THE UNIVERSITY OF MANITOBA

THE CONCEPT OF LONG RUN AVERAGE COST: THEORY AND MEASUREMENT

By

George E. Gore

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE

OF MASTER OF ARTS

DEPARTMENT..ECONOMICS..................

WINNIPEG, MANITOBA

May 1973

Abstract

This thesis is concerned with the relationship between the theory of long run average cost and the cost-output relationship observed in practice. The specific purpose of the study is to assess the usefulness of the theoretical definition of long run average cost and the empirical verifiability of the hypothesis that long run average cost declines, reaches a minimum, and rises thereafter.

The first chapter examines the assumptions and derivation of the long run average cost curve in theory. It is shown that the long run average cost curve may be interpreted alternatively as the envelope of the short run average cost curves,or as the locus of points for which the ratio of the marginal productivities of the inputs equal the ratio of their marginal expenses. The importance of rising long run average costs for profit maximization under perfect competition is then demonstrated, followed by consideration of the factors determining the shape of the cost-output relationship in practice. It is argued that increasing complexity of the managerial function will tend to increase costs, but that this may be offset by forces making for economies of scale. A conflict between the strictly theoretical definition of long run average cost and certain observed sources of economies and diseconomies of scale is also noted. However, the contention that constant returns to scale necessarily follows from the theoretical definition of long run average cost is shown to be unjustified.

The above analysis is carried out under the assumption of perfect competition. It is argued that imperfect competition has three main implications for the analysis. First, pecuniary economies and pecuniary diseconomies associated with changing factor prices would be accounted for. However, it is pointed out that pecuniary economies may vitiate the prediction that the ratio of the marginal productivities of the inputs equals the ratio of their marginal expenses. Second, profit maximization and declining costs per unit of output are no longer incompatible. Third, in relaxing the assumption of perfect competition the existence of selling costs can be recognized. However, possible non-reversibility and variation of the sales-cost relationship with price are advanced as significant obstacles to the incorporation of selling costs.

Finally, consideration is given to whether the long run average cost curve should be revised to conform more closely to the real world. It is argued that while the inclusion of factor price changes, and indivisibilities leave the underlying logic of marginal productivity theory unchanged, more fundamental problems are posed by stochastic economies and selling costs. Nonetheless, the adequacy of theory is argued not to depend upon the realism of its' assumptions. Rather, the criterion employed to evaluate

the theory of long run average cost is whether the application of the concepts involved yield consistent and accurate predictions.

Following the theory of long run average cost the empirical techniques used to determine the nature of returns to scale are examined. For each of the different cost estimation techniques - statistical production and cost analysis, the survivor technique, the questionnaire and interview method, and engineering estimates - the general methodology involved is analyzed, followed by several specific illustrative studies. The studies selected were chosen so as to reveal points of methodology; either general problems involved in using a particular method of investigating costs or the ability to circumvent problems arising in the use of other techniques.

How the theoretical long run average cost concept has been modified in investigating the scale-cost relationship is then considered. The revisions of theory implied in the use of each method of determining cost are discussed. Through a comparative evaluation of the efficacy of the different cost estimation techniques the U-shaped long run average cost curve is concluded not to be representative of cost conditions found in industry. Finally, it is argued that from the viewpoint of determining whether the cost-output relationship is U-shaped marginal productivity theory provides an adequate conceptual framework.

Table of Contents

ACKNOWLEDGEMENTS

The author would like to express his appreciation to Professor Ralph F. Harris of the Department of Economics of the University of Manitoba for his support in the preparation of this thesis.

Introduction

One of the main economic foundations of the anti-combines laws is based on the shape of the long run average cost relationship. The rise in unit costs at small levels of output, predicted by the theory of pure or perfect competition, constitutes a strong reason for the preservation of competition. While cases of continually declining long run average costs have been recognized in the case of public utilities the traditionally accepted hypothesis has been that long run average cost declines, reaches a minimum, and rises thereafter. However, the results of new empirical techniques incorporating advanced statistical analysis do not support the existence of diseconomies of scale. Rather costs which fall sharply at first followed by constant costs or an assymptotic cost-output relationship are more typically encountered.

Proposed revisions of competition policy in Canada contained in the Interim Report on Competition Policy stress the need to consider economies of scale. Ideally an industry structure would be created such that firms would be large enough to exploit scale economies, but where a sufficiently large number of firms would exist to ensure the transmission of benefits to the consumer. To assess whether greater concentration in any particular industry is desirable there is clearly a need to quantify the extent of economies of scale. For different methods of investigating costs certain problems may be identified. Accurate measurement of scale economies will then require evaluating the strengths and weaknesses of different cost estimation techniques in relation to the technical conditions specific to each industry. However, in addition to the problem of selecting that empirical tool which yields statistically valid results, the existing studies show that there is a need to ensure that the economic meaning of the relationships is not violated.

CHAPTER I

THE THEORY OF LONG RUN AVERAGE COST AND ARGUMENTS FOR ECONOMIES AND DISECONOMIES OF SCALE

The concept of long run average cost attempts to show the effects on costs per unit attributable solely to increases in output when all inputs are variable, and combined so as to minimize costs for each level of output. The nature of the cost-output relationship is described in terms of internal economies or internal diseconomies of scale. Internal economies and internal diseconomies should be distinguished from external economies and external diseconomies; the latter external cost effects resulting from changes in the growth of the industry as a whole. Internal economies and internal diseconomies will be referred to respectively simply as economies and diseconomies of scale. Economies of scale as defined here will be said to exist when costs per unit of output are falling. Conversely, diseconomies of scale occur where costs per unit are rising.

The type of U-shaped short run average cost curve for the firm, when the capital stock is taken as given, is

* The method of citing references is adopted from The Government of Canada style manual for writers and editors. Ottawa, Queens Printer, 1962.

a textbook commonplace.[1] It is assumed that the firm produces a single homogeneous product. Consequently, costs are composed of production and distribution expenditures with selling costs being excluded. Moreover, cost per unit of factor input refer to the opportunity cost involved which is defined as the best rate of return the input could obtain in alternative employments. The relationships of average total cost, average variable cost, and marginal cost in the short run are shown by the curves ATC, AVC, and MC.

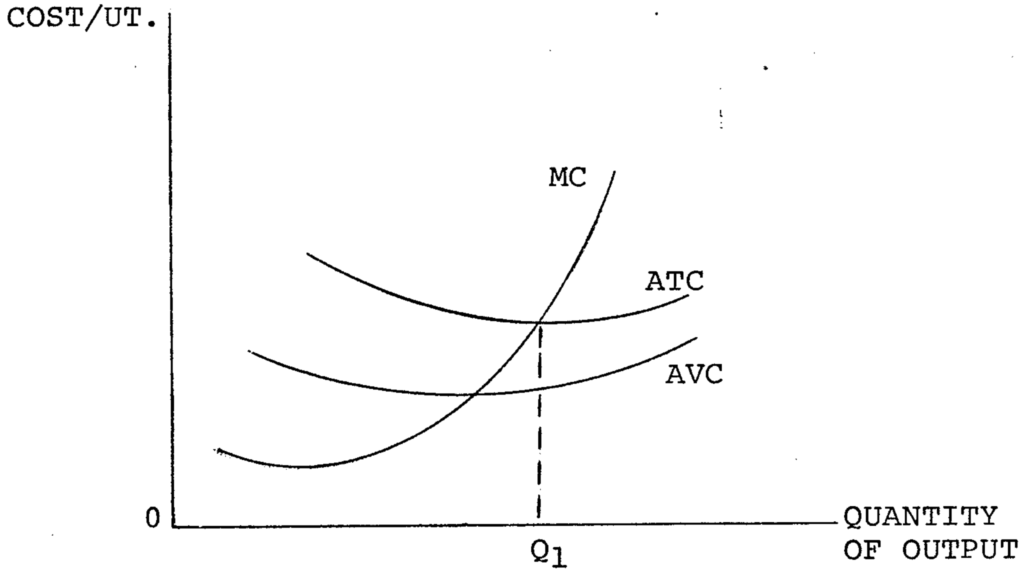

Figure 1-1

In the short run average cost would be minimized at Q_1, equal to capacity output, where marginal cost equals average cost.

[1] See Ferguson, C.E. Microeconomic theory. Homewood, Irwin Co., 1969. p. 187-198.

For each level of capital stock, which will be taken to be synonymous with size of plant, there results a different SAC curve. In the long run the entrepreneur will choose that plant size which minimizes the cost of production for his expected level of output. Moreover, cost minimization will require that unlike the short run the entrepreneur consider how factor prices change with variation in the scale of output, where there exists imperfect competition.

In figure 1-2 consider the case where the entrepreneur has only three plant sizes designated SAC^1, SAC^2, and SAC^3 from which to choose. In the real world the assumption that the entrepreneur has only a limited range of plant sizes or size of machine from which to choose, may be quite realistic due to the indivisibility of factor inputs. The long run average cost curve would then consist of those portions of the short run curves labelled AB, BC, CD which give minimum unit costs for each level of output.

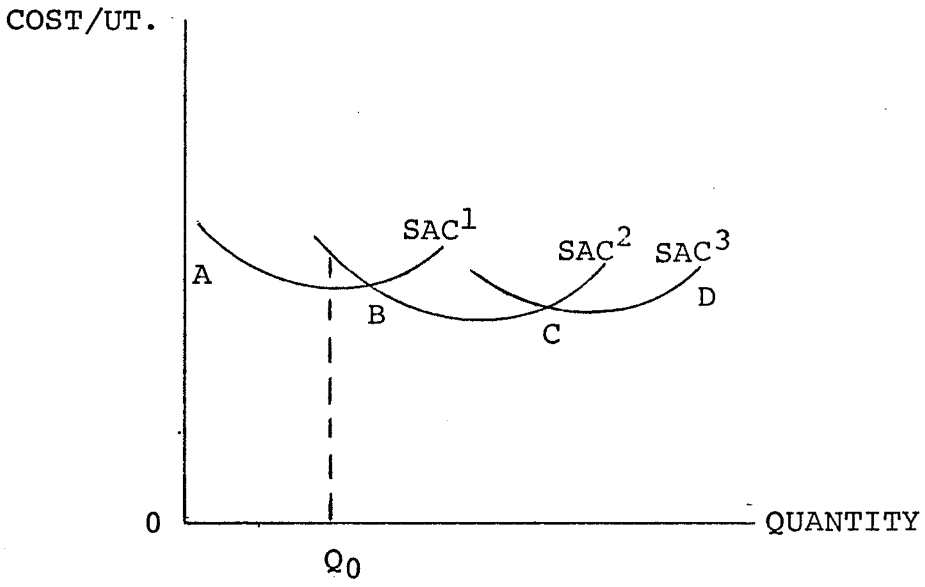

Figure 1-2

If the number of plant possibilities are expanded so as to become a continuous variable, the contribution of each plant segment to the LAC curve is miniscule. Hence, in figure 1-3 the LAC curve assumes a smooth U-shape.

In Viner's pioneering article in cost theory "Cost Curves and Supply Curves", it is stated that for all points on the long run average cost curve each plant size must be operated to capacity.[2] Assume long run average cost is constant with respect to output as in figure 1-4.

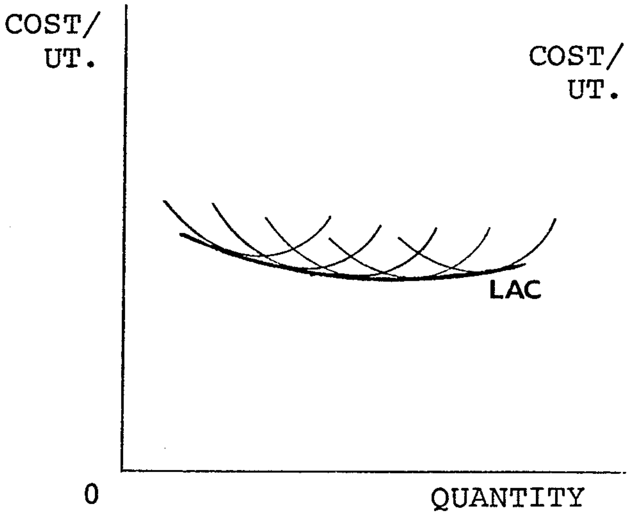

Figure 1-3

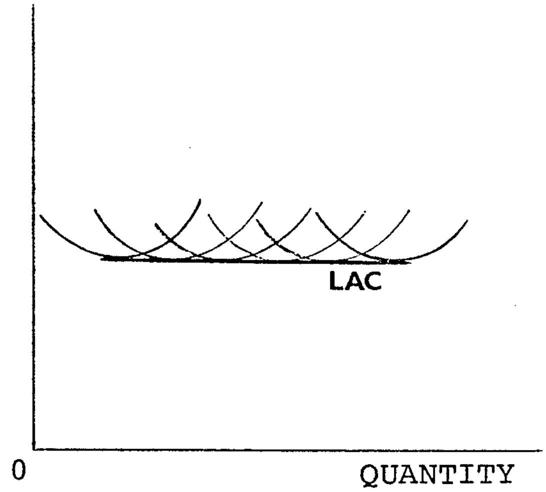

Figure 1-4

Given perfect divisibility of inputs, and thus an infinite series of plant possibilities, each point on the LAC curve corresponds exactly to minimum short run average cost.

[2] Viner, Jacob. In Readings in price theory. Edited by G.J. Stigler and K.E. Boulding. Homewood, Irwin Co., 1952. p. 198-232.

Where long run average costs are U-shaped, however, for levels of output less than that associated with minimum long run average costs there would be under capacity utilization. In figure 1-2 producing a quantity of output equal to Q_0 makes it more profitable to operate the larger plant SAC^2 at less than capacity rather than produce at capacity output with size of plant SAC^1. Similarly, it can be shown that above minimum long run average costs there would be over capacity utilization.

The long run average cost curve and short run average cost curves can be alternatively derived through isoquant analysis given information on substitution possibilities among the inputs and relative factor prices. Assuming perfectly divisible inputs for all points on the LAC curve it will be seen that the least cost input combination satisfies the condition whereby the marginal productivities of the inputs equal the ratio of the marginal expenses per unit of input in perfect competition.

In describing the substitution possibilities among the inputs the isoquant may be used. An isoquant or equal product curve shows those combinations of factor inputs, which yield a constant level of output. Assume there are only two homogeneous and perfectly divisible inputs, capital (K) and labour (L), used in the production of (Q) units of output

$$Q = f \quad (K,L)$$

If the marginal products of the inputs are positive, that is

$$\frac{\partial f}{\partial K} > 0 \; ; \quad \frac{\partial f}{\partial L} > 0 \; , \qquad (1\text{-}1)$$

for a given increase in capital there must be a corresponding reduction in the quantity of labour if output is to remain constant.[3] This may be expressed as

$$\frac{\partial f}{\partial K} \, dK = \frac{-\partial f}{\partial L} \, dL \qquad (1\text{-}2)$$

for all points along the isoquant.

The marginal rate of technical substitution (MRTS) is the precise term for the rate at which inputs can be substituted when output remains unchanged. This rate of trade off at a point is given by the absolute value of the slope of the isoquant at that point.[4] Thus from equation (1-2) it follows that the marginal rate of technical substitution of capital for labour equals the ratio of the

[3] Marginal productivity may also be negative. However, the rational producer will not operate where an increase in one input necessitates a further increase in other inputs just to maintain output constant.

[4] The total differential of the production function is

$$dQ = \frac{\partial f}{\partial K} \, dK + \frac{\partial f}{\partial L} \, dL$$

Since output is constant along an isoquant dQ = 0, and substituting there results

$$\frac{\partial f}{\partial K} \, dK + \frac{\partial f}{\partial L} \, dL = 0$$

The marginal rate of technical substitution is defined as dK/dL, hence

$$MRTS_{K \text{ for } L} = \frac{dK}{dL} = \frac{\frac{\partial f}{\partial L}}{\frac{\partial f}{\partial K}} \, .$$

marginal productivity of labour relative to the marginal productivity of capital

$$\text{MRTS} = \frac{dK}{dL} = \frac{\frac{\partial f}{\partial L}}{\frac{\partial f}{\partial K}} . \qquad (1-3)$$

In figure 1-5 an isoquant curve is drawn convex to the origin. For combinations of capital and labour K_1L_1, K_2L_2, and K_3L_3 quantity of output (Q) remains constant at 100 units.

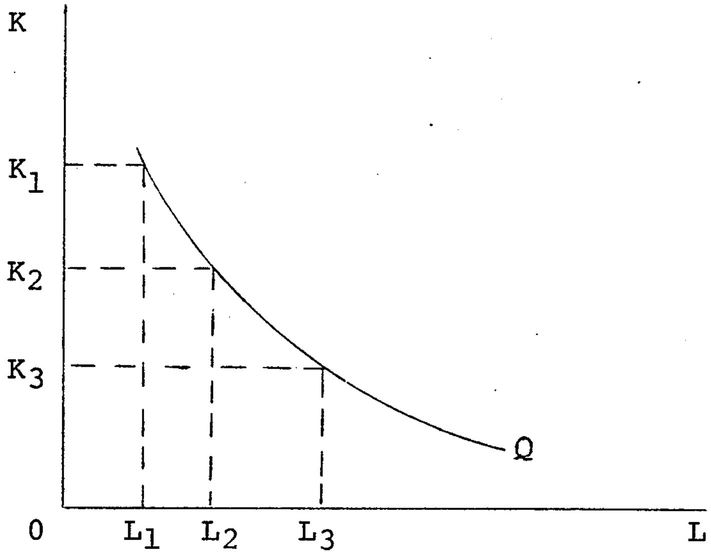

Figure 1-5

The convex shape of the isoquant indicates the existence of some degree although imperfect substitutability.[5] For given

[5] Alternative assumptions might also be made concerning the degree of input substitutability. Production subject to fixed proportions would indicate an increase in the quantity of one input alone adds nothing to output. Perfect substitutability between inputs suggests an increase in the quantity of one input will leave the marginal productivity of that input unchanged. Graphically fixed proportions and perfect substitutability imply the isoquants become respectively right angled or downward sloping straight lines.

increments of capital less and less labour can be traded off with output held constant. In figure 1-5 equal increases in the quantity of capital, K_1K_2 and K_2K_3, would require reductions in the quantity of labour inputs by the amounts L_1L_2 and L_2L_3 respectively, where $L_1L_2 < L_2L_3$. Thus as the quantity of labour used falls, its' marginal product rises, while the increase in quantity of capital causes the marginal product of capital to fall.

The nature of returns to scale can be determined by completing the isoquant mapping for all levels of output. Since inputs are assumed to be perfectly divisible from equation (1-1) it follows that isoquants become everywhere dense. However, from equation (1-2) it can be shown that isoquants never cross. The implications of intersecting isoquants may be examined with respect to figure 1-6. Isoquants Q_1 and Q_2 are shown to intersect at position B.

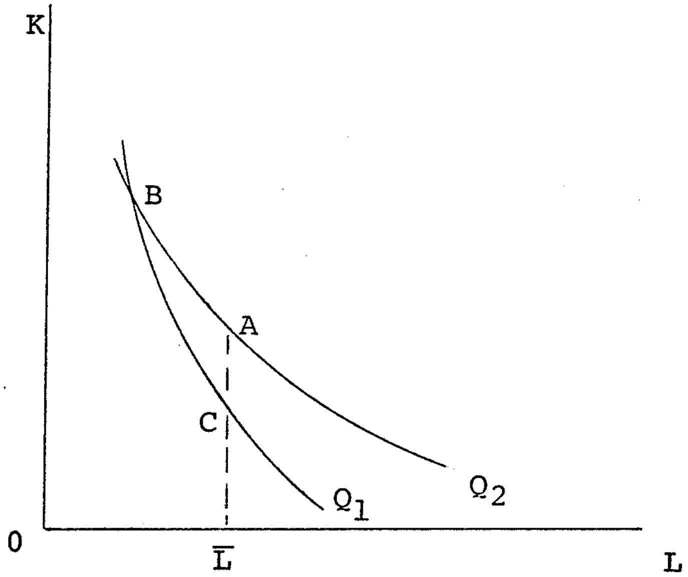

Figure 1-6

Since B yields the same output as A and B yields the same output as C, then it follows that A should give the same output as C. However, at position A more capital is being used with the same amount of labour which suggests that the marginal productivity of capital must be zero. That the marginal productivity of capital is non-zero is indicated by the curvature of the isoquants; implying a logical inconsistency.

If there exists increasing returns to scale the increase in inputs required for equal increments to output diminishes with higher levels of output. In figure 1-7 isoquants designated Q_1, Q_2, ..., Q_5 are drawn, representing successive increases in output of 100 units.

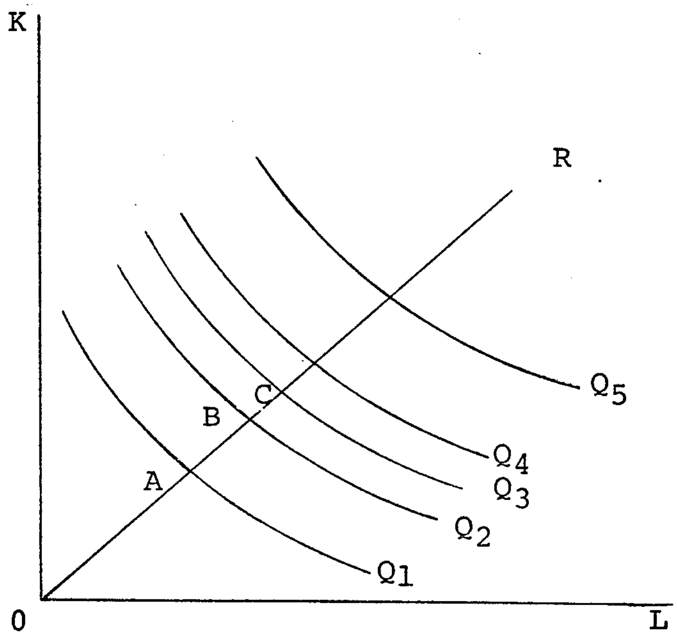

Figure 1-7

Points for which the marginal rate of technical substitution are constant are shown by the scale line OR. The existence of increasing returns implies that along OR isoquants become increasingly close together. Thus, as production is increased from Q_2 to Q_3 the distance AB > BC. In the case of constant returns successive increases in output by the same amount necessitate equal increases in inputs. This indicates the distance between the isoquants remains constant which would result if output was increased from Q_2 to Q_4. Finally, the existence of decreasing returns would imply that the isoquants become increasingly farther apart as over the range of output from Q_3 to Q_5.

It has been argued by Hahn that if the marginal rate of technical substitution remains unchanged there must exist constant returns to scale.[6] Thus, Hahn states that if two divisible inputs are combined in a proportion a/b and both multiplied by some factor K ($k>1$) then

$$a/b = c/d$$

where $c = ka$ and $d = kb$. It is concluded there must be constant returns to scale, since the rate at which c can be substituted for d is the same as that at which a can be

[6] Hahn, F.H. Proportionality, divisibility, and economies of scale: two comments. Quarterly Journal of Economics. 62: 132-133. 1948.

substituted for b, where c and d are equiproportionate amounts of a and b.[7]

The Hahn thesis may be examined with reference to the Cobb-Douglas form of the production function which has been used extensively in empirical research. The Cobb-Douglas function may be expressed as

$$X = AL^{a}K^{b} \qquad \begin{array}{lll} X>0 & , a\geq 0 & , \\ K>0 & , b\geq 0 & , \\ L>0 & , A\geq 0 & . \end{array} \qquad (1\text{-}4)$$

where X equals output, and A, a, and b are parameters.[8]

Equation (1-4) should be interpreted as follows:

1. The parameters a and b are the elasticities of production with respect to labour and capital respectively.
2. The function is homogeneous of degree a + b implying increasing, constant, or decreasing returns to scale depending upon whether the sum of a + b exceeds, equals, or is less than unity.
3. The marginal product of capital is

$$\frac{\partial X}{\partial K} = \frac{bAL^{a}K^{b}}{K} = \frac{bX}{aK} \qquad (1\text{-}5)$$

[7] Ibid. p. 133.

[8] Walters, A.A. Production and cost functions: an econometric survey. Econometrica. 31: 5-6. 1963.

which declines with greater capital inputs if b>1 since $\partial^2 X/\partial K^2 = b(b-1)X/K^2 > 0$. Similarly, the marginal product of labour is

$$\frac{\partial X}{\partial L} = \frac{aL^aK^b}{L} = \frac{aX}{L}\,. \qquad (1\text{-}6)$$

4. The marginal rate of technical substitution of labour for capital from equations (1-5) and (1-6) is

$$\text{MRTS} = \frac{\partial X/\partial K}{\partial X/\partial L} = \frac{bL}{ak}\,.$$

For multiples of each input combination the sum of a + b is not restricted to unity, which would be necessary for constant returns.

To choose that combination of inputs which minimizes cost the entrepreneur must take the prices of the inputs into account as well as their productivities. The prices of inputs may be represented by iso-cost or equal cost lines, showing those combinations of inputs which may be purchased for a given expenditure.

Assume there exists only two inputs, capital and labour, whose prices remain constant for all levels of output. The isocost lines may then be described by the equation

$$TC = rK + wL$$

where TC equals total cost, r equals rent, and w equals the wage rate. In figure 1-8 the iso-cost line K^1L^1 is super-imposed on the isoquant mapping. For the given level of expenditure on inputs cost minimization is equivalent to maximizing output.

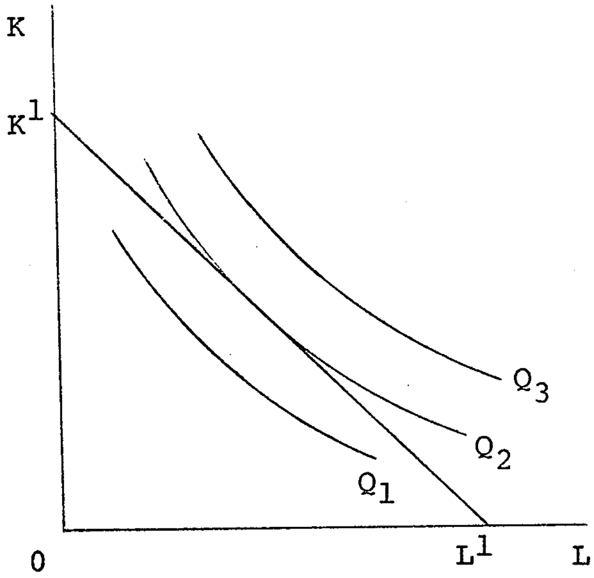

Figure 1-8

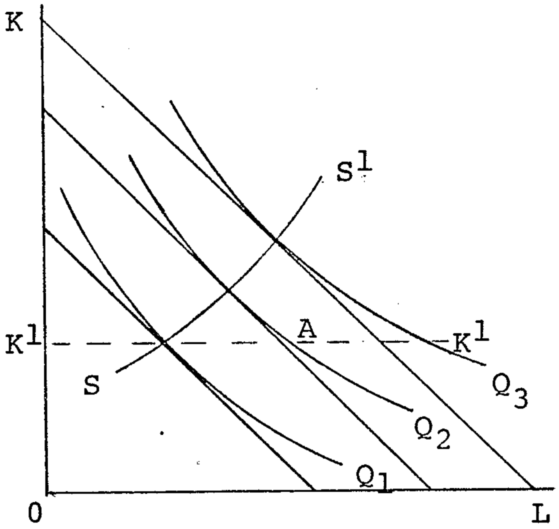

Figure 1-9

The highest level of output obtainable in figure 1-8 is Q_2, where isoquant Q_2 is just tangent to the iso-cost line.

At the tangency position, corresponding to minimum average cost for each level of output, it can be shown that the ratio of the input prices must equal the ratio of the marginal productivities of the inputs. In figure 1-8 the slope of the iso-cost curve can be defined as OK^1/OL^1. Since OK^1 equals $\frac{TC}{r}$ and OL^1 equals $\frac{TC}{w}$ it follows that the slope of the isocost curve is $\frac{TC}{r}/\frac{TC}{w}$ or w/r. However, at the point of tangency slopes of isoquant and isocost line are equal, so that from equation (1-3) one obtains

$$\text{MRTS} = \frac{dK}{dL} = \frac{MPP_L}{MPP_K} = \frac{w}{r} .$$

This condition may be alternatively interpreted as stating that for each dollar of expenditure on inputs cost minimization requires that the marginal productivities of the inputs be equal.[9]

For each amount that the producer has to spend on inputs there will result a different isocost curve. If input prices remain constant as expenditures increases, the isocost curves will shift parallel in a north east direction. Given isoquants which are everywhere dense there results a series of tangency solutions SS^1 in figure 1-9. From the curve SS^1 both the production and cost functions can be easily derived since for each level of output total cost is minimized and hence the amount of inputs given constant factor prices.

The difference between the short run and long run situations can also be examined in reference to figure 1-9. In the short run the capital stock would be fixed at K^1. Expansion would proceed not along SS^1 but along K^1K^1. Since the slope of the isoquants is less than the isocost curve at position A

$$\frac{MPP_L}{MPP_K} < \frac{w}{r} .$$

[9] Ferguson. op. cit. p. 165-169.

This implies that the substitution of capital for labour would increase output. However, restrictions on the size of the capital stock will cause the above inequality to persist.

Similarly, where there exists indivisibilities the ratio of the marginal productivities of the inputs would not equal the ratio of their relative prices. In figure 1-10 assume there exists only two plant sizes K^1 and K^2.

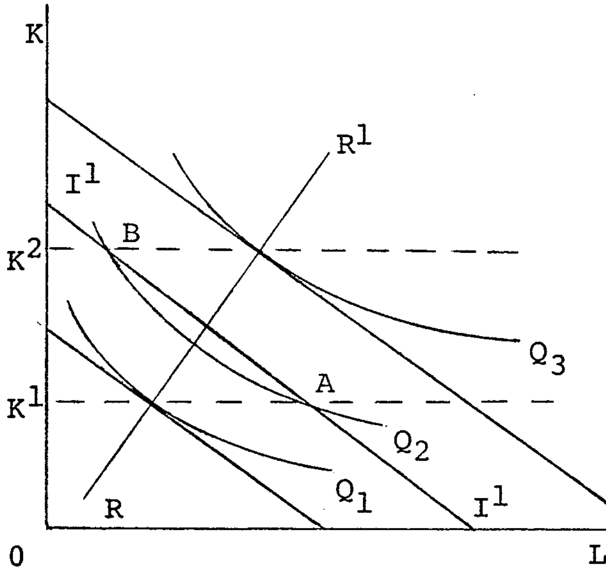

Figure 1-10

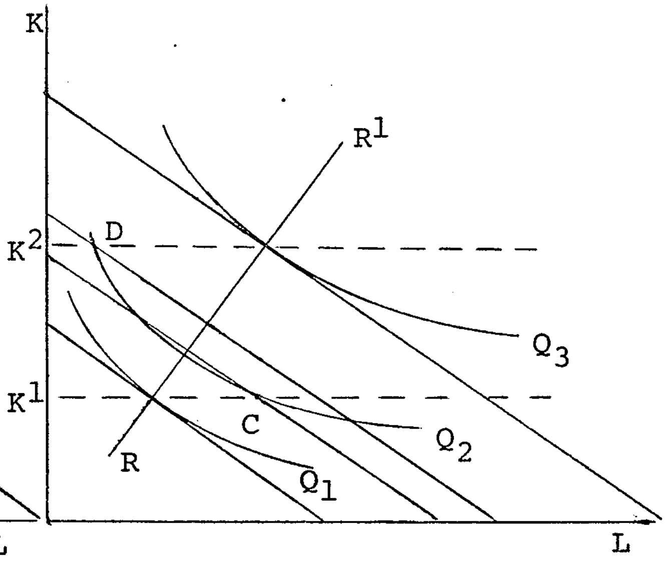

Figure 1-11

For levels of output between Q_1 and Q_3 the entrepreneur may expand with an excessively large usage of capital or an excessively large usage of labour relative to those input combinations shown by RR^1 giving maximum efficiency. For an increase in output from Q_1 to Q_2 the entrepreneur can choose to produce at either points A or B. The entrepreneur would be indifferent to the available input combinations. Thus total costs would be the same in either case since an isocost line can be drawn through both points A and B, as indicated by I^1I^1.

However, in figure 1-11 an increase in output from Q_1 less than Q_2 would be produced by using capital inputs K^1 and an excess amount of labour. For example, with an increase in output from Q_1 to Q_2 the entrepreneur could produce at positions C or D. Since the isocost line through position D lies above the isocost line through position C, the entrepreneur would choose position C. By similar logic, as output increased by an amount greater than Q_1Q_2 there would occur a transition to plant size K^2 and excess capital inputs.

In pure or perfect competition profit maximization requires that the long run average cost curve must rise. Moreover, the increase in unit costs must supplant any scale economies at low levels of output. The theory of the firm suggests that producers will expand output until the marginal revenue of an addditional unit equals its' marginal cost. In perfect or pure competition the existence of a large number of small producers implies that price is unaffected by the quantity of output supplied by each firm. Since price equals marginal revenue then price must equal marginal cost in equilibrium.

Assume the long run average cost curve falls. Declining average cost, and smaller and smaller increments to total cost, imply that marginal cost must be less than average cost. Producing where price equals marginal cost the firm would be suffering losses, obtaining a rate of return below that which could be obtained in other industries. Alternatively, assume long run average cost is constant with respect to output. When average cost is neither rising nor falling then marginal cost must equal average cost. Since both marginal cost and price are constant, given perfect knowledge and perfect resource mobility there results a multitude of solutions consistent with profit maximization. The producer could then expand to large size with no reductions in the level of profits. However, where long run average cost rises there would exist a brake on firm expansion. Where average costs are rising the increments to total cost becoming larger and larger, then marginal cost must be rising as well. Since price remains constant increases in output would then decrease profit.

Based on the following explanations for economies and diseconomies of scale, it is traditionally hypothesized that the long run average cost curve is U-shaped. First, indivisibilities may give rise to scale economies. If labour and equipment are available in only a limited range of sizes, and production takes place below capacity, output can be increased without additional outlays. Moreover, the existence of bottlenecks or the inability to properly synchronize equipment at low levels of output, may magnify scale economies due to indivisibilities. For example, suppose there are only two types of machines, one producing, and the other loading the product for shipment. If the first machine can produce 20,000 units per day, and the second machine could load 50,000 units per day, output would need to be at least 100,000 units for full capacity utilization of each machine.[10]

Second, specialization and division of labour may result in economies of scale. A larger plant employing a greater number of workers can enable each worker to specialize in one job. Adam Smith, in considering the manufacture of pins was the first to emphasise how through one man drawing the wire, another straightening it, a third cutting it, etc., each worker gains in efficiency, through repetition of the same task and eliminating time consuming interchanges of plant and equipment.[11]

Third, qualitative changes in inputs may generate scale economies as various forms of automation are introduced which were not profitable at smaller scales of output. One job illustrating such qualitative changes is ditchdigging which would be performed initially by simply adding men and shovels, but once a certain scale is reached it becomes profitable to employ a modern ditchdigging machine.[12]

[10] Ferguson. op. cit. p. 211.

[11] Robinson, E.A.G. The structure of competitive industry. Cambridge, Cambridge University Press, 1958. p. 13.

[12] Chamberlin, E.H. The theory of monopolistic competition. 7th ed., Cambridge, Harvard University Press, 1956. p. 235-236.

Fourth, economies may be of a stochastic nature. This results from the fact that the variance of sales fluctuations or the number of expected breakdowns in plant equipment expand less than proportionately to changes in scale. With respect to the former decreased variance of sales would lower costs as firms are able to adjust more fully to the level of present output.

Finally, geometric economies may occur in the utilization of equipment such as pipes and containers. The material required for their construction depends on surface area, whereas capacity depends on volume. For example, doubling the linear dimensions of a storage tank would increase surface area four times original size but expand capacity eight times over.

Difficulties in maintaining control and co-ordinating the operations of various departments within the firm has been the traditional explanation given for the existence of diseconomies of scale. With increasing size management has to delegate authority to lower echelon employees. This increases the number of hierarchical levels over which information and instructions must pass thereby decreasing the quality of communication. Thus, mistakes are not only less easily discovered but when revealed it is unclear where responsibility lies.

The importance of increasing complexity of the managerial function has been a controversial issue. It may be argued that if a hierarchical form of organization was replaced by decentralization and the appointment of managers with equal powers management diseconomies would be avoided. However, while within any plant the orders of the manager may be efficiently carried out, poorer quality of communication within the management sector itself may result in the wrong orders being given. Thus Chamberlin has observed a residual of authority must remain in central hands.[13] Conditions encountered in independent units may not be entirely reproducible since the firm as a control unit can not divest itself completely of control over its' component parts. Nonetheless, one may well question whether the effects of such factors as specialization and division of labour, qualitative input changes, and indivisibilities do not offset the tendency for higher costs per unit resulting from difficulties of management.

A further criticism of the hypothesized U-shaped cost-output relationship points to a conflict between the explanations for economies and diseconomies of scale and the assumptions of the theory of long run average cost. It has been argued that the major source of economies and diseconomies of scale is indivisibility of inputs.[14] Accordingly, if inputs are defined to be homogeneous and perfectly divisible the theory of long run average cost strictly interpreted can only relate to the case of constant returns.

[13] Chamberlin. op. cit. p. 248.

[14] Hahn. op. cit. p. 133-135.

In addition to indivisibilities of plant and equipment it has been suggested that specialization and division of labour are a type of indivisibility. Hahn has observed that with perfect divisibility it is solely a matter of subdividing any single productive process into a large number of stages and this by definition is possible irrespective of the absolute amount of factors employed.[15] With respect to qualitative changes in inputs the only possible explanation for the greater range of technical possibilities with increases in scale must be in the indivisibility of these "technical possibilities".[16] While independent of the divisibility or indivisibility of inputs stochastic economies are also inconsistent with the static theory of long run average cost. Thus, the long run average cost concept refers to a single point in time and does not properly relate to arguments concerning the duration for which demand conditions are expected to prevail and the frequency over time of repairs. Nonetheless, consideration of geometric economies as well as diseconomies resulting from greater complexity of the managerial function indicate that the long run average cost curve is not necessarily restricted to the case of constant returns.

A final criticism is that perfect competition constitutes an extreme and unlikely description of market

[15] Ibid. p. 133-135.

[16] Ibid. p. 134.

structure. Obviously, perfect knowledge and perfect resource mobility do not exist in the real world. Moreover, firms seldom produce a single homogeneous product. Imperfectly competitive elements in factor and product markets have important implications both for the method of deriving and shape of the long run average cost function.

Economies and diseconomies of scale may be either technological or pecuniary in nature consisting of decreasing technical coefficients of production or changes in the price paid for the factors.[17] Variation of input prices with the quantities purchased results where the firm is a large purchaser of inputs relative to total market demand and the factor supply curve rises or falls. Nonetheless, depending upon the context and the problem being investigated constancy or non-constancy of factor prices may be more appropriate. Since factor price changes may be caused by monopsonistic exploitation of factor inputs, if one is investigating the optimal level of output from the point of view of welfare economies, economies with respect to factor price changes may be suitably excluded. On the other hand, if one is concerned with assessing, for example, barriers to entry - a problem of what is; rather than what should be - all economies including those attributable to changing input prices should be considered.

The existence of pecuniary economies and diseconomies will require the entrepreneur to consider what is the marginal

[17] Viner. op. cit. p. 210

expense of inputs. For example, if the firm hires additional labour causing the wage rate to rise, expenses will increase by more than the wage bill of the additional labourors, because all workers employed now receive the new higher price for their services. The cost minimization condition in the case of two inputs capital and labour then becomes

$$\frac{MPP_K}{MEI_K} = \frac{MPP_L}{MEI_L}$$

where MEI_K is the marginal expense of capital and MEI_L is the marginal expense of labour.[18] In the above example the increase in employment would be smaller than under competitive conditions. Graphically, the isocost lines would no longer be straight lines but become concave to the origin.

Difficulties may be encountered in the analysis where factor prices fall with increases in the quantity of inputs. Reductions in input prices with increasing size could occur where suppliers grant quantity discounts due to the fear of losing large contracts. In addition, economies may be obtained in the financing of investment since generally as the firm grows and becomes more well known, it needs to offer a smaller yield on bonds and stocks to borrow from the public. In figure 1-12 assume the price of labour remains constant but that with increases in the

[18] For a mathematical proof see Ferguson. op. cit. p. 406-408.

quantity of capital the price of capital falls. The iso-cost lines indicated by the heavily shaded curvedlines II^1, II^2, and II^3 are then drawn convex to the origin.

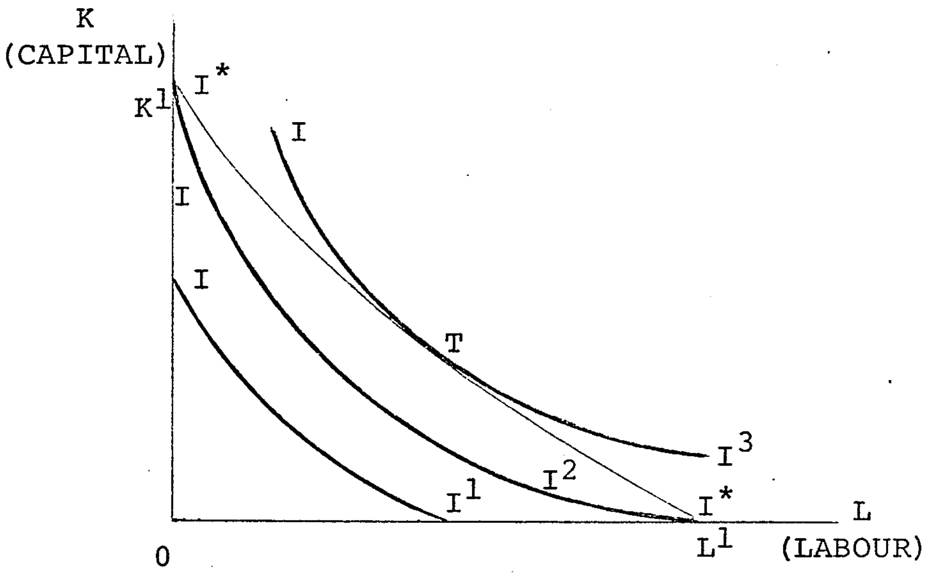

Figure 1-12

Assume the entrepreneur wishes to minimize cost for a level of output indicated by the isoquant labelled I^*I^*. The position of tangency between isoquant and iso-cost curves is shown by position T. However, position T lies on the iso-cost curve II^3 which is farthest to the right. The optimal input combinations are "corner solutions" such as K^1 or L^1 on the iso-cost curve II^2. Thus those input combinations satisfying the condition that the ratio of the marginal expenses of the inputs equal the marginal productivity of the inputs do not minimize costs. However, this would only occur where the degree of convexity of the iso-cost curves is greater than that of the isoquants. In figure 1-12 if the isocost curves were now to become isoquants and vice-versa, for an expenditure shown by I^*I^* the tangency position T would represent maximum output.

With imperfect competition in the product market there exist large firms which may compete on a price and/or non-price basis. Large changes relative to total market supply in the firms' output then violate the assumption that price remains constant. Thus, Sraffa argued in the Laws of Returns under Competitive Conditions that the firm faces individual diminishing cost but that the entrepreneur would be prevented from continually expanding his business due to reductions in the price of his product.[19]

In the traditional theory of long run average cost selling outlays are totally excluded. The distinction between selling costs and costs of production and distribution has been stated by Chamberlin to be one of whether inputs are being used to change consumer wants or to make available a product to satisfy given wants.[20] In a market structure of pūre or perfect competition it follows from the assumptions of product homogenity and the existence of a large number of small sized firms that selling costs will be nil. However, in the case of oligopoly the influence of the firms advertising on its' own demand will not be negligible, while, in a situation of monopolistic competition the disincentive to advertise due to large numbers may be offset by the unique characteristics of each product,

[19] Sraffa, Piero. In readings in price theory. Edited by G.J. Stigler and K.E. Boulding. Homewood, Irwin Co., 1952. p. 180-197.

[20] Chamberlin. op. cit. p. 117.

which can be magnified by advertising so as to better direct consumer response.

That greater emphasis should be given to selling costs may be justified by the following arguments supporting the existence of economies and diseconomies of scale. First, reductions in sales promotion costs with increases in firm size may result, as it becomes feasible to utilize the more efficient national advertising media rather than local advertising.[21] Second, to the extent that increases in firm size are associated with a greater number of product lines, scale economies may result from joint advertising of a series of related products being more efficient than advertising individual products.[22] Both changes to different media and advertising of multiple products are equivalent to the argument considered above with respect to production and distribution: that large size brings with it a qualitative as well as quantitative change in input requirements. Third, there is no reason a priori to expect the gains in efficiency from specialization and division of labour to be any less in the area of selling costs, than those encountered in production and distribution. Fourth, advertising may have a cumulative impact, increases

[21] Ibid. p. 134.

[22] Stigler, G. The economies of scale. Journal of Law and Economics. 1:54. 1958.

in sales being small until consumer resistance is finally broken down.[23]

While some doubt exists with respect to the eventual increase in unit costs in production and distribution, the case for diseconomies of scale becomes much stronger in the area of selling costs. In addition to rising input prices, and increased difficulties of management and co-ordination, with increases in sales volume there develops increased consumer resistance. Such increasing resistance will result due to the fact that buyers are not equally accessible; some possessing more direct needs for the product or service than others. In addition, amongst the same group of consumers as quantity sold increases advertising must induce the sacrifice of continually more important alternative needs.

The exclusion of selling costs might be defended on three grounds. First, advertising may conflict with welfare considerations through its' being non-informative in content, and merely diverting sales from one product group to another. Nonetheless, close examination of the nature of advertising inputs will be necessary before conclusions as to the effects on social welfare can be established. It may be that advertising satisfies a general desire on the part of consumers for variety, although no new information on the qualities of the product is being imparted. Moreover, as

[23] Chamberlin. op. cit. p. 133. It should be noted that the cumulative impact of advertising may be a function not only of scale but of time.

was pointed out in the case of factor price changes, if one is investigating barriers to entry, for example, all economies and diseconomies of scale are relevant.

Second, the shape and position of the curve of selling costs will vary depending on the play of other variables. Especially important is price which must be held constant if the effects of selling costs on sales are to be determined. Graphically, this implies that although advertising increases the demand at all prices, a single price must be chosen, and the increase in demand at that price examined.[24] In figure 1-13 the cost-output relationships for production and selling are combined to show the overall effects of size on costs per unit.[25] The independent variable now becomes quantity of output produced and sold. It is

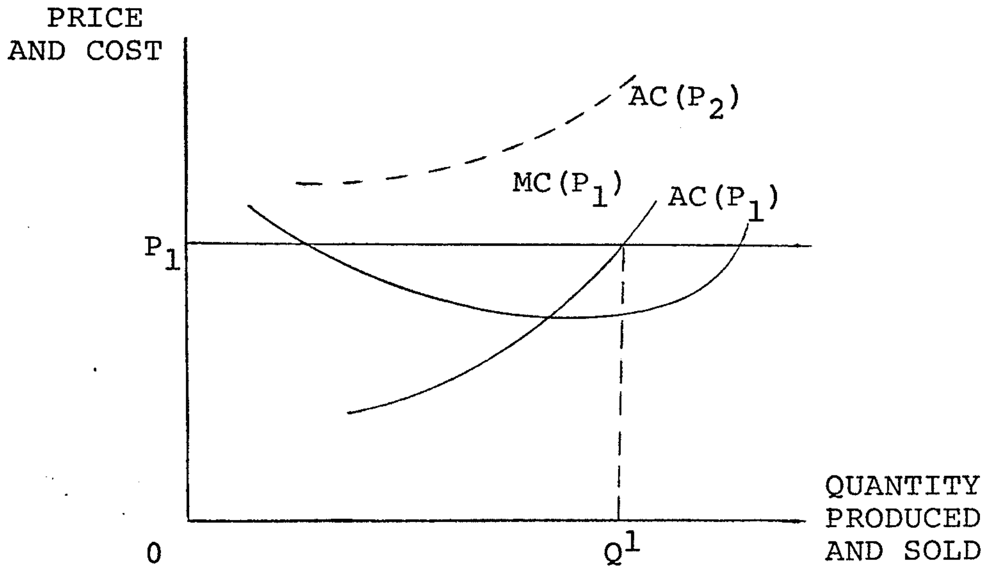

Figure 1-13

[24] Chamberlin. op. cit. p. 130

[25] Ibid. p. 142.

assumed the price is set by custom at OP_1. If price is constant at OP_1 the average and marginal costs of producing and selling varying quantities of output are $AC(P_1)$ and $MC(P_1)$ respectively. The profit maximizing firm would produce Q_1 units of output, where the last unit is just worth the cost of producing and selling it.

The question now arises whether advertising has the same proportionate effect at all prices. Chamberlin, has argued that while the rate of increase or decrease in selling cost per unit and the point at which decreasing returns set in may vary for different prices, the same general stages will be gone through.[26] Thus the contention is that if at one price the curve of selling and production costs is U-shaped, at another price again a U-shaped relation will prevail. However, if prices increased substantially it may be economies of scale obtained when price was lower would be completely eliminated due to increased consumer resistance. Thus, in figure 1-13 if price rose from P_1 to P_2 the average cost of producing and selling may be as described by the curve AC^2. How consumer resistance to advertising responds to price changes cannot be determined a priori.

[26] Ibid. p. 131.

A third difficulty is that the relationship of selling costs to sales volume may be non-reversible. If a firm increases advertising expenditure, and then cuts expenditure by the same amount, sales volume will not return to its' original level.[27] Thus, large increases in advertising expenditures may be initially required to attract the attention of consumers. However, thereafter advertising can be reduced to original levels since appeals designed to serve only as reminders may be all that is necessary to sustain the higher level of sales volume.

In conclusion, severe difficulties appear to exist in describing the cost-output relationship observed in practice. Part of the difficulty arises from the interpretation of the theory of long run average cost. Indivisibilities, factor price changes, and dynamic considerations may all be inconsistent with interpreting the long run average cost concept as requiring that the ratio of the marginal productivities of the inputs equals the ratio of their prices or marginal expenses. However, this prediction was obtained under one set of assumptions and through the same process of deduction the cost curves can be alternatively derived under different assumptions, which are more applicable to the real world. Thus, the assumption

[27] It should be noted that certain economies or diseconomies in production and distribution may also not be fully reversible. For example, in imperfectly competitive labour markets an expansion in demand for labour may generate pecuniary diseconomies but when demand falls back workers may refuse to accept reductions in wages.

of perfectly divisible inputs,[28] and constant factor prices may be suitably relaxed.[29]

Despite incorporation of indivisibilities and factor price changes certain sources of economies or diseconomies of scale will be excluded. In the area of selling costs the need to examine whether the cost-output relationship holds for different prices as well as increases or decreases in sales volume cannot be considered simply differences in interpretation. Moreover, stochastic economies pose more fundamental difficulties for the analysis since the theory of long run average cost is inherently static in nature. Nonetheless, it should be emphasised that the conclusions not the assumptions of theory are tested against reality. While the theory of long run average cost may exclude certain sources of economies or diseconomies of scale, the question arises whether the results are significantly altered.

[28] See below pp. 14 - 16.

[29] See below pp. 22 - 24.

Chapter II

The Empirical Investigation of Economies and Diseconomies of Scale

The existing methods of investigating economies and diseconomies of scale include statistical production and cost analysis, the questionnaire and interview method, engineering estimates, and the survivor technique. Statistical production and cost analysis utilize the familar cross section and time series approaches, which like the questionnaire and interview method have been used extensively in other empirical research. Engineering estimates are obtained from a building up of the relationship between inputs and output from individual pieces of equipment or process areas. Finally, the survivor technique attempts to determine efficiency from changes in the firms' market share over time.

Of the various cost estimation techniques a general classification can be made into ex post and ex ante studies. Statistical production and cost analysis, as well as the survivor method are ex post attempts to determine the nature of returns to scale while engineering estimates are ex ante in approach. Where the questionnaire and interview method have been used businessmen were asked to predict how costs would respond to changes in output. Since causality was implied the existing questionnaire and interview studies may also be classed as ex ante. The distinction between ex post and ex ante studies has relevance for the type of problems encountered in empirical analysis. Studies of an ex ante nature have less difficulties with respect to the problems introduced by product differentiation, external influences, and the determination of causality in the size-cost or input-output relationship. However, avoidance of such difficulties may have resulted in exclusion of relevant economic variables or reliance upon subjective and inaccurate information.

Statistical Production and Cost Analysis

In measuring economies and diseconomies of scale of the firm cross section analysis involves a comparison of cost-output or input-output data for different firms within the same industry at a particular point in time. Alternatively, time series analysis might be undertaken for the same firm over time. The following problems are encountered in both time series and cross section analysis.

First, the firm seldom produces a single homogeneous product. Where the firm produces a number of different products in varying proportions two basic approaches have been used. Either output is treated as multi-dimensional and attempts made to allocate costs among individual products, or a composite measure of output is constructed.

A second difficulty occurs in attempting to eliminate the influence of numerous external factors. Where the effects of such external factors vary with the size of firm, determining how costs respond to changes in size alone becomes especially difficult. In this context the controversial issue of whether large firms undertake greater research and development should be considered. Key factors in examining the relationship between technological innovation and firm size are competitive pressures, size of investment funds, and divisibility of research and development inputs. Further, differences in the quality of factor inputs particularly managerial efficiency may vary with size. With increases in size it may be argued a reduction in the quality of managerial inputs will occur due to a divorce of ownership and management. On the other hand, formal training and recruitment programs instituted by large firms may increase the quality of managerial ability. Whether such forces would tend to cancel out making the disturbance truly random is then subject to question.

An additional complication is that it may be difficult to distinguish which factors are external. This may occur, for example, in the case of qualitative changes in inputs which may be the result of size or the firms' location in a particular geographic locale. Similarly, in investigating the cost function adjusting costs for different price levels at the time of purchase may also eliminate pecuniary economies and diseconomies of scale.

A third problem is that the series of observations will reveal firms in various stages of disequilibrium. Due to imperfections in mobility substantial time will be required for the placing of orders, and for the production, delivery, and installation of equipment. The observations will then be influenced by short run factors. Moreover, while firms may have completely adjusted to the planned level of output, errors in forecasting sales and a resultant divergence between planned and actual output may cause firms to be not only out of long run equilibrium but short run equilibrium as well. Consequently, the fitting of a production or cost function to unprocessed data will clearly yield biased estimates of the level of the curves. Finally, it has been argued that large firms are more likely to be in long run equilibrium causing the slope to be also affected.

Fourth, consideration must be given to the range of output. While for the observed range of output no indication of rising long run average cost exists, it can not be conclusively established that the long run average cost curve is not U-shaped. The possibility exists that study is being made of only a portion of the long run average cost curve. Also, the converse porposition applies. Where there exists indivisibilities a U-shaped relation for the observed range of output may be atypic for the long run average cost curve as a whole.

Fifth, the relationship among the variables may be one of multi-lateral rather than uni-lateral causation. The production function and cost function are parts of a simultaneous system of equations. Not only does output determine costs, but the theory of profit maximization suggests output is influenced by cost. Similarly, for the production function the quantity of inputs determines the quantity of output, but the quantities of various inputs are influenced by the level of output through the marginal productivity conditions.

The Statistical Production Function

Considerable study has been made of the aggregate production function applicable to an industry or for a sector of the national economy. Such research has attempted to determine the role of technology and economies of scale in promoting growth, as well as the constancy of labours' share of output. However, relatively few studies have been made of the production function of the firm although the theory of production is strictly applicable only at the micro-level. Moreover, in investigating the nature of returns to scale pecuniary economies and pecuniary diseconomies would be excluded. Hence, only in perfect competition could the results of statistical studies of the production function be interpreted as an overall measure of economies and diseconomies of scale.

Regression analysis may be used to determine the nature of returns to scale by the fitting of an appropriate equation to input-output data. The Cobb-Douglas function is frequently used as the general form of such equations. This may be expressed in stochastic form as

$$X_{it} = AL_{it}{}^{a}K_{it}{}^{b}V_{it} \qquad (2\text{-}1)$$

where V_{it} is a random disturbance and the subscripts $i = 1, 2, \cdots, n$ indicate the number of firms, with

$t = 1, 2, \ldots, n$ referring to the number of time periods.[1] Equation (2-1) can be expressed as

$$\bar{X} = \bar{A} + a\bar{L} + b\bar{K} + \bar{V} \qquad (2\text{-}2)$$

where the superscript - indicates natural logarithms and the subscripts have been dropped for convenience of notation.

The parameters $\bar{A}$, a, and b could then be estimated from the principle of least squares. The least squares method requires that the sum of squared deviations of observed from expected values be minimized. Assume the estimated values from the sample regression function are given by

$$\hat{\bar{X}} = \hat{\bar{A}} + \hat{a}\bar{L} + \hat{b}\bar{K} \qquad (2\text{-}3)$$

where the superscript ^ designates estimated values. Subtracting (2-3) from the values of the underlying population in (2-2) it is necessary that

$$\Sigma(\bar{X} - \hat{\bar{A}} - \hat{a}\bar{L} - \hat{b}\bar{K})^2 \qquad (2\text{-}4)$$

[1] The equations in cross section and time series analysis are respectively

$$X_{i1} = AL_{i1}{}^{a}K_{i1}{}^{b}V_{i1}$$

and

$$X_{1t} = AL_{1t}{}^{a}K_{1t}{}^{b}V_{1t} \quad .$$

be minimized for all observations 1, 2, $\cdots$, n. Through analysis of variance tests it could then be determined whether the parameter estimates were significant of correlation. Finally, through comparison of the size of the residual variance from equation (2-4) with the total variance the correlation coefficient (r^2) could be computed to determine the explanatory power of each of the independent variables.

The ability of the least squares principle to yield best linear unbiased estimates and the applicability of analysis of variance tests depends upon the following assumptions. First, the relevant variables must be observed without error. Second, the disturbance is assumed to be a random normal variable with zero mean for all t. Third, the explanatory variables must be independent of the disturbance. Fourth, the disturbances are homoscedastic or constant at any given time t. Fifth, the disturbance in period t is independent of the disturbances that emerged in period t-1, t-2, etc. Sixth, there must not exist multicollinearity or dependence among the explanatory variables.

An analysis or survey of the results of sampling experiments attempting to assess the errors resulting from the failure of the above assumptions is beyond the scope of this paper. However, it can be shown that the regression coefficients will be biased if there exists either measurement errors, absence of normality, or correlation between the explanatory variables and the disturbance.[2] The existence of heteroscedasticity, serial correlation among the disturbances, or multicollinearity among the independent variables will vitiate the minimum variance property of the estimators.[3] Finally, unless the disturbances are normally distributed with a mean of zero and constant variance, standard F and t-tests will no longer be strictly valid.

A test on the homoscedasticity of the disturbances could be performed by marking off arbitrary intervals on the input axes and calculating the variance about the regression surface within each interval. In addition, a test for serial correlation among the disturbances is provided by the Durbin-Watson d statistic.[4] Theoretically, certain

[2] See Johnston, J. Statistical cost analysis. New York, McGraw-Hill, 1960. p. 31-43.

[3] Ibid. p. 31-43.

[4] Let Z_t (t = 1, ..., n) denote the residuals from a fitted least squares regression. The d-statistic is calculated as

$$d = \sum_{t=1}^{m} \frac{(Z_t - Z_{t-1})^2}{Z_t{}^2}$$

Comparison is then made of the theoretical d values associated with random disturbances, and those calculated.

transformations of the data could then be made to randomize the disturbance. However, in practice little information is generally available on the form of the heteroscedasticity or serial correlation to suggest the appropriate transformation.

In addition to the Cobb-Douglas equation other forms of the production function may be desirable due to the existence of specification error. One disadvantage of the Cobb-Douglas production function is the inability to incorporate various degrees of input substitutability. In lieu of the marginal rate of substitution such substitutability will be more appropriately described by the elasticity of substitution concept; the latter being independent of the units of measurement. The elasticity of substitution may be defined as the proportionate change in the capital/labour ratio induced by a given proportional change in the factor price ratio. For the Cobb-Douglas function it can be shown that the elasticity of substitution is equal to unity.[5] However, many processes may have very low elasticities of substitution, or zero elasticity of substitution as in the case of fixed proportions. Thus, errors may result from attempting to force the data into a mould that stipulates unitary elasticity.

[5] See Ferguson, C.E. The neo-classical theory of production and distribution. London, Cambridge University Press, 1969. Ch. 5.

A more general specification of the production function is the constant elasticity of substitution (C.E.S.) equation. This may be expressed as

$$X = c\left[aK^{-p} + (1 - a)\ L^{-p}\right]^{-v/p}$$

or in logarithmic form

$$\log\ X = \log c\ - {}^{v}/p\ \log\left[aK^{-p} + (1 - a)L^{-p}\right].$$

The degree of homogeneity or the nature of returns to scale is indicated by V. The efficiency parameter is c determining the size of output for given quantities of inputs. Relative intensity of capital and labour for each level of output is shown by the parameter $a (0 \leq a \leq 1)$. Finally, it can be shown that the parameter p is obtained by

$$p = \frac{1}{b} - 1 \qquad (2\text{-}5)$$

where b is the elasticity of substitution.[6] Equation (2-5) suggests that as the elasticity of substitution tends to infinity p approaches -1, while as the elasticity of substitution approaches zero p approaches infinity. However, the presence of the parameter p also indicates that the C.E.S. equation has the disadvantageous feature of being non-linear in the logarithms. No simple method of classical

[6] Ibid. Ch. 5.

linear estimation, in which the properties and assymptotic tendencies of the estimators are well known, can be used. Resort must then be made to ad hoc methods.[7]

The C.E.S. form of the production function will reduce but not eliminate specification error. First, for any particular equation of the C.E.S. or Cobb-Douglas type, the elasticity of substitution is assumed to remain constant. However, with increasing output inputs undergo a qualitative change due to an expansion of technical possibilities. Second, a single equation of the C.E.S. or Cobb-Douglas form could only be used to test whether the hypothesized patter of increasing, constant, and then decreasing returns, as in the U-shaped long run average cost curve does not exist. Thus, in the Cobb-Douglas function the parameters a and b are independent of output, and similarly for the parameter estimate V in the C.E.S. equation. Where such independence does not exist specification error would result unless distinct groupings of observations over small levels of output, and large levels of output were constructed and separate functions fitted to each. However, studies in which the production function is described by a single equation can be used to investigate the major criticism of the assumed U-shaped cost-output relationship, i.e., whether diseconomies of scale are ever encountered in

[7] See above p. 59 .

practice. While in the nature of a partial analysis of the problem of cost estimation certain important features of the relevant relationships may be revealed.

The simultaneous nature of the production function relations indicates the direct application of ordinary least squares methods yield biased and inconsistent estimates. This can be illustrated with reference to the Cobb-Douglas production function. The marginal productivity conditions for labour and capital are respectively

$$\frac{\partial X}{\partial L} = a\frac{X}{L} = \frac{P_1}{P_0}. V_1 \qquad (2\text{-}6)$$

$$\frac{\partial X}{\partial K} = b\frac{X}{K} = \frac{P_2}{P_0}. V_2 \qquad (2\text{-}7)$$

where the Vi(i = 1, 2) indicate the disturbances and P_0, P_1, and P_2 denote respectively the prices of output, labour, and capital. Expressing (2-6) and (2-7) in logarithmic form there results

$$\log a + \log X + \log P_0 = \log P_1 + \log V_1 + \log L$$

$$\log b + \log X + \log P_0 = \log P_2 + \log V_2 + \log K$$

Again using the superscript - to denote natural logarithms in perfect competition one can write

$$\bar{X} = \bar{C}_1 + \bar{L} + \bar{V}_1 \qquad (2\text{-}8)$$

$$\bar{X} = \bar{C}_2 + \bar{K} + \bar{V}_2 \qquad (2\text{-}9)$$

where $\overline{C}_1$ and $\overline{C}_2$ are constants with

$$\overline{C}_1 = \log P_1 - \log P_0 - \log a$$

$$\overline{C}_2 = \log P_2 - \log P_0 - \log b.$$

Substituting (2-8) into (2-2) and solving for L gives

$$\overline{L} = \frac{\overline{A} + b\overline{K} + \overline{U} - \overline{V}_1 - \overline{C}_1}{1 - a} .$$

The explanatory variable $\overline{L}$ is dependent on the disturbance $\overline{U}$ and by similar logic it can be shown $\overline{K}$ also depends on $\overline{U}$. Thus, if the disturbance in the production function is positive and output above normal, the greater marginal productivities of the inputs would result in greater quantities of inputs being used.*

To eliminate simultaneity bias possible methods of estimation include calculation of reduced form equations and two stage least squares.[8] The use of reduced form

* It should be noted that the extent of simultaneity bias will be affected by the state of competition in product and factor markets. If a positive disturbance caused output to be abnormally high but a steeply falling demand curve and sharply rising input supply curve caused large increases in the real wage rate the feedback effect on the quantity of input used would be small.

[8] An additional approach is that of Marschak and Andrews which involves the use of a prior restrictions on the parameter values obtained from profit maximizing conditions and economic interpretations of the residuals to achieve identification. However, this approach has not been used at all extensively, since rather than unique estimates of the parameters of the production function, one is only able to narrow the range of admissible values.

equations involves solving multiple equation systems so that each equation contains only one current endogenous variable. The resulting coefficients of the reduced form would be consistent and perhaps unbiased. It would then be necessary to unscramble the structural parameters in the original equations from the reduced form estimates.[9] Alternatively, through two stage least squares methods, first, the least squares regression of the explanatory variable on some specified exagenous variable is calculated. The explanatory variable expressed as a function of this exogenous variable would then be substituted back into the original relation and ordinary single equation least squares regression performed.[10]

For the production function relations described by equations (2-2) (2-8) and (2-9) difficulties are encountered in estimation of the structural parameters since all variables are endogenous. In two stage least squares regression it would be impossible to remove the correlation between the explanatory variables and the disturbance by the first least squares step. The use of reduced form

[9] See Johnston J. Econometric methods. New York, McGraw-Hill, 1963, Ch. 9.

[10] Ibid. Ch. 9.

equations would result in problems of identifying the production surface from the combination of the marginal productivity equations.

The problem of identification arises since a linear combination of the marginal productivity conditions yields an equation that has the same form as the production function. Multiplying equation (2-8) by m and equation (2-9) by (1 - m) and adding gives

$$m\bar{X} = m\bar{C}_1 + m\bar{L} + m\bar{V}_1$$

$$(1 - m)\bar{X} = (1 - m)\bar{C}_2 + (1 - m)\bar{K} + (1 - m)\bar{V}_2$$

$$\bar{X} = m\bar{C}_1 + (1 - m)\bar{C}_2 + m\bar{L} + (1 - m)\bar{K} + m\bar{V}_1 + (1 - m)\bar{V}_2$$

The resulting equation like the production function is linear in the logarithms of output and the two inputs. Consideration must then be given to the interpretation and size of the disturbances. Unless, the disturbances in the marginal productivity conditions are independent of and large relative to the disturbance in the production function identification would be impossible.[11]

[11] It is always the function which is subject to the smallest variance which is identified. The standard example given is the identification of the supply function from a series of price-output observations. If large changes in income have occurred the demand curve would shift tracing out observations along the supply schedule.

Cross Section Analysis

In industries where market conditions approximate those of pure or perfect competition cross section analysis may be restricted in usefulness. Given that prices of input and output will remain constant to the firm, if entrepreneurship and the production function were identical for all firms, it has been argued there exists no observable independent force capable of generating different levels of output.[12] Rather all firms would produce at the minimum point on the long run average cost curve.

Despite constancy of input and output prices in pure or perfect competition arguments have been advanced attempting to show that the production function can be determined. First, it has been maintained that mistakes in the form of producing too much or too little will yield a series of identifiable observations.[13] However, if mistakes are large enough to generate a wide range of outputs, mistakes will also be large enough to permit of wide variation in the degree to which inputs are properly adjusted. Since the existence of mistakes will affect the variance of the

[12] Walters, A.A. An introduction to econometrics. London, Macmillan, 1968. p. 288-289.

[13] Ibid. p. 289.

disturbance in both production and marginal productivity equations identification will not be possible. Second, it has been argued that the ownership of specialized resources, particularly entrepreneurship, will identify the production function.[14] However, this would result in shifts in the production function with the result that one would be measuring the marginal productivity equations.

One means of identification, which avoids the above difficulties, results from imperfect mobility of factor inputs. In pure competition one may postulate that firms will face random differences in factor prices. There would then result a series of different points on the production surface as in figure 2-1. Assume the actual production function for firms in the industry is given by the locus of points RR^1. The relative prices of labour and capital

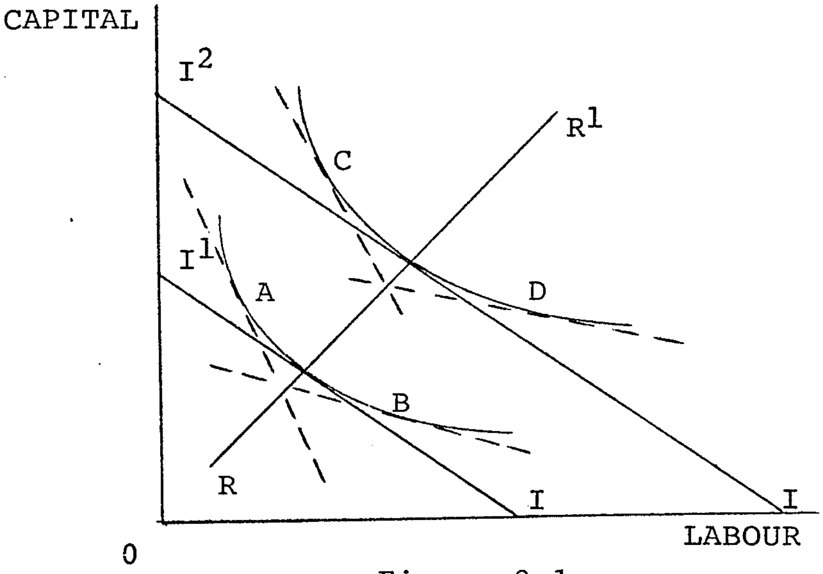

Figure 2-1

[14] Ibid. p. 292.

facing the firm are given by the parallel iso-cost lines II^1, II^2. From a cross section sample differences in relative factor prices would result in the iso-cost lines indicated by dashes. Since in pure competition each firms' purchases would have a negligible effect on input prices, differences in factor prices could be assumed to be random with respect to firm size. A group of points such as A, B, C, and D would be evenly distributed around RR^1, resulting in an average version of the production function.

In imperfect competition firms would be faced with differing elasticities of the demand for output, and differing supply elasticities of factor inputs. The marginal productivity conditions of the firm are

$$a\frac{X_0}{L} = \frac{P_1}{P_0}\left[\frac{1 + (1/E_L)}{1 + (1/E_x)}\right]$$

$$b\frac{X_0}{K} = \frac{P_2}{P_0}\left[\frac{1 + (1/E_K)}{1 + (1/E_x)}\right]$$

where E_x denotes demand elasticity, and E_L and E_K indicate the supply elasticities of labour and capital respectively. Whether sufficient variance will be imparted to the marginal productivity equations to trace out the production function depends on the movement of factor prices as compared to the variation in the price of output over the cross section.

However, while one may be able to derive unbiased estimates of the parameters of the production function, the economic meaning of the results becomes uncertain. Where firms are producing differentiated products attempting to infer the production function of any particular firm within the product group, from the resulting observations becomes especially questionable.

Time Series Analysis

In time series analysis the problems resulting from the lack of an independent force generating different levels of output in perfect competition are considerably lessened. The firms' level of output will change with variations in the price of output resulting from changes in consumer tastes or income. However, additional problems will result in isolating internal and external economies or diseconomies of scale as well as removing the effects of differences in technology.

Growth of the industry may cause changes in factor prices or the physical productivities of some inputs. For example, external diseconomies of a pecuniary nature may result where industry expansion requires inputs which must be bid away from other industries. Also, external diseconomies could occur in the exploitation of natural resources such as logging where resort must be made to progressively less favourable stands of timber. On the other hand, manpower training programs paid for by public funds and falling input supply curves may cause external economies. Graphically, external economies or diseconomies would respectively raise or lower the level of the firms production function.

To remove the effects of changing technology and possibly external economies and diseconomies a multiplicative trend term is usually inserted in the production function. In the case of the Cobb-Douglas production function the amount of output in time (t) resulting from the contemporaneous employment of labour and capital would become

$$X(t) = Ae^{at} L(t)^{b} K(t)^{c}$$

or in logarithmic form

$$\log X(t) = \log A + at + b\log L(t) + c\log K(t).$$

If the quantity of labour and capital were held constant output would change at the rate of "a" over time.[15] The

[15] The effects of changes in technology and industry growth are specified as being neutral. The marginal rate of technical substitution is unaffected since

$$\frac{\partial X/\partial K}{\partial X/\partial L} = \frac{bX}{K} \cdot \frac{L}{aX} = \frac{bL}{bK} .$$

use of a trend term assumes that the advancement of technological knowledge or industry growth is a linear, logarithmically linear, or some other regular function of time. However, bias may well remain if such changes occur sporadically. An explicit measure of the above factors would then be required. For example, in labour intensive industry Niitamo's introduction of a variable called the level of knowledge, defined as the ratio of each years graduating class from lower secondary schools to the size of the work force, may be a superior measure of technological progress.[16]

Measurement and the Quality of Data

One of the basic assumptions of regression analysis is that the variables are observed without error. With respect to the measurement of labour differences in the quality of labour inputs preclude the simple adding of the number of persons employed or the number of man-hours. However, a standardized unit of labour could be obtained by referring to the workers' marginal productivity. The quantities of different types of labour could then be weighted since in

[16] Niitamo, O. The development of productivity in Finnish industry 1925-1952. Productivity Measurement Review. 15: 1-12. 1958.

equilibrium the marginal productivity of the worker equals his wage.[17]

Especially intractable problems occur in the measurement of capital. The appropriate concept of capital inputs is one of the capital services provided, since wide fluctuations can occur in the degree of capacity utilization. However, no satisfactory method exists for ensuring that capital inputs are standardized. It has been suggested that the capital stock deflated by the percentage of the labour force employed would approximate the quantity of capital services.[18] Nonetheless, the distinction between fixed and variable costs suggests that such a method would underestimate capital services.

A further problem in the measurement of capital occurs in attempting to aggregate different kinds of machines, buildings and inventories at different stages of their life cycle and the process of technological change. Unlike the case of labour inputs the price of capital cannot be used in the aggregation process. This results from the fact that the relative prices of equipment are determined by future profit expectations.

[17] Walters, A.A. An introduction to econometrics. op. cit. p. 829.

[18] Solow, R.W. Technological change and the aggregate production function. R. Ec. and Stats. 39: 312-20. 1957.

With respect to the measurement of output firms will seldom be producing perfectly homogeneous products. Changes in the quantity of inputs required may result simply from the inability to obtain a standard unit of output. Moreover, although the product may be roughly homogeneous consideration must be given to the product mix. In the multi-product firm a variation in the facilities for producing one commodity or service may change the production of other goods and services in the same direction. Where the firm produces a relatively small number of products the production function might be derived for each output simultaneously through the use of multiple correlation techniques. However, if the firm produces a multitude of different products the only practical possibility may be to construct a composite measure of the value of different outputs. It would then be necessary to correct the data for average price changes due to imperfections in competition in the case of cross section studies or changes in demand for the firms product in time series analysis.

In addition to measurement difficulties unprocessed data may be subject to the problem of multi-collinearity whereby the explanatory variables labour and capital are correlated. In cross section studies a trend would exist for large firms to employ a greater number of workers. The extreme case of perfect multicollinearity would suggest that the associated variances could not be determined. Where there exists less than perfect multicollinearity the variance estimates will be biased upwards by an amount depending upon the strength of correlation between capital and labour. While in time series analysis similar intercorrelation would exist the relationship would be cyclical in nature and hence may be less pronounced than in cross section studies.

Illustrative Studies of the Production Function

A cross section study was undertaken by Klein for the production of railroad services in the United States.[19] Unlike most studies Klein was able to take into account the simultaneous equation effect. Moreover, a check on the normality and independence of the logarithmic disturbances in the simultaneous equation system was undertaken.

The fact that the railroad industry is a regulated sector of the economy, in which each carrier accepts the traffic as given, can be shown to simplify the estimation procedure. The central problem of each firm is then to minimize costs of the existing or given level of traffic.

[19] Klein, L.R. Econometrics. Evanston,Illinois, Row Peterson. 1953. p. 226-236.

For the i'th firm or carrier the production function for net passenger miles X_{1i} is expressed as

$$X_{1i} = AX_{2i} \cdot n_i \cdot c_i \cdot d_i \cdot u_i \cdot \qquad (2\text{-}10)$$

where X_{2i} is net ton miles, n_i is man hours, c_i is tons of fuel consumed, and d is train hours utilized.[20] The total cost of the inputs is

$$wn_i + qc_i + rd_i \qquad (2\text{-}11)$$

where w equals average hourly earnings, q equals average fuel costs, and r equals average cost of capital services. Minimizing total cost (2-11) subject to (2-10) the resulting marginal conditions are

$$\frac{q^{c}i}{wn_i} = -V_{1i} \qquad (2\text{-}12)$$

$$\frac{rd_i}{wn_i} = -V_{2i} \qquad (2\text{-}13)$$

where Vii (i = 1, 2) is a random disturbance.

It is seen in equations (2-12) and (2-13) that contrary to the competitive model of profit maximization output does not appear. Assuming that firms face random differences in factor prices this implies that the system of equations (2-10) (2-12) and (2-13) is recursive, since each can be estimated in turn through two stage least squares methods.

[20] The specification of the function may be criticized due to possible multicollinearity between the input variables and the output measure - ton miles.

With respect to the quality of data Klein was able to obtain a measure of the flow of capital services. The measure chosen was train hours. Unfortunately data on train hours neglect varying length of trains. To ensure that passenger and ton miles were homogeneous output variables, average length of haul (Z_{1i}) was introduced as an added explanatory variable, as well as a measure of the type of product being transported (Z_{2i}). Finally, as might be expected problems are encountered in the measure of the price of capital services. This was obtained by dividing non-wage maintenance outlays by the number of train hours. However, non-wage maintenance outlays as a measure of the total costs of capital services are deficient since such outlays would be partly determined by the usage of equipment in prior years. Moreover, the measure of total costs would neglect equipment for which it was considered unprofitable to undertake repairs.

The results appear to indicate increasing returns to scale. For seventy-eight carriers producing both freight and passenger service, the regression equation is

$$\log n_i + 0.1349 \log c_i + 0.3124 \log d_i$$

$$\underset{(0.2404)}{.8410} + \underset{(0.0422)}{1.1220} \log X_{1i} + \underset{(0.0208)}{0.1807} \log X_{2i} \tag{2-14}$$

$$- \underset{(0.1057)}{0.3864} \log Z_{1i} - \underset{(0.0904)}{0.2788} \log Z_{2i} .$$

The numbers in parenthesis indicate standard errors. The regression coefficients are all relatively high multiples of standard errors and the multiple correlation coefficient was reported as 0.99. Transforming equation (2-14) into the original exponential form of the production function yields

$$X_{1i} = 5.62\ X_{2i}^{-0.16} . n_i^{0.89}\ . c_i^{0.12} . d_i^{0.28}\ . Z_{1i}^{0.34} . Z_{2i}^{0.25}$$

Thus, the exponents are seen to exceed unity. Whether, in fact there exist increasing returns, however, may be debated since the disturbances showed large departures from normality and correlation significantly different from zero was found between the disturbances in the marginal productivity equations and the production function.

It has been observed that a large number of studies have been made of public utilities and railroads.[21] The findings generally indicate either constant or falling long run average costs. Nonetheless, these results are not inconsistent with traditional theory since public utilities and railroads have always been considered exceptions to the normal hypothesis of a U-shaped long run average cost curve.

[21] Walters, A.A. Production and cost functions: an econometric survey. op. cit. p. 50-51.

A study has been undertaken of halibut fishing using both C.E.S. and Cobb-Douglas forms of the production function.[22] An especially significant feature of this study is the attempt to take into account differences in managerial skills.

As a measure of capital services the market value of the boat was divided by fifty and multiplied by the number of days at sea during a year. Labour input is measured simply by multiplying the number of fishermen aboard by the number of days fished. A variable referred to as "catch per skate", c, is also introduced to account for differences in the density of the fish population.

Cross section estimates for thirty-two boats in each year from 1958 to 1964 are obtained using an ad hoc approximation to the C.E.S. function. The equation fitted is of the form

$$\log q = a_0 + a_1 \log K + a_2 \log L + a_3(\log \frac{K}{L})^2 + a_4 \log C + V$$

where q equals output, v is the disturbance, and a term log c added.[23] The regression coefficients a_1 and a_2, which

[22] Comitini, Salvatore and Huang, David S. Production and factor shares. Journal of Political Economy, 75:366-372. 1967.

[23] For the details regarding statistical properties of the C.E.S. parameter estimates see Kmenta, John, On estimation of the c.e.s. production function. International Economic Review. Vol. VIII, 1967.

indicate the nature of returns to scale, are in almost all cases insignificantly different from zero. This may be attributable to high collinearity between labour and capital. Moreover, the authors state that basing the measure of K on the market value of the boat at the end of 1964, will result in larger and larger errors of observation as one goes back in time.[24]

To more adequately measure capital services time is included as an additional explanatory variable. The Cobb-Douglas function is fitted to the entire sample of seven cross sections. The resulting regression equation is

$$\log q = \underset{(.300)}{.525} + \underset{(.0100)}{0.111t} + \underset{(.0620)}{.214 \log K} + \underset{(.0770)}{.621 \log L} \qquad (2\text{-}15)$$

$$+ \underset{(.0610)}{.576 \log C} \qquad R^2 = .759$$

In addition a pairing of the questionnaire and interview method, by which the skills of the captain were ranked, and multiple regression analysis was used.[25] The following

[24] Comitini, Salvatore, and Huang, David S., op. cit., p. 370.

[25] The authors state that an individual who had been with the halibut industry for many years and was thoroughly familar with the boats and their captains evaluated entrepreneurial skill.

equation was estimated for the seven cross sections

$$\log q = \underset{(.00986)}{.00595t} + \underset{(.0670)}{.121 \log K} + \underset{(.0790)}{.702 \log L} + \underset{(.321)}{.993M_1} \qquad (2\text{-}16)$$

$$+ \underset{(.314)}{.897\, M_2} + \underset{(.300)}{.773\, M_3} + \underset{(.0659)}{.495 \log C} \qquad R^2 = .781$$

where M_1 = 1 if the captain is excellent and = 0 otherwise, M_2 = 1 if the captain is good and = 0 otherwise, and M_3 = 1 if the captain is average and = 0 otherwise. Both equations (2-15) and (2-16) indicate slightly decreasing returns to scale, the sum of the coefficients for labour and capital being .835 and .823 respectively. However, the findings are subject to the criticism that single equation least squares estimates may be biased by the simultaneous equation effect.

The above studies both encountered difficulties in obtaining an adequate measure of capital services. In attempting to rectify this problem some studies have distorted the meaning of the production function. In a study of coal production in Great Britain by Lomax capital inputs were measured by the amount of coal cut by machinery and obtained independently by pneumatic picks over the period 1927 to 1943.[26] It is stated that there is obvious danger in taking partial output figures as an independent variable but this

[26] Lomax, K.S. Coal production functions for Great Britain. J.R.S.S. 113:346-51. 1950.

is outweighed by the advantages of an index which signifies actual use and does so efficiently.[27] The conclusion of this study is that a one percent change in capital and labour would result in an approximately similar one percent change in output. However, since seventy-five percent of output was mechanically cut and thus entered into the measure of capital the nature of returns to scale has not been determined.

Statistical Cost Analysis

A major advantage in the analysis of the cost function is the ability to describe in a single equation cases where the nature of returns to scale is changing. Three major hypotheses concerning the slope of the cost-output relationship may be described by the inclusion of first, second, and third degree terms in output. The cost function may then be specified as

$$Y_t = a_0 + a_1X_{1t} + a_2X_{2t} + a_3X_{3t} + \dots + a_kX_{kt} + U_t$$

$$t = (1, 2, \dots, n)$$

where Y_t denotes total cost, X_{1t} represents the rate of output, while X_{2t} and X_{3t} may designate squared or cubed

[27] Ibid. p. 346.

terms in output. The remaining X's would be external factors, whose influence one is trying to hold constant.

The application of least squares techniques to the above cost function requires consideration of how the level of output was determined. Let the demand function be of the form

$$P_t = a_0 - a_1X_t$$

where P_t is the price associated with output X_t in period t.

The total cost function is assumed to be

$$IIt = B_0 + B_1X_t + B_2X_t{}^2 + U_t$$

where U_t suggests that costs may vary from period to period about the expected value of the polynomal. On the assumption that businessmen are attempting to maximize profits the level of output would be

$$X_t = \frac{a_0 - B_1}{2(a_1 + B_2)} + V_t \qquad (2\text{-}17)$$

where V_t is the divergence between actual and desired output. In perfect competition equation (2-17) would be expressed as

$$X_t = \frac{a_0 - B_1}{2B_2} + V_t \qquad (2\text{-}18)$$

since a_1 equals zero.

Equation (2-18) indicates that with perfect competition little variation in output would be observed in cross section analysis since the only possible source of output change would be random disturbances about the profit maximizing position. In time series analysis while changes in price may generate different output levels there remains the additional problem of correlation between the disturbance terms in the cost function and the output determination function. Consequently there would also be a lack of independence between the explanatory variable and the disturbance in the cost function since

$$E(X_tU_t) = \frac{a_0 - B_1}{2B_2} \quad E(U_t) + E(U_tV_t)$$

$$= E(U_tV_t).$$

where E represents expected value.

Independence of the disturbances in the output determination function and the cost function is not a likely result. Rather if costs are above expected levels output will be below that planned. For example, a disturbance such as a machine breakdown which elevates costs above expected levels will result in a reduction in output below that initially planned.[28] Consequently, unless a check has been

[28] Johnston, J. Statistical cost analysis. op. cit. p. 41.

made of the independence of the disturbances, the findings of statistical cost studies are methodologically suspect.

Cross Section versus Time Series Analysis

Cross section studies of the cost function and production function in pure or perfect competition encounter similar problems with respect to a lack of independent stimuli generating different levels of output. Also, common to both studies of the production and cost functions will be the problem of eliminating differences in entrepreneurial ability and the quality of factor inputs. Finally, the source of information on costs has been accounting data. In cross section analysis it may be unlikely that the accounting records of a group of firms are comparable. The author of a well known text on financial statement analysis has cautioned:

> "the figures of one enterprise may be compared with those compiled for another only with great care. The combination of the financial statement data of different enterprises for statistical studies is usually unsatisfactory."[29]

[29] Smith, Caleb A. Statistical cost functions. In Cost Behaviour and Price Policy, NBER, Princeton, Princeton University Press, 1955, p. 216.

Especially, important in this context may be differences in the depreciation methods used by firms. Depreciation charges may be allocated according to the straight line, double declining balance, or sum of the years digits techniques. For example, the straight line method would uniformly allocate costs over time, whereas the double declining balance or sum of the years digits technique charge a greater proportion of costs to the early years in which the equipment is being used.

The regression fallacy has been a common criticism of cross section studies of the cost-output relationship. Friedman attempts to explain the regression fallacy by means of the following example:

> "Suppose a firm produces a product the demand for which has a known two year cycle, so that it plans to produce 100 units in year one, 200 in year two, 100 in year three, etc. Suppose also that the best way to do this is by an arrangement that involves identical outlays for hired factors in each year (no variable costs). If outlays are regarded as total costs, average cost per unit will obviously be twice as large when output is 100 as when it is 200. If instead of years one and two we substitute firms one and two, a cross section study would show sharply declining

> average costs. When firms are classified by actual output essentially this kind of bias arises. The firms with the largest output are unlikely to be producing at an unusually low level; on the contrary they are likely to be producing at an unusually high level and conversely for those which have the lowest output."[30]

Considerable confusion appears to exist in interpreting the regression fallacy as elaborated by Friedman. It has been stated by Borts that the regression fallacy as it applies to scale economies has never received an unambiguous definition.[31] Friedman would appear to be emphasizing the normal rate of output as opposed to temporary fluctuations in determining the quantities of inputs to be used. However, it has been argued by Walters that in discussing a known two year cycle of production Friedman neglects the fact that for all firms in a given industry in a given year the state of the business cycle will be approximately a constant factor.[32] While

[30] Friedman, M. Comment, In Conference on Business Concentration and Price Policy, Princeton, Princeton University Press, 1955.

[31] Borts, G.H. The estimation of rail cost functions. Econometrica. 28:108-131. 1960.

[32] Walters, A.A. Expectations and the regression fallacy in estimating cost functions. Rev. Ec. and Stats. 42:210-215. 1960.

the observations will reflect short run differences in capacity utilization Friedman does not show that such differences will be related to size of firm.

An alternative interpretation of the regression fallacy suggested by Borts emphasizes the divergence of the observed output rate due to unforseen changes in demand.[33] In addition to the short run and long run cost curves, one may then define a third cost curve the "short run maladjustment cost function".[34] This latter cost function indicates actual costs which deviate from the plant curve depending upon the extent to which the planned level of output is achieved. The envelope of the short run maladjustment cost curves would then give the plant cost curve. However, again as in the Friedman version there is no explanation given as to whether the effects of unplanned changes in output will be related to firm size.

To extract the long run average cost curve from the observations capacity output and planned output might be introduced as explanatory variables in the cost-output function. It has been suggested that an equation may be fitted as follows:

$$C_t = aQ_t{}^c + B(Q_t{}^p - Q_t{}^c) + c\,(Q_t - Q_t{}^p)$$

[33] Borts. op. cit. p. 114.

[34] Wilson, T.A. and O. Eckstein, Short run productivity behaviour in U.S. manufacturing. Rev. Ec. and Stats. 46;41-54.

where C_t is total cost in time t, $Q_t{}^c$ is capacity output, $Q_t{}^p$ is planned output, and Q_t equals actual output.[35] However, the following criticisms may be made. First, only where there exists constant returns to scale will the parameter B measure solely the effects of changes in capacity utilization peculiar to the short run. Where the nature of returns to scale varies the parameter B will partially include the effects of changes in capacity which are implied by the theory of long run average cost. Second, stochastic economies resulting from a proportionately smaller variance of sales fluctuations with increases in scale would be excluded. Even where there exists constant returns treating differences in the degree of capacity utilization and the inflationary effects on costs associated with unplanned output changes as completely external to the cost-output relationship may only result from a rigid adherence to static concepts. Thus, in both cases there arises the problem of distinguishing internal from external factors.

Time series analysis encounters problems of differences in technology, and removing the effects of external economies and diseconomies as in studying the production

[35] Ibid. p. 43.

function. However, it also becomes necessary to correct the data for factor price changes. The method of adjustment used has been to super-impose some particular set of base factor prices to the actual factor inputs in each period. This procedure assumes that changes in factor prices during the period have not resulted in a shift in the physical proportions of the factors employed. However, if factor substitution did occur the combination of factors employed would be more expensive at the prices of the correcting period than the different combination which would have been used had the prices of the correcting period actually prevailed.

Bias may also result in time series analysis due to the method of allocating depreciation costs. This would occur where there exists a divorce of ownership and management. During periods of slack demand management may charge a smaller proportion of depreciation costs to that periods operation in order to appear to be making profits. Thus, low levels of output and below normal profits may result in an understatement of actual costs.

Measurement and the Quality of Data

Where the firm or plant produces many differentiated products it becomes difficult to identify individual costs. The usual method employed is to construct an output index by weighting quantities of differing output with estimates of average direct costs. However, this amounts to determining output by costs, i.e., to introducing a spurious dependence where measurement of an independent relationship is wanted. Other weights used include relative produce prices, or the amount of raw materials entering into the different products. Relative factor prices would be inadequate as one cannot simply assume that a higher price for a particular good or service will cause more of that good or service to be produced. The amount of raw materials utilized would also be unsuitable due to the difficulties in summing physically diverse inputs. However, an alternative method where the number of products is small would be to again use more than one measure of output as independent variables in multiple correlation analysis.

In addition to the problems of differences in depreciation methods and changing factor prices an intricate processing of accounting cost data may be necessary due to the following difficulties. First, the concept of cost used in economic theory is opportunity costs. To translate accounting costs to costs as perceived in economics a value will have to be imputed to those productive factors supplied by the owner. The most significant component of imputed costs would generally be the value of services provided by the owner, which should be measured by the highest interest rate such funds could obtain elsewhere. In addition, the value of services of an owner-manager should be included in opportunity costs but in accounting data would generally appear under profits. While accounting costs may underestimate opportunity costs due to failure to include productive factors supplied by the owner, over-estimation may occur if payments to owners of productive services, which are specific to the firm, and worthless if not employed by that firm are included. Examples of productive factors specific to the firm would be local monopoly rights and public carrier licenses granted by the government, or a highly specialized entrepreneurial skill which is a natural endowment.[36]

Second, the unit period for accounting purposes will generally be longer than the unit economic period.[37] For each accounting period observed output will not be produced under the theoretically desirable condition of a uniform rate of production within the period. While the extent of such variations in output may be lessened if the length of the period of observation could be

[36] Walters, A.A. Production and cost functions: an econometric survey. op. cit. p. 42.

[37] Johnston, J. Statistical cost analysis. op. cit. p. 26-27.

shortened, problems would occur in matching cost and output figures. Thus, it may be argued that changes in the rate of output within the interval of observation will inevitably bias the observed relationship due to averaging effects.

Illustrative Studies of the Cost Function

A study of electricity generation in Britain has been undertaken by Johnston using both cross section and time series methods.[38] Moreover, firms produce under the directions of the Central Electricity Board and were not in the position of adjusting output in the search for maximum profits. Since the level of output is an exogenous variable the simultaneity bias is thereby avoided.

For the period 1928 to 1947 time series analysis was used giving total corrected working expenses as a linear function of output, measured in Kilowatt-hours, and time. Working expenses were used as a proxy for the variable costs of economic theory and included: (1) fuel costs, (2) salaries and wages and (3) repairs and maintenance. To maintain a constancy of absolute and relative factor prices each component of working expenses was deflated with a selected

[38] Ibid. p. 44-73.

price index number. It is necessary to consider whether the implied assumption of unchanging factor proportions is realistic. If machines are in poor working order fuel costs may be increased. A rise in the price of fuel may result in greater repairs being undertaken indicating factor proportions may well vary. Finally, an additional source of bias may result from pecuniary economies of large firms being excluded.

For twenty three firms the predominating type of equation shows total costs as a linear function of output with or without the inclusion of time as an explanatory variable. For eight firms the trend term proved significant while for six firms a quadratic cost function significantly improved the goodness of fit. The existence of a quadratic cost function for six firms may result from greater variation in the range of output since in a few cases the highest plant level is as much as seven times as great as the lowest, but in most cases the ratio is about 2:1.

A cross section study of the variation of working expenses with the level of output was conducted for the year 1946-1947. Johnston states that while there exist differences in the type, age, absolescence of plant, and nearness to coal fields among firms, the influence of such factors will average out with large sample size.[39] For forty

[39] Ibid. p. 51.

firms with output ranging from 1.1 to 1,150.5 units a simple linear regression was obtained

$$Y = 57.6 + 1.3298X \qquad (2\text{-}19)$$

$$R^2 = .9534$$

where Y equals total working costs, and X equals output. Finally, it is stated that terms in X^2 and X^3 were both non-significant.[40]

A logarithmic function incorporating thermal efficiency as an additional explanatory variable was also fitted for approximately the same forty firms. Through the introduction of thermal efficiency attempt was made to account for differences in the age, type, and efficiency of different plants. The logarithmic form was chosen since a given change in thermal efficiency may be expected to exert a constant proportional change, rather than a constant absolute change in total costs. The resulting equation was

$$Y = 8.301 \quad X^{0.7919} E^{-0.0175V}$$

where Y and X are defined as before, and V equals thermal efficiency. Holding V constant the relationship of average working expense to the level of output is given by

$$\frac{Y}{X} = 5.8956 \quad X^{-0.2081} \qquad (2\text{-}20)$$

[40] Ibid. p. 66.

The linear function of equation (2-19) and logarithmic function in equation (2-20) yield contradictory results: suggesting constant and increasing returns respectively. Unfortunately, neither the standard errors are included or in equation (2-20) the correlation coefficient so that one cannot determine which more accurately describes the variation of working costs with changes in output.

With respect to capital charges again a linear as well as logarithmic function was tested. Unfortunately the unavailability of British statistics necessitated resorting to American data. Moreover, it is noted that plant costs will be influenced by the date of installation; plant and construction costs being high in the period from 1920 to 1930, low from 1930 to 1940, and high again in the post-war era.[41] To account for variation in load factors a measure of the percent of full rated capacity utilization (P) was included in the logarithmic function. For 73 firms the following equations were drived:

$$Y = 382 + 1.8030\,X + .0003674\,X^2 \qquad (2\text{-}21)$$

$$R = .9301$$

and

$$Y = 8.898X^{0.9746}\,E^{-\,0.11459P} \qquad (2\text{-}22)$$

[41] Ibid. p. 68.

In equation (2-21) a term in X^2 proved significant while in equation (2-22) holding capacity utilization constant average capital charges were reported to fall sharply at first and then level off.

It is concluded that examination of working expenses and capital charges indicate that long run average costs falls quickly and steeply thereafter approximating a straight line. However, the following criticisms may be made. First, output is not perfectly homogeneous. One kilowatt produced for one thousand hours is qualitatively different to the buyer as well as the supplier than a thousand kilowatts produced for one hour. Second, either factor price changes have not been accounted for, or the method of holding factor prices constant, itself, leads to error. Finally, the use of data on capital and working expenses from two completely diverse sources conflicts with the marginal conditions underlying the theory of long run average cost.

A second study of life assurance in Great Britain may be viewed as a direct test of the hypothesis that increasing complexity of the managerial function will result in diseconomies of scale.[42] Cross section analysis was undertaken for the year 1952 of the relationship between total

[42] Ibid. p. 106-110.

annual premiums and the sum of management expenses and commission, expressed as a percentage of premium income. The correlation coefficient between the expense ratio and the logarithm of total annual premiums is -0.3702 which for 61 observations is significant at the 5 percent level.

It is stated that the results may be biased due to two factors. First, due to the payment of the initial commission and other expenses associated with the issuance of the policy, firms with a greater proportion of new business may be expected to have a higher expense ratio. Second, total annual premiums may consist of individual policy business or schemes business such as group life or group endowments.[43] Schemes business is serviced at much lower expense so that the inverse relationship between the expense ratio and total premiums may also be explained by a greater proportion of schemes business among large firms.

To check the validity of the above relationship the data was stratified into three groups, according to the amount of schemes business, and the percentage of new premiums to total premiums was included as an additional explanatory variable. For each of the three groups the partial correlation coefficient was calculated between the expense ratio and the logarithm of total premiums with new

[43] Ibid. p. 107.

premiums as a percent of total premiums held constant. The partial correlation coefficients in order of increasing schemes business were -.5546, -.4482, and -.7868 which are significant at the 5 percent, 10 percent, and 1 percent level respectively. It is concluded that allowing for the effects of schemes and new business again the expense ratio declines with increasing total premiums.

The following factors should be considered in evaluating the observed relationship of expense ratio to total premiums. First, the quality of management and technological efficiency may vary with size. The latter may be especially important as increasing complexity of the managerial function is allayed by the introduction of data processing systems and computerization. Second, there may exist simultaneity bias since part of the observed cost-output relationship may be the result of entrepreneurs adjusting output in the search for maximum profits.

Conclusions on Statistical Production and Cost Analysis

The statistical approach used in the analysis of production and cost functions is based on the mathematical theory of statistics, a theory suggesting how inferences may be drawn from a random normally distributed sample. In the natural sciences where one can conduct controlled experiments a relationship, subject to experimental error, may be readily investigated with statistical techniques. However, in economic applications it is less obvious that the assumptions of theoretical statistical models are satisfied due to the existence of a multitude of simultaneous events, and relationships associated with the problem being investigated. The situation has been described as one in which:

> "the economic system grinds out its' complex convolusions; the myriad of actors, - consumers, firms regulatory agencies, and governmental units act and interact; a more or less imperfect collection of statistical agencies records, with various degrees of errors and omissions, partial, quantitative measures of this evolutionary working process, and the poor econometrician comes along in the wake of the monster, gathering what data he can in an attempt to test various hypotheses about the aspects of economic activity."[44]

The Survivor Technique

While both the survivor technique and statistical cost analysis are of an ex post nature unlike statistical cost analysis the survivor technique makes no reference to the actual cost records of a firm. Rather, the level of

[44] Ibid.

costs is inferred by changes in the percent of industry output supplied by the firm.

The fundamental assumption of the survivor principle is that only efficient firms will be able to survive in a competitive market structure. Expressing this principle in Darwinian terms Marshall states that as a general rule the law of substitution - which is nothing more than a limited and special application of the law of survival of the fittest - tends to make one method of industrial organization supplant another when it offers a direct and immediate service at a lower price.[45] The survivor technique then proceeds to solve the problem of determining the optimum firm size as follows: classify the firms in an industry according to size and calculate the share of industry output contributed by each class over time.[46] Size classes experiencing increasing shares of industry output are assumed to be efficient, while evidence of a declining class share is taken to mean relative inefficiency. However, such efficiency is defined in a broader context than the theoretical concept of efficiency in production and distribution, which assumes a given set of demand conditions. An efficient firm would be one capable of meeting the problems of the total economic environment including

[45] Stigler, G. Economies of scale, Journal of Law and Economics. 1:54-66. 1958.

[46] Ibid. p. 56.

prediction of future demand, introducing new products, unstable foreign markets, general recessions, etc. Such a broadening of the definition of efficiency implies, moreover, that since the total economic environment will seldom be the same for all firms even in the same industry, there will be a range of optimal firm sizes.

The importance of competition as an eliminating mechanism underscores three key questions. First, how is the industry to be defined? This decision will establish what data are included in the estimation procedure. The solution suggested by economic theory whereby all firms producing "closely substitutable" products constitute an industry is fraught with difficulties in a world where firms produce multiple differentiated products. Second, given the different forms competition can take, how is it to be determined whether there exists sufficient competition to result in a proper test of efficiency? No exact rules exist for judging at what point spatial or product differentiation barriers render firms non-competitive. Further with respect to the number of firms in an industry, dissension exists in the literature as to whether the survival principle operates under oligopoly. Third, in terms of different objectives of firm behaviour will firms in an industry be equally aggressive in attempting to expand their share of the market? For example with the divorce of ownership and management sales revenue maximization rather than profit maximization may be the desired goal. Even where firms produce homogeneous products evidence of an increase in market share by one firm does not necessarily indicate greater competitive efficiency. The expansion in market share may reflect the willingness on the part of management to accept lower profits, to the extent that it would not compromise the "normal" rate of return expected by the owners.

Illustrative Survivorship Measures

The pioneering study of long run average cost by means of the survivor principle was conducted by G. Stigler, who applied the technique to the production of steel ingot.[47] The percentage of industry capacity contributed by both differing firm and individual plant sizes was calculated for the years 1930, 1938, and 1951. To better ensure that all firms were supplying a common market and producing the same quality of steel ingot, analysis was restricted to firms using similar production processes. A fairly large number of firms were producing ingot; the average for the years observed being 52 firms. For individual firms with a capacity of from 2½ percent to 25 percent of industry capacity their market share grew or remained constant. This range was then concluded to be consistent with optimum size. While it cannot be established how much greater than the minimum are the costs of firms experiencing declining market shares, since the share of firms with less than a

[47] Ibid. p. 57.

half percent of total capacity fell more than firms having greater than 25 percent capacity, the former is stated as being subject to greater diseconomies of scale.

A total of 117 individual plants were also studied. The smallest size groups again are shown to decline. Plants having a capacity less than 3/4 of a percent of the total capacity experienced reduced market shares, while the remaining plants with capacity from 3/4 of a percent to 10 percent of the total exhibited no systematic tendency towards smaller market shares.

Two main difficulties are associated with the above findings for steel ingot. First, it is recognized that shifting industry boundaries make the identification of whether the market is national or regional in scope difficult to determine. The solution to this problem is not satisfactory, as Stigler simply asserts that a national classification probably does less violence to the facts than a sharp regional classification.[48] Second, the question arises of how representative are the findings. The possibility that the economic environment and forces affecting the size distribution are untypical, in terms of the underlying trend is large when focusing upon only three points in time. Also, many size classes were characterized by a paucity of firms as well as plants. For example, the largest 4 firm sizes included on average only 2 or 3 companies while the largest 2 plant sizes contained a maximum of 3 plants over the years 1930, 1938, and 1951.

[48] Ibid, p. 57.

A second study, undertaken by Stigler, for companies in the automobile industry reduces the problem concerning the extent of the market and refers to annual data for the entire period from 1936 to 1955.[49] Companies were classified by expressing actual production (rather than capacity) as a percent of the total national output of automobiles. It was observed that changes in the size distribution during different periods varied with price controls introduced in the immediate post war period and two years after the start of the Korean War. Thus, long run average cost was believed to rise for the largest outputs in inflationary times, when price controls existed, although no such tendency was thought to result in other times. Finally, although subject to fairly erratic movements the share of smaller companies over a longer span of time falls.

The major deficiency in applying the survivor principle to the automobile industry is the restrictions on competition deriving from the small number of firms in the industry. Stigler regards this as only a statistical problem, which reduces sample size, and no consideration is given to the more fundamental issue of how the nature of competition will be affected. The justification for this treatment it would probably be argued is the broader context in which efficiency is defined. Thus, applying the Darwinian rationale the economic environment would be seen as providing

[49] Ibid. p. 61.

sufficient tests of fitness or efficiency in the form of such factors as unstable foreign markets, strained labour relations, government regulation that even if collusive agreements existed unpredictable factors introduced by a changing economic environment would dissolve checks on competition and firms less able to cope with the new circumstances would be extirpated. Nonetheless, it has been maintained that large firms have the power to significantly refashion the environment according to their own liking. Galbraith contends the giant corporation through advertising, increasing complexity of the managerial function, and re-investment of profits is immune from the control of consumers, stockholders, and the capital market. Further, through vertical integration and the monopolization of inputs required to produce in an industry inefficiency will remain. That many of such elements are present in the auto industry is evident and suggests the inappropriateness as a testing ground for the survivor technique.

Having examined the studies by Stigler it is seen that the effectiveness of competition in the industries selected is uncertain. Further, to explain why some firms are more able to compete effectively, it is necessary to go quite beyond the survivor technique. Thus, survival may be attributed to different goals of firms behaviour, technological change, or circumvention of the law. Finally, to clarify the relationship between competitive effectiveness and low costs, greater attention should be given to whether plants are operated by one plant firms or multi-plant firms so as to eliminate internal cross subsidization.

Questionnaire and Interview Method

Investigating 20 manufacturing industries Bain asked businessmen to estimate the minimum physical production capacity of plant required for lowest unit costs of production and distribution.[50] Also, businessmen were questioned on the percentage by which total unit costs would be higher below minimum efficient scale (M.E.S.). For two industries plant scale economies were classed as "very important" as (M.E.S.) exceeded 10 percent of total market capacity and unit costs were elevated by 5 percent or more at half optimal scale. Five industries were thought to have "moderately important" plant scale economies, M.E.S. being 4 to 6 percent of market capacity and unit costs raised by at least 5 percent at half optimal scale. For nine industries a small M.E.S. and relatively flat scale curve indicated unimportant plant scale economies.[51] The greater frequency of a small M.E.S. is consistent with traditional theory. However, rather than a U-shaped relationship unit costs may eventually become constant. Finally, no systematic relationship between plant scale economies and concentration was found.

[50] Bain, J.S. Barriers to new competition. Cambridge, Harvard University Press, 1956. p. 71-93.

[51] Ibid. p. 103-105.

A distinctive future of this study is the clear separation between economies of the large plant and economies of the large scale firm. Questions on multi-plant economies revealed either economies did not exist or where present were of slight magnitude.

The validity of results obtained through a questionning of businessmen will vary according to the nature of the industry in which firms operate and the skill of the investigator in formulating questions. An initial requirement is that demand conditions not act to constrain firms from attaining large size. If due to the nature of demand only small scale operations could be sustained firms may lack any knowledge of minimum efficient scale considering the question to be irrelevant in terms of present output. In the opposite case where production takes place significantly in excess of minimum efficient scale businessmen may be reluctant to disclose information fearing the dissolution of the existing industry structure. Further, in the formulation of questions difficulties in translating ideas from the economists language to that of the businessman may introduce error. For example, if information is sought on the variation of production and distribution costs with scale, the distinction between distribution and selling costs as perceived in economics must be made clear.

Engineering Estimates

Investigation of economies and diseconomies of scale from engineering data offers two main advantages.

First, the assumptions of engineering data are consistent with those of the theoretical long run average cost relationship. Capacity is varied while supply conditions, product design and location are all held constant. With respect to technology changes in technique do enter but are confined to the existing knowledge of the state of the arts. Second, unlike statistical cost and production analysis, and the survivorship technique observations are not restricted to the narrow range of outputs which producers consider commercially profitable.

Most engineering estimates are based on input-output type models. Data on the relationship of inputs to outputs is obtained from engineering text-books where such relationships have been calculated based on the laws of physics and chemistry. It is generally individual pieces of equipment or process areas for which data is available. For example, to determine the input requirement for an unspecified model the hardness, tensile strength and resistance to shear necessary to produce a given result may be given. The units of study then are separate physical processes which must be combined to give the overall input-output function.

The manner in which the engineering production function relates to the traditional production function of economy theory has been described as follows:[52] The production

[52] Chenery, H.B. Engineering production functions. Q.J.E. 63:507-531. 1949.

function of economic theory may be written

$$X = f\ (U_1, \ldots, U_m) \qquad (2\text{-}23)$$

where X is output per unit of time and U (i = 1,...,m) is the quantity of each physical input. A number of engineering variables (V_i) will describe each unit of physical input yielding

$$U_i = U_i\ (V_1, \ldots , V_n) \quad .$$

The engineering production function would then be

$$X = \emptyset\ (V_1, \ldots, V_n) \qquad (2\text{-}24)\ .$$

since the U's in equation (2-23) may be expressed in terms of $V_1, \ldots, V_n$.

Through a similar transformation the cost function can also be obtained. If the price per unit of physical inputs is P_i then total costs are

$$C = \sum_{1}^{m} U_i P_i \qquad (2\text{-}25)$$

where

$$P_i = P_i\ (V_1, \ldots, V_n) \quad .$$

Since both quantities and costs are functions of the engineering variables, equation (2-23) may be expressed as

$$C = g(V_1, \ldots, V_n) \quad . \qquad (2\text{-}26)$$

The usual mathematical procedure for minimizing cost for any given output then follows from equations (2-24) and (2-26).

Illustrative Engineering Estimates

Ferguson has developed a multidimensional marginal cost function for air transportation.[53] Explicit examination was given to the effects on costs of changes in the quality of output, technology, as well as the rate of output.

For each product characteristic, which is considered significant in terms of costs, output is considered to have another dimension. In air transportation output may be specified as depending upon hours, speed, and weight. This is written as

$$X = 3600\ H_h\ \ V\ W\ .$$

where

X = output measured in ft. - lbs. produced per month

H_h = number of hours flying time

V = velocity in ft./sec.

W = gross weight

Weight and hours are interpreted as quantitative dimensions of output while speed is considered a qualitative characteristic.

For changes in each of the above variables changes in the amount of fuel consumption is then determined. Fuel consumption depends, first, upon the power required to maintain an airplane in level equilibrium flight, and for

[53] Ferguson, A.R. Empirical determination of a multidimensional marginal function. Econometrica. 18: 217-235. 1950.

ground operations (taxiing, take off, and landing). Second, fuel consumption will vary according to the power produced as determined by the efficiency of converting fuel into useful power. Fuel consumption (F) expressed in terms of engineering variable is

$$F = \frac{3600\ Hh(b_1 s_1 + b_2 s_2\ pV^3 + \dfrac{2W^2}{b_2 s_1 pV} - T) + b\ M_b + E}{c \cdot e_p \cdot e_t} \qquad (2\text{-}27)$$

where

$$\left.\begin{array}{l}\left.\begin{array}{l} b_1 = \text{drag coefficient for zero lift}\\ s_1 = \text{wing area in ft.}^2\\ s_2 = \text{parasite area in ft.}^2\\ b_2 = \text{factor of induced drag}\\ T = \text{power equivalent of jet thrust}\end{array}\right\}\begin{array}{l}\text{determinants}\\ \text{of power}\\ \text{required}\end{array}\\ \\ \left.\begin{array}{l} c = \text{combustion energy in ft. lbs./}\\ \quad\ \ \text{lb. of fuel}\\ e_p = \text{propulsive efficiency}\\ e_t = \text{thermal efficiency}\end{array}\right\}\begin{array}{l}\text{determinants of}\\ \text{power produced}\\ \text{/hr. fuel consumed}\end{array}\end{array}\right\}\begin{array}{l}\text{total fuel}\\ \text{consumption}\\ \text{- flying}\\ \text{time}\end{array}$$

$$\left.\begin{array}{l} B = \text{number of landings}\\ Mb = \text{fuel consumed per landing}\\ E = \text{fuel consumed in other}\\ \quad\ \ \text{ground operations}\end{array}\right\}\begin{array}{l}\text{determinants of}\\ \text{total fuel}\\ \text{consumed in}\\ \text{ground operations}\end{array}$$

It is stated that the form of equation (2-27) is independent of technological change, type of aircraft employed, the conditions of operation, or institutional factors.

From equation (2-27) the marginal cost of qualitative changes is then obtained by substituting X/(3600 W V) for Hh and taking the partial derivative of fuel consumption with respect to speed:

$$\frac{\partial F}{\partial V} = \frac{X}{c\ e_p\ e_t\ w}\left[(b_1 s_1 + s_2)\ pV - \frac{4w^2}{b_2 s_1 pV^3} + \frac{T}{}\right].$$

Similarly the marginal cost of technological changes such as changes in wing design could be determined:

$$\frac{\partial F}{\partial S_1} = a\left[(b_1 pV^3/2) - (2W^2/b_2 pVs_1{}^2)\right] .$$

Finally, the marginal cost associated with each of the quantitative dimensions of output is calculated. To eliminate qualitative changes speed and the number of landings is held fixed. The relationship between changes in hours and fuel consumption is

$$\frac{\partial F}{\partial H_h} = \frac{3600\ (\frac{b_1 s_1 + s_2}{2}\ pV^3 + \frac{2W^2}{b_1 s_1 pV} - T) + k\ m_b}{c\ .\ e_p\ .\ e_t} .$$

Since a curvilinear relationship results, assuming factor prices constant, partial support is given to the hypothesis of a U-shaped long run average cost. With respect to the second quantity variable, changes in gross weight of the airplane, the effects on fuel consumption in landing could not be established. However, the relationship between changes in gross weight and fuel consumed in the air is calculated as

$$\frac{\partial F}{\partial W} = \frac{(4)\ (3600)\ H_h}{b_2 s_1 pV} \cdot \frac{W}{c.e_p.e_t}$$

yielding a linear function.

The above study has made no use of standard statistical analysis. With respect to the ability to directly control for qualitative changes in output, and changes in technology, the engineering approach is clearly superior. However, Ferguson states engineering studies must be supplemented by or considered ancillary to statistical investigations since the amount of effort that must be expended to obtain a quantitative statement of the determinants of each type of input in a complex industry is very great and in some cases impossible.[54] Nonetheless, used in combination with statistical studies a knowledge of engineering relationships may be highly useful. Engineering data may suggest the relevant variables, and the shape of the equation so that only the values of the parameters must be determined from the observations.

A second study by Moore combines the engineering approach with statistical techniques.[55] For the equipment studied engineers predict geometrical relationships will exist. This suggests an equation of the form

$$C = aX^b$$

provides an appropriate basis for fitting a least squares line to cost-capacity data. The relationship between capital costs C and output capacity X will be one of

[54] Ibid. p. 233.

[55] Moore, F.T. Economies of scale: some statistical evidence. Q.J.E. 73: 232-245. 1959.

increasing, constant, or decreasing returns depending upon whether the scale coefficient b is less than, equals, or is greater than unity.

Moore made a deliberate attempt to ensure the data was homogeneous with respect to the different methods of expanding plant and equipment. Study was restricted to complete new plants and balanced additions.

In all cases where statistical tests have been applied aluminum reduction, aluminum rolling, aluminum drawing - the scale coefficient is not significantly different from 1 at the .05 significance level. (see Table 1) It is also stated that the same applies to cement although the standard deviation is not given.[56] The findings may be criticised due to failure to measure strength of correlation and standard errors. Moreover, almost no information is given on the range of output considered. Extrapolation of the regression line, however, may lead to serious error. For example, in the building of fractionating towers, it has been stated that an economical limit is reached at about twenty foot diameters beyond which very heavy beams are necessary.[57]

Through similar methods Haldi and Whitcomb have obtained various estimates of the scale coefficient.[58] This

[56] Ibid. p. 242.

[57] Ibid. p. 235.

[58] Haldi, J., and Whitcomb, D. Economies of scale in industrial plants. J.P.E. 75: 373-385. 1967.

Table 1

Economies of Scale in

Plant and Equipment

ITEM OR PROCESS	b	r	σ_b
1. ALUMINUM REDUCTION			
Total plant and equipment	.93	.98	.06
Total equipment	.95	.99	.03
2. ALUMINUM ROLLING			
Total plant	.88	.95	.16
Equipment	.81	.93	.18
3. ALUMINUM DRAWING			
Total plant	1.00	.99	-
Equipment	.92	.92	.13
4. CEMENT			
Equipment	1.06	-	-
Total plant	.77	-	-
5. OXYGEN COMPRESSION			
Equipment	.54	-	-

b = value of scale coefficient

study has the advantage of being able to collect data on a greater proportion of a firms activity than is contained in most engineering studies. Thus investigation was made of: (1) the cost of individual units of industrial equipment, (2) the initial investment in plant and equipment, and (3) operating costs namely labour, raw materials, and utilities. Unfortunately, neither the standard error in the scale factor nor the correlation coefficient is given. Instead scale factors between .90 and 1.10 were arbitrarily classified as not significantly different from one.

For basic industrial equipment out of a total 687 estimated scale coefficients 618 (90.0 percent) display increasing returns and 50 (7.3 percent) show constant returns. Decreasing returns were observed for only 19 scale coefficients or 2.8 percent. Separate analysis was given to equipment which would be likely to exhibit geometrical relationships. For various types of containers the median value of the scale coefficient of between .60 and .69, is consistent with the simple mathematics of surface area-volume problems.

In other areas scale economies were thought to be substantial for the construction of plants. Rather than aggregating equipment data investment economies were estimated from data by engineers on building costs, equipment costs, and labour involved in installation for complete plants. Out of 221 scale coefficients 186 showed increasing returns. Finally, the median scale coefficient is reported to be .73.

Within operating costs for both labour and management expenses scale coefficients considerably below one were observed. Scale economies were explained by the fact that workers were employed in process plants where typical jobs are watching gauges, adjusting valves, and making repairs. For such tasks expansions in plant capacity generally require a less than proportionate increase in labour input.

Two main difficulties are encountered in engineering studies. First, engineering data generally encompasses only technical aspects of the firms' operations. The accuracy of a production function derived from engineering data then varies inversely with the amount of labour input. Moreover, while the nature of returns to scale may be determined for a particular process or at the plant level this does not apply to the firm. Higher management costs, and selling expenses of the firm may offset increasing returns at the plant level. Second, in combining physical processes the functions must be independent and additive. However, where it is necessary to synchronize the flow of output resulting from different production processes so as to avoid "bottlenecks" such independence may not exist. As a result engineering studies will be most useful where a small number of principle processes determine the basic cost structure of the plant.

Chapter III

Conclusions

Different interpretations may be made of the long run average cost concept. At one extreme long run average cost may be perceived as the relationship of production and distribution costs to size when the ratio of the marginal productivities of the inputs equals the ratio of their input prices. This would require the existence of perfectly homogeneous inputs and outputs, complete divisibility, perfect competition, and an absence of fluctuations in demand.

In engineering studies the cost-output relationship investigated tends toward the strictly theoretical interpretation since capacity is varied while supply conditions, product design, location, and technology are all held constant. Moreover, since the cost curve is obtained independent of demand conditions there is a complete adjustment to present output. However, modifications of simplified conventional price theory are generally necessary to relate to actual conditions of the scale-cost relationships. As another extreme case Stigler suggests that the long run average cost concept relates to the effects of size on the ability to meet the problems of the total economic environment such as inputs available in limited size, obtaining inputs of sufficiently high quality, instability of demand, and ability to successfully differentiate the product through advertising. Between the two extremes one may classify the questionnaire and interview method and statistical production and cost analysis. Use of the questionnaire and interview method has included stochastic economies of scale associated with greater demand stability, qualitative changes in inputs, and factor price changes. In statistical analysis of the production function the specification of the equations exclude qualitative changes in inputs but such changes would enter into statistical cost analysis. Finally, pecuniary economies and pecuniary diseconomies of scale would be included in statistical cost analysis but excluded from the statistical production function.

An economic theorist anxious to maintain the convenience of working with models of pure or perfect competition would not have difficulty in finding grounds upon which to criticize the various cost estimation techniques. With respect to engineering estimates he might point out that non-technical aspects are given insufficient treatment so that the major source of diseconomies of scale, increasing complexity of the managerial function is not included. The findings of statistical production and cost analysis may be criticized on the grounds that the production and cost functions are part of a simultaneous system of equations. The survivor technique may be criticized as having been applied to industries in which the degree of competition is grossly deficient, and in which a multitude of external factors affect survival. Finally, with reference to the questionnaire and interview method our pertinacious theorist might contend that it is ludicrous to ask businessmen if they are efficient, especially those in highly concentrated industries fearing dismemberment by anticombines authorities.

However, evaluating the existing empirical evidence by examining each method of cost estimation in isolation ignores the following important points. If engineering estimates have neglected non-technical aspects such as selling costs this does not apply to the survivor technique. If statistical production and cost analysis or the survivor technique have been unable to indicate the direction of causality in the size-cost relationship or control for the effects of external factors engineering studies fulfill this need. Further, while an absence of competition detracts from the validity of survivor estimates, imperfect mobility and variation of factor prices in cross section analysis of the production function may be an advantage in avoiding the identification problem. Finally, while the questionnaire and interview method may rely upon subjective information the findings of engineering studies are based on the laws of physics and chemistry. In fact it was seen that a combining of different cost estimation techniques has occurred in the case of statistical cost analysis and the questionnaire and interview technique. Moreover, engineering studies and statistical cost analysis were profitably used in conjunction.

The results of testing the hypothesis whereby unit costs decline, reach a minimum, and rise thereafter are generally consistent. Of the studies examined only two instances may be cited as supporting the traditional hypothesis of a U-shaped long run average cost curve.[59] Only study of the relationship between hours of flying time and fuel consumption by Ferguson and halibut fishing by Comitini and Huang indicate the likelihood of diseconomies of scale. While the number of industries studied is small, the results agree well with those of an extensive survey conducted by Wiles where in only 32 percent of the cases was a U-shaped relationship in evidence.[59] One may state then that the U-shaped long run average cost curve generally does not exist although the evidence is often non rigorous in nature.[60]

The results indicate that the theoretical interpretation of long run average cost does not require major revision. Whether changes in factor prices, instability of demand, advertising expenditures, or qualitative changes in inputs were or were not present did not affect the conclusion that the U-shaped cost curve is infrequently encountered. However, no conclusions regarding the importance of indivisibilities are warranted, since the

[59] Wiles, P.J.B. Price, cost, and output. Oxford, Oxford University Press, 1956.
See also Walters, A.A. Production and cost functions: an econometric survey. op. cit. p. 50-51.

[60] This must be qualified to the extent that the questionnaire and interview method was not designed to test the overall shape of the cost function. Also, the fact that small outputs were frequently associated with minimum efficient scale could be argued to increase the likelihood of diseconomies of scale, although costs may also simply level off and become constant.

effects of indivisibilities were included in the results of each cost estimation technique. Thus, it does not follow that the long run average cost curve can be precisely defined in terms of the marginal conditions. Nonetheless, the adequacy of the traditional theoretical analysis is not impaired since the effects of indivisibilities can be analyzed by the use of marginal productivity theory.

Selected Bibliography

1. Bain, J.S. Barriers to new competition. Cambridge, Harvard University Press, 1956.

2. Borts, G.H. The estimation of rail cost functions. Econometrica. 28:108-131. 1960.

3. Barzel, Yoram. Productivity in the electric industry 1929-1955. Rev. Ec. and Stat. 46:395-408. 1964.

4. Chamberlin, E.H. The theory of monopolistic competition. 7th ed., Cambridge, Harvard University Press, 1956.

5. Chenery, H.B. Engineering production functions, Q.J.E. Vol. 63. 1949.

6. Comitini, Salvatore and Huang, David S. Production and factor shares. Journal of Political Economy. 75:366-373. 1967.

7. Ferguson, A.R. Empirical determination of a multidimensional marginal cost function. Econometrica. 18: 217-235. 1950.

8. Ferguson, C.E. Micro-economic theory. Homewood, Irwin Co., 1969.

9. Ferguson, C.E. The neo-classical theory of production and distribution. London, Cambridge, University Press, 1969.

10. Friedman, M. Comment. In Conference on Business Concentration and Price Policy, Princeton, Princeton University Press, 1955.

11. Hahn, F.H. Proportionality, divisibility and economies of scale: two comments. Q.J.E. 62: 132-133. 1948.

12. Haldi, J. and Whitcomb, D. Economies of scale in industrial plants. J.P.E. 75: 373-385. 1967.

13. Johnston, J. Statistical cost analysis. New York, McGraw-Hill, 1960.

14. Johnston, J. Econometric methods. New York, McGraw-Hill, 1963.

15. Klein, L.R. Econometrics. Evanston, Illinois, Row Peterson, 1953.

16. Lomax, K.S. Coal production functions for Great Britain. J.R.S.S. 113: 346-51. 1950.

17. Niitamo, O. The development of productivity in Finnish industry: 1925-1952. Productivity Measurement Review. 15: 1-12. 1958.

18. Solow, R.W. Technological change and the aggregate production function. R. Ec. and Stats. 39: 312-20. 1957.

19. Smith, Caleb A. Statistical cost functions. In Cost Behavior and Price Policy, N.B.E.R., Princeton, Princeton University Press, 1955.

20. Stigler, G. The economies of scale. Journal of Law and Economics. 1:54-71. 1958.

21. Viner, Jacob. In Readings in price theory. Edited by G.J. Stigler and K.E. Boulding. Homewood, Irwin Co., 1952.

22. Walters, A.A. An introduction to econometrics. London, Macmillan, 1968.

23. Walters, A.A. Expectations and the regression fallacy in estimating cost functions. Rev. Ec. and Stats. 42: 210-215. 1960.

24. Walters, A.A. Production and cost functions: an econometric survey. Econometrica. 31: 1-52. 1963.

25. Wiles, P.J.B. Price, cost, and output. Oxford, Oxford University Press, 1956.

26. Wilson, T.A. and Eckstein O. Short run productivity behaviour in U.S. manufacturing. Rev. Ec. and Stat. 46: 41-54. 1964.