Working Paper Series

Economic Education

No. 001

Infinitesimal Firms and Increasing Cost Industries

Richard M. Peck, Associate Professor
University of Illinois at Chicago

Center for Economic Education

Department of Economics (m/c144)
University of Illinois at Chicago
601 South Morgan Street
Chicago, Illinois 60607
Tel 312-996-2684
Fax 312-996-3344
http://cee.econ.uic.edu

Infinitesimal Firms and Increasing Cost Industries

Richard M. Peck

Economics Department
University of Illinois at Chicago

I wish to thank William Becker, Hirschel Kasper and three anonymous referees for helpful comments that significantly improved this paper.

Abstract

Infinitesimal Firms and Increasing Cost Industries

This article presents a rigorous version of the basic model of an increasing cost competitive industry found in many textbooks. In the model, firms are infinitesimal, which justifies price- taking behavior and a continuous industry supply curve. The industry supply curve slopes upwards because of dispersion in the efficiency of firms. In this framework, the role of the marginal firm is emphasized. This role is not clearly emphasized in many textbook presentations of the increasing cost industry.

Key words: perfect competition, increasing cost industries, infinitesimal firms

JEL codes: A2, D4

Infinitesimal Firms and Increasing Cost Industries

Economic theorists have long regarded the assumption of a continuum of firms as the appropriate way to model the competitive process. This approach was introduced by Aumann[1] (1964, 1966) and further developed by Hildenbrand (1974) and his students (see Trockel 1984). Novshek and Sonnenschein (1979a, 1979b) have also made contributions; indeed this formulation owes much to Hugo Sonnenschein's graduate lecture notes (Sonnenschein 1979).[2] Surprisingly, Aumann's contribution has had little impact on economic education. This is mainly because of the highly technical way in which the subject has evolved; a mathematically rigorous and complete treatment, requiring the integration of correspondences and measure theory, is indeed daunting. Nonetheless, the basic ideas are quite simple, particularly when presented in the context of special cases and examples. This article considers the special case where the available technology uses only one input and there is a continuum of firm types. The examples are accessible to undergraduates who have had two semesters of calculus, that is, are acquainted with integration and the fundamental theorem of calculus.

Why is a continuum of firms necessary? Two problems with the standard Marshallian partial equilibrium model are eliminated when a continuum of firms is assumed. In the standard presentation, firms have nonconvex production technologies, that is, they have U-shaped average cost curves. This implies that each firm's supply curve is discontinuous so that if there are a finite number of firms, the industry supply curve is discontinuous. Equilibrium may fail to exist. Despite this, most intermediate textbooks routinely draw continuous industry supply curves. Second, firms are assumed to be price takers. If there are a finite number of firms, each firm, given the actions of other firms, faces a downward-sloping demand curve for output.[3] Hence the price-taking assumption is not consistent with the economic environment in which firms operate. Both problems disappear when there is a continuum of firms.

Consider the case of an increasing cost competitive industry. Most intermediate textbooks explain long-run upward-sloping supply curves by an appeal to pecuniary diseconomies of scale: Entry pushes up the price of inelastically supplied inputs, which shifts up average cost curves. A second explanation, the one used in this article, is somewhat less common in intermediate textbooks.[4] It assumes there are differences in efficiency between firms. In other words, there is dispersion in the average cost curves of firms and, accordingly, dispersion in the entry price of firms. If all firms have identical average cost curves, then the long-run supply curve is flat, but, if there is dispersion, the supply curve is upward sloping. Of the two explanations for an increasing cost industry, this is the most compelling, because casual empiricism along with more careful studies suggests that within an industry, profit levels often differ across firms. On the other hand, well-documented examples of persistent pecuniary diseconomies are few.

The role of the marginal firm is an important determinant of equilibrium in an increasing cost industry but is often neglected. It emerges clearly in this framework. The role of the marginal firm is slightly paradoxical because the common untutored perception is that equilibrium prices are set by low cost, high efficiency producers. In fact, the competitive equilibrium price equals the average cost of the least efficient producer; the equilibrium price is "just high enough to cover the costs of the highest cost firm that is producing" (Friedman 1990, 321), that is, cover the costs of the marginal firm. Leone (1986, 47) puts it somewhat differently: "It is a basic principle of supply and demand that prices are set to reflect the costs of the marginal producer in an industry - not the highest cost producer, not the lowest-cost firm, but that last or "marginal" producer that can barely justify production at a price consumers are willing to pay." This basic point is absent from most of the textbooks mentioned in endnote four.

Of course, the assumption of a continuum of firms has intellectual and pedagogical costs. Not only is the mathematics more difficult, but the interpretations that one must use are different and unusual; this will become evident as the article proceeds. For example, assuming a continuum of firms means that the number of firms is equal to the number of points in the unit interval which is an uncountably infinite set. This is clearly counter-factual: Whereas the number of firms producing wheat is vastly larger than the number of firms manufacturing cigarettes, it is still a finite number.[5] The standard defense, which dates back to Aumann (1964), is that a continuum of firms captures an "idealized" notion of perfect competition. Aumann argues that the assumption should be taken in the same spirit as the assumption by physicists that a fluid is a continuum, even though it is in fact made up of a large but finite number of particles. Moreover, Aumann pointed out that the assumption of a continuum in economics is less novel than it first appears because economists have routinely assumed, beginning in the nineteenth century, that price and quantity can take on a continuum of values. The continuum approach provides an exact model of perfect competition which, if the model of perfect competition is appropriate, provides an approximation that is useful in certain situations. But the model exacts a price in terms of diminished realism and greater abstraction.

The Basic Model

Each firm produces output using just one input, labor, that is denoted ℓ. There are different types of firms indexed by the parameter β. Firms differ in that they require different amounts of ℓ to produce a given amount of output, that is, firms vary in efficiency. A firm of type β has a production function $f(\ell;\beta)$ and given the price of output, p, and the wage rate, w, determines ℓ to solve:

$$\max\ pf(\ell;\beta) - w\ell$$

The solution to this problem, denoted $\ell(p,w;\beta)$, determines the firm's competitive supply, $s(p,w;\beta)$, that is,

$$s(p,w;\beta) = f\{\ell(p,w;\beta);\beta\}$$

Production functions are assumed to be such that the corresponding average cost curves are U-shaped. Denoting minimum average cost as AC_{min}, this assumption implies that firms have piecewise continuous supply curves, that is, the firm produces zero output when $p < AC_{min}$, but at $p = AC_{min}$, output jumps to Q_{min} and subsequently corresponds to the firm's marginal cost curve for $p > AC_{min}$. This is depicted in Figure 1. The supply function can be written formally as

$$s(p,w;\beta) = \begin{cases} \tilde{s}(p,w;\beta), & if\ p/w \geq \dfrac{\tilde{p}}{w} \\ 0, & if\ p/w < \dfrac{\tilde{p}}{w} \end{cases} \tag{3}$$

where $\tilde{s}(p,w;\beta)$ is nondecreasing and continuous in p and positive for $p/w \geq \widetilde{p/w}$. The expression $\widetilde{p/w}$, which depends on β, is the output price p relative to w, which just allows the firm to break even.

Put Figure 1 about here

The model is completed by assuming that there is a density function $g(\beta)$ of firm types, with $\beta \in [\alpha,\infty)$, $\alpha \geq 0$. Loosely, $g(\beta)$ is interpreted as a measure of the number of firms of type β, or if $g(\beta)$ is a probability density, that is, $g(\beta)$ integrates to unity, then $g(\beta)$ gives the proportion of firms that are of type β. Market supply is determined by adding up the supply of individual firms. Holding w fixed, at a given output price p, the output of β type firms is multiplied by the number of β type firms, $g(\beta)$. With a continuum of firms the appropriate summation is given by the integral

$$S(p,w) = \int_{\alpha}^{\infty} s(p,w;\beta)g(\beta)d\beta \tag{4}$$

If g is a probability density, then $S(p,w)$ is the average or mean supply curve; in this case, industry supply could be obtained by scaling the mean supply curve by an amount that reflects the number or, more formally, the measure of firms in the industry, if this is well defined. If g is continuous and s is piecewise continuous, then S is well defined, and a continuous function of p. We have just noted that there are two separate interpretations of the integral and each has advantages and disadvantages.[6] Operationally, there is little to distinguish the two interpretations. One interpretation views this integral as total market supply,[7] which is a natural generalization of the process of summing individual firm supply curves to obtain the market supply curve. Here, one regards the amount supplied by a single firm as small when compared to the amount supplied by the entire industry, and so, to make sense out of this approach, one thinks of integration as representing a change in scale. In other words, the amount provided by the market, $S(p,w)$, and the amount provided a single firm, $s(p,w;\beta)$, are measured in different, noncomparable units. This approach is consistent with the view taken implicitly by Aumann. By way of analogy, the height of the curve at a given point and the area under the curve are both expressed as real numbers in casual discourse, but implicitly the units are different, the height of the curve is in units of length and the area of the curve is measured in units of length squared. What may make this interpretation problematic for some is that there is no exact empirical counterpart to "change of scale" when it comes to measuring actual industry output. Total output of the widget industry is the sum of the output of individual firms; the output of individual firms and the industry output are typically measured in the same units. Finally, in this interpretation, $g(\beta)$ is regarded as the absolute number of firms of type β, which some may find problematic because typical applications do not restrict the function g to integer values.

Another perspective, associated with Hildenbrand (1974) and his students, is to view $S(p,w)$ as average or mean supply. Most instructors, however, present market demand and supply to their intermediate classes as total, not mean, demand and supply curves, which is a simpler approach that is much more natural because it corresponds directly with data on industry output. A principal advantage is that the approach provides an appealing interpretation of the density function $g(\beta)$ as a proportion of firms of type β. In advocating the mean supply approach, Hildenbrand implicitly argues that individual firm supply should be regarded as finite; he avoids the word "infinitesimal" used by Aumann. But, if there is an infinite number of firms and each supplies a positive amount, then total supply may not be well defined, that is, it may be infinite.[8] To get around this problem, Hildenbrand resorts to average or mean supply, which can be positive and finite even though total supply is infinite. Note that infinite total supply is obviously counterfactual, and is an artifact of the assumption of a continuum of firms.[9]

Assume throughout the remainder of this article that there is a continuous function h such that for all $\beta > h(p/w)$, $s(p,w;\beta) = 0$, and for β greater than α but less than or equal to $h(p/w)$, $s(p,w;\beta) > 0$. The function h identifies the marginal firm, that is, the firm that can just break even at price p/w; for $\beta = h(p/w)$ profits are zero. Firms for which $\beta > h(p/w)$ have not entered the industry because negative profits would result whereas firms with $\beta < h(p/w)$ are producing and are enjoying positive profits. The presence of such a function h means that (4) can be rewritten as

$$S(p,w) = \int_{\alpha}^{\infty} s(p,w;\beta)g(\beta)d\beta = \int_{\alpha}^{h(p/w)} s(p,w;\beta)g(\beta)d\beta \qquad (5)$$

The examples of the next section are of this form. If there is an exogenous increase in p (because of an increase in aggregate demand), and the aggregate supply curve is given by equation (5), then the effect on aggregate supply is given by

$$\partial S(p,w)/\partial p = \int_{\alpha}^{h(p/w)} [\partial s(p,w;\beta)/\partial p]g(\beta)d\beta + s(p,w;h(p,w))g(h(p/w))\partial h(p/w)/\partial p \qquad (6)$$

This equation is determined by applying Liebnitz's rule to equation (5) which states how to differentiate an integral with respect to a variable that appears in both the integrand and in the limits of integration.[10]

Expression (6) was first derived in Novshek and Sonnenschein (1979a) and has a straightforward interpretation. The expression indicates that price changes affect industry supply on two different margins. The total effect of a higher p is shown in Figure 2 as the sum of areas A and B. The first right-hand term of equation (6) is always nonnegative and represents the response of firms which are already in the industry to a rise in price from p' to p''. This "output effect" is marked as region A in Figure 2.[11] The second term, also nonnegative, arises because as the relative price of output rises additional firms will enter, that is, some firms producing zero output will now choose to produce positive quantities. This "entry effect" is marked as region B in Figure 2. The decomposition of the industry response into an output and entry effect is well known; for instance, Milton Friedman (1962, 78) states, "The actual expansion in supply ... is in general a result of both expansion in the output of each firm separately and an increase in the number of firms."

Put Figure 2 about here

One attractive feature of this approach is that it is possible to geometrically represent the aggregate supply response. The industry supply can be geometrically represented as the area under the surface $s(p,w;\beta)g(\beta)$ depicted in Figure 3; at a given price industry supply is the area of a slice under the curve perpendicular to the p axis. In other words, with w fixed, Figure 2 is a slice from the three-dimensional graph shown in Figure 3. As price, p, varies, the supply response is the area of the corresponding slice, the size of which is determined by the supply response of active firms and the number of active firms. When p increases the slice is taken at a point further from the origin and there are two effects: The slice becomes taller as firms produce more output (the output effect), and the slice becomes wider as more firms enter the market (the entry effect).

Put Figure 3 about here

Equilibrium p, of course, is determined by the interaction of demand and supply, that is,

$$D(p) = S(p,w) \tag{7}$$

where $D(p)$ is aggregate demand. In this formulation, if a firm of type β' unilaterally alters its output level, aggregate supply remains unchanged. In Figure 4, firms of type β' have decreased their output level, generating a discontinuity at point β'. The resulting curve in Figure 4 shows the supply response of all firms. It is piecewise continuous but the area under this curve is defined and, what is more important, is unchanged from when the discontinuity is not present.[12] This is because the weight or measure assigned to β' is zero. Because aggregate supply is not altered, equation (7) is satisfied and the equilibrium price is unchanged. Hence, firms are indeed price takers.

Put Figure 4 about here

Examples

These examples demonstrate the application of the framework and illustrate the points that I have made. There are two things to note. First, in the examples, the supply curves of individual firms are discontinuous because of fixed costs, but the industry supply curve is continuous. In other words, the process of summing the actions of a continuum of firms smooths out the industry supply curve. Formally, this comes about as a consequence of Richter's Theorem (Hildenbrand 1974), but our examples do not require the invocation of this theorem. Intuitively, the approach works "because the discontinuity in the behavior of any individual [firm] has infinitesimal weight in the aggregate"(Arrow and Hahn 1971, 182). Second, these examples illustrate the important role that marginal firms, or entry effects, can play in determining the shape of the industry supply curve.

Example One:

The production of a type β firm is given by

$$f(\ell;\beta) = \begin{cases} 1, & \text{if } \ell \geq \beta \\ 0, & \text{if } \ell < \beta \end{cases} \tag{8}$$

A graph of this production function for a given value of β is shown in Figure 5. Note that as β increases, more of the labor input is required to produce one unit of output, which means that a firm with a higher β is less efficient than a firm with a lower β. The individual supply curve is

$$s(p,w;\beta) = \begin{cases} 1, & \text{if } \beta \leq p/w \\ 0, & \text{if } \beta > p/w \end{cases} \tag{9}$$

This is determined for a firm β by setting profits, which are given by the expression $p - w\beta$, equal to zero and solving for β to get $\beta = p/w$. If p/w exceeds β then the firm is earning positive profits and produces one unit of output. If p/w is less than β, then the firm earns a negative profit if it produces one unit and so chooses, instead, to shut down. What is noteworthy about this example is that at the level of an individual firm, there is little flexibility: a firm either shuts down or inelastically supplies one unit of output.

Put Figure 5 about here

If $g(\beta) = 1/\beta^2$, with support $[1,\infty)$ then

$$S(p,w) = \int_1^{p/w} 1(1/\beta^2)d\beta = -1/\beta \Big|_1^{p/w} = 1 - w/p \tag{10}$$

This example illustrates in a striking manner how the supply curves of individual firms can be discontinuous, yet the aggregate supply curve is not only continuous but differentiable. Second, even though each individual firm has a perfectly inelastic supply curve, the industry supply curve is upward sloping. Here, entry effects alone are sufficient to smooth out the aggregate supply curve and ensure that it is upward sloping. Put another way, marginal firms are playing a significant role in the determination of the market supply function. Note that the slope of the supply curve, w/p^2, declines as p rises. This is because the number of firms entering as the price rises is declining, that is, it reflects the attenuated density of low efficiency firms.

Example Two: Suppose the production function of a β type firm is given by

$$f(\ell;\beta) = \begin{cases} 2(\ell-\beta)^{1/2}, & \text{if } \ell \geq \beta \\ 0, & \text{if } \ell < \beta \end{cases} \tag{11}$$

The β can be thought of here as a fixed labor cost that is required before the firm can begin producing positive output. Thus less efficient firms require a larger fixed input of labor, that is, have a larger β, than more efficient firms. This production function gives rise to a U-shaped average cost curve and the following supply curve:

$$s(p,w;\beta) = \begin{cases} 2p/w, & \text{if } \beta \leq p^2/w^2 \\ 0, & \text{if } \beta > p^2/w^2 \end{cases} \tag{12}$$

For a specified p/w, the marginal firm has β equal to p^2/w^2. More efficient firms, those with β less than p^2/w^2, have entered the industry, whereas less efficient firms, those with β greater than p^2/w^2, are producing zero output, that is, they have not entered the industry.

If the distribution of firms is given by $1/(4\beta^{3/2})$, with support $[1,\infty)$, then the aggregate supply is

$$S(p,w) = \int_1^{p^2/w^2} \{2p/w\}(1/4\beta^{3/2})\,d\beta = \frac{p}{w} - 1 \tag{13}$$

The slope of the supply curve is $1/w$. Using equation (6), the slope decomposes into an entry effect which is $1/p$ and an output effect which is $1/w - 1/p$. As p increases, the entry effect goes to zero, and the output effect approaches $1/w$.

Conclusion

The continuum approach has distinct advantages and disadvantages. The prime advantage of the approach is that it allows the presentation of a tight, coherent and complete model of perfect competition. In particular, the assumption of a continuum of firms means that firms are true price takers and that the industry supply curve is continuous, even though individual firms are Marshallian, that is, have U-shaped average cost curves. In addition, the approach captures the notion that marginal firms and the process of entry play a significant role in determining the shape of the industry supply curve, which, of course, is another significant part of the Marshallian partial equilibrium story. The model outlined in this article is, in some sense, a more rigorous and complete version of what most of us teach in intermediate theory courses. A prime pedagogical disadvantage of the continuum approach is that even the stripped down version presented here requires the use of calculus. The conceptual disadvantage is that there is a sacrifice of realism: the model assumes an uncountable number of firms and, as we noted, the interpretation of industry supply itself is somewhat novel and potentially problematic. To provide a final perspective, though, we give Solow the last word: "All theory depends on assumptions which are not quite true. That is what makes it theory" (Solow 1956, 65).

NOTES

1. The use of a continuum of players in a game theoretic context dates back to Shapley (1953).

2. A textbook treatment is available in Ellickson (1992), Chapter 3.

3. See Stigler (1966) for a demonstration. Of course, in the limit as the number of firms increases, the demand curve faced by individual firms becomes infinitely elastic.

4. Browning and Browning (1992), Frank (1994), Hyman (1993), Mansfield (1991), Pindyck and Rubinfeld (1994) explain increasing cost industries as arising from pecuniary diseconomies. David Friedman (1990), Landsburg (1992) and Pashigian (1995) along with the industrial organization textbook of Carlton and Perloff (1994) use efficiency differences between firms to explain increasing cost industries.

5. Robinson (1934) touches on these issues in an early, very illuminating article. "Competition can only be absolutely perfect, given rising marginal costs, if the number of firms is infinite. Absolute perfection of competition is therefore an impossibility" (p. 119).

6. Aumann (1964, 1966) and Hildenbrand (1974) focus primarily on demand; the different approaches to supply outlined here are implicit in their work.

7. From a technical perspective, Aumann's (1964, 1966) framework is also consistent with Hildenbrand's mean demand approach because Aumann identifies the set of agents with the closed unit interval. However, the interpretation provided by Aumann is more in keeping with the interpretation suggested here.

8. This point is made in Chapter 4 of Hildenbrand and Kirman (1976).

9. Another reason for interpreting (4) as mean supply, aside from obtaining a finite integral, is

that mean supply is generally better behaved mathematically than total supply. In particular, mean supply may be continuous and the underlying mean production set may be convex. For a brief but informal discussion on this issue see Trockel (1984, 30-32).

10. A statement and proof of Liebnitz's rule can be found in Bartle (1964, 245-246).

11. Novshek and Sonnenschein (1979a) call this the "substitution effect."

12. See Lang (1983, 211 - 226) for a discussion of the integrals of piecewise continuous functions.

REFERENCES

Arrow, K., and F. Hahn. 1971. *General competitive analysis*. San Francisco: Holden Day.

Aumann, R. J. 1964. Markets with a continuum of traders. *Econometrica* 32 (January-April): 39-50.

Aumann, R. J. 1966. Existence of competitive equilibria in markets with a continuum of traders. *Econometrica* 34 (January): 1-17.

Bartle, R. 1964. *The elements of real analysis*. 2d ed. New York: John Wiley and Sons.

Browning, E., and J. Browning. 1992. *Microeconomic theory and applications*. 4th ed. New York: Harper Collins.

Carlton, D., and J. Perloff. 1994. *Modern industrial organization*. 2d ed. New York: Harper Collins.

Ellickson, B. 1993. *Competitive equilibrium*. Cambridge, United Kingdom: Cambridge University.

Frank, R. 1994. *Microeconomics and behavior*. 2d ed. New York: McGraw Hill.

Friedman, D. 1990. *Price theory*. 2d ed. Cincinnati, OH: South Western.

Friedman, M. 1962. *Price theory: a provisional text*. Chicago: Aldine.

Hildenbrand, W. 1974. *Core and equilibria of a large economy*. Princeton, NJ: Princeton University Press.

Hildenbrand, W., and A. Kirman. 1976. *Introduction to equilibrium analysis*. New York: North Holland.

Hyman, D. 1993. *Modern microeconomics: analysis and applications*. 3d ed. Homewood, IL: Irwin.

Lang, S. 1983. *Undergraduate analysis.* New York: Springer Verlag.

Landsburg, S. 1992. *Price theory and applications*. 2d ed. Fort Worth, TX: Dryden.

Leone, R. 1986. *Who profits?* New York: Basic Books.

Mansfield, E. 1991. *Microeconomics*. 8th ed. New York: Norton.

Novshek, W., and H. Sonnenschein. 1979a. Supply and marginal firms in general equilibrium. *Economic Letters* 3 (No. 2): 109-14.

Novshek, W., and H. Sonnenschein. 1979b. Marginal consumers and neoclassical demand theory. *Journal of Political Economy* 87 (December): 1368-82.

Pindyck, R., and D. Rubinfeld. 1994. *Microeconomics***.** 3d ed. Englewoods Cliff, NJ: Prentice Hall.

Pashigian, P. 1995. *Price theory and applications*. 1st ed. New York: McGraw Hill.

Robinson, J. 1934. What is perfect competition? *Quarterly Journal of Economics* 49 (November): 104-20.

Shapley, L. 1953. Stochastic games. *Proceedings of the National Academy of Sciences* 39 (October): 1095-100.

Solow, R. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70 (February): 65-94.

Sonnenschein, H. 1979. Unpublished Lecture Notes: Economics 501. Princeton, NJ: Princeton University.

Stigler, G. 1966. *The theory of price*. 3d ed. New York: Macmillan.

Trockel, W. 1984. *Market demand: an analysis of large economies with non-convex preferences*. Lecture Notes in Economics and Mathematical Systems 223. Berlin: Springer Verlag.

Varian, H. 1992. *Microeconomic analysis*. 3d ed. New York: Norton.

FIGURE 1

Average and Marginal Cost Curves of the Typical Firm

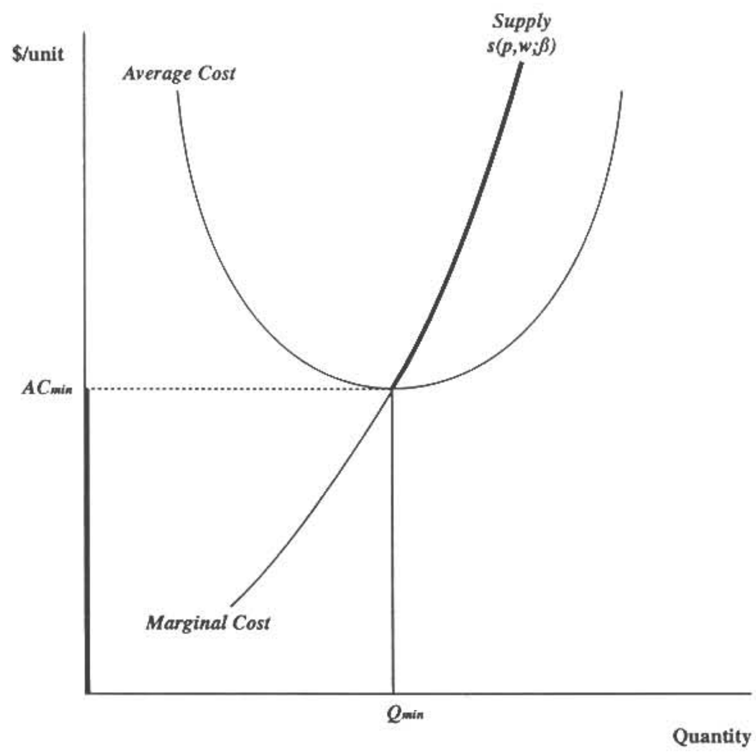

FIGURE 2

The Effect of a Price Increase on Market Supply

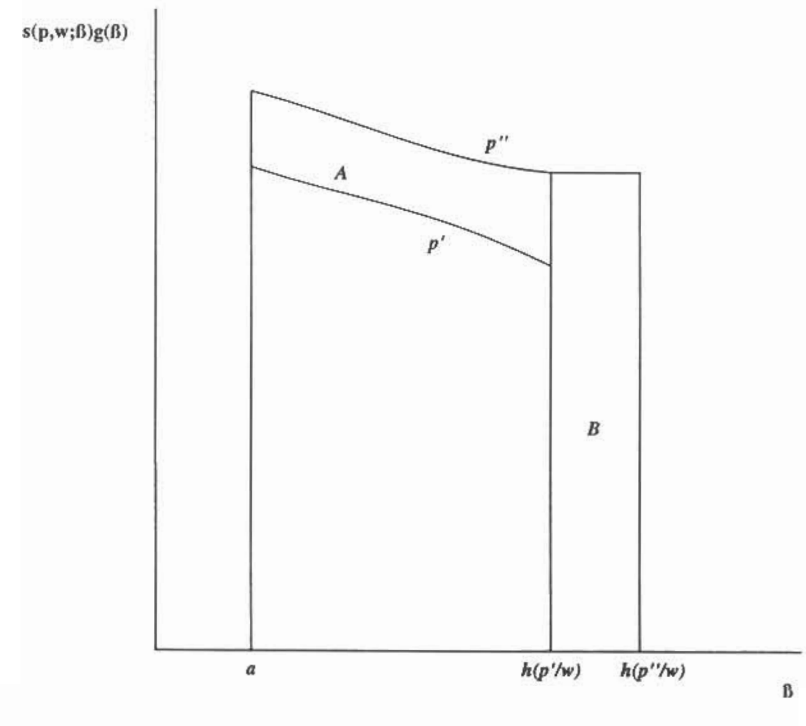

FIGURE 3

Market Supply "Slices"

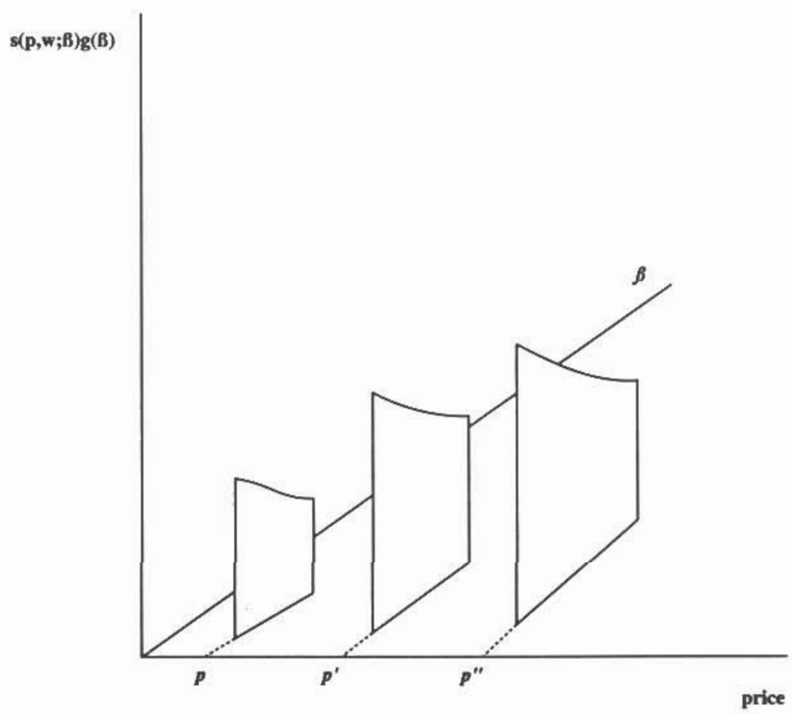

FIGURE 4

The Negligible Effect of a Single Firm on Market Supply

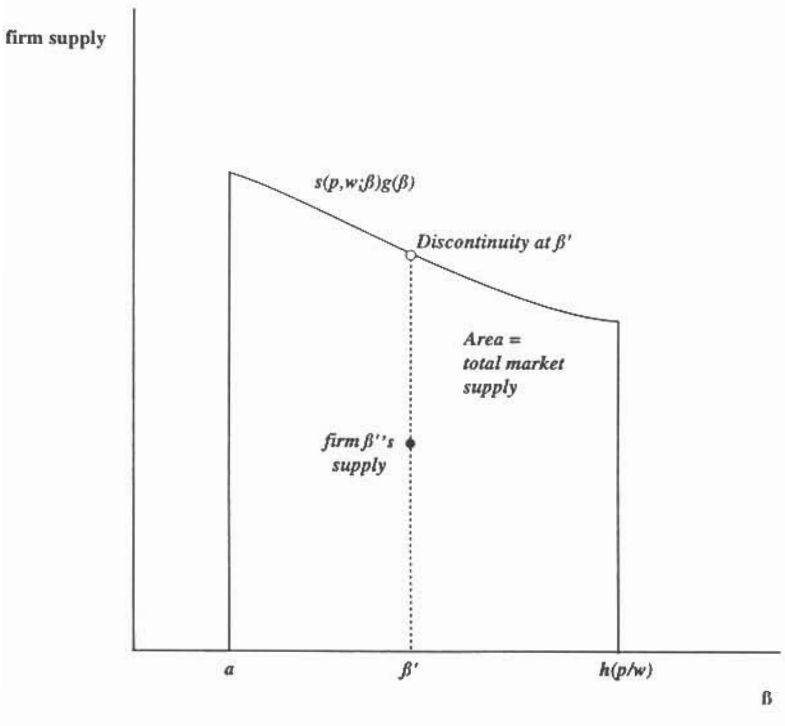

FIGURE 5

Production Technology of Example One