2

Regulatory Mechanisms to Induce Optimal Outcomes for One-Product Natural Monopolies

2.1 Introduction

The Averch-Johnson model indicates that rate-of-return regulation does not induce a firm to choose the optimal inputs and output. The same method can be used, however, to identify other mechanisms that do induce optimality. In this chapter we examine three types of regulation that are similar to rate-of-return regulation. Each places a limit on the profits the firm is allowed to earn but differs from ROR regulation in the factor on which allowed profits is calculated. Under ROR regulation, allowed profits rise with capital. Under the three mechanisms described in this chapter, allowed profits rise with output, sales (that is, revenue), and costs, respectively. It is shown that, under certain circumstances, these regulatory mechanisms induce the firm to choose a level of output and inputs that is arbitrarily close to the second-best outcome. Bailey's (1973) work in this area is particularly useful in identifying these results.

A mechanism is also introduced that induces the firm to choose the first-best outcome. This mechanism is different in form from ROR regulation in that it does not place a limit on the firm's profit. A regulator might, therefore, consider this form of regulation inappropriate for equity reasons. However, the concepts embodied in this form of regulation—the forces that drive the optimality—are important and serve as the basis for the design of other, more equitable regulatory mechanisms.

To facilitate the discussions of alternative forms of regulation, recall the concepts of first- and second-best outcomes. Total surplus is maximized when price is equal to marginal cost. The first-best outcome consists therefore of using the cost-minimizing input combination to produce the level of output that is demanded when price equals mar-

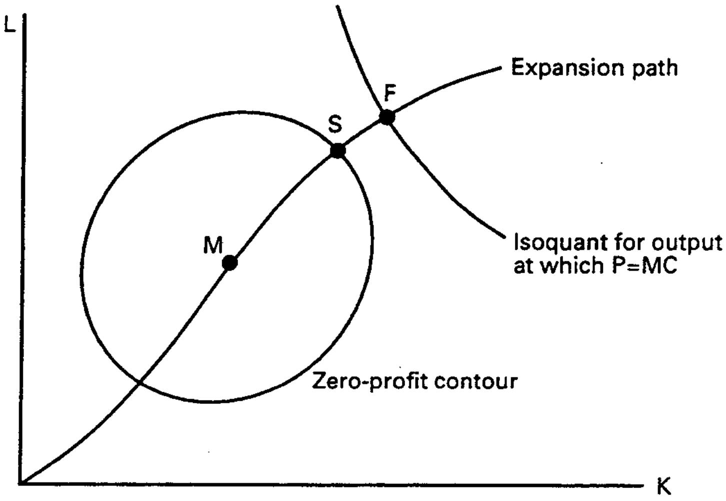

Figure 2.1
First- and second-best outcomes

ginal cost. This outcome is illustrated in figure 2.1 as point *F*, where the isoquant for the level of output at which price equals marginal cost intersects the expansion path.[1]

This first-best outcome might not be feasible. For many public utilities, fixed costs are so high that average cost exceeds marginal cost at relevant demand levels, such that pricing at marginal cost causes the firm to lose money. If the firm's losses cannot be subsidized, then price must be raised to average cost. The second-best outcome consists therefore of using the least-cost input combination to produce the level of output demanded when price equals average cost. This outcome is point *S*, where the expansion path intersects the zero-profit contour.

Four types of regulation are described. The findings for each can be summarized as follows:

1. Return-on-output (ROO) regulation. The firm is allowed to earn a certain amount of profit on each unit of output it sells. The firm is free to choose its inputs, output level, and price as long as its profits

---

[1] Point *F* is necessarily further out along the expansion path, representing more output, than the profit-maximizing point *M*. Profits are maximized when marginal revenues equal marginal cost. Because demand is downward sloping, price exceeds marginal revenue (the firm must lower its price in order to sell more output). Therefore, at the profit-maximizing point, price exceeds marginal cost. For price to equal marginal cost, price must be lowered, thereby increasing output.

do not exceed the allowed amount per unit of output. Under this form of regulation, the firm increases its output beyond the level it would choose if it were not regulated. The firm also chooses an efficient input combination for its level of output and does not waste inputs. If the regulator sets the allowed return on output sufficiently low, the firm can be induced to expand output practically to the second-best level. The second-best outcome cannot be achieved exactly; however, it can be approached arbitrarily closely.

2. Return-on-sales (ROS) regulation. The firm is allowed to earn a certain amount of profit on each dollar of revenue. If marginal revenue is positive up to the second-best output (that is, if the second-best output is in the elastic portion of demand), then the firm under ROS regulation behaves exactly the same as under ROO regulation. This equivalence is due to the fact that when marginal revenue is positive, revenues rise with output such that tying allowed profits to revenues is the same as tying them to output. With positive marginal revenue, the firm can be induced to move arbitrarily closely to the second-best outcome. Unlike ROO regulation, however, the firm under ROS regulation will not expand its output into the inelastic portion of demand where marginal revenue is negative. In this region, an expansion of output decreases revenues and hence allowed profit under ROS regulation. ROO and ROS regulation differ when demand is inelastic because expanding output increases allowed profit under ROO regulation, but decreases it under ROS regulation. If the second-best output is in the inelastic portion of demand, ROS regulation can induce the firm to move only part of the way to second-best, namely, only to the point where marginal revenue starts to be negative.

3. Return-on-cost (ROC) regulation. The firm is allowed a certain amount of profit on each dollar it expends. The firm behaves the same under ROC regulation as under ROS regulation. Specifically, the firm expands output, using least-cost production, but will not enter the inelastic portion of demand. The reasons for this behavior, however, are somewhat different than with ROS regulation. Under ROC regulation, the firm increases its allowed profit by increasing its costs. As long as marginal revenue is positive, the firm benefits from increasing its output along with its costs, because the extra revenues obtained from the extra output help to offset the costs. That is, by expanding outputs, feasible profits rise along with its allowed profit, or at least fall by less than if output were not expanded. However, if marginal

revenue is negative, the firm obtains more revenues by *not* increasing output. Consequently, at the point at which marginal revenue starts to be negative, the firm increases cost (to increase its allowed profit) but does not increase its output (so as to keep its feasible profit as high as possible). In short, the firm starts wasting at this point.

4. Price discrimination. Primary price discrimination occurs when the firm charges a different price to each customer. The firm that is allowed to engage in primary price discrimination will attain the first-best optimum. The reason is simple. The firm charges each customer the maximum the customer is willing to pay for the good, thereby extracting all surplus. Because all surplus accrues to the firm in the form of profit, the firm maximizes its own profit by choosing the surplus-maximizing output, which, by definition, is the first-best. The firm earns large profits; however, in theory, this profit can be taxed and redistributed to consumers in a way that will not affect the firm's behavior.

2.2 Return-on-Output Regulation

Consider a regulatory mechanism that ties the allowed profits of the firm to the firm's output. Under return-on-output (ROO) regulation, the firm is free to choose its input and output levels, but is not allowed to earn (economic) profits in excess of a "fair" return per unit of output. The fair return is set by the regulator and stated in terms of dollars of profit per unit of output. For example, the regulator of an electric utility might state that the firm can earn up to one-tenth of a cent of profit per kilowatt-hour sold.

The profit constraint is expressed as $\pi \leq kQ$, where k is the allowed profit per unit of output.[2] At any input combination, the maximum allowed profit is k times the maximum output that can be produced with the inputs. The constraint surface therefore takes the same shape as the production function, rescaled by k. Figure 2.2 depicts this surface. Each isoquant, which represents input combinations whose maximum output is the same, is a contour on the constraint surface,

---

[2] The constraint can also be expressed in terms of the firm's price. Because $\pi = PQ - wL - rK$, the profit constraint can be rewritten as $PQ \leq kQ + wL + rK$. Dividing through by Q, we have $P \leq k + (wL + rK)/Q$, or $P \leq k + AC$, where AC is average cost. That is, the firm can mark up price above average cost by at most the amount k—the allowed profit per unit of output.

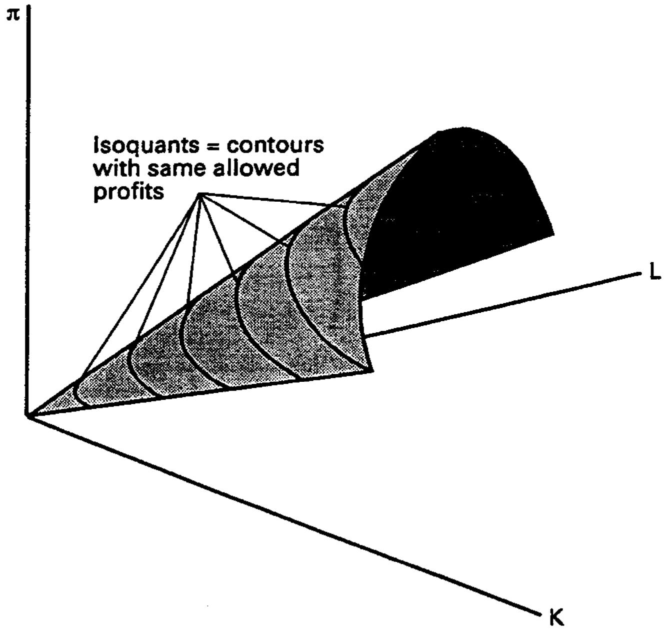

Figure 2.2
Maximum allowed profits under return-on-output regulation

representing the input combinations with the same maximum allowed profits.[3]

The constraint surface slices the profit hill, as depicted in figure 2.3. The firm chooses the highest point on the sliced-off profit hill, which is point *E*.

The position of point *E* can be visualized more readily by transposing the information in figure 2.3 onto the two-dimensional graph of input combinations, figure 2.4. The intersection of the constraint surface and the profit hill in figure 2.3 is the constraint curve in figure 2.4. This curve is the set of input combinations at which the maximum profit the firm is able to earn, given technology and demand, is equal to the maximum profit the firm is allowed to earn. At any input combination inside this constraint curve, feasible profit exceeds allowed profit; and vice versa for points outside the constraint curve.

---

[3] Note that the constraint surface as defined here gives the maximum allowed profits at each input combination. Allowed profits will be less than this maximum amount if the firm produces less output than is maximally feasible with given inputs, that is, if the firm wastes. Result 3 states that the firm under ROO regulation does not waste, such that the maximum allowed profits can legitimately be considered the constraint surface.

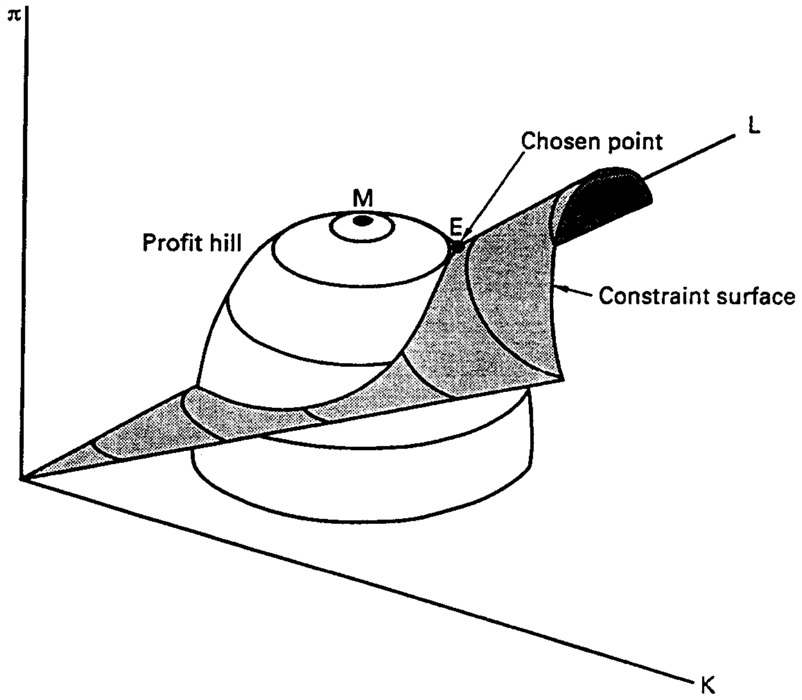
π
L
Chosen point
M
E
Profit hill
Constraint surface
K

Figure 2.3
Profit hill and constraint for firm under ROO regulation

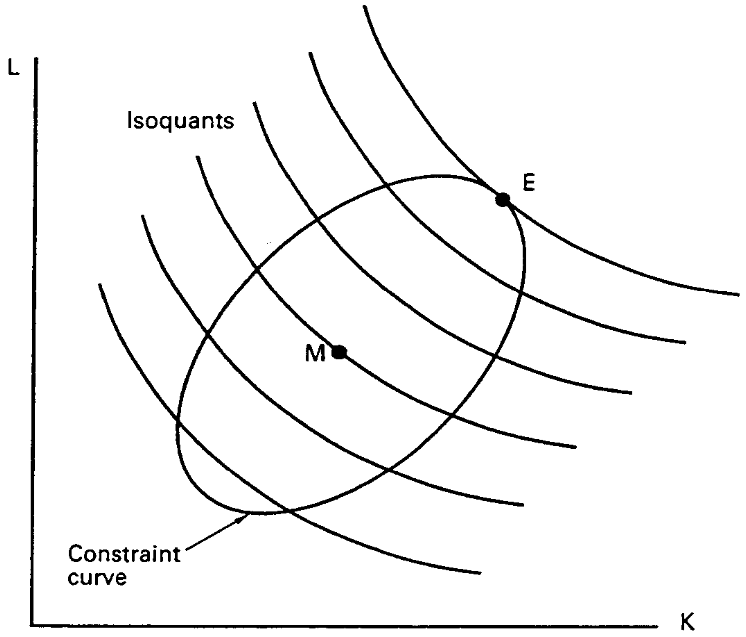
L
Isoquants
E
M
Constraint curve
K

Figure 2.4
Constraint curve and chosen point under ROO regulation

All points on the constraint curve provide the firm with the same profit per unit of output, namely k. Because profit per unit is constant on the constraint curve, absolute profits increase as output increases. The firm therefore chooses the point on the constraint curve with the greatest output, which is E. At this point there is tangency between the constraint curve and an isoquant. If the firm were to increase output beyond this point, going outside the constraint curve, the firm would be allowed to earn more profits but would not be able to.

Several results are now apparent.

Result 1: A firm under ROO regulation produces more output than if unregulated.

For the regulation to be binding, the constraint surface must slice off the top of the profit hill. The point M is therefore inside the constraint curve of figure 2.4. Because the regulated firm chooses the point on the constraint curve with the highest output, it necessarily chooses greater output than is produced at any point within the constraint curve.

Result 2: A firm under ROO regulation uses the efficient input combination for its level of output. That is, the firm produces on the expansion path.

Suppose the contrary, that the chosen point is not on the expansion path. This supposition is depicted in figure 2.5, in which the chosen point E is not on the expansion path. Consider point G, where the isoquant through E intersects the expansion path. Profits at G exceed profits at E because costs are lower at G (by definition of the expansion path) and revenues are the same at both points. However, G is outside the constraint curve, meaning that profits at G are less than k per unit of output. Because output is the same at G and E, and profits per unit of output are k at E and less than k at G, absolute profits at G are less than at E, contradicting the first comparison. Therefore, E cannot be off the expansion path.

Result 3: A firm under ROO regulation does not waste. That is, the firm produces as much output as possible with its inputs.

If the firm starts with no inputs and moves out the expansion path, its feasible profits increase until it reaches the top of the profit hill, after which feasible profits decrease. Figure 2.6 illustrates the relation between feasible profits and output with and without waste. The upper curve gives the profit the firm obtains at each level of output if it uses the cost-minimizing set of inputs. The lower curve gives the profit

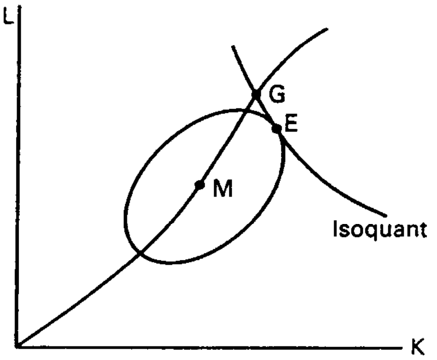

Figure 2.5
Firm under ROO regulation chooses point off expansion path: impossible

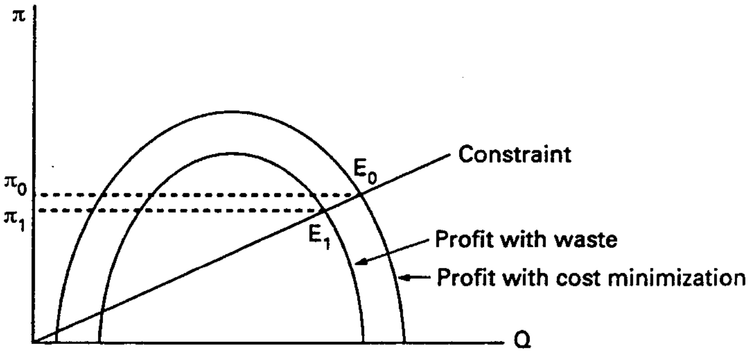

Figure 2.6
Firm under ROO regulation will not waste

the firm obtains if it uses more inputs than necessary to produce the output, that is, if it wastes inputs. Allowed profit, as represented by the constraint plane, increases with output and is the same for each level of output whether or not the firm uses more inputs than necessary to produce the output. Without wasting inputs, the firm chooses point E_0 and earns profit π_0. If the firm wastes, it chooses point E_1 and earns profit π_1. Since $\pi_0 > \pi_1$, the firm chooses not to waste.

Unlike the analogous result for ROR regulation, the fact that a firm under ROO regulation does not engage in pure waste does not depend on marginal revenue being positive. In fact, ROO regulation might induce the firm to increase output sufficiently such that it produces in the inelastic region of demand where marginal revenue is negative. Consider figure 2.7 in which Q_0 denotes the output at which marginal revenue is zero. For all output levels above Q_0, marginal

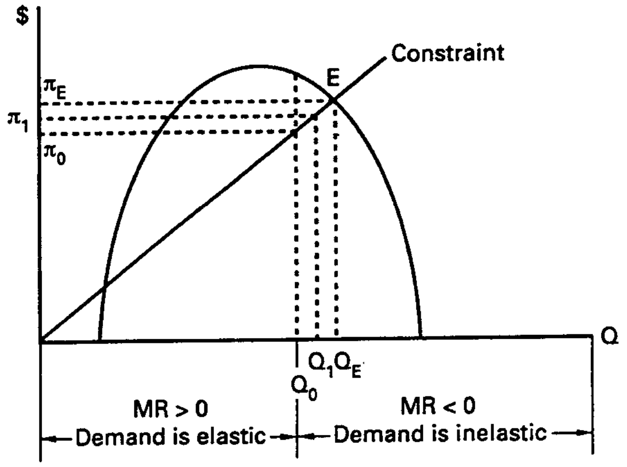

Figure 2.7
Firm under ROO regulation could produce in inelastic portion of demand

revenue is negative. The graph is constructed such that Q_0 is less than the output at which the constraint curve intersects the profit hill. (If Q_0 were beyond E, marginal revenue would be positive throughout the relevant range, such that the issue of negative marginal revenue does not arise.) We can show that the firm increases output beyond Q_0 into the inelastic region of demand. At Q_0, the firm is able to earn more profits than it is allowed to earn. If it remains at Q_0, it must therefore waste an amount of money equal to the difference between its allowed profit and its feasible profit, and ends up earning the allowed profits, π_0. If the firm were to increase its output to Q_1, its feasible profit would decrease because its revenues would decrease (marginal revenue becomes negative) and its costs would increase. However, its allowed profits would increase. The profit that the firm could keep would rise to π_1. The firm would therefore increase output until it reached the point at which its feasible profits equaled its allowed profits. Increasing output beyond this point would increase allowed profits but decrease feasible profits to a point *below* allowed profits, such that the firm would not choose to increase output further.

Another way of stating this argument is perhaps more straightforward. The most profit the firm could make if it stayed within the elastic portion of demand is π_0 in figure 2.7. By expanding output into the inelastic portion of demand, the firm can make greater profits, with a maximum of π_E.

Result 4: If the allowed rate of profit on output is lowered, the output of the regulated firm increases.

As *k* is lowered, the constraint surface slices off more of the profit hill. The new constraint curve therefore encompasses the original constraint curve, as depicted in figure 2.8. The firm moves out the expansion path to the intersection with the new constraint curve, increasing output and using inputs efficiently.

Result 5: If the allowed rate of profit on output is set at zero, the firm is indifferent among many input and output combinations, including the option of not using any inputs or producing any output.

The result is essentially the same as the analogous statement regarding ROR regulation, namely, that if the firm earns zero profit over a range of output and input levels, and cannot earn more than zero profit, it has no incentive to choose one outcome over another.

Recall that the goal of the regulator is to induce the regulated firm to operate at point *S* in figure 2.9, where the zero-profit contour intersects the expansion path. Result 4 indicates that the regulator, by lowering *k* toward zero, can induce the firm to move out the expansion path toward *S*. Result 5 implies that by lowering *k* all the way to zero, the regulator will not necessarily induce the firm to take the final step to *S*; the firm might choose to close down instead. Taken together, the results on ROO regulation indicate that the regulator can induce the firm to move arbitrarily close to the desired price and output level and to use the cost-minimizing input combination. That

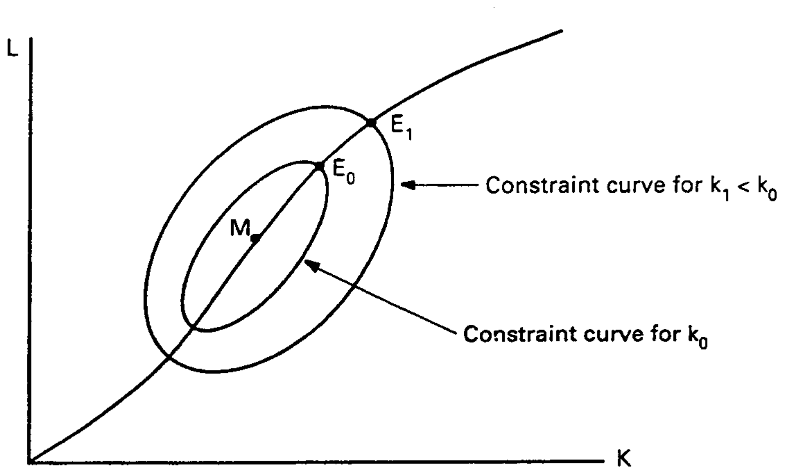

Figure 2.8
Output increases when *k* is lowered

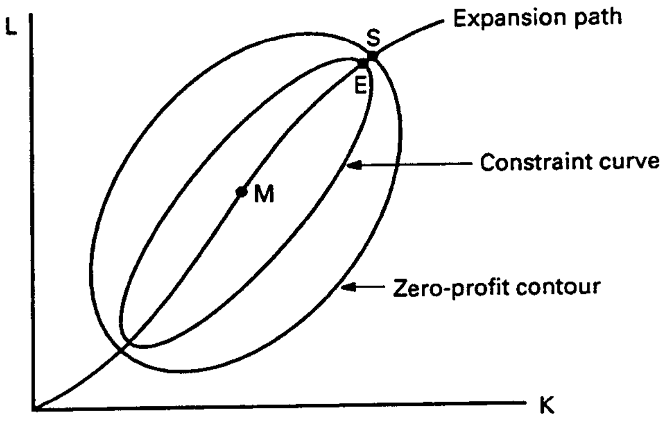

Figure 2.9
Firm's chosen point can be arbitrarily close to second-best

is, the firm can be induced to choose a point very close to *S* even though S itself cannot necessarily be attained.

This ability of ROO regulation to move the firm arbitrarily close to the desired input and output levels contrasts with the situation under ROR regulation. With ROR regulation, the firm cannot be induced to enter the inelastic portion of demand. Consequently, if point *S* is in the inelastic portion of demand, ROR regulation is not able to induce the firm to move close to *S* no matter what rate of return is allowed. Furthermore, ROR regulation does not move the firm along the expansion path, but rather induces the firm to operate with an inefficient input mix.

A difficulty with ROO regulation arises if the firm has the ability to influence its demand curve. If the firm can use advertising or other means to increase its demand, ROO regulation establishes an incentive to engage in these demand-stimulating activities. Conversely, if the firm has the ability to reduce its demand, ROO regulation gives it an incentive not to do so even if demand reductions are desirable from a social perspective. Conservation is an important case in point. Under ROO regulation, the firm would have an incentive not to undertake conservation programs that induce consumers to reduce their consumption even if these programs were cost-effective from a social perspective.[4]

---

[4] Cost effective in this context means that total surplus is greater if resources are expended on the programs that reduce demand than on producing the output needed to meet current demand.

2.3 Return-on-Sales Regulation

The revenues generated by a firm are often called its sales, or, more precisely, its dollar volume of sales. If the sales of a firm are easier to measure than its quantity of output, the regulator might want to use sales as the basis for determining allowed profit. Return-on-sales (ROS) regulation allows the firm to choose its outputs and inputs under the constraint that its profits do not exceed a portion of its revenues: $\pi \leq kPQ$, where k is the allowed proportion of revenues that can be retained as profit.

If marginal revenue is positive over the relevant output levels, then allowed profit increases with the quantity of output, because revenues increase. Consequently, the analysis of ROS regulation when marginal revenue is positive is essentially the same as that for ROO regulation. The conclusions are the same: if marginal revenue is positive over the relevant output range, ROS regulation induces the firm to increase output, not waste, and to choose the efficient input mix for its level of output. Furthermore, output increases as the allowed proportion of revenues that can be retained in profits decreases, such that the firm can be induced, by lowering the allowed proportion toward (but not to) zero, to produce arbitrarily close to the second-best output level, using cost-minimizing inputs.

If marginal revenue is negative, allowed profit drops when output rises, because revenue decreases. Consequently, ROS regulation differs radically from ROO regulation if the optimal output level is in the inelastic portion of demand. Under ROO regulation, the firm will not expand output into the inelastic region (where marginal revenue is negative) because doing so would decrease its allowed profit.

Figure 2.10 illustrates the situation. The profit hill denotes the maximum profits the firm can attain at each output level given demand and technology. Q_S is the optimal output, where price equals average cost such that profit is zero. Q_0 is the output level at which marginal revenue is zero. Revenues, and hence allowed profit, rise with output up to Q_0 (because marginal revenue is positive in this range) and then decline. Because the firm is allowed to earn more profit at Q_0 than at any other output, it chooses Q_0. Two conclusions are implied. First, the firm produces less than the optimal level of output:[5] Q_0 is less

---

[5] The allowed proportion k determines the height of the allowed profit surface at each output level, but does not affect the location of the top of this surface: revenues and

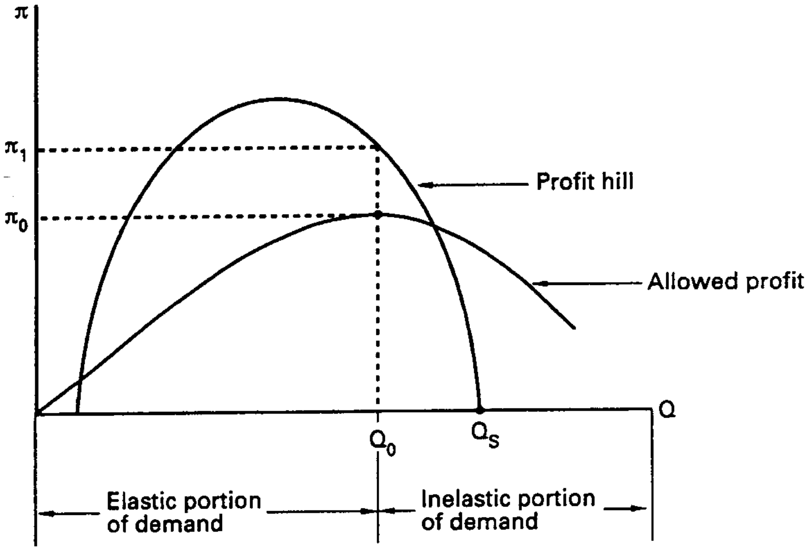

Figure 2.10
ROS regulation when marginal revenue becomes negative

than Q_S. Second, the firm does not cost-minimize in its use of inputs. At Q_0 the firm is able to attain profit of π_1 if it uses inputs efficiently; however, it is only allowed to earn π_0. To ensure that its profit does not exceed the allowed amount, the firm must use inputs inefficiently, through pure waste (that is, by producing less than is possible with the inputs) and/or by choosing an inefficient input mix. Costs exceed their minimizing level by the difference between π_0 and π_1. Furthermore, because k affects the level of allowed profits but not the firm's choice of output, lowering k simply increases the inefficiency costs.

The general conclusion is that ROS regulation induces desirable behavior on the part of the firm if marginal revenue is positive throughout the relevant range of output, but not if the optimal output is in the inelastic portion of demand.

2.4 Return-on-Cost Regulation

Allowed profit can also be based on the costs of the firm. Return-on-cost (ROC) regulation imposes a constraint on the firm of the form

---

thus allowed profit are highest where marginal revenue is zero for any value of k (insofar that k is sufficiently low such that allowed profits has a maximum within the profit hill).

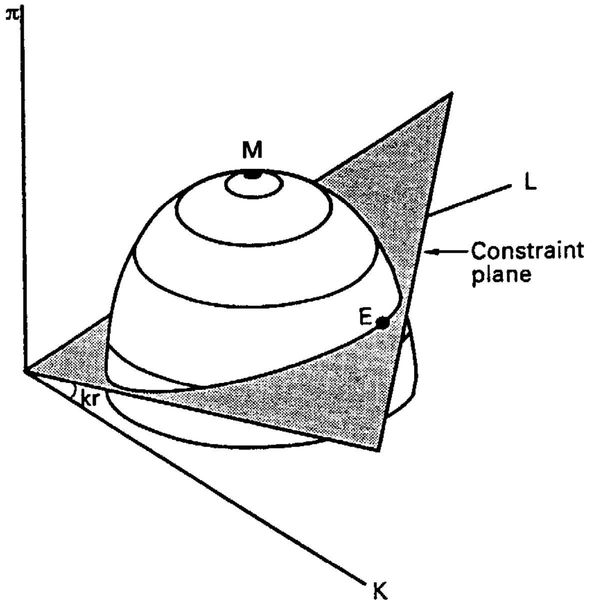

Figure 2.11
Firm under ROC regulation

$\pi \leq k(wL + rK)$, where k is the proportion of costs the firm is allowed to retain as profit.[6] As with ROS regulation, the implications of this form of regulation are very different if marginal revenue is consistently positive than if the optimal output falls in the inelastic portion of demand. Consider first the situation of positive marginal revenue throughout the relevant range.

The constraint surface is a plane with a slope kr in the capital direction and slope kw in the labor direction. The contours of this surface are the isocost lines, with "higher" isocost lines corresponding to greater costs and hence greater allowed profits. As shown in figure

---

[6] This form of regulation is equivalent to allowing the firm to mark up price over average costs by the proportion k:

$$\begin{aligned} \pi &\leq k(wL + rK) \\ PQ - (wL + rK) &\leq k(wL + rK) \\ PQ &\leq (1 + k)(wL + rK) \\ P &\leq (1 + k)AC, \end{aligned}$$

where AC is average cost. Note that this markup is different from that discussed in section 2.2 regarding ROO regulation. Under ROO regulation, the firm is allowed to mark up its price by a certain dollar amount over average cost. Under ROC regulation, the firm can mark up by a given proportion of average costs, such that the dollar amount of markup varies.

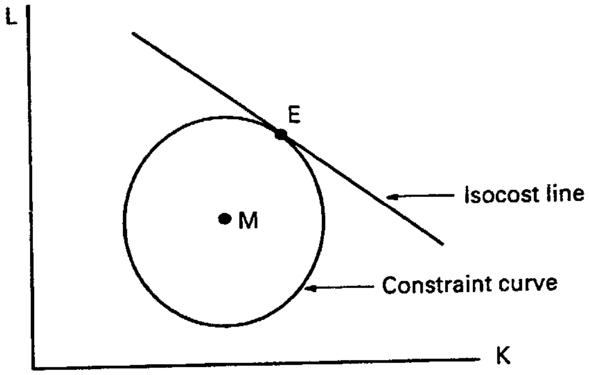

Figure 2.12
Choice of firm under ROC regulation: tangency of constraint curve with an isocost line

2.11, the constraint surface slices the profit hill. The firm chooses point E, the highest point on the sliced-off profit hill.

Figure 2.12 depicts the chosen point with the π dimension suppressed. Profits as a proportion of costs are the same for all points on the constraint curve, such that absolute profits are higher for those points with higher costs. The firm therefore chooses the point on the constraint curve that touches the highest isocost line, at which there is tangency between the isocost line and the constraint curve.

Knowing that the firm chooses this point of tangency allows us to readily demonstrate several results.

Result 1: A firm under ROC regulation and facing positive marginal revenue produces on the expansion path, using the efficient input mix for its level of output.

Suppose the contrary, as illustrated in figure 2.13. The firm chooses point E, which is not on the expansion path. Point H is the intersection of the expansion path with the constraint curve. Point G is the intersection of the constraint curve with the isoquant that goes through H. (Because the isocost at H cuts the constraint curve instead of being tangent, the isoquant also cuts the constraint curve. This isoquant therefore intersects the constraint curve at a second point, G.) Because G and H are on the same isoquant but H is on the expansion path, $\pi_H > \pi_G$. We can also show, however, that $\pi_G > \pi_H$. Because H is on the expansion path, which represents cost minimization, costs are lower at H than G. Both G and H are on the constraint curve, such that profit at each point is the same proportion of costs at each point.

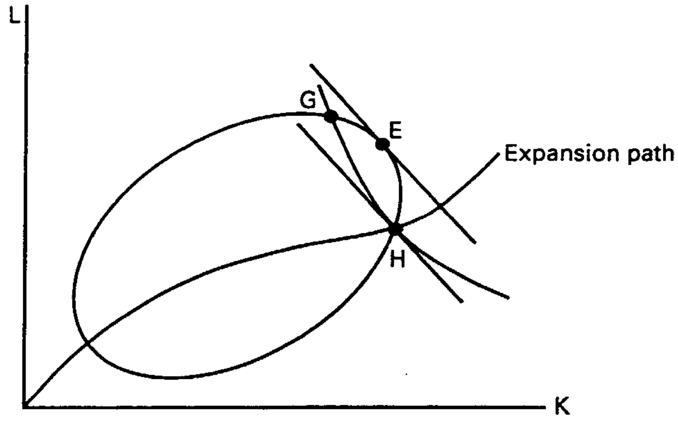

Figure 2.13
Firm chooses a point off the expansion path: impossible

Therefore, absolute profit is lower at *H* than *G*. This contradiction implies that the firm's chosen point must be on the expansion path.

Result 2: The firm under ROC regulation and facing positive marginal revenue produces more output than the unregulated firm.

Result 3: The firm under ROC regulation and facing positive marginal revenue does not waste inputs.

Result 4: If the allowed proportion k *of costs to be retained as profit is lowered toward (but not to) zero, then the firm under ROC regulation and facing positive marginal revenue increases output, using inputs efficiently.*

These results are straightforward applications of previous concepts. They are illustrated in figures 2.14, 2.15, and 2.16, respectively. It is interesting that the firm does not increase its costs through wasting inputs even when it is allowed to earn more profits by doing so. This result is critically dependent on marginal revenue being positive. If the firm purchases nonproductive inputs (that is, if it wastes), then its allowed profit increases but its feasible profit decreases (because costs increase without an increase in revenues). If instead the firm uses the same amount of money to purchase inputs but uses them productively, allowed profit rises by the same amount and yet feasible profit either rises or drops by less (because revenues increase, at least partially offsetting the cost of the extra inputs). ROC regulation gives the firm an incentive to increase costs, but as long as marginal revenue is positive, the firm earns greater profit by increasing output as much as possible along with costs.

If marginal revenue becomes negative within the relevant range of output, then the cost-based incentive does not translate into a quantity-based incentive. If marginal revenue is negative, the firm loses revenue by selling extra output and gains revenue by selling less output. As a result, the firm is able to earn greater profits by selling less even without reducing inputs: its allowed profit does not change and its feasible profit increases. Conversely, the firm can increase its allowed profit by purchasing inputs (whether productive or not), and yet its feasible profit decreases less when inputs are purchased without increasing output than when using the inputs to produce more.

The firm will increase output beyond its unregulated level only to the point that marginal revenue is zero. If allowed profit exceeds feasible profit at this point, the firm purchases nonproductive inputs, increasing allowed profit while decreasing feasible profit as little as possible. (If the firm used these extra inputs to produce extra output, its feasible profit would decline even more.) Figure 2.17 illustrates the situation. In this graph, the level of labor is assumed constant. As capital is increased, output and revenue increase until marginal revenue is zero. This point is labeled K_0. If capital is increased beyond this point and the capital is used to produce and sell extra output, the firm's feasible profits decrease along the downside of the profit hill. However, if the firm purchases extra capital but does not produce more output, its feasible profits decrease by less. (By not selling extra output, the firm's revenues do not decrease.) The profit that the firm can obtain by using extra capital but not selling extra output is given by the downward sloping line that starts at capital level K_0. The slope of this line is the cost of capital r: for each extra unit of capital purchased, profits decrease by exactly r. If extra output were produced with the extra capital and sold, then profit would decrease by r plus the decrease in revenues that results from selling the extra output.

The firm in this situation chooses to produce the output at which marginal revenue is zero and yet purchase K_E capital. Because only K_0 capital is needed to produce the level of output, the firm wastes the difference between K_E and K_0. The same type of waste also occurs with labor.

In summary, if the optimal output is in the inelastic portion of demand, then the ROC regulation can be used to induce the firm to increase output and use cost-minimizing inputs only to the point at which marginal revenue is zero. Any attempt to induce the firm to

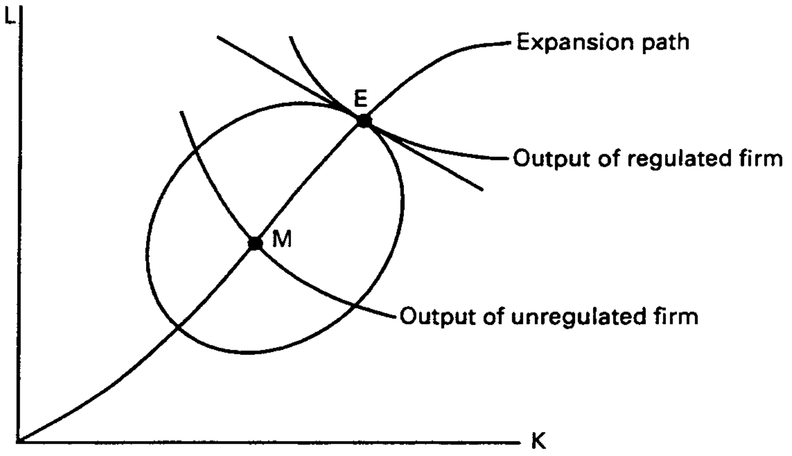
L
E
Expansion path
Output of regulated firm
M
Output of unregulated firm
K

Figure 2.14
Firm under ROC regulation will produce more output than if unregulated

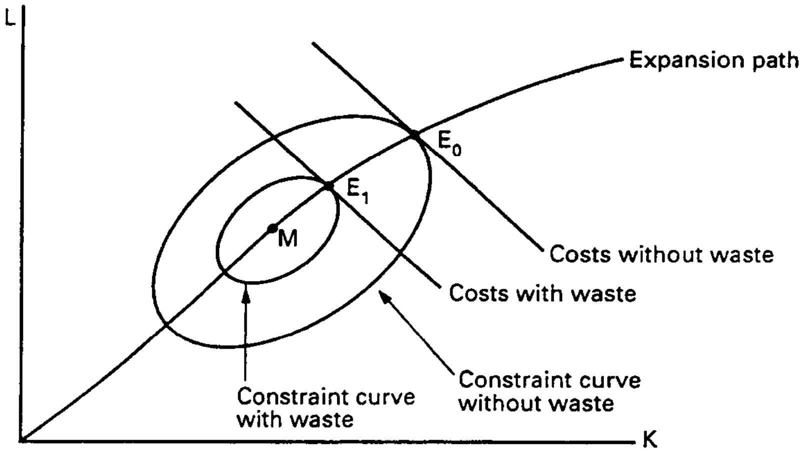
L
Expansion path
E_0
E_1
M
Costs without waste
Costs with waste
Constraint curve with waste
Constraint curve without waste
K

Figure 2.15
Firm under ROC regulation will not waste inputs

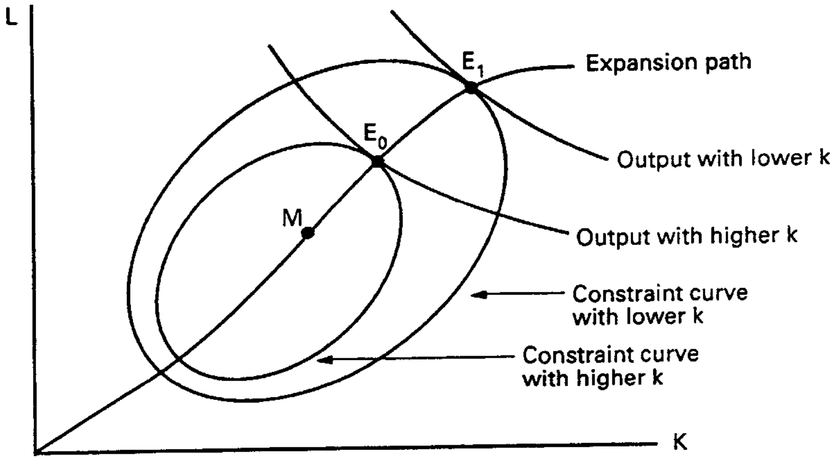
L
K
E$_1$
E$_0$
M
Expansion path
Output with lower k
Output with higher k
Constraint curve with lower k
Constraint curve with higher k

Figure 2.16
Firm under ROC regulation will increase output when allowed profit is lowered

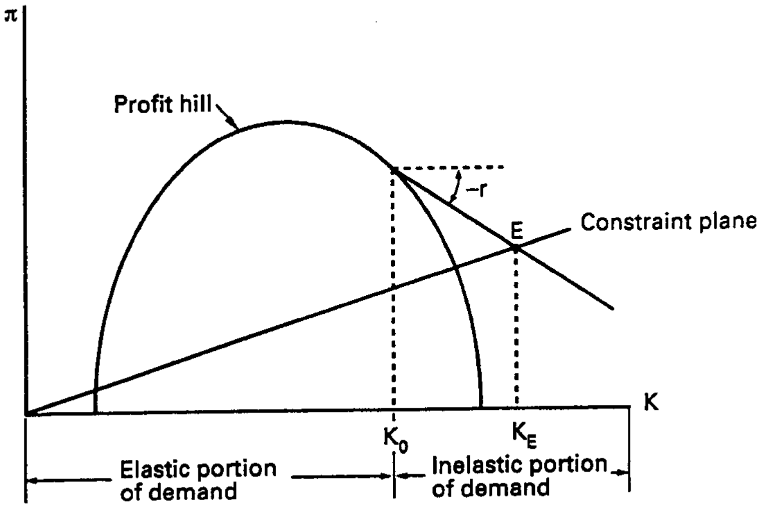
π
Profit hill
-r
E
Constraint plane
K
K$_0$
K$_E$
Elastic portion of demand
Inelastic portion of demand

Figure 2.17
ROC regulation

increase output further, by lowering the allowed profits, simply induces the firm to waste.

2.5 Price Discrimination

Price discrimination exists when a firm charges different prices to different customers and/or for different units of output (e.g., one price for consumption up to a certain quantity and then another price for additional units). Primary price discrimination (also called perfect price discrimination) is defined as a situation in which the firm charges each unit of output at exactly the amount that a consumer is willing to pay for that unit.

Under the regulatory mechanisms described so far, the regulated firm is assumed to charge one price to all of its customers and for all units of output. It has long been recognized (for example, Robinson 1933)[7] that price discrimination by a monopolist results in greater total surplus than if the firm charges only one price. In fact, as shown below, a firm that is able to engage in primary price discrimination chooses the first-best output level and uses cost-minimizing inputs. This fact suggests that a potentially effective form of regulation is for the regulator to allow and assist the firm in price discrimination.

Price discrimination is not always possible, and even if possible, it might violate goals that the regulator holds in addition to the objective of inducing optimal input and output levels. For example, primary price discrimination results in all surplus accruing to the firm and none to consumers, which might not be considered equitable. Furthermore, different customers are required to pay different prices for the same good or service, which can also be considered inequitable. However, before addressing these limitations, let us demonstrate the fact that primary price discrimination leads to the first-best outcome in an efficiency sense.

Consider first the decision process of a non-price-discriminating monopolist, as depicted in figure 2.18. At any level of output, the firm must lower its price in order to sell additional units of output, because market demand is downward sloping. Consequently, marginal revenue is below price at each level of output. The firm maximizes profit by expanding output whenever marginal revenue exceeds marginal cost, eventually choosing the output at which marginal revenue equals

---

[7] More recent treatments are provided by Schmalensee (1981) and Varian (1985).

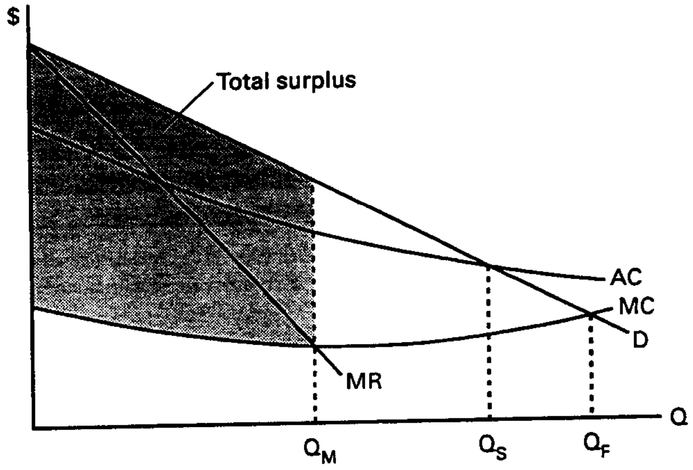

Figure 2.18
Non-price-discriminating monopolist

marginal cost. This output is labeled Q_M, and at this output total surplus (excluding fixed costs) is the shaded area.

The monopolists' output is below the socially optimal level and total surplus is less than maximal. The first-best output is Q_F, at which price equals marginal cost. At this output, the total surplus is the area below the demand curve and above the marginal cost curve up to Q_F. If the firm is a natural monopoly, marginal cost is below average cost, and the firm would lose money at the first-best output. The second-best output is Q_S, which is the largest output consistent with non-negative profits.

Suppose that the firm is able to engage in primary price discrimination, pricing each unit separately. For each unit of output, the firm charges the maximum that any customer is willing to pay for the unit. The firm sells its first unit of output to the customer most willing to pay for that unit. In figure 2.19, this price is P_1, because at a price of P_1 one unit of output is demanded. Then the firm sells its second unit to the customer who is second-most willing to pay; the price for this unit is P_2. Note that the firm still charges P_1 for the first unit: it does not have to lower its price for previous units in order to sell more units. The firm sells extra units whenever the price it can receive from an extra unit, as indicated by the demand curve, exceeds the extra cost incurred in producing the extra unit; that is, whenever demand exceeds marginal cost. The firm chooses the output at which demand equals marginal cost, which is the first-best output Q_F.

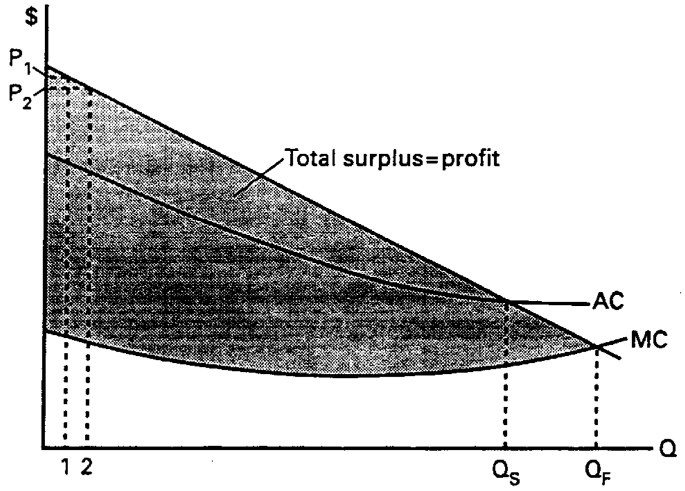

Figure 2.19
Monopolist with primary price discrimination

The firm in both situations sells extra output whenever the extra revenue it obtains exceeds the marginal cost of producing the extra unit. When only one price is charged for all units, the extra revenue from an extra unit is less than the price the firm can obtain for that one unit, because price must be lowered on all units of output, not just the marginal unit. However, when each unit can be priced separately, the extra revenue from selling extra output is exactly the price the firm receives for the one extra unit, since the prices for other units do not change. Essentially, the demand curve is the marginal revenue curve under primary price discrimination. The firm maximizes profits by choosing output at which marginal revenue equals marginal cost, which in the case of price discrimination is where demand equals marginal cost.

The basic task of regulatory economics is to ensure consistency between the firm's goal of maximizing profit and the regulator's goal of maximizing surplus. The manner by which primary price discrimination attains this consistency constitutes a fundamental solution to the problem. Under primary price discrimination, the firm extracts all surplus: it charges each customer exactly the customer's willingness to pay. Because the firm obtains all surplus as profit, profit maximization and surplus maximization are identical: the firm naturally chooses the optimal outcome. This transfer of all surplus to the firm is clearly the most straightforward way (at least theoretically) to provide consistency between the regulator's goal and the firm's profit

Note that primary price discrimination, unlike the other regulatory mechanisms discussed in this chapter, brings the firm to the first-best, not second-best, output. The possibility of attaining first-best is an interesting consequence of price discrimination. Without price discrimination, the largest output the firm can produce and not lose money is the second-best level, at which price equals average cost and profits are zero. However, with a different price for each unit, the firm is able to earn larger profits at any level of output sold. Consequently, the firm earns positive profits at this second-best output. (Its profits are the area below demand and above marginal cost up to this output level.) Because profits are positive at this output, the firm can expand output without its profits becoming negative. In fact the firm makes *more* profit by expanding output beyond the second-best level. At Q_S, the price the firm can charge for an extra unit (as given by the demand curve) exceeds the cost of producing the extra unit (as given by the marginal cost curve). As a result, primary price discrimination allows, and induces, the firm to produce more output than would be possible (given that the firm cannot lose money) under regulatory mechanisms that require one price for all units of output.

The potential advantages of price discrimination are obvious. There are, however, some limitations, both practical and ethical.

1. The existence of price discrimination provides an incentive to customers to establish resale markets, which undermine the monopolist's attempt to price discriminate. A customer that is charged a low price by the monopolist would, if possible, sell the units to a customer who is being charged a higher price by the monopolist. The monopolist would find itself selling only to the low-priced customer, because the customer that is charged the higher price would buy at resale from the low-priced customer. The attempt to charge different prices would therefore result in sales at only one price, namely the lowest price offered to any customer.

The monopolist might be able to prevent resale. For example, a monopolist in trash collection can reduce the potential for resale by placing a limit on the number of barrels of trash that will be collected from each customer.[8] However, in many situations, resale cannot be readily prevented by the monopolist.

---

[8] Without this limit, any customer that is charged a lower per-barrel fee than his neighbors has an incentive to charge his neighbors to put their barrels in his collection area, with the charge being below the per-barrel charge of his neighbors but above his own per-barrel charge. With the limit on number of barrels, this resale of service would be restricted.

The prevention of resale is generally more feasible for regulated than unregulated firms. In a regulated setting, the regulator—usually an arm of the government—can establish and, more important, enforce rules against resale to an extent not possible by the firm itself. Consequently, a regulator that wishes to use price discrimination as a way to induce first-best output and input levels might find it possible to enforce the discrimination even if the firm itself could not.

2. To extract the entire surplus, the firm must know the willingness to pay of each customer for each unit of output. This amount of information is usually beyond the scope of most firms. However, the firm need not actually extract all surplus in order to produce the first-best output. Suppose the firm knows whether or not it can sell an extra unit of output for more than its marginal cost, though it does not necessarily know the maximum that it can obtain for the unit. In this case, the firm sells extra output whenever a customer is willing to pay more than the marginal cost. The firm therefore produces the first-best output level even though it attains less than the full surplus.

3. Assuming the firm has full information, primary price discrimination results in the firm making large profits, consisting of the entire surplus. Thus, even though the optimal output is attained, the benefits of attaining this output all accrue to the monopolist. The regulator might consider this distribution of benefits to be inequitable.

In theory, it is possible to tax the monopolist at a rate that is a fixed proportion of its profits, and then refund the tax revenues to the firm's customers. In this way, the surplus is shared between the firm and its customers. The firm's actions would be the same with or without the tax since the output and input levels that maximize its profits also maximize, say, 50% of its profits. Of course, the issue then arises of whether the regulator has the authority to tax the firm and how the tax funds can be distributed to consumers without the distribution essentially changing the price consumers pay for the firm's output.[9]

4. It might be considered inequitable for different customers to pay different prices for the same goods or services. That is, the basic premise of price discrimination might conflict with the regulator's goals re-

---

[9] If the size of the refund to each customer is based on the number of units purchased or the total amount paid by the customer, then the refund constitutes a change in price. If, however, an equal refund is made to all of the firm's customers independent of consumption level or payments, then people who would not buy the firm's output without the refund would choose to buy a small quantity in order to obtain the refund.

garding equity, even though the regulator's goals regarding input and output are met.

Regulators usually allow different prices to be charged for different groups of customers. For example, electricity is priced differently for residential customers than for commercial and industrial customers; and among residential customers, different prices are often charged for electrically heated households than for those with gas or other heat. This implies that regulators often consider it equitable to price discriminate on the basis of some factors. The question is therefore not whether price discrimination is equitable per se, but rather the extent to which the regulator can equitably distinguish customers for the purpose of price discrimination.

The practical and ethical difficulties of primary price discrimination are formidable. Our purpose in describing price discrimination is not necessarily to recommend it as a form of regulation. The point is rather to illustrate the concept that consistency between social goals and the firm's profit drive can most simply be attained by enabling the firm to secure as profit *all* social benefits (and then, perhaps, taxing and redistributing these benefits). In chapter 6, regulatory mechanisms are introduced that utilize this concept in a fashion that can be more equitable and yet just as effective from an efficiency perspective.