)$ must represent the value in boards of a tree that is t years old. (The parameter p may be set to one.) The only amendment to the rule for cutting down trees is that the most profitable point in time to cut down the tree occurs slightly before the time when the value (in boards) of the tree is growing at the interest rate.

For example, lobsters become more valuable as they grow; the profit-maximizing time to harvest lobsters is governed by the same equation, where $b(T)$ is the value of a lobster of age T . Prohibiting the harvest of lobsters under age T is a means of insuring the profit-maximizing capture of lobsters, and preventing over-fishing, a topic considered in section 6.3.6.

The implementation of the formula is illustrated in Figure 4-23. The dashed line represents the growth rate $b'(T)/b(T)$, while the solid line represents the discount rate, which was set at 5%. Note that the best time to cut down the trees is when they are approximately 28.7 years old, and at that time, they are growing at 6 1/2 %. Figure 4-23 also illustrates another feature of the optimization – there may be multiple solutions to the optimization problem, and the profit-maximizing solution involves $b'(T)/b(T)$

cutting $\frac{\rho}{1 - e^{-\rho T}}$ from above.

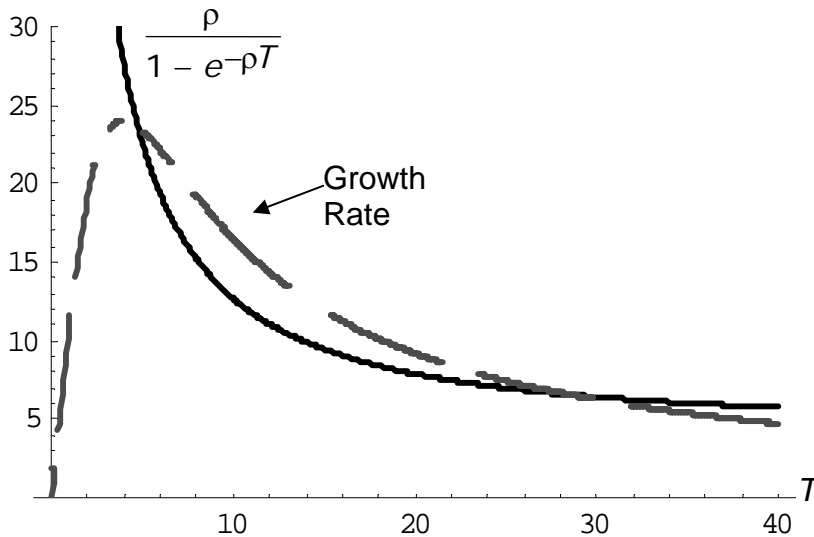


Figure 4-23: Optimal Solution for T

The U.S. Department of the Interior is in charge of selling timber rights on federal lands. The Department uses the policy of *maximum sustainable yield* to determine the time that the tree is cut down. Maximum sustainable yield maximizes the long-run average value of the trees cut down, that is, it maximizes

$$\frac{b(T)}{T}.$$

4.3.5.1 (Exercise) Show maximum sustainable yield results in cutting down the tree when it is T years old, where T satisfies

$$\frac{b'(T)}{b(T)} = \frac{1}{T}.$$

Maximum sustainable yield is actually a special case of the policies considered here, and arises for a discount factor of 0. It turns out (thanks to a formula known variously as L'Hôpital's or L'Hospital's rule) that

$$\lim_{\rho \rightarrow 0} \frac{\rho}{1 - e^{-\rho T}} = \frac{1}{T}.$$

Thus, the rule $\frac{b'(T)}{b(T)} = \frac{\rho}{1 - e^{-\rho T}} \rightarrow \frac{1}{T}$ as $\rho \rightarrow 0$, and this is precisely the same rule that arises under maximum sustainable yield.

Thus, the Department of the Interior acts as if the interest rate is zero, when it is not. The justification given is that the Department is valuing future generations at the same level as current generations, that is, increasing the supply for future generations, while slightly harming the current generation of buyers. The major consequence of the

Department's policy of maximum sustainable yield is to force cutting of timber even when prices are low during recessions.

4.3.5.2 (Exercise) Suppose the growth rate of trees satisfies $\frac{b'(T)}{b(T)} = te^{-t}$. Numerically

approximate the efficient time to cut the tree if $\rho=0.1$. How does this compare to the solution of maximum sustainable yield?

4.3.6 Collectibles

Many people purchase durable goods as investments, including Porsche Speedsters, Tiffany lamps, antique telephones, postage stamps and coins, baseball cards, original Barbie dolls, antique credenzas, autographs, original rayon Hawaiian shirts, old postcards, political campaign buttons, old clocks and even Pez dispensers. How is the value of, say, a 1961 Porsche Speedster or a \$500 bill from the confederacy, which currently sells for over \$500, determined?



Figure 4-24: The Porsche Speedster

The theory of resource prices can be adapted to cover these items, which are in fixed supply. There are four major differences that are relevant. First, using the item doesn't consume it; the goods are durable. I can own an "I Like Ike" campaign button for years, then sell the same button. Second, these items may depreciate. Cars wear out even when they aren't driven, and the brilliant color of Pez dispensers fades. Every time a standard 27 ½ pound gold bar, like the kind in the Fort Knox depository, is moved, approximately \$5 in gold wears off the bar. Third, the goods may cost something to store. Fourth, the population grows, and some of the potential buyers are not yet born.

To understand the determinants of the prices of collectibles, it turns out to create a major simplification to perform the analysis in continuous time. Let t , ranging from zero to infinity, be the continuous time variable. If the good depreciates at rate δ , and q_0 is the amount available at time 0, the quantity available at time t is

$$q(t) = q_0 e^{-\delta t}.$$

For simplicity, assume that there is constant elasticity of demand ε . If g is the population growth rate, the quantity demanded, for any price p , is given by

$$x_d(p, t) = ae^{gt} p^{-\varepsilon},$$

for a constant a which represents the demand at time 0. This represents demand for the good for direct use, but neglects the investment value of the good – that the good can be resold for a higher price later. In other words, x_d captures the demand for looking at Pez dispensers or driving Porsche Speedsters, but does not incorporate the value of being able to resell these items.

The demand equation can be used to generate the lowest use value to a person owning the good at time t . That marginal use value v arises from the equality of supply and demand:

$$q_0 e^{-\delta t} = q(t) = x_d(v, t) = ae^{gt} v^{-\varepsilon}$$

or

$$v^\varepsilon = \frac{a}{q_0} e^{(\delta+g)t}.$$

Thus, the use value to the marginal owner of the good at time t satisfies

$$v = \left(\frac{a}{q_0} \right)^{1/\varepsilon} e^{\frac{\delta+g}{\varepsilon} t}.$$

An important aspect of this development is that the value to the owner is found without reference to the price of the good. The reason this calculation is possible is that the individuals with high values will own the good, and the number of goods and the values of people are assumptions of the theory. Essentially, we already know that the price will ration the good to the individuals with high values, so computing the lowest value individual who holds a good at time t is a straightforward “supply equals demand” calculation. Two factors increase the marginal value to the owner – there are fewer units available because of depreciation, and there are more high-value people demanding them, because of population growth. Together, these factors make the marginal use value grow at the rate $\frac{\delta+g}{\varepsilon}$.

Assume that s is the cost of storage per unit of time and per unit of the good, so that storing x units for a period of length Δ costs $sx\Delta$. This is about the simplest possible storage cost technology.

The final assumption that we make is that all potential buyers use a common discount rate r , so that the discount of money or value received Δ units of time in the future is $e^{-r\Delta}$. It is worth a brief digression why it is sensible to assume a common discount rate, when it is evident that many people have different discount rates. Different discount rates induce gains from trade in borrowing and lending, and create an incentive to have

banks. While banking is an interesting thing to study, this section is concerned with collectibles, not banks. If we have different discount factors, then we must also introduce banks, which would complicate the model substantially. Otherwise, we would intermingle the theory of banking and the theory of collectibles. It is probably a good idea to develop a joint theory of banking and collectibles given the investment potential of collectibles, but it is better to start with the pure theory of either one before developing the joint theory.

Consider a person who values the collectible at v . Is it a good thing for this person to own a unit of the good at time t ? Let p be the function that gives the price across time, so that $p(t)$ is the price at time t . Buying the good at time t and then selling what remains (recall that the good depreciates at rate δ) at time $t+\Delta$ gives a net value of

$$\int_0^{\Delta} e^{-ru}(v-s)du - p(t) + e^{-r\Delta} e^{-\delta\Delta} p(t+\Delta).$$

For the marginal person, that is, the person who is just indifferent to buying or not buying at time t , this must be zero at every moment in time, for $\Delta=0$. If v represents the value to a marginal buyer (indifferent to holding or selling) holding the good at time t , then this expression should come out to be zero. Thus, dividing by Δ ,

$$\begin{aligned} 0 &= \lim_{\Delta \rightarrow 0} \frac{1}{\Delta} \int_0^{\Delta} e^{-ru}(v-s)du - \frac{p(t)}{\Delta} + \frac{e^{-(r+\delta)\Delta} p(t+\Delta)}{\Delta} \\ &= \lim_{\Delta \rightarrow 0} v-s + \frac{p(t+\Delta) - p(t)}{\Delta} - \frac{1 - e^{-(r+\delta)\Delta}}{\Delta} p(t+\Delta) = v-s + p'(t) - (r+\delta)p(t). \end{aligned}$$

Recall that the marginal value is $v = \left(\frac{a}{q_0}\right)^{1/\varepsilon} e^{\frac{\delta+g}{\varepsilon}t}$, which gives

$$p'(t) = (r+\delta)p(t) + s - v = (r+\delta)p(t) + s - \left(\frac{a}{q_0}\right)^{1/\varepsilon} e^{\frac{\delta+g}{\varepsilon}t}.$$

The general solution to this differential equation is

$$p(t) = e^{(r+\delta)t} \left(p(0) + \frac{1 - e^{-(r+\delta)t}}{(r+\delta)} s - \left(\frac{a}{q_0}\right)^{1/\varepsilon} \frac{1 - e^{-\left(r+\delta - \frac{\delta+g}{\varepsilon}\right)t}}{r+\delta - \frac{\delta+g}{\varepsilon}} \right).$$

It turns out that this equation only makes sense if $r + \delta - \frac{\delta + g}{\varepsilon} > 0$, for otherwise the present value of the marginal value goes to infinity, so there is no possible finite initial price. Provided demand is elastic and discounting is larger than growth rates (which is an implication of equilibrium in the credit market), this condition will be met.

What is the initial price? It must be the case that the present value of the price is finite, for otherwise the good would always be a good investment for everyone at time 0, using the “buy and hold for resale” strategy. That is,

$$\lim_{t \rightarrow \infty} e^{-rt} p(t) < \infty.$$

This condition implies

$$\lim_{t \rightarrow \infty} e^{\delta t} \left(p(0) + \frac{1 - e^{-(r+\delta)t}}{(r+\delta)} s - \left(\frac{a}{q_0} \right)^{1/\varepsilon} \frac{1 - e^{-\left(r + \delta - \frac{\delta + g}{\varepsilon} \right) t}}{r + \delta - \frac{\delta + g}{\varepsilon}} \right) < \infty$$

and thus

$$p(0) + \frac{1}{(r+\delta)} s - \left(\frac{a}{q_0} \right)^{1/\varepsilon} \frac{1}{r + \delta - \frac{\delta + g}{\varepsilon}} = 0.$$

This equation may take on two different forms. First, it may be solvable for a non-negative price, which happens if

$$p(0) = \left(\frac{a}{q_0} \right)^{1/\varepsilon} \frac{1}{r + \delta - \frac{\delta + g}{\varepsilon}} - \frac{1}{(r+\delta)} s \geq 0.$$

Second, it may require destruction of some of the endowment of the good. Destruction must happen if the quantity of the good q_0 at time 0 satisfies

$$\left(\frac{a}{q_0} \right)^{1/\varepsilon} \frac{1}{r + \delta - \frac{\delta + g}{\varepsilon}} - \frac{1}{(r+\delta)} s < 0.$$

In this case, there is too much of the good, and an amount must be destroyed to make the initial price zero. Since the initial price is zero, the good is valueless at time zero, and destruction of the good makes sense – at the current quantity, the good is too costly

to store for future profits. Enough is destroyed to insure indifference between holding the good as a collectible and destroying it. Consider, for example, the \$500 confederate bill pictured in Figure 4-25. Many of these bills were destroyed at the end of the US Civil War, when the currency became valueless, burned as a source of heat. Now, an uncirculated version retails for \$900.

The amount of the good that must be destroyed is such that the initial price is zero. As q_0 is the initial (pre-destruction) quantity, the amount at time zero after the destruction is the quantity $q(0)$ satisfying

$$0 = p(0) = \left(\frac{a}{q(0)} \right)^{1/\varepsilon} \frac{1}{r + \delta - \frac{\delta + g}{\varepsilon}} - \frac{1}{(r + \delta)} s.$$

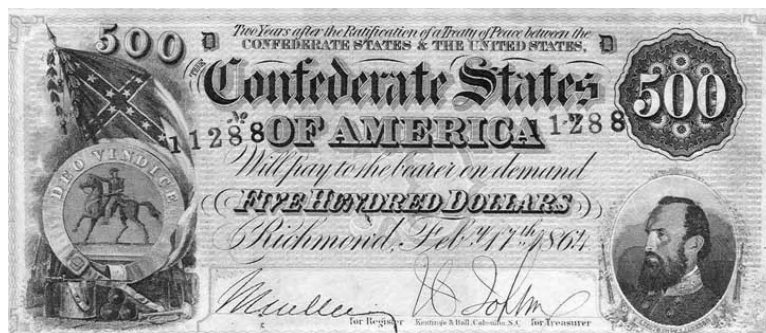


Figure 4-25: \$500 Confederate States Bill

Given this construction, we have that

$$p(0) + \frac{1}{(r + \delta)} s - \left(\frac{a}{q(0)} \right)^{1/\varepsilon} \frac{1}{r + \delta - \frac{\delta + g}{\varepsilon}} = 0,$$

where either $q(0) = q_0$ and $p(0) \geq 0$, or $q(0) < q_0$ and $p(0) = 0$.

Destruction of a portion of the stock of a collectible, followed by price increases, is actually a quite common phenomenon. In particular, consider the “Model 500” telephone by Western Electric illustrated in Figure 4-26. This ubiquitous classic phone was retired as the US switched to tone dialing and push-button phones in the 1970s, and millions of phones – perhaps over 100 million – wound up in landfills. Now, the phone is a collectible and rotary phone enthusiasts work to keep them operational.



Figure 4-26: Western Electric Model 500 Telephone

The solution for $p(0)$ dramatically simplifies the expression for $p(t)$:

$$\begin{aligned}
 p(t) &= e^{(r+\delta)t} \left(p(0) + \frac{1 - e^{-(r+\delta)t}}{(r+\delta)} s - \left(\frac{a}{q(0)} \right)^{1/\varepsilon} \frac{1 - e^{-\left(r+\delta - \frac{\delta+g}{\varepsilon}\right)t}}{r+\delta - \frac{\delta+g}{\varepsilon}} \right) \\
 &= e^{(r+\delta)t} \left(\frac{-e^{-(r+\delta)t}}{(r+\delta)} s + \left(\frac{a}{q(0)} \right)^{1/\varepsilon} \frac{e^{-\left(r+\delta - \frac{\delta+g}{\varepsilon}\right)t}}{r+\delta - \frac{\delta+g}{\varepsilon}} \right) \\
 &= \left(\frac{a}{q(0)} \right)^{1/\varepsilon} \frac{e^{\frac{\delta+g}{\varepsilon}t}}{r+\delta - \frac{\delta+g}{\varepsilon}} - \frac{s}{r+\delta}
 \end{aligned}$$

This formula lets us compare different collectibles. The first insight is that storage costs enter linearly into prices, so that growth rates are approximately unaffected by storage costs. That gold is easy to store, while stamps and art require control of humidity and temperature to preserve value and are hence more expensive to store, affects the level of prices but not the growth rate. However, depreciation and the growth of population affect the growth rate, and they do so in combination with the demand elasticity. With more elastic demand, prices grow more slowly and start at a lower level.

4.3.7 Summer Wheat

Typically, wheat harvested in the fall has to last until the following harvest. How should prices evolve over the season? If I know that I need wheat in January, should I buy it at harvest time and store it myself, or wait and buy it in January? We can use a theory analogous to the theory of collectibles developed in Section 4.3.6 to determine the evolution of prices for commodities like wheat, corn, orange juice, and canola oil.

Unlike collectibles, buyers need not hold for their personal use, since there is no value in admiring the wheat in your home. Let $p(t)$ be the price at time t and suppose that the year has length T . Generally there is a substantial amount of uncertainty regarding the size of wheat harvests and most countries maintain an excess inventory as a precaution. However, if the harvest were not uncertain, there would be no need for a precautionary holding, and instead we would consume the entire harvest over the course of a year, at which point the new harvest comes in. It is such a model that is investigated in this section.

Let δ represent the depreciation rate (which for wheat includes the quantity eaten by rodents) and s be the storage cost. Buying at time t and reselling at $t+\Delta$ should be a break-even proposition. If one purchases at time t , it costs $p(t)$ to buy the good. Reselling at $t+\Delta$, the storage cost is about $s\Delta$. (This is not the precisely relevant cost, but rather it is the present value of the storage cost, and hence the restriction to small values of Δ .) The good depreciates to only have $e^{-\delta\Delta}$ left to sell, and discounting reduces the value of that amount by the factor $e^{-r\Delta}$. For this to be a breakeven proposition, for small Δ ,

$$0 = e^{-r\Delta} e^{-\delta\Delta} p(t+\Delta) - s\Delta - p(t),$$

or

$$\frac{p(t+\Delta) - p(t)}{\Delta} = \frac{1 - e^{-(r+\delta)\Delta}}{\Delta} p(t+\Delta) - s.$$

Taking the limit as $\Delta \rightarrow 0$,

$$p'(t) = (r + \delta)p(t) - s.$$

This arbitrage condition insures that it is a break-even proposition to invest in the good; the profits from the price appreciation are exactly balanced by depreciation, interest and storage costs. We can solve the differential equation to obtain:

$$p(t) = e^{(r+\delta)t} \left(p(0) + \frac{1 - e^{-(r+\delta)t}}{r+\delta} s \right) = e^{(r+\delta)t} p(0) + \frac{e^{(r+\delta)t} - 1}{r+\delta} s.$$

The unknown is $p(0)$. The constraint on $p(0)$, however, is like the resource extraction problem – $p(0)$ is determined by the need to use up the harvest over the course of the year.

Suppose demand has constant elasticity ε . Then the quantity used comes in the form $x(t) = ap(t)^{-\varepsilon}$. Let $z(t)$ represent the stock at time t . Then the equation for the evolution of the stock is $z'(t) = -x(t) - \delta z(t)$. This equation is obtained by noting that

the flow out of stock is composed of two elements: depreciation δz and consumption x . The stock evolution equation solves for

$$z(t) = e^{-\delta t} \left(q(0) - \int_0^t e^{\delta u} x(u) du \right).$$

Thus, the quantity of wheat is consumed exactly if

$$\int_0^T e^{\delta u} x(u) du = q(0).$$

But this equation determines the initial price through

$$q(0) = \int_0^T e^{\delta u} x(u) du = \int_0^T e^{\delta u} a p(u)^{-\varepsilon} du = \int_0^T e^{\delta u} a \left(e^{(r+\delta)u} p(0) + \frac{e^{(r+\delta)u} - 1}{r + \delta} s \right)^{-\varepsilon} du$$

This equation doesn't lead to a closed form for $p(0)$ but is readily estimated, which provides a practical means of computing expected prices for commodities in temporarily fixed supply.

Generally, the price equation produces a “saw-tooth” pattern, which is illustrated in Figure 4-27. The increasing portion is actually an exponential, but of such a small degree that it looks linear. When the new harvest comes in, prices drop abruptly as the inventory grows dramatically, and the same pattern is repeated.

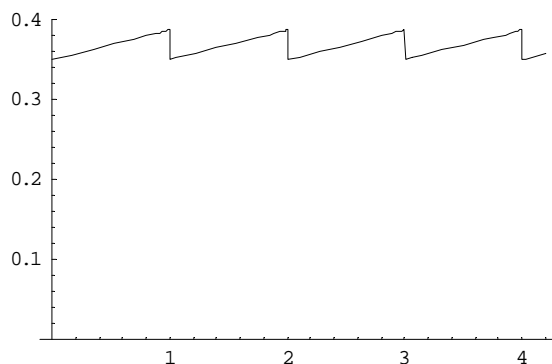


Figure 4-27: Prices over a Cycle for Seasonal Commodities

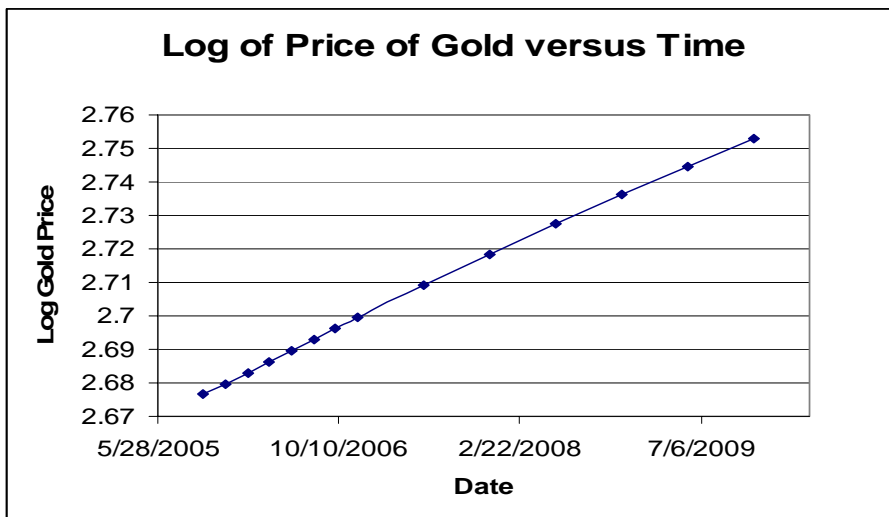


Figure 4-28: Log of Price of Gold over Time

How well does the theory work? Figure 4-28 shows the log of the future price of gold over time. The relevant data comes from a futures market which establishes, at one moment in time, the price of gold for future delivery, and thus represents today's estimate of the future price of gold. These data, then, represent the expected future price at a particular moment in time (the afternoon of October 11, 2005), and thus correspond to the prices in the theory, since perceived risks are fixed. (Usually in the real world, risk plays a salient role.) We can observe that prices are approximately an exponential, because the log of prices is approximately linear. However, the estimate of $r+\delta$ is surprisingly low, at an annual level of less than 0.03, or 3% for both discounting and depreciation. Depreciation of gold is low, but this still represents a very low interest rate.