1

Production

Production is possibly the basic economic activity. Without it there would be nothing to consume, so the theory of demand would not be much of an issue. Consequently we begin our introduction to contemporary economic concepts with the choices people face when producing goods or services. In addition to introducing you to a particular body of theory, we also begin here in exposing you – gradually though – to the terminology of contemporary economics. Much of it is intuitive, but at just enough of an oblique angle to daily meanings of the identical words that you should pay careful attention. Our beginning point is the relationship between the things people use to produce other things and the things they produce with them – called inputs and outputs in the economic lexicon. The concept of the production function (sections 1.1 and 1.2) makes aspects of these relationships somewhat more precise than their use in casual conversation, but the degree of precision can vary according to the need for precision, which is a pleasant characteristic of this body of theory. The production function characterizes the technology – the actual physical and engineering relationships among inputs and outputs – in a fashion that constrains the choices people find it useful to make as well as the consequences of any choices they do make. Correspondingly, changes in technology can change both choices and results (section 1.9).

One of the more important insights that contemporary economics uses, time and again, is that there is generally more than one way to do just about anything. Economics calls this aspect of life "substitution" or "substitutability" (sections 1.3 and 1.4). It characterizes consumption as well as production, but in this chapter we'll focus on its role in production choices. One of the critical capacities of contemporary production concepts in economics is the ability to attribute proportions of products to the inputs that helped produce them. This attribution is called income distribution, and it involves attributing the product(s) produced to the inputs that produced them (or their owners, more precisely) in the form of income (section 1.6). This process may actually feel quite intuitive to scholars of the ancient world who are accustomed to thinking of many workers, particularly in the Near Eastern and Aegean palatial and temple economies, being paid in the form of rations or a comparable part of what they produced. It's the same thing, basically. (As an historical accident of intellectual development, the term "income distribution" has also come to name a different, but certainly not unrelated, concept – that of how a total

Economic Theory and the Ancient Mediterranean, First Edition. Donald W. Jones.

income in an economy is distributed among its members. This has become called the "personal distribution of income" to distinguish it from the "functional distribution of income," which refers to how output is attributed, if not necessarily actually distributed, to the inputs that produced it; section 1.11.)

Throughout this introduction to concepts about the economics of production – the choices people make in production – we have woven both actual and hypothetical examples from times and places in the ancient Mediterranean region. We close the chapter with a more extended example of how the use of concepts from production theory can illuminate the interpretation, and possibly even the translation, of ancient texts.

Economic concepts are prescriptive, as well as descriptive, in the sense that they identify the choices people could make that would make them the best off, in their own assessments, in terms of their own goals. Accordingly, the concept of efficiency emerges (section 1.7). With the further step of a widespread belief that most people at most times and places haven't willingly left "food on the table," these descriptive prescriptions also yield predictions of how people will behave – the choices they'll make – in a wide range of circumstances (sections 1.8 and 1.9).

1.1 The Production Function

The workhorse concept of the theory of production is the production function, which relates the quantity of a product produced to the quantities of things used to produce it. The "things used to produce it" are called "factors of production" (sometimes "factors" for short) or "inputs." For expositional purposes it is common (because it is simple) to study production functions with two inputs. Suppose we consider cotton (an output) to be produced with labor and land as the inputs, or the factors of production. Introducing some simple notation, we could use the shorthand $Q = f(L, N)$, where Q represents the quantity of cotton produced, L is the quantity of land used, N is the quantity of labor used, and f stands for the technological relationship between the inputs and the output.[1] The expression $Q = f(L, N)$ is read as "Q equals (or "is") a function of L and N," not "Q equals f times L or N."

Assume that all units of labor are equivalent to one another (that is, no big strong fellows and small weak fellows), all units of land are identical (fertility, slope, and so forth), and that all units of the cotton are of the same kind and quality. Otherwise, how could we compare units with one another? If you wanted to distinguish between, say, two categories of labor, one small and weak, the other big and strong, you would just specify two different labor inputs. This is the first example of a simplifying assumption in economic analysis (most assumptions do simplify; life is complicated enough without assuming that it is more so). The second example is in the assumption that the production function has just two inputs in it. This is a commonly used assumption designed to highlight the behavior of an individual factor. We could have called one of the factors "labor" and the other "all other inputs." A two-factor designation serves to demonstrate most – but admittedly not all – of the behavior we want to investigate in production. The same simplification to just two items will appear commonly throughout this survey.

The relationship between each input and the output is precisely defined. To get more cotton, if the quantity of land is held fixed at the amount $\overline{L}$, we must increase the quantity of labor used. Conversely, if labor is fixed at $\overline{N}$, to get more cotton we must increase the amount of land we use. To get more output, at least one of the inputs must be increased in number. Further, production functions commonly – but not necessarily always – have the property that if the quantity of any one of the inputs used (we are not restricted to only two inputs; this is just for expositional convenience) is zero, the output is zero. Thus, $Q = f(0, N) = f(L, 0) = f(0, 0) = 0$.

Production functions contain considerably more information about the technology of production than just that more inputs are required to produce more of any output. They describe (i) exactly how much more of each input is required to produce another unit of output, and how this quantitative relationship can be expected to change as quantities of inputs and production change; (ii) the ways that other inputs affect the relationship between any particular input and output; (iii) relationships among inputs such as substitutability and complementarity; and (iv) the effects, if any, of overall scale of production on the productivity of inputs. They help predict the employment decisions of producers and how producers will respond to cost changes and various technological changes.

Even if ancient data are scarce or missing altogether, the concept of the production function is useful, simply for collecting and clarifying your thoughts about what was used in production and what factors might have caused production to differ among locations or times. When we want to use the production function concept to think about a particular line of production at a particular time and place, there is absolutely no difficulty in adding more factors of production than the two we've talked about so far. To think about the economics of, say, pottery production, we certainly would want to include labor time, and for a relatively large potting operation, possibly several skill levels of labor. On the other hand, we might decide that land used in pottery production is so insignificant that we could just ignore it; or alternatively, we might have a case of ceramic production in a city such as fifth-century Athens, where finding space to let freshly turned pots dry before firing, as well as space for kilns and fuel inventories, would have been a non-negligible concern. Next, we might have some capital equipment – wheels, brushes, various tools for smoothing and scraping. Then there is the clay itself, which may be quite specialized. The kilns for firing the pots are a type of capital equipment, and the fuel for the fire is a material input. Each of these inputs would have required decisions that the remainder of the chapter will examine: how much to use, proportions relative to one another, technically possible and economically (even aesthetically) acceptable substitutions among one another.

The pottery example is a case of a production function for a product. We can develop production functions for processes as well, such as different types of industrial heat generation (for ceramics, metallurgy, baking, and preparation of various materials) and chemical processes such as dyeing and oil purification. Some of these production functions could be thought of as nested, in the sense that many of the chemical processes require controlled heat as well as other inputs combined with the heat. Economics has developed the "engineering production function," which uses chemical, mechanical, and other engineering knowledge to develop empirical relationships between "economic" inputs such as quantities of materials and sizes (capacities) of capital equipment and quantities of these process outputs, such as the magnitude of processed oil, dyed textiles, or quantity of heat output (Chenery 1948; Smith 1961, Chapter 2; Marsden *et al.* 1974). Much of the literature on ancient technologies that addresses such topics as the techniques of firing pottery and related ceramic materials such as faience and glass, smelting metals, and the production and use of various chemicals such as cosmetics and dyes, focuses on the material components of recipes, frequently on steps in processes, and occasionally on firing temperatures.[2] Much of the recent, physical science analysis of metals and ceramics is essentially reverse engineering from slags in the case of metals and the actual pots in the ceramic cases, to infer firing temperatures and technological innovations in materials that permitted desired transformations to occur at lower temperatures.[3] While considerable technological knowledge has derived from these investigations, they tend to yield impressions of (i) unique methods used at particular places and times, with deviations representing errors and (ii) different technologies in use to produce similar or identical products at different locations or times. The element of choice of technique within a given technology, which was capable of alternative implementations, gets downplayed in these approaches. This is not a criticism per se, since each analytical methodology offers a certain range of insights; overcoming such restrictions presumably is the motivation for continual calls for interdisciplinary analysis of the ancient world.

Smith's example of "multiple-pass regeneration processes" illustrates the types of choices emphasized by the production function construct (Smith 1961, 42–44). In this type of process, a mixture of reactants, such as a vegetable oil, is passed over a bed composed of some catalytic substance such as fuller's earth. The filtering operation saturates the clay adsorbent but it can be regenerated by washing and burning in a furnace, although the clay's adsorbing capacity falls with each regeneration. Eventually, after a number of these regenerations, the adsorbent declines sufficiently in efficiency that it pays to begin operations with a new adsorbent charge. Smith uses the chemical engineering parameters relating number of passes and subsequent regenerations to adsorbent capacity, then, through a series of substitutions involving quantities of adsorbent (clay) and equipment capacity, derives a production function that says that for a given capacity of filtering equipment, the adsorbent input to the process per year can be reduced only by increasing the number of passes per cycle, which entails using the clay at a lower level of efficiency. A given quantity of filtered vegetable oil can be produced in a year with alternative combinations of equipment capacity and throughput of fuller's earth. This example speaks to findings of alternative material recipes and process steps in ancient industries. There is no necessary implication of different technologies; archaeologists may be observing different choices of production techniques within a given technology. Why they might make those different choices is the subject of section 1.7.

In the meantime, before leaving this introduction to the production function, let's listen to Moorey (1994, 144) on the variability in the ancient use of kilns:

> Pottery kilns were always adapted to the peculiar circumstances of the situation, the resources available, and the type of pottery to be produced . . . Throughout, into modern times, "open" and "kiln" pottery firing, in single- or double-chamber structures, might be found side by side in the same workshop or settlement for the production of different types of vessels or various ceramic fabrics.

Moorey's first observation focuses on the choices available to the ancient potters in choosing the combination of capital and other inputs (primarily fuel, probably, but possibly clay as well). The second observation may be a case of either coexistence of different technologies or simply of different ratios of capital to other inputs within a single technology, with the choice of that ratio depending on clay quality (which we could translate into alternative inputs) or even specific products to be produced, with the input ratio possibly influenced by the relative prices different fabrics or vessel types could command. This last interpretation takes us beyond the concepts we've introduced so far, so with this we return to the development of production theory.

1.2 The "Law" of Variable Proportions

Consider the issue of how output changes with changes in the quantities of inputs applied. Figure 1.1 shows how total output increases as the quantity of labor (N) increases, with the quantity of land (L) fixed. As drawn, the total product (the curve labeled TP) increases moderately at first, then increases more steeply, then has its increase begin to slow down, eventually go to zero, and finally turn down. In Figure 1.2, consider that we

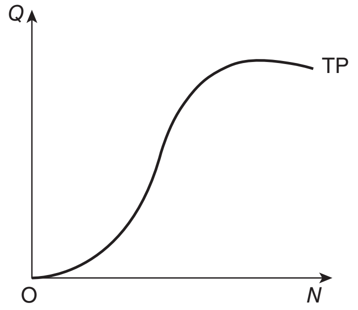

Figure 1.1 The total product curve.

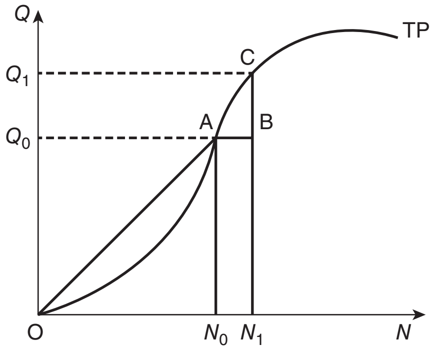

Figure 1.2 Average and marginal products.

have employed labor in the amount N_0. The average product of labor (output Q_0 divided by labor N_0) can be represented by the slope line from the origin to point A on the TP curve ($Q_0 \div N_0$, or Q_0/N_0). Now, suppose we increase labor from N_0 to N_1. Output increases from Q_0 to Q_1, or to point C on the TP curve. The incremental output attributable to the incremental labor input is distance BC. This incremental output is called the *marginal product* of labor. (The definition of the marginal product of labor is $\Delta Q/\Delta N$, where the symbol Δ represents a change in the variable following it.) TP has some degree of curvature between points *A* and *C*, so we cannot draw any straight line to represent the marginal product. But suppose we contemplate making the difference between N_1 and N_0 smaller and smaller, until N_1 is just a tiny bit larger than N_0 – so close together that it looks like we are at a single point on the TP curve. The slope of the TP curve at point A (actually not a point, but the infinitesimal distance between N_0 and N_1 as we've shrunk the increment so much that we can approximate the difference by the point A) represents the marginal product of *N* at the quantity of labor N_0. (The marginal product of labor at N_1 would be the slope of the TP curve at point C.)

The steepest line from the origin to a point on the TP curve will indicate the quantity of *N* per unit of *Q* (actually the Q/N ratio, which is the average product) that gives the largest average product of *N*. Figure 1.3 shows this line. The slope of this line equals the slope of the tangent to the TP curve at this point. So, at the maximum value of average product, average product (AP) equals marginal product (MP). Figure 1.3 depicts the average product and marginal product curves corresponding to the total product curve and shows the intersection of the AP and MP curves below the intersection on the TP curve of the line from the origin that has the greatest slope. Figure 1.3 also marks out three stages of production on the basis of the relationship between average and marginal product. In Stage 1, the average product of the "variable factor" is increasing. Symmetrically, the marginal product of the "fixed" factor is negative. The boundary between Stages 1 and 2 is the maximum point of average product. In Stage 3, marginal product of the variable factor is negative. The boundary between Stages 2 and 3 is the point of maximum total product, indicated by the horizontal line tangent to TP. Producing at any ratio of the variable factor to the fixed factor contained in Stage 1, the producer could get a larger average product by adding more of the variable factor, and he or she would be irrational not to add more of the variable factor. Consequently, production in Stage 1 is irrational. In Stage 3, the producer has added so much of the variable factor that the units are literally tripping over one another; they actually lower total product, which is the meaning of a negative marginal product. Production in that stage is also irrational. Stage 2 contains the only ratios of factors (inputs) that it is rational to employ. One of the thoughts to take away from this exposition is that producers will always produce in a range (of input ratios) of decreasing marginal product, for all inputs. Explanations of people's actions as being efforts to get away from, or avoid, decreasing marginal productivity are incorrect.

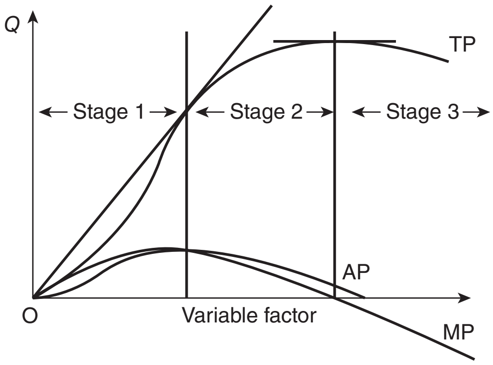

Figure 1.3 Total, average, and marginal product curves.

In Book XI of *De Re Rustica*, ll. 17–18, Columella notes that a specific area of land, an *iugerum*, can be trenched for a vineyard to a depth of 3 feet by 80 laborers working for one day, to 2½ feet by 50 laborers, or to 2 feet by 40 laborers. Notice the constant marginal returns, in terms of depth dug, between the application of 40 and 50 laborers and the decreasing returns when he increases the number of laborers to 80: 80 laborers can dig less than twice as deep as can 40 laborers (Forster and Heffner 1955, 79). This, of course, is not an empirical observation but, possibly even more important, it is a recognition, or expectation, of decreasing marginal returns to increasing applications of labor to a fixed quantity of land.

1.3 Substitution

The next technological relationship specified by the production function that we will discuss is the array of ways that different combinations of the inputs (two in this case) can produce a given quantity of the output. You also can think of this topic as how the inputs relate to one another. In Figure 1.4, the quantity of labor (N) is measured on the abscissa (the horizontal axis) and the quantity of land is measured on the ordinate. The curved line labeled Q_0 represents a constant quantity of output, say 100 bales; it can be produced with any of the combinations of land and labor represented by coordinates lying on it. Thus, the labor-land combinations represented by A (N_0, L_0) and B (N_1, L_1) will both yield 100 bales of cotton (Q_0). The curve Q_0 is called an *isoquant*, because each point on it represents the same quantity of output. Isoquant Q_1 represents a larger quantity of cotton, say 200 bales. Combinations of labor and land represented by points C (N_3, L_3) and D (N_1, L_4) will both produce 200 bales of cotton. Notice that, as these isoquants are drawn, it is not necessary to use larger quantities of both inputs to produce a larger output; in fact, we can produce 200 bales at point D using no more labor than we used at point B to produce 100 bales (N_1) if we are willing to increase our use of land to L_4 from L_1. This concept ("there's more than one way to skin a cat," begging my own cats' pardon for the expression) is known as "substitution." Specifically, the production function represented by the family of curves Q in Figure 1.4 indicates that there is substitutability between land and labor in the production of cotton. Empirically, most production technologies embody substitutability between (among) inputs.

The alternative – nonsubstitutability – can be represented graphically as the L-shaped curves in Figure 1.5. We can combine N_0 units of labor and L_0 units of land to produce Q_0 units of output. If we add some labor, say to N_1, but keep land unchanged at L_0, we still get Q_0 units of output, so we just wasted labor in the amount $N_1 - N_0$. Only land-labor combinations along the line labeled R will be efficient; above R, we're using land that contributes nothing to output, below it we're using labor that contributes nothing. Such a production technology commonly is called a "fixed-coefficients" technology. Why even consider a production function with such a characteristic? Several reasons. First, it is one logical end of the continuum of degrees of substitutability between inputs. Second, for very short periods of analysis, in which it is difficult to substitute among inputs, many technologies with flexibility over longer periods can be studied as if they were fixed-coefficient technologies. The technique known as input-output analysis generally specifies fixed-coefficients technologies.

Let's return to the isoquant diagram and the issue of substitutability among inputs. Figure 1.6

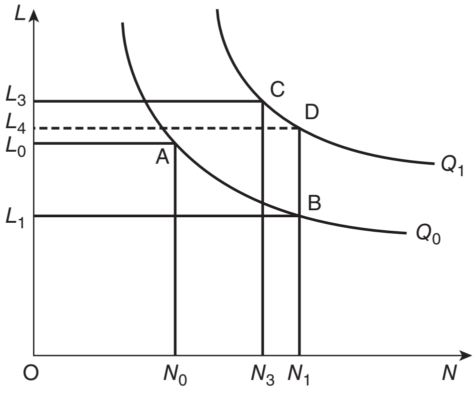

Figure 1.4 An isoquant with substitution between inputs in the production technology.

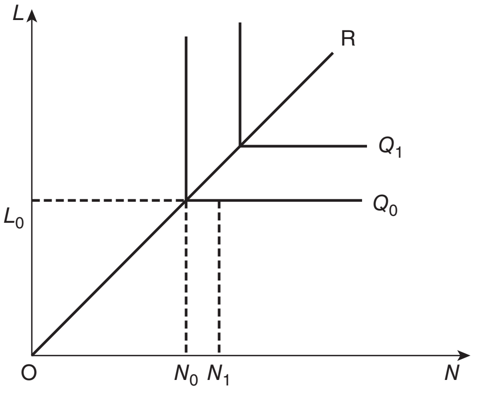

Figure 1.5 An isoquant with no substitution between inputs in the production technology.

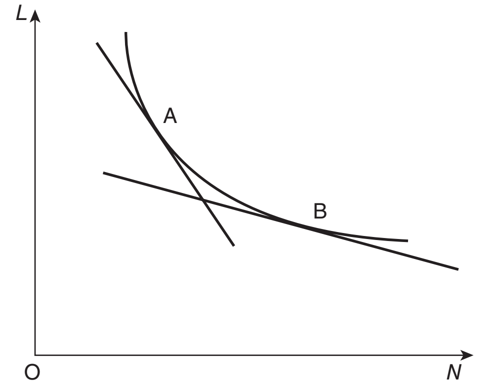

Figure 1.6 Marginal rate of technical substitution (MRTS).

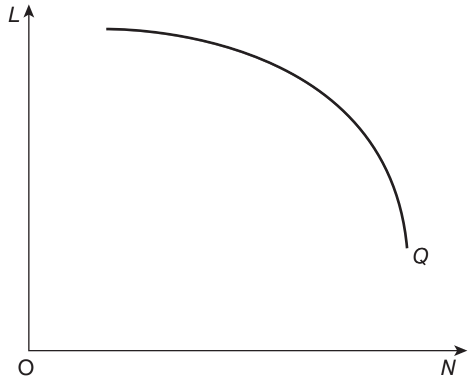

Figure 1.7 An isoquant with increasing MRTS – impossible.

reproduces isoquant Q_0 with points A and B from Figure 1.4. The two lines drawn tangent to points A and B have marginal interpretations analogous to the tangent to the total product curve (TP) in Figure 1.2. The slope of the line tangent to the isoquant at point A represents the number of units of land (L) that have to be substituted for a single unit of labor (N) at that point (the change in L divided by the change in N). The slope is steep relative to the slope of the line tangent through point B. Point A represents a labor-land input combination that uses relatively few units of labor. At such a point, substituting even more land for another unit of labor is relatively difficult. At a labor-land combination like point B, where the ratio of labor to land is high, substituting a unit of land for labor is not nearly so difficult. The slope of the isoquant (actually, the negative of the slope) at any point is called the *marginal rate of technical substitution* (which itself is, in fact, the ratio of the marginal products of the two inputs at that ratio of inputs; we will discuss the concept of the marginal product shortly).

The reader may have wondered why the curvature of the isoquant that allows substitution between inputs is shaped the way it is. Specifically, why is it convex, as Figure 1.4 shows, rather than concave, as in Figure 1.7? We have already presented the information to answer this question, but it may be useful to reassemble it here. The convex isoquant of Figure 1.4 indicated diminishing marginal rates of technical substitution as we moved toward either axis. That is, as more labor is substituted for land (to the right end of the abscissa, or N-axis), it takes progressively more labor to replace a unit of land and still produce a constant output. Viewing this corner of Figure 1.4 alternatively (moving from right to left instead of from left to right), when we are already using a lot of labor, the amount of land required to replace a unit of labor and keep output constant isn't very large. If we had a concave isoquant such as Figure 1.7 shows, our technology would be characterized by increasing marginal rates of technical substitution: as we replaced more land with labor, we could substitute away units of land more and more easily. As we will see below when we introduce the role of input prices in determining input ratios in production, a concave isoquant would encourage the use of higher proportions of the relatively more expensive input.

Figure 1.8 shows an isoquant that possesses infinite substitutability between land and labor. At any location along the isoquant, a unit of land can substitute for x units of labor (where the value of x is determined by the slope of the isoquant). Perfect substitutability does not play a large role in economic analysis, probably because it is not important empirically. We present it simply to show the limiting case of substitutability in production.

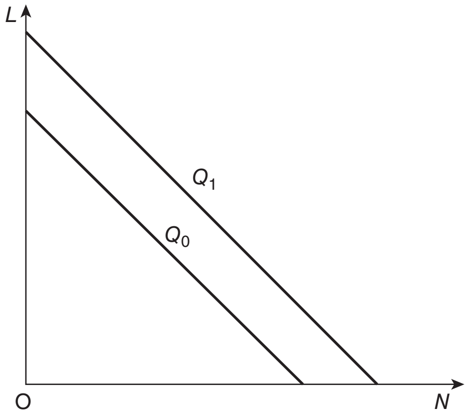

Figure 1.8 An isoquant with perfect substitutability between inputs – unlikely.

Much of the practical agricultural advice contained in the Roman texts such as Columella's *Res Rustica* and *De Arboribus* and portions of Pliny the Elder's *Natural History* is written as if the combinations of resources used in various crops and husbanding were required in very specific proportions, very much as would be implied by fixed-coefficients production functions. Nonetheless, even in these texts we can find discussions of alternative ways of doing things. Pliny, in Book XVIII of the *Natural History*, l. 35, notes that, at least in older times, it was considered better to sow less land and plough it better – clearly a substitution of labor for land (Rackham 1950, 213).

1.4 Measuring Substitution

Recall from Figure 1.2 that we can calculate the marginal products of both inputs – and consequently the ratio of their marginal products – from knowledge of the ratio of the quantities of the two factors (with, of course, knowledge of the "functional form" of the production function, which we will discuss below). A summary measure of the degree of substitutability between inputs in producing a constant quantity of output, called the *elasticity of substitution* (between inputs), is the percentage change in the ratio of inputs divided by the percentage change in the ratio of marginal products. It is always measured positively; frequently the lower case Greek letter σ (or σ_{ij} – read as "sigma-sub ij" – for the elasticity of substitution between inputs i and j when there are more than two inputs in the production function) is used to denote it. In a more mathematical treatment than we will use here, there are a number of ways of deriving formulae for the elasticity of substitution, some using strictly characteristics of the production function, others using input prices; none is "wrong," but different measures illuminate different aspects of substitution and different circumstances. Another, fairly intuitively appealing formula defines the elasticity of substitution between two inputs as the negative of percentage change in the ratio of the quantities used divided by the percentage change in the ratio of their costs. The elasticity of substitution – indeed any elasticity – is a pure, dimensionless number. That is, it does not have the dimensions of output/input or cost/quantity, or whatever; it will have the dimensions of input/input or cost/cost, such that the measured units cancel. (If, in modeling some problem yourself, you find occasion to construct an elasticity and you find that it has the dimensions of, say, distance over time, or some such, you've made an error.)

When a production function has only two inputs, those inputs are always substitutes for each other. In the cases of three or more inputs it is possible for some pairs of inputs to be complements. In the case of substitutes, when the relative price of one input goes up – call it input A – the ratio of input A to substitute input B would fall as the producer substitutes B for A. If inputs A and C are complements to each other, when the use of one of those inputs falls because of a rise in its relative price, the use of the complement also will fall; whether the ratio of the two complementary inputs falls, rises, or remains constant is an empirical matter. Nevertheless, the ratios of both those inputs to input B, for which they must be substitutes, will fall when the price of one of them rises relative to the price of B. The issue of substitutability or complementarity is important in the subject of the demand for inputs, which we will discuss below.

1.5 Specific "Functional Forms" for Production Functions

Since we have brought up the concept of the "functional form" of a production function, let's discuss it somewhat further. We introduced the concept of the production function with general, functional notation $f(■,■)$, where "f" (it could have been any letter, Roman, Greek, or otherwise) deliberately avoids spelling out exactly what the equation looks like. Recalling junior high school algebra, some function $y = f(x)$ could represent a specific equation like $y = a + 2x$, where x is the "independent" variable and y is the "dependent" variable. (On a Cartesian graph, such as we've used here to describe the behavior of production functions, y is on the ordinate and x is on the abscissa.) Several specific functional forms have been extremely popular for production functions, because of both their theoretical properties and their ability to find empirical correspondence in data on production.

The simplest functional form that allows substitutability between (among) inputs is the Cobb–Douglas function: $Q = AN^{\alpha}L^{\beta}$, in which A is simply a constant term, which turns out to be handy to represent such events as technical change. First, note that if the value of either input (N or L in our cotton case) is zero, the value of Q will be zero. The exponential parameters α and β, called "output elasticities," are positive and generally add up to a value close to 1.0. We've run into the term "elasticity" already, in reference to substitutability. Elasticities are widely used in economics to describe the percentage change in one quantity (the one in the numerator of the ratio) caused by a 1% change in another quantity; the elasticity is the percentage change in the "dependent" variable divided by the percentage change in the "independent" variable. An output elasticity is the percentage change in output attributable to a 1% change in the corresponding input. The sum of the output elasticities in the Cobb–Douglas function has an important physical interpretation: it is the degree of returns to scale in production. A sum of output elasticities exactly equal to 1.0 implies constant returns to scale (sometimes abbreviated CRS): a 1% increase in all inputs will yield exactly a 1% increase in output. A sum of output elasticities greater than 1.0 implies increasing returns to scale, and a sum less than 1.0 gives decreasing returns to scale. An example of increasing returns to scale would be if a 1% increase in all inputs yielded a 1.05% increase in output. For decreasing returns to scale, a 1% increase in all inputs would yield, say, a 0.95% increase in output. A restrictive feature of the Cobb–Douglas function is that its elasticity of substitution between each pair of inputs is exactly 1.0, and the elasticity of substitution has exactly that value at all points on the isoquant. (As such, the Cobb–Douglas function is one of a class of production functions called "constant elasticity of substitution" functions. This is in contrast to production functions that allow the elasticity of substitution to vary at different points along an isoquant, an apparently "nice" characteristic when one wants to study the effects of substitutability quite closely but one that adds enormous mathematical complexity to any analysis.) Consider the magnitudes of the output elasticities α and β. Under CRS, reasonable values of these two parameters would be $\alpha = 0.8$ and $\beta = 0.2$. By "reasonable," we mean that considerable empirical investigation of agricultural production with the Cobb–Douglas production function has yielded statistically estimated values of closely equivalent parameters around this pair of values. Now, what does it mean to say that the output elasticity of labor is 0.8? A 1% increase in the use of labor, holding constant the amount of land used, will increase output by 0.8%. Doubling your labor alone *won't* double your output: such a proposition ignores the fact that labor isn't the only thing that contributes to the production. It would increase it by 80%. Correspondingly, increasing your land by 1% would increase output by 0.2%; doubling your land input would get an additional 20% of your output.

Another especially popular functional form for production functions is the so-called constant elasticity of substitution, or CES, function: $Q = A[\delta N^{-\rho} + (1-\delta)L^{-\rho}]^{-\nu/\rho}$, where the elasticity of substitution between inputs N and L is

$\sigma = 1/1 + \rho$, and the value of ρ is between positive infinity and -1.0. The A term is comparable to the A term in the Cobb – Douglas function. The parameter v indicates the returns to scale ($v = 1.0$ for CRS). The δ coefficients represent the intensity of use of the inputs, but are not exactly comparable to the output elasticities of the Cobb – Douglas; in fact the output elasticities for the CES function are quite complicated formulae rather than single parameters. The CES is a much more difficult functional form to use for analytical (as contrasted with empirical) study. Nevertheless, this functional form allows the elasticity of substitution between each pair of inputs (all elasticities are constrained to be the same value) to be greater or less than unity, which can have significant implications for the demands for inputs as their relative costs change. (We have not discussed demands for inputs yet – or demands for products for that matter; the concept, applied to inputs, describes how much of the input a producer will want to use, according to its productivity and cost. The issue is of critical importance in determining the distribution of income in an economy among the owners of various factors of production.) When the elasticity of substitution in the CES function is unity ($\sigma = 1.0$ when $\rho = 0$), the form collapses to the Cobb – Douglas form; when $\sigma = 0$ (as $\rho \rightarrow \infty$; in other words, "goes to infinity"), it collapses to the fixed-coefficients production function.

Considering the limitations of these two production functions, we have to say a few words explaining why they maintain their popularity. Contemporary empirical (econometric) study of production favors more sophisticated functions, such as the transcendental logarithmic ("translog"), which allows any degree of substitutability (or complementarity) between any pair of inputs and allows substitutability to vary along isoquants. This functional form has a large number of parameters, which requires a correspondingly large data base for statistical estimation. In circumstances where data are less readily available, the CES and even the Cobb – Douglas are still used. In analytical uses (just writing equations and diagrams with pencil and paper), both the Cobb – Douglas and the CES can demonstrate many interesting theoretical issues while offering considerable mathematical tractability (particularly the Cobb – Douglas). The translog function would be quite difficult to manipulate for heuristic purposes, and would offer little in the way of additional insights to compensate for the greater trouble. The engineering production functions we introduced in section 1.1 generally are far more intricate than any of these functional forms designed for analytical or empirical research.[4]

1.6 Attributing Products to Inputs: Distributing Income from Production

After this brief excursion into functional forms, let's return to the issue of marginal products of inputs. We've seen that the marginal (physical) product (MPP) of an input is the contribution that an increment of the input makes to total output. Under conditions of constant returns to scale, total output can be decomposed into a sum of MPPs: in our case of producing cotton with labor and land, $Q = \text{MPP}_\text{N}N + \text{MPP}_\text{L}L$. Now, think of the cost of producing Q: we have to pay for labor and land. Let's put the cotton in terms of its value by multiplying the entire equation by the price of cotton, $p : pQ = p\text{MPP}_\text{N}N + p\text{MPP}_\text{L}L$. Now, thinking in terms of "wages" and "rents" for labor and land (terms to which we will return shortly), we can express the revenue from the cotton we produced as $pQ = wN + rL$. The wage rate (or the "price" paid for labor, by any other name) is equal to the marginal physical product of labor (which is actually in cotton) times the price of cotton; and similarly for the rental rate (or the "price" paid for using land this season). If we were working in a barter economy (that is, one in which money doesn't exist and people purchase one good directly with another), the payments to labor and land (or to the people who own those factors of production) are made directly in the output, cotton. (What happens to this simple equation when there are either increasing or decreasing returns to scale? With decreasing returns to scale, payment according to marginal productivity will more than exhaust the output – that is, there won't be enough to go around; with increasing returns to scale, there'll be product left over after paying all the factors their marginal products. This doesn't cause as severe a problem for marginal productivity theory of factor pricing – and the income distribution theory based on that – as it might seem, but we'll have to come back to why.)

We can obtain more information out of this cost relationship. We can divide our cotton revenue-cost equation by the value of the cotton output to get an equation in terms of cost shares: $1 = wN/\mathrm{p}Q + rL/\mathrm{p}Q$, where wN/pQ is the proportion of the cost of cotton production that can be attributed to labor and rL/pQ is the proportion attributable to land. These are commonly called "cost shares" or "factor shares." However, it can be shown mathematically that these cost shares are equivalent to the output elasticities of their respective inputs: the percentage change in output divided by the percentage change in input, or the ratio of the marginal product to average product of each input. Recall that *w,* the wage rate, is the marginal physical product of labor, times the price of the output, *p*; since we have w/p, the *p*s cancel and we're left with just the marginal physical product of labor. This is multiplied by N/Q, which is one over the average product of labor; so the entire "share" expression is the marginal product of labor divided by the average product, which is the definition of the output elasticity of labor in the production function.

Having introduced the concept of the factor share, this is a good place to note that the elasticity of substitution gains particular interest for its role in determining the distribution of income among the owners of factors of production. Suppose for the moment that we have two principal factors in our economy (or at least in our model of our economy) – labor and land – and that our economy produces a single good – food. An abstraction, admittedly. If the elasticity of substitution between land and labor in the food production function is unity (1.0), a change in the relative price of land and labor, caused possibly by technological change, population growth, expansion of arable, or some other major event, will leave the factor shares unchanged. However, if $\sigma > 1$, the share of the factor whose relative price has fallen will increase at the expense of the other factor. For example, with $\sigma = 1.5$, say, if the relative price of land falls, land will be substituted for labor to an extent that the relative share of total income going to land will increase; since there are only two factors, that of labor will fall. If $\sigma < 1$, the relative income share of the factor whose relative price has increased will rise at the expense of the other factor.

1.7 Efficiency and the Choice of How to Produce

Let's return to our isoquant version of the production function. Why should we pick one point on it for our input combination rather than any other? In Figure 1.6, the slope of the isoquant at any point represented the rate at which we could substitute land for labor (or labor for land) and still produce the same amount of output. That described our technological capabilities. The negatives of sloped lines in that diagram also represent the cost of land in terms of labor – either minus the rental rate on land divided by the wage rate of labor if we want to use a monetary numeraire, or the number of units of land we could rent if we were to trade a unit of labor for it in the case in which there is no money to use for a numeraire. Either way – with money or without – the (negative of the) slope of a line "in $L–N$ space" represents the availability of land and labor to our producer. The isoquant represents the technical ability to substitute land for labor and still produce the same output, and a "price" or "cost" line represents our producer's ability to secure the services of those two inputs. At a point of tangency between such a price line and an isoquant, the producer can substitute between labor and land in production at the same rate at which he or she can "hire" or "rent" them. In general, higher costs of land relative to labor will prompt producers to use higher ratios of labor to land; similarly for ratios of any two inputs in proportion to their relative costs.

This description of the conditions of efficiency in production may sound fine as theory, but it is legitimate to ask how real people might discover such efficient allocations of their resources for themselves. First, agents directing production operations for themselves or for others can be expected to have a good, first-hand idea of what their input costs are. Even if they do not hire inputs on an open market in an easily measured numeraire such as money, they can be expected to have a good, working idea of what they would have to pay in kind or cash for additional units of each of their inputs. Next, how do they find out about the rates of technical substitution in their production technologies? Two ways: experience and the pressures of competition. Experience is self-explanatory by and large. Competition can come from the interactions of a large number of other individuals interested in bidding away resources for other activities or in supplying the same products as our agent under consideration. Alternatively, staying a step or so ahead of the grim reaper (competition with nature) can have a similar effect in, as Dr. Johnson expressed it, concentrating the mind wonderfully. Does this mean that all societies at all times are perfectly efficient? The answer is, naturally and obviously, "No," but neither can they be expected to leave a lot of so-called "low-hanging fruit" around to rot. Efficiency in any real conditions depends on the users' understanding of their technology and, to some extent, on their understanding of how their own societies operate and respond to opportunities and incentives.

It is important for students of economies, ancient and modern, to distinguish between efficiency and productivity. Ancient agriculture used low-productivity technologies, but chances are excellent that ancient farmers used those low-productivity technologies highly efficiently. The ancient land transportation industry similarly is invariably characterized as inefficient, a quite unlikely state of affairs. Efficiency is a matter of how close the marginal rate of technical substitution (along an isoquant) is to the marginal rate of substitution of inputs as represented by a relative price line in our diagrams or, more generally, by producers' ability to acquire an extra unit of one input in exchange for some quantity of another input. Productivity is represented by how far from the origin of our diagrams an isoquant representing a particular quantity of output is located: a unit isoquant (representing the quantity of inputs required to produce one unit of output) closer to the origin uses fewer inputs than one farther away, hence representing greater productivity. Efficiency refers to the behavioral choice of *where* on that isoquant to produce – that is, given a relative price of inputs and the input substitutability within a technology, how close to the maximum possible output the producer gets from his resources. The difference in contemporary scholars' attitudes toward the people of antiquity, depending on whether we view them as having been inefficient – with all the other pejorative characteristics associated with that unfortunate state of being – or efficient but burdened with unproductive technologies, could have broad consequences for our own studies.

Economic efficiency is not a product of the modern, industrial world, but is simply getting the most out of one's resources that one can, subject to the institutional constraints one faces. In Chapter 6, we'll discuss the role of constraints in modifying an absolute efficiency concept to various forms of conditional efficiency. For a consumption-oriented example, the absence or poor development of information markets to support the Roman housing market, as noted by Frier (1977),[5] probably did retard the rapid matching of people wanting to occupy housing with those having units available, but information is a tricky good to produce, economically speaking, as we will learn in Chapter 7. Given the limited information available on housing, there is little reason to suspect that people knowingly made less of their resources in housing than they believed they could. In pursuing the issue of inefficiency in the Roman housing market further, the tendency to execute long-term contracts and the institutionalized payment after occupancy rather than before or during both could be ascribed to the limited production of information. Introducing concepts from four subsequent chapters in the quasi-empirical discussion of efficiency is not a deliberate tease, but rather a demonstration of the intricacy of the empirical application of the efficiency concept. When ancient institutions supporting some activity do not demonstrate the same capacities of flexibility and overall productivity that typically accompany corresponding activities in the post-World War II period in the Western, industrialized nations, it is simplistic, as well as just plain wrong, to adopt the fallback position that those people did not act economically or that their activities were simply governed by social restraint. Better to investigate the economic reasons for the ancient constraints, as Stambaugh has done regarding the public services that were and weren’t offered in Roman cities.[6]

1.8 Predictions of Production Theory 1: Input Price Changes

Let’s exercise the theory a bit, using this last set of relationships about picking the optimal input ratios according to the prevailing price or cost ratios. Figure 1.9 has a lot of lines in it, but we can walk through them and take away the information they convey. The production technology is characterized by the family of isoquants Q_i, of which we have drawn just three. The amount of output associated with the isoquants increases as we move outward from Q_0 to Q_3. We begin with the situation in which the relative price of land and labor is characterized by line AA', which is tangent to isoquant Q_0 at point 1. Our producer (this “producer” might be an individual, a firm, a family farm, a temple, or an entire region or country) finds that it can produce the most output with its technology by using L_1 amount of land and N_1 labor. The line from the origin, R_A, is called an expansion path; it describes the combinations of land and labor that this technology would employ if it were to expand at the constant set of relative prices described by line AA (refer to A' as “A prime”). Let’s consider a change in this situation: the relative price of labor drops from AA' to AB. But before we proceed, how do we know that such a counterclockwise pivot of the price line around its intersection with the ordinate (the land axis) represents a cheapening in the relative cost of labor? Here’s one way. Suppose that the actual intercepts of price line AA with both axes represent the real resources available to the producer: if the producer decided to put all available resources into the acquisition of land and none into hiring labor, OA is the quantity of land that could be acquired (rented) at the relative prices described by AA'. Alternatively, if she were to devote all her resources to hiring labor at the same relative price ratio, she could hire the services of OA' labor. (There’s no good reason why any producer would want to put all resources into just one input; this is just a method of demonstrating a point.) Now, the relative price changes to the line AB. With the same resources, the producer could still rent OA units of land but could hire OB units of labor, which is considerably more than she could hire under the relative prices of AA'. Consequently, labor is cheaper relative to land under AB than under AA'.

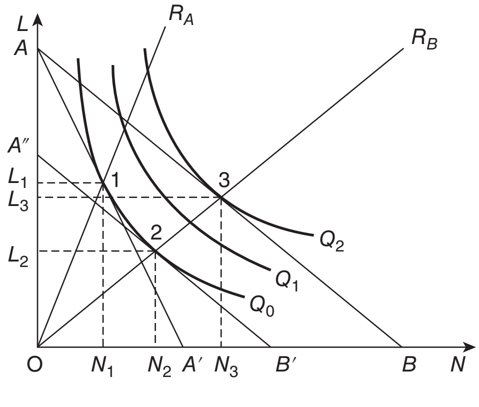

Figure 1.9 Production responses to input price changes.

Now, the relative cost of labor has fallen, and the production technology has remained unchanged. The highest isoquant our producer can reach with the resources characterized by the intercepts of relative price line AB is Q_2. The movement from the input combination (L_1, N_1) to input combination (L_3, N_3) includes a substantial decrease in the ratio of land to labor represented by the shift from expansion path R_A to expansion path R_B. This move includes both a substitution effect and a scale-of-production effect. If we were to change the relative price from AA' to AB but restrict the producer to the same level of production, the input combination would still move toward more labor and less land; the same relative price of AB is reproduced in $A''B'$ (refer to A'' as “A double-prime”), which is tangent to Q_0 at point 2. Here the producer uses less land than before ($L_2 < L_1$) and more labor ($N_2 > N_1$), but still produces the same amount of output. Since we’re letting the change in the relative price reflect a real change in the resources available to the producer, she can expand her scale of production to the point where some isoquant will be just tangent to the new relative price line AB. The producer *could* produce only the amount of output described by isoquant Q_1, but she is able to reach as great a scale of production as that associated with Q_2. The movement from point 2 on Q_0 to point 3 on Q_2 represents the scale effect; land used rises from L_2 to L_3, and labor hired rises from N_2 to N_3. This change in relative price could represent an actual change over time (or even instantaneously) in a single location or a comparison of production choices involving the same technology but different locations with different resource availabilities.

Let's consider a couple of applications of this concept. Agricultural technology in Egypt and Lower Mesopotamia during the middle of the second millennium had much in common, to avoid saying outright that it was identical. Similar arrays of crops were grown with pretty much the same array of tools and animals, and with comparable biological understanding on the parts of the two societies' farmers. Water supply in Egypt was primarily by inundation, with various lifting equipment, while the Mesopotamians supplemented with a more extensive system of canal irrigation, supplemented with comparable lifting equipment. The different water-supply systems may have altered the price of water relative to other inputs, such as seed, labor, animal traction, and hand-held equipment between the two regions. Different population densities would have altered the relative availabilities (and hence costs) of labor and land. We can expect that these differences in relative prices would have had some impacts on the ratios of a number of these inputs, making Egyptian and Mesopotamian agriculture look more different than they actually were at a fundamental, technological level.

Correspondingly, dry-land agriculture in Upper Mesopotamia, with its different relative cost of water (relative to the other inputs such as labor, land, and equipment) than existed in Lower Mesopotamia, would have conferred a considerably different appearance to agricultural practices in the two regions. The seed/land ratios would have responded to the land/water cost ratio, and if more plentiful availability of water enhanced the value of land in Lower Mesopotamia, we would expect to have seen lower ratios of labor to land in Upper Mesopotamia.

1.9 Predictions of Production Theory 2: Technological Changes

Consider another possible change or difference. Figure 1.10 can represent either a technological change facing a given producer or a difference in technologies faced by producers at different locations. Using the technology associated with isoquant Q_0, a producer facing relative prices represented by AA would choose input combinations along expansion path R_0. Facing the same relative prices but using a different technology, represented by isoquant Q_1 – in which the substitution of labor for land is more difficult at each land-labor combination – a producer would use a higher ratio of labor to land, represented by input choices along expansion path R_1.

Consider an example of technological change later in antiquity – the Roman use of pozzolana for a hard, strong cement. In addition to the possibility of producing entirely new products (structures) such as true arches, the increased material strength would have permitted the substitution of land used in structures to actual construction material: buildings would have been able to cover larger floor spaces because the relative price of material strength to land's price had fallen. Additionally, the ratio of usable space per unit of land would have increased as the relative price of its provision fell.

Not all changes in the way things are done are technological changes. Some observations

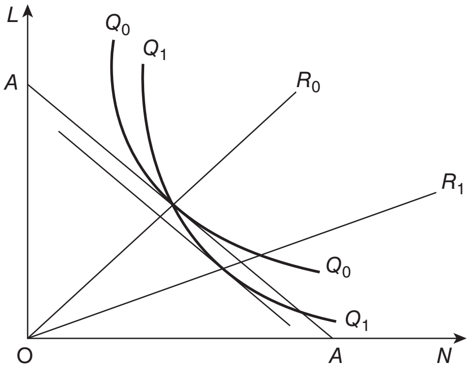

Figure 1.10 Production responses to a change in production technology.

in Laurence's (1999, Chapter 5) recent study of Roman roads can be used to illustrate this point. Laurence notes that the apparent growth in the use of basalt (*silex* or *selce*) in road paving from the third century B.C.E. into the first century C.E. may have been limited primarily (with the glaring exception of the Via Domitiana) to short stretches of "showcase" roads and some city streets. He does not attempt to attribute any part of the increase in its use to technological improvements in quarrying this stone, and indeed it is difficult to see the increase in use of *selce* as representing a technological change. As Laurence notes, the weathering properties of this stone may have made it more conducive to foot and pack animal traffic than to wheeled vehicles (Laurence 1999, 72). Most roads in Italy continued for some time to be surfaced with compacted gravel (*glarea*), but Laurence notes that these roads appear to have been gradually upgraded in quality: "the nature of the road surface and its associated structures were upgraded and altered to reflect changes in the available technology, with a marked improvement in terms of the speed of travel or the weight of goods . . . that might have been transported" (Laurence 1999, 73). It is not clear that the gradual upgradings cited here were indeed permitted by improvements in construction or materials technology, as opposed to having entailed simply more extensive applications of the same technology in response to demands for more roads to accommodate the growth of traffic, in terms of both flow volumes and weight of goods carried.[7] It is possible that technological changes in vehicle technology and tackle for the animal prime movers could have contributed to increased demands for road durability, width, and speed qualities, but no evidence is suggested that changes in actual road construction technology occurred in these particular upgrades. This is not to imply that technological improvements in Roman road construction did not occur, such as modifications of substructural support, use of sand and pebbles for grading under paving stones, and changes in preferred materials because of their properties of cementation, durability, flexibility, and so forth – changes that would have altered isoquants for roads. The advances in bridge construction that permitted wider valleys to be spanned (Laurence 1999, 73–75) probably do represent what would be represented by changes in an isoquant (quantities and proportions of, and substitutabilities between, labor, equipment, and materials required to build a bridge of specified dimensions). This discussion is not a semantic or nomenclatural cavil, but an emphasis of the distinction between a change in the isoquant involved in constructing particular types of infrastructure and growth in the quantity and quality of infrastructure, because of an increased demand for it, using a constant isoquant. The importance of the distinction should be obvious: first, no one would want to confuse growth without technological change for technological change; and second, the implications for who benefitted from the observed events are different as the incomes of the owners of different factors would have fared differently under the alternative circumstances.

1.10 Stocks and Flows

In our discussion of factors, we have referred to "renting" land, "hiring" labor, or the ambiguous term "acquiring the services" of either. It is intuitive to visualize the productive input called "labor" as individual people and to think of their use in a productive activity as occupying their entire persons. Correspondingly with land: why not just "buy" it and "use" it as much as you want? We have deliberately avoided including physical capital, or items of equipment, among our factors of production, for reasons to be discussed below, but the same issue arises very clearly with equipment. If we use a hammer among our inputs, what is the relation between "owning" the hammer and "using" it?

For each type of input, it is useful to distinguish between the stock of the input and the flow of services derived from a unit of it in each time period. The form in which a stock of labor appears is the individual person; even if the person happened to be owned by the agent organizing the production (a case of slavery), only the *services* of the person are used in production during any period. Moving to a less potentially controversial case, consider the hammer we just introduced. Treated reasonably carefully, a hammer will last several periods; we can use it this period, next period, the period after, and so on. In each period, we use the services of the hammer. The services are derived from the "stock" of hammer, and it might be possible to eventually "use up" the stock of hammer – i.e., either wear it out gradually or break it all of a sudden, just through regularly conscientious use. Consider land. Land sometimes is (was) thought of as an "original and indestructible" factor. However, land can be "overused" and its fertility exhausted. Farmers routinely conduct maintenance of one kind or another on their land to keep it from washing away through erosion, burning out through salt accumulation, or otherwise becoming less productive.

Production theory works with flows of services of factors during a given period of time. These flows are derived from a stock that embodies the factor. The acquisition cost of a flow of services from a factor for one time period generally is substantially lower than the acquisition cost of the stock of the factor. It will be helpful to maintain a conscious distinction between stocks and flows in many contexts; ignoring or confusing the two concepts can lead to serious analytical errors.

1.11 The Distribution of Income

We should say a few words about the distribution of income at this point. First, what do we mean by the distribution of income? There are two principal interpretations of this expression, the functional and the personal distributions. The latter is possibly more intuitive: it refers to how total income in the economy is distributed among individuals or families. Frequently it is measured by the Gini coefficient, whose numerical values can be interpreted as degrees of skewness. For instance, in a number of Latin American countries in the 1950s and 1960s (and later as well), the 1 to 2% of families with the highest incomes in the country received around 20 to 30% of total national income, and the top 10% about 50%, while the bottom 60% received around 20% (Jain 1975, 24 Table 13, 89 Table 57; Fishlow 1976, 61 Table 1, 72 Table 5; Webb 1976, 12 Table 1, 13 Table 2; Weisskoff 1976, 35 Table 1, 38–39 Table 2). They also held a similar proportion of national wealth (a stock-flow distinction). The 20% of families receiving the lowest incomes pulled in about 5% of national income. Together with the people somewhere in the middle, these numbers represent the personal distribution of income in these countries. The personal distribution of income has considerable political importance, as it is easy to imagine. These countries may be a reasonable benchmark against which to gauge personal income distribution in antiquity.

The functional distribution of income describes the proportions of total income in the economy going to particular factors of production (land, labor, capital), or more specifically, to the owners of those factors. A functional distribution of income underlies each personal distribution of income: the factors still produce the income, regardless of who owns them. Following the neoclassical model of the functional distribution of income, the factor shares (proportion of input costs accounted for by each factor of production) from individual production functions can be aggregated across the economy to reach aggregate factor shares, which amounts to the functional income distribution. One of the very handy features of the way we have studied production is that, under constant returns to scale, the output elasticity of each factor (of production) equals its share of income from a production process. Proceeding to the level of the aggregate economy, it is not difficult to find that, with a land output elasticity of 0.15, a capital output elasticity of 0.05, and a labor output elasticity of 0.8, labor claims 80% of total income in the economy, the owners of land 15%, and the owners of capital 5%. These output elasticities are characteristic of the output elasticities of "traditional" agriculture (i.e., the agricultural sector that uses animal and human power rather than fossil fuels and natural rather than chemical fertilizers) over the past several decades. Of course, if 5% of families end up with 40% of income, we need to look into how some of the labor income of the bottom 95% of families is being captured by the 5%.

The neoclassical theory of income distribution has been criticized from several directions, primarily for its use of the construct of aggregate capital rather than a plethora of individual items of equipment and because "market imperfections" (a term we have not discussed yet, generally used to refer to departures of industry structure from that of perfect competition; see Chapter 4) cause the incomes to factors to differ from their values of marginal product (which equality is what lets us associate the output elasticities with factor shares). The most important of these disagreements about neoclassical income distribution theory is known in the economics literature as the Cambridge Controversy, or sometimes the Cambridge–Cambridge Controversy; the leading critics of neoclassical income distribution theory have come from Cambridge University, in England, while its most cogent defenders were at the Massachusetts Institute of Technology, in Cambridge, Massachusetts in the United States. We will not devote much space to that discussion; the most telling criticisms of neoclassical distribution theory appear to have their strongest force in situations that are not particularly important empirically, and abandoning the simplicity and power of the neoclassical theory would leave us effectively without an alternative theory of functional income distribution. So, as is the case with many theories in science, while it may not be perfect, it will get us by until a superior theory emerges, which to date has not occurred.

It is easy to see that neoclassical distribution theory relies on both supply and demand influences to arrive at a distribution of income. Output elasticities in production functions are clearly and purely technological parameters. However, it is the value of marginal product that determines factors' shares; value of marginal product of any factor is the price (value however expressed) of the product times the marginal physical product of the factor – the physical amount of what it produces. Prices reflect individual and group valuations – that is, the foundations of product demand, which we discuss in Chapter 3.

Using neoclassical income distribution theory gives us a baseline, functional distribution of income that "should have" appeared in many of the ancient Mediterranean and Aegean economies, considering their production technologies. To the extent that we can get occasional glimpses from either textual or artifactual records that the personal distribution of income (which may be easier to observe for those places during the long periods of antiquity) that differ substantially from the 15-5-20 distribution we noted in the previous paragraph, there is a need for explanation. Neoclassical income distribution theory gives us an implicit baseline that needs explanation. Some combination of taxation, tribute, slavery, and imperfect markets (monopolization of some productive activities is one such imperfection) are obvious candidates. More subtle possibilities derive from other market imperfections, including the possible absence of markets for such items as insurance of various sorts (see Chapter 3, on consumption, for a discussion of risk and insurance, although there is no treatment there of the absence of insurance provision). Other theories of the personal distribution of income would assist in this explanation, but since some of the most incisive of those theories rely on concepts we have not introduced yet, we defer further discussion of them to a later chapter.

Before leaving the subject here, however, we offer a brief preview of what to expect in the way of the analytical treatment of income distribution. So far, we have treated production in a partial-equilibrium approach. The alternative is a general-equilibrium approach. Partial equilibrium describes situations in which either the problem under study has few and small enough connections to other parts of the economy that we can safely ignore them, or that the extension to general equilibrium brings sufficient complications that we need to walk before we run and learn what we can on the *assumption* that those interactions are insignificant. One thing that happens in general equilibrium analyses that generally doesn't in partial equilibrium is that the distribution of income can change as a consequence of some of the changes under study. A change in the income distribution can lead to a change in aggregate demand because the consumption patterns of major groups of individuals differ. For instance, a shift in income from labor to owners of capital might precipitate a shift in demand from basic foods to "luxuries" or from consumption to saving. We are getting a bit ahead of ourselves now, because we have yet to introduce the study of demand, but we believe there is sufficient intuition about what "demand" and "consumption" involve for the reader to take away a satisfactory preview impression of what to expect in the general equilibrium analysis of income distribution.

1.12 Production Functions in Achaemenid Babylonia

Matthew Stolper's analysis of tablets from the Murašû Archive (Stolper 1985, Chapter VI) contains a number of tables showing various inputs into agricultural operations and some indicators of outputs: numbers of oxen or cows; equipment such as plows and harness; rental prices, in terms of grain, of these variable inputs and land, and the apparent rent on land; outputs of barley; influence of a plot's location adjacent to a canal, which of course offers a more convenient supply of water as well as the possibility of transportation of harvested crop. He infers the considerable value of location next to a canal, but otherwise a generally low price of land relative to the costs of the moveable inputs. These are the classic ingredients of a production function, but analyzed without the benefit of the production function as an organizing concept. At the risk of using an interesting and excellent piece of work as a negative example, Stolper relates outputs to the quantity of a single input at a time, which doesn't take advantage of the information on how the presence of one input affects the productivity of another – with the exception of the water in the canals, which he doesn't really acknowledge as another input. Also the production function framework for thinking about everyday work offers an adding-up discipline that is useful – it helps the student account for everything that goes into the production and relate those things to everything that comes out. In a particularly interesting subset of these texts, Stolper ran into this adding up issue and intuitively recognized it but did not appreciate the full implications of his conclusions. We turn to this case.

Reinforcing, in his judgment, the conclusion of typically low land prices are four texts recording what Stolper calls "agreements to cultivate land in partnership" (130). It could be called a share-rental agreement, such as is common throughout both the developing and industrialized world today. These four texts describe agreements (contracts) between owners and renters of land, and Stolper interprets the agreement in such a way that the land owner furnished his land, both parties supplied animals, equipment, laborers, and so on, in equal quantities, and then they shared the crop equally. Stolper noticed something odd about this arrangement – that it didn't leave any return for the people supplying the land – but did not pursue the matter other than to interpret the case as further evidence of cheap land.

With the application of some simple production theory, it's easy to show that this interpretation of the tablets implies that the land owners were letting the renters use their land rent-free. Some contemporary scholars would be inclined to favor such an interpretation if they actually worked it out themselves, but the issue of "free" land outside a distant frontier region, which this wasn't, raises more questions than it answers satisfactorily. An alternative interpretation of these results is that the tablet evidence was incomplete but was translated as if it were complete.

If the land owner supplies the land and half of everything else and the other fellow supplies half of everything else, how do you decide what the income share of land, labor, equipment, and so forth, are? If 50% goes to half of the labor input (the "other fellow") and 50% goes to half the labor income and all the rental (land) income, what are the shares of labor and land in production? If you've looked at these numbers and thought that something was funny, you're right. Follow this: Let s_N be the share of labor's contribution to the output and s_L be land's share (the "share" concepts from production theory – they refer to the share of the output produced by the specified inputs; the shares will add up to 1). Start with the "other fellow": he gets half of labor's share of output and that's it; that is equal to half of the total product. In other words, $0.5s_N R = 0.5R$. Now, the land owner gets the other half of the output, while his contributions are half of the labor and all of the land. So he has a claim on half of labor's contribution to output plus all of land's contribution to output. Expressed as an equation, that is: $0.5s_N R + s_L R = 0.5R$. Now, from the other fellow's equation we get the solution that $s_N = 1.0$; plug that into the land owner's equation and we get the result that $s_L = 0$. We could set this up as a set of two simultaneous equations (s_N and s_L are the variables) and use matrix algebra to get numerical solutions for s_N and s_L, and we also get $s_N = 1$ and $s_L = 0$. The social interpretation of this is that the contribution of land to production was zero – at least if people were able to claim what they had produced. (You may ask: "but how did they know what they, or the inputs they supplied, produced?" Answer: observation and passing down the information in a social information storage and retrieval system.) Either the land in question was at the absolute spatial edge of economically usable land (along the lines of the von Thünen model; see Chapter 11), leaving zero revenue net of transportation costs for land rent, or the observations are questionable. Actually, one of the observations simply has to be wrong; what's questionable is which one it is – that they split the output down the middle, or that they each supplied half of everything else.

References

Belfiore, C.M., P.M. Day, A. Hein, V. Kilikoglou, V. La Rosa, P. Mazzoleni, and A. Pezzino. 2007. "Petrographic and Chemical Characterization of Pottery Production of the Late Minoan I Kiln at Haghia Triada, Crete." *Achaeometry* 49: 621 – 653.

Boni, M., G. Di Maio, R. Frei, and M. Villa. 2000. "Lead Isotopic Evidence for a Mixed Provenance for Roman Water Pipes from Pompeii." *Archaeometry* 42: 201 – 208.

Chenery, Hollis B. 1948. "Engineering Production Functions." *Quarterly Journal of Economics* 63: 507 – 531.

Fishlow, Albert. 1976. "Brazilian Size Distribution of Income." In *Income Distribution in Latin America*, edited by Alejandro Foxley. Cambridge: Cambridge University Press, pp. 59 – 75.

Forbes, R.J. 1954. "Chemical, Culinary, and Cosmetic Arts." In *A History of Technology*, Vol. I. *From Early Times to the Fall of Ancient Empires*, edited by Charles Singer, E.J. Holmyard, and A.R. Hall. New York: Oxford University Press, pp. 238 – 298.

Forster, E.S., and Edward H. Heffner, translators. 1955. *Lucius Junius Moderatus Columella on Agriculture and Trees*, Vol. III. *Res Rustica X – XII, De Arboribus*. Loeb Classical Library. Cambridge MA: Harvard University Press.

Freestone, Ian C. 1995. "Ceramic Petrography." *American Journal of Archaeology* 99: 111 – 115.

Frier, Bruce Woodward. 1977. "The Rental Market in Early Imperial Rome," *Journal of Roman Studies* 67: 27 – 37.

Hein, A., and H. Mommsen. 1999. "Element Concentration Distributions and Most Discriminating Elements for Provenancing by Neutron Activation Analyses of Ceramics from Bronze Age Sites in Greece." *Journal of Archaeological Science* 26: 1053 – 1058.

Jain, Shail. 1975. *Size Distribution of Income; A Compilation of Data*. Washington, D.C.: International Bank for Reconstruction and Development.

Laurence, Ray. 1999. *The Roads of Roman Italy*. London: Routledge.

Lucas, A., and J.R. Harris. 1962. *Ancient Egyptian Materials and Industries*, 4th edn. London: Edward Arnold.

Marsden, James, David Pingry, and Andrew Whinston. 1974. "Engineering Foundations of Production Functions." *Journal of Economic Theory* 9: 124 – 140.

Moorey, P.R.S. 1994. *Ancient Mesopotamian Materials and Industries; The Archaeological Evidence*. Oxford: Clarendon Press.

Quinn, P.S., and P.M. Day. 2007. "Calcareous Microfossils in Bronze Age Aegean Ceramics: Illuminating Technology and Provenance." *Archaeometry* 49: 775 – 793.

Rackham, H., translator. 1950. *Pliny, Natural History*, Vol. V. *Libri XVII-XIX*. Loeb Classical Library. Cambridge MA: Harvard University Press.

Smith, Vernon L. 1961. *Investment and Production; A Study in the Theory of the Capital-Using Enterprise*. Cambridge MA: Harvard University Press.

Stambaugh, John E. 1988. *The Ancient Roman City*. Baltimore MD: Johns Hopkins University Press.

Stolper, Matthew W. 1985. *Entrepreneurs and Empire; The Murašû Archive, the Murašû Firm, and Persian Rule in Babylonia*. Istanbul: Nederlands Historisch-Archaeologisch Instituut te Istanbul.

Stos, Zofia A., and Noel H. Gale. 2006. "Lead Isotope and Chemical Analysis of Slags from Chrysokamino." In *Hesperia Supplements* 36, *The Chrysokamino Metallurgy Workshop and its Territory*, edited by P.P. Betancourt. Princeton NJ: American School of Classical Studies at Athens, pp. 229 – 319.

Stos-Gale, Zofia. 2001. "Minoan Foreign Relations and Copper Metallurgy in MM III-LM III Crete." In *The Social Context of Technological Change: Egypt and the Near East, 1650–1550*, edited by A.J. Shortland. Oxford: Oxbow Books, pp. 195–210.

Vandiver, Pamela, and Charles S. Tumosa. 1995. "Xeroradiographic Imaging." *American Journal of Archaeology* 99: 121–124.

Vaughan, Sarah J. 1995. "Ceramic Petrology and Petrography in the Aegean." *American Journal of Archaeology* 99: 115–117.

Webb, Richard C. 1976. "The Distribution of Income in Peru." In *Income Distribution in Latin America*, edited by Alejandro Foxley. Cambridge: Cambridge University Press, pp. 11–25.

Weisskoff, Richard. 1976. "Income Distribution and Economic Growth in Puerto Rico, Argentina and Mexico." In *Income Distribution in Latin America*, edited by Alejandro Foxley. Cambridge: Cambridge University Press, pp. 27–58.

Suggested Readings

Beattie, Bruce R., and C. Robert Taylor. 1985. *The Economics of Production*. New York: John Wiley & Sons, Inc.

Becker, Gary S. 1971. *Economic Theory*. New York: Knopf. Chapters 7–8.

Pindyck, Robert S., and Daniel L. Rubinfeld. 2001. *Microeconomics*, 5th edn. Upper Saddle River NJ: Macmillan. Chapter 6.

Varian, Hal R. 1999. *Intermediate Microeconomics; A Modern Approach* ["Baby Varian"], 5th edn. New York: Norton. Chapter 18.

Varian, Hal R. 1992. *Microeconomic Analysis*, 3rd edn. New York: Norton. Chapter 1.

Notes

[1] Section 1.5 shows the full expressions of some popular production functions. These formulas are called "functional forms."

[2] Lucas and Harris (1962, 150–154) on dyeing; Forbes (1954, 249–250) on dyeing; Moorey (1994, 144–150 on firing pottery, 240–301 on base metals). Moorey (1994, 150) notes the fact that kilns permit more efficient use of fuel; translated into the language of the production function, there is a tradeoff between the use of capital embodied in a kiln and the amount of fuel used in the production of a given quantity of pottery. Lucas and Harris (1962, 371–372) hint at the capital–fuel tradeoff as Egyptian potters moved from simply covering the pots to be baked with a heap of animal dung in Pre-Dynastic times and using straw, chaff, reeds, and so forth, for fuel, to surrounding the heap with a low, clay wall and the dung covering replaced by clay, to finally a true kiln, the use of which must have been well established by the Early Kingdom.

[3] A sample of metals analyses: Boni *et al.* 2000; Stos-Gale 2001; Stos and Gale 2006. A sample of ceramic analyses: Freestone 1995; Vaughan, 1995; Vandiver and Tumosa 1995; Hein and Mommsen 1999; Belfiore *et al.* 2007; Quinn and Day 2007.

[4] For instance, the production function Smith (1961, 44) derived for the multiple-pass regeneration process (use of fuller's earth as a catalyst in purifying vegetable oil) is $y = Ax_1[1 - Br^{\gamma X_2/x_1}]$, where $A = \alpha/1 - \mathrm{r}$, $B = r^{1-[\theta_\mathrm{f}/(\theta_\mathrm{f}+\theta_\mathrm{r})]}$, and $\gamma = H/\beta(\theta_\mathrm{f} + \theta_\mathrm{r})$. The output level of purified oil is y, x_1 is the quantity of fuller's earth, and X_2 is the capacity of the adsorptive equipment (the capital). The definitions of the engineering parameters are: r is the capacity of the adsorbent after its regeneration relative to before; θ_f is the hours per pass in the filtering phase of the process and θ_r is the regeneration time per pass; H is the hours per year of operation; α is a proportional constant representing the ratio of output in the initial pass to the size of the adsorbent charge in that pass; and β is the corresponding ratio of the adsorbent facility to the initial adsorbent charge. The specialization of the engineering production function to the process (or product) it describes is responsible for this complication. While the Cobb–Douglas, CES, and translog functions can be used to approximate this process as well as innumerable others, the engineering production function can yield insights into the choice behavior toward one specific process or product, contingent upon the technology. Whether the engineering information available on many ancient production processes, such as various metallurgical operations, is sufficient to develop engineering production functions along these lines or not is less important than the alternative perspective on ancient production behavior that this concept opens. The production function focuses our attention on the choices available to ancient producers, within the confines of the

technologies they used, somewhat expanding our horizons beyond the relatively rigid combinations of materials and time incidentally implied by strict readings of many of the physical science studies of these ancient technologies.

5 I say "consumption-oriented" because the production of housing is implicit in the example, as well as consumption. We deal with some of the peculiarities of housing as a good in Chapter 12.

6 Stambaugh (1988, Chapter 8) identifies the public services offered in contemporary (implicitly United States) and ancient Roman cities and considers a number of reasons for the absences of public (or even private, sometimes) provision of some of them in the ancient cities. He explicitly declines to apologize for what he believes some readers may consider the use of "modernizing" concepts.

7 This is a matter of the "derived demand" for roads – a demand derived from people who want to travel and carry things. We'll introduce derived demand in Chapter 2.