The Law of Increasing Return to Scale, Decreasing Average Cost Function and Negatively Sloped Supply Curve

Xiaolin Zhong[1]

Abstract: Among various returns to scale, only the law of increasing return to scale qualifies for general rule. Under this law, the average cost function is a negatively sloped curve. So is the supply curve. The rationale for such supply curve is as follows. As output expands, both price and average cost decrease. As long as the fall of price does not exceed the drop of average cost, it is profitable for the firm to reduce price in exchange for more sales. The firm's profit is maximized at the point where marginal revenue equals marginal cost, total revenue equals total cost, price equals average cost, and marginal price equals marginal average cost. At this point, fixed cost is kept in a level that cannot be reduced without hindering effective administration of production. Fiscal stimulus policy forces the firm to reduce fixed cost, hence production.

Keywords: Return to scale, average cost, supply curve, price, fiscal policy

1 University of Lethbridge, Email: wwzz2@rogers.com

The "law" of diminishing return to scale is the underlying assumption of the positively sloped supply curve of a commodity developed by Alfred Marshall. This curve renders it possible for Marshal to construct the market equilibrium between demand and supply by having it intercept the negatively sloped demand curve. As Mark Skousen indicates, however, under the law of constant return to scale, the supply curve will be a horizontal strait line, and negatively sloped under the law of increasing return to scale, respectively. This invites the question whether market equilibrium still exists with such supply curves (Skousen 2009: 211).

The problem seems to go beyond this point. Marshal's market equilibrium appears to contain some other drawback as well. First of all, it does not incorporate in it the rule of profit maximization, that is, marginal revenue equals marginal cost. It seems to suggest that demand and supply intercept one another at the point where price equals marginal cost. Price, however, equals marginal revenue only if all suppliers are price takers, an implication which causes more difficulties. Such implication does not only contradicts the notion Marshal holds himself that it is the manoeuvre between supply and demand that determines price level, but also brings it into doubt as to how prices are formed, one of the major difficulties the notion of complete competition faces. To overcome this difficulty, some economists have gone so far as to invent an imaginary "arbitrator" who calls the shots, literally turning economics as a science into myth. More significantly, given that no laws of return to scale have been proven to be general, there has yet to be any foundations for constructing any supply curves, hence any market equilibriums. Marsha's market equilibrium therefore is virtually a house of sand fog. Needless to stress how inadequate it is to have such a foundational economic theory remain so problematic. Economists, however, have been reluctant to address this issue, according to Skousen (2009: 211). Consequently, the ground work to rebuild the house has remained in the state of adjournment. As such, I set out to bring this adjournment to an end. Presented here is the result of this endeavour.

The reason for the reluctance of economists, including Marshall himself, to deal with the above problem perhaps rests on the difficulty to conceive a negatively sloped supply curve. Given the conventional belief that higher prices cause supply to increase, which is often easily translated, unconsciously, into the mentality that the firm only responds to higher price with increasing supply, it is indeed hardly conceivable that the firm would respond to falling price with increasing production, despite price discrimination, that is, the more the quantity of goods purchased, the lower the price, is seen everywhere and all the time.

If the origin of the law of increasing return to scale had been identified and therefore the law had been established, this mentality could have been removed. Under this law, average cost continuously decreases as output increases. It can be easily seen that the firm would love to lower price in exchange for higher volumes in sales, as long as the fall of price does not exceed the drop of average cost. The lack of the comprehension of the origin of the law of increasing return to scale, hence the law itself, however, hinders the acceptance of the decreasing average cost by economists as the commonly applied cost function. This is not to say that economists have made no efforts to figure out the origin of the law of increasing return to scale. Adam Smith, for instance, discovered through observing the production process of pins at a factory that a doubling of scale results in a greater division of labour, hence higher productivity of labour. Hal R. Varian presented the case of an oil pipe line. When all inputs are doubled, points out Varian, the output may more than double, for an increase of the area surface of the pipe line by 2 will increase the volume by a factor of more than 2 (1992: 15). The results of these examples, however, can hardly be generalized.

As we can see, Smith and Varian only consider direct inputs in their analysis of the law of increasing return to scale, ignoring indirect inputs altogether. They are not alone in this regard. It appears that no economists have ever paid attention to the relationship between indirect inputs and return to scale. This is perhaps because they see nothing marginal about indirect inputs. The

methodology of marginality is of course sound. And it is true that the change in indirect inputs hardly causes change in output. The absence of indirect inputs, however, renders the view of the whole picture incomplete.

Indirect inputs largely consist of administrative activities, such as management, research and development, finance and accounting, marketing, information technology, etc. There is little doubt that, in modern production, the significance of such activities has been continuously increasing. Some physical facilities, such as non-productive buildings, and warehouses, are also a part of indirect inputs. Their importance in modern production can hardly be ignored. It would not seem reasonable therefore to say that indirect inputs have nothing to do with production, hence return to scale. The reason for economists not to see this, as indicated earlier, might have to do with the difficulty to apply marginal analysis to the change in indirect inputs, for there does not exist numerical relationship between indirect inputs and output. This difficulty can easily be overcome. All it takes is to think about this relationship the other way around. Economists seem to have used to considering the impact of the change in indirect inputs on output and therefore never tried to see things from the other side, that is, the impact of the change in output on that in indirect inputs. If they had, it would have been obvious that, while it is true that a numerical relationship between the changes in indirect inputs and output can be hardly established, it is also true that the change in output causes little change in indirect inputs. As such, indirect inputs do not even nearly double when output doubles. Is this not precisely the definition of the law of increasing return to scale? It thus can be seen that indirect inputs clearly exhibit an increasing return to scale. They are the greatest origin of the law of increasing return to scale, for they exist universally and all the time.

It is worth noting that marginal analysis can be also applied to the change in indirect inputs. While remaining constant in absolute terms, indirect inputs decrease in relative terms. That is, as output increases, the ratio of indirect inputs to each unit of output diminishes. Such ratio is a perfect objet for marginal analysis.

So we have just identified the greatest origin of the law of increasing return to scale. Now we need to turn our attention to direct inputs. Obviously, the total return to scale is determined by the combination of returns to scale for both direct and indirect inputs. Any discussion on return to scale for direct inputs cannot be conducted without connecting them with certain production functions that reflect different technologies. Leontief production function is the best candidate in this regard. This is so because, as indicated in a text book, such technology is vastly employed in the real-world production. Exceptions are rare (Nicholson 1997: 155). Conclusions based on the analysis of Leontief production therefore can be largely generalized.

Leontief production function depicts the technology that the ratios among all direct inputs, such as machine and direct labour hours, production buildings, raw materials, etc., are fixed. As such, the marginal productivity of any single input is zero without the proportional increase of all other inputs. This clearly exhibits a constant return to scale. Given that production without machine is exceedingly rare now days, it seems hardly surprising that this technology appears everywhere in modern production. Even in agriculture production where manual operation still constitutes a significant element of production, the number of labours a given piece of land can contain is still largely fixed. In theory, despite the marginal productivity of labour is diminishing, more labours can still produce more. Thus, it might still be worthwhile to employ more labours in a given piece of land after their marginal productivity begins diminishing. In reality, however, it does not make much sense to do so. Such practice would increase marginal cost. This increase can hardly be compensated by raising prices, for prices can hardly been raised without hurting sales. This is why farmers might not even response to higher prices, for higher prices often lead to diminishing demand. As such, it seems safe to conclude that direct inputs exhibit a constant return to scale. Combining a constant return to scale for direct inputs with an increasing return to scale for indirect inputs, we

prove that the law of increasing return to scale is not *a*, but *the*, general rule.

Now let us consider the kind of cost corresponding to indirect and direct inputs. It should not be too difficult to see that fixed cost and constant marginal cost correspond to indirect inputs and direct inputs, respectively. Indirect inputs are largely constant, hence the cost. And a constant return to scale exhibited by direct inputs entails a constant marginal cost. Leontief production function does not include indirect inputs. This is inevitable. For no numerical relationship can be established between output and indirect inputs, the presence of the latter is not only unnecessary, but disallowable. It would be gravely mistaken, however, to exclude cost of indirect inputs, or fixed cost, from cost functions, for it constitutes a great chunk of entire cost. The cost functions associated with Leontief production function are thus

$$TC=f(Q)=MCQ+F \quad (1)$$
$$AC=f(Q)Q=MC+FQ \quad (2)$$
$$MC = C \quad (3)$$

Where *TC* represents total cost, *MC* marginal cost, *Q* output, *F* fixed cost, and *AC* average cost, respectively.

Equation (2) clearly demonstrates that, when *Q* increases indefinitely, *AC* decreases and approaches *MC* indefinitely. Calculus can also be employed here. By deriving *AC* we have

$$AC'= -FQ2 \quad (4)$$

Clearly, the right side of equation (2) is always negative.

These cost functions entail some interesting extensions of the rule of profit maximization. To demonstrate this we start with the expression of the established rule with the specific total cost function expressed in equation (1) (note that all models of the rule of profit maximization to date are based on general function of *TC)*. Denoting profit, we write profit function

$$\pi=PQ-MCQ-F \quad (5)$$

Deriving π, we get

$$\pi'=P'Q+Q'P-MC \quad (6)$$

The first order condition is met when. Thus, we have

$$P'Q+Q'P=MC \quad (7)$$

The left side of equation (7) is marginal revenue. Replacing it with *TR*' we rewrite equation (7)

$$\int TR'=\int MC \quad (8)$$

This equation gives rise to the first extension of the rule of profit maximization, that is, *total revenue equals total cost*.

From equation (8) we know

$$\int TR'=\int MC \quad (8)$$

Since is *TR*, thus

$$TR=\int MC \quad (9)$$

Since

$$\int MC=MCQ+C \quad (10)$$

We have

$$TR=MCQ+C \quad (11)$$

Mathematically, *C* could be any constant. As equation (5) shows, however, it has to equal the sum of *F* and At the point of maximized profit. At this point is fixed and therefore can be treated as a constant. For mathematical simplicity, we combine it with *F*. Thus,

$$TR=MCQ+F=TC \quad (12)$$

It ought to be emphasized that the justification for combining with *F* goes beyond mathematical simplicity. As an independent entity, the firm has the obligation to deliver to its investors a profit whose ratio to its equity reaches at least the rate of return

to the investment expected by the investors. Given that the equity is largely fixed, the expected profit is a fixed obligation, hence a part of fixed cost.

With equation (12) comes the second extension of the rule of profit maximization, *price equals average cost*. Dividing both sides of the equation by *Q* results in

$$\frac{TR}{Q} = \frac{TC}{Q} \qquad (13)$$

Obviously, the left side of this equation is price and the right side average cost. It can thus be rewritten

$$P=AC \qquad (14)$$

It follows that

$$P' = AC' \qquad (15)$$

We have just seen from equation (15) the third extension of the rule of profit maximization: *marginal price equals marginal average cost*. It can also be derived from equation (7). Rearranging it as follows

$$P'Q=MC-Q'P \qquad (16)$$

From this equation we get

$$P' = \frac{MC-Q'P}{Q} \qquad (17)$$

Since *Q'* equals 1 and *P* equals *AC*, equation (17) can be rewritten

$$P' = \frac{MC-AC}{Q} = \frac{(MC-MC-\frac{F}{Q})}{Q} = -\frac{F}{Q^2} \qquad (18)$$

From equation (4) we know the right side of equation (18) is marginal average cost. The intuition of this extension is obvious. When output increases, both price and average cost fall. There is always more profit to make as long as the fall of price does not exceed that in average cost. At the point where the two become equal, and the fall of price will be greater than that in average cost beyond this point, there can be no more profit to make. Profit is thus maximized at this point.

These extensions clearly indicate that the firm literally reduces price to generate sales, until the fall of price matches that of average cost. Thus, supply curve is downwardly sloped, similar to average cost curve. It meets with demand curve at the point where *P= AV*.

The firm's short-run supply curve intercepts the demand curve at the point where price equals p' and output equals q'. In the long run, as the firm seeks yet more profit, its supply curve moves to S2, and intercepts demand curve at the point where price equals p" and output equals q". These two curves demonstrate clearly that it is the firm's pursuit of profit maximization that grows economy, brings prices down and outputs up.

Now we have just constructed new market equilibrium. It differs from the one developed by Marshal in three significant perspectives, precisely because it irons out the three wrinkles in Marshal's. First, with the proof of the increasing return to scale as the general rule, it is built on a solid foundation. Second, it specifies the way how demand and supply interact to form prices. Third, it is reached precisely at the point where the firm's profit is maximized and, therefore, is consistent with the rule of profit maximization.

As can be expected, this newly constructed market equilibrium has tremendous implications on economic theories, as well as economic policies. It is impossible, however, to discuss all of them in this paper. Considered here is its implication on fiscal stimulus policy. The ultimate cause for recession is a topic that has been debated for decades. The misallocation of resources, however, is the indisputable direct cause for recession. It is obvious that a paper such as this is inadequate to debate it further. Thus, it takes the latter for granted and analyzes the effects of fiscal stimulus policy on it.

When market is efficient, all inputs are transformed into outputs. Denoting *I* inputs and *O* outputs, we have

$$I=O \qquad (19)$$

The consequence of the misallocation of resources is that some of inputs are not transformed into outputs. They are wasted. Outputs therefore are

less than *O*. Denoting *O'* reduced outputs and *W* waste, such a consequence can be articulated as the following equation

$$I=O'+W \qquad (20)$$

Keep in mind that *I* is a constant, for not only inputs are limited, but the productivity to transform them into outputs during a given period is limited as well. From equation (20) we know

$$O'= I-W \qquad (21)$$

It can easily be seen that the less the *W*, the greater the *O'*.

Clearly, *W* has no value and therefore cannot be used to exchange for anything. Consequently, demand for *O* diminishes.

The story does not stop here. Since there is still demand for *O'*, fiscal stimulus policy is to create demand for *W*. For this purpose the state collects taxes from producers of *O'* to purchase *W.* In doing so the state literally transfers resources from *O'* to *W*. The immediate consequence of such policy can be clearly demonstrated by deriving *O'* in equation (20)

$$O'= \qquad (22)$$

That is, resources employed to produce *O'* decreases, dollar for dollar, as stimulus spending increases. These resources are of course employed to produce *W*. This fosters even more misallocation of resources. Consequently, the supply of O' is further reduced, hence demand.

The firm is a living thing. It responds to outside stimuli. As indicated earlier, the firm has to deliver to the investor proper return to his investment. The increased tax burden caused by fiscal policy hurts profit. The firm has no choice but to cut other fixed costs to prevent average cost from rising. At the point of profit maximization, however, all other fixed costs have been kept at the level that cannot be lowered without hindering the effective administration of production. Consequently, the firm has to reduce production, lay off direct labours and dispose productive facilities. Such practice raises average cost. This rise is partially offset by reduced taxes because of decreased profit. The firm will continue cutting other fixed cost and reducing production until the decrease in average cost is entirely offset. Consequently, supply curve moves from SS2 to SS3 and intercept DD2 at point *P'"O"*. In doing so, the firm certainly suffers a drop in profit. It is still be able to, however, maintain proper rate of return to investment, for investment is reduced as well. As we can see, fiscal policy actually causes demand to fall by a magnitude far more grandeur than that of its spending.

We have established that, as general rule, only the law of increasing return to scale exists. This law entails a negatively sloped supply curve, indicating that the firm literally drops price in exchange for more sales. Such practice is profitable for as output increases, average cost decreases. As long as the fall in price does not exceed that in average cost, the firm will continue to do so. As a result, the firm grows economy, brings down prices and raises outputs, hence demand. Stimulus fiscal policy is supposed to create demand. Its consequence, however, is precisely the opposite.

References

Skousen, Mark. *The Making of Modern Economics, 2nd ed*. Armonk: M.E. Sharp, 2009.

Varian, Hal R. *Microeconomic Analysis, 3rd ED.* New York: Norton, 1992.

Nicholson, Walter. *Intermediate Microeconomics and its application*. Fort Worth: The Dryden Press, 1997.