22 Regulating Prices in Natural Monopolies

This and the following four chapters are concerned with economic regulation of the utilities sector—including communications, energy, transport, and water. These are often described as network industries, because they rely upon some form of physical distribution network, such as gas or water pipes, electricity or telecommunications wires, train tracks or a postal delivery round.[1] This chapter focuses on the nature and implications of natural monopoly in such industries. Natural monopoly can be defined as a situation in which the market can most cheaply be supplied by a single firm. (A gas distribution network is a good example.) A natural monopolist, left to itself, would for reasons discussed below be likely to charge excessive prices, and there is accordingly a need for some form of regulation of price and quality, and scope for debate about the kind of regulation that is appropriate.

Historically, utility companies in Europe and elsewhere (excluding the United States) were typically government departments or public enterprises, undertaking all aspects of production of the service in question in a vertically integrated manner. Thus the same gas enterprise often owned the gas fields, landed the gas, piped it to homes and businesses, and retailed it to customers.

When privatization occurred, as it has in many countries, the starting point for regulatory discussion was often how to regulate a privately owned vertically integrated firm which either was a monopoly by law, because other firms were prohibited from entering the relevant market, or was a monopoly in practice because competitors were unable to establish themselves in the face of the market power wielded by the 'historic monopolist' or incumbent. In these circumstances, regulators saw their task as controlling the prices of a vertically integrated monopoly, by setting a retail price the firm could not exceed.

It has been a key insight of utility regulation in the past twenty or so years that not all of these activities undertaken by a utility are natural monopolies. Thus while the gas distribution network may be a natural monopoly, other activities in the 'value chain' such as retailing or extracting gas, need not be. Once this is recognized, decisions have to be made about the degree to which

[1] Networks can be defined more broadly to include virtual networks, such as the set of users of a particular computer operating system. Here we focus on physical networks.

the separate activities undertaken by utilities can be broken down or 'unbundled', about where the market can be liberalized by allowing competitors to enter, and about the terms on which competitors are permitted to have access to the incumbent's natural monopoly assets. This may lead to a reorganization of production so that different firms or divisions are created or separated to perform different tasks.

Whether we are looking at the problem of regulation as one of controlling retail prices paid by a vertically integrated monopolist, or the prices at which competitors have access to the incumbent's facilities, there is a general problem of setting prices in a context with elements of natural monopoly. This chapter contrasts the outcome in the absence of regulation with what price regulation can achieve. It shows how closely such regulation can get to efficient pricing based on marginal costs. The next chapter deals with the use of competition in network industries. The following three chapters then deal, respectively, with issues of contestability and separation (Chapter 24), the implementation of price controls (Chapter 25), and questions of efficiency and innovation (Chapter 26).

What is a Natural Monopoly?

A natural monopoly arises when the market is served most cheaply by a single firm, rather than by a multiplicity of competing firms. In cases where the firm is producing a single product or service—for example, the distribution of gas to homes in a particular location—the situation can be represented as in Figure 22.1, which shows how average cost (AC) per unit house served falls as the number of homes served increases. The fundamental reason is that laying down the gas distribution network accounts for the majority of the costs; connecting an individual home to it involves a relatively small additional expense.

Thus the declining average cost of service derives from the fact that the marginal cost of serving a new customer is less than the average cost. This situation is shown in Figure 22.1 by the marginal cost curve (MC).

Possible sources of declining unit costs are many.[2] In the case of pipelines, the capacity of the pipe can be increased without a commensurate increase in investment cost. Firms with a larger scale of operation may also be able to reduce costs by having proportionately lower overheads or by being able to employ more specialized and efficient personnel. In the case of distribution

[2] See D.A. Hay and D.T. Morris, *Industrial Economics and Organisation* (2nd edn, Oxford, 1991), ch. 2.

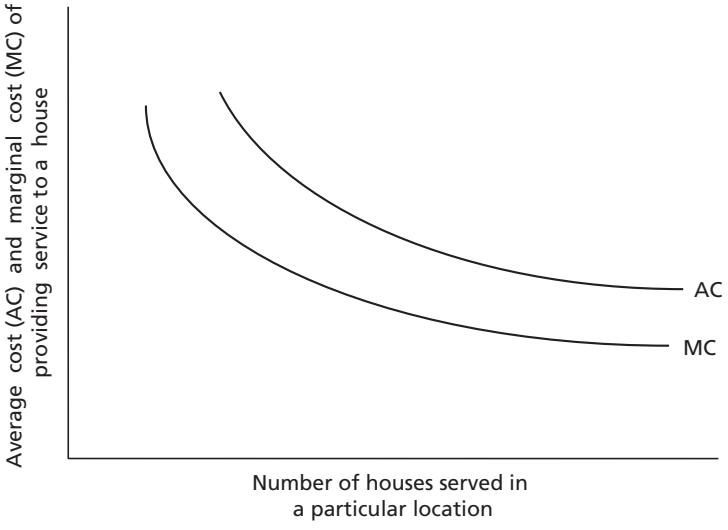

Figure 22.1. Average and marginal cost in a natural monopoly
As the number of houses served grows, there is a fall in the average cost (AC) of service—both operating costs and investment or capital costs. This implies that the cost of serving an additional customer (the marginal cost or MC) always lies below the average cost: what drags the average down is the (low) additional cost of serving a new home.

networks for electricity, gas, telecommunications, posts, and water, the reason for declining costs is what can be called an economy of density. Thus, it is cheaper on a per household basis for a single distribution company to deliver gas to all the houses in an area than to have two competing networks each serving half of them. This is because the latter arrangement requires unnecessary duplication of a major part of the distribution network.

In other areas of utilities, such as high-voltage electricity transmission, capacity can be increased without a commensurate increase in costs. Thus the same pylons can carry more or less electricity, and the costs of burying cable are insensitive to the number of kilowatts transmitted.

The degree to which an activity exhibits the characteristics of a natural monopoly depends not only on engineering factors, but also upon management processes and the operation of social and economic factors within the enterprise. It might theoretically be possible for a monopolist to serve a market at a lower unit cost than two or more competing firms can achieve. But incentives to efficiency under a monopoly may be very weak, and as a result, it may in practice be cheaper to have the market supplied by two competitors in spite of the theoretical advantage of the monopoly.

The declining unit costs associated with economies of scale of the kind described above are one aspect of natural monopolies. A second factor is economies of scope, which are encountered in many industries when it is cheaper for one firm to provide two or more related products and services together, than for each of them to be provided by a separate firm. A good example from the communications industry is provided by telecommunications and cable television networks, each of which can deliver voice calls, broadband, and video entertainment services.[3] Economies of scope, which typically arise from the use of common assets to produce separate products, are likely over time to have the effect of reducing the number of firms in an industry.

The tendency towards natural monopoly is most pronounced when economies of scale and density are combined with economies of scope. The former reduce the number of firms producing each service individually, while the latter encourage each firm in the market to produce a range of services. Acting in combination, they may generate a situation in which a significant number of markets are served by the same monopolist.

Determining whether a particular area of activity is a natural monopoly is a complex process. Natural monopolies are vulnerable to technological development. Thus, the argument that telecommunications, particularly the access network or local loop which connects households and firms to the local exchange, is a natural monopoly has been significantly weakened by the development of new technologies based on wireless distribution. These give customers access to the exchange without the necessity to construct fixed-link networks. The natural monopolies of energy and water distribution systems, however, appear to be well rooted.

Basic Concepts for Regulated Pricing

The implications for pricing of services provided by a natural monopoly can be tackled by asking two questions: what price would emerge in the absence of intervention, and what prices should regulation try to attain? The first

[3] Telecommunications services (at that time mainly voice calls) and cable TV services were initially provided, for technical reasons, over quite separate networks. From about 1990, technical developments permitted a single network to provide both.

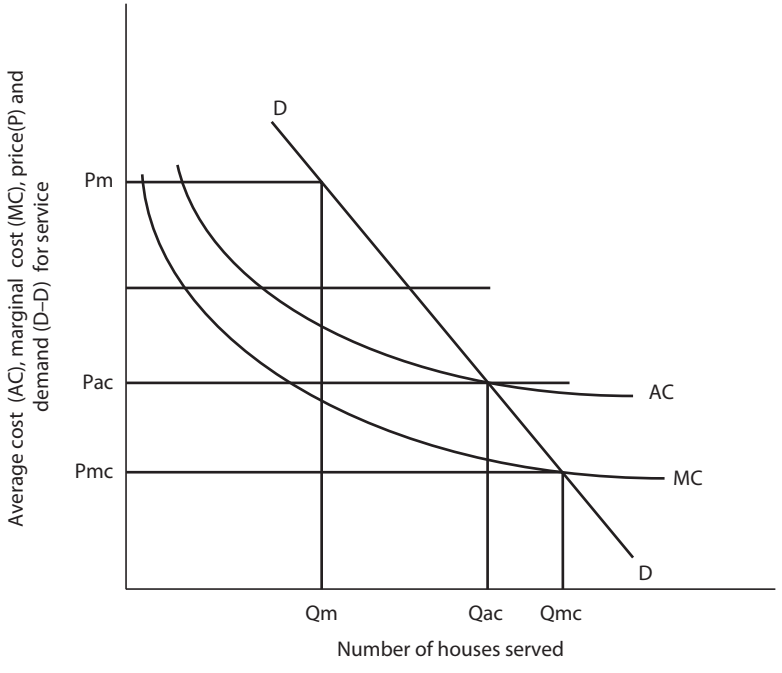

Figure 22.2. Pricing options for a natural monopolist
An unregulated monopolist would charge a high price (Pm), thus raising the price of service to the household. The ideal price would be Pmc, where the demand curve (D–D) cuts the marginal cost curve (MC). If this price were charged, the price to the household would be the true marginal cost to the economy of providing the service. A household making the decision whether to connect would thus face a price reflecting the resource cost to the economy of making the connection. However, a price equal to Pmc would fail to cover the firm's average cost (AC); hence the firm could not survive in the long run. The lowest price consistent with the firm breaking even is Pac. If the regulator must ensure that the firm breaks even—to meet the expectation of its private-sector owners—and if more sophisticated pricing strategies are not available, this is the best feasible price.

question can readily be answered in relation to Figure 22.2. This reproduces the shape of the curves in Figure 22.1, with the addition of a demand curve DD, which shows how demand for, say, gas distribution services varies with the price charged.

If the price of the gas distribution is higher, the implied price of gas delivered to consumers will be higher, and gas consumption will diminish. As the price of gas distribution services falls, this will be reflected in lower prices at the retail level, and demand for gas will rise.

In these circumstances, a monopolist controlling a distribution network will maximize its profits by setting a relatively high price, Pm, which lies

above average cost and hence delivers a monopoly profit.[4] As a result, gas prices paid by consumers will be high, and those consumers will suffer, to the benefit of shareholders in the monopoly, who will enjoy excess profits.

This unsatisfactory state of affairs can clearly be mitigated by the regulation of prices, but what price for gas distribution should the regulator set? Ideally the prices of goods and services sold in the economy should be set at their marginal costs,[5] whether they apply to final demand such as gas purchased by households or to an intermediate product such as gas distribution. This is desirable because at a price (Pmc in Figure 22.2) where the demand curve cuts the marginal cost curve, output has been expanded up to the point where the buyer's willingness to pay for an additional unit of the service provided, shown by the height of the demand curve, exactly equals the marginal cost to the economy of producing that final unit of output. At a price higher than this, the buyer's willingness to pay would exceed the marginal cost of providing an extra unit. At a price lower than this, the marginal cost to the economy of providing the last unit of output is greater than the buyer's willingness to pay for it. The best price for the service is, therefore, a price equal to marginal cost.

As inspection of Figure 22.2 demonstrates, however, if the service were priced at Pmc, then the price charged would fail to cover average cost. As a result, the firm would make a loss.

If it were a public enterprise, that loss could be made up from general taxation. A privately owned single product firm which did not receive state aid or another form of subsidy would, however, go out of business. If the firm is constrained to avoid losses and break even, then the most appropriate regulated price is shown by Pac in Figure 22.2. This is more satisfactory than the monopoly price Pm, but less efficient than a price equal to marginal cost, Pmc.

The implication is that a regulator who is setting prices for a single-product firm which has to break even should seek to drive prices down to average costs.

Most regulated firms, however, produce several services, and this gives more flexibility in the pricing process. A traditional example comes from telecommunications, where end users pay a monthly rental for connection to the network and then pay additional sums for making calls. It is not possible in the case of a multi-product firm to identify individual average costs for the separate services, because those services will typically have common inputs such as switches in the local exchange, and as a result it will not be possible to attribute all costs unambiguously to individual services. It will, however, be

[4] For how that price is determined, see D. Begg, G. Fischer, and R. Dornbusch, *Economics* (8th edn, London, 2005), Ch. 8.

[5] See Armstrong, Cowan, and Vickers, *Regulatory Reform: Economic Analysis, and British Experience*, 14–18.

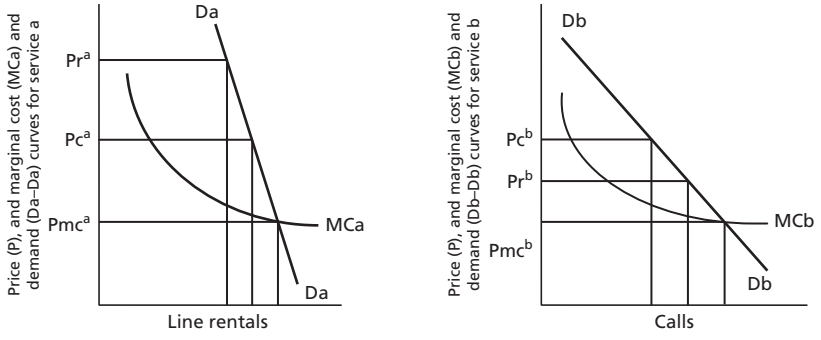

Figure 22.3. Efficient pricing for a multi-service utility
The ideal prices are where the demand curves (Da–Da and Db–Db) cut the marginal cost curves (MCa and MCb for each service). A firm charging such prices would, however, make losses, as both products are produced in conditions of economies of scale. Hence the need for a mark-up over marginal costs. One possibility would be to have an equal mark-up to cover common and fixed costs; i.e. to choose Pc^a and Pc^b. Such prices, though, have different distorting effects on demand for the two services; it falls much more for calls than it does for line rental. A preferred option is so-called Ramsey pricing, which involves a high proportionate mark-up on line rentals, where demand is unresponsive to price (Pr^a), and a low proportionate mark-up on calls, where demand is more responsive to price (Pr^b).

possible to establish the marginal cost of each service, by identifying the increase in overall costs associated with increasing the output of any service when the output of other services is held constant.

As before, the most efficient price for each service occurs where the demand curve Da Da or Db Db crosses the marginal cost curve (MCa or MCb), as illustrated by Pmc^a and Pmc^b in Figure 22.3.

We face once again, however, the problem that, if the firm sells each service at a price equal to its marginal cost, which lies below average cost (not shown), the firm will incur losses. In order to break even, it must, therefore, charge a mark-up above marginal costs.

One simple way of achieving this objective would be to fix prices that contain an equal proportionate mark-up on each service, of a size which just allows the firm to break even. These prices are shown in Figure 22.3 as Pc^a and Pc^b, respectively. This is the solution consistent with average cost pricing for the single-product firm. It is preferable in most circumstances, however, to set a proportionate mark-up over marginal cost for each service which varies from service to service in accordance with demand conditions.

These preferred prices are illustrated in Figure 22.3 by Pr^a and Pr^b; they are also known after their inventor as Ramsey prices.[6] The logic behind them is as

[6] Armstrong, Cowan, and Vickers, *Regulatory Reform: Economic Analysis, and British Experience*, 47–51; S.J. Brown and D.S. Sibley, *The Theory of Public Utility Pricing* (Cambridge, 1986), 39–44.

follows. In the case of line rentals, demand is relatively unresponsive to price, and a high mark-up can be charged without that mark-up having a major effect on consumption, compared with the case where price is equal to marginal cost. Demand for calls, by contrast, falls much more as price rises. A high mark-up on calls will lead to a major distortion of the amount consumed.

To express this principle more generally, when prices are being set for a regulated monopoly which produces a variety of services and which—because it is privately owned—is required to break even, they should embody the minimum mark-ups over marginal costs that are necessary to allow the firm to break even. Services where demand is relatively responsive to price should generally have a lower than average proportionate mark-up, while services where demand is relatively unresponsive to price should have a higher than average mark-up over their marginal costs. This enables common costs to be recovered in a way that reduces to a minimum the harmful effects of distortion in output caused by the mark-up over marginal cost.[7]

As it happens, for many decades telecommunications regulators did the opposite of what the principle set out above implied. That is, they imposed a low mark-up on line rentals, and a high mark-up on calls. Why was this? At least part of the reason was a desire to encourage the spread of telecommunications, to be achieved by making as many households subscribe as possible. This universal service objective, discussed in Chapter 24 below, in effect took priority over the requirements of economic efficiency. This is a useful reminder of the variety of goals which affect pricing by regulators.

The institutional setting for the kind of price control described is one in which the regulatory body is implicitly in centre stage. Its job is to find a set of permitted prices which balances the interest of investors in utilities and consumers of their services, and we have set out above some of the concepts which can be adopted. We return in Chapter 25 to consider how this might be done in practice, and, also to describe an alternative approach in which price setting is accomplished by a form of negotiation between suppliers and customers or their representatives.

Conclusions

This chapter has defined natural monopoly, which occurs when a market is most cheaply served by a single producer. Natural monopolies arise from

[7] For an implementation of Ramsey pricing to the telecommunications sector, see Brown and Sibley, *Theory of Public Utility Pricing*, ch. 7.

economies of scale, which mean that the largest firm has a cost advantage over its competitors, and is hence likely to become a monopolist. Where two or more products or services are produced more cheaply by a single firm than separately by two firms, economies of scope are in question. A combination of economies of scale and economies of scope is likely to lead to dominance of the market by a single multi-product firm.

Such a firm has the market power to charge prices which generate excessive profit. The natural regulatory response is to control prices. In the case of a single-product firm, if the firm is required to break even, the most satisfactory regulated price which can be imposed at any point in time is equal to average cost. In the case of a multi-product firm, a break-even constraint should lead to differential mark-ups on services. It has been suggested that such mark-ups should be greater where demand is relatively unresponsive to price and smaller where it is relatively responsive.

This analysis has allowed us to identify what might be efficient pricing rules for a natural monopoly. Utility regulators in practice have to undertake the prior, and crucial, process of determining whether regulation of price and other aspects is in fact necessary. Our discussion has also assumed that regulators know the costs of the firms they regulate. In practice they do not, and they need to develop incentives for firms to show how they can reduce their costs and keep them down. These issues are discussed in the following chapters.