4.1 Introduction

Throughout this chapter and the next the production function is assumed to be single valued, everywhere continuous, and well defined over the range of inputs yielding nonnegative outputs. A second assumption, which increases generality without sacrificing content, is that the production function belongs to a class of functions with continuous first and second partial derivations. Third, it is assumed that the inputs are continuously variable, i.e. the inputs are real variables defined over the nonnegative domain of real numbers.

A fourth assumption, which is sometimes not invoked, concerns the unchanging character of the production function and of the inputs. Specifically, unless otherwise stated, the production function is given and fixed; there is neither technological progress nor retrogression. Similarly, the inputs remain unchanged in character and are assumed to be homogeneous within themselves. Finally, our attention is chiefly focused upon the production of a single commodity and the behavior of a single-product firm.

As mentioned in chapter 1, it is both convenient and instructive to examine the behavior of output when only one input is varied, all other inputs being held constant. In the conventional terminology of economics such investigations are called short-run analyses; the inputs held constant are called ‘fixed inputs’, the one allowed to vary is called a ‘variable input’. Otherwise, one may simply regard such investigations as the study of cross-sections of the production surface or of the behavior of a single partial derivative.

Finally, it might be well to note that inputs are sometimes classified as ‘general’ and ‘specific’. General inputs are those used in producing a variety of different outputs; specific inputs are those used to produce the...

\[\text{For a more thorough discussion of the concept of a production function, see Smithies [1935], Smithies [1936], and Carlson [1939, pp. 14-16].}\]
commodity under consideration but not used in the production of any other good. For example, 'unskilled labor' is a general input used in some phase of production of most goods, while 'neurosurgeon' is a specific input efficiently utilized only in cerebral operations. Often the category in which an input is placed depends upon the stage of processing it has reached. Thus the input 'cotton' is widely utilized, whereas 'percale' is virtually specific to the fabrication of bed linens.

This dichotomy of inputs is not important so far as the technical theory of production is concerned. It is relevant only when economic considerations are introduced, especially in determining input demand functions. For the most part it is assumed that all inputs are 'general'. Thus their prices may be regarded as parametric data to the firm. In chapter 8, however, the economic theory of production with variable input prices is discussed.

4.2 The production function

Production functions must generally be given an algebraic representation; however, if the number of inputs is sufficiently small, graphical and tabular representations are possible. In this section primary emphasis is placed upon algebraic formulations; graphical and tabular forms are sometimes shown for purposes of illustration.

4.2.1 Algebraic and tabular representations

In principle the variety of equations that may validly represent a production function is virtually limitless. Three special features, however, deserve notice. First, the production function might have an additive constant term; however, this would imply that a positive level of output could be obtained without the use of inputs. This is a 'something for nothing' condition that violates one's established notions of the real world, not to mention the real economic world. Hence we assume that the production function does not contain an additive constant.

Second, the variables could enter singly, in pairs, etc., or they may enter only in universally multiplicative groups. The more reasonable assumption seems to be that a positive usage of all inputs simultaneously is required to produce a positive output. Let the production function be written

\[ q = f(x_1, x_2, ..., x_n), \]  \hspace{1cm} (4.2.1)

where \( q \) is output and \( x_i \) is the usage of input \( X_i \). Thus, we assume that

\[
\begin{align*}
  f(0, 0, ..., 0) &= f(0, x_2, ..., x_n) = f(x_1, 0, x_3, ..., x_n) = ... \\
  &= f(x_1, x_2, ..., x_{n-1}, 0) = 0. \hspace{1cm} (4.2.2)
\end{align*}
\]
CONTINUOUS PRODUCTION FUNCTIONS

Third, the production function may be homogeneous or inhomogeneous. Chapter 5 is devoted to a special analysis of homogeneous functions. In this chapter our attention is directed chiefly to the more general class of inhomogeneous functions.

Specific examples of production functions are easy to come by. Perhaps the simplest form is

\[ q = \sum_{i=1}^{n} a_i x_i \quad (0 < a_i). \]  

However, this function does not satisfy our general requirement concerning second partial derivatives. Perhaps the two best known production functions are the Cobb–Douglas function and the Arrow–Chenery–Minhas–Solow function shown, respectively, by:

\[ q = A x^\alpha y^\beta \quad (0 < \alpha, \beta < 1), \]

and

\[ q = \gamma \left( \delta x^{-p} + (1 - \delta) y^{-q} \right)^{-1/p} \quad (0 < \gamma, \delta). \]

Other examples of homogeneous functions are listed below:

\[ q = 2\alpha x y - bx^2 - cy^2 \quad (0 < a, b, c \text{ and } a^2 > bc), \]

\[ q = 2\alpha x y - bx^2 - cy^2 \frac{dx}{dx + cy} \quad (0 < a, b, c, d, e \text{ and } a^2 > bc), \]

\[ q = a_1 x^\alpha y^\beta + a_2 x^\gamma y^\delta \quad (0 < a_1, a_2, \beta_1, \beta_2, a_3, a_4), \]

\[ q = a \left[ \frac{h x^2 y^2 + x^2 y^2}{x^4 + y^4} \right] \quad (0 < a, b). \]

An example of an inhomogeneous function, which is used illustratively in this chapter, is:

\[ q = a \left[ x^{\gamma} + x y^{\alpha} - \frac{1}{2} x^{\delta} - \frac{1}{2} y^{\delta} \right] \quad (0 < a). \]

The function above satisfies all the criteria introduced above, and it is uniquely defined over the range from (0, 0) to (1.5, 1.5). A tabular representation of this production function (for \( a = 10,000 \)) is presented in table 3. At present two features of this function merit notice. First, for small values of either input, the usage of the other input can become so great that output is zero. Second, the production surface rises to a unique maximum at \( x = y = 1.125 \) and declines thereafter. In point of fact, output is zero when \( x = y = 1.50 \).

1 Cobb and Douglas [1928] and Arrow, Chenery, Minhas and Solow [1961]. For another derivation of the CES function, see Kendrick and Sato [1963].

2 I am indebted to Mr Earl Sasser for the programming and calculations upon which the numerical examples are based.
4.2 THE PRODUCTION FUNCTION

TABLE 3 Tabular representation of equation (4.2.10)
with \( a = 10,000 \)

<table>
<thead>
<tr>
<th>Y</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1.0</th>
<th>1.1</th>
<th>1.125</th>
<th>1.13</th>
<th>1.14</th>
<th>1.15</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>1657</td>
<td>2019</td>
<td>2183</td>
<td>2053</td>
<td>1509</td>
<td>417</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.6</td>
<td>2183</td>
<td>2995</td>
<td>3125</td>
<td>2862</td>
<td>2069</td>
<td>695</td>
<td>404</td>
<td>101</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.7</td>
<td>2053</td>
<td>3125</td>
<td>4069</td>
<td>4102</td>
<td>3633</td>
<td>2499</td>
<td>2343</td>
<td>2092</td>
<td>205</td>
<td>1823</td>
<td>1632</td>
</tr>
<tr>
<td>0.8</td>
<td>2053</td>
<td>3125</td>
<td>4069</td>
<td>4102</td>
<td>3633</td>
<td>2499</td>
<td>2343</td>
<td>2092</td>
<td>205</td>
<td>1823</td>
<td>1632</td>
</tr>
<tr>
<td>0.9</td>
<td>1509</td>
<td>2862</td>
<td>4105</td>
<td>5135</td>
<td>5812</td>
<td>6050</td>
<td>5605</td>
<td>5355</td>
<td>5450</td>
<td>5401</td>
<td>5297</td>
</tr>
<tr>
<td>1.0</td>
<td>417</td>
<td>2069</td>
<td>3033</td>
<td>5003</td>
<td>6595</td>
<td>6667</td>
<td>6673</td>
<td>6534</td>
<td>6534</td>
<td>6470</td>
<td>6398</td>
</tr>
<tr>
<td>1.1</td>
<td>0</td>
<td>595</td>
<td>2499</td>
<td>4229</td>
<td>6665</td>
<td>6673</td>
<td>7009</td>
<td>7103</td>
<td>7095</td>
<td>7088</td>
<td>7069</td>
</tr>
<tr>
<td>1.2</td>
<td>0</td>
<td>404</td>
<td>2343</td>
<td>4110</td>
<td>5585</td>
<td>6634</td>
<td>7103</td>
<td>7112</td>
<td>7110</td>
<td>7106</td>
<td>7091</td>
</tr>
<tr>
<td>1.3</td>
<td>0</td>
<td>101</td>
<td>2092</td>
<td>3961</td>
<td>5450</td>
<td>6561</td>
<td>7095</td>
<td>7110</td>
<td>7119</td>
<td>7118</td>
<td>7113</td>
</tr>
<tr>
<td>1.4</td>
<td>0</td>
<td>0</td>
<td>2005</td>
<td>3847</td>
<td>5401</td>
<td>6533</td>
<td>7088</td>
<td>7106</td>
<td>7118</td>
<td>7118</td>
<td>7113</td>
</tr>
<tr>
<td>1.5</td>
<td>0</td>
<td>0</td>
<td>1823</td>
<td>3702</td>
<td>5297</td>
<td>6470</td>
<td>7069</td>
<td>7091</td>
<td>7110</td>
<td>7113</td>
<td>7112</td>
</tr>
</tbody>
</table>

4.2.2 Graphical representation

To represent a production function graphically, the number of variables must be restricted so that dimensionality is not a problem. To that end, let the production function be

\[ q = f(x, y), \]  

a specific form of which is given by equation (4.2.10).

A portion of a typical production surface corresponding to equation (4.2.11) is shown in figure 17. The production surface is \( OYQX \) (note that only the rising portion of the surface is shown in the figure). If the surface is intersected at \( Y = Y_1 \), the curve \( Y_1PF \) shows a cross-section of the surface in the \( Q-X \) plane. Similarly, holding \( X \) constant at \( X_1 \), \( X_1PD \) traces the outputs obtainable for a given input of \( X \) and variable usage of \( Y \).

4.2.3 Isoquants

Cross-sections of the production surface in the \( Q-X \) and \( Q-Y \) planes are useful because they show how total product varies with the variable usage of one input, the other input held constant. Thus these cross-sections illustrate one fundamental property of continuously differentiable production functions: output varies in response to a change in any one input, all other inputs held constant. The change in input usage may be in either direction; output always responds unless the point of absolute saturation has been attained (e.g. at the input point \((1.5, 1.5)\) in equation (4.2.10)).
CONTINUOUS PRODUCTION FUNCTIONS

This contrasts sharply with the case of fixed-proportion production functions, in which output responds only to a negative variation in a single limitational input.

The single-input variation property is very important; yet there is another property of equal importance: various combinations of inputs may be used to produce the same quantity of output. That is, one input may be substituted for another while maintaining a constant level of output. Another cross-section, one in the $X$-$Y$ plane, serves to illustrate this property.

Again consider figure 17. Intersect the production surface $OYQX$ by a plane $RSTV$, representing the level of output $RR' = PP' = BB' = YS$. The resulting cross-section is the curve $APB$. Projecting this cross-section onto the $X$-$Y$ plane yields the curve $A'P'B'$, which shows all combinations of $X$ and $Y$ that may be used to produce $RR'$ units of output. This projection of the cross-section onto the $X$-$Y$ plane is called an isoquant.

**Definition:** An isoquant is a locus of input combinations each of which is capable of producing the same level of output.
4.2 THE PRODUCTION FUNCTION

Every possible level of output may be represented by an isoquant. Since output is a continuous variable, there are infinitely many isoquants in any isoquant map, i.e. the isoquant map is everywhere dense. Furthermore, the isoquants do not intersect; this property follows immediately from the assumption that the production function is single-valued.

Various isoquant maps are illustrated in figures 18–20. First, consider figure 18, which depicts typical isoquants from production functions such as those in equations (4.2.4) and (4.2.5). \( Q_o \) and \( Q_1 \) are isoquants; since \( Q_1 \)

![Diagram](image)

is located to the northeast of \( Q_o \), the output level associated with \( Q_1 \) exceeds the output associated with \( Q_o \).

On an isoquant map any ray from the origin, such as \( OR \) in figure 18, represents a constant input ratio, i.e. \( (y_i/x_i) = (y_2/x_2) \). Thus an isoquant map enables one simultaneously to observe the way in which the input ratio changes for a given level of output and the way in which output changes for a given input ratio. To reemphasize, output is constant and the input ratio changes for movements along an isoquant, whereas the input ratio is constant and output changes for movements along a ray from the origin.

It is only in special cases that isoquant maps resemble the map in figure 18. Generally, the \textit{full} isoquant map consists of a set of concentric

65
CONTINUOUS PRODUCTION FUNCTIONS

ellipses, as shown in figure 19. The more interior the ellipse, the greater the associated output. Isoquant maps have this elliptical shape when there is a unique maximum output that is physically attainable. The maximum is simply a dot \((M)\) on the map at the point where the ellipses degenerate. For the production function in equation (4.2.10), the maximum is attained at \(x = 1.125, y = 1.125\). A fair representation of an elliptical isoquant may be obtained by plotting the four input combinations yielding 7,110 units of output according to equation (4.2.10), see table 3.

![Diagram of ellipses]

If the production function does not have a unique maximum, the isoquants are not ellipses. In this case they are hyperbolas or parabolas similar to those shown in figure 20. As we shall later see, when the isoquant map resembles the one in figure 20, it is of great interest to determine the locus of points \((OS)\) along which the slopes of the isoquants are infinite and the locus \((OS')\) along which the slopes are zero.

4.3 Single input variations

In section 4.2 two different types of cross-sections of the production surface were examined: (a) the type showing the variation in output associated

\(^1\) Except when the production function is homogeneous. For a numerical illustration, see Cassels [1936].

66
4.3 SINGLE INPUT VARIATIONS

with a variation of one input (and hence the input ratio), and (b) the type showing the variations in the input combination associated with a constant level of output. In this section our attention is directed to the results of single-input variation, while the following section is concerned with input substitution when output is held constant.

Another view of the cross-section for a single-input variation is shown in figure 21. This graph depicts equation (4.2.11), and the section $X_1PS$ shows the particular form $q = f(y | x = x_1)$. (4.3.1)

This cross-section illustrates the behavior of total output over the entire range from $f(0 | x = x_1) = 0$ to the saturation point $f(y_1 | x = x_1) = 0$.

4.3.1 Marginal product

The concept of marginal product is introduced by the following

Definition: Consider a production function of the form

$$q = f(x_1, x_2, ..., x_n)$$ (4.3.2)

that satisfies the conditions stated in section 4.1. The marginal product of the $ith$ input is the partial derivative of the production function with
Continuous Production Functions

respect to the input under consideration:

\[ MP_i = \frac{\partial q}{\partial x_i} = f_i(x_1, x_2, \ldots, x_n). \]

In general, the marginal product of an input is a function of all inputs. This is emphasized in equation (4.3.3) by writing \( f_i(x_1, x_2, \ldots, x_n) \). If \( f_i \) is a continuous function of all the inputs at a specified input combination, \( X_i \) is said to be a continuity factor at that point.\(^1\)

![Diagram](image)

Fig. 21

Now consider the exact increment of equation (4.3.2):

\[ \Delta q = df + \sum_{i=1}^{n} \eta_i dx_i, \]

where \( \eta_i \to 0 \) as \( dx_i \to 0 \) if all inputs are continuity factors at an input point. In this case the increments \( dx_i \) may be chosen so small that \( \Delta q \to df \). Hence

\[ df = \sum f_i dx_i \]

is a suitable approximation of the actual increment \( \Delta q \) when all inputs are continuity factors. In other words, the total increment of output is equal to the sum of the input increments each multiplied by its marginal product.

Specific examples of marginal product functions may be obtained from the production functions in equations (4.2.4)–(4.2.10). For example, the

\(^1\) For a detailed discussion of several alternative meanings of ‘marginal product’, see Machlup [1936].

\(^a\) Frisch [1965, p. 54] presumably introduced this definition.
4.3 SINGLE INPUT VARIATIONS

Marginal products for equations (4.2.4) and (4.2.5) are, respectively,

\[
\frac{\partial q}{\partial x} = \frac{a q}{x} \quad \text{and} \quad \frac{\partial q}{\partial y} = \frac{b q}{y},
\]

(4.3.6)

and

\[
\frac{\partial q}{\partial x} = \gamma^p \delta \left(\frac{q}{x}\right)^{1+p} \quad \text{and} \quad \frac{\partial q}{\partial y} = \gamma^{-p}(1-\delta) \left(\frac{q}{y}\right)^{1-p}.
\]

(4.3.7)

Similarly, the marginal products from equation (4.2.10) are

\[
\frac{\partial q}{\partial x} = a(2xy+y^2-x^2) \quad \text{and} \quad \frac{\partial q}{\partial y} = a(x^2+2xy-y^3).
\]

(4.3.8)

Numerical values for the marginal products in equations (4.3.8) are shown in table 4.

4.3.2 Marginal returns and input relations

An examination of the second partial derivatives of equation (4.3.2) enables one to determine a good bit about the behavior of the marginal product function. Denote the second partial derivatives by

\[
\frac{\partial^2 q}{\partial x_i \partial x_i} = f_{ii} \quad (i, j = 1, 2, \ldots, n).
\]

(4.3.9)

The \( f_{ii} \) terms are sometimes called the direct acceleration coefficients and the \( f_{ij} \) terms cross-acceleration coefficients.\(^1\)

The direct acceleration coefficient shows whether the marginal product function increases or decreases at any point. That is, if \( f_{ii} > 0 \) at a point, \( f_i = MP_i \) is an increasing function at that point. Similarly, if \( f_{ii} < 0 \) at a point, \( f_i = MP_i \) is a decreasing function at that point. Consequently, at a given input point there are increasing or diminishing marginal returns from the \( i \)th input according as \( f_{ii} \geq 0 \).

In certain production functions, such as equations (4.2.4) and (4.2.5), the direct acceleration coefficient is always negative. If this is so, the marginal product function itself must always be a nonnegative function. In other types of production functions, such as equations (4.2.9) and (4.2.10), the direct acceleration coefficient is sometimes positive and sometimes negative. In this case marginal returns increase over certain ranges of inputs and diminish over others. The fact that every acceptable production function is characterized by a range of diminishing marginal returns is usually called the Law of variable proportions. With a given state of technology, if the quantity of one productive service is increased by equal increments, the quantities of the other productive services remaining fixed, the resulting

\(^1\) Frisch [1965, pp. 58-61].
<table>
<thead>
<tr>
<th>Inputs</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1.00</th>
<th>1.100</th>
<th>1.125</th>
<th>1.130</th>
<th>1.140</th>
<th>1.150</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>4167</td>
<td>6267</td>
<td>8667</td>
<td>11067</td>
<td>13767</td>
<td>16667</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>2740</td>
<td>5040</td>
<td>7540</td>
<td>10340</td>
<td>13140</td>
<td>16240</td>
<td>19540</td>
<td>19881</td>
<td>20396</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.7</td>
<td>353</td>
<td>2853</td>
<td>5553</td>
<td>8453</td>
<td>11553</td>
<td>14853</td>
<td>18353</td>
<td>18714</td>
<td>19260</td>
<td>19442</td>
<td>19809</td>
</tr>
<tr>
<td>0.8</td>
<td>-3153</td>
<td>-453</td>
<td>2447</td>
<td>5547</td>
<td>8847</td>
<td>12347</td>
<td>16047</td>
<td>16428</td>
<td>17003</td>
<td>17196</td>
<td>17583</td>
</tr>
<tr>
<td>0.9</td>
<td>-7940</td>
<td>-5040</td>
<td>-1940</td>
<td>4860</td>
<td>8660</td>
<td>12460</td>
<td>12861</td>
<td>13466</td>
<td>13669</td>
<td>14076</td>
<td>14485</td>
</tr>
<tr>
<td>1.00</td>
<td>-14167</td>
<td>-11067</td>
<td>-7767</td>
<td>-4267</td>
<td>-567</td>
<td>3333</td>
<td>7433</td>
<td>7854</td>
<td>8490</td>
<td>8702</td>
<td>9129</td>
</tr>
<tr>
<td>1.10</td>
<td>-18693</td>
<td>-15193</td>
<td>-11493</td>
<td>-7593</td>
<td>-3493</td>
<td>807</td>
<td>1248</td>
<td>1913</td>
<td>2136</td>
<td>2583</td>
<td>3032</td>
</tr>
<tr>
<td>1.110</td>
<td>-19550</td>
<td>-16030</td>
<td>-12310</td>
<td>-8390</td>
<td>-4270</td>
<td>50</td>
<td>493</td>
<td>1161</td>
<td>1385</td>
<td>1834</td>
<td>2285</td>
</tr>
<tr>
<td>1.125</td>
<td>-20869</td>
<td>-17319</td>
<td>-13569</td>
<td>-9619</td>
<td>-5469</td>
<td>-1119</td>
<td>-673</td>
<td>0</td>
<td>225</td>
<td>677</td>
<td>1131</td>
</tr>
<tr>
<td>1.130</td>
<td>-17757</td>
<td>-13097</td>
<td>-10037</td>
<td>-5877</td>
<td>-1517</td>
<td>-1070</td>
<td>-396</td>
<td>-170</td>
<td>283</td>
<td>738</td>
<td></td>
</tr>
<tr>
<td>1.140</td>
<td></td>
<td>-18648</td>
<td>-14868</td>
<td>-10888</td>
<td>-6708</td>
<td>-2328</td>
<td>-1879</td>
<td>-1202</td>
<td>-975</td>
<td>-520</td>
<td>-63</td>
</tr>
<tr>
<td>1.150</td>
<td></td>
<td></td>
<td>-19557</td>
<td>-14757</td>
<td>-7557</td>
<td>-3175</td>
<td>-2706</td>
<td>-2025</td>
<td>-1798</td>
<td>-1341</td>
<td>-882</td>
</tr>
</tbody>
</table>
4.3 SINGLE INPUT VARIATIONS

increment of product will decrease after a certain point. That is, \( f_{ii} < 0 \) \((i=1, 2, \ldots, n)\) over some range of inputs.¹

The direct acceleration coefficient shows how the marginal product of input \( X_i \) varies when the usage of \( X_i \) varies. The cross-acceleration coefficient \( f_{ij} \) shows how the marginal product of \( X_i \) varies when the usage of \( X_j \) changes. If \( f_{ij} > 0 \), the marginal product of \( X_i \) increases when the input of \( X_j \) increases. The inputs are accordingly said to be complementary at the input point under consideration. On the other hand, if \( f_{ij} < 0 \), the marginal product of \( X_i \) declines when the usage of \( X_j \) increases. In this situation the inputs are said to be competitive or alternative.² The complementarity relation generally prevails. Furthermore, if there is a range of competitive-ness (i.e. input points for which \( f_{ij} < 0 \)), there must also exist a range of input complementarity.

4.3.3 Average product

The concept of average product is introduced by the following

Definition: The average product of an input at any factor point is the quantity of output per unit of the input used, i.e. the output-input ratio. Given the production function in equation (4.3.2), the average product of the \( i \)-th input is

\[
AP_i = \frac{q}{x_i} = \frac{f(x_1, x_2, \ldots, x_n)}{x_i},
\]  

(4.3.10)

where the right-most expression emphasizes that the average product of one input is typically a function of all inputs.

A measure closely related to average product is sometimes called the fabrication coefficient of the \( i \)-th input.³ The fabrication coefficient shows, on average, the number of units of input \( X_i \) used per unit of output. Thus it is the input-output ratio, the inverse of average product. Denoting the fabrication coefficient of the \( i \)-th input by \( \theta_i \), we have

\[
\theta_i = \frac{x_i}{q} = \frac{1}{AP_i}.
\]  

(4.3.11)

4.3.4 Average returns

Changes in the level of output resulting from input variations are often called ‘returns to something’. When all inputs are varied in the same proportion, one speaks of ‘returns to scale’. When a single input is varied, the

¹ For a historical account of the development of the law of variable proportions, see Stigler [1946]. For an interesting recent contribution, see Tangri [1966].
² The complementarity-competitive relation should not be confused with the more important substitution relation. The latter concerns the size of the marginal products at a point, the former concerns the change in the marginal products in a neighborhood of the point.
³ This is the terminology of Frisch [1965, p. 62].
CONTINUOUS PRODUCTION FUNCTIONS

appropriate concept is 'marginal' or 'average' returns. In subsection 4.3.2 it was shown that there are increasing or diminishing marginal returns according as $f_{ii} > 0$. The reason for this classification, of course, lies in the fact that $f_{ii}$ is the first derivative of the marginal product function $f_i$.

In the same way, for a single input variation there exist increasing or diminishing average returns according as

$$\frac{\partial (q/x_i)}{\partial x_i} \geq 0. \quad (4.3.12)$$

Performing the differentiation, one obtains

$$\frac{\partial (q/x_i)}{\partial x_i} = \frac{1}{x_i} \left( f_i - \frac{q}{x_i} \right) = \frac{1}{x_i} (MP_i - AP_i). \quad (4.3.13)$$

From the last expression in equation (4.3.13) it follows immediately that there are increasing or diminishing average returns to the $i$th input at a given input point according as the marginal product is greater than or less than the average product.

Another familiar relation is obtainable from equation (4.3.13). The average product is a maximum when

$$\frac{\partial (q/x_i)}{\partial x_i} = 0 \quad \text{and} \quad \frac{\partial^2 (q/x_i)}{\partial x_i^2} < 0. \quad (4.3.14)$$

Equations (4.3.13) and (4.3.14) taken together imply that the average product function attains its maximum at the point where marginal and average products are equal. The sufficient condition for a true maximum is

$$\frac{\partial^2 (q/x_i)}{\partial x_i^2} = \frac{1}{x_i} f_{ii} - \frac{2}{x_i^2} \left[ f_i - \frac{q}{x_i} \right] = \frac{1}{x_i} f_{ii} < 0. \quad (4.3.15)$$

Since $x_i > 0$ at any factor point, the existence of a true maximum on the average product function requires the marginal product function to be in the stage of diminishing marginal returns.

4.3.5 Relations among the product curves

At this point it is useful to summarize the relations that exist among the total, marginal, and average product curves. The typical cross-section shown in figure 21 is reproduced in figure 22 as the total product curve OCE. The corresponding marginal and average product curves are $OA'C'D'E$ and $OBE$ respectively (at this time, ignore the lower panel of the figure).

By condition (4.2.2) imposed on the production function, the total product curve emanates from the origin. Since the function is zero, its first derivative is too. Hence the marginal product curve begins at the
4.3 **SINGLE INPUT VARIATIONS**

origin. By applying L'Hospital's Rule it may be shown that the average product curve also begins at the origin. Consequently, all three product curves emanate from the origin.

*Total and marginal product curves.* As previously seen, the marginal product curve may decrease monotonically throughout its entire range,

or it may have rising and falling segments. In the latter case, there may or may not be points of inflection along the marginal product curve. The marginal product function depicted in figure 22 illustrates the more general case in which the marginal product curve both rises and falls and in which it contains points of inflection.

Fig. 22
CONTINUOUS PRODUCTION FUNCTIONS

It is now convenient to state the following mathematical
Criterion: Consider any function \( \eta = \phi(x) \). The function is said to have a point of inflection\(^1\) at \( x = x \) if \( \phi_{xx} = 0 \). If \( \phi_{xxx} < 0 \), the function changes from concave from above to concave from below; if \( \phi_{xxx} > 0 \), concavity changes in the opposite direction.\(^2\)

The total product curve in the cross-section under consideration is

\[
q = f(x_i, x_j) \quad \text{for} \quad j \neq i. \tag{4.3.16}
\]

Hence marginal product, or the slope of the total product curve, is

\[
\frac{\partial q}{\partial x_i} = f_i. \tag{4.3.17}
\]

The marginal product function attains an extreme value when

\[
\frac{\partial^2 q}{\partial x_i^2} = \frac{\partial^2 f}{\partial x_i^2} = f_{ii} = 0; \tag{4.3.18}
\]

and the extremum is a maximum or a minimum according as

\[
\frac{\partial^2 q}{\partial x_i^2} = \frac{\partial^2 f}{\partial x_i^2} = f_{ii} \leq 0. \tag{4.3.19}
\]

If the marginal product curve attains a maximum at a point such as \( A' \), conditions (4.3.18) and (4.3.19) must hold. Hence by the criterion stated above, the total product curve has a point of inflection \( A \) at the input combination for which the marginal product curve is a maximum. The direction of concavity changes from ‘above’ to ‘below’.

In general production laws such as equation (4.2.10), the cross-section total product curve rises from zero to a maximum and then declines to zero again. The maximum point on the total product curve \( C \) is attained when

\[
\frac{\partial q}{\partial x_i} = f_i = 0 \quad \text{and} \quad \frac{\partial^2 q}{\partial x_i^2} = f_{ii} < 0. \tag{4.3.20}
\]

Hence marginal product must be zero and declining when total product reaches its maximum. These conditions are satisfied by point \( C' \) in figure 22.

The remaining relation of interest concerns the points \( D \) and \( D' \). First note that since the marginal product curve must be decreasing at \( C' \), it must become negative beyond \( C' \). Next, since total product is zero at \( E \), so is its derivative. Now let us summarize; \( a \) by assumption, the marginal product function is continuous; \( b \) marginal product is zero at \( C' \) and \( E \). Hence by Rolle’s Theorem, the marginal product function must have at

---

\(^1\) A point of inflection is a point at which the direction of concavity changes.

\(^2\) A function is said to be concave from above (below) at a point if it lies above (below) its tangent in a neighborhood of the point.
4.3 SINGLE INPUT VARIATIONS

least one extreme point between \( C' \) and \( E \). By assumption, the production function increases monotonically over the range from \( A \) to \( C \) and decreases monotonically over the range from \( C \) to \( E \). Therefore, the marginal product function has only one extreme point over the range; and since the marginal product is negative within the range, the extremum must be a minimum.

Suppose the minimum occurs at \( D' \). Thus at this point \( f_{ii} = 0 \) and \( f_{iii} > 0 \). Applying the criterion, the total product curve has a point of inflection at \( D \), where its concavity changes from ‘below’ to ‘above’.

Marginal and average products. The chief relation between the marginal and average product curves was established in section 4.3.4. Specifically, the average product curve attains its maximum at the point \( (B) \) where marginal product equals average product. For lower levels of input usage, marginal product exceeds average product; at greater input usages, average product exceeds marginal product. Beyond its maximum point, the average product curve declines monotonically, reaching zero when total product is zero.

The only other relation of interest concerns the concavity of the average product function. The average product curve may or may not have a point of inflection; but if it has one, it will have two. If a point of inflection exists, it will occur at the point where

\[
\frac{\partial^2(q(x_i))}{\partial x_i^2} = \frac{1}{x_i} f_{ii} - \frac{2}{x_i^2} \left( f_i - \frac{q}{x_i} \right) = 0.
\]  

(4.3.21)

Since \( f_{ii} = 0 \) at \( A' \) and \( f_i - \frac{q}{x_i} \neq 0 \) there,

the point of inflection must occur to the left of \( A' \) (and to the left of \( D' \)).

Next, observe that if a point of inflection exists to the left of \( A' \), the concavity of the average product curve must change from ‘above’ to ‘below’. Consider the third derivative:

\[
\frac{\partial^3(q(x_i))}{\partial x_i^3} = \frac{1}{x_i} f_{ii} - \frac{2}{x_i^2} f_{ii} + \frac{6}{x_i^3} \left( f_i - \frac{q}{x_i} \right).
\]  

(4.3.22)

In light of equation (4.3.21) and the known direction of concavity change if it changes, the existence of a point of inflection corresponding to equation (4.3.21) requires

\( f_{iii} < 0. \)  

(4.3.23)

Now consider the marginal product curve in the interval from \( 0 \) to \( A' \). At \( A', f_{iii} < 0 \); and \( f_{iii} \) may be less than zero throughout the range. Thus the average product curve may have a point of inflection in the range from \( 0 \) to \( A' \) whether or not the marginal product curve does. If the latter curve does have a point of inflection, \( f_{iii} \) must equal zero at some point between \( 0 \) and \( A' \). Further, since \( f_{iii} < 0 \) at \( A' \), \( f_{iii} \) must change from positive, to
CONTINUOUS PRODUCTION FUNCTIONS

zero, to negative over the range. Hence in light of inequality (4.3.23), if both the marginal and average product curves have a point of inflection in the range from 0 to $A'(A')$, the point of inflection on the average product curve corresponds to a greater input usage than does the point of inflection on the marginal product curve. If the point of inflection on the marginal product curve is associated with the input of OF units of $X_i$, the point of inflection on the average product curve must lie between $F$ and $G$. Therefore, the point of inflection on the average product curve occurs before the point of inflection on the total product curve. Similar statements apply, mutatis mutandis, to the point of inflection that occurs to the left of $D'$.

4.3.6 Output elasticity and the elasticity of average product

At any given input point, a comparison of marginal and average products permits one to determine the output elasticity of the input in question. Formally, the concept is stated in the following

Definition: The output elasticity of the $i$th input is the proportional change in output induced by a change in the $i$th input relative to the given proportional change in this input. For the production function in equation (4.3.2), the output elasticity of the $i$th input ($e_i$) is

$$e_i = \frac{\partial q}{\partial x_i} \frac{x_i}{q} = \frac{\delta q}{\delta x_i} \frac{x_i}{q} = f_i \frac{1}{q/x_i} = \frac{MP_i}{AP_i}. \quad (4.3.24)$$

Thus the output elasticity at an input point is the ratio of the marginal product to the average product at the point in question.

As examples, the output elasticities for equations (4.2.4)-(4.2.6), respectively, are

$$e_x = \alpha, \quad e_y = \beta; \quad (4.3.25)$$

$$e_x = \gamma^{1-\delta} \left( \frac{q}{x} \right)^\rho, \quad e_y = \gamma^{1-\delta} \left( \frac{q}{y} \right)^\rho; \quad (4.3.26)$$

$$e_x = \frac{2(x(y-bx))}{2axy-bx^2-cy^2}, \quad e_y = \frac{2y(ax-cy)}{2axy-bx^2-cy^2}. \quad (4.3.27)$$

A point worth noting from these examples and from equation (4.3.24) is that the output elasticity is typically a function of all the inputs. Thus the output elasticity generally depends upon the particular input point at which it is measured. In one special case, however, the output elasticities are constants independent of the factor point. This situation, of course, is represented by the Cobb–Douglas production function [equation (4.2.4)]; the output elasticities are shown in equation (4.3.25)].

The output elasticity can now be used to define the elasticity of average
4.3 SINGLE INPUT VARIATIONS

product. Using the conventional elasticity formula,

\[
\frac{\partial (q/x_i)}{\partial x_i} = \frac{x_i f_i q}{x_i^2} = \frac{f_i}{q/x_i} - 1 = \frac{MP_i}{AP_i} - 1 = \epsilon_i - 1.
\]

(4.3.28)

The elasticity of the average product curve is equal to the output elasticity minus one. Or, using the concept of average returns introduced in subsection 4.3.4, there exist increasing or decreasing average returns at a given point according as \( \epsilon_i \geq 1 \).

The behavior of the output elasticity is shown graphically in figure 22. At the origin both marginal and average products are zero. Using L'Hospital's Rule, it may be shown that \( \epsilon_i = 1 \). Moving away from the origin, marginal product exceeds average product, so \( \epsilon_i > 1 \). At point \( B \), where marginal and average products are equal, \( \epsilon_i \) becomes one again. Beyond \( B \), \( \epsilon_i \) diminishes, becoming zero when marginal product is zero and negative when the marginal product is negative. As constructed in figure 22, \( \epsilon_i \) reaches its minimum when marginal product does; however, this is not necessarily the case. The minimum of \( \epsilon_i \) does not have to coincide with the minimum marginal product. After attaining its minimum, \( \epsilon_i \) rises to zero at point \( E \), where all product curves are zero.

4.3.7 The stages of production

Since the appearance of Cassels' noted paper, it has become traditional to classify the 'stages of production' relative to single-input variations.\(^1\) According to Cassels, the stages of production are as shown in figure 23. The tasks are to show that production will occur only in Stage II and that point \( A \) divides Stages I and II.

First, we should emphasize that the problem at hand concerns the short run only. All inputs are fixed except one. Changes in output can be accomplished only by changes in this one variable input. It is quite obvious that production will not occur at a point to the right of \( B \), where total product is a maximum and marginal product is zero. In Stage III, any additional usage of the variable input diminishes, rather than augments, output. In older terminology, point \( B \) corresponds to the intensive margin beyond which production will not take place.

It is clear that the conditions of production undergo a fundamental change at the intensive margin; and it is reasonable to designate the portion of the product graph to the right of \( B \) as a separate 'stage of production'. To establish the dividing point between Stages I and II requires somewhat more explanation.

\(^1\) Cassels [1936, esp. pp. 226–8].
CONTINUOUS PRODUCTION FUNCTIONS

To begin, note that the point of division is established at $A$, corresponding to maximum average product or the *extensive margin*. In the case of functions homogeneous of degree one, this is clearly the point of demarcation (because the marginal product of the fixed factor is negative to the left of $A$). Otherwise, it is necessary to use the concept of output elasticity developed in subsection 4.3.6. above.

Output elasticity ($\epsilon_i$) is the proportional change in output relative to a proportional change in input. If $\epsilon_i > 1$, a given increase in usage of the variable input results in a proportionately greater increase in output. The point is that production would never take place over the range of input values for which $\epsilon_i > 1$ because an increase in output could be secured by refusing to utilize all available units of the fixed inputs. Over this range, the entrepreneur has two methods of expanding output: (a) expand the usage of the variable input, and (b) reduce the utilization of fixed inputs. The latter would always be chosen if a volume of output corresponding to Stage I were to be produced. But in this eventuality, the product curves
4.3 SINGLE INPUT VARIATIONS

would have to be redrawn because the effective amount of fixed input has changed.

Relative to the original amount of fixed input, production would never occur over the range for which \( \epsilon_i > 1 \). It would occur over the range for which \( 0 \leq \epsilon_i \leq 1 \). Since \( \epsilon_i = 1 \) when marginal product equals average product, the division point \( A \) is established. There are indeed three distinct stages of production; and a rational entrepreneur will produce only in Stage II, between the extensive and intensive margins.

4.4 The function coefficient\(^1\)

Attention in this chapter has so far been directed to the behavior of output when one input is varied, all other inputs being held constant at some specified level. The same study was undertaken in chapter 2; a comparison of results shows the essential differences between fixed and variable proportions production functions for single input variations. Additionally, in chapter 2 the behavior of output for proportional variations of all inputs was analyzed. It is to a similar task that the present section is devoted.

4.4.1 Proportional input variations and the function coefficient

The concept of the function coefficient is introduced by the following

**Definition:** The function coefficient is the elasticity of output with respect to an equiproportional variation of all inputs. Thus the function coefficient is the proportional change in output relative to the proportional changes in the inputs for movements along a ray from the origin in input space.

Consider any initial input point \( x^0_1, x^0_2, \ldots, x^0_n \). By equation (4.2.1), output is

\[
q^0 = f(x^0_1, x^0_2, \ldots, x^0_n).
\]

(4.4.1)

Let each input be increased by the same proportion \( \lambda \). Thus

\[
x_i = \lambda x^0_i = x_i(\lambda) \quad (i = 1, 2, \ldots, n),
\]

(4.4.2)

and

\[
q = f(\lambda x^0_1, \lambda x^0_2, \ldots, \lambda x^0_n) = g(\lambda).
\]

(4.4.3)

Using equation (4.4.3), the definition of the function coefficient (\( \epsilon \)) may be given more precisely by

\[
\epsilon = \frac{dq}{q} \frac{d\lambda}{\lambda} = \frac{dg(\lambda)}{d\lambda} \frac{\lambda}{g(\lambda)} = \frac{d\ln q}{d\ln \lambda}.
\]

(4.4.4)

where \( \ln \) denotes the natural logarithm.

\(^1\) The terminology for the concept in question varies. The term ‘function coefficient’ is used by Carlson [1939]. Schneider [1934] uses ‘Ergebigsgrad der Produktion’ for this concept and ‘Ertragsfunktion’ to refer to equiproportional variation of inputs. Johnson [1913] uses ‘elasticity of production’ to describe the concept, while Allen [1938] uses ‘elasticity of productivity’. Finally, Frisch [1965] uses ‘passus coefficient’.
CONTINUOUS PRODUCTION FUNCTIONS

One feature of the function coefficient is immediately apparent from equation (4.4.4). Specifically, \( \lambda \) is a scale factor; hence the function coefficient is the elasticity of output with respect to scale. Thus in a neighborhood of any input point, production is subject to increasing, constant, or decreasing returns to scale according as \( e \geq 1 \).

As shown in chapter 5, the function coefficient is a constant when the production function is homogeneous; indeed, \( e \) is the (constant) degree of homogeneity of the function. If the production function is not homogeneous, however, the function coefficient is a variable that depends upon (a) the specific factor proportion (ray) at which it is measured, and (b) the scale of input usage. Thus for inhomogeneous functions, the magnitude of the function coefficient depends not only upon the ray along which it is measured but also the point at which it is measured. Further, as should be obvious from equation (4.4.4), the function coefficient has a range of negative values if the production surface has a maximum point.

Total output and the function coefficient for equation (4.2.10) are shown in table 5. It is interesting to note that the function coefficient is zero at the input point corresponding to maximum total product. This relation follows immediately from equation (4.4.4) inasmuch as \( dq \) must be instantaneously zero at the maximum point.

<table>
<thead>
<tr>
<th>( x = y )</th>
<th>Output</th>
<th>Function coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.50</td>
<td>1667</td>
<td>2.500</td>
</tr>
<tr>
<td>0.60</td>
<td>2592</td>
<td>2.333</td>
</tr>
<tr>
<td>0.70</td>
<td>3659</td>
<td>2.125</td>
</tr>
<tr>
<td>0.80</td>
<td>4779</td>
<td>1.857</td>
</tr>
<tr>
<td>0.90</td>
<td>5832</td>
<td>1.500</td>
</tr>
<tr>
<td>1.00</td>
<td>6667</td>
<td>1.000</td>
</tr>
<tr>
<td>1.10</td>
<td>7099</td>
<td>0.250</td>
</tr>
<tr>
<td>1.110</td>
<td>7112</td>
<td>0.154</td>
</tr>
<tr>
<td>1.125</td>
<td>7119</td>
<td>0.000</td>
</tr>
<tr>
<td>1.130</td>
<td>7118</td>
<td>-0.054</td>
</tr>
<tr>
<td>1.140</td>
<td>7111</td>
<td>-0.167</td>
</tr>
<tr>
<td>1.150</td>
<td>7097</td>
<td>-0.286</td>
</tr>
<tr>
<td>1.500</td>
<td>0</td>
<td>-2.000</td>
</tr>
</tbody>
</table>

TABLE 5 Output and function coefficients from equation (4.2.10)
4.4 FUNCTION COEFFICIENT

4.4.2 Derivation of the function coefficient

The total differential of the production function in equation (4.2.1) is

\[ dq = f_1 dx_1 + f_2 dx_2 + \ldots + f_n dx_n, \]  

or

\[ dq = f_1 x_1 \frac{dx_1}{x_1} + f_2 x_2 \frac{dx_2}{x_2} + \ldots + f_n x_n \frac{dx_n}{x_n}. \]  

(4.4.6)

When all inputs are varied equipropotionately by the factor \( \lambda \),

\[ \frac{dx_i}{x_i} = \frac{d\lambda}{\lambda} \quad (i = 1, 2, \ldots, n). \]  

(4.4.7)

Substituting equation (4.4.7) in equation (4.4.6) yields

\[ dq = \left( \sum_i f_i x_i \right) \frac{d\lambda}{\lambda}. \]  

(4.4.8)

Using definition (4.4.4), one obtains

\[ \epsilon = \frac{\sum_i f_i x_i}{q}, \]  

(4.4.9)

the desired expression for the function coefficient.

While the derivation given above is the most straightforward, the expression for the function coefficient in equation (4.4.9) may be derived in two alternative ways.\(^1\) To obtain the first alternative, differentiate equation (4.4.3) with respect to \( \lambda \), obtaining

\[ \sum_i \frac{\partial q}{\partial x_i} \frac{dx_i}{d\lambda} = \frac{dg(\lambda)}{d\lambda}. \]  

(4.4.10)

From equations (4.4.2) it follows that

\[ \frac{dx_i}{d\lambda} = \frac{d(\lambda x_i^0)}{d\lambda} = x_i^0 \quad (i = 1, 2, \ldots, n). \]  

(4.4.11)

Using equation (4.4.11) in equation (4.4.10), one obtains

\[ \sum_i f_i x_i^0 = \frac{dg(\lambda)}{d\lambda}, \]  

(4.4.12)

from which it follows that

\[ \sum f_i(\lambda x_i^0) = \sum f_i x_i = \lambda \frac{dg(\lambda)}{d\lambda} = \left( \frac{dg(\lambda)}{d\lambda} \frac{\lambda}{g(\lambda)} \right) q. \]  

(4.4.13)

Finally, using the definition \( \epsilon \) from equation (4.4.4) reduces equation (4.4.13) to equation (4.4.9).

\(^1\) Both these derivations are presumably attributable to Frisch [1965, pp. 75–7].
CONTINUOUS PRODUCTION FUNCTIONS

The final derivation depends upon the fact that the function coefficient is generally dependent upon the particular input point at which it is measured. That is, for any specified ray, the function coefficient depends upon the scale factor. Symbolically,
\[ e = h(x_1, x_2, \ldots, x_n) = h(\lambda x_1^0, \lambda x_2^0, \ldots, \lambda x_n^0) = e(\lambda). \quad (4.4.14) \]

Substituting equation (4.4.14) into equation (4.4.4), one obtains
\[ d \ln g(\lambda) = e(\lambda) d \ln \lambda. \quad (4.4.15) \]

Next, integrate, equation (4.4.15) between two limits, say \( \lambda = 1 \) and \( \lambda = M \), obtaining
\[ \ln g(M) - \ln g(1) = \int_1^M e(\lambda) d \ln \lambda. \quad (4.4.16) \]

Taking the antilogarithm of equation (4.4.16) yields
\[ g(M) = g(1) \exp \left( \int_1^M e(\lambda) d \ln \lambda \right). \quad (4.4.17) \]

Inserting equation (4.4.3) in equation (4.4.17), one obtains
\[ f(Mx_1^0, Mx_2^0, \ldots, Mx_n^0) = f(x_1^0, x_2^0, \ldots, x_n^0) \exp \left( \int_1^M e(\lambda) d \ln \lambda \right). \quad (4.4.18) \]

Next, differentiate equation (4.4.18) with respect to \( M \). This results in
\[ \sum_i \frac{\partial f}{\partial (Mx_i^0)} \frac{d(Mx_i^0)}{dM} = q \frac{d \exp \left( \int_1^M e(\lambda) d \ln \lambda \right)}{dM}, \quad (4.4.19) \]

or since \( d \ln \lambda = \frac{1}{\lambda} d\lambda \) and \( \frac{d(Mx_i^0)}{dM} = x_i^0 \),
\[ \sum_i \frac{\partial f}{\partial (Mx_i^0)} x_i^0 = q \exp \left( \int_1^M \frac{1}{\lambda} e(\lambda) d\lambda \right) \frac{e(M)}{M}. \quad (4.4.20) \]

Since equation (4.4.18) holds for any \( M \), one may set \( M = 1 \) in equation (4.4.20). In this case
\[ \int_1^{M-1} \frac{1}{\lambda} e(\lambda) d\lambda = 0, \quad (4.4.21) \]

so equation (4.4.20) becomes
\[ \sum_i \frac{\partial f}{\partial x_i^0} x_i^0 = qe(1). \quad (4.4.22) \]

Finally, using equation (4.4.14) in equation (4.4.22), one obtains
\[ \sum f_i x_i^0 = qe(x_1^0, x_2^0, \ldots, x_n^0). \quad (4.4.23) \]
4.4 Function Coefficient

In this expression all variables are related to the specific input point \( x_0, x_1, x_2, \ldots, x_n \). However, since this point is arbitrary, equation (4.4.9) follows immediately from equation (4.4.23).

4.4.3 Relation of the function coefficient to output elasticities and the elasticity of average product

The general expression for the function coefficient in equation (4.4.9) may be expanded and written as

\[
e = \frac{\partial f}{\partial x_1} \frac{x_1}{q} + \frac{\partial f}{\partial x_2} \frac{x_2}{q} + \cdots + \frac{\partial f}{\partial x_n} \frac{x_n}{q}.
\]  

(4.4.24)

Using the definition of output elasticity from equation (4.3.24), equation (4.4.24) becomes

\[
e = \varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_n.
\]  

(4.4.25)

Thus the function coefficient is the sum of all output elasticities. In a way, this merely reemphasizes a proposition familiar from the calculus: the variation of a function resulting from the simultaneous variation of all arguments equals the sum of the variations in the function ascribable to independent variations of the arguments. Yet it also provides an important connection between simultaneous and independent input variations. In particular, the ultimate proportional change in output is the same whether all inputs are varied simultaneously or one at a time.

An even more meaningful relation may be established between returns to scale and average returns. Consider the expression for the elasticity of the average product of the \( i \)th input with respect to the scale factor \( \lambda \):

\[
\frac{d(q/x_i)}{d\lambda} \frac{\lambda x_i}{q} = \frac{dq}{d\lambda} \frac{\lambda}{q} \frac{dx_i}{d\lambda} \frac{\lambda}{x_i} \quad (i = 1, 2, \ldots, n).
\]  

(4.4.26)

First, note that the elasticity of any input with respect to the scale factor is unity. Using this information, together with equation (4.4.4), in equation (4.4.26) yields

\[
\frac{d(q/x_i)}{d\lambda} \frac{\lambda x_i}{q} = \varepsilon - 1 \quad (i = 1, 2, \ldots, n).
\]  

(4.4.27)

The results contained in equation (4.4.27) are of the first order of importance. For small variations along an input ray, all average products increase, remain unchanged, or decrease according as \( \varepsilon \geq 1 \). In other words, for proportional input changes, there exist increasing, constant, or diminishing average returns for each input according as production occurs in the region of increasing, constant, or decreasing returns to scale.
CONTINUOUS PRODUCTION FUNCTIONS

4.5 Simultaneous input variations

To this point our attention has been directed exclusively to the variations in output attributable to variations in the use of a single productive input, all other inputs held constant. This study is important because it emphasizes one aspect of production under conditions of variable proportions: output changes in response to a change in factor proportions. Yet this is but one of the essential features of variable proportions production functions. The other is that the same output can be produced by a variety of factor proportions, i.e., one input may be substituted for another without affecting the level of output. To emphasize this important aspect of production, one must analyze the effects of simultaneous input variations.

4.5.1 The marginal rate of technical substitution

As stated above, one input may sometimes be substituted for another without affecting the level of output. In other words, over a certain range inputs may be substituted along an isoquant so that factor proportions vary while output remains unchanged. In effect, one input is substituted for another within this range. It is of considerable interest to measure the rate at which such a substitution can take place.

Definition: The marginal rate of technical substitution of input \( j \) for input \( j(MRTS) \) is the number of units by which the usage of \( j \) may be reduced when the usage of \( i \) is expanded by one unit so as to maintain a constant level of output.

To get at a precise measure of the marginal rate, consider the isoquant \( Q_0 \) between the points \( A \) and \( B \) in figure 24. Suppose production is initially organized at point \( P(x_1, y_1) \) and subsequently moves to point \( S(x_2, y_2) \). In the change \( X \) is substituted for \( Y \) without affecting the level of output. The rate at which \( X \) is substituted for \( Y \) is

\[
-\frac{Oy_1 - Oy_2}{Ox_1 - Ox_2} = \frac{PR}{RS}.
\]

(4.5.1)

Equation (4.5.1) shows the (positive) rate of substitution for a finite move along \( Q_0 \). However, the closer is \( P \) to \( S \), the more nearly is the rate of substitution approximated by the tangent to \( Q_0 \) at \( S \), i.e., \(-\tan TS\). And in the limit, when \( P \) is arbitrarily close to \( S \), the marginal rate of technical substitution of input \( Y \) for input \( X \) is given by the negative of the slope of the isoquant at \( S \).

Now consider the production function in equation (4.2.1) and suppose only two inputs change, \( X_1 \) and \( X_2 \). Further, impose the restriction that output does not change. Taking the (relevant) total differential, one obtains

\[
dq = f_1dx_1 + f_2dx_2 = 0.
\]

(4.5.2)
4.5 SIMULTANEOUS INPUT VARIATIONS

Equation (4.5.2) is the differential equation for an isoquant whose slope is, accordingly,
\[ \frac{dx_j}{dx_i} = \frac{f_i}{f_j}, \]  
(4.5.3)

Using the previous results,
\[ MRTS_{ij} = -\frac{dx_j}{dx_i} = \frac{f_i}{f_j}, \]  
(4.5.4)

Thus the marginal rate of technical substitution of \( j \) for \( i \) is the ratio of the marginal product of \( i \) to the marginal product of \( j \).

The result (4.5.4) has been obtained under the assumption that input variations are restricted to the arc \( AB \) in figure 24. Over this region, inputs are truly substituted one for another inasmuch as the use of additional units of one input permits some units of the other input to be released. The same is not true beyond points \( A \) and \( B \). At point \( A \), the marginal product of \( Y \) is zero, and it is negative beyond. Thus if the usage of \( Y \) is expanded beyond \( A \), the usage of \( X \) must also expand to offset the depressive output effect of the negative marginal product. The same statement applies, mutatis mutandis, for a movement beyond \( B \), at which point the marginal product of \( X \) is zero. Beyond points \( A \) and \( B \), the inputs cannot be substituted inasmuch as both must be increased to maintain a constant level of output. Hence the marginal rate of technical substitution is properly
CONTINUOUS PRODUCTION FUNCTIONS

defined only over the range in which both marginal products are non-negative. Thus over the substitution range,

\[ MRTS_{it} = -\frac{dx_i}{dx_t} = \frac{f_t}{f_i} \leq 0. \]  (4.5-5)

4.5.2 Production isoclines

Since an isquant is smoothly continuous by assumption, its slope changes continuously for movements along it. Hence in light of equation (4.5-4), the marginal rate of technical substitution also varies continuously for movements along an isquant. Indeed, over the substitution range, the marginal rate of technical substitution diminishes from an arbitrarily large number (at \( A \)) to zero (at \( B \)). Since this is true for one isquant, it is true for all isocurves. Consequently, given an isquant map and any stipulated values of the marginal rate of technical substitution, there exists precisely one point on each isocline at which the marginal rate of technical substitution has the stipulated value.

Definition: A locus of points along which the marginal rate of technical substitution is constant is called an isocline.

On a purely formal level, the isocline corresponding to the marginal rate of technical substitution whose value is \( \bar{c} \) (a constant) is given by the following partial differential equation:

\[ f_t(x_t, x_i) = \bar{c} f_i(x_t, x_i), \]  (4.5-4)

where there are no constants of integration since \( f_t(0, 0) = f_t(0, 0) = 0 \).

The graphical derivation of an isocline is illustrated in figure 25. \( Q_0, Q_1, \) and \( Q_2 \) are isocurves and \( CF, CF', \) and \( CF'' \) are lines whose slopes are identical. Hence the slopes of \( Q_0, Q_1, \) and \( Q_2 \) are the same at \( P, R, \) and \( S \) respectively. Since the (negative) slope of an isocline is the marginal rate of technical substitution, these rates are the same at \( P, R, \) and \( S \). Connecting all points such as \( P, R, \) and \( S \) generates an isocline as that shown in figure 25. It is worth noting that since the integral of equation (4.5-6) does not contain a constant term, all isocurves emanate from the origin.

4.5.3 The substitution region

In subsection 4.5.1 it was noted that the marginal rate of technical substitution could only be defined for the negatively sloped portion of an isocline. In certain cases, such as the production functions shown in

\footnote{As previously noted, certain production functions are such that the marginal products never become zero. However, even in these cases, the marginal products approach zero asymptotically, and the statement in the text holds in the limit.}
4.5 SIMULTANEOUS INPUT VARIATIONS

equations (4.2.4) and (4.2.5), the isoquants are negatively sloped throughout. More generally, however, the isoquant maps resemble those shown in figures 20 and 26. In the former case, the entire map comprises the substitution region; in the latter cases, illustrated by figures 20 and 26, however, the substitution region is only a portion of the entire input space.

Definition: 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 may be substituted for another while maintaining a constant level of output.

The substitution region is determined (see figure 26) by finding the two isoclines corresponding to infinite and zero marginal rates of technical substitution. In figure 26 these two isoclines are given by $OS$ and $OS'$ respectively. $OS$ is the locus along which the marginal product of $Y$ is zero; hence the marginal rate of technical substitution is indefinitely large. The opposite characterizes $OS'$, which is the locus along which the marginal
CONTINUOUS PRODUCTION FUNCTIONS

product of $X$ is zero. Hence the marginal rate of technical substitution is also zero.

The substitution region lies between these two limiting isoclines, which are sometimes called ridge lines. This terminology is presumably based upon the fact that the cross-section total product curves attain maxima in the $Q-Y$ plane for input combinations associated with $OS$ and in the $Q-X$ plane for input combinations associated with $OS'$. 

![Diagram](image)

Fig. 26

4.5.4 The diminishing marginal rate of technical substitution

By equation (4.5.3), the slope of an isoquant is

$$\frac{dx_i}{dx_i} = -\frac{f_i}{f_i}.$$  \hspace{1cm} (4.5.7)

Hence the concavity of the isoquant depends upon the second derivative:

$$\frac{d^2x_i}{dx_i^2} = \frac{d\left(-\frac{f_i}{f_i}\right)}{dx_i} = -\frac{f_i}{f_i^2} \left( f_{ii} + f_{ij} \frac{dx_j}{dx_i} - f_i f_{ij} \frac{dx_j}{dx_i} \right).$$  \hspace{1cm} (4.5.8)
4.5 SIMULTANEOUS INPUT VARIATIONS

Substituting from equation (4.5.7) yields

\[ \frac{d^2 x_i}{d x_i^2} = -\frac{1}{f_i^2} \left( f_i^2 f_{ii} - 2f_i f_{ij} + f_j^2 f_{ij} \right). \]  

(4.5.9)

If the right-hand side of equation (4.5.9) is positive, the isoquant is concave from above; if negative, it is concave from below.

Over the substitution region, all isoquants are concave from above. Within the region, \( f_i > 0 \) and \( f_{ii} < 0 \) for all \( i \). In almost all cases, \( f_{ij} > 0 \) in the substitution region. But regardless of the sign of \( f_{ii} \), the term in parentheses on the right in equation (4.5.9) must be negative; hence the right-hand side is positive, and the isoquant is concave from above in the substitution region.¹

The fact that isoquants are concave from above in the substitution region enables one to establish the following

**Proposition:** as \( X_i \) is substituted for \( X_j \) so as to maintain a constant level of output, the marginal rate of technical substitution declines. Thus throughout the substitution region, there is a *diminishing marginal rate of technical substitution.*

¹ If the production function is homogeneous of degree one, \( f_{ii} \) and \( f_{ij} \) necessarily have opposite signs. Otherwise, \( f_{ii} \) and \( f_{ij} \) may have identical signs. In this case, the concavity of the isoquant cannot be inferred *a priori* from the technical theory of production. However, as shown in chapter 6, economic efficiency can be attained if, and only if, production isoquants are concave from above within the substitution region.
CONTINUOUS PRODUCTION FUNCTIONS

4.5.5 The elasticity of substitution

Consider the isoquant $Q_a$ in panel a, figure 27, and the $(Y:X)$ input ratio represented by the ray $OP$, i.e. $y/x = \tan \theta$. At this input ratio and scale of operation, the slope of the isoquant at point $P$ gives the marginal rate of technical substitution. Now transfer to panel $b$ and plot this pair of values, i.e. the $X:Y$ ratio and the associated marginal rate of technical substitution. The point $P$ on $Q_a$ might, for example, give rise to the point $P'$ in panel $b$.

Next, select another input ratio, say $OR$, where $y/x = \tan \phi$. This gives rise to another value of the marginal rate of technical substitution. Plot this point, and many similar points, on Panel $b$ and connect all such points. The resulting curve, called the substitution curve, shows how factor proportions change in response to a change in the marginal rate of technical substitution. The elasticity of this curve, called the elasticity of substitution, is very important in the neoclassical theory of distribution.\footnote{This graphical treatment of the elasticity of substitution is presumably attributable to Lerner [1933]. It was also used in Ferguson [1944].}

The elasticity of substitution is a relatively new concept, having been introduced by Hicks in 1932.\footnote{Hicks [1932, p. 117 and pp. 244-5].} Hicks himself did not give a precise definition in the text of his book; ‘...the “elasticity of substitution” is a measure of the ease with which the varying factor can be substituted for others’ [1932, p. 117]. He did give an exact mathematical formulation of the concept in an appendix [pp. 244-5], but the mathematical definition applies only to production functions homogeneous of degree one.

Shortly after Hicks’ Theory of Wages appeared, Mrs Robinson gave the concept a more precise definition: ‘...the proportionate change in the ratio of the amounts of the factors divided by the proportionate change in the ratio of their marginal physical productivities.’\footnote{Mrs Robinson also defined the elasticity of substitution in terms of input prices (p. 256). However, she states (p. 330) that the definition given above is the more fundamental one.} Subsequent to its introduction by Hicks and Mrs Robinson, the elasticity of substitution became the subject of intense investigation, possibly because the concept was initially misunderstood by many economists.\footnote{For a sampling of this literature, see Champernowne [1935], Friedman [1936], Hicks [1933], Hicks [1936], Kahn [1933], Kahn [1935], Kaldor [1937], Lerner [1932], Lerner [1934], Machlup [1936], Machlup [1936a], Meade [1934a], Pigou [1934], Robinson [1936], Sweezy [1933], and Tarshis [1934].} However, it would seem that by 1938,\footnote{Allen [1938, pp. 340-3].} the concept was well established and widely understood.
4.5 SIMULTANEOUS INPUT VARIATIONS

By equation (4.5.3) the marginal rate of technical substitution (denoted by $s$ below) is

$$s = -\frac{dx_i}{dx_t} = \frac{f_i}{f_t}, \quad (4.5.10)$$

Further, let $y$ represent the input ratio $x_i/x_t$. The elasticity of substitution ($\sigma$) is given by the following.

**Definition:** The elasticity of substitution of $X_i$ for $X_t$ is

$$\sigma = \frac{dy}{y} \frac{ds}{s}, \quad (4.5.11)$$

where the differentials are restricted to variations along an isoquant. Thus the elasticity of substitution refers only to input substitutions associated with a constant level of output.

The formula (4.5.11) may be written directly in terms of the partial derivatives of the production function. First, note that

$$dy = x_i dx_t - x_t dx_i, \quad (4.5.12)$$

and

$$ds = \frac{\partial s}{\partial x_t} dx_t + \frac{\partial s}{\partial x_i} dx_i, \quad (4.5.13)$$

Since

$$dx_i = -\frac{f_i}{f_t} dx_t = -s dx_t, \quad (4.5.14)$$

equations (4.5.12) and (4.5.13) may be written

$$dy = -\frac{s x_t + x_i}{x_i^2} dx_t, \quad (4.5.15)$$

and

$$ds = \left( s \frac{\partial s}{\partial x_t} - \frac{\partial s}{\partial x_i} \right) dx_i. \quad (4.5.16)$$

Next,

$$\frac{\partial s}{\partial x_t} = \frac{\partial (f_i f_t)}{\partial x_t} = \frac{f_i f_{tt} - f_t f_{ti}}{f_t^2}, \quad (4.5.17)$$

and

$$\frac{\partial s}{\partial x_i} = \frac{\partial (f_i f_t)}{\partial x_i} = \frac{f_i f_{tt} - f_t f_{ti}}{f_t^2}. \quad (4.5.18)$$

Substituting equations (4.5.15)-(4.5.18) in formula (4.5.11) yields an expression for the elasticity of substitution based directly upon the production function:

$$\sigma = -\frac{f_i f_t (x_i f_t + x_t f_t)}{x_i x_t (f_i f_t^2 - 2 f_t f_i f_t + f_i f_t^2)}, \quad (4.5.19)$$

In this form it is readily seen that there is no 'problem of symmetry' at all.\(^1\)

\(^1\) The 'problem of symmetry' was posed by Tarshis [1934] and Machlup [1935]; the 'problem' was disposed of by Lerner [1938].
CONTINUOUS PRODUCTION FUNCTIONS

The elasticity of substitution of \( X_i \) for \( X_j \) is precisely the same as the elasticity of substitution of \( X_j \) for \( X_i \). Considering an alternative expression for the elasticity of substitution enables one to obtain additional information. Substituting equations (4.5.15) and (4.5.16) in formula (4.5.11) yields:

\[
\sigma = \frac{s}{x_i x_j} \frac{x_i s + x_j}{s (\partial s / \partial x_i) - (\partial s / \partial x_j)}. \tag{4.5.20}
\]

The change in the slope of an isoquant is given by

\[
\frac{d^2 x_j}{dx_i^2} = -\frac{\partial s}{dx_i} = \frac{\partial s}{\partial x_j} \frac{\partial^2 s}{\partial x_i \partial x_j}. \tag{4.5.21}
\]

Hence the elasticity of substitution may also be written

\[
\sigma = \frac{s}{x_i x_j} \frac{x_i s + x_j}{(d^2 x_j / dx_i^2)}. \tag{4.5.22}
\]

Since \( s, x_i \), and \( x_j \) are positive, the elasticity of substitution is inversely proportional to the change in the slope of the isoquant. Further, since an isoquant must be concave from above over the substitution region, \( d^2 x_j / dx_i^2 > 0 \). Thus the elasticity of substitution is a nonnegative magnitude. Two limiting values of the elasticity of substitution may be found by using equation (4.5.22). First, if the two inputs are perfect substitutes, the isoquants are a series of straight lines. The second derivative is accordingly zero, and the elasticity of substitution is infinite. The other extreme occurs when the inputs cannot be substituted for one another while maintaining a constant output. This is the case of fixed proportions between limited inputs. The isoquants are right angles, the second derivative approaches a limiting value of infinity, and the elasticity of substitution approaches a limiting value of zero. Generally, the inputs must be mutually limitative at a point if the elasticity of substitution is to be a nonzero magnitude.\(^1\)

In summary, the elasticity of substitution is a nonnegative measure of the relative ease with which one input may be substituted for another while maintaining a constant output. Typically, the elasticity is zero when the inputs are mutually limitational and strictly positive when the inputs are mutually limitative.\(^2\) The elasticity is a pure number independent of the units in which inputs and outputs are measured. Finally, it is a symmetrical relation that is typically a function of the input point at which it is measured.

\(^1\) For exceptions, see subsection 2.4.2.

\(^2\) An exception in the case of mutual limitativeness is given by the following production function:

\[ q = ax + by \ (a \leq 0, b \leq 0). \]