PART III

Theory of Production and Costs

CHAPTER 10

MARSHALL'S ANALYSIS OF SUPPLY

Marshall begins his analysis of supply by introducing the agents, or factors, of production—land, labor, and capital:

> The agents of production are commonly classed as Land, Labour, and Capital. By Land is meant the material and the forces which Nature gives freely for man's aid, in land and water, in air and light and heat. By Labour is meant the economic work of man, whether with the hand or the head. By Capital is meant all stored-up provision for the production of material goods, and for the attainment of those benefits which are commonly reckoned as part of income. It is the main stock of wealth regarded as an agent of production rather than as a direct source of gratification. [Marshall 1972: 115]

He then undertakes a detailed analysis of each agent, especially with respect to their supply available for production, their specifications as an agent of production, and their use as a conceptually homogeneous category.

Agents of Production

Land

Since land is a non-produced agent of production, it cannot be augmented by labor. Therefore the amount or *supply of land available for production* is given:

> The area of the earth is fixed: the geometric relations in which any particular part of it stands to other parts are fixed. Man has no control over them; they are wholly unaffected by demand; they have no cost of production, there is no supply price at which they can be produced. [Marshall 1972: 120]

Land contributes to the production of material goods in as many ways as there are different kinds of land. In the case of agriculture, the kinds of land used in the production of agricultural goods depend on the natural qualities of the soil. That is, land used in agriculture is denoted or *specified by its natural properties* and, thus its use in production depends on the kinds of agricultural crops being grown. In the same respect, fisheries, mines, and other natural resources can also be denoted or specified according to their natural properties.

Because land is delineated according to its natural specifications or properties, it appears as many different agents of production that have no common attribute. However, Marshall argued that any agent of production that was non-produced, irrespective of its natural properties, could be classified as land if it could be appropriated by some individuals and, at the same time, leaving other individuals without it. Thus, land, as a conceptually homogeneous category for naturally inhomogeneous non-produced agents of production, is a *social-economic* concept that captures the social-economic relationships between the "landed" and the "landless". It should be noted that in the process of making land a conceptually homogeneous category, Marshall has smuggled in the idea of scarcity in that there exists less land than the amount desired by the individuals in the economy.[1]

Labor

In discussing the amount or supply of labor available for production, it must be kept clear that two distinct but closely related phenomena are involved—the absolute amount or supply of laboring bodies available for production and the amount or supply of labor effort or services available for production. Since the former provides the base from which the latter works, it will be considered first.

The amount of laboring bodies available for production at a given point in time is given and known; however the amount depends firstly on the natural excess of births over deaths a generation back

[1] For further discussion, see Levine (1977), pp. 241 – 248.

and secondly on migration. In turn, the number of births depends on the habits relating to marriage and the number of deaths on the abundance of food. Regarding marriage, the age of marriage varies with climate, with younger marriages taking place in warm climates and older marriages in colder climates; and given climate, the average age of marriage depends on the ease with which the individuals can gain employment and support themselves "comfortably":

> Early in this century [19th century], when wages were low and wheat was dear, the working classes generally spent more than half their income on bread: consequently a rise in the price of wheat diminished marriages very much among them: that is, it diminished very much the number of marriages by banns [public announcement, especially in churches]….But as time went on, the price of wheat fell and wages rose, till now the working classes spend on the average less than a quarter of their incomes on bread; and in consequence the variations of commercial prosperity have got to exercise a preponderating influence on the marriage-rate. [Marshall 1972: 157-8]

Since death rates depends on the abundance of food, the amount or supply of laboring bodies available for production at any point in time is not independent of the activities undertaken to satisfy wants a generation(s) past. Consequently, when viewing the supply of laboring bodies available for production from a secular time advantage it is not possible to state that it is knowable in terms of an absolute amount or in terms of a natural growth rate. However, within the economic time periods of the short or long period, the supply or absolute number of laboring bodies is given and therefore is known .

Given the supply of laboring bodies, we can now determine the amount of labor effort or services available for production. Marshall argued that labor effort was a displeasure or disutility since the worker's unwillingness to work, to work long hours, or to work in unappealing places indicated that work was a disutility [Introduce David Spencer's work]. Therefore as a worker's total amount of labor effort or services increased so did his total disutility and it did so at an increasing rate. That is, if V is taken to represent the total disutility of the amount of labor services L, then $dV/dL > 0$ represents the marginal degree of disutility, with $d^2V/dL^2 > 0$. Then Marshall argued that:

> …it is broadly true that the exertions which any set of workers will make, rise or fall with a rise or fall in the remuneration which is offered to them…the price required to call forth the exertion necessary for producing any given amount of a commodity, may be called the *supply price* for that amount during the same time. [Marshall 1972: 118]

That the supply price of labor is related to the marginal disutility of labor, meaning that higher supply prices (wages) are needed to call forth greater amounts of a laborer's labor service. This can be seen in the following manner:

> Let dU/dM be the marginal utility of money income to the worker; let W be the total wages paid to the worker for supplying the total amount of labor services L; and let dW/dL be the wage rate (w); then $dW/dL \times dU/dM = dV/dL$ or $w \times dU/dM = dV/dL$. Since Marshall tacitly argued that dU/dM is a constant, it is clear that the amount of labor services available for production is an increasing function of the wage rate, that is the supply curve for a worker's labor services is positively sloped throughout.[2]

The specification of labor as an agent of production depends on physical vigor and industrial training. That is, the efficiency which a worker's labor service can be employed in the production of various goods is dependent on his physical vigor and industrial training. In turn, physical vigor is dependent on climate, race, food, clothing, housing, rest, freedom, and the occupation's working conditions. Given physical vigor, the form it takes in production depends on the kind of industrial

---

[2] However as in the case of consumer demand, if dU/dM is allowed to change, then the shape of the labor services supply curve can vary, as in Lionel Robbins's backward bending supply curve for labor services. [Robbins, 1930]

training the worker has received. In turn, the worker's industrial training depends on his general ability and specialized ability. The former denotes those faculties, general knowledge and intelligence that are common property of all higher grades of industry and that depend on upbringing and general education, and the latter denotes manual dexterity and acquaintance with particular materials and processes that are required for the special purposes of individual trades and which depend on technical education and on the job training either by parents or employers. Thus labor services as an agent of production can be delineated by its natural vigor and industrial training; hence not only is each kind of labor service technically-naturally specific but also inhomogeneous. It should also be noted that the specification of labor services is dependent ultimately on a given fixity of wants and on the given fixity of activities undertaken to satisfy those wants.

Because labor services are inhomogeneous, the general concept of labor appears to be unsustainable. However there is a common link between the different kinds of labor services. For a labor service to be in demand, that is to be part of the productive activities that lead to the satisfaction of wants, it must be useful. To be useful means that the labor service can be used to help satisfy the wants of a variety of individuals, including those that do not possess the particular labor service. Thus all labor services that are useful, that is form part of the productive activities that lead to the satisfaction of the wants of the society as a whole, can be grouped together into a conceptually homogeneous category called labor.

Capital

Marshall's presentation of capital as an agent of production is complicated by the fact that he sees capital as both produced goods and money. His presentation is further complicated by including organization and knowledge as capital as well. To obtain a clear understanding of Marshall's discussion of capital, first consider the following chart:

Chart 1

```
                              |----consumption capital
                              |                              |----fixed capital
Capital as produced goods----|                              |
                              |----instrumental capital------- |
                                                             |
                                                             |----circulating capital
```

where consumption capital consists of goods in a form to satisfy wants directly;

instrumental capital consists of all goods that aid labor in production;

fixed capital consists of all goods that exist in a durable shape and the return to which is spread over a period of corresponding duration; and

circulating capital consists of all goods used directly and only once in the production of another good (usually material inputs).

In his discussion of capital as an agent, Marshall concentrated on instrumental capital and organization, leaving aside consumption capital. In fact, Marshall considered all produced goods as capital since they could either be used in the production of other goods or as consumption goods required by labor used in production. However, since instrumental capital plays the most significant part in the production of material goods that satisfy wants, Marshall concentrated on it. But in concentrating on instrumental capital with respect to production, Marshall did not simply deal with machines qua machines. Rather he saw the productivity of the instrumental capital intimately tied up with the manner in which they are organized and used. Thus to Marshall, capital as an agent of production consists of instrumental capital and the manner in which it is used in production, or more succinctly instrumental capital and knowledge.

To determine the amount or supply of capital, that is instrumental capital, available for production, Marshall first considered the growth of wealth. Up until the 18th century, he argued, wealth increased very slowly because the forms of instrumental capital were very primitive. However, with the advent of the industrial revolution and more sophisticated instrumental capital, wealth increased at a greater rate. Moreover, the growth of wealth stimulates the development of new wants, thus providing new openings for the investment of money into instrumental capital. This creative effect of wealth not only increases the rate of growth of wealth, but also indicates that capital can be seen as self-expanding. That is, instrumental capital creates greater amounts of wealth that, in turn, creates new wants and therefore new opportunities for the investment of new instrumental capital; thus instrumental capital creates itself with no determined end. Consequently, over a secular period of time, the amount of instrumental capital is not knowable in terms of an absolute amount or in terms of a natural growth rate. However, within the economic time periods of short period or long period, the absolute amount of capital available for production is knowable, but in a very peculiar way, as will be noted below.

Marshall approached the specifications of capital as an agent of production in terms of instrumental capital and its organization for production. Therefore, he concentrated on the relationship of industrial organization to specialized plant and equipment, and to the division of labor. Marshall started his analysis by linking specialized instrumental capital and the specialization of labor to the scale of production. That is, as the firm's and/or industry's scale of production increased, the methods of organizing production on larger scales would require that the division of labor become more intense and systematic and that the instrumental capital become more specialized in the tasks it performs:

> It is the largeness of markets, the increased demand for great number of things of the same kind, and in some cases of things made with great accuracy, that leads to subdivision of labor; the chief effect of the improvement of machinery is to cheapen and make more accurate the work which would anyhow have been subdivided. [Marshall 1972: 212]

More particularly, given the scale of production there is a relationship between specialized machinery and division of labor in that the creation of new specialized machinery created the need for new kinds of specialized labor. This incestuous relationship between machinery and the division of labor not only undermines the giveness of labor in the long period, but also enables any particular organization of production (or method of production) to be identified by its division of labor, its particular categories of labor services, and its set of specialized machinery.

In developing his view of the relationship between the scale of production of output and specialized machinery and the organization of production, Marshall relied on the concepts of external and internal economies. *External economies* are economies that arise from an increase in industry output when the industry and its subsidiary industries are localized. In such a situation, a small firm may employ highly specialized machinery by specializing in the production of an intermediate input that belongs to one stage of the production process:

> …the economic use of expensive machinery can sometimes be attained in a very high degree in a district in which there is a large aggregate production of the same kind, even though no individual capital employed in the trade be very large. For subsidiary industries devoting themselves each to one small branch of the process of production, and working it for a great many of their neighbors, are able to keep in constant use machinery of the most highly specialized character, and to make it pay its expenses, though its original cost may have been high, and its rate of depreciation very rapid. [Marshall 1972: 225]

On the other hand, *internal economies* are economies that arise from an increase in the scale of production of the individual firm and are dependent on the firm's resources, its organization, and the efficiency of its management. The advantages of production on a large scale are best shown in manufacturing, which includes all businesses engaged in working up material into forms that will be adopted for sale in distant markets. The chief advantages of large-scale production are *economy of skill*, *economy of materials*, and *economy of machinery*. The first advantage refers to the possibility of a firm with large output to be able to keep each of its employees constantly engaged in the most difficult work of which he is capable, and yet so to narrow the range of his work that he can attain the faculty and excellence which comes from long-continued practice. The second advantage a firm with large output has is that it can utilized scrap material and be more precise when cutting material. The last advantage is that firms with large scale production are knowledgeable about the kinds of machinery that will best suit them, can use highly specialized machinery programmed to perform a high specialized task, and can build their own specialized machinery.

While large scale production promotes economy of skill and materials, its most interesting and important relationship is with specialized machinery. By specifying a unique relationship between the scale of production and specialized machinery, internal economies has, in effect, stipulated that each scale of production has its own unique *method of production* including specialized labor services. Thus when the scale of production increases, new specialized machinery and labor services must be employed. Thus it can be concluded that instrumental capital consists of technically specific materials, plant, and equipment. In particular, the kinds of machinery and plants that make up instrumental capital are not only technically specific, their specification also ties them to a specific scale of production. Hence not only are the produced goods that fall in the province of instrumental capital technically different, some of them are output-specific different.

To say that the heterogeneous technically specific types of instrumental capital can be denoted simply as capital, they must have a common denominator. One possible common denominator is that they are produced goods that are used indirectly to satisfy wants. This not only makes all instrumental capital useful capital, but it also makes them consumption goods of a slightly different form. A second possible common denominator is to view all instrumental capital in terms of money. A third possible common denominator is to view all instrumental capital in terms of their *real costs*, that is in terms of the efforts and sacrifices that went into them. Of the three possibilities, only the second and third can be used to determine the amount of capital as a homogeneous mass that is available for production, as will shown below.

Before doing that, let us revisit the notion of the supply of capital. In the short period, the various kinds of instrumental capital are given. Because of their heterogeneous nature, the amount of capital used in production cannot be determined except in value terms or in terms of sacrifices. In the long period, instrumental capital becomes "free form"; therefore Marshall had to delineate the conditions under which the individual would save, that is forgo consuming all their income and convert the savings into instrumental capital. To initiate the analysis, Marshall first had to name the unit by which the free floating capital and savings are measured. Choosing money as the unit, Marshall then argued that security was a condition that must exist if an individual is to save and the principle reason for saving is family affection, that is for providing his/her family with a larger stock of wealth/capital on which to build. Since such a desire is as much of a want as reading a book, the individual will naturally distribute his money income in a manner in which the ratio of marginal utility of savings to the price of saving is equal to the marginal utility of money ($\lambda = MU_s/p_s$). Because the individual prefers present consumption to future consumption, the satisfactions he receives from the future consumption must be greater than that he receives from present consumption. That is, the individual discount future pleasures to the present: future pleasures/ (1 + i) = present pleasures, where i = Δ pleasure/present pleasure and is called the discount rate.[3] While it is not strictly possible to estimate the quantity of future satisfaction, it can be roughly estimated by the interest rate found in the money market under the following assumptions: (a) the individual expects to be about as rich in the future as he is presently, and (b) his capacity for deriving benefit from the goods money will buy will on the whole remain unchanged.

[3] As Marshall stated: "The unwillingness to postpone enjoyment, and thus to save for future use, is measured by the interest on accumulated wealth which just affords a sufficient incentive to save for the future." [Marshall 1972: 18]

Therefore, we can conclude that the amount of savings an individual will undertake is directly related to the interest rate in the money markets and that the higher (lower) the interest rate, the higher (lower) the amount of savings will be:

> And human nature being what it is, we are justified in speaking of the interest on capital as the reward of the sacrifice involved in the waiting for the enjoyment of material resources….The greater the rate of gain from the present sacrifice the greater will often be the saving….So the higher the rate of interest the greater the saving as a rule…. [Marshall 1972: 193-5]

Assuming that all savings are converted into instrumental capital, the supply of capital available for production in the long period is determined by the interest rate.[4]

Production

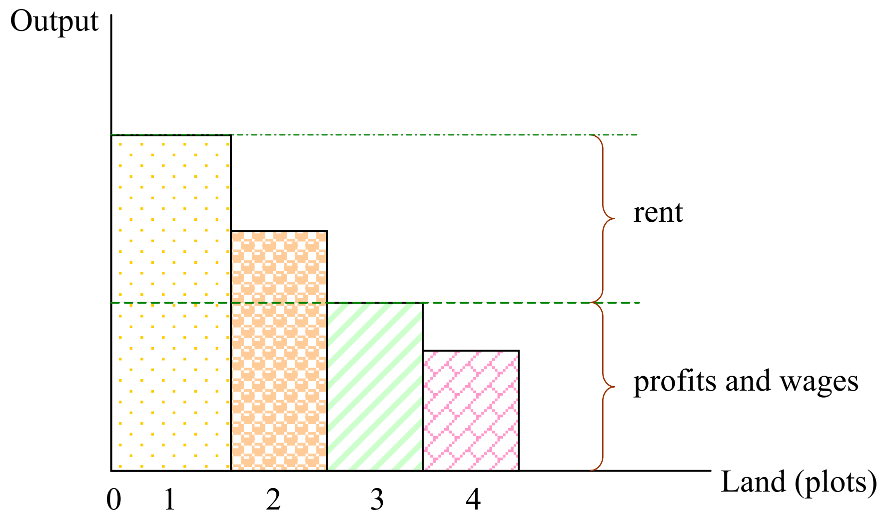
Marshall's analysis of production was based on the concepts of *diminishing returns*, *returns to scale*, and *substitution*. He began the analysis by initially considering diminishing returns with respect to land and then extended it to capital and labor. He then dealt with returns to scale with respect to capital and labor.

Land and Diminishing Returns

In classical political economy, the concept of diminishing returns was not part of value theory; rather it was used to explain rent. That is, when discussing production and output with respect to land, the Classical economists dealt with the idea of simultaneous cultivation of lands of different fertilities or qualities. Given this "photographic" view of production and land, the differential productivities of land and rents arising there from are directly worked out on the basis of a single observation. Thus the no rent (or marginal) land is concretely identifiable as the plot that produces no surplus above profits:

[4] For further discussion on the supply of capital, see Garegnani (1978).

Figure 10.1



where 3 is the no rent land;

each plot of land employs the same amount of labor and capital; and

the rent obtained in this situation is called *extensive rent*.

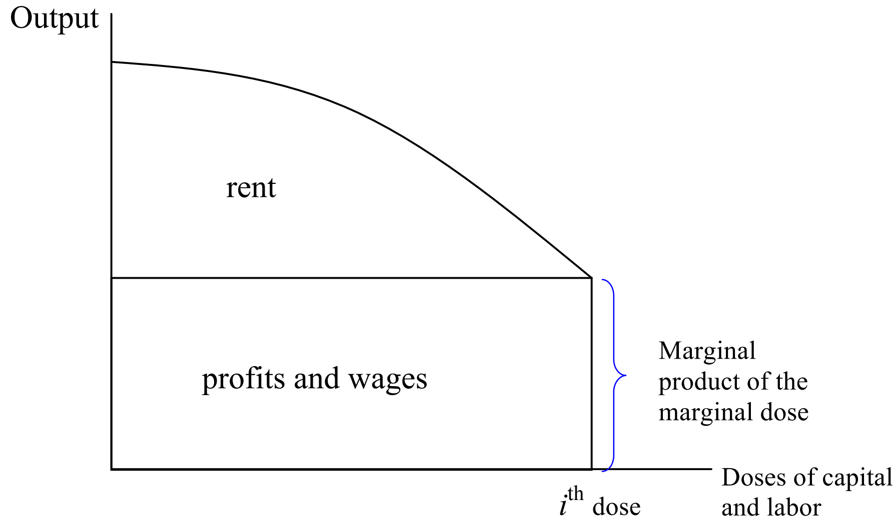
However, for Marshall to place diminishing returns within his theory of value (price) [short discussion in a footnote on the difference/same], he had to introduce a new idea to production—the idea of hypothetical (as opposed to real) change. That is, to talk about the increment of output that emerges from successive applications of doses of capital and labor to land, the marginal output must be due to the additional dose; thus potential or hypothetical change is needed to talk about marginal output and to derive a function between the number of doses of capital and labor and output.

Marshall defined diminishing returns with respect to land as follows:

An increase in the capital and labour applied in the cultivation of land causes in general a less than proportionate increase in the amount of produced raised, unless it happens to coincide with an improvement in the arts of agriculture. [Marshall 1972: 125]

The *dose* is always a combination of capital and labor and they are always applied in equal amounts. The dose that only just remunerates the cultivator is the *marginal dose* and the increase in output is the *marginal return* or the *marginal product* which is measured in physical terms:

Figure 10.2



To obtain a better understanding of Marshall's argument let us consider the following points. Marshall argued that the law of diminishing returns was a technical relationship that could be verified by commonsense:

> Were it not for this tendency every farmer could save nearly the whole of his rent by giving up all but a small piece of his land, and bestowing all his capital and labour on that. If all the capital and labor which he would in that case apply to it, gave as good a return in proportion as that which he now applies to it, he would get from that plot as large a produce as he now gets from his whole farm; and he would make a net gain of all his rent save that of the little plot that he retained. [Marshall 1972: 126]

In spite of the law's *prima facie* respectability, there are two theoretical problems with it. First, the law assumes that the productivity of different doses of capital and labor to the same lot of land are independent of each other. However, this is not easy to ascertain. The ordering of different uses according to productivity would be ambiguous if the utilization of a dose of capital and labor in one particular use affected the marginal product of the dose in the successive use. That is, if the return on any particular dose is affected by the past history of specific uses made of the preceding doses, then there would be no unique ordering possible on the basis of purely the number of doses. [Bharadwaj 1994]

A second problem concerns the doses of capital and labor. One may construct examples of situations where diminishing returns is attributed a physical and tangible meaning. The ideal case is (Figure 3) when a variable input (or an assortment of inputs in fixed composition—Figure 4) is applied (in a single use) in successive uniform doses to land, producing a homogeneous product, and it is assumed that the successive doses so applied have a diminishing productivity:

Figure 10.3

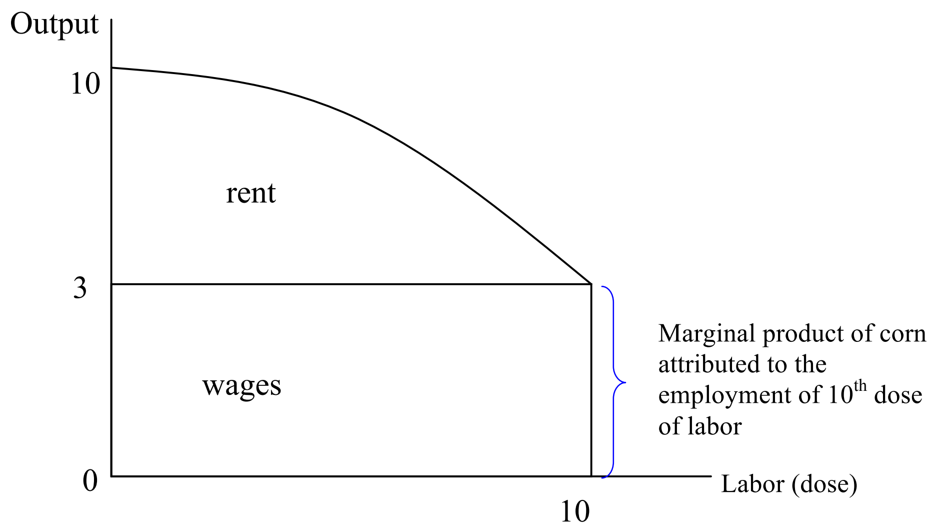
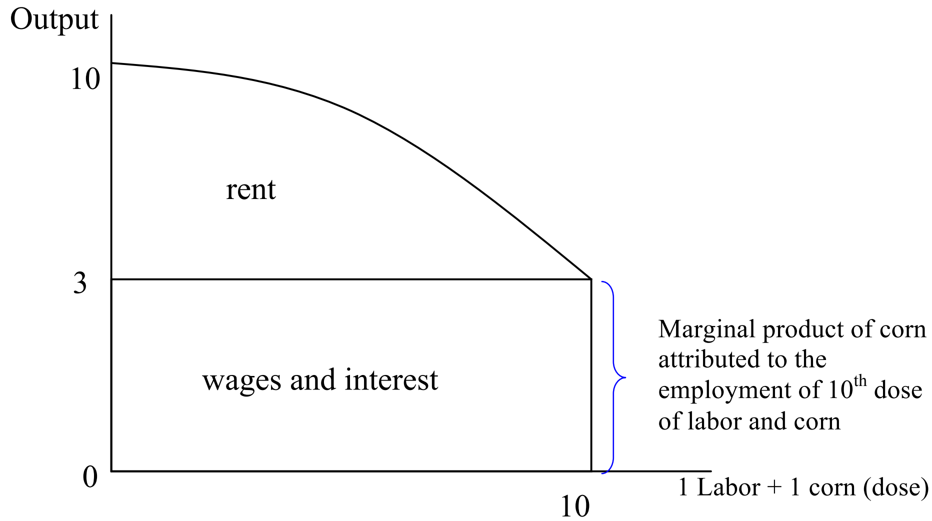


Figure 10.4



However, such an exposition of the law of diminishing returns cannot be extended to situations in which different kinds and composition of labor and instrumental capital are used. In this case, the doses must be measured in terms of money (or reduced to homogeneous efforts and sacrifices) and so must the marginal product. The exposition cannot also be extended when the doses of capital and labor have alternative possibilities of employment. This point will be expanded upon when the concept of substitution is examined.

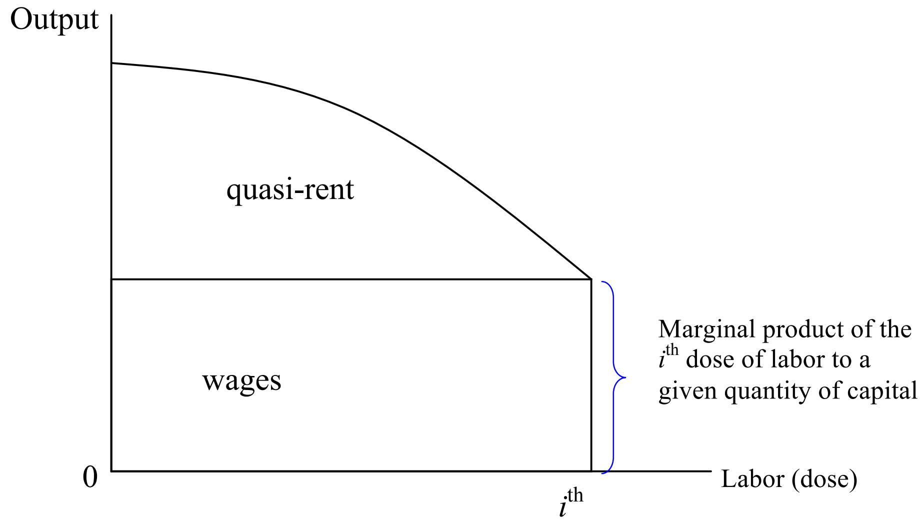
Labor, Capital, and Diminishing Returns

Marshall generalized his analysis of diminishing returns to capital and labor; but in doing so he ran into a number of problems. In an effort to discuss diminishing returns with respect to labor, he created his famous sheep example (Marshall 1972: 427–30) in which additional shepherds were added to the production of sheep that did not use instrumental capital. While such an example permits a clear delineation of the law of diminishing returns and identification of labor's marginal product, it is inconsistent with the general tenor of Marshall's analysis in which capital and labor are used jointly in production. Thus he created another example using both labor and capital in which labor was applied to a fixed amount of instrumental capital:

> If a manufacturer has, say, three planting machines there is a certain amount of work which he can get out of them easily. If he wants to get more work from them he must laboriously economize every minute of their time during the ordinary hours, and perhaps overtime. Thus after they are once well employed, every successive application of effort to them brings him a diminishing return. At last the net return is so small that he finds it cheaper to buy a forth machine than to force so much work out of his old machine: just as a farmer who has already cultivated his land highly finds it cheaper to take in more land than to force more produce from his present land. [Marshall 1972: 140]

However, again, this example violates Marshall's view of the relationship between labor services and specialized machinery. That is, let us assume that we have a particular kind of machine or a fixed set of machines. According to Marshall these machine(s) require a particular kind of labor service; thus it would nearly be impossible to apply an additional unit of specialized labor service to the machine(s). To surmount this difficulty, Marshall recognized that the instrumental capital had to be expressed in terms of money: "It is true that when the tendency to diminishing returns is generalized, the return is apt to be expressed in terms of value, and not quantity" (Marshall 1972: 142). Thus the analysis of diminishing returns with respect to labor, given a fixed amount of capital, reduces to the application of successive units of homogeneous labor to homogeneous capital:

Figure 10.5



However there are two problems with this solution. Although the value (or dollar) amount of capital does not change with the successive additions of labor, the actual instrumental capital must change.[5] Thus it becomes questionable whether we can say that capital is fixed. As a corollary, the labor services applied each time must also change with each application; consequently, if each dose of labor is to be the same, they must be applied in money form. Hence the above example should be changed to where the vertical axis is denoted as value of output, the horizontal axis as value of labor, and the margin as the value of the marginal product. The second problem is that if instrumental capital is allowed to change, then the law of diminishing returns takes on a long period perspective and thus undermines its use as the basis for the short period supply curve.

---

[5] An example of this is Denis Robertson's beer example:

> If ten men are to be set to dig a hole instead of nine, they will be furnished with ten cheaper spades instead of nine more expensive ones; or perhaps, if there is no room for him to dig comfortably, the tenth man will be furnished with a bucket and sent to fetch beer for the other nine. Once we allow ourselves this liberty, we can exhibit in the sharpest form the principle of *variation*,--the principle that you can combine varying amounts of one factor with a fixed amount of all the others; and we can draw, for labour or for any other factor, a perfectly definite descending curve of marginal productivity. [Robertson 1931: 47]

Reversing the analysis to where labor is fixed and capital is applied in successive doses, the same problems above arise. Capital, labor, and output must be in value terms, while the fixity of labor and homogeneity of the doses of capital become questionable. These problems are accentuated when alternative possibilities of employment for either capital and labor exist. This point will be examined next when the relationship between substitution and diminishing returns is broached.

<u>Substitution and Diminishing Returns</u>

Up until now, only single possibilities for use of variables inputs have been considered when dealing with diminishing returns. However under more general conditions, the application of the variable inputs must be such as to maximize the firm's profits in all directions. But for the firm to be able to do this, the returns from the alternative uses must be comparable, which in the last instance means they must be in terms of money (or homogeneous efforts and sacrifices). Consider the following example where the firm has two alternative uses for its variable input, labor:

Doses of Labor	Alternative 1 Value of Marginal Product	Alternative II Value of Marginal Product
1	20	30
2	15	25
3	10	8
4	5	3

The first point that must be noted is that the returns can only be compared in terms of money (otherwise it would be like comparing apples and oranges). Secondly, the doses of labor themselves must be in value terms for the reasons noted above (and the same also for capital). Lastly, we find that diminishing returns appears not as a technical necessity, but a prearranged ranking arrived at by the firm seeking to maximize its returns. Thus, it is not so much that the successive doses of labor cause diminishing returns; rather diminishing returns is simply a descriptive outcome of the ranking of the different returns.[6]

Diminishing Returns: Conclusion

The above analysis of diminishing returns indicates that it cannot be extended beyond the simple parable of the application of identical physical doses of a variable input to land and in which no alternative uses exist. In all other cases, the fixity of the given input, the homogeneity of the doses, and the functional relationship between variable input and output becomes questionable. In addition, further arguments can be made indicating that a fixed amount of money capital cannot be seen as fixed at different interest rates, thus undermining diminishing returns with respect to capital. However, discussion of this point will be put off until distribution is dealt with in Part VI.

Capital, Labor, and Returns to Scale

Marshall's discussion of returns to scale focused on *increasing returns*. Restricting the use of the concept to manufacturing industries where instrumental capital was freely reproducible, Marshall defined the *law of increasing returns* as

> an increase of labour and capital leads generally to improved organization, which increases the efficiency of labour and capital. Therefore in those industries which are not engaged in raising raw produce an increase of labour and capital generally gives a return increased more than in proportion;…. [Marshall 1972: 265]

The definition implies that if capital and labor is increased by the same percentage, output will increase by more than that percentage due to the improvement in the organization of production. However there are some problems with it. Keeping with Marshall's view of the relationship between machinery, division of labor, and the scale of production, the following situation could occur:

---

[6] See Bharadwaj, 1994, pp. 44-50.

1 machine A + 2 laborers A → 3 corn

2 machine B + 1 laborer A + 2 laborers B → 7 corn

While it is obvious that output has more than doubled, it is not obvious that capital and labor has doubled. This problem can be overcome if the inputs and outputs are put in value form. But this "solution" brings problems of its own, especially with respect to comparing two different amounts of capital that consist of different kinds of instrumental capital.

A second way of dealing with the problem is to convert the instrumental capital and labor to a homogeneous effort and sacrifice (or waiting). But since different instrumental capital and specialized labor services consist of different efforts and sacrifices, this solution is doomed to theoretical failure. Still Marshall preferred it to the valuation approach above:

> Increasing Return is a relation between a quantity of effort and sacrifice on the one hand, and a quantity of product on the other. The quantities cannot be taken exactly, because changing methods of production call for machinery, and for unskilled and skilled labour of new kinds and in new proportions. But, taking a broad view, we may perhaps say vaguely that the output of a certain amount of labour and capital in an industry has increased by perhaps a quarter or a third in the last twenty years. To measure outlay and output in terms of money is a tempting, but a dangerous resource: for a comparison of money outlay with money returns is apt to slide into an estimate of the rate of profit on capital. [Marshall 1972: 266]

A third solution to the problem is to simply double the fixed combination of instrumental capital and labor exactly:

1 machine A + 2 labor A → 3 corn

2 machines A + 4 laborers A → 7 corn

In this case, increasing returns is clearly evident and discernable completely in physical quantities or efforts and sacrifices. However this solution violates Marshall's view of the relationship between instrumental capital, division of labor, and scale of output. Thus, although Marshall defined increasing returns, he could not provide a sound theoretical foundation for it. Rather he was forced into assuming homogeneous efforts and sacrifices.

Costs

Marshall begins his discussion of costs by first introducing the concepts of *short period* and *long period*. He defined short period as the period of time in which the stock of appliances of production are fixed and supply is adjusted to demand by varying their utilization. It is within this context that the law of diminishing returns operates since there exists some fixed agents of production (fixed instrumental capital) on which circulating instrumental capital and labor can be applied. The long period, on the other hand, is defined as the period of time in which the stock of appliances of production can be freely adjusted to demand. It is in this context that increasing returns to scale operate since both capital and labor are freely adjustable. Next Marshall introduced two different views of costs—one view deals only with the efforts (or services) of labor and the sacrifices of capitalist forgoing consumption and is called the *real cost of production*; and the second deals with the money counterpart to the real cost of production and is called *expenses of production*. To obtain a more concrete understanding of the costs of production, the expenses of production will be considered first.

The expenses of production consist of two major categories, *prime costs* and *supplementary costs*. Prime costs consists of those input costs that go directly into the production of the output and includes direct material costs (circulating instrumental capital) and direct labor costs. Supplementary costs consists of those input costs that go indirectly into the production of the output and generally includes salaries of upper level employees, depreciation of fixed instrumental capital, and general expenses of the firm including marketing costs, interest on capital, and the gross earnings of management. Given the cost categories, the firm will, argued Marshall, estimate its expenses of production by estimating at normal output its total prime and supplementary costs and adding them together to get total expenses of production at normal output. This can be shown in the following manner. Let $x_1,\ldots,x_n$ be direct material inputs and $p_1,\ldots,p_n$ be their (supply) price; $l_1,\ldots,l_n$ be direct labor inputs and $w_1,\ldots,w_n$ be their (supply) price; E be the effort and organization of management and p_e be its (supply) price; and K be the value of instrumental capital, i the rate of interest and its (supply) price, and δ its rate of depreciation (with all other costs being ignored). Thus the total expenses of producing a normal volume of output is derived as $x_1p_1 + \ldots + x_np_n + l_1w_1 + \ldots + l_nw_n + Ep_e + K(\delta + i) =$ total expenses of production. From total expenses of production, *marginal expenses of production* and *average total expenses of production* are derived in the usual fashion.

Before investigating the movement of expenses of production with respect to changes in output both in the short and long period, its relationship with real costs of production needs to be delineated. Marshall defined *supply price* as the price required to call forth a particular amount of effort and sacrifice. Thus each input price is a supply price that calls forth a particular amount of effort and sacrifice sufficient to produce the amount of output and the output price is the supply price that is made up of all input supply prices. Hence to get at the real cost of producing a good, that is the efforts and sacrifices that really make up the output supply price, the supply prices of the material inputs have to be resolved into supply prices that make them up. But such a pursuit of real costs would never come to a end in an economic system made up of just instrumental capital goods and labor and where production is a circular process.[7] Hence Marshall took the material inputs as well as all other inputs as the "ultimate facts" in that each input consisted only of congealed efforts and sacrifice. As a result, each input supply

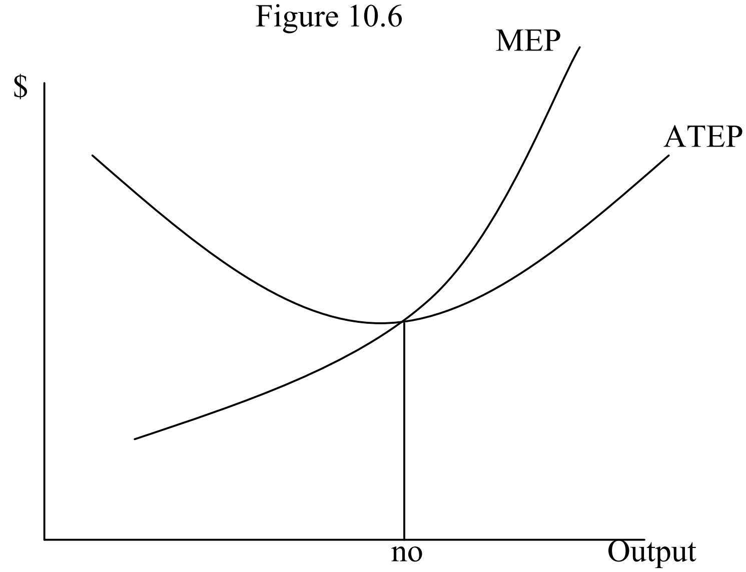
[7] Footnote on Sraffa, circular production, etc.

price could be completely resolved into its constituent parts. Marshall then called the inputs *factors of production*.[8]

Cost Curves and Supply Curves

In the short period, the stock of instrumental capital is fixed and the input supply prices are given. Hence as more direct material and labor inputs are applied diminishing returns sets in, and hence marginal expenses of production increases:

Figure 10.6



no – normal output where MEP = ATEP;

ATEP – average total expenses of production; and

---

[8] An example of this is the following. Assume a simple economic system in which labor and sacrifice are the primary inputs; the production scheme would look as follows:

$l_1 \rightarrow ICG_1$ (instrumental capital good one)

$ICG_1 + l_2 \rightarrow ICG_2$

$ICG_2 + l_3 \rightarrow CG$ (consumption good).

The input supply price of ICG_2, for example, consists of the labor efforts denoted by l_1 and l_2 and the sacrifice of not immediately consuming ICG_1 and ICG_2 when immediately produced. Thus each input supply price can be resolved into efforts and sacrifices which in turn means that each output supply price can be resolved into efforts and sacrifice. Consequently, it is possible, given Marshall's factor assumption, to go from the expenses of production to the real costs of production. (Of course the problem still exists concerning the possibility of aggregating inhomogeneous efforts and sacrifices; a problem Marshall never really solved.)

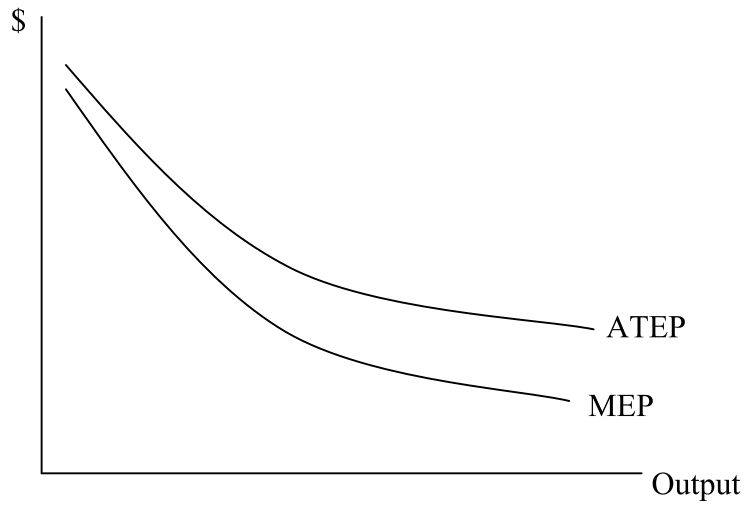
MEP – marginal expenses of production or marginal prime costs.

Because the MEP represents the expenses the firm must recover in the short period if it is going to produce a given amount of output, the MEP curve traces out the firm's supply curve in the short period.[9]

In the long period, the stock of instrumental capital is variable in that they become "free capital" and hence can be invested into any kind of fixed capital. Consequently the amount of capital invested in the production of any good represents the sacrifices by the capitalist and which can be denoted as K. In addition, labor services become liberated in the long period and thus can seek out the highest wage rate; hence the amount of labor services (including management services) used in the production of any good, hence the amount of effort used in production, can be denoted as L. Because of increasing returns to scale, increasing both K and L by the same percentage will increase output by more than that percentage. Converting these real costs of production to expenses of production, we find that because of increasing returns to scale in the long period, both the ATEP and MEP decrease with output:

[9] The MEP curve has the functional form of MEP = f(output). To have a supply curve, its function form must be output = F(MEP) where MEP is equal to the supply price. For the MEP to trace out the supply curve, the independent variable of the MEP curve must become the dependent variable of the supply curve without changing axis. This is one of Marshall's distinctive contributions to economics.

Figure 10.7



CHAPTER 11

MODERN THEORY OF PRODUCTION[10]

Introduction

Before delving directly into production and costs, a number of terms and concepts must first be defined and delineated, starting with the concept of *relative scarcity* (or scarcity in short) followed by time, production function and limitational and limitative inputs. [More development]

<u>Relative Scarcity</u>

Relative scarcity is defined as the relation of given individual desires to the available world of objects that are capable of fulfilling those desires.[11] Such objects are 'relatively scarce' when the available supply is inadequate to completely satisfy the desires of individuals. That is, given a world of objects ($x_1, \ldots, x_n, x_{n+1}, \ldots, x_z$) and given *v* consumers each with a preference (utility) function [$\mu_1(y_1, \ldots, y_z), \ldots, \mu_v(y_1, \ldots, y_z)$], the objects become relatively scarce when the consumers' demand for them is greater then the amount in existence. Assuming $x_{iD} > x_{iS}$ $i = 1, \ldots, n$, the first n objects can be appropriated and exchanged; in fact the properties of ownership and exchanged follow from the notion of relative scarcity. Thus, the utility function for each consumer is truncated to n goods. However, since the array of given objects does not have to match the preferences of consumers, which means what is demanded is not matched by any object—therefore, none of the objects would be relatively scarce and there would be no exchange.

This simple "exchange" view of scarcity becomes more complicated when production is involved, that is the use of scarce objects to produce an increased amount of objects that are also scarce. This could involve combining the existing scarce factors to produce more of an existing scarce object

---

[10] Much of the material in this chapter and the following chapters come from Ferguson (1979).

that is used for consumption. Or it could involve the combining of the scarce factors to produce a new but still scarce consumption good. All of these produced goods are scarce and hence are *economic goods*; this means that they are appropriable and have exchange value. Thus produced economic goods (whether new or not) are original scarce objects but in a different form and therefore do not only have the same properties as them, but also have the same end point, the satisfaction of wants:

```
                                          |→ existing scarce objects for consumption
                                          |
Given Scarce Objects → Production----|→ new scarce consumption goods
          |                               |
          |→ immediate consumption        |→ new scarce capital goods → scarce consumption goods
```

In characterizing given scarce objects in this manner, they take on the attributes of being non-produced, that is in Marshall's view they are either land or labor. Moreover, the produced economic goods do not replenish the given scarce objects; hence if all the scarce goods are used up in production, then once consumption takes place the economic system ceases to exist.[12] Given this discussion of relative scarcity the following terms can be defined:

> *output* – any consumption good or service whose fabrication or creation requires one or more scarce resources; and
>
> *inputs (factors of production)* – any scarce resource used in the production of any consumption good or service.

There are two broad groups of inputs: *material* inputs – those inputs that are used up or consumed in the process of production; and *service* inputs – those inputs that render a flow of services that are consumed in the production processes but the inputs themselves are not consumed.

---

[11] For neoclassical economics, the concept of absolute scarcity has no meaning. Either an object is relatively scarce or not scarce that is it is a free good. However, the extent to which a good is relatively scarce can vary and is revealed in the magnitude of the good's price. [Robbins Essay]

[12] MORE DISCUSSION HERE.

Time

The *short period (run)* is that period of analytical time during which the quantity of one or more inputs can not be changed. The quantity of the input or of its maximum flow of services is fixed and can be neither augmented nor diminished. Such an input is called a *fixed input*. The *long period (run)* is that period of time in which all inputs are variable. We shall assume that a firm's productive activities are so arranged that the production of one time period is entirely separate from the production of the proceeding and subsequent time periods. In addition we shall assume that the firm is interested in the activity of only one period at a time, and that this activity is determined exclusively by the conditions prevailing in that period and is independent of any other conditions.[13]

Production Function

A production function shows the maximum output attainable from any specific set of inputs, that is any set of quantities of material inputs and flow of services from other inputs. There are two broad classes of production functions: *variable-proportions production function* in which the same level of output may be produced by two or more combinations of inputs and its essential technological feature is input substitutability; and *fixed-proportions production functions* in which each level of output technologically requires a unique combination of inputs. If the technological determined input-output ratio is independent of the scale of production for each input, the production process is characterized by *fixed input coefficients*. If the input-output ratios are not independent of scale, but if all pairs of input ratios are constant, then the production process is characterized by *fixed proportional input coefficients*.[14]

---

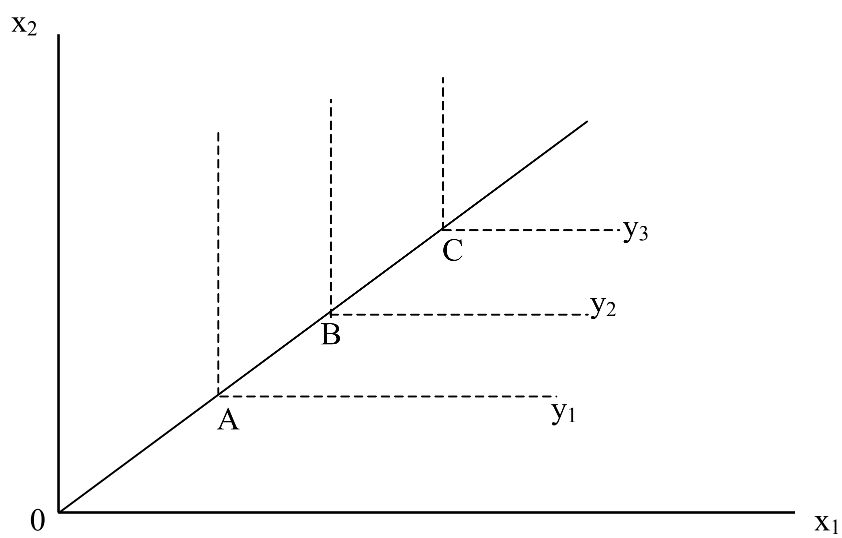
[13] The implication of this for a capital-labor model economy is that in the short period the produced outputs are different from their inputs with respect to time; hence the inputs take on the characteristics of being given scarce objects. Such a dichotomy, however, is not possible in the long period.

[14] A third kind of production function is conceivable in which the pairs of input ratios are technically fixed for each level of output, but they change as output changes. It is, however, not discussed in what

<u>Limitational and Limitative Inputs</u>

For a *limitational input* an increase in its usage is necessary, but not sufficient, condition for an increase in output. Such a situation is found in fixed-proportions production function. A limitational input can be shown in terms of right-angle isoquants:

Figure 11.1



Limitative input, on the other hand, is an input in which the increase in its usage is both necessary and a sufficient condition for an increase in output. Such an input occurs in a variable-proportions production function.

---

follows because it implies that some of the inputs change qualitatively when the scale of production changes, hence violating the comparability conditions for proportional input variation. Moreover, it undermines the notions of fixed inputs and given array of inputs so essential for the law of diminishing returns and the discussion of production *per se*. It should be noted that this perverse view of production is quite similar to Marshall's discussion of production. [Ferguson 1979: 8]

Variable-Proportions Production Function

The variable-proportions production function is written as $y = f(x_1, \ldots, x_n)$ where y is the output and x_i is a factor input. The assumptions and properties underlying the production function are the following (Ferguson 1979 60):

(1) it is singled valued everywhere, continuous, and well-defined over the range of inputs yielding non-negative outputs;

(2) it is a monotonic function, so that if x_i increases, y increases;

(3) it is a strictly concave (strong assumption) or a strictly quasi-concave function (weak assumption);

(4) it has strictly convex level curves (or isoquants);

(5) it has first and second derivatives and its level curves are continuous and twice differentiable; consequently the inputs are internally homogeneous and hence continuously variable;

(6) it is fixed, that is there is no technical change; and

(7) all $x_i > 0$, $i = 1 \ldots n$ if $y > 0$ and x_i and y are bounded from below by zero.

There are two kinds of production functions: *homogeneous* and *inhomogeneous* production function. A production function is homogeneous of degree r if multiplication of each of its independent variables by a constant k will alter the value of the function by the proportion k^r, that is, if $f(kx_I, \ldots, kx_n) = k^r f(x_1, \ldots, x_n)$. For example, $A(kx_1)^{\alpha}(kx_2)^{\beta} = k^{\alpha+\beta}(Ax_1^{\alpha}x_2^{\beta}) = k^{\alpha+\beta}Q$—this production function is called a Cobb-Douglas production function—see Chapter 13. Otherwise a production function is inhomogeneous if it does not adhere to the above manipulation. Being more general, only inhomogeneous production functions are utilized in this chapter (although homogeneous functions are used in examples)—More.

Properties of the Production Function: Single Input Variation

The first property of the production to be examined concerns variations in output due to a variation of a single input.

Marginal Product

The *marginal product* of a input is the addition to total product attributable to the addition of one unit to the production process, with all other inputs remaining unchanged:

$$\frac{\partial y}{\partial x_i} = \frac{\partial f(x_1, \ldots, x_n)}{\partial x_i} > 0$$

Considering the second partials to determine the direction the marginal product is "moving":

$$\frac{\partial^2 y}{\partial x_i^2} = \frac{\partial^2 f(x_1, \ldots, x_n)}{\partial x_i^2} < 0 \text{ if the production is strictly concave;}$$

$$\gtreqless 0 \text{ if the production function is strictly quasi-concave.}$$

The second partial is also called a *direct acceleration coefficient* since it shows whether the marginal product is increasing or decreasing at any point. Finally, there is the second cross partial, also called a *cross-acceleration coefficient* since it shows how the marginal product of x_i varies when the usage of x_j varies:

$$\frac{\partial^2 y}{\partial x_i \partial x_j} = \frac{\partial^2 f(x_1, \ldots, x_n)}{\partial x_i \partial x_j} \lesseqgtr 0.$$

If $\partial^2 y/\partial x_i\partial x_j > 0$, the inputs are complementary; and if $\partial^2 y/\partial x_i\partial x_j < 0$, then the inputs are competitive. This leads us to the *law of variable proportions*:

> With a given technology, if the quantity of one productive service is increased by equal increments, the quantities of the other productive services remaining fixed, the resulting increment of product will decrease after a certain point: (a) if the production function is strictly concave, then $\partial^2 y/\partial x_i^2 < 0$ throughout or the *law of diminishing returns* is pervasive; and (b) if

the production function is strictly-quasi concave, then $\partial^2 y/\partial x_i^2 \leq >0$, but over some range $\partial^2 y/\partial x_i^2 < 0$.

In fact every acceptable production function is characterized by a range of diminishing returns.

Average Product and Returns

The *average product* is defined as the quantity of output per unit of the input used, that is it is an output-input ratio: $AP_i = y/x_i = f(\mathbf{x})/x_i$. Average returns, that is the shape of the average product curve, is the following:

$$\frac{\partial AP_i}{\partial x_i} = \frac{\partial (y/x_i)}{\partial x_i} = \frac{(MP_i - AP_i)}{x_i} \begin{cases} < 0 \text{ if } MP_i < AP_i \\ > 0 \text{ if } MP_i > AP_i \\ = 0 \text{ if } MP_i = AP_i. \end{cases}$$

The average product is at a maximum if $\frac{\partial (y/x_i)}{\partial x_i} = 0$ and $\frac{\partial^2 (y/x_i)}{\partial x_i^2} < 0$.[15]

Relationship Between Total Output, Average Product, and Marginal Product

To discuss the relationship between total output, average product, and marginal product, let us introduce the following terms:

output elasticity of the ith input – the proportional change in output induced by a change in the ith input relative to the given proportional change in this input:

$$\varepsilon_i = \frac{\partial y}{y} \div \frac{\partial x_i}{x_i} = \frac{\partial y x_i}{\partial x_i y} = \frac{MP_i}{AP_i}$$

The output elasticity of the ith input can now be used to define the elasticity of the average product:

elasticity of the average product - $\dfrac{\partial (y/x_i) x_i}{(\partial x_i)(y/x_i)} = \dfrac{MP_i}{AP_i} - 1 = \varepsilon_i - 1.$

Now we are in a position to discuss the relationships:

(1) if, $\varepsilon_i > 1$, we have increasing average returns and total output is increasing;

(2) if $\varepsilon_i = 1$, we have zero average returns, average product is at its maximum, and total output is increasing;

(3) if $0 < \varepsilon_i < 1$, we have decreasing average returns, the marginal and average product decreasing with the former decreasing faster than the latter, and total output increasing but at a decreasing rate;

(4) if $\varepsilon_i = 0$, we have decreasing average returns, marginal product is zero, and total output is at its maximum; and

(5) if $\varepsilon_i < 0$, we have decreasing average returns, negative marginal product, and declining total output.

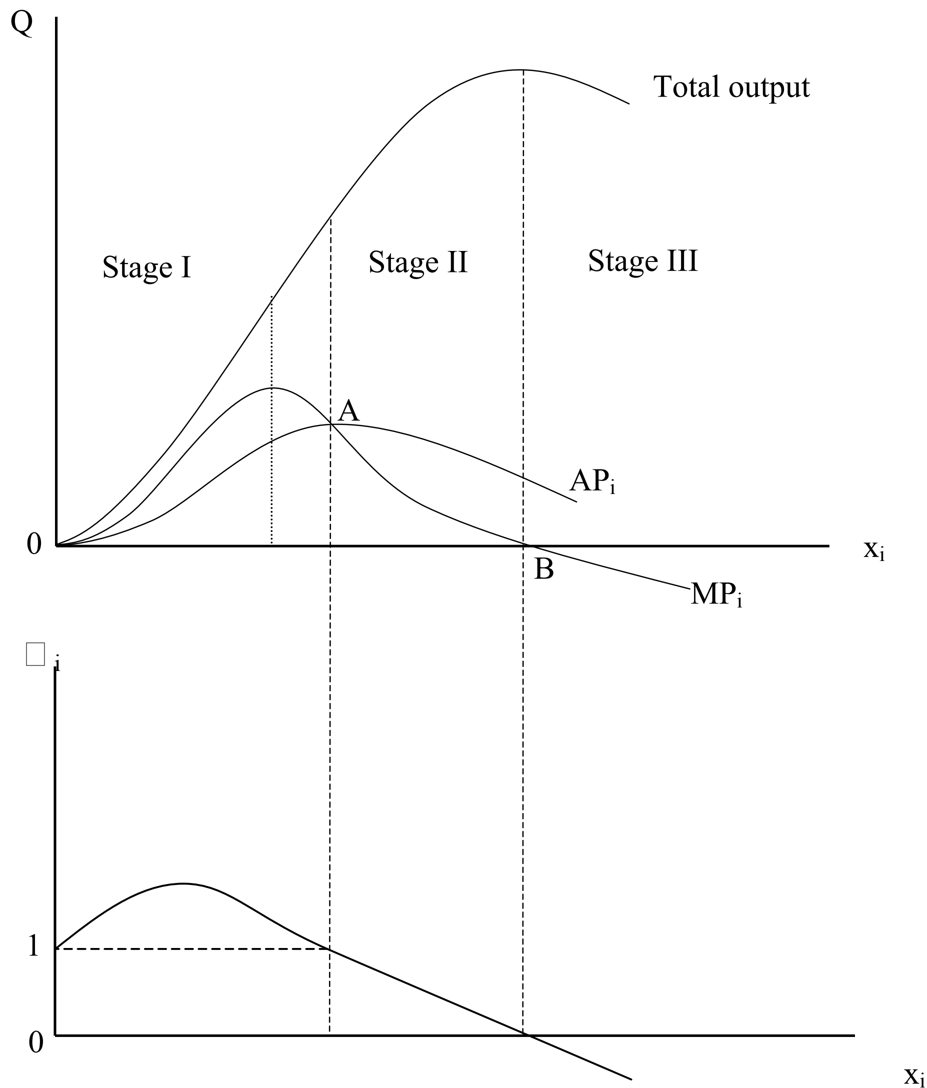
The above relationships can be expressed in terms of *stages of production*—see Figure 1 below.

In stage I of production, the fixed inputs are in excess of what is needed. That is, if $\varepsilon_i > 1$, a given increase in usage of x_i results in a proportionately greater increase in y. Consequently an increase in output could be secured by not utilizing all the fixed inputs. Specifically, production would take place in a manner akin to returns to scale. But in this case, the product curves would have to be redrawn because the effective amounts of fixed inputs have changed. Hence production will not take place in stage I of production (Maxwell 1965). Turning to stage III of production we find that $MP_i < 0$ and total output is declining. Hence a firm will not operate in this region, assuming that it wants to maximize its output given inputs. Finally, since production is not taking place in stage I and III of production, it has to take place in stage II. Here we have $MP_I < AP_I$, diminishing marginal and average products, and increasing total output. Thus, production takes place where the production function is locally strictly concave, with marginal product positive but declining and the law of diminishing returns operates.

---

[15]Proof: taking the second derivative we have $\partial^2(y/x_i)/\partial x_i^2 = \partial^2 y/\partial x_i^2/x_i - (2/x_i)^2[\partial y/\partial x_i - y/x_i] = \partial^2 y/\partial x_i^2/x_i < 0$. Thus AP_i can only be at a maximum if the MP_i is decreasing, that is when the law of diminishing returns is operative.

Figure 11.2



<u>Properties of the Production Function: Proportional Input Variation</u>

We are now concerned with proportional variations in all inputs and the impact on output. This is known as *returns to scale*. Because of the assumptions made above that inputs cannot change qualitatively as the scale of production changes, returns to scale cannot include Marshall-like production relationships. Rather by reducing returns to scale to a quantity (vs. a quality) relationship, it is reduced to comparing proportional changes in quantities of inputs and outputs. Thus returns to scale is discussed solely in physical terms. So let us consider the following input point $\lambda x_1, \ldots, \lambda x_n$ where $\lambda$is a multiplier and the corresponding output is $y = f(\lambda \mathbf{x})$. Thus the elasticity of output with respect to an equi-proportional variation of all inputs is:

function coefficient $= \varepsilon_f = \underline{dy\lambda}/d\lambda y > 1$ – increasing returns to scale

$= 1$ – constant returns to scale

< 1 – decreasing returns to scale.

Now let us consider the relationship between the function coefficient, the output elasticity of the ith input, and the elasticity of the average product. If we totally differentiate the production function we get

$$dy = \frac{\partial f}{\partial x_1} dx_1 + \ldots + \frac{\partial f}{\partial x_n} dx_n = \frac{\partial f\, x_1}{\partial x_1 x_1} dx_1 + \ldots + \frac{\partial f\, x_n}{\partial x_n x_n} dx_n$$

Since $dx_i/x_i = d\lambda/\lambda$ we have

$$dy = \left[\frac{\partial f\, x_1}{\partial x_1} + \ldots + \frac{\partial f\, x_n}{\partial x_n} \right] \frac{d\lambda}{\lambda} \text{ or}$$

$$\frac{dy\lambda}{d\lambda} = \frac{\partial f\, x_1}{\partial x_1} + \ldots + \frac{\partial f\, x_n}{\partial x_n}$$

Dividing by y, we get

$$\varepsilon_f = \frac{dy\lambda}{d\lambda y} = \frac{\partial f\, x_1}{\partial x_1 y} + \ldots + \frac{\partial f\, x_n}{\partial x_n\, y}$$

$$= \varepsilon_1 + \ldots + \varepsilon_n \text{ since } \partial f/\partial x_n = \partial y/\partial x_n.$$

Thus the function coefficient is the sum of all output elasticities. Moreover, the ultimate proportional change in output is the same whether all outputs are varied simultaneously or one at a "time". Finally, let us consider the relationship between returns to scale and average returns:

$$\frac{d(y/x_i)\lambda x_i}{d\lambda\, y} = \frac{dy\lambda}{d\lambda\, y} - \frac{dx_i\lambda}{d\lambda\, x_i} = \varepsilon_f - 1$$

Therefore if $\varepsilon_f > 1$, then the average product is increasing on that portion of the production function where increasing returns to scale exists; if $\varepsilon_f = 1$, the average product is constant on that portion of the production function where constant returns to scale exists; and if $\varepsilon_f < 1$, the average product is decreasing on that portion of the production function where decreasing returns to scale exists.

Properties of the Production Function: Simultaneous Input Variation

Because a strictly quasi-concave production function has strictly convex level curves (or hyperplanes), the concept of *isoquant* can be introduced. An isoquant is a locus of input combinations each of which is capable of producing the same amount of output; and it is denoted as $y^o = f(x_1, \ldots x_n)$ where y^o is a constant. The slope of the isoquant for x_i and x_j (keeping all other inputs constant) is

$$dy^o = \frac{\partial f}{\partial x_i} dx_i + \frac{\partial f}{\partial x_j} dx_j = 0 \text{ or}$$

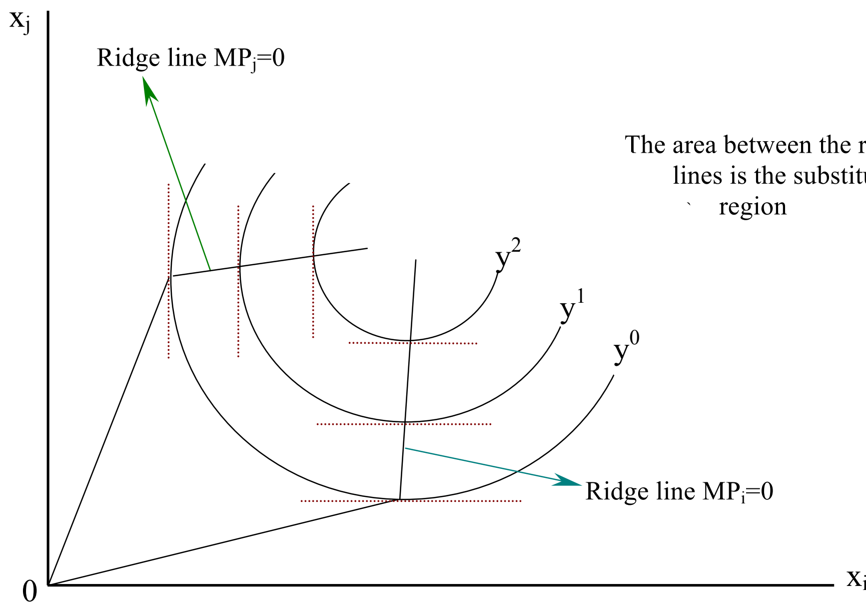
$$-\frac{dx_j}{dx_i} = -\frac{\partial f/\partial x_i}{\partial f/\partial x_j} = -\frac{MP_i}{MP_j} = MRTS_{ji}$$

The slope of the isoquant is called the *marginal rate of technical substitution* of input j for input i ($MRTS_{ji}$). It is the number of units by which the usage of x_j is reduced when the usage of x_i is expanded by one unit so as to maintain a constant amount of output. Because of the assumption of a strictly quasi-concave production function, the $MRTS_{ji}$ diminishes as x_i is substituted for x_j

Substitution Region

The substitution region is that portion of input space in which all isoquants are negatively sloped. It is thus the region in which one input is substituted for another while maintaining a constant amount of output—see Figure 2. In addition, diminishing marginal products exists throughout the region. The degree of convexity of the isoquant is an indication of the ease with which one input can be substituted for another in the production process. The more convex the isoquant, the greater is the effect that a change in the input ratio along the isoquant has upon the MRTS. Consequently, the more convex the

Figure 11.3

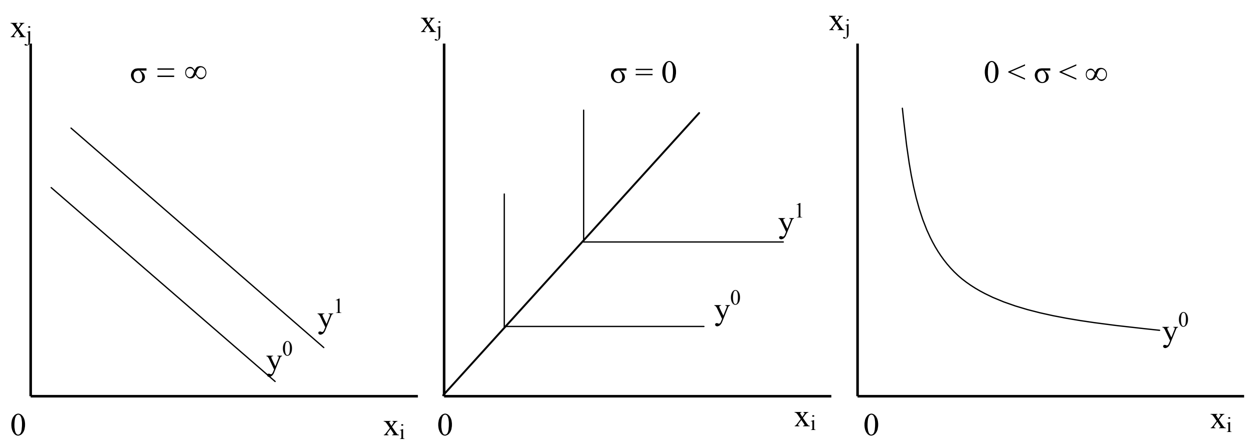


isoquant, the faster will the required substitution of one input for another rise as the amount of the latter input utilized is decreased. This is what is meant by difficulty in substitution. Conversely, the less convex is the isoquant, the smaller is the effect of changes in input ratios upon the MRTS. Consequently one input can be substituted for another without a rapid increase required in one input to hold the output level constant when the amount of the other input is diminished. This can summarized in terms of the *elasticity of substitution* that is defined as

$$\sigma = \frac{dx_i}{x_i} \frac{MRTS_{ji}}{dMRTS_{ji}} \geq 0$$

There are three general and limiting cases of elasticity of substitution:

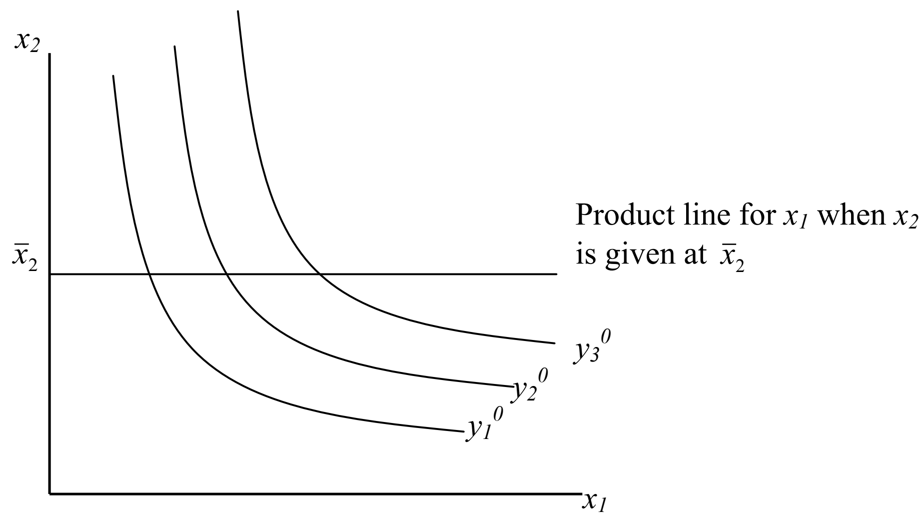
Figure 11.4



Product Lines and Isoclines

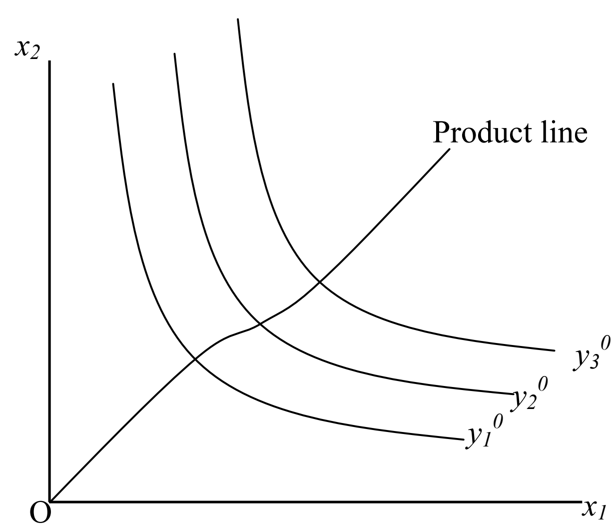
A *product line* shows the movement from one isoquant to another as we change either a single input or both inputs. For single input variation and product line see Figure 4:

Figure 11.5



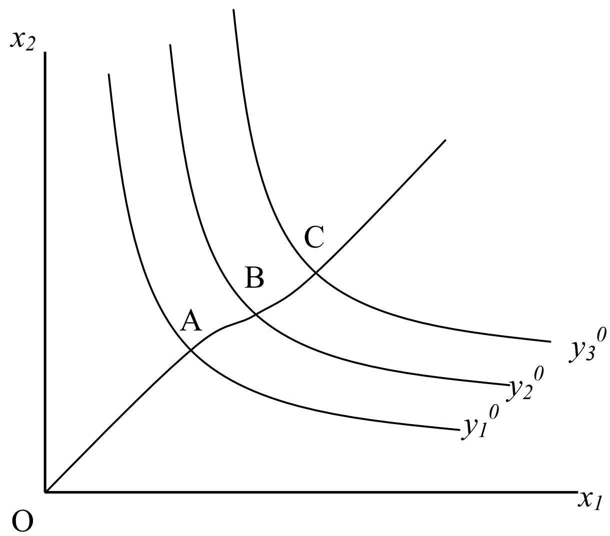
For proportional input variation and product line see Figure 5. As long as both inputs are variable the product line will pass through the origin.

Figure 11.6



<u>Isocline</u> is a locus of points of different isoquants at which the MRTS is constant, see Figure 6. Thus an isocline is a special kind of product line.

Figure 11.7



line OABC is an isocline since the MRTS at points A, B, and C are all the same

CHAPTER 12

MODERN THEORY OF COSTS – CHANGES IN THE LEVEL OF OUTPUT

To convert production theory into a cost theory, the prices of the factor inputs have to be introduced. This is done in terms of the *total costs* of using a particular factor input combination: Total Cost (TC) = $p_1x_1 + \ldots + p_nx_n = \mathbf{px}$ where **px** is the combination of factor inputs and their prices. Because the partial equilibrium framework of analysis with its explicit *ceteris paribus* methodology underpins the theory of costs, all input prices are exogenously fixed, that is they are parameters.

Optimal Input Combination – Cost Minimization Approach

With the introduction of TC = **px**, an *isocost curve* can be derived which is defined as all the combinations of the factor inputs that may be purchased or hired for a given expenditure of funds. Its slope at every point is the negative of the factor input-price ratio. From this it follows that when factor input prices are fixed, isocost curves are straight lines. Given the isocost curve, it is now possible to determine the optimal input combination using the cost minimization approach.[16] The objective of this approach is to minimize the cost of producing a given amount of output. This is formulated as minimizing TC = $p_1x_1 + \ldots + p_nx_n$ subject to $f(\mathbf{x}) = y^o$. Setting up the Lagrangian function, we have

$$L = p_1x_1 + \ldots + p_nx_n + \lambda[y^o - f(\mathbf{x})] = \mathbf{px} + \lambda[y^o - f(\mathbf{x})].$$

The first order conditions are

$$L_1 = p_1 - \lambda\partial f(\mathbf{x})/\partial x_1 = 0$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$L_n = p_n - \lambda\partial f(\mathbf{x})/\partial x_n = 0$$
$$L_\lambda = y^o - f(\mathbf{x}) = 0$$

Interpreting the FOC:

[16] There is also an output maximization approach that gives nearly the same results. It is delineated in the Appendix to the chapter.

(1) $\partial f(\mathbf{x})/\partial x_i$ is the marginal product of the ith factor input

(2) $\frac{\partial f(\mathbf{x})/\partial x_i}{\partial f(\mathbf{x})/\partial x_j} = \frac{MP_i}{MP_j} = MRTS_{ji} = \frac{p_i}{p_j}$

which says that the optimal input combination is attained at the production point where the MRTS between every pair of inputs is equal to the price ratio prevailing between those pairs of inputs.

(3) $\frac{\partial f(\mathbf{x})/\partial x_1}{p_1} = \ldots = \frac{\partial f(\mathbf{x})/\partial x_n}{p_n} = \frac{1}{\lambda}$

which says that the optimal input combination is attained at the production point where the marginal productivity of the last dollar spent on resources ($1/\lambda$) is the same in every use. That is, the marginal productivity of a dollar's worth of any one resource must be equal to the marginal productivity of a dollar's worth of any other resource.

(4) $y^o = f(\mathbf{x})$

which says that the combination of inputs will produce the given amount of output.

To see if a cost minimizing position has been attained, the second order conditions need to be examined:

$L_{11} = -\lambda\partial^2 f(\mathbf{x})/\partial x_1^2; \ldots, L_{1n} = -\lambda\partial^2 f(\mathbf{x})/\partial x_1\partial x_n; L_{1\lambda} = -\partial f(\mathbf{x})/\partial x_1$

………………………………………………………………..

$L_{\lambda 1} = -\partial f(\mathbf{x})/\partial x_1; \ldots, L_{\lambda n} = -\partial f(\mathbf{x})/\partial x_n; L_{\lambda\lambda} = 0$

Putting this into a bordered Hession matrix and then taking its determinant, we have

$$\begin{vmatrix} -\lambda\partial^2 f(\mathbf{x})/\partial x_1^2 & \ldots & -\lambda\partial^2 f(\mathbf{x})/\partial x_1\partial x_n & -\partial f(\mathbf{x})/\partial x_1 \\ \ldots & \ldots & \ldots & \ldots \\ -\partial f(\mathbf{x})/\partial x_1 & \ldots & -\partial f(\mathbf{x})/\partial x_n & 0 \end{vmatrix} < 0$$

because the production function is strictly quasi-concave. Hence we have cost minimization. Solving the FOC, we get *constant output factor input demand functions* and the Lagrangian multiplier function:

$x_i^e = \psi_i(p_1, \ldots, p_n, y^o)$

$\lambda^e = \psi_\lambda(p_1, \ldots, p_n, y^o)$.

Total Cost Function

The constant output factor input demand functions are the basis on which the total cost function and subsequently the various cost curves (and eventually supply curves) are derived. These demand functions show the changes in quantities demanded of inputs resulting from *price changes* (when equilibrium adjustments are restricted to movements along the isoquant) and from *output changes*. Thus to construct a total cost function in which output is one of its parameters, the constant output factor input demand functions are substituted into TC = **py** and we get $TC = p_1x_1^e + \ldots + p_nx_n^e = TC^*(\mathbf{p}, y^o)$ which is defined at the *total cost function* and gives the minimum costs for producing any given amount of output.[17] From the total cost function, we can derive our analysis of short and long period costs.[18]

Short Period Theory of Costs

The short period is defined in terms of there being at least one factor input that is fixed in its amount. In particular, we shall assume that inputs $x_1 \ldots x_h$ are variable in amount and are represented by their factor input demand functions; and that inputs $x_{h+1} \ldots x_n$ are fixed in amount. Therefore the constant output factor input output demand functions and the Lagrangian multiplier function have the following form:

$x_i^e = \psi_i(p_1, \ldots, p_h, y^o)$

---

[17] The properties of the total cost function are the following (Varian 1992: 72):

(1) non-decreasing in **p**; if $\mathbf{p}' \geq \mathbf{p}$, then $TC^*(\mathbf{p}', y^o) \geq TC^*(\mathbf{p}, y^o)$.
(2) homogeneous of degree 1 in **p**: $TC^*(t\mathbf{p}, y^o) = tTC^*(\mathbf{p}, y^o)$ for $t > 0$.
(3) concave in prices.
(4) is continuous as a function of prices.
(5) $TC^*(\mathbf{p}, y^o) >$ for $\mathbf{p} \geq 0$ and $y^o > 0$.
(6) concavity in prices is assured as long as the cross second partials in prices are equal.

[18] Conventional cost theory assumes that factor input prices are parameters and then carries out a comparative static analysis of costs with respect to output. Such a procedure means that each point on the cost curves are unrelated, resulting in cost curves being made up of equilibrium points that depict the

$\lambda^e = \psi_\lambda(p_1, \ldots, p_h, y^o)$.

Thus, the short period total cost function becomes $TC = p_1x_1^e + \ldots + p_hx_h^e + p_{h+1}x_{h+1}^F + \ldots + p_nx_n^F$.

From this we can derive the following categories of short period costs:

(1) variable costs: $VC = p_1x_1^e + \ldots + p_hx_h^e$

(2) average variable costs: $AVC = \underline{p_1x_1^e + \ldots + p_hx_h^e}/y^o$

(3) fixed costs: $FC = p_{h+1}x_{h+1}^F + \ldots + p_nx_n^F$

(4) average fixed costs: $AFC = \underline{p_{h+1}x_{h+1}^F + \ldots + p_nx_n^F}/y^o$

(5) average total costs: $ATC = \underline{p_1x_1^e + \ldots + p_hx_h^e + p_{h+1}x_{h+1}^F + \ldots + p_nx_n^F}/y^o$

To construct a short period theory of costs, the behavior of marginal costs with respect to output must be delineated. To derive *marginal costs*, total costs is differentiated with respect to output:

$$MC = \frac{\partial TC}{\partial y} = \frac{\partial[p_1x_1^e + \ldots + p_hx_h^e + p_{h+1}x_{h+1}^F + \ldots + p_nx_n^F]}{\partial y} = \frac{\partial[\, p_1x_1^e + \ldots + p_hx_h^e]}{\partial y}$$

From the first order conditions, we know that $p_1 = \lambda^e\partial f(\mathbf{x}^e)/\partial x_1$. Therefore, substituting, we get

$$MC = \frac{\partial TC}{\partial y} = \lambda^e\left[\frac{\partial f(\mathbf{x}^e)\partial x_1^e}{\partial x_1\ \partial y} + \ldots + \frac{\partial f(\mathbf{x}^e)\partial x_h^e}{\partial x_h\ \partial y}\right]$$

Furthermore, from first order conditions, we know that $L_\lambda = f(x_1^e, \ldots, x_h^e, x_{h+1}^F, \ldots, x_n^F) - y^o \equiv 0$;

differentiating by y, we get

$$\frac{\partial f(\mathbf{x}^e)\partial x_1^e}{\partial x_1\ \partial y} + \ldots + \frac{\partial f(\mathbf{x}^e)\partial x_h^e}{\partial x_h\ \partial y} \equiv 1.$$

Substituting into the above, we have

$$MC = \frac{\partial TC}{\partial y} = \lambda^e$$

relationship of costs to alternative amounts of output. Hence it is not possible to view cost curves as showing the relationship between costs and output through time.

To determine the shape of the marginal cost curve, we have to return to the FOC, substitute in the demand and Lagrangian multiplier functions and differentiate with respect to y^o:

$$L_{1\,y} = \frac{-\lambda^e \partial^2 f(\mathbf{x^e}) \partial x_1^e}{\partial x_1^2 \; \partial y} - \ldots - \frac{\lambda^e \partial^2 f(\mathbf{x^e}) \partial x_h^e}{\partial x_h^2 \; \partial y} - \frac{\partial f(\mathbf{x^e}) \partial \lambda^e}{\partial x_1 \; \partial y} \equiv 0$$

………………………………………………………………..

$$L_{h+1\,y} = \frac{-\partial f(\mathbf{x^e}) \partial x_1^e}{\partial x_1 \; \partial y} - \ldots - \frac{\partial f(\mathbf{x^e}) \partial x_h^e}{\partial x_h \; \partial y} + 1 \equiv 0$$

Putting into matrix form and letting $f_{ij} = \underline{\partial f(\mathbf{x^e})}/\partial x_i \partial x_j$:

$$\begin{bmatrix} -\lambda^e f_{11} & \ldots\ldots\ldots & -f_1 \\ \ldots\ldots\ldots & & \ldots\ldots\ldots \\ -f_1 & \ldots\ldots\ \ldots\ldots & 0 \end{bmatrix} \begin{bmatrix} \partial x_1^e/\partial y \\ \ldots\ldots\ \ldots \\ \partial \lambda^e/\partial y \end{bmatrix} \equiv \begin{bmatrix} 0 \\ \ldots. \\ -1 \end{bmatrix}$$

Using Cramer's rule, $\partial \lambda^e/\partial y$ can be solved for:

$$\frac{\partial \lambda^e}{\partial y} = \frac{(-1)D_{h+1\ h+1}}{D} > 0$$

since $D < 0$ and $D_{h+1\ h+1} > 0$ because production is restricted to the segment of the production function that is locally strictly concave, that is where the marginal products are declining. Thus the short period marginal cost curve is upward sloping due to the law of variable proportions—specifically the law of diminishing returns.

Using the marginal cost curve, we can now investigate the other cost categories, starting with the average total cost curve. Let us first consider changes in ATC with respect to output:

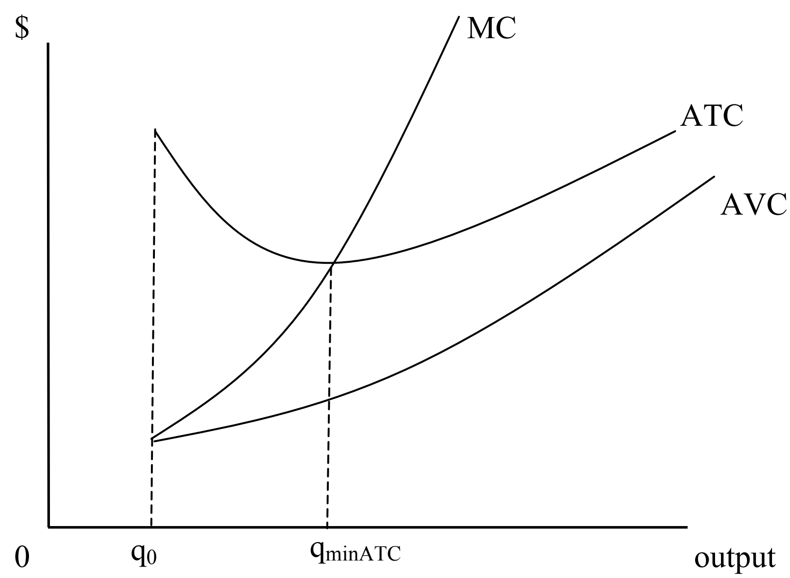
$$\frac{\partial ATC}{\partial y} = \frac{y\partial TC/\partial y - TC}{y^2} = \frac{yMC - TC}{y^2}$$

Since $\partial MC/\partial y > 0$, MC = ATC at its minimum point; hence when MC < ATC, the ATC is falling and when MC > ATC, ATC is rising. Turning to the average variable cost curve, changes in AVC with respect to output is:

$$\frac{\partial AVC}{\partial y} = \frac{y\partial VC/\partial y - VC}{y^2} = \frac{yMC - VC}{y^2}$$

Since $\partial MC/\partial y > 0$, MC = AVC at its minimum point; hence when MC > AVC, the AVC is increasing; moreover since $\partial MC/\partial y = \partial^2 VC/\partial y^2 > 0$, AVC = MC at the initial point of production. Thus we can conclude that the AVC curve increases because of the law of diminishing returns and that the ATC curve is generally U-shaped due to the counter-acting influences of increasing AVC and declining AFC.

Figure 12.1



Long Period Theory of Costs

The long period is defined in terms of all inputs being variable. Therefore the constant output factor input demand functions and the Lagrangian multiplier function have the following form:

$x_i^e = \psi_i(p_1, \ldots, p_n, y^o)$

$\lambda^e = \psi_\lambda(p_1, \ldots, p_n, y^o)$.

Thus the long period total cost function becomes $TC = p_1x_1^e + \ldots + p_nx_n^e = TC^*(\mathbf{p}, y^o)$. From this we can derive average total costs as $ATC = \underline{p_1x_1^e + \ldots + p_nx_n^e}/y^o$. To construct a long period theory of costs, the behavior of ATC and MC with respect to output needs to be connected to laws of return to scale and the function coefficient. To do this let us start with total costs:

$$TC = \lambda^e \left[\frac{\partial f(\mathbf{x^e})}{\partial x_1} x_1^e + \ldots + \frac{\partial f(\mathbf{x^e})}{\partial x_n} x_n^e \right]$$

since $p_i = \lambda^e \partial f(\mathbf{x^e})/\partial x_i^e$. Now because the function coefficient, ε_f, is defined as

$$\varepsilon_f = \frac{dy\lambda}{d\lambda y} = \frac{\partial f\, x_1}{\partial x_1 y} + \ldots + \frac{\partial f\, x_n}{\partial x_n\, y}$$

we can rewrite total costs as $TC = \lambda^e y \varepsilon_f$; and converting to average total costs, we have $ATC = \lambda^e \varepsilon_f$ or $ATC = MC\varepsilon_f$. This latter relationship can be rewritten as $ATC/MC = \varepsilon_f$ which links the laws of returns to scale to long period costs. In particular, if we have increasing returns to scale, $\varepsilon_f > 1$, then MC < ATC which means that the ATC curve is declining; if we have constant returns to scale, $\varepsilon_f = 1$, then MC = ATC which means that the ATC curve is horizontal; and if we have decreasing returns to scale, $\varepsilon_f < 1$, then MC > ATC which means that the ATC curve is increasing.

Let us take a closer look at the relationship between marginal costs and the function coefficient. To do so, we need to introduce the following cost elasticities:

elasticity of total cost: $\kappa = \frac{dTCy}{dyTC} = \frac{MC}{ATC}$

elasticity of average total costs: $\gamma = \frac{dATCy}{dyATC} = \frac{MC}{ATC} - 1 = \kappa - 1$

elasticity of marginal costs: $MC_e = \kappa - 1 + \kappa_e$ where κ_e is the elasticity of κ.

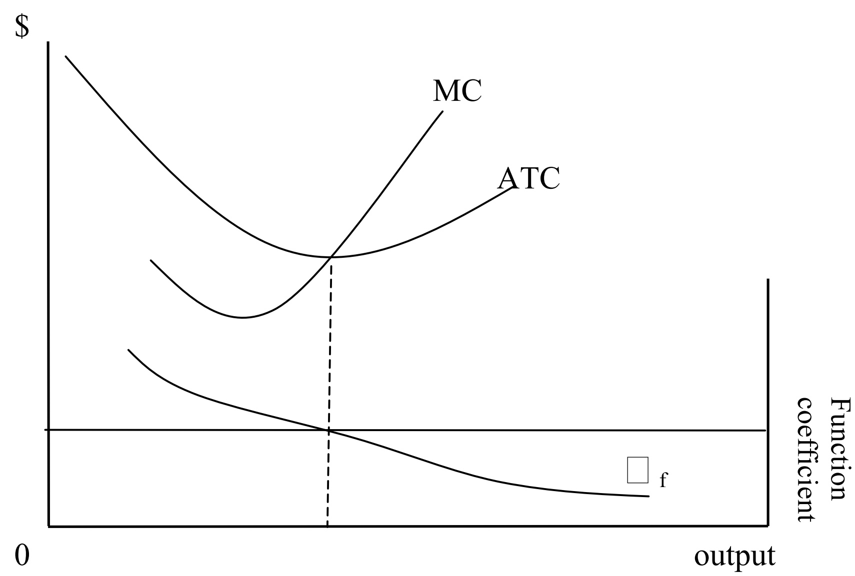
Working with the elasticity of marginal costs, we have

$$\begin{aligned} MC_e &= \kappa - 1 + \kappa_e \\ &= (1/\varepsilon_f) - 1 + (1/\varepsilon_f)_e \text{ since } \kappa = 1/\varepsilon_f \\ &= \underline{(1 - \varepsilon_f)}/\ \varepsilon_f - (\varepsilon_f)_e \text{ where } (\varepsilon_f)_e \text{ is the elasticity of the function coefficient.}^{19} \end{aligned}$$

[19] For a more complete presentation, see Ferguson, 1979, pp. 159 – 163.

So if $\varepsilon_f > 1$, then marginal costs may be decreasing if increasing returns to scale are very strong, so that the negative term $(1 - \varepsilon_f)/ \varepsilon_f$ dominants the necessarily positive term $-(\varepsilon_f)_e$. However, the marginal cost will turn upward at some output rate while $\varepsilon_f > 1$ when the impact of increasing returns to scale becomes dominated by $-(\varepsilon_f)_e$. The relationships between ATC, MC and the function coefficient (or returns to scale) are illustrated in Figure 2 below (Ferguson 1979: 162]:

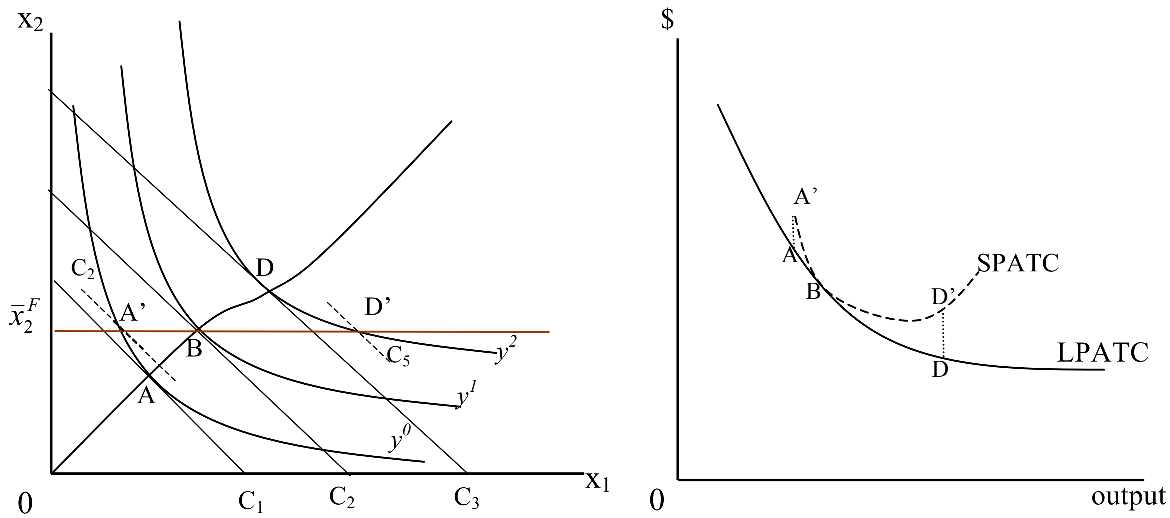
Figure 12.2



Relationship between Short Period Costs and Long Period Costs

Let us assume a two input production function in which x_2 is the fixed input in the short period. Given prices p_1 and p_2, the following figure can be drawn [Ferguson 1979: 155-56]:

Figure 12.3



y^0, y^1, y^2 are isoquants associated with different levels of output;

c_1, c_2, c_3, c_4, c_5 are the isocost curves associated with the different levels of output;

A'BD' generates the firm's short period average total cost curve; and

ABD is the *expansion path* from which the long period average cost curve is derived.[20]

To understand the above, lets suppose for the moment that we are in a short period situation in which the quantity of x_2 is fixed at x_2^F. If production occurs at point B, the short run and long period costs are identical. However, any short period alternative level of output along x_2^F will give rise to higher costs than are entailed when factor input proportions are optimally adjusted (at point B). For example, if the y^0 level of output is produced, the firm must operate at A' rather than A. Total cost is represented by c_2

[20] An expansion path is the locus of input combinations for which the MRTS equals the input-price ratio.

$> c_1$; hence short period average total cost exceeds long period average total costs. A similar statement applies to production at D' on y^2 with short period total cost represented by c_5. Since the isoquants y^0 and y^2 are arbitrary, the above analysis applies to all levels of output above and below y^2. Hence the short period average total cost curve lies above the long period average total cost curve at point other than B, and is precisely tangent at B. By hypothetically altering the quantity of x_2, the *envelope theorem* is established – that is the long period average total cost curve is an envelope of short period average total cost curves.[21]

Constant Output Factor Input Demand Curves

As noted above, constant output factor input demand functions are derive from the first order conditions of the cost minimization approach. Now we are in a position to investigate them. Assuming a long period position substituting x_i^e and λ^e into the forst order conditions and differentiating with respect to p_i, we get the following:

$$\begin{bmatrix} \lambda^e f_{11} & \ldots & -f_1 \\ . & \ldots & . \\ \lambda^e f_{i1} & \ldots & -f_i \\ . & \ldots & . \\ -f_1 & \ldots\ -f_n & 0 \end{bmatrix} \begin{bmatrix} \partial x_1^e / \partial p_i \\ \ldots \\ \partial x_i^e / \partial p_i \\ \ldots \\ \partial \lambda^e / \partial p_i \end{bmatrix} \equiv \begin{bmatrix} 0 \\ . \\ -1 \\ . \\ 0 \end{bmatrix}$$

Using Cramer's rule, $\partial x_1^e / \partial p_i$ (or the shape of the constant output factor input demand function) can be derived:

$$\frac{\partial x_i^e}{\partial p_i} = \frac{(-1)D_{ii}}{D} < 0 \text{ since both } D_{ii} \text{ and } D \text{ are negative.}$$

Hence the constant output factor input demand function for the ith input slopes downward. This is due to changes in quantity demanded being restricted to the original isoquant since output is constant. Using

[21] There is criticism here. In particular, the long run and short run points actually do not touch because they are in different analytical time periods. So illegitimate use of a diagram. MORE

Cramer's rule again, changes in the usage of x_1^e induced by a change in the price of the ith input, output held constant can be solved for:

$$\frac{\partial x_1^e}{\partial p_i} = \frac{(-1)D_{i1}}{D} \lesseqgtr 0 \text{ since } D_{i1} \text{ is a non-principle minor.}$$

If $\frac{\partial x_1^e}{\partial p_i} < 0$ then x_1 and x_i are complements; if > 0 then they are substitutes; and if = 0 then they are independents.[22] Now if the first order conditions are differentiated with respect to p_1, we find the following:

$$\frac{\partial x_i^e}{\partial p_1} = \frac{(-1)D_{1i}}{D} \lesseqgtr 0$$

Since D is symmetric, $D_{i1} = D_{1i}$,

$$\frac{\partial x_1^e}{\partial p_i} = \frac{\partial x_i^e}{\partial p_1}$$

That is, when output is held constant, the change in the usage of x_1^e induced by a change in the price of the ith input is exactly the same as the change in the usage of x_i^e induced by a change in the price of the first input.

Changes in Output and Demand for Factor Inputs

Returning to the first order conditions and now differentiating with respect to output, we get the following equation:

$$\begin{bmatrix} \lambda^e f_{11} & \cdots & -f_1 \\ \vdots & & \vdots \\ \lambda^e f_{i1} & \cdots & -f_i \\ \vdots & & \vdots \\ -f_1 & \cdots \; -f_n & 0 \end{bmatrix} \begin{bmatrix} \partial x_1^e / \partial y \\ \vdots \\ \partial x_i^e / \partial y \\ \vdots \\ \partial \lambda^e / \partial y \end{bmatrix} \equiv \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \vdots \\ -1 \end{bmatrix}$$

Using Cramer's rule, we can see how the demand for a factor input varies with output:

$$\frac{\partial x_i^e}{\partial y} = \frac{(-1)D_{n+1,i}}{D} \lesseqgtr 0 \text{ since } D_{n+1\ i} \text{ is a non-border preserving minor.}$$

[22] In a 2-factor input case, the two factors must be substitutes.

The expression says that if the output level is raised, the factor input levels can either increase or decrease. This is due to the possibility that the factor may be an *inferior factor*. That is, if $\partial x_i^e / \partial y < 0$, then x_i is an *inferior factor input*; if $\partial x_i^e / \partial y > 0$, then x_i is a *normal factor input*. It is also of interest to note the following:

$$\frac{\partial \lambda^e}{\partial p_i} = \frac{\partial x_i^e}{\partial y} \gtrless$$

since $D_{n+1\ i}$ and $D_{i\ n+1}$ are symmetrical and non-border preserving minors.

This means that the rate of change of the marginal cost function with respect to a factor price is equal to the magnitude of the output effect for that factor.

Total Cost Function and the Constant Output Factor Input Demand Function

Let the total cost function be denoted as $TC = p_1x_1^e + \ldots + p_nx_n^e = TC^*(\mathbf{p},y)$. Hence, the constant output factor input demand function of the ith factor input is derived by differentiating with respect to its price:

$$\frac{\partial TC^*(\mathbf{p},y)}{\partial p_i} = \frac{\partial (p_1x_1^e + \ldots + p_nx_n^e)}{\partial p_i} = x_i^e$$

The procedure is called Shephard's Lemma. Now differentiating again with respect to p_i we get:

$$\frac{\partial^2 TC^*(\mathbf{p},y)}{\partial p_i^2} = \frac{\partial x_i^e}{\partial p_i} < 0.$$

Appendix

Optimal Input Combination – Output Maximization Approach

The objective of this approach is to maximize output given total costs. This can be formulated as maximize $y = f(\mathbf{x})$ subject to $TC^o = \mathbf{px}$. Setting up the Lagrangian function we have

$$L = f(\mathbf{x}) + \mu(TC^o - \mathbf{px}).$$

The first order conditions are

$$L_1 = \partial f(\mathbf{x})/\partial x_1 - \mu p_1 = 0$$
$$\cdots\cdots\cdots\cdots$$
$$L_n = \partial f(\mathbf{x})/\partial x_n - \mu p_n = 0$$

$$L_\mu = TC^o - \mathbf{px} = 0$$

Interpreting the FOC:

(1) $\partial f(\mathbf{x})/\partial x_i$ which is the marginal product of the ith factor input

(2) $$\frac{\partial f(\mathbf{x})/\partial x_i}{\partial f(\mathbf{x})/\partial x_j} = -\frac{MP_i}{MP_j} = MRTS_{ji} = -\frac{p_i}{p_j}$$

which says that the optimal input combination is attained at the production point where the MRTS between every pair of inputs is equal to the price ratio prevailing between those pairs of inputs.

(3) $$\frac{\partial f(\mathbf{x})/\partial x_1}{p_1} = \ldots = \frac{\partial f(\mathbf{x})/\partial x_n}{p_n} = \mu = \frac{1}{\lambda}$$

which has the same meaning under the cost minimization approach.

(4) $TC^o = \mathbf{px}$

which says that the combination of inputs, given prices, must equal the given total costs. Since the above equilibrium conditions are the same as in the cost minimization approach, the optimal input combinations are the same in both approaches. To see if a output maximization has in fact been reached, the second order conditions need to be examined:

$L_{11} = \partial^2 f(\mathbf{x})/\partial x_1^2; \ldots, L_{1\mu} = -p_1$

$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$

$L_{\mu 1} = -p_1; \ldots, L_{\mu n} = -p_n; L_{\mu\mu} = 0$

Putting this into a bordered Hession matrix and then taking its determinant, we have

$$\begin{vmatrix} \partial^2 f(\mathbf{x})/\partial x_1^2 & \ldots & \partial^2 f(\mathbf{x})/\partial x_1 \partial x_n & -p_n \\ \ldots & \ldots & \ldots & \ldots \\ -p_1 & \ldots & -p_n & 0 \end{vmatrix} > 0$$

because the production function is strictly quasi-concave. Hence we have output maximization. Solving the FOC, we get *constant cost factor input demand functions* and the Lagrangian multiplier function:

$x_i^e = \phi_i(p_1, \ldots, p_n, TC^o)$

$\mu^e = \phi_\mu(p_1, \ldots, p_n, TC^o)$.

With the ith factor input not being a function of output, constant cost demand functions cannot be used for developing a theory of costs that is based on changes in output.

CHAPTER 13

SPECIAL TOPICS IN PRODUCTION AND COST THEORY

Linearly Homogeneous Production Functions

Properties – Single Input Variation

Assuming a two factor input production function, we can make the following argument:

$$y = f(x_1, x_2) = x_2\, f(x_1/x_2, 1)$$

Hence, the *marginal products* of x_1 and x_2 are:

$$\frac{\partial y}{\partial x_1} = x_2 \frac{\partial f(x_1/x_2, 1)}{\partial (x_1/x_2)} \frac{\partial (x_1/x_2)}{\partial x_1} = \frac{\partial f(x_1/x_2, 1)}{\partial (x_1/x_2)} = f_1' > 0$$

$$\frac{\partial y}{\partial x_2} = f(x_1/x_2, 1) + x_2 \frac{\partial f(x_1/x_2, 1)}{\partial x_2} = f(x_1/x_2, 1) - \frac{x_1}{x_2} f_1' > 0$$

In short, the marginal products are a function of the input ratio x_1/x_2. Moreover, the marginal products are declining.

The *average products* are

$$AP_1 = y/x_1 = (x_1/x_2)\, f(x_1/x_2, 1)$$

$$AP_2 = y/x_2 = f(x_1/x_2, 1)$$

The magnitudes of the average products depend only on the input ratio x_1/x_2.

The output elasticity of the input is then:

$$\varepsilon_1 = \frac{MP_1}{AP_1} = \frac{f_1'}{(x_1/x_2)\, f(x_1/x_2, 1)} = \frac{x_1 f_1'}{y}$$

$$\varepsilon_2 = \frac{MP_2}{AP_2} = 1 - \frac{x_1 f_1'}{x_2\, f(x_1/x_2, 1)} = 1 - \frac{x_1 f_1'}{y}$$

Properties – Proportional Input Variations

Function coefficients--let us first consider the original input point, x_1 and x_2: $y = f(x_1, x_2)$.

Now consider a new input point, λx_1 and λx_2:

$y^* = f(\lambda x_1, \lambda x_2) = \lambda f(x_1, x_2) = \lambda y$

The *function coefficient* is:

$$\varepsilon_f = \frac{y^* - y}{y} \div \frac{\lambda - 1}{1} = \frac{y(\lambda - 1)}{y} \div (\lambda - 1) = 1$$

That is a lineally homogeneous production function exhibits constant returns to scale.

The relationship between function coefficient and output elasticities is:

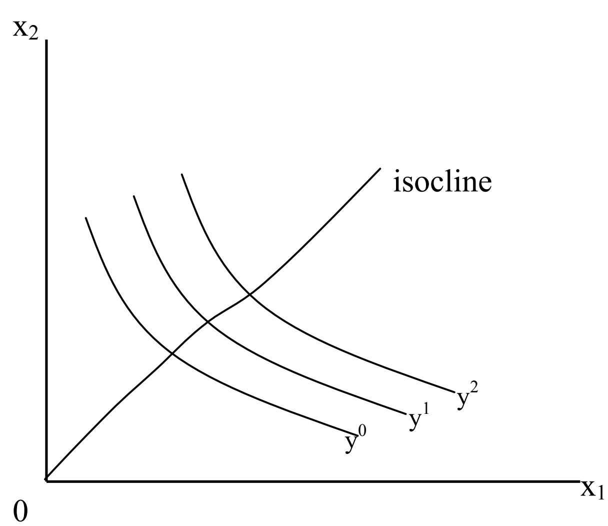
$$\varepsilon_f = \varepsilon_1 + \varepsilon_2 = \frac{x_1 f_1'}{y} + \frac{y - x_1 f_1'}{y} \qquad \text{or}$$

$$\varepsilon_f = \frac{x_1 f_1' - x_1 f_1' + y}{y} = 1$$

Properties – Simultaneous Input Variations

Since each marginal product is a function of the input ratio, the marginal rate of substitution is also a function of the input ratio. Consequently for any given input ratio, the MRTS is the same irrespective of the scale of input usage. Therefore the *isocline* is a ray from the origin.

Figure 13.1



Example - Cobb-Douglas Production Function

The form of the Cobb-Douglas production function is $y = A\, x_1^{a}\, x_2^{1-a}$ where A is a constant and equal to 1. The *marginal products* are then:

$$\frac{\partial y}{\partial x_1} = \frac{\partial (x_1^{a}) x_2^{1-a}}{\partial x_1} = a x_1^{a-1} x_2^{1-a} = a(x_1/x_2)^{a-1} > 0$$

$$\frac{\partial y}{\partial x_2} = \frac{x_1^{a}\, \partial x_2^{1-a}}{\partial x_2} = (1\text{-}a) x_1^{a} x_2^{-a} = (1–a)(x_1/x_2)^{a} > 0$$

$$\frac{\partial^2 y}{\partial x_1^{2}} = a(1–a)\ (x_1/x_2)^{a-2} < 0 \qquad \text{since } (a – 1) < 0$$

$$\frac{\partial^2 y}{\partial x^{2}} = (1–a)(\text{-}a) x_1^{a} x_2^{a-1} < 0 \qquad \text{since } (1–a)(\text{-}a) < 0$$

Thus marginal products are positive and the law of diminishing returns prevails. The *average products* are:

$$\frac{y}{K} = \frac{A\ K^{a}\ L^{1-a}}{K} = A\ K^{a-1}\ L^{1-a} = A\ (K/L)^{a-1}$$

$$\frac{y}{L} = \frac{A\ K^{a}\ L^{1-a}}{L} = A\ K^{a}\ L^{-a} = A\ (K/L)^{a}$$

Now the *average returns*, that is the shape of the average product curves, are:

$$\frac{\partial AP_K}{\partial K} = \frac{MP_K – AP_K}{K} = \frac{A\ a}{K}\ (K/L)^{a-1} - \frac{A}{K}\ (K/L)^{a-1} = (a – 1)\ \frac{A}{K}\ (K/L)^{a-1} < 0$$

$$\frac{\partial AP_L}{\partial L} = \frac{MP_L – AP_L}{L} = \frac{A\ (1\text{-}a)}{L}\ (K/L)^{a} - \frac{A}{L}\ (K/L)^{a} = (\text{-}a)\ \frac{A}{K}\ (K/L)^{a} < 0$$

and the *output elasticity of the inputs* is:

$$\varepsilon_K = \frac{MP_K}{AP_K} = A\ a\ (K/L)^{a-1} \div A\ (K/L)^{a-1} = a \qquad \text{which is constant}$$

$$\varepsilon_L = A\ (1\text{- } a)\ (K/L)^{a} \div A\ (K/L)^{a} = 1 \text{ - } a \qquad \text{which is constant}$$

Regarding the *stages of production*, since the average and marginal product curves are decreasing throughout, production can only take place in the second stage of production, irrespective of the initial endowments of capital and labor.

Turning to the *function coefficient*

$$y = A\,(\lambda K)^a(\lambda L)^{1-a} = \lambda A K^a\,L^{1-a}$$

$$\varepsilon_f = \frac{dy\lambda}{d\lambda y} = (AK^a\,L^{1-a})(\lambda/\,\lambda AK^a\,L^{1-a}) = 1$$

That is, Cobb-Douglas production functions exhibit constant returns to scale. If we consider the *relationship between function coefficient and output elasticities*, we have

$$\varepsilon_f = \varepsilon_K + \varepsilon_L = a + 1 - a = 1$$

Finally, with regard to simultaneous input variation, the *marginal rate of technical substitution* is

$$-\frac{dK}{dL} = \frac{A\,(1 - a)\,K^a\,L^{-a}}{A\,a\,K^{a-1}\,L^{1-a}} = \frac{(1 - a)\,K}{a\,L} = MRTS_{KL}$$

Since the MRTS is dependent solely on the capital-labor ratio, it is not affected by the scale of the input usage. Hence the isocline is a ray from the origin. The *elasticity of substitution* is

$$\frac{d(K/L)}{dMRTS_{KL}}\quad\frac{MRTS_{KL}}{(K/L)} = \frac{1 - a}{a}\;\frac{a}{1 - a}\,(K/L)\,(L/K) = 1$$

Deriving Cobb-Douglas Cost Function, Cost Curves, and Constant Output Factor Input Demand Function for the Short Period

Working with the *optimal input combination – cost minimization approach*, we first consider the short period. Assuming the amount of capital is fixed at K^F, the optimal input combination can be derived in the following manner (A = 1):

$$\mathcal{L} = p_1L + p_2K^F + \lambda\,(y - L^{1-a}\,K^a)$$

The first order conditions are

$$\mathcal{L}_L = p_1 + \lambda\,(1 - a)\, L^{-a} K^a = 0$$

$$\mathcal{L}_\lambda = y + L^{1-a} K^a = 0$$

From this we can obtain

$$1/\lambda = ((1\text{-}a)\,(K/L)^a / p_1$$

The second order conditions are

$$\mathcal{L}_{LL} = -\lambda\,(1 - a)\,(-a)\, L^{-a-1} K^a \qquad \mathcal{L}_{L\lambda} = -(1 - a)\, L^{-a} K^a$$

$$\mathcal{L}_{\lambda\lambda} = 0 \qquad \mathcal{L}_{\lambda L} = -(1 - a)\,(K/L)^a$$

or we have

$$\begin{bmatrix} -\lambda\,(1 - a)\,(-a)\, L^{-a-1} K^a & -(1 - a)\, L^{-a} K^a \\ -(1 - a)\,(K/L)^a & 0 \end{bmatrix}$$

Taking the determinant we have

$$D = -[-(1 - a)\, L^{-a} K^a]^2 < 0$$

Solving the first order conditions, we get a constant output factor input demand function for labor and the Lagrangian multiplier function:

$$L^e = [y\, K^{-a}]^{1/(1-a)}$$

$$\lambda^e = \frac{p_1}{1 - a}\,(y/K^a)^{a/(1-a)}$$

Turning to cost functions and costs, to derive the cost function L^e, is substituted for L in

$$TC = p_1 L + p_2 K^F$$

$$TC = p_1 L^e + p_2 K^F = p_1\,[y\, K^{-a}]^{1/(1-a)} + p_2 K^F$$

From the cost function, the following cost relationships can be derived

average fixed costs - $p_2 K^F / y$

average variable costs - $p_1[y\, K^{-a}]^{1/(1-a)} / y$

marginal costs - $\dfrac{\partial TC}{\partial y} = \dfrac{p_1 K^{-a/1-a} y^{a/1-a}}{1\text{-}a}$ which, it should be noted is equal to λ^e.

shape of the marginal cost curve $\frac{\partial^2 TC}{\partial y^2} = \frac{p_1 a}{(1 - a)^2} (y/K^a)^{a/(1 - a)} > 0$

Since the constant output factor input demand function for labor is $L^e = [y\ K^{-a}]^{1/(1 - a)}$, then $\partial L^e/\partial p_1 = 0$, which means that there is no substitution between labor and capital. Hence the demand for labor is not a function of its price. This particular result is restricted to the short period case in which there is only one variable factor input and one fixed factor input.

Homogeneous and Homothetic Production and Cost Functions[23]

A homothetic production function is a monotonically increasing transformation of a homogeneous production functions. Thus any homogeneous production function is homothetic (but homothetic functions are not necessarily homogeneous). Thus if y = f(**x**) is a homogeneous function, then y = F[f(**x**)] is a homothetic function if F' > 0. Moreover, it is easily shown that for a homogeneous production function

$$\frac{df_i(x_1, \ldots x_n)}{df_j(x_1, \ldots x_n)} = \frac{df_i(tx_1, \ldots tx_n)}{df_j(tx_1, \ldots tx_n)}$$

That is, the slopes of the level curves are the same along every point of a given ray out of the origin. This property also holds for a homothetic production function. Thus the isoquants of homothetic production functions are all radical expansions of one another, hence the expansion path is a straight line from the origin.

A homogeneous cost function derived from a homogeneous production function has the following form:

$$TC = y^{1/t} A(p_1, \ldots p_n)$$

in which costs, output, and input prices are related in a multiplicatively separable fashion. In addition ATC are always proportional to marginal costs, the factor of proportionality being the degree of homogeneity t, that is ATC = t MC. In the case of cost functions derived from homothetic production functions, they can be written as the product of two functions, that is as a function y and a function of input prices TC = J(y) A(p_1, . . . p_n). Hence, for average total costs we have ATC = [MC G(y)] / y = MC G(y) y^{-1}.

Duality of Cost and Production Functions

We started out by describing technology by means of production functions. Then we used the cost minimization approach to define the concept of the cost function, followed by an extensive examination of the behavior of the cost function, first regarding costs as a function of output with input prices held fixed. We have seen how restrictions on technology imposed certain restrictions on cost behavior: for example, fixed factors of production tend to produce short period U-shaped average total cost curves, and a constant-returns-to-scale technology produces long period constant average total cost curves. In addition, we have seen how the behavior of the cost function provides information about technology. For example, the derivative property of the cost function is that optimal combinations of inputs and outputs can be found by looking at the derivatives of the cost function with respect to input prices. In this section we take up this latter theme and ask to what extent we can recover information about a production function (technology) by investigating its cost function.

Suppose we are given a cost function and are asked to derive the production function that is consistent with it. That is, suppose one was given a cost function that satisfied the properties listed in footnote 17 on page 4, is it possible to identify with that cost function some unique production function that would generate that cost function? The answer in general is yes because there is a duality between production and cost functions. Consider the general case. Given a well behaved cost function C*(p_1, p_2, y°), Shephard's lemma can be used to derive the constant output factor input demand functions

---

[23] For further discussion, see Silberberg and Suen (2001). THIS SECTION HAS TO BE DEVELOPED

$x_1^e = \Psi_1(p_1, p_2, y°)$

$x_2^e = \Psi_2(p_1, p_2, y°)$

Since x_1^e and x_2^e are homogeneous of degree zero, they can be rewritten as

$x_1^e = x_1^*(1, p_2/p_1, y°) \equiv g_1(p^*, y°)$

$x_2^e = x_2^*(1, p_2/p_1, y°) \equiv g_2(p^*, y°)$

Now we have two equations and four variables x_1, x_2, p, and y°. Assuming that the Jacobian of these two equations is nonzero, that is

$$J = \begin{vmatrix} g_1 p & g_1 y° \\ g_2 p & g_2 y° \end{vmatrix} \neq 0$$

these equations can be used to eliminate the variable p. This will leave one equation in, x_1, x_2, and y°, say $g(x_1, x_2, y°) = 0$. Solving this equation for $y = f(x_1, x_2)$ yields the production function. For example, let us consider the following total cost function

$TC^*(p_1, p_2, y°) = y\, p_1^a\, p_2^{1-a}$

Using Shephard's lemma we get the following constant output factor input demand functions

$x_1^e = a\, y\, p_1^a\, p_2^{1-a} = a\, y\, (p_2/p_1)^{1-a}$

$x_2^e = (1 - a)\, y\, p_1^a\, p_2^{1-a} = (1 - a)\, y\, (p_2/p_1)^{-a}$

We want to eliminate p_2/p_1 from these two equations and get an equation y° for in terms of x_1 and x_2. Rearranging each equation gives

$p_2/p_1 = [x_1^e / (a\, y°)]^{1/(1-a)}$

$p_2/p_1 = [x_2^e / ((1 - a)\, y°)]^{-1/a}$

Setting these equal to each other and raising both sides to the -a (1 – a) power we get

$$\frac{(x_1^e)^{-a}}{a^{-a}\,(y°)^{-a}} = \frac{(x_2^e)^{1-a}}{(1-a)^{1-a}\,(y°)^{1-a}} \qquad \text{or}$$

MORE—AND AN EXAMPLE IS NEEDED.

$$y = \frac{1}{a^{-a}(1-a)^{1-a}} (x_1^e)^{-a} (x_2^e)^{1-a} = A (x_1^e)^{-a} (x_2^e)^{1-a}$$

which is a Cobb-Douglas production function.[24]

We can conclude that as long as the production function and cost function are well behaved, one implies the other. That is, if the cost function is well behaved (that is has the properties listed in footnote 17, page 44), it will summarize all of the economically relative aspects of the production function and will generate a production function which has the properties listed on page 31. Thus duality of cost and production functions is important for reasons other than mathematical elegance. Neoclassical economists use two different approaches to estimate constant output factor input demand functions and cost functions. One is to estimate the underlying production functions for some activity and to then calculate the constant output factor input demand functions. This, however, is a very arduous procedure since production functions are largely unobservable.[25] The data points used for the estimate represent a sampling of input and output levels that have taken place at different times and with different input and output prices. And of what use is knowledge of the production function itself? Largely, it is to derive implications regarding factor usage and cost considerations when various parameters change. It would seem to make more sense to start with estimating the cost functions or the constant output factor input demand functions directly. However, this procedure is always subject to the criticism that the estimated cost or demand functions are beasts without parents. However the duality results assure that if the derived cost function is well behaved (that is, is linear homogeneous and concave in prices), then there in fact is some real, unique underlying production function. Since the cost function is a function of output and prices that are potentially observable, it can be grounded in the real

[24] It should also be noted that a production function could be derived directly from the constant output factor input demand functions, assuming, of course that they are well behaved. THIS SECTION NEEDS TO BE DEVELOPED MORE.

world. Thus duality connects the real world with theory, with the unobservable production function. But this is open to real questions as to whether this is really knowledge in the first place. NEED TO WORK ON.

Fixed Proportions Production and Cost Theory

[25] This raises the issue of whether they actually exist in the first place. MORE ON THIS POINT.

CHAPTER 14

CRITICISMS

Relative to demand, the supply side of neoclassical microeconomic theory is more complex because of the pre-conditions that need to be specified before any analysis of production and costs take place; and the latter has to occur before discussion about supply curves take place, and that discussion requires the introduction of an additional set of assumptions. Therefore, in this chapter attention will be focus on production and costs.

Technology and the Production Function

As in the textbooks, we start with a firm production function of the general form:

(1) $y = f(\mathbf{x})$

where y is the output and the vector of factor inputs $\mathbf{x} = (x_1, \ldots, x_n) > 0$ and divisible. The production function also has three additional definitional properties: it consists only of technology that ensures for any technique of production represented by the factor input combination $\mathbf{x^i}$, y is maximized; it and its technology is considered exogenous datum and fixed; and the factor inputs, x_i, are scarce factor inputs. However, these definitional properties generate three problems. The first concerns the technology itself in that the technology creators draw upon technological, economic, and social influences (all of which are external to the production function and hence cannot be restricted) to create technology for a specific valued end which the influences also define. Consequently, the range (which may be great or small) of technology that the firm can choose to include in its production function can have fixed production coefficients where the increase in a single input is necessary but not sufficient for an increase in output, variable production coefficients where the increase in a single input is both necessary and sufficient for an increase in output, or a combination of both. And since the valued end can be the maximization of output given inputs or something else, there can be only one or quite many $\mathbf{x^i}$ that produce the same y.

In particular, for neoclassical economists, the objectives of the technology creators at this level of analysis are outside of consideration and investigation. Hence it is possible and even plausible that the technology available to the firm is consciously engineered to not maximize output from given inputs. It is also possible that the technologists do not separate technology from valued end output objective; thus $\mathbf{x}^i \rightarrow y$ can be based on objectives completely outside $\mathbf{x}^i \rightarrow$ maximum y. Finally, there is no reason not to suppose that technological, economic, and social influences on the technology creators are constructed and altered through the use of the technology to produce goods and services—that is, technology can change through usage, hence making it endogenous. These possibilities render incoherent the assumption that technology is fundamental datum and the definition of neoclassical economics that requires the technology to be separate from the ends.

Given the array of technologies and corresponding techniques of production available to the firm, the second problem arises over the choice of technology and techniques to be included in its production function. Assuming that the firm prefers technology that maximizes output given inputs, the firm's choice algorithm, as in consumer choice theory, can include many influences concerning the nature and usage of the factor inputs relative to what is meant by maximizing output. Consequently, choices of $\mathbf{x}^i$ (and its technology) can be cyclical and hence cannot arrive at a single $\mathbf{x}^i$ that maximizes y; the firm can have a cyclical interpretation of maximum y relative to $\mathbf{x}^i$ hence also making its choice of $\mathbf{x}^i$ indeterminate; or the factor inputs in different $\mathbf{x}^i$ are different thus making it impossible for the firm to compare and chose between the different technologies relative to a given y. Moreover, the choice of technology combined with the "curse of dimensionality" implies that the firm may not be able to choose a range of technology for its production function that is complete, singled valued in that for any $\mathbf{x}^i$ there is a single y, and for any given $\mathbf{x}^i$ the resulting y is maximized; and these specific shortcomings render the conception of a production function incoherent as well as preventing it from being a useful guide and tool for cost minimization.

However, assuming that the firm does choose technology for its production function and given its choice algorithm, the resulting production function, in conjunction with the issues raised in the first problem, could plausibly have the following properties:

(1) each of the techniques of production has fixed production coefficients which implies that y is not a monotonic in x_i, the marginal product of x_i and the *marginal* rate of technical substitution do not exist, and there is no distinction between fixed and variable inputs;

(2) there is a single technique of production with fixed production coefficients which has all the implications of (1) above as well as no technical substitution at all;

(3) scale dependent inputs linked with output such that for any y there is a single $\mathbf{x^i}$ (with fixed production coefficients) and for y + 1 there is a single $\mathbf{x^j}$ (with fixed production coefficients) where $\mathbf{x^i} \neq \mathbf{x^j}$ in that there is at least one input in $\mathbf{x^j}$ that is not in $\mathbf{x^i}$; such a production schema violates continuity and convexity and eliminates proportional changes in inputs and outputs, which means there are no isoquants especially convex isoquants, marginal rate of technical substitution, and laws of returns to scale; and

(4) variable production coefficients that are constant or decline until the fixed factor input is fully utilized and ceases to take on any more of the variable inputs, which means that marginal products do not decline.

Since the influences on the creation and choice of technology is unrestricted, the resulting "production function" created by the firm may have none of the usual properties and characteristics associated with strictly quasi-concave production function (differentiable or not, homogeneous or homothetic) with strictly convex technology.[26] In short, incoherent, useless as a guide and tool for cost minimization, and lacking traditional production properties, the neoclassical production function is neither a sensible or sustainable delineation of production.

The third problem concerns scarcity as a definitional property of the production function. Given the lack of restrictions on the technology available to the firm and the firm's choice of technology, it is possible that its production function contains inputs that are produced by other firms and does not include constraints on production such as declining marginal products or decreasing returns to scale. In addition, the produced input connection between firms, when taken across all firms, could generate a system of production where all firms use produced and non-produced inputs in production. Thus, the production of produced inputs can be represented, as is overwhelming empirically the case, in terms of an input-output model with circular production and one or more non-produced inputs. With the lack of production constraints combined with producibility, reproducibility, and circular production, the produced inputs in the production function cease to have the properties of a scarce factor input;[27] and more significantly, so do the non-produced inputs, as will be elaborated on below.[28] With perhaps none of the inputs in the production function scarce, although with production still taking place, the production function is not only an incoherent concept, it also does not exist. So just because production

---

[26] The empirical evidence does support this possibility—see Lee (1986).

[27] Neoclassical economists have tried to circumvent this problem by defining goods according to time periods. Thus, because they represent different time periods, an input is conceptually different from an output even when they have the same technical characteristics. This converts all produced inputs into relatively scarce factor inputs (assuming demand for their usage is sufficient). But this intertemporal equilibrium approach to production discards long run methodology utilized in virtually all neoclassical textbooks. More significantly, this definitional-based distinction has no sense in that no substantial reason is given for why time will make technically identical goods different; and in the real world example of wheat being an input into its own production, this approach is simply nonsense.

[28] Since produced inputs and circular production presuppose the prior existence of social activities engaged in production, production is also fundamentally a social process where output is a result of common, complementary, and coordinated effort; and social production is incompatible with the notion of scarcity.

takes place and output is related to inputs, this does not allow one to conclude that production functions exist or to insert faith in place of scientific inquiry. Yet, for the neoclassical faithful, it can be said that those who believe in production functions with all their will are Fergusonians still.[29] [Ferguson 1979; Varian 1992; Mas-Colell, Whinston, and Green 1995; Lee 1998; Bortis 1997]

Cost Curves, Demand for Factor Inputs, and Partial Equilibrium

Without a production function or even a production function with marginal products, proportional input variation, and convex technology, it is not possible to derive cost minimizing constant output factor input demand function and its derivative properties or total cost function and its derivative properties of short and long period cost curves, marginal cost curve and their shape. That is, if, as is quite possible, the production function does not exist or exists but with properties noted above in chapters one and two, then there would be no basis for cost minimization, isoquants, marginal rate of technical substitution, and total cost functions and the standard cost curves. Moreover, since both the short and long period marginal cost curves are based on marginal products and proportional input variation, they would not exist (irrespective of their shape). Finally, without technology restricted to generating at some point declining marginal products and decreasing returns to scale, there would be no necessary reason for increasing short and long run marginal cost curves to exist at all.[30] In short, without a production function and its traditional properties, it is not possible to establish in neoclassical economics a functional relationship between output and costs, which implies that *cost elasticity*, *duality* between the production function and the total cost function, and the long run average total cost curve being an envelope of short run average total cost curves are meaningless concepts.

---

[29] There are those who still believe in aggregate production functions in spite of well-known aggregation problems because they “work in practice”. But this has been shown not to be the case at all—see McCombie (1998, 2000-2001, and 2001) and Felipe and McCombie (2001).

Turning to the constant output factor input demand function and working in the long period where traditionally the substitution of factor inputs is permitted, when differentiating $x_i^e = \psi_i(\mathbf{p}, y^o)$ with respect to p_i, we find that the demand for factor input x_i^e is inversely related to its own price, That is the demand curve for a factor input slopes downward because changes in quantity demanded of the factor input is restricted to the original isoquant since output is constant. But, without a spectrum of techniques, marginal products, isoquants, and bordered Hessian matrix, cost minimizing constant output factor input demand functions do not exist and there is no functional relationship between x_i^e and its own price—thus no law of demand for factor inputs.[31] However, the shape of the constant output factor input demand function poses even more significant issues once produced inputs and circular and complementary production are considered. It was established, in the context of the capital controversies, that for a system of production in which circular production takes place and all inputs are reproducible except labor, a reduction in a factor's input price would not necessarily increase its demand nor result in its substitution for the relatively higher-price factor input. More detailed research has reinforced these results as well as extending them to include more than one non-produced factor input, which means that non-produced inputs are "acting" liked produced inputs.[32] The research also shows that the results emerge because an arbitrary change in an input price, p_i, in a system of produced inputs and circular production has collateral effects that are non-negligible, such as affecting other input prices that are presumed to be constant and putting other firms out of equilibrium hence requiring them to make adjustments (that also have collateral effects) to get back to equilibrium. The existence of collateral

[30] Empirical evidence on short run "marginal cost curves" suggests that they are in general not upward sloping; and if they are upward sloping the explanation is not based on marginal products—see Lee (1986) and Blinder, et. al., (1998)

[31] For methodological, theoretical, and empirical evidence supporting the absence of the law of demand for factor inputs, see Fleetwood (2002), Michl (1987), and Bewley (1999).

effects invalidates the *ceteris paribus*, partial equilibrium methodology underpinning the derivation of the slope of the constant output factor input demand function, hence making it meaningless. Thus it calls into question **any** partial equilibrium analysis (short period or long period) that allows for some price and quantities changes and input substitutions and yet does not take into account their possible disequilibrium impact on other firms and their actions to regain equilibrium. Without partial equilibrium methodology, the traditional market analysis articulated in neoclassical textbooks is rendered incoherent. [Ferguson, 1972; Pasinetti, 1977; and Steedman, 1985, 1988, and 2002]

[32] The research is carried out at the industry level, but it is equally applicable to the individual firm whose production function contains multiple technologies since the change in input price is arbitrary and in principle extends to all firms in the economy if the law of one price is to prevail.