1

The Averch-Johnson Model of Rate-of-Return Regulation

1.1 Purpose

Averch and Johnson (1962) initiated one of the earliest and most influential investigations into the effects of regulation on the behavior of a regulated firm. They argue that the most prevalent form of regulation currently applied to public utilities, rate-of-return regulation, induces the firm to engage in inefficiencies. These inefficiencies are the natural result of the regulation, in that a firm that is attempting to maximize profits is given, by the form of the regulation itself, incentives to be inefficient. Furthermore, the aspects of monopoly control that regulation is intended to address, such as high prices, are not necessarily mitigated, and could be made worse, by the regulation.

Averch and Johnson conducted their analysis within a relatively restrictive model that abstracts from many real-world issues. Their model and conclusions have been questioned from a number of perspectives,[1] and, in fact, some errors in their logic have been discovered (though these errors do not affect their essential conclusions).[2] Their work is nevertheless invaluable, not only for its specific conclusions but, more generally, because it introduces a fundamental criterion for evaluating regulatory mechanisms plus a method for applying this criterion. In the case of rate-of-return regulation, their method shows that the regulatory procedure does *not* induce the firm to choose the socially optimal outcome. However, the method can be used to identify other types of regulation that do.

---

[1] For example, Bailey and Coleman (1971), Davis (1973), Klevorick (1973), Joskow (1974), Bawa and Sibley (1980), and Logan et al. (1989) show that the inefficiencies are mitigated or even eliminated when the analysis is changed to allow for a time lag in the regulatory process. In these models, the firm, during the period between price reviews, takes its price as given and retains whatever profit it earns. The firm therefore has less incentive to produce inefficiently than when, as in Averch-Johnson's model, the firm's profits are constrained continuously. Similarly, various authors (see note 3 to the introduction) show that different results obtain if the firm is assumed to maximize some variable other than profits (such as output or return on shareholder equity).

[2] Takayama 1969; Baumol and Klevorick 1970, p. 168.

The following sections describe the Averch-Johnson (or A-J) model and its implications for rate-of-return regulation. Section 1.2 describes the behavior of an *unregulated* firm, using a method that facilitates comparison with the firm's behavior when regulation is imposed. Section 1.3 defines rate-of-return regulation, identifying exactly the form of regulation that is imposed on the firm. Section 1.4 determines the behavior of a firm that is subject to this rate-of-return regulation and compares this behavior with that of an unregulated firm and with the behavior the regulator would like to induce. Throughout, the discussion draws on clarifications of the A-J model provided especially by Zajac (1970), Baumol and Klevorick (1970), and Bailey (1973).

Before entering the substance of this and each subsequent chapter, we summarize the major results and conclusions in the chapter introduction. This summary provides both a preview of what is to follow and a concise reference for later review. The statements will not always be completely clear prior to reading the chapter itself. However, on returning after completing the chapter, the reader may find the summary a useful reminder and guide.

The findings of chapter 1 can be summarized as follows. Under rate-of-return (ROR) regulation, the firm is allowed to earn no more than a "fair" rate of return on its capital investment. The firm is free to choose its price, output level, and inputs as long as its profits do not exceed this fair rate. We show that this form of regulation provides perverse incentives that operate against optimality.

Suppose first that the regulator sets the fair rate of return above the cost of capital. In this case:

- The regulated firm will utilize more capital than if it were unregulated.
- The regulated firm will use an inefficiently high capital/labor ratio for its level of output. That is, the firm's output could be produced more cheaply with less capital and more labor.
- It is possible that the firm will produce less output and charge a higher price than if it were not regulated.
- The firm might increase its output above the level it would produce if not regulated. However, the firm will always produce in the elastic

portion of demand. That is, ROR regulation will not induce the firm to expand output so far that it moves into the inelastic portion of demand. Insofar as the optimal output is in the inelastic portion of demand, ROR regulation cannot induce the firm to produce the optimal output (and may, as noted in the previous point, induce the firm to reduce output).

- There is one bright spot. Contrary to popular notions, a firm under ROR regulation will not waste capital. The firm will produce as much output as possible given its inputs. The firm will choose an inefficient mix of inputs (this is the second point above), but it will use the inputs that it has chosen efficiently.

In short, ROR regulation with the fair rate of return above the cost of capital induces the regulated firm to use an inefficient input mix and does not necessarily induce it to increase output.

Suppose instead that the regulator sets the fair rate of return equal to the cost of capital. In this case, the regulated firm becomes indifferent between many possible outcomes, and its choice is indeterminant. In particular, the firm would earn the same profit whether it increased or decreased output, used an efficient or inefficient input mix, and wasted inputs or not. In fact, the firm would make the same profit if it closed down and sold off its capital. Because the firm's profits are the same in each of these cases, the firm is as likely to choose one as the other. Consequently, ROR regulation with the fair rate set at the cost of capital cannot be relied upon to induce the firm to act in any particular way.

Suppose finally that the fair rate is set below the cost of capital. In this case, the firm makes more profit by shutting down and selling its capital than by remaining in operation. If it is legally able to do so, the firm will choose this option. Otherwise, it will reduce its capital as much as possible, which could result in less output and a higher price.

The overall picture is quite damaging. The basic problem with ROR regulation is that it provides incentives based on the amount of capital that the firm invests, whereas the goal of the regulator is not to increase capital per se. The regulator's goal is to induce the firm to increase output, decrease price, and produce at minimum cost. Other forms of regulation that match incentives more closely to the regulator's goals are more likely to be successful. Some of these are explored in chapter 2.

To derive our results regarding ROR regulation, we first consider the behavior of an *unregulated* firm. We then examine how behavior changes when the firm is subjected to ROR regulation.

1.2 Behavior of the Unregulated Firm

To describe the behavior of an unregulated firm, a method and some terms are employed that are somewhat different than those used in standard microeconomics textbooks. While the behavior of the firm is the same as in the standard presentation (that is, the firm behaves the same in either case), this alternative representation facilitates analysis of how the firm's behavior changes when regulation is imposed. In fact, part of the value of the A-J model is the development of this alternative way of describing the behavior of the unregulated firm, which generalizes more readily than the standard method to situations with regulation.

Consider a monopolist that produces only one output (such as electricity), the quantity of which is denoted Q. Assume, for convenience, that the firm produces this output with only two inputs, capital K and labor L. The price of capital (interest rate) is r per unit, the price of labor (wage rate) is w, and the firm takes these input prices as given.

The input possibilities of the firm are summarized by the familiar isocost-isoquant mapping, as in figure 1.1. The axes represent the two

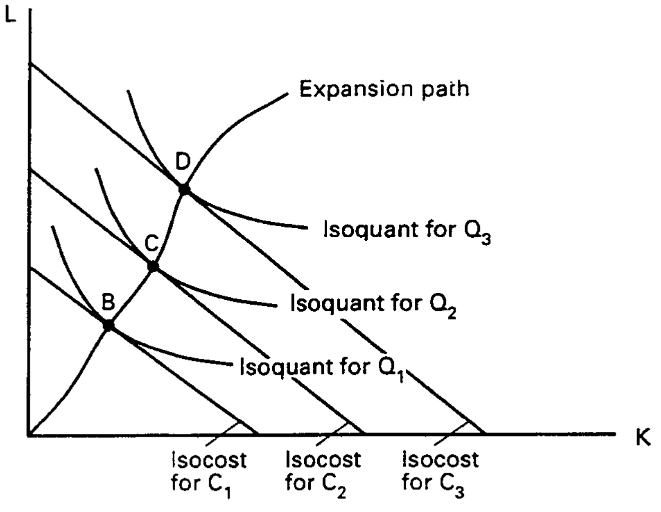


Figure 1.1
Isocost-isoquant mapping

inputs, such that each point in the graph denotes a certain quantity of each input. An isocost line is defined as a set of input combinations that cost the same amount. For example, all of the K and L combinations on the isocost for C_1 cost exactly C_1; that is, the input levels times their prices sum to C_1: $rK + wL = C_1$. Rearranging, we find the equation for the isocost line as $L = (C_1/w) - (r/w)K$, namely, a line with a y-intercept of C_1/w and a slope of $-r/w$. There is an infinite number of isocost lines, one for each possible level of cost. They all have the same slope and differ only in their distance from the origin (i.e., their y-intercepts). Higher isocost lines (that is, those further from the origin) represent higher costs. Three isocost lines are shown in the graph representing three levels of costs: $C_3 > C_2 > C_1$.

An isoquant is defined as the set of input combinations that can be used to produce a given level of output. For example, output level Q_1 can be produced using any of the input combinations on the isoquant for Q_1.[3] The shape of the isoquant depends on the technology available to the firm, as summarized in its production function. The slope of an isoquant at any point is the negative of the marginal rate of technical substitution (*MRTS*), which is the extra quantity of one input that must be used to continue producing the same level of output if one unit of the other input is foregone. For our inputs, *MRTS* is the amount of extra labor required to continue producing the same level of output with one less unit of capital. The isoquants bend away from the origin, such that *MRTS* decreases as more capital and less labor is used to produce a fixed level of output. This shape reflects diminishing marginal product of inputs.

An infinite number of isoquants exist, one for each possible level of output. Because more inputs are required to produce more outputs, "higher" isoquants represent greater levels of output. Three of these isoquants, representing increasing levels of output $Q_3 > Q_2 > Q_1$, are shown in figure 1.1.

Given that the firm is producing a certain level of output, the firm minimizes its costs (and, because revenues are fixed by its output level, maximizes profits) by choosing the input combination at which the isoquant for that level of output is tangent to an isocost line. Sup-

---

[3] It is always possible to produce less than maximally possible (that is, to waste inputs). Consequently, output level Q_1 can be produced with any input combination either on or beyond the isoquant, where "beyond" means more of either input than a point on the isoquant. To account for this fact, an isoquant can be more precisely defined as the set of input combinations whose maximal output level is a given quantity.

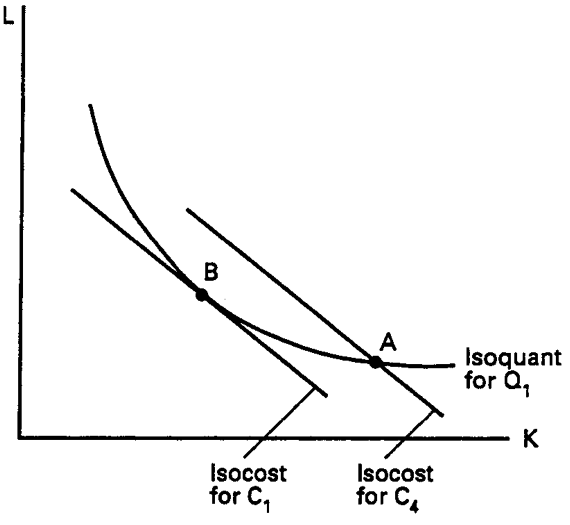


Figure 1.2
Cost-minimization occurs at point of tangency

pose, for example, that the firm is producing output level Q_1 using an input combination at which the isoquant is *not* tangent to the isocost line, say point *A* in figure 1.2. The firm's costs are C_4, because point *A* is on the isocost line for input combinations that cost C_4.

The firm can reduce its costs by moving from point *A* to point *B*. At point *B*, output is the same (because *A* and *B* are on the same isoquant) but costs are lower: the isocost line through *B* is closer to the origin than that through *A*. This same result is obtained whenever the isoquant is not tangent to the isocost line: the firm will be able to reduce costs by moving to the point at which they *are* tangent. When at this point of tangency (point *B* in our graph), the firm has the lowest possible costs for its level of output.

For each level of output, there is one input combination that is cost minimizing, that is, at which the isoquant is tangent to the isocost line.[4] Because there are numerous possible output levels, there are numerous input combinations that are cost minimizing for *some* level of output. Consider the set of all such input combinations. This set is called the expansion path. For output level Q_1 in figure 1.1, input combination *B* is cost minimizing; *C* is cost minimizing for Q_2; and *D* is cost minimizing for Q_3. Connecting these points plus all the other points that are obtained for other levels of output gives the expansion path.

---

[4] If the isoquant has a linear segment, there can be more than one cost-minimizing point. However, we will ignore this possibility because it does not affect the basic results.

There are numerous ways to view, or define, the expansion path:

1. The expansion path is the set of input combinations that are cost minimizing for some level of output. Production at any point that is *not* on the expansion path is inefficient.

2. The unregulated profit-maximizing firm will necessarily choose an input combination on the expansion path. *Which* combination on this path is chosen depends on what level of output the firm chooses to produce. The term "expansion path" denotes the idea that as the firm expands its output, it moves along the expansion path. For example, when expanding its output from Q_1 to Q_2 and further to Q_3, the firm moves along the expansion path from *B* to *C* to *D*.

3. The slope of each isocost line is $-r/w$ and the slope of the isoquant at any point is the (negative of) *MRTS* at that point. Because the expansion path consists of input combinations at which the isocost line is tangent to the isoquant, *MRTS* equals r/w at each point on the expansion path. The expansion path can therefore also be defined as the set of input combinations at which $MRTS = r/w$.

The analysis of isocosts and isoquants provides *some* information about the firm's choice of inputs: it demonstrates that the firm chooses an input combination on the expansion path. However, it does not allow a complete determination of the input choice of the firm. In particular, *which* of the points on the expansion path does the firm choose? This cannot be determined with isocosts and isoquants alone.

The firm's choice among the various input combinations on the expansion path is equivalent to its choice of output. Given an output level, the firm chooses the input combination on the expansion path that corresponds to that output level. For example, in figure 1.1, if the firm chooses output level Q_2, then it also chooses input combination *C*. Conversely, once the firm chooses an input combination, its output level is determined: if the firm chooses input combination *C*, then its output is Q_2. Because of this equivalence, the firm's behavior is fully described (that is, its input and output levels are completely determined) by its choice of where to locate along the expansion path.

There is a direct relation, elaborated below, between the input combination the firm chooses and the profits it can earn. This relation

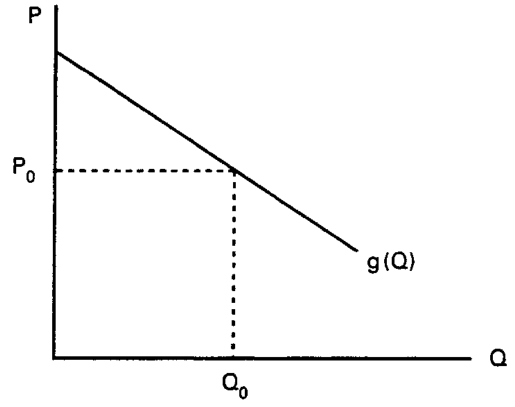


Figure 1.3
Demand curve of firm

provides the information required to locate the firm on the expansion path. The relation between inputs and profit is based on the fact that, once the firm chooses an input level, the maximum profits it can earn are set. To see this, start with inputs. With a given level of inputs, the firm can produce a certain maximum quantity of output. This quantity is given by the firm's production function: $Q = f(K,L)$. The maximum price at which the firm can sell this output depends on the demand for the firm's output. The demand function, denoted $P = g(Q)$, gives the maximum price that the firm can charge and sell quantity Q of output. For example, in figure 1.3, at quantity Q_0 the maximum price the firm can charge is P_0: if it tried to charge a higher price, it would not be able to sell all of its output.[5] Given the output level denoted by the production function and the price denoted by the demand function, the firm's profits are fully determined.

The relation between inputs and profits can be shown functionally. Profits, π, are the difference between revenues and costs:

$$\pi = PQ - rK - wL.$$

Substituting in the demand function for P:

$$\pi = g(Q)Q - rK - wL.$$

---

[5] The demand function also can be considered in its inverse form as giving the quantity demanded at each price. Though it is sometimes easier to think of demand in this latter way, it is more fundamental and, in our analysis, more useful to consider price as a function of quantity—that is, the maximum price at which the firm can sell a given quantity.

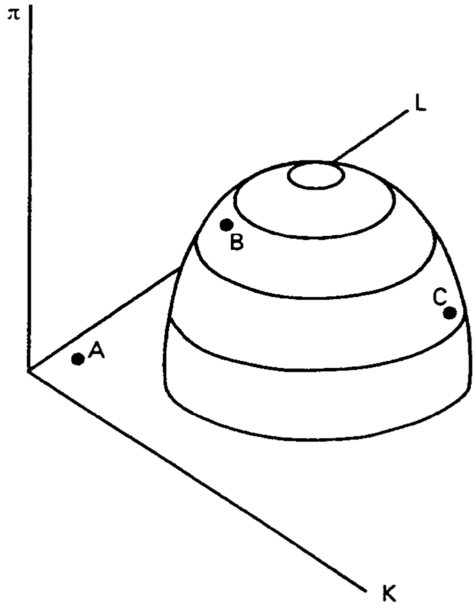


Figure 1.4
Profit hill

Substituting the production function for Q:

$$\pi = g(f(K,L)) \cdot f(K,L) - rK - wL,$$

which is a function of K and L only. Profits depend only on the levels of inputs, such that profits are determined once input levels are determined. (Stated succinctly, the argument is simply: given inputs, output is determined; given output, price is determined; and given inputs, output, and price, profits are determined.)

Because profits are set once input levels are chosen, profits can be expressed directly as a function of inputs only: $\pi = h(K,L)$. This function takes the general shape of a hill, as shown in figure 1.4, reflecting the fact that profits increase at first and then decrease as the use of inputs expands. Consider a low level of inputs, say point A. A firm must usually incur setup costs that are independent of the level of output; that is, it must use some inputs before it is able to produce any output. As a result the firm usually loses money at low levels of inputs, because the revenues that can be obtained from the small amount of output produced are insufficient to cover the setup costs. If more inputs are used, as at point B, output is higher and the firm is able to earn positive profits. Profits increase as the scale of production (the quantity of inputs, and hence, output) expands. However,

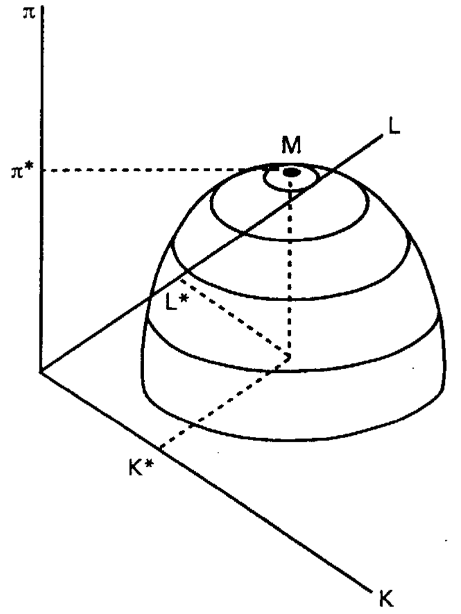


Figure 1.5
Chosen input levels for unregulated firm

this rise in profits does not continue indefinitely. When the firm expands its inputs (and, hence, its output), it must lower its price in order to sell the additional output. Eventually, the firm must lower its price so much that profits drop (the market becomes saturated, in a sense). For example, at point *C*, with more inputs (and, hence, output) than *B*, profits are lower. Because profits increase and eventually decrease as input levels increase, the relation between profits and inputs is called the "profit hill."[6]

The behavior of the firm can be visualized with this profit hill. The firm chooses the inputs that give it the highest possible profits. These are the input levels associated with the top of the profit hill, that is, the inputs that provide the greatest profit. In figure 1.5, the firm chooses input levels K^* and L^* and makes profits π^*, which is the top of the profit hill. Given these inputs, the firm produces output $Q^* = f(K^*, L^*)$ and sells it at price $P^* = g(Q^*)$.

We want to combine the information contained in the profit hill with the isocost-isoquant mapping so as to locate the firm's chosen

---

[6] This relation is sometimes called the profit function. However, this latter term is more widely used to denote the relation, important in duality theory and econometrics, between profits and the prices of inputs and output. The term "profit hill" avoids confusion, while also signifying the function's shape.

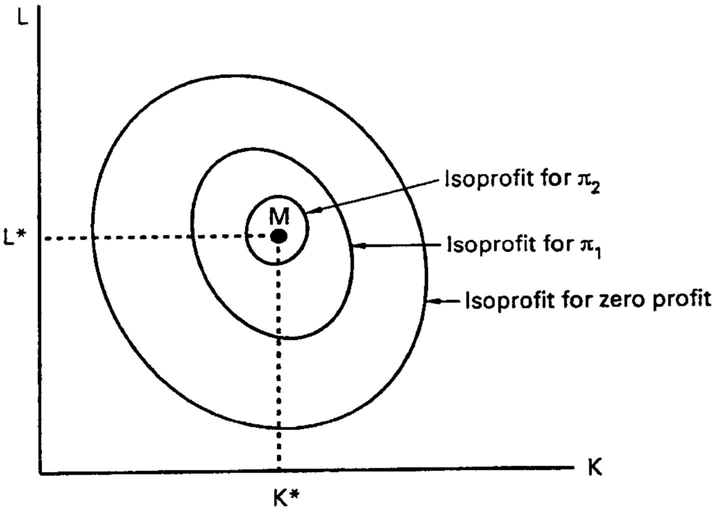


Figure 1.6
Isoprofit contours

point on the expansion path. To do this, we make a topological map of the profit hill. That is, we represent the three-dimensional profit hill in two dimensions, as in figure 1.6. Each contour on this map denotes a given level of profit. Each of these contours is called an "isoprofit contour," defined as the set of input combinations that result in a given level of profits. For example, all of the input combinations on the isoprofit contour for π_1 result in the firm earning π_1 profits. The isoprofit contours are concentric, with inner contours representing higher profits than outer contours. The outermost contour is the base of the profit hill, representing zero profits; it is called the zero-profit contour.[7] The top of the hill is point *M*. The firm chooses this input combination.

To locate the firm along the expansion path, the isoprofit contours are superimposed on the expansion path, as in figure 1.7. The firm chooses point *M*,[8] which represents capital level *K**, labor *L**, and

---

[7] The contours could be extended outward to represent various levels of negative profits. Except for a few situations, these negative-profit contours are not relevant.

[8] Point *M* is necessarily on the expansion path. If it were not, then there would be another point at which profits are higher than at *M*, namely the point on the expansion path where the isoquant through *M* intersects the expansion path. (At this point, revenues are the same because both points are on the same isoquant, but costs are lower because this point is on the expansion path and *M* is not.) Because *M* is defined as the top of the profit hill, there can be no point with higher profits; hence *M* must be on the expansion path.

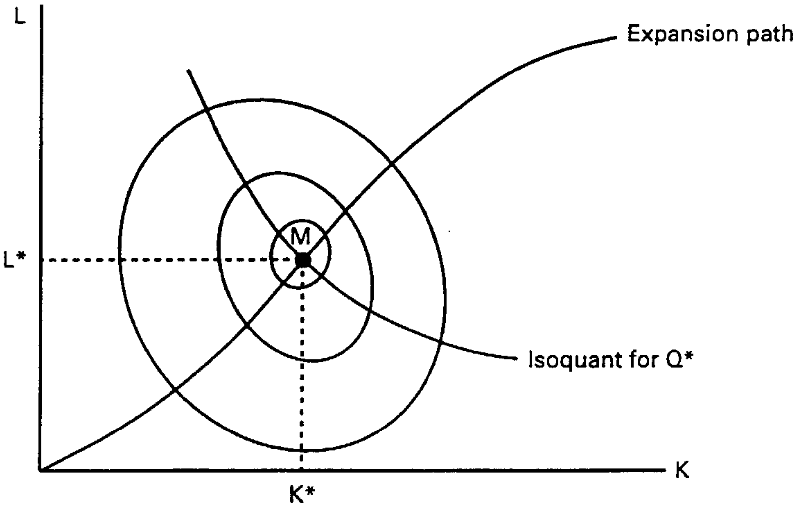


Figure 1.7
Isoprofit contours superimposed on expansion path

resultant output level Q^*. The input and output choices of the firm are now fully determined.

It is important to note, especially in relation to the effects of rate-of-return regulation, that at these chosen input and output levels, marginal revenue is positive. Recall from microeconomics that marginal revenue is the extra revenue the firm obtains from expanding output by one unit. Suppose that marginal revenue is negative. This means that expanding output decreases revenues, or, conversely, that reducing output increases revenues. If marginal revenue is negative, the firm would earn more profit by reducing its output: revenues increase and, because less output is being produced, costs decrease. With higher revenue and lower costs, profits are higher.

Whenever marginal revenue is negative, the firm will decrease its output. Consequently, the firm's final choice of output necessarily occurs at a point where marginal revenue is positive. Generally, marginal revenue is positive for lower levels of output and is negative for sufficiently higher output levels. The reason for this is clear. Marginal revenue consists of two components: (1) the extra revenue that the firm obtains from the sale of one extra unit of output, *minus* (2) the loss in revenue that occurs because the firm has to reduce its price to sell the extra unit. Because the price reduction applies to all goods sold, the size of the second component depends on the output level

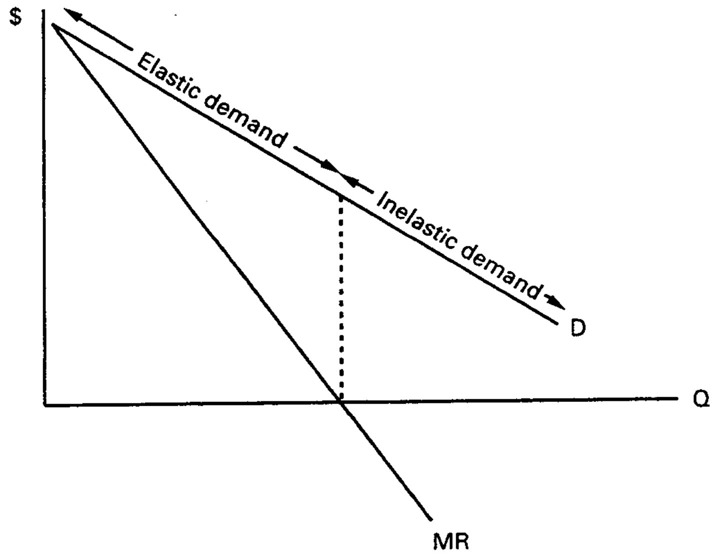


Figure 1.8
Relation of marginal revenue to elasticity of demand

of the firm. For low levels of output, the second component is small, such that the first component (which is positive) exceeds the second (which is negative), for a net effect that is positive. For sufficiently high levels of output, the second component exceeds the first, for a negative net effect.

Figure 1.8 illustrates the situation. Marginal revenue is positive at first, indicating that the firm can increase its revenues by selling additional output. Eventually marginal revenue becomes negative, such that the firm's revenues would decrease if it attempted to sell additional output. The fact that marginal revenue is positive at the firm's chosen output level means that the firm will never expand output beyond the range for which marginal revenue is positive.

We can relate these concepts to the elasticity of demand. Marginal revenue is positive when the elasticity of demand is greater than one (in magnitude). Suppose, for example, that elasticity is two, indicating that a 1% increase in price results in a 2% decrease in quantity demanded. Stated equivalently, a 2% expansion of output requires a 1% decrease in price. Because total revenue is price times quantity, a 2% increase in output coupled with a 1% decrease in price results in an increase in total revenue (the increase in quantity is greater than the decrease in price). Therefore, expanding output increases revenue. Similarly, if elasticity is less than one, marginal revenue is negative.[9]

At low levels of output, the elasticity of demand generally exceeds one, such that marginal revenue is positive. This is called the elastic portion of demand and is labeled as such in figure 1.8. Eventually, at higher levels of output, elasticity falls below one and marginal revenue becomes negative. This is called the inelastic portion of demand.

The fact that marginal revenue is positive at the firm's chosen output level means, equivalently, that the firm will never increase its output beyond the elastic portion of demand. That is, the firm will never choose to operate in the inelastic portion of demand, where marginal revenue is negative.

We can use another method to demonstrate that an unregulated firm will necessarily produce in the elastic portion of demand. This alternative proof ties more readily to analogous statements, made below, about the behavior of a regulated firm. We use proof by contradiction. Suppose the top of the profit hill, which is the firm's chosen point, is in the *inelastic* portion of demand. We can show that this supposition leads to a contradiction and consequently cannot occur. Figure 1.9 illustrates the situation. The isoquant that is shown is for the output level at which marginal revenue is zero. Points on the expansion path below this isoquant represent production in the elastic portion of demand (where elasticity exceeds one and marginal revenue is positive), while points above the isoquant represent production in the inelastic portion of demand. Point *M* is placed above the isoquant to represent the supposition that the top of the profit hill occurs in the inelastic portion of demand. We will show that this is not possible, that *M* cannot be above the isoquant for zero marginal revenue.

Consider point *J*. Profits at *M* necessarily exceed those at *J*, because *M* is the top of the profit hill. However, because marginal revenue is negative along the expansion path past the designated isoquant, revenues are higher at *J*, which represents less output, than at *M*. Costs are lower at *J* than *M*, because output is lower. Consequently, profits at *J* exceed profits at *M*. Because this contradicts the fact that profits at *M* exceed profits at *J*, the situation depicted in figure 1.9 is impossible. Point *M* can only occur in the elastic portion of demand, on the

---

[9] Consider an elasticity of one-half, meaning that a 1% increase in price results in a .5% decrease in quantity demanded. Stated alternatively, a .5% increase in quantity necessitates a 1% decrease in price. Because price drops by more than the rise in quantity, total revenues decrease.

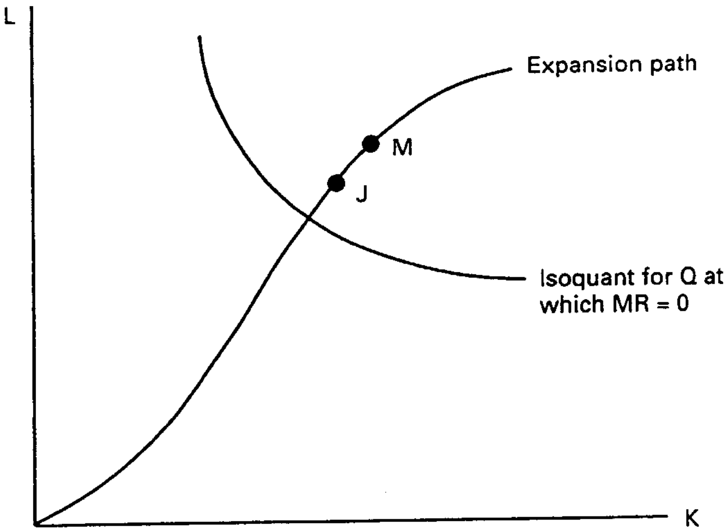


Figure 1.9
Unregulated firm will move to elastic portion of demand

expansion path below the isoquant associated with zero marginal revenue.

1.3 Rate-of-Return Regulation

Under rate-of-return regulation, the regulated firm is allowed to earn a "fair" return on its investment in capital, but is not allowed to make profits in excess of this fair rate of return. The firm can freely choose its levels of inputs, its output level, and its price as long as the chosen levels do not result in profits in excess of the fair return.[10]

The rate of return on capital is defined as revenues minus costs for noncapital inputs, divided by the level of capital investment. With only one noncapital input, L, the rate of return is $(PQ - wL)/K$. This rate must, by the terms of ROR regulation, be no greater than the fair rate of return, labeled f, that the regulator has previously announced.[11] Therefore, the firm can choose any K, L, Q, and P as long as

---

[10] In reality, the regulator has oversight control over the firm's choices and can, for example, disallow costs for unneeded inputs. The A-J model's characterization of ROR regulation abstracts from this aspect of real-world regulation. The extent to which the regulator can effectively exercise this oversight function is unclear. (If the regulator could identify precisely efficient input and output levels, there would be no need to have ROR regulation: the regulator could simply mandate the efficient levels.) The A-J model can be viewed as a worst-case situation, in which the regulator is unable to distinguish between efficient and inefficient behavior. The lessons from this model can be expected to hold to a degree when the regulator has some but less than perfectly effective oversight ability.

$$f \geq (PQ - wL)/K.$$

The firm must operate in a way that satisfies this inequality, that is, that does not result in too high a rate of return.

Economic profits, or what economists call excess profits, are the difference between the firm's revenues and its costs for *all* inputs, including capital: $\pi = PQ - wL - rK$.[12] The maximum return the regulated firm is allowed to earn can be expressed in terms of economic profit. Subtract the price of capital from both sides of the above inequality and rearrange:

$$f - r \geq ((PQ - wL)/K) - r;$$

$$f - r \geq (PQ - wL - rK)/K;$$

$$f - r \geq \pi/K.$$

$$\pi \leq (f - r)K.$$

That is, the maximum (economic) profit that the firm is allowed to earn is $(f - r)K$.

If the fair return is 10% and the price of capital is 8%, then the firm is allowed to earn no more than 2% of its invested capital. For example, if the firm invests $100 million, it is allowed to earn no more than $2 million in profits.[13]

---

[11] The regulator establishes the fair rate on the basis of a variety of factors. For the A-J model, the rate is assumed to be established prior to the choices of the firm and not adjusted on the basis of the firm's choices.

[12] If the firm borrowed its capital, economic profits are the profits it earns after making the interest payments on the borrowed capital. If the firm uses its own capital, it incurs an opportunity cost per unit of capital, which is the return the firm could obtain by lending out the funds. In this case, economic profits are the profits the firm earns after it subtracts out the profits it could obtain by lending the money.

[13] In this example, K is measured in dollars and r in percentages, when it is more customary to measure inputs in physical units and price in dollars per physical unit. While less intuitive, the example can be reworded in terms of the more traditional measuring conventions for inputs. Capital is 100 million units. The fair rate of return is set at 10 cents per unit of capital. The price of capital is 8 cents per unit, meaning that the firm must pay 8 cents in interest for each unit of borrowed capital or can receive 8 cents for each unit of lent capital. The firm is allowed to make economic profits of 2 cents per unit of capital, or a total of $2 million dollars.

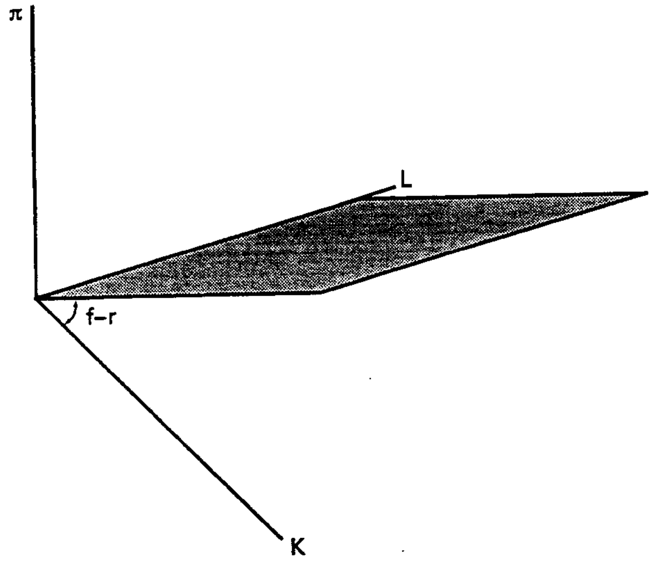


Figure 1.10
Constraint plane

1.4 Behavior of the Regulated Firm

ROR regulation restricts the options of the firm. If unregulated, the firm can choose any K, L, Q, and P. Under ROR regulation, the firm can choose only among those levels that do not result in profits in excess of the allowed amount. That is, there is a constraint on the behavior of the regulated firm. Our goal is to determine the behavior of the firm under this constraint.

Assume for now that the fair rate of return exceeds the price of capital, $f > r$, such that the allowed economic profit is positive for any positive amount of capital. The other possibilities, with f equaling r and r exceeding f, are considered later in the chapter.

The economic profits the firm is allowed to earn are represented graphically in figure 1.10. Recall that the firm is not allowed to make more than $(f - r)K$ profit. This amount is represented by a plane that is hinged on the L-axis and increases linearly with K, with a slope of $(f - r)$. As K increases, the firm is allowed to earn more profits in absolute terms; that is, the plane increases in K. For example, if the fair rate exceeds the price of capital by 2%, the firm is allowed to earn a maximum of \$200,000 in economic profits if \$10 million in capital is utilized (invested) and \$400,000 if \$20 million is utilized. (The *rate* of

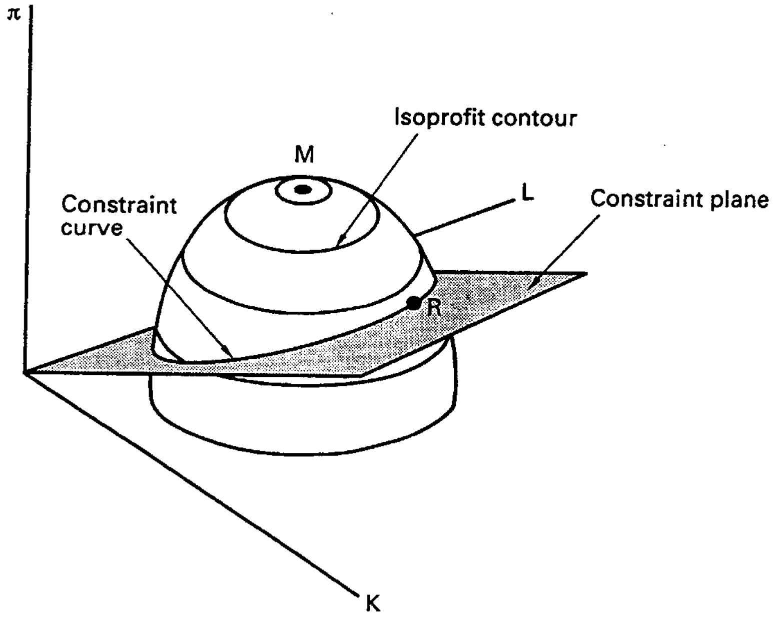


Figure 1.11
Constraint plane and profit hill

return is the same, but the absolute profits are higher.) The amount by which allowed profits increase for each extra unit of capital is $f - r$, such that the slope of the plane is $f - r$. If K is zero, allowed profit is also zero. Furthermore, allowed profits are not affected by the amount of labor that the firm uses; for a given K, the firm is allowed to make the same rate of return on this K no matter how much L it uses. The plane is therefore hinged on the L-axis. (The amount of labor the firm uses affects the profits that the firm is *able* to earn, but does not affect the amount of profits it is *allowed* to earn.)

The firm is not allowed, by the terms of ROR regulation, to make profits in excess of that represented by the plane in figure 1.10. To reflect this fact, the plane is called the "constraint plane," because the firm is constrained to make profits that are on or below this plane.

Given its technology (as embodied in the production function) and the demand for its product, the maximum profits the firm is able to earn at any input combination are given by the profit hill. Figure 1.11 shows both the profit hill and the constraint plane. The profit hill represents the profits the firm is *able* to earn given technology and demand, while the constraint plane depicts the profits that the firm is *allowed* to earn. To distinguish these two concepts of profit, the maximum profit the firm is able to earn given its technology and demand is called the "feasible" profit, while the maximum profit the firm is allowed to earn under its regulation is called the "allowed" profit. Feasible profit is given by the profit hill, and allowed profit is given by the constraint plane.

The constraint plane slices through the profit hill. The parts of the profit hill above the constraint plane correspond to input combinations with which feasible profits exceed allowed profits. The profits that are available to the regulated firm, that is, that are both feasible given technology and demand and allowed by the regulator, are given by the "sliced-off" profit hill: the part of the profit hill that remains after the part above the constraint plane is removed.

The firm maximizes its profits by choosing the highest point on the sliced-off profit hill; or, stated more accurately, by choosing the input combination that provides the greatest profits on the sliced-off profit hill. This is point *R* in figure 1.11.[14]

The exact location of point *R* can be visualized more readily when the profit hill and constraint plane are shown in two dimensions, on the *K–L* graph. Consider the intersection of the profit hill and the constraint plane in figure 1.11. The set of input combinations at which this intersection occurs is called the constraint curve and is mapped on the *K–L* graph of figure 1.12. The input combinations on the constraint curve are those with which the profits that the firm can feasibly earn given demand and technology are the same as the profits that the firm is allowed to earn. With any input combination inside the constraint curve, the firm can feasibly earn more than it is allowed to earn. If the firm chooses one of these input combinations, it must waste resources in some way (that is, produce less than is maximally possible) in order not to exceed the allowed profit level. With any input combination outside the constraint curve, the maximum profits the firm can feasibly earn given demand and technology is less than the allowed profits. That is, the firm is allowed to earn more profits than it is able to earn at these input combinations.

For each input combination on the constraint curve, the firm earns

---

[14] As stated above, the unregulated firm chooses the input combination associated with the top of the profit hill, point *M*. If point *M* is not above the constraint plane (that is, if the constraint plane does not slice off the top of the profit hill), the regulated firm also chooses point *M*, behaving the same as if it were not regulated. In this case, there is essentially no regulation. We assume that the regulated firm is truly regulated such that the unconstrained profit maximizing point is not allowed under regulation.

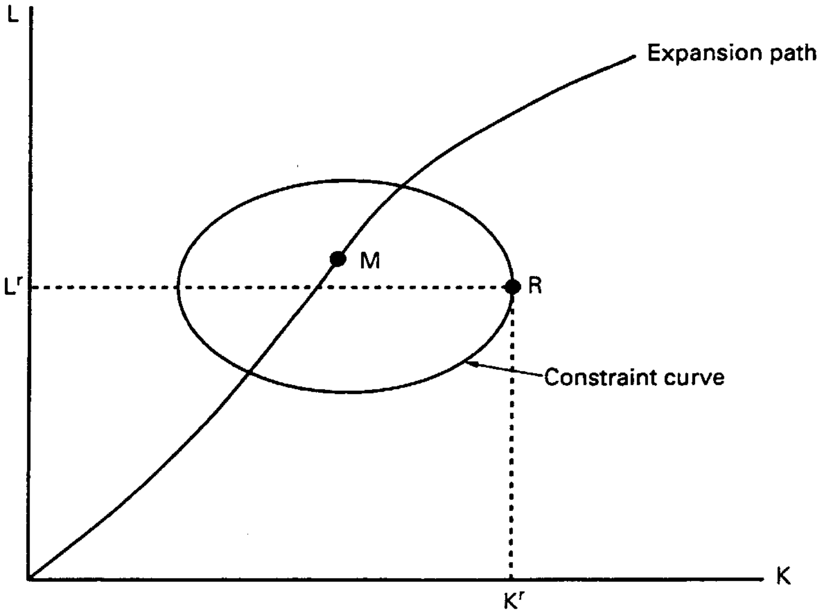


Figure 1.12
Constraint curve

the same rate of return, the fair or allowed rate. However, the absolute level of profits increases as the firm utilizes more capital. That is, points farther to the right on the constraint curve (representing more *K*) represent greater absolute profits than those farther to the left, though profits as a rate of return is the same. For example, a 2% return on $100 million is $2 million profits, while on $10 million it is only $200,000.

The firm chooses the input combination that results in the greatest absolute profits. This is the point on the constraint curve, labeled *R*, with the greatest amount of capital, that is, farthest to the right on the *K–L* graph. Any other point on the constraint curve provides less profit, because any other point represents the same rate of return but on a smaller amount of capital.

Point R also represents more absolute profits than any point inside or outside the constraint curve. At all points inside the constraint curve, the firm is utilizing less capital (all these points are to the left of *R*). The firm can feasibly earn a higher rate of return at these points than at *R*; however, the firm is not allowed to earn a higher rate of return. The firm is allowed to earn the same rate of return as at *R*, but, because the rate is applied to a smaller quantity of capital, absolute profits are lower.

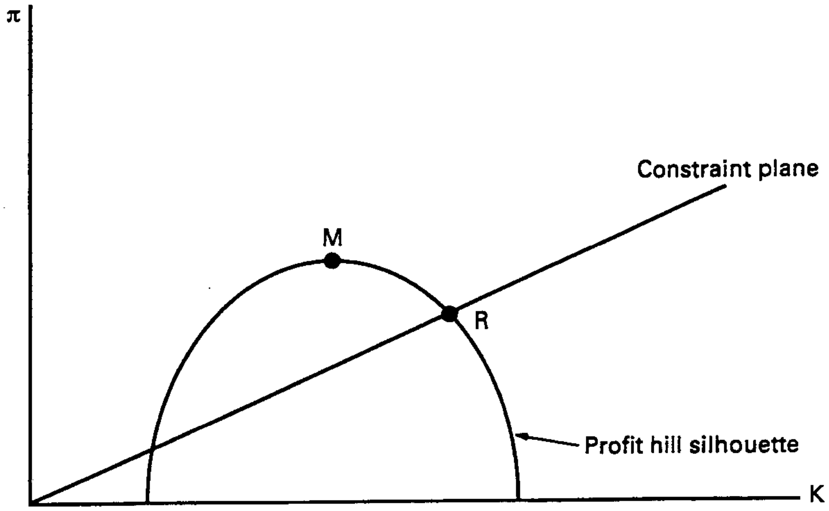


Figure 1.13
Constraint plane and profit hill

For points outside the constraint curve, the firm cannot feasibly earn as high a rate of return as the regulator allows. For points to the left of K_R, the firm earns a lower rate of return on a smaller amount of capital than at *R*, such that its absolute profits are clearly lower. For points to the right of K_R, the firm earns a lower rate of return but on a larger amount of capital. Recall, however, that point *M* is the top of the profit hill and that profits drop as the firm moves down the profit hill away from *M* in any direction. All points to the right of *R* are farther from *M*, and hence farther down the profit hill, representing less profit than at *R*.

Another graph is useful for visualizing the relation of the profit hill to the constraint plane. In figure 1.13, the profit hill and constraint plane are represented in the dimensions of *K* and π, with the *L* dimension suppressed. The profit hill in this graph gives the maximum profits that are feasible at each level of capital if labor is adjusted appropriately. That is, it is the silhouette of the profit hill as viewed from the K-π plane. The constraint plane shows the maximum allowed profit for each level of capital. As capital increases, the firm is allowed to make greater absolute profits, though the rate of return is the same. The regulated firm chooses point *R*, that is, the point on the intersection of the constraint plane and the profit hill where capital is greatest. If the firm were to increase its use of capital beyond this point, it would be allowed to earn more profits, but it would not be able to. That is, the profits that are feasible for the firm to earn given technology and demand would decrease from using more capital even though its allowed profits increase.

Using the fact that the regulated firm chooses the input combination on the constraint curve that has the most capital, several important results can be shown.

Result 1: The regulated firm uses more capital than the unregulated firm.

This result is essentially definitional at this point. In using the terms "regulated firm" and "unregulated firm," we refer to two firms that are exactly the same except that one is subject to ROR regulation and the other is not. Equivalently, we can think of the same firm when it is under regulation compared to when it is not. Furthermore, the firm is considered regulated only if the constraint plane passes below (that is, cuts off) the top of the profit hill; otherwise the regulation would not be effective and the firm would behave the same as when unregulated. For a regulated firm, therefore, the constraint curve in figure 1.12 must encircle point M, such that the point on the constraint curve with the greatest amount of capital, point R, is necessarily to the right of point M, and therefore represents more capital than M.

The impact of regulation on the firm's use of labor, on the other hand, is not definite. Depending on the shape on the profit hill, the regulated firm can use either more, less, or the same amount of labor as the unregulated firm. The three possibilities are shown in figure 1.14.[15]

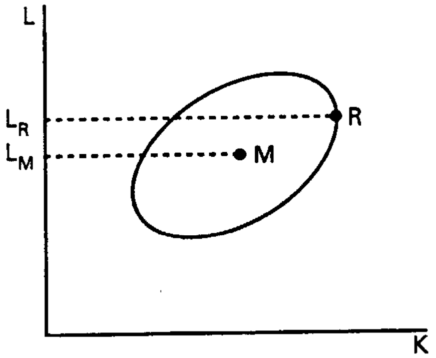
Result 2: The capital/labor ratio of the regulated firm is inefficiently high for its level of output. That is, the output that the regulated firm produces could be more cheaply produced with less capital and more labor than the regulated firm chooses.

This result is the primary, and most famous, conclusion of the A-J model. The term "A-J effect" has come to be known as the bias induced by ROR regulation toward using too much capital relative to labor. To demonstrate the result, we first consider what the result implies about the position of the firm's chosen input combination relative to the expansion path, and then show why this relative position is necessary.

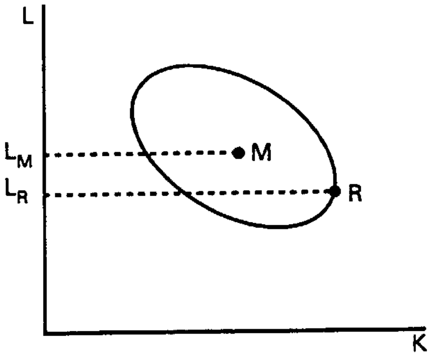
Figure 1.15 illustrates a situation that conforms to the result. The

---

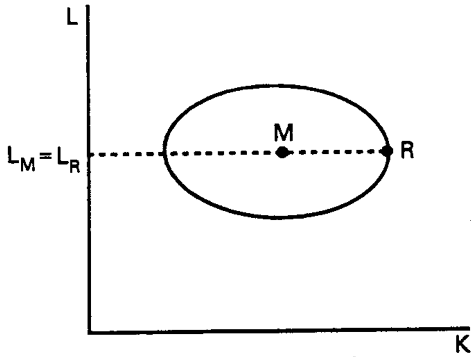
[15] While the firm might increase its use of labor, as in the first panel of figure 1.14, there is a limit on this increase. This limit is an implication of result 3 and is discussed below.



More labor under regulation



Less labor under regulation



Same labor under regulation

Figure 1.14
Impact of ROR regulation on firm's use of labor

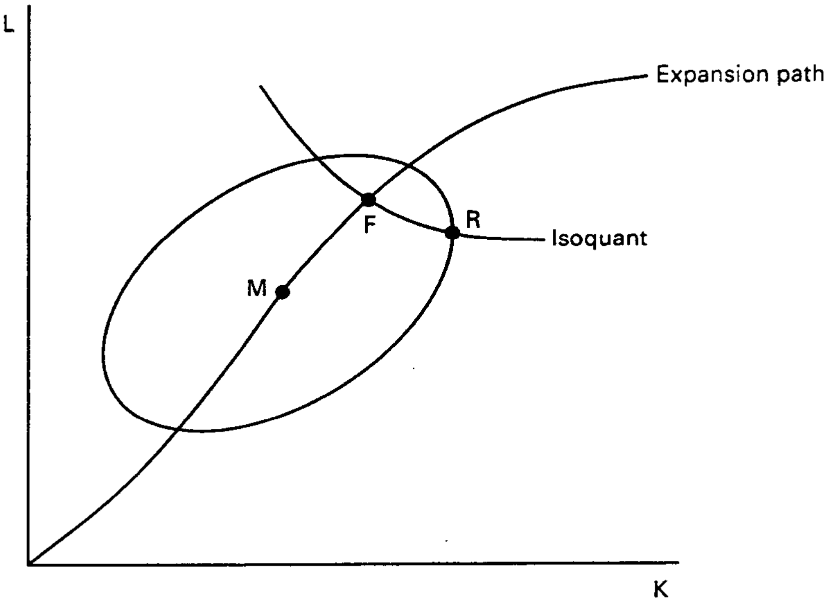


Figure 1.15
Regulated firm's *K/L* ratio is inefficiently high

regulated firm chooses input combination *R*. With this input combination, the firm produces the level of output given by the production function. The isoquant through *R* gives the set of input combinations that can be used to produce the same level of output as is produced at *R*. This isoquant intersects the expansion path at *F*. By definition of the expansion path, costs are lower at *F* than at any other point on the isoquant, including *R*. Point *F* represents greater use of labor and less use of capital than at point *R*. The cost of producing the regulated firm's output could therefore be reduced by using more labor and less capital. Stated another way, the regulated firm's capital/labor ratio is inefficiently high: the firm uses too much capital relative to labor for its level of output.

In figure 1.15, the regulated firm's chosen input combination, *R*, is "below" the expansion path. Result 2 states that this always occurs. To demonstrate the result, therefore, we must show that the regulated firm will never choose a point above or on the expansion path.

Consider figure 1.16 in which the constraint curve and expansion path are drawn such that the firm chooses a point *above* the expansion path. In this case, the firm would choose an inefficient input mix, but the inefficiency is in the opposite direction than stated in the result:

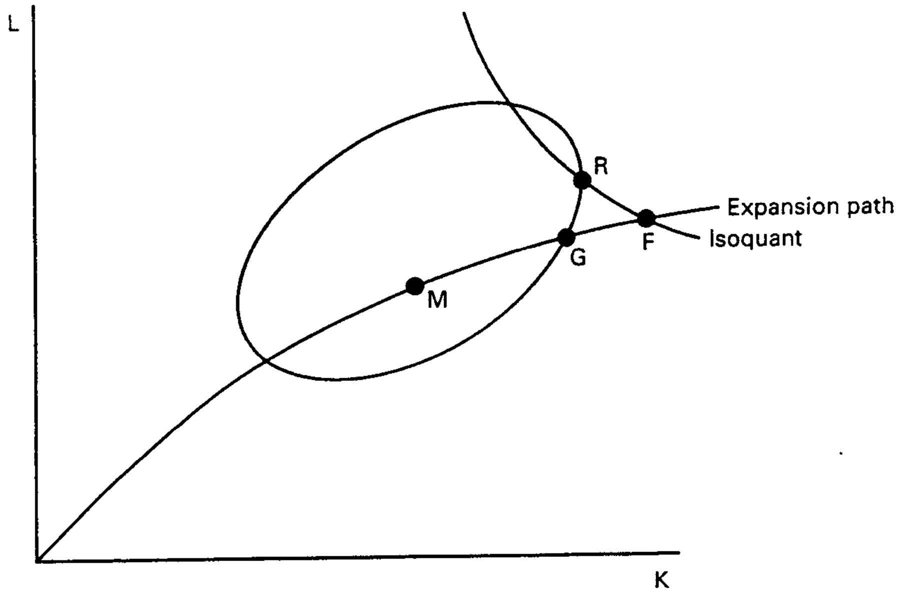


Figure 1.16
Regulated firm's *K/L* ratio is inefficiently low: an impossible situation

the firm uses an inefficiently *low* capital/labor ratio. We can show that this situation is impossible. Consider point *G*, where the expansion path intersects the constraint curve. Absolute profits at *R* are necessarily greater than at *G*: $\pi_R > \pi_G$. This is true because the rate of return is the same at both *R* and *G*, because they are both on the constraint curve representing the allowed rate of return, and yet *R* represents more capital than G and hence more absolute profits. Essentially, this comparison is simply a restatement of the fact that the firm chooses the point on the constraint curve that provides the greatest absolute profits, such that π_R exceeds profits at any other point on the constraint curve by definition of R.

It can also be shown, using a different line of logic, that $\pi_R < \pi_G$. Consider point *F*, where the isoquant through *R* intersects the expansion path. As discussed above, $\pi_F > \pi_R$, because costs are lower at F and output (and hence revenues) are the same. Furthermore, $\pi_G > \pi_F$ because *G* is closer to point *M*, which is the top of the profit hill. (Because the profit hill is indeed shaped like a hill with its top at *M*, profits decrease as one moves along the expansion path beyond *M*. *F* is farther from *M* along the expansion path than *G*, meaning that it is lower on the profit hill.) Because $\pi_G > \pi_F$ and $\pi_F > \pi_R$, it must be the

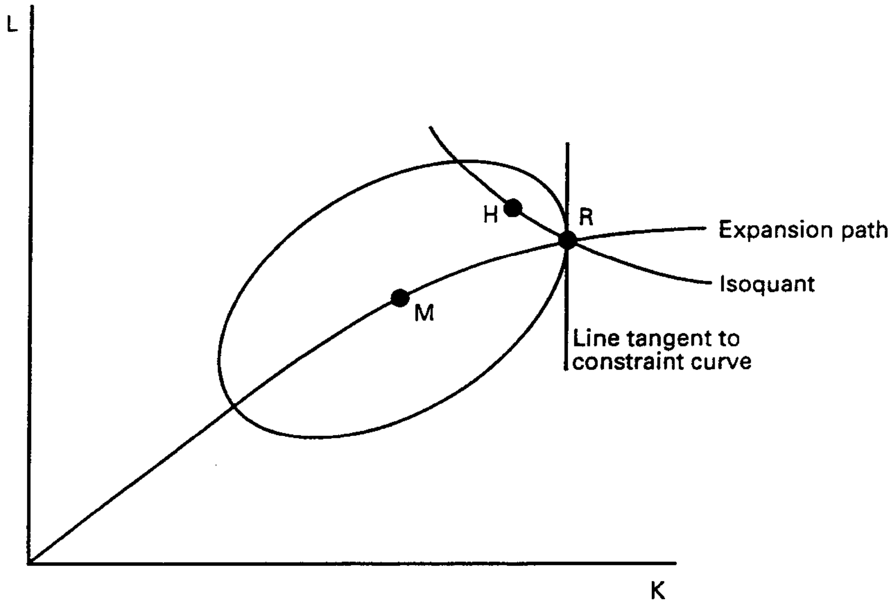


Figure 1.17
Regulated firm's *K/L* ratio is efficient: an impossible situation

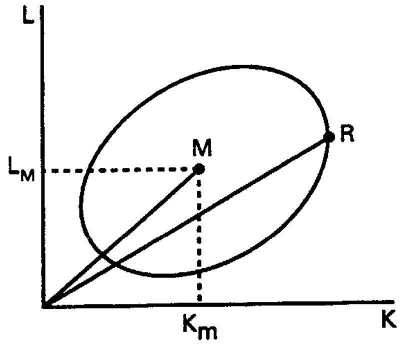
case that $\pi_G > \pi_R$. However, this contradicts the fact that $\pi_R > \pi_G$. Because profits at *R* cannot both exceed and be less than profits at *G*, the situation depicted in figure 1.16 is impossible.

Now consider the situation depicted in figure 1.17 in which the firm's chosen input combination is on the expansion path, such that the firm is choosing the efficient input combination for its level of output. We can show that this situation is also impossible. Recall that *R* is the point on the constraint curve with the greatest amount of capital. Because of this, any point on the constraint curve near *R* is necessarily to the left of *R*, meaning that the slope of the constraint curve at *R* is infinite (that is, the line tangent to the constraint curve at *R* is vertical.) The isoquant through *R* is downward sloping but not vertical, reflecting the fact that labor can be substituted for capital without affecting output.[16] Therefore, the isoquant cuts and passes inside the constraint curve.

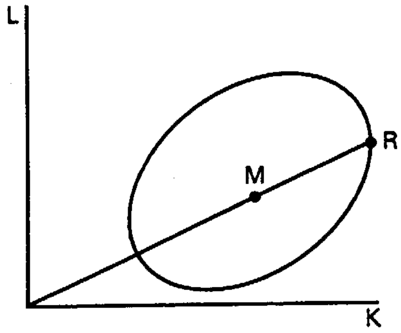
Consider points that are inside the constraint curve and near *R*.

---

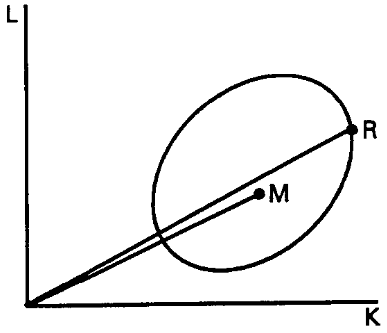
[16] If the isoquant is vertical (that is, output depends on capital only and does not increase with labor), then the situation in figure 1.17 is possible. In fact, it is the only situation possible, meaning that the firm necessarily chooses the efficient input combination. This extreme situation is implicitly excluded for the A-J result.



K/L ratio is higher for regulated firm



K/L ratio is same for regulated firm



K/L ratio is lower for regulated firm

Figure 1.18
K/L ratio of regulated and unregulated firm

Feasible profits at these points are higher than profits at R, because these points represent parts of the profit hill that have been sliced off by the constraint plane (they are higher up the profit hill, closer to M).[17] Now consider a specific point, H, which is inside the constraint curve near R but is also on the isoquant through R. By the above argument, $\pi_H > \pi_R$. However, because R is on the expansion path, $\pi_R > \pi_H$. (Again, costs are lower at R than at H by definition of the expansion path, and yet output and revenues are the same.) Because H cannot obtain profits that are simultaneously greater than and less than the profits at R, the situation depicted in figure 1.17 is impossible.

Result 2 has caused some confusion in the field. The confusion arises from a false syllogism: the capital/labor ratio of the regulated firm is inefficiently high and the capital/labor ratio of the unregulated firm is at the efficient level; "therefore" the capital/labor ratio of the regulated firm is higher than that of the regulated firm. Actually, the capital/labor ratio of the regulated firm can be either greater than, less than, or equal to that of the unregulated firm. In the first panel of figure 1.18, the regulated firm's capital/labor ratio exceeds that of the unregulated firm. To see this, consider the ray from the origin to point M. The slope of this ray is L_M (the "rise") divided by K_M (the "run"). That is, the slope of the ray from the origin to M is the inverse of the capital/labor ratio at M. Similarly, the slope of the ray from the origin to R is the inverse of the capital/labor ratio at R. Because the slope of the ray to R is lower than the slope of the ray to M, the inverse of the capital/labor ratio is lower at R, which is equivalent to saying that the capital/labor ratio is higher.

Contrary to the false syllogism, the capital/labor ratio need not be greater at R than M. The second and third panels depict situations in which the regulated firm has the same and lower capital/labor ratio than the unregulated firm.

The problem with the logic that led to the false syllogism is that it ignores output. Result 2 states that the unregulated firm has an inefficiently high capital/labor ratio *for its level of output*. However, the output of the regulated firm is not generally the same as that of the unregulated firm. It is therefore possible that the regulated firm uses a capital/labor ratio that is inefficiently high for its *own* output level and yet is nevertheless lower than the efficient ratio for the unregulated firm's output level. Stated graphically, result 2 requires that R be below the expansion path; however, the ray to R can be steeper than the ray to M even though R is below the expansion path. Figure 1.19 depicts this possibility.

---

[17] Note that the points must be near R for their profits to be higher. The sliced-off part of the profit hill represents points for which the rate of return exceeds the allowed rate. For points sufficiently far from R, absolute profits can be less even though the rate of return is greater.

For particular types of production processes, it is possible to state definitely the relation between the capital/labor ratios of the regulated and unregulated firms. "Homothetic" production functions, which are widely used in theoretical and econometric work, are an important case. A production function is defined as homothetic if the expansion path associated with the function is a ray. That is, under homothetic production, the cost-minimizing capital/labor ratio is the same for all levels of output.[18]

If the production function is homothetic, then the capital/labor ratio of the regulated firm is necessarily higher than that of the unregulated firm. Figure 1.20 illustrates this situation. The expansion path passes though M and, by homotheticity, is a ray from the origin. By

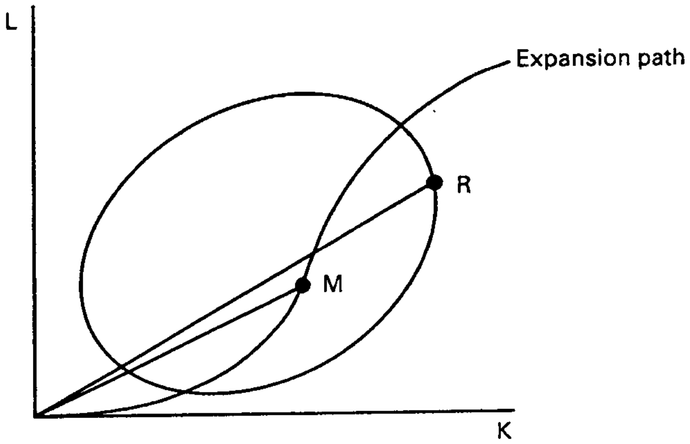


Figure 1.19
K/L ratio of regulated firm lower than for unregulated firm

---

[18] A homothetic production function can, but need not, exhibit constant returns to scale. Constant returns to scale exist when output expands proportionately to inputs (e.g., doubling all inputs results in a doubling of output). Homotheticity requires that, when output expands, the cost-minimizing level of each input expands by the same proportion as that of each other input, but not necessarily by the proportion by which output expands. For example, if output doubles, homotheticity is met if the cost-minimizing levels of capital and labor each increase by, say, half.

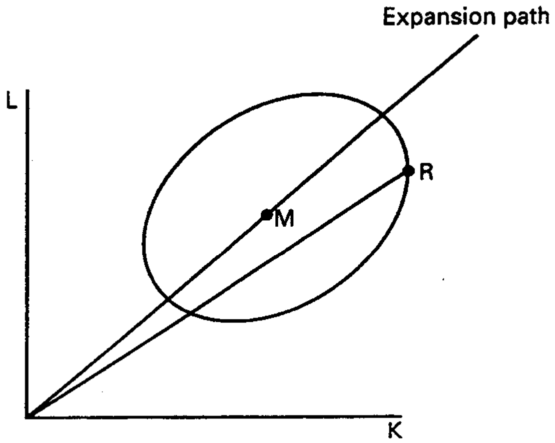


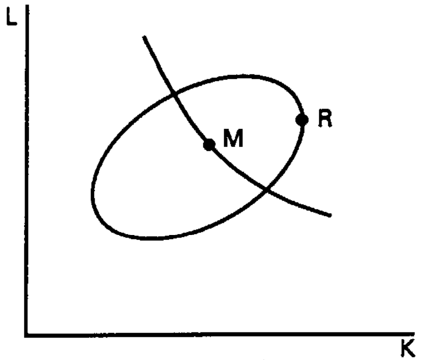
Figure 1.20
Homothetic production function

result 2, point *R* is necessarily below the expansion path. The slope of the ray from the origin to *R* is therefore less than the slope of the ray from the origin to *M*, because the latter is the expansion path that passes above *R*. Since the slope of the ray is the inverse of the capital/labor ratio, this ratio is higher at *R* than *M*.

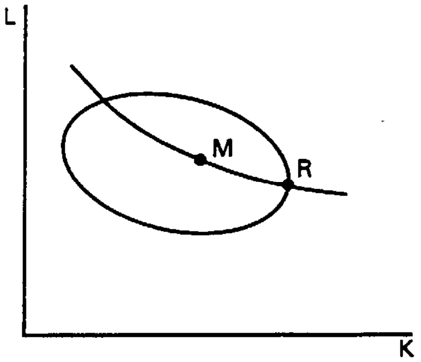
Consider now the output level of the regulated firm. One of the basic results of economic theory is that an unregulated monopolist produces too little output by setting price above marginal cost. A purpose of regulation is to induce public utilities to increase output and lower price. Unfortunately, ROR regulation does not necessarily achieve this objective. The regulated firm might, depending on the shape of its profit hill, produce more, the same, or less output than the unregulated firm. Figure 1.21 illustrates the three possibilities.

Result 1 states that the regulated firm uses more capital than the unregulated firm. With more capital, one might expect that the regulated firm would produce more output. However, this expectation ignores labor. Recall that the firm might increase or decrease its use of labor. If it uses less labor, output could decrease even though capital increases. This is the situation depicted in the third panel of figure 1.21. Furthermore, because demand is downward sloping, the firm in this situation would raise price to be consistent with its lower level of output, such that the price charged by a regulated firm might be higher than if unregulated.

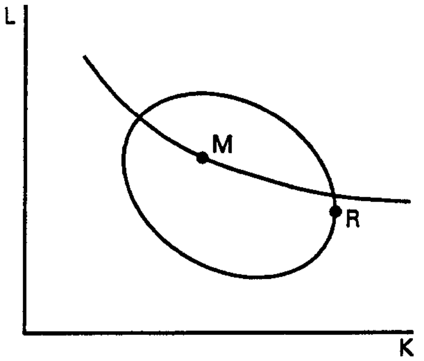
The firm might raise output, as shown in the first panel. However,



Regulated firm
produces more



Regulated firm
produces same



Regulated firm
produces less

Figure 1.21
Effect of ROR regulation on output level

as shown in the next result, there is a limit on how far output can expand under ROR regulation.

Result 3: The regulated firm necessarily operates in the elastic portion of demand, where marginal revenue is positive. That is, the regulated firm never increases its output beyond the point at which marginal revenue is zero.

Suppose the contrary, that the regulated firm chooses an output level at which marginal revenue is negative. This supposition leads to a contradiction. Figure 1.22 illustrates the situation. As discussed in section 1.2, marginal revenue is generally positive for low levels of output and eventually becomes negative at higher levels of output. The isoquant in the graph is for the output level at which marginal revenue is zero. Consequently, points below this isoquant represent output levels at which marginal revenue is positive, and points above it correspond to output levels at which marginal revenue is negative. Point R is placed above the isoquant, reflecting the supposition that marginal revenue is negative at the firm's chosen output level.[19]

Consider point H, which represents the same capital as R but less labor. With less labor, output is lower at H than at R.[20] Because marginal revenue is negative over these levels of output, revenue is therefore higher a H than at R. Because less labor is used at H than at R (without a change in capital), costs are lower at H than R. Because revenues are higher and costs are lower, $\pi_H > \pi_R$. However, R is on the constraint curve while H is outside of it. This means that the rate of return at R is the allowed rate, while that at H is below the allowed rate. Because both R and H represent the same amount of capital, the lower rate of return at H means that $\pi_H < \pi_R$, contradicting the first comparison.

The reason the regulated firm produces in the elastic portion of demand, where marginal revenue is positive, is intuitively meaningful. At any level of capital, if the firm finds that its marginal revenue is negative, then it can increase its profits by decreasing its use of labor. With less labor, its costs are lower, and it produces less such that its revenues are higher. Because the level of capital is not changed, the firm's allowed profits are the same and yet its feasible profits are

---

[19] As shown in section 1.2, marginal revenue is necessarily positive at point M, the top of the profit hill. Therefore, the graph represents a situation in which the firm increases output from the elastic into the inelastic portion of demand when subjected to regulation.

[20] Given, as usual, that the marginal product of labor is positive.

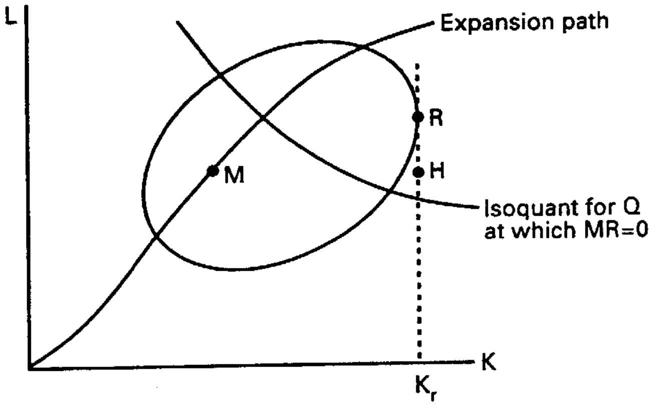


Figure 1.22
Regulated firm operates in inelastic portion of demand: impossible

higher. If feasible profits are less than allowed profits, then increasing feasible profits clearly helps the firm. If feasible profits exceed allowed profits after labor has been reduced, capital can be expanded to increase allowed profits. This expansion of capital decreases feasible profits, but because feasible profits exceed allowed profits, the firm is not able to keep all of its feasible profits anyway. With higher allowed profits, the firm is better off. In either case, the firm will continue to decrease its use of labor until it enters the elastic portion of demand, where marginal revenue is positive.

Result 2 states that the regulated firm will choose a point below the expansion path. Result 3 states that the firm's chosen point will also be below the isoquant for output at which marginal revenue is zero. The complete picture is shown in figure 1.23.

Compare now the firm's chosen point with the socially optimal outcome. As discussed in section I.3, pricing at marginal cost provides the first-best output. However, with a natural monopoly facing continuously declining average cost, pricing at marginal costs results in the firm losing money. If the regulator cannot subsidize the firm, then the second-best outcome becomes the goal. The second-best outcome consists of the firm's pricing at average cost (such that profits are zero) and using the cost-minimizing inputs for the level of output demanded at that price. The expansion path represents points that are cost minimizing, and the zero-profit contour represents points that result in zero profit. The second-best outcome occurs, therefore, at the intersection of the expansion path and the zero-profit contour,

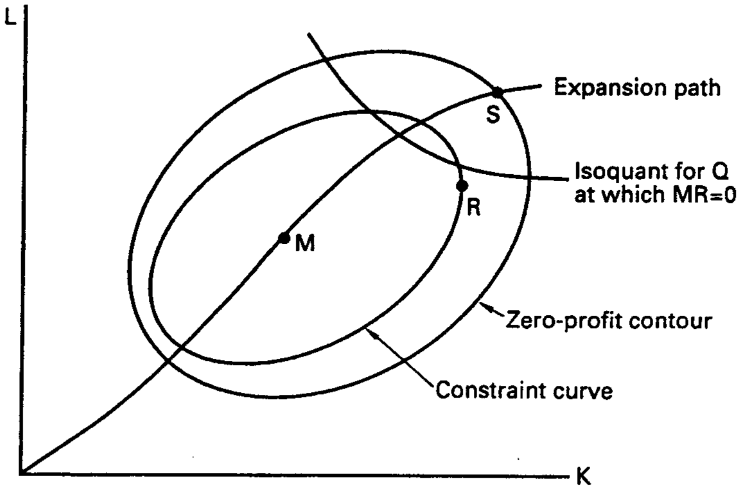


Figure 1.23
Regulated firm chooses input combination below expansion path and the isoquant for $MR=0$

where profits are zero (such that price equals average cost) and costs are minimized. This intersection is point S in the graph.

Ideally, the regulator would like to establish a form of regulation that induces the firm to move out the expansion path from M to (or at least toward) S, using inputs efficiently and increasing output and reducing price. Result 2 shows that under ROR regulation, the firm will not move out the expansion path but rather will produce with an inefficient input combination. In fact, the firm need not increase output and may even decrease it. If the firm increases output, result 3 shows that there is a limit to how far the firm would possibly increase its output; in particular, it would never move into the inelastic portion of demand. The A-J critique of ROR regulation is therefore quite damaging: ROR regulation induces the firm to be inefficient and yet does not necessarily induce it to increase output and decrease price.

The problems with ROR regulation essentially arise from the fact that it gives the firm incentives based on capital while capital per se is not what the regulator is wanting the firm to increase. We investigate below whether the situation is improved by reducing the allowed rate of return, thereby reducing the profits the firm earns on capital. But first we consider a result that shows that ROR regulation does not induce one type of inefficiency, despite a widely held myth to the contrary.

Result 4: The regulated firm produces as much output as possible given its capital and labor.

This result states that the regulated firm will not indulge in pure waste in the sense of producing less than is maximally possible given its inputs. With respect to capital, the result means that the firm will not acquire nonproductive capital, that is, capital that does not serve a productive function. The result contradicts a commonly held belief that firms under ROR regulation have an incentive to purchase capital that is not used. The result is demonstrated as follows. At any input levels K_0 and L_0, the maximum amount of output the firm can produce is given by its production function evaluated at these input levels: $f(K_0, L_0)$. The firm can choose to waste inputs and produce a lower level of output: $Q_0 < f(K_0, L_0)$. By result 3, the firm necessarily chooses to operate in a region of demand where marginal revenue is positive. The firm therefore earns less revenue if it produces less output. Because the cost of K_0 and L_0 is the same whether or not the firm uses the inputs productively, feasible profits decrease if the firm produces less than is maximally possible with these inputs.

Figure 1.24 depicts the situation in the graph of profits and capital. If the firm uses its capital to produce as much output as possible, then it chooses point R_0 and earns profits π_0. If the firm wastes capital, its profit hill is lower: at any level of capital it earns less profit than if the capital were used productively. With waste, the best the firm can do is operate at point R_1 and earn profits π_1. Because $\pi_0 > \pi_1$, the firm chooses not to waste.

This result does not contradict result 2, which states that the firm chooses an inefficiently large amount of capital. Two different kinds of inefficiency are being considered. The firm chooses an input combination that is inefficient for its level of output (result 2), but *given its inputs* it produces as much output as possible (result 4). Stated in terms of isoquants: the firm will choose the wrong point on the isoquant (costs would be lower if it moved to the expansion path), but it nevertheless produces the full amount of output designated by that isoquant.

The use of nonproductive capital is often called "goldplating," referring to the idea that a firm might plate its building and fixtures in gold simply because gold is more costly than other materials. It is widely thought that the A-J effect means that a firm under ROR reg-

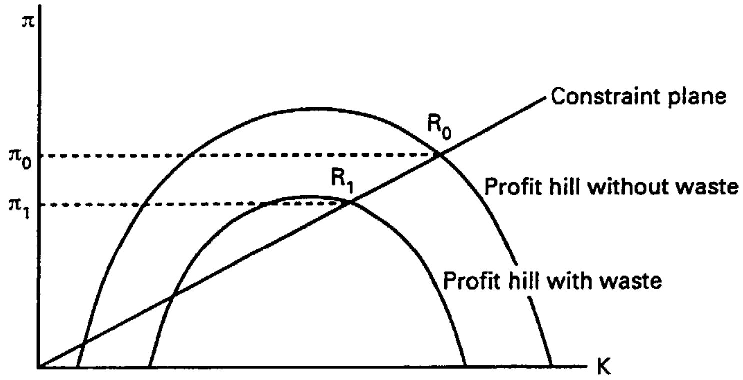


Figure 1.24
Effect of wasting inputs

ulation will goldplate. This conception is incorrect. The A-J effect (result 2) states that the firm will use too much capital relative to labor but does not state that the firm will waste capital. In fact, result 4 implies that the regulated firm will not purchase capital that does not serve a productive purpose, nor will it purchase a less productive type of capital when it could purchase a more productive type instead.

The concept that a firm under ROR regulation will not purchase nonproductive capital is somewhat subtle. If, for example, the firm is allowed to earn 10% profits on any extra capital it purchases and the cost of extra capital is 8%, why would the firm not buy extra capital, even if the capital were unproductive?

The answer depends on a careful differentiation of *allowed* profits and *feasible* profits. Consider figure 1.25. Suppose the firm starts with K_L amount of capital. At this level of capital, the firm is able to earn π_{L1} but is only allowed to earn π_{L2}. If the firm were to stay at that capital level, then it would need to waste inputs to reduce its profits by $\pi_{L1} - \pi_{L2}$, so that its actual profits do not exceed the allowed amount. However, instead of wasting and earning profits of π_{L2}, the firm would earn more profits by increasing its amount of capital toward K_R. With K_R, the profits the firm is able to earn are exactly the same as the profits it is allowed to earn; that is, by using its inputs as productively as possible, the firm with K_R is just able to earn the maximum profits that it is allowed to earn. Now, consider an increase in capital beyond K_R to K_H. The amount of profit the firm is allowed to earn increases, because the firm has more capital. However, the amount that the firm is able to earn decreases, even if the extra capital is used

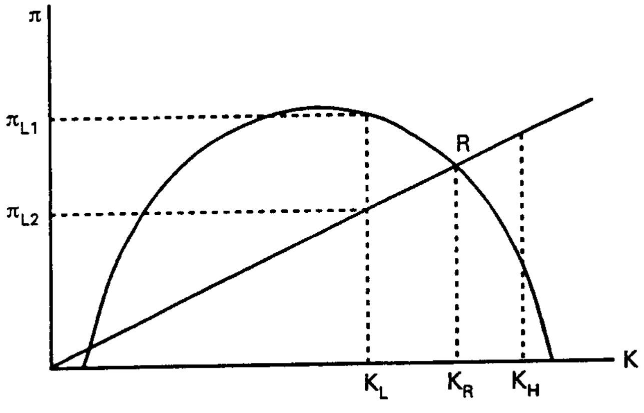


Figure 1.25
Allowed versus feasible profits

as productively as possible. If nonproductive capital is purchased, or the extra capital is not used as productively as possible, the profits the firm is able to earn decrease even more.

To repeat: At levels of capital lower than K_R, the firm must waste so as to reduce profits to the allowed level. However, rather than wasting, the firm is better off by purchasing more capital, such that its allowed profits rise. Once the firm has reached K_R, the maximum profits it is able to earn equal the profits it is allowed to earn. If the firm wasted at this point, it would earn less than the allowed amount. Also, if the firm increased its capital beyond K_R, it would be allowed to make extra profits but would not be able to, whether the extra capital were productive or nonproductive.

Return now to the question about the firm that faces a 10% allowed rate and yet can purchase capital at 8%. At the firm's chosen capital level, the firm would be allowed to earn more profit if it purchased more capital, but it would not be able to, even if it used the capital productively. If the firm cannot earn extra profits with productive capital, it will certainly not purchase *nonproductive* capital.[21]

---

[21] In the extreme case of fixed-proportions production (i.e., no substitution between labor and capital), the distinction fades between wasting capital and using an inefficiently high capital/labor ratio. Isoquants are L-shaped under fixed-proportions production, with the cost-minimizing input combination being at the kink. Under regulation, the firm will choose a point on the leg of the isoquant representing more capital than is needed for that output. Result 4 is still correct: the firm produces as much output as designated by the isoquant. However, the same output could be produced with less capital and no extra labor. In that sense, the excess capital is indistinguishable from pure waste.

As stated in relation to results 2 and 3, the basic problem with ROR regulation is that the firm makes profit on capital, whereas the goal of regulation is not to increase capital per se. Let us examine therefore whether the distortions from ROR regulation can be alleviated if the firm is not allowed to make as much, or even any, profits on capital. We obtain two more results.

Result 5: When the fair rate of return is reduced toward the cost of capital (that is, when the allowed rate of economic profit is lowered toward zero), the regulated firm increases its use of capital.

Figure 1.26 illustrates this result. The slope of the constraint plane is $f-r$, that is, the allowed rate of economic profit. As f drops toward r, this slope becomes less positive. More of the profit hill is sliced off with the lower f. The firm chooses the profit-maximizing point on the more severely sliced-off profit hill, increasing capital from K_0 to K_1. Figure 1.27 illustrates the same ideas in a somewhat more discernible fashion, by collapsing figure 1.26 to only the K-π dimensions.

The firm still uses an inefficiently high capital/labor ratio, because result 2 holds whenever f exceeds r. Furthermore, no matter how close f is to r, the firm still does not produce in the inelastic region of demand. If necessary, the firm reduces its labor sufficiently such that the increase in capital does not move it into this inelastic region.

Clearly, reducing the amount of profit that the firm is allowed to earn on capital does not solve the basic problems. Consider, however, not allowing the firm to make *any* economic return on capital.

Result 6: If the fair rate of return is set equal to the cost of capital (that is, if the allowed rate of economic profit is zero), then the firm is indifferent among many levels of output and many input combinations, including the option of closing down.

If f is set equal to r, then the slope of the constraint plane is zero such that it is flat at the base of the profit hill, as shown in the first panel of figure 1.28. The constraint curve in the K-L dimensions is therefore the set of input combinations that result in exactly zero profits. That is, the constraint curve is the same as the isoprofit contour for zero profits, as given in the second panel.

The firm is not allowed to make more than zero profits and is indifferent among the various options available to it for making zero profits. The firm could choose input combinations on the constraint curve, produce as much output as possible with these inputs, and earn zero

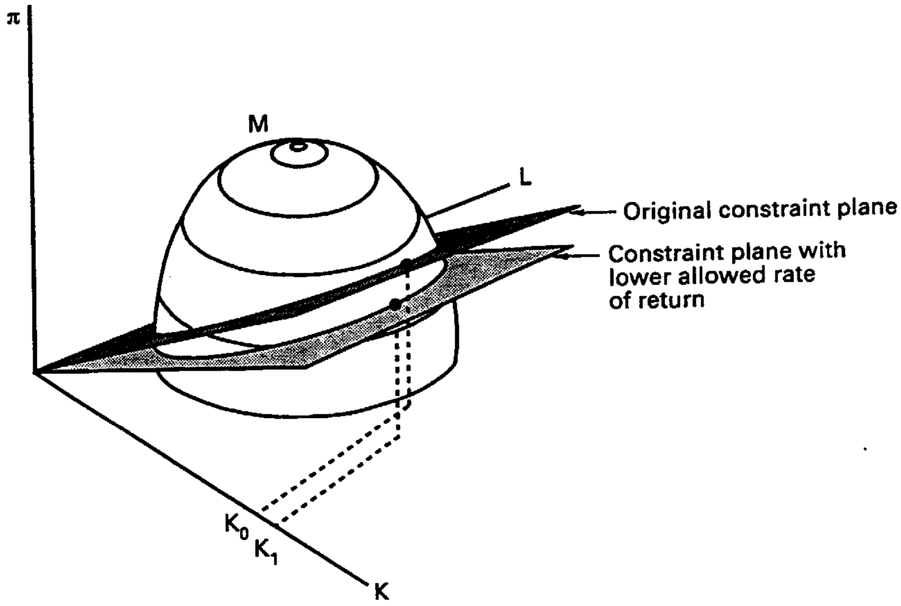

π
M
L
Original constraint plane
Constraint plane with lower allowed rate of return
K_0
K_1
K

Figure 1.26
Effect of lowering the fair rate of return

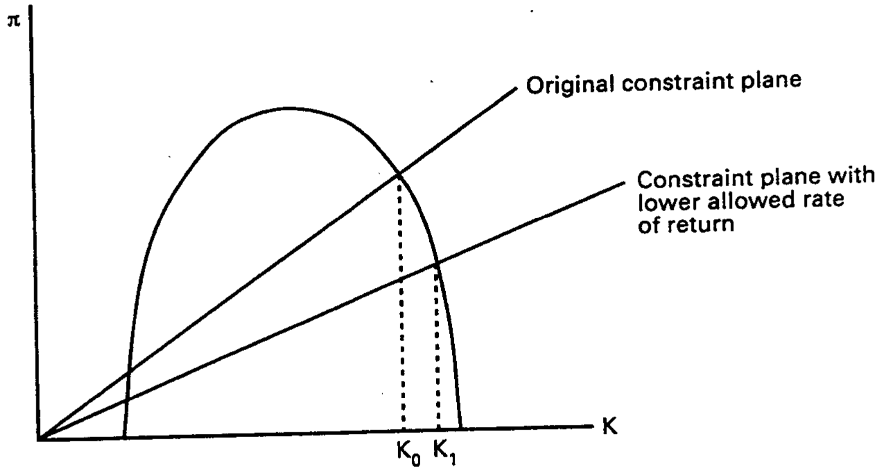

π
Original constraint plane
Constraint plane with lower allowed rate of return
K
K_0
K_1

Figure 1.27
Effect of lowering the fair rate of return

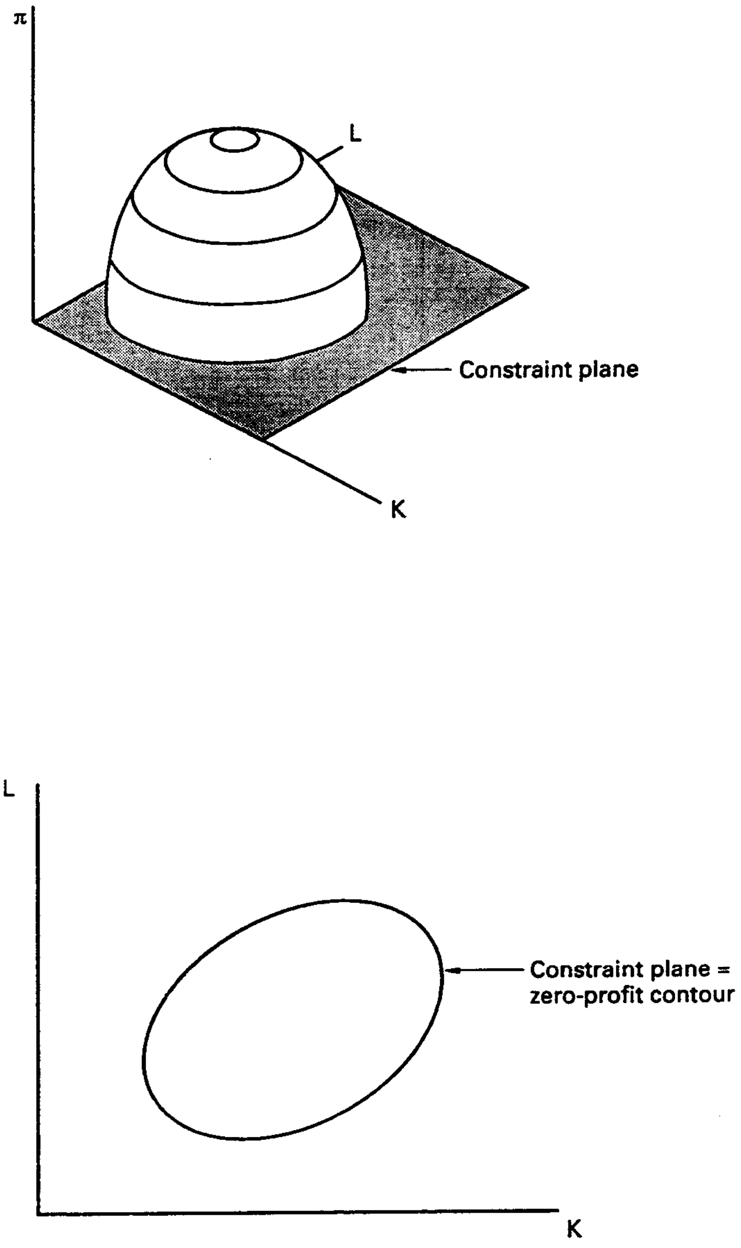

π
L
Constraint plane
K
L
Constraint plane =
zero-profit contour
K

Figure 1.28
Fair rate of return set equal to cost of capital

profits. Alternatively, the firm could choose an input combination inside the constraint curve. Even though it would be able to earn positive profits with these inputs, it would not be allowed to earn more than zero profits. Consequently, it would waste inputs until its profits fell to zero. As a third alternative, the firm could earn zero profits by choosing no inputs and halting production. The firm is indifferent, therefore, among producing without waste at any input combination on the constraint curve, producing with waste at any input combination inside the constraint curve, and not producing at all.

Result 6 contains an important lesson that generalizes beyond ROR regulation. The problem with ROR regulation is that it establishes a mechanism under which the firm earns profits on capital when the regulator is not interested in increasing the use of capital per se. However, the solution is not simply to prevent the firm from earning any profits. If the firm earns zero profit no matter what it does (at least within a range), the firm becomes indifferent in its choice of input, outputs, and whether to waste. In that case, there is no reason to expect the firm to make choices that satisfy the regulator's goals. The solution is rather to establish a situation in which the firm earns *more* profit at the socially optimal outcome than at any other. Profits might be zero at this outcome, but, for the firm to choose it, profits must then be negative at all other outcomes.

Consider now the possibility of lowering the allowed rate of return *below* the cost of capital. An important result obtains.

Result 7: If the fair rate of return is set below the cost of capital, then the regulated firm will choose to utilize no inputs and produce no output.

If r exceeds f, the constraint plane is downward sloping, as shown in figure 1.29. Allowed economic profits are negative for any positive amount of capital. If the firm continues to produce, it minimizes its losses at K_0. At higher levels of capital, the firm is able to earn greater profits, but its allowed profits are lower; while with less capital (but still a positive amount), the firm is allowed to earn more profits but is not able to.

Rather than produce at a loss, the firm will choose to go out of business, that is, use no inputs and produce no outputs. Profits are zero under this option, and because zero profits are greater than the negative profits the firm earns at K_0, the firm exercises this option.

In reality, the firm might not legally be able to go out of business. For example, it is doubtful that the electric utility for an area would

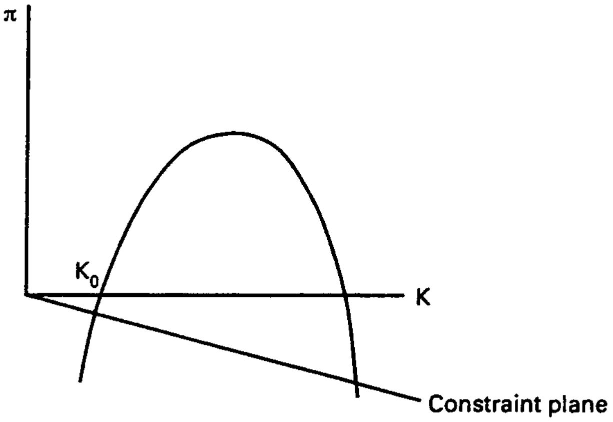


Figure 1.29
Fair rate of return set below the cost of capital

be allowed to stop production. Furthermore, the concept embedded in the A-J model that capital can be bought and sold at a fixed rate *r* does not reflect the fact that the utility's capital (e.g., a hydroelectric plant) is highly specialized such that markets for it do not exist. The utility might not be able to dispose of its capital, in which case it might make more profits (smaller losses) by producing output than by shutting down. It is probably even more realistic to expect that the price the utility can obtain for its capital depends on the rate of return that the regulator sets as "fair." That is, the most that capitalists would be willing to pay for the utility's capital, which can only be used for the production of the utility's output, is the return the capitalists expect they would be able to earn with the capital. Once the regulator sets a fair rate of return, the capitalists would be willing to pay at most this fair return. Thus, if the regulator attempts to set *f* below *r*, *r* will simply drop until it is below or equal to *f*.[22]

The point of result 7 is simply that the firm has an incentive to reduce its capital if it is allowed to make more profits (smaller losses) with less capital. The same types of distortions occur as when the firm is allowed to make more profits by using more capital, only in reverse. As discussed in relation to result 6, the solution is to develop an alternative form of regulation under which the firm faces incentives that are consistent with the regulator's goals.

---

[22] If there are a sufficient number of capitalists, *r* would not drop below *f*, because if *r* were below *f*, it would be bid up to *f*. However, it is doubtful that the number of capitalists who are willing and able to purchase a utility's specialized capital is sufficiently large for the market for this capital to be efficient in this sense.

1.5 Empirical Evidence on the A-J Effect

The A-J model abstracts from many important aspects of real-world regulation. The regulated firm is assumed to be prohibited from earning more than the allowed profits in each time period, even though in reality regulated firms' profits are allowed to fluctuate above and below the allowed level as long as the average profit over a number of periods is within the allowed amount. The firm is assumed to face a capital cost independent of the fair rate of return that the regulator sets, whereas in reality the cost of borrowing funds depends on the fair rate, and the regulator sets the fair rate based, in part, on the cost of capital to the firm. The firm is assumed to know its demand and costs exactly, when usually a firm has only partial information. Strategic considerations, by which the firm might make choices in one period in order to affect the regulator's decisions in later periods, are also omitted.

The A-J model has value, even if its results do not generalize to the more complex situation of the real world. Its value lies in focusing the concept that any particular form of regulation induces the firm to act in a particular way that may be consistent or inconsistent with the regulator's goals.

It is useful, however, to recognize that the A-J model can be tested, because the procedure for such testing further elucidates the model. A study by Courville (1974), one of the first empirical tests of the A-J model, is particularly well suited to elucidating the method and difficulties of such testing.[23]

The main proposition of the A-J model is that the regulated firm employs too much capital relative to labor given its level of output. This proposition can be stated in terms of a hypothesis that is empirically testable. Consider figure 1.30. According to the A-J model, the regulated firm chooses point R while the cost-minimizing input combination for this level of output is *F*. At *F*, the isocost line is tangent to the isoquant, as required for cost minimization. At *R*, the isocost is not tangent to the isoquant. All isocost lines have a slope equal to the negative of the ratio of input prices: $-\ r/w$. The slope of the isoquant at any point is the negative of the marginal rate of technical substitution *(MRTS)* at that point. Recall that *MRTS* is the amount of extra labor that must be employed to keep output constant if capital is re-

---

[23] Other empirical studies have been conducted by Spann (1974) and Peterson (1975).

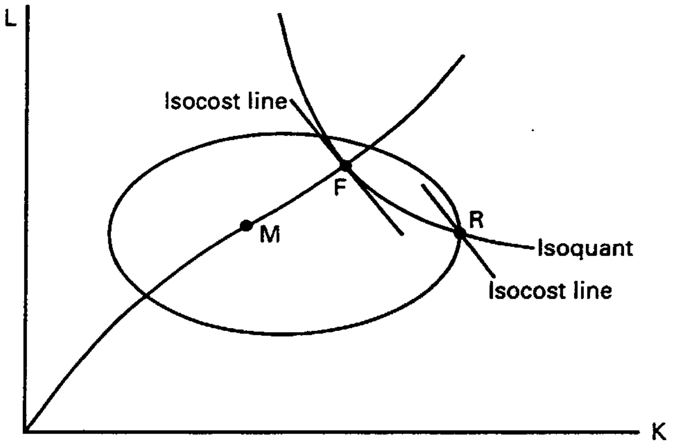


Figure 1.30
Observable consequence of the A-J effect

duced by one unit. By definition, *MRTS* is the ratio of the marginal products of capital and labor; that is, *MRTS* = *MPK/MPL*.[24] Because marginal products change as the firm uses more or less of an input, *MRTS* is different at different points on the isoquant. At the cost-minimizing input combination, *F*, *MRTS* equals the ratio of input prices because the isocost line and isoquant are tangent. However, at *R*, more capital and less labor is used, such that *MRTS* is lower than at *F* and, in particular, is less than the ratio of prices. Stated succinctly, because the isocost line is steeper than the isoquant at *R*, $r/w > MRTS$. Equivalently, because *MRTS* equals the ratio of marginal products, $r/w > MPK/MPL$.

This fact constitutes a testable hypothesis. The A-J model states that the regulated firm will choose point R, where the ratio of input prices exceeds the ratio of marginal products. By observing the input prices and marginal products for a regulated firm, the validity of this hypothesis for that firm can be assessed.

---

[24] This equality can be easily demonstrated. If the firm uses one less unit of capital, its output decreases by *MPK*. To keep output constant, the firm must increase its use of labor enough to boost output to its original level, that is, to increase output by *MPK*, which is the amount lost with the reduction in capital. Each unit of labor increases output by *MPL*. Therefore, the amount of extra labor the firm must use is the amount of extra output required (*MPK*) divided by the amount of output obtained with each extra unit of labor (*MPL*). For example, suppose the marginal product of capital is six and that of labor is three. If capital is reduced by one unit, output decreases by six units. To regain those six units of output, the firm must employ two extra units of labor, because each unit of labor boosts output by three units. Hence, *MRTS* (that is, the amount of labor required to keep output constant when capital is reduced by one unit) equals $MPK/MPL = 6/3 = 2$.

The difficulty, of course, is that marginal products are not directly observable. Rather, marginal products are approximated by estimating the firm's production function and deriving marginal products. The ratio of these estimated marginal products are then compared with the ratio of input prices, and statistical tests are performed to determine whether the two ratios are significantly different.

Courville performs this test using data on electric generation plants. He considers three inputs: capital, labor, and fuel (labeled F, with price v). Because fuel costs are treated like labor costs under ROR regulation, the A-J model implies that $r/v > MPK/MPF$, just as $r/w > MPK/MPL$. To derive marginal products, Courville assumes that the production function for electric generation is Cobb-Douglas, which takes the form:

$$\log Q = \alpha + \beta \log K + \mu \log F + \psi \log L + \epsilon,$$

where ϵ is an error term. Under this specification, the marginal products are:[25]

$$MPK = \beta(Q/K);$$

$$MPF = \mu(Q/F);$$

$$MPL = \psi(Q/L).$$

By estimating the parameters β, μ, and ψ, and observing the firm's levels of Q, K, F, and L, the marginal products are calculated.

To reflect differences in the production technologies among plants, two terms are added to the production function. (1) Different-sized plants generally use different technologies and have different levels of efficiency. To reflect this fact, Courville includes the capacity of the plant, denoted C, as an explanatory variable in the production function. Capacity is defined as the maximum output the plant is capable of producing. (2) Output Q is measured as total kilowatt-hours produced in the year. However, each plant's output at any point in time varies over times of day and season (for example, output is often greater in the peak afternoon period, when customers use their air conditioners, than in the morning or evening). This variation affects the estimation of the production function because the efficiency with which a plant operates depends on the degree of variability in its output. To reflect this fact, Courville includes a variable that partially captures the variation in output: capacity utilization, denoted U, which is the annual output of the plant expressed as a percentage of the plant's capacity.[26]

---

[25] $MPK = \Delta Q/\Delta K = (\Delta \log Q/\Delta \log K)(Q/K) = \beta(Q/K)$. Similarly for MPF and MPL.

With these additions, the equation to be estimated is

$$\log Q = \alpha + \beta \log K + \mu \log F + \psi \log L + \zeta C + \theta U + \epsilon.$$

The equation is estimated with annual data on the inputs, outputs, and other characteristics of 134 electricity-generation plants. To reflect the fact that technology changes over time, the 134 plants are grouped into four categories on the basis of when they were built (1948–50, 1951–55, 1956–59, and 1960–66), and the production function is estimated separately for each group.

In preliminary estimation, the variable for labor consistently entered with the wrong sign and without statistical significance. Courville therefore eliminates labor from the equation for final estimation. As a result, he is able to test whether $r/v > MPK/MPF$, but not whether $r/w > MPK/MPL$.

The estimation results for plants built in the period 1960–66 are the following:

$$\begin{array}{lccccc} \log Q = & 0.73 & + 0.10\log K & + 0.97\log F & + 0.00012C & + 0.34U, \\ & (3.4) & (3.1) & (17.4) & (0.13) & (3.0) \end{array}$$

with the t-statistic for each estimated parameter given in parentheses.[27] Results for plants built in other periods are qualitatively similar, except for those built in 1955–59, for which the coefficient of capital is estimated to be implausibly negative. Eliminating the plants built in 1955–59 left 110 plants for which to compare ratios of capital and fuel prices with ratios of marginal products.

For each plant, the ratio of marginal products is calculated:

$$\begin{aligned} MPK/MPF &= \beta(Q/K) \,/\, \mu(Q/F) \\ &= (\beta/\mu)\,(F/K). \end{aligned}$$

---

[26] The use of this variable might be problematic econometrically. Because U is defined as Q/C, entering U as an explanatory variable is nearly the same as entering the dependent variable, $\log Q$, as an explanatory variable.

[27] A t-statistic indicates the precision with which the parameter is estimated, with higher t-statistics representing greater precision. As a reference point, a t-statistic exceeding 2.0 indicates that the hypothesis that the parameter is zero can be rejected with 95% confidence.

For plants built in 1960–66, the estimates of β and μ indicate that $MPK/MPF = (0.10/0.97)\,(F/K) = 0.103(F/K)$. By inserting the levels of F and K for each plant, the ratio of marginal products for each plant is calculated. Comparison with the ratio of input prices determines whether, as suggested by the A-J model, the ratio of input prices exceeds the ratio of marginal products such that the firm is using too much capital relative to fuel for its level of output.

Courville's results are striking. Using capital expressed in real terms (that is, deflated by a price index) and fuel consumption in the first year of each plant's operations, he finds that for *all* 110 plants the ratio of input prices exceeds the ratio of marginal products as the A-J model suggests. Furthermore, this comparison is statistically significant at the 95% confidence level for 105 of the 110 plants.

Courville repeats the tests using different measures of capital and fuel consumption, because the most appropriate measure of these variables is not clear (for example, fuel consumption in the first year of operation versus the year of the plant's greatest output). In each of these sets of tests the A-J effect is confirmed. In the "worse" case (i.e., using the set of measures that least supports the A-J model), he finds the A-J effect to be confirmed with 95% confidence for 71 of the 110 plants. The ratio of input prices exceeds the ratio of marginal products in even more than these 71 plants, though not significantly so.

The same concepts are used to calculate the extra cost that results from overcapitalization. Refer again to figure 1.30. The isocost line through R represents the cost of producing the firm's output with its chosen inputs. The isocost through F represents the cost of producing the same output with the cost-minimizing inputs. The difference between these two costs is the loss due to the inefficiency induced by ROR regulation.

Using the estimated production functions, Courville calculates this loss for each plant. He finds that costs are as much as 40.6% higher than minimum because of the plants' inefficiently high capital/labor ratios. Averaged over all plants, the loss is 11.4%.

Courville's method of testing the A-J model focuses on result 2, which deals with the input mix of the firm. Bailey (1973) suggests another way to test the A-J model, based on the firm's output. Result 3 states that the regulated firm produces a level of output at which marginal revenue is positive. Since marginal revenue being positive implies that the demand elasticity exceeds one (in magnitude), Bailey

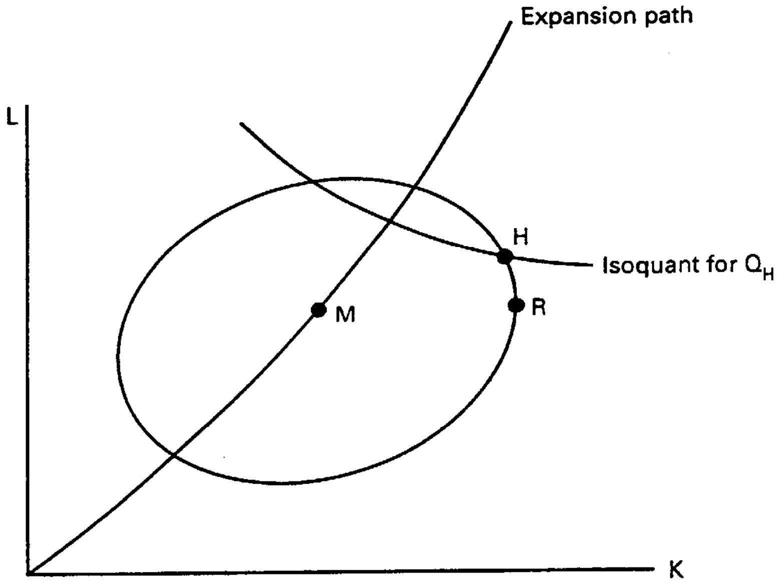


Figure 1.31
Regulated firm under ROR regulation and a maximum price

observes that a test of whether elasticity exceeds one constitutes as test of the A-J model (though not of the A-J *effect*, which is result 2).

The two methods of testing are complementary. Courville's method tests whether the firm uses an inefficient input combination, while Bailey's method tests whether the firm keeps output in the elastic region of demand. It is possible that the regulated firm overcapitalizes in accordance with result 2 but is nevertheless somehow induced by its regulator to increase output into the inelastic portion of demand (where marginal revenue is negative). This inducement might occur, for example, if the regulator applies ROR regulation but also provides oversight on prices. In this situation, the firm is not able to choose its price unilaterally within only the constraint that profits do not exceed the allowed amount. Rather, there is a maximum price the regulator is willing to approve. If demand at this price is so high that marginal revenue is negative, then the firm necessarily produces in the inelastic portion of demand. Yet the firm still uses too much capital relative to labor for its level of output. Figure 1.31 illustrates the situation. Under the type of ROR regulation described by the A-J model, the firm chooses *R* where the capital/labor ratio is inefficiently high and output is sufficiently low that marginal revenue is positive. Suppose, however, that the regulator also oversees price directly and that the highest price the regulator will approve is P_H. At this price, the quantity demanded is Q_H, which, for illustration, is assumed to be sufficiently large that marginal revenue is negative. The firm is restricted in its choice to any input combination on or above the isoquant for Q_H, because it cannot restrict output below this level by raising price. The firm would choose point H, which contains the most capital, and hence profits, of those points on the constraint curve and above the isoquant. At H, the firm has an inefficiently high capital/labor ratio for its level of output and yet is producing in the inelastic portion of demand.

In this situation, the method used by Courville is expected to detect the presence of overcapitalization, whereas that suggested by Bailey is expected to indicate that output is greater than would occur without direct price control by the regulator. Thus, applying the two tests together assists in identifying the aspects of the A-J model that are applicable in a particular setting.