1
The Determinants of Market Outcomes

A solid knowledge of both econometric and economic theory is crucial when designing and implementing empirical work in economics. Econometric theory provides a framework for evaluating whether data can distinguish between hypotheses of interest. Economic theory provides guidance and discipline in empirical investigations. In this chapter, we first review the basic principles underlying the analysis of demand, supply, and pricing functions, as well as the concept and application of Nash equilibrium. We then review elementary oligopoly theory, which is the foundation of many of the empirical strategies discussed in this book. Continuing to develop the foundations for high-quality empirical work, in chapter 2 we review the important elements of econometrics for investigations. Following these first two review chapters, chapters 3–10 develop the core of the material in the book. The concepts reviewed in these first two introductory chapters will be familiar to all competition economists, but it is worthwhile reviewing them since understanding these key elements of economic analysis is crucial for an appropriate use of quantitative techniques.

1.1 Demand Functions and Demand Elasticities

The analysis of demand is probably the single most important component of most empirical exercises in antitrust investigations. It is impossible to quantify the likelihood or the effect of a change in firm behavior if we do not have information about the potential response of its customers. Although every economist is familiar with the shape and meaning of the demand function, we will take the time to briefly review the derivation of the demand and its main properties since basic conceptual errors in its handling are not uncommon in practice. In subsequent chapters we will see that demand functions are critical for many results in empirical work undertaken in the competition arena.

1.1.1 Demand Functions

We begin this chapter by reviewing the basic characteristics of individual demand and the derivation of aggregate demand functions.

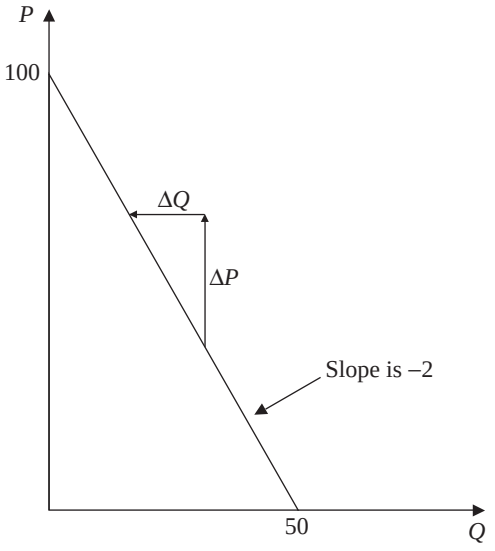

Figure 1.1. (Inverse) demand function.

1.1.1.1 The Anatomy of a Demand Function

An individual's demand function describes the amount of a good that a consumer would buy as a function of variables that are thought to affect this decision such as price P_i and often income y. Figure 1.1 presents an example of an individual linear demand function for a homogeneous product: $Q_i = 50 - 0.5P_i$ or rather for the inverse demand function, $P_i = 100 - 2Q_i$. More generally, we may write $Q_i = D(P_i, y)$.[1] Inverting the demand curve to express price as a function of quantity demanded and other variables yields the "inverse demand curve" $P_i = P(Q_i, y)$. Standard graphs of an individual's demand curve plot the quantity demanded of the good at each level of its own price and take as a given the level of income and the level of the prices of products that could be substitutes or complements. This means that along a given plotted demand curve, those variables are fixed. The slope of the demand curve therefore indicates at any particular point by how much a consumer would reduce (increase) the quantity purchased if the price increased (decreased) while income and any other demand drivers stayed fixed.

In the example in figure 1.1, an increase in price, ΔP, of €10 will decrease the demand for the product by 5 units shown as ΔQ. The consumer will not purchase any units if the price is above 100 because at that point the price is higher than the value that the customer assigns to the first unit of the good.

One interpretation of the inverse demand curve is that it shows the maximum price that a consumer is willing to pay if she wants to buy Q_i units of the good. While a

[1] This will be familiar from introductory microeconomics texts as the "Marshallian" demand curve (Marshall 1890).

consumer may value the first unit of the good highly, her valuation of, say, the one hundredth unit will typically be lower and it is this diminishing marginal valuation which ensures that demand curves typically slope downward. If our consumer buys a unit only if her marginal valuation is greater than the price she must pay, then the inverse demand curve describes our consumer's marginal valuation curve.

Given this interpretation, the inverse demand curve describes the difference between the customer's valuation of each unit and the actual price paid for each unit. We call the difference between what the consumer is willing to pay for each unit and what he or she actually pays the consumer's surplus available from that unit. For concreteness, I might be willing to pay a maximum of €10 for an umbrella if it's raining, but may nonetheless only have to pay €5 for it, leaving me with a measure of my benefit from buying the umbrella and avoiding getting wet, a surplus of €5. At any price P_i, we can add up the consumer surplus available on all of the units consumed (those with marginal valuations above P_i) and doing so provides an estimate of the total consumer surplus if the price is P_i.

In a market with homogeneous products, all products are identical and perfectly substitutable. In theory this results in all products having the same price, which is the only price that determines the demand. In a market with differentiated products, products are not perfectly substitutable and prices will vary across products sold in the market. In those markets, the demand for any given product is determined by its price and the prices of potential substitutes. In practice, markets which look homogeneous from a distance will in fact be differentiated to at least some degree when examined closely. Homogeneity may nonetheless be a reasonable modeling approximation in many such situations.

1.1.1.2 The Contribution of Consumer Theory: Deriving Demand

Demand functions are classically derived by using the behavioral assumption that consumers make choices in a way that can be modeled as though they have an objective, to maximize their utility, which they do subject to the constraint that they cannot spend more than they earn. As is well-known to all students of microeconomic theory, the existence of such a utility function describing underlying preferences may in turn be established under some nontrivial conditions (see, for example, Mas-Colell et al. 1995, chapter 1). Maximizing utility is equivalent to choosing the most preferred bundle of goods that a consumer can buy given her wealth.

More specifically, economists have modeled a customer of type (y_i, θ_i) as choosing to maximize her utility subject to the budget constraint that her total expenditure cannot be higher than her income:

$$V_i(p_1, p_2, \ldots, p_J, y_i; \theta_i) = \max_{q_1, q_2, \ldots, q_J} u_i(q_1, q_2, \ldots, q_J; \theta_i)$$
$$\text{subject to } p_1 q_1 + p_2 q_2 + \cdots + p_J q_J \leqslant y_i,$$

where p_j and q_j are prices and quantities of good j, $u_i(q_1, q_2, \ldots, q_J; \theta_i)$ is the utility of individual i associated with consuming this vector of quantities, y_i is the disposable income of individual i, and θ_i describes the individual's preference type. In many empirical models using this framework, the "i" subscripts on the V and u functions will be dropped so that all differences between consumers are captured by their type (y_i, θ_i).

Setting up this problem by using a Lagrangian provides the first-order conditions

$$\frac{\partial u_i(q_1, q_2, \ldots, q_J, y_i; \theta_i)}{\partial q_j} = \lambda p_j$$
$$\iff \quad \frac{\partial u_i(q_1, q_2, \ldots, q_J, y_i; \theta_i)/\partial q_j}{p_j} = \lambda \quad \text{for } j = 1, 2, \ldots, J,$$

together with the budget constraint which must also be satisfied. We have a total of $J + 1$ equations in $J + 1$ unknowns: the J quantities and the value of the Lagrange multiplier, λ.

At the optimum, the first-order conditions describe that the Lagrange multiplier is equal to the marginal utility of income. In some cases it will be appropriate to assume a constant marginal utility of income. If so, we assume behavior is described by a utility function with an additively separable good q_1, the price of which is normalized to 1, so that $u_i(q_1, q_2, \ldots, q_J; \theta_i) = \tilde{u}_i(q_2, \ldots, q_J; \theta_i) + q_1$ and $p_1 = 1$. This numeraire good q_1 is normally termed "money" and its inclusion provides an intuitive interpretation of the first-order conditions. In such circumstances a utility-maximizing consumer will choose a basket of products so that the marginal utility provided by the last euro spent on each product is the same and equal to the marginal utility of money, i.e., 1.[2]

More generally, the solution to the maximization problem describes the individual's demand for each good as a function of the prices of all the goods being sold and also the consumers' income. Indexing goods by j, we can write the individual's demands as

$$q_{ij} = d_{ij}(p_1, p_2, \ldots, p_J; y_i; \theta_i), \quad j = 1, 2, \ldots, J.$$

A demand function for product j incorporates not only the effect of the own price of j on the quantity demanded but also the effect of disposable income and the price of other products whose supply can affect the quantity of good j purchased. In figure 1.1, a change in the price of j represents a movement along the curve while a change in income or in the price of other related goods will result in a shift or rotation of the demand curve.

[2] This is called a quasi-linear demand function and gives the result because the first-order condition for good 1 collapses to

$$\lambda = \frac{\partial u_i(q_1, q_2, \ldots, q_J, y_i; \theta_i)/\partial q_1}{p_1} = \frac{\partial u_i(q_1, q_2, \ldots, q_J, y_i; \theta_i)}{\partial q_1},$$

which is the marginal utility of a monetary unit. That in turn is equal to one.

The utility generated by consumption is described by the (direct) utility function, u_i, which relates the level of utility to the goods purchased and is not observed. We know that not all levels of consumption are possible because of the budget constraint and that the consumer will choose the bundle of goods that maximizes her utility. The *indirect* utility function $V_i(p, y_i; \theta_i)$, where $p = (p_1, p_2, \ldots, p_J)$, describes the maximum utility a consumer can feasibly obtain at any level of the prices and income. It turns out that the direct and indirect utility functions each can be used to fully describe the other.

In particular, the following result will turn out to be important for writing down demand systems that we estimate.

For every indirect utility function $V_i(p, y_i; \theta_i)$ there is a direct utility function $u_i(q_1, q_2, \ldots, q_J; \theta_i)$ that represents the same preferences over goods provided the indirect utility function satisfies some properties, namely that $V_i(p, y_i; \theta_i)$ is continuous in prices and income, nonincreasing in price, nondecreasing in income, quasi-convex in (p, y_i) with any one element normalized to 1 and homogeneous degree zero in (p, y_i).

This result sounds like a purely theoretical one, but it will actually turn out to be very useful in practice. In particular, it will allow us to retrieve the demand function $q_i(p; y_i; \theta_i)$ without actually explicitly solving the utility-maximization problem.[3] Computationally, this is an important simplification.

1.1.1.3 Aggregation and Total Market Size

Individual consumers' demand can be aggregated to form the market aggregate demand by adding the individual quantities demanded by each customer at any given price. If $q_{ij} = d_{ij}(p_1, p_2, \ldots, p_J; y_i; \theta_i)$ describes the demand for product j by individual i, then aggregate (total) demand is simply the sum across individuals:

$$Q_j = \sum_{i=1}^{I} q_{ij} = \sum_{i=1}^{I} d_{ij}(p_1, p_2, \ldots, p_J, y_i; \theta_i), \quad j = 1, 2, \ldots, J,$$

where I is the total number of people who might want to buy the good. Many potential customers will set $q_{ij} = 0$ at least for some sets of prices $p_1, p_2, \ldots, p_J$ even though they will have positive purchases at lower prices of some products. In some cases, known as single "discrete choice" models, each individual will only buy at most one unit of the good and so $d_{ij}(p_1, p_2, \ldots, p_J, y_i; \theta_i)$ will be an indicator variable taking on the value either zero or one depending on whether individual i buys the good or not at those prices. In such models, the total number of people

[3] This result is known as a "duality" result and is often taught in university courses as a purely theoretical equivalence result. For its very practical implications, see chapter 9, where we describe the use of Roy's identity to generate empirical demand systems from indirect utility functions rather than the direct utility formulation.

who may want to buy the good is also the total potential market size. (We will discuss discrete choice models in more detail in chapter 9.) On the other hand, when individuals can buy more than one unit of the good, to establish the total potential market size we need to evaluate both the total potential number of consumers and also the total number of goods they might buy. Often the total potential number of consumers will be very large—perhaps many millions—and so in many econometric demand models we will approximate the summation with an integral.

In general, total demand for product j will depend on the full distribution of income and consumer tastes in the population. However, under very special assumptions, we will be able to write the aggregate market demand as a function of aggregate income and a limited set of taste parameters only:

$$Q_j = D_j(p_1, p_2, \ldots, p_J, Y; \theta),$$

where $Y = \sum_{i=1}^{I} y_i$.

For example, suppose for simplicity that $\theta_i = \mu$ for all individuals and every individual's demand function is "additively separable" in the income variable so that an individual's demand function can be written

$$d_{ij}(p_1, p_2, \ldots, p_J, y_i; \theta_i) = d_{ij}^*(p_1, p_2, \ldots, p_J; \mu) + \alpha_j y_i,$$

where α_j is a parameter common to all individuals, then aggregate demand for product j will clearly only depend on aggregate income. Such a demand function implies that, given the prices of goods, an increase in income will have an effect on demand that is exactly the same no matter what the level of the prices of all of the goods in the market. Vice versa, an increase in the prices will have the same effect whatever the level of income.[4]

The study of the conditions under which we can aggregate demand functions and express them as a function of characteristics of the income distribution such as the sum of individual incomes is called the study of aggregability.[5] Lessons from that literature motivate the use of particular functional forms for demand systems in empirical work such as the almost ideal demand system (AIDS[6]). In general, when building empirical models we may well want to allow market demand to depend on other statistics from the income distribution besides just the total income. For example, we might think demand for a product depends on total income in the population but also the variance, skewness, or kurtosis of the income distribution. Intuitively, this is fairly clear since if a population were made up of 1,000 people

[4] If consumer types are heterogeneous but are not observed by researchers, then an empirical aggregate demand model will typically assume a parametric distribution for consumer types in a population, $f_\theta(\theta; \mu)$. In that case, the aggregate demand model will depend on parameters μ of the distribution of consumer types. We will explore such models in chapter 9.

[5] For a technical discussion of the founding works, see the various papers by W. M. Gorman collected in Gorman (1995). More recent work includes Lewbel (1989).

[6] An unfortunate acronym, which has led some authors to describe the model as the nearly ideal demand system (NIDS).

making €1bn and everyone else making €10,000, then sales of €15,000 cars would be at most 1,000. On the other hand, the same total income divided more equally could certainly generate sales of more than 1,000. (For recent work, see, for example, Lewbel (2003) and references therein.)

1.1.2 Demand Elasticities

Elasticities in general, and demand elasticities in particular, turn out to be very important for lots of areas of competition policy. The reason is that the "price elasticity of demand" provides us with a unit-free measure of the consumer demand response to a price increase.[7] The way in which demand changes when prices go up will evidently be important for firms when setting prices to maximize profits and that fact makes demand elasticities an essential part of, for example, merger simulation models.

1.1.2.1 Definition

The most useful measurement of the consumer sensitivity to changes in prices is the "own-price" elasticity of demand. As the name suggests, the own-price elasticity of demand measures the sensitivity of demand to a change in the good's own-price and is defined as

$$\eta_{jj} = \frac{\%\Delta Q_j}{\%\Delta P_j} = \frac{100(\Delta Q_j / Q_j)}{100(\Delta P_j / P_j)}.$$

The demand elasticity expresses the percentage change in quantity that results from a 1% change in prices. Alfred Marshall introduced elasticities to economics and noted that one of their great properties is that they are unit free, unlike prices which are measured in currency (e.g., euros per unit) and quantities (sales volumes) which are measured in a unit of quantity per period, e.g., kilograms per year. In our example in figure 1.1 the demand elasticity for a price increase of 10 leading to a quantity decrease of 5 from the baseline position, where $P = 60$ and $Q = 20$, is $\eta_{jj} = (-5/20)/(10/60) = -1.5$.

For very small variations in prices, the demand elasticity can be expressed by using the slope of the demand curve times the ratio of prices to quantities. A mathematical result establishes that this can also be written as the derivative with respect to the logarithm of price of the log transformation of demand curve:

$$\eta_{jj} = \frac{P_j}{Q_j}\frac{\partial Q_j}{\partial P_j} = \frac{\partial \ln Q_j}{\partial \ln P_j}.$$

[7] The term "elasticity" is sometimes used as shorthand for "price elasticity of demand," which in turn is shorthand for "the elasticity of demand with respect to prices." We will sometimes resort to the same shorthand terminology since the full form is unwieldy. That said, we do so with the caveat that, since elasticities can be both "with respect to" and "of" anything, the terms elasticity or "demand elasticity" are inherently ambiguous and therefore somewhat dangerous. We will, for example, talk about the elasticity of costs with respect to output.

Demand at a particular price point is considered "elastic" when the elasticity is bigger than 1 in absolute value. An elastic demand implies that the change in quantity following a price increase will be larger in percentage terms so that revenues for a seller will fall all else equal. An inelastic demand at a particular price level refers to an elasticity of less than 1 in absolute value and means that a seller could raise revenues by increasing the price provided again that everything else remained the same. The elasticity will generally be dependent on the price level. For this reason, it does not usually makes sense to talk about a given product having an "elastic demand" or an "inelastic demand" but it should be said that it has an "elastic" or "inelastic" demand at a particular price or volume level, e.g., at current prices. The elasticities calculated for an aggregate demand are the market elasticities for a given product.

1.1.2.2 Substitutes and Complements

The *cross-price elasticity* of demand expresses the effect of a change in price of some other good k on the demand for good j. A new, higher, price for p_k may, for instance, induce some consumers to change their purchases of product j. If consumers increase their purchases of product j when p_k goes up, we will call products j and k demand *substitutes* or just substitutes for short.

Two DVD players of different brands are substitutes if the demand for one of them falls as the price of the other decreases because people switch across to the now relatively cheaper DVD player. Similarly, a decrease in prices of air travel may reduce the demand for train trips, holding the price of train trips constant.

On the other hand, the new higher price of k may induce consumers to buy less of good j. For example, if the price of ski passes increases, perhaps fewer folk want to go skiing and so the demand for skiing gear goes down. Similarly, if the price of cars increases, the demand for gasoline may well fall. When this happens we will call products j and k demand *complements* or just complements for short. In this case, the customer's valuation of good j increases when good k has been purchased:[8]

$$\eta_{jk} = \begin{cases} \dfrac{P_k}{Q_j}\dfrac{\partial Q_j}{\partial P_k} > 0 \quad \text{and} \quad \dfrac{\partial Q_j}{\partial P_k} > 0 & \text{if products are substitutes,} \\ \dfrac{P_k}{Q_j}\dfrac{\partial Q_j}{\partial P_k} < 0 \quad \text{and} \quad \dfrac{\partial Q_j}{\partial P_k} < 0 & \text{if products are complements.} \end{cases}$$

[8] Generally, this terminology is satisfactory for individual demand functions but can become unsatisfactory for aggregate demand functions, where it may or may not be the case that $\partial Q_j/\partial P_k = \partial Q_k/\partial P_j$ since in that case the complementary (or substitute) links between the products may be of differing strengths. See, in particular, the discussion in the U.K. Competition Commission's investigation into Payment Protection Insurance (PPI) at, for example, www.competition-commission.org.uk/inquiries/ref2007/ppi/index.htm. In that case, some evidence showed that loans and insurance covering unemployment, accident, and sickness were complementary only in the sense that the demand for insurance was affected by the credit price while the demand for credit appeared largely unaffected by the price of the accompanying PPI. That investigation (chaired by one of the authors) found it useful to introduce a distinction between one-sided and two-sided complementarity. An analogous distinction could be made for asymmetric demand substitution patterns.

1.1.2.3 Short Term versus Long Term

Most demand functions are static demand functions—they consider how consumers allocate their demand across products at a given point in time. In general, particularly in markets for durable goods, or goods which are storable, we will expect to have important intertemporal linkages in demand. The demand for cars today may depend on tomorrow's price as well as today's price. If so, demand elasticities in the long run may well be different from the demand elasticities in the short run. In some cases the price elasticity of demand will be higher in the short run. This happens for instance when there is a temporary decrease in prices such as a sale, when consumers will want to take advantage of the temporarily better prices to stock up, increasing the demand in the short run but decreasing it at a later stage (see, for example, Hendel and Nevo 2006a,b). In this case, the elasticity measured over a short period of time would overestimate the actual elasticity in the long run. The opposite can also occur, so that the long-run elasticity at a given price is higher than the short-run elasticity. For instance, the demand for petrol is fairly inelastic in the short run, since people have already invested in their cars and need to get to work. On the other hand, in the long run people can adjust to higher petrol prices by downsizing their car.

1.1.3 Introduction to Common Demand Specifications

We often want to estimate the effect of price on quantity demanded. To do so we will typically write down a model of demand whose parameters can be estimated. We can then use the estimated model to quantify the impact of a change in price on the quantity being demanded. With enough data and a general enough model our results will not be sensitive to this choice. However, with realistic sample sizes, we often have to estimate models that impose a considerable amount of structure on our data sets and so the results can be sensitive to the demand specification chosen. That unfortunate reality means one should choose demand specifications with particular care. In particular, we need to be clear about the properties of the estimated model that are being determined by the data and the properties that are simply assumed whatever the estimated parameter values. An important aspect of the demand function will be its curvature and how this changes as we move along the curve. The curvature of the demand curve will determine the elasticity and therefore the impact of a change in price on quantity demanded.

1.1.3.1 Linear Demand

The linear demand is the simplest demand specification. The linear demand function can be written $Q_i = a - bP$ with analogous inverse demand curve

$$P = \frac{a}{b} - \frac{1}{b}Q_i.$$

In each case, a and b are parameters of the model (see figure 1.2).

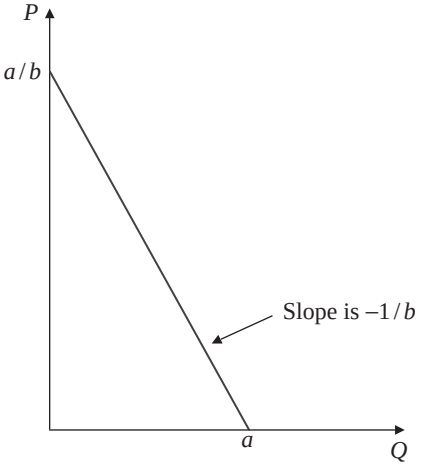

Figure 1.2. The linear demand function.

The slope of the inverse demand curve is

$$\frac{\partial P}{\partial Q} = \frac{-1}{b}.$$

The intercepts are a/b at $Q = 0$ and a at $P = 0$. The linear demand implies that the marginal valuation of the good keeps decreasing at a constant rate so that, even if the price is 0 the consumer will not "buy" more than a units. Since most analysis in competition cases happens at positive prices and quantities of the goods, estimation results will not generally be sensitive to assumptions made about the shape of the demand curve at the extreme ends of the demand function.[9] The elasticity for the linear demand function is

$$\eta = (-b)\frac{P}{Q}.$$

Note that, unlike the slope, the elasticity of demand varies along the linear demand curve. Elasticities generally increase in magnitude as we move to lower quantity levels because the variations in quantity resulting from a price increase are larger as a percentage of initial sales volumes. Because of its lack of curvature, the linear demand will sometimes produce higher elasticities compared with other demand specifications and therefore sometimes predicts lower price increases in response to mergers and higher quantity adjustments in response to increases in price. As an extreme example, consider an alternative inverse demand function which asymptotes as we move leftward in the graph toward the price axis where $Q = 0$. In that case, only very large price increases will drive significant quantity changes at low levels of

[9] We rarely get data from a market where goods have been sold at zero prices. As we discuss below, calculations such as consumer surplus on the other hand may sometimes be very sensitive to such assumptions.

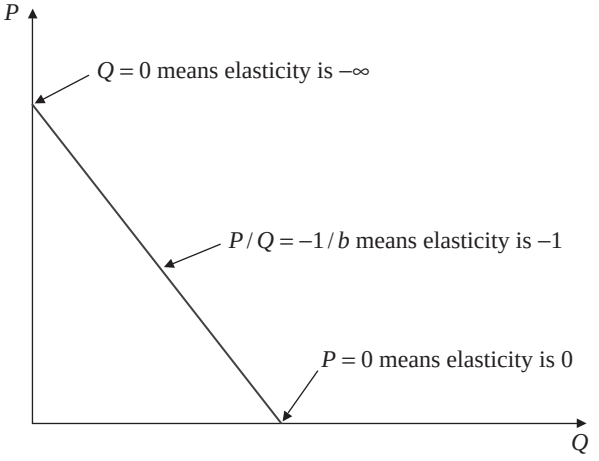

Figure 1.3. Demand elasticity values in the linear demand curve.

output or, analogously, small price changes will drive only small quantity changes, i.e., a low elasticity of demand. An example in the form of the log-linear demand curve is provided below. In contrast, the linear demand curve generates an arbitrarily large elasticity of demand (large in magnitude) as we move toward the price axis on the graph (see figure 1.3).

1.1.3.2 Log-Linear Demand

The one exception to the rule that elasticities depend on the price level is the log-linear demand function, which has the form

$$Q = D(P) = \mathrm{e}^{a} P^{-b}.$$

Taking natural logarithms turns the expression into a demand equation that is linear in its parameters:

$$\ln Q = a - b \ln P.$$

This specification is particularly useful because many of the estimation techniques used in practice are most easily applied to models which are linear in their parameters. Expressing effects in terms of percentages also provides us with results that are easily interpreted. The inverse demand which corresponds to figure 1.4 can be written

$$P = P(Q) = (\mathrm{e}^{-a} Q)^{-1/b}.$$

When prices increase toward infinity, if $b > 0$ then the quantity demanded tends toward 0 but never reaches it. An assumption embodied in the log-linear model is that there will always be some demand for the good, no matter how expensive it is. Similarly, the demand tends to infinity when the price of the good approaches 0.

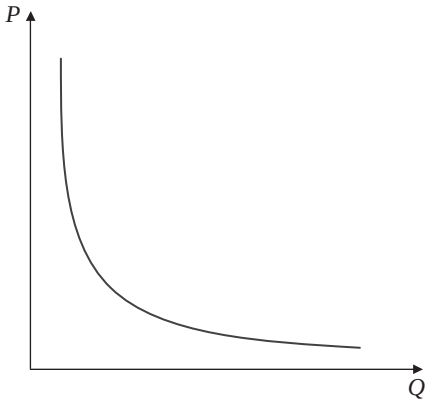

Figure 1.4. The log-linear demand curve.

As a product approaches the zero price, consumers are willing to have an unlimited amount of it:

$$\lim_{P\to\infty} D(P) = \mathrm{e}^{a} \lim_{P\to\infty} P^{-b} = 0,$$
$$\lim_{Q\to\infty} P(Q) = \lim_{Q\to\infty} (\mathrm{e}^{-a} Q)^{-1/b} = 0.$$

The log-linear demand also has a constant elasticity over the entire demand curve, which is a unique characteristic of this functional form:

$$\eta = \frac{\partial \ln Q}{\partial \ln P} = -b.$$

As a result the log-linear demand model is sometimes referred to as the constant elasticity or iso-elastic demand model. Price changes do not affect the demand elasticity, which means that if we have one estimate of the elasticity, at a given price, this estimate will—rather conveniently but perhaps optimistically—be the same for all price points. Of course, if in truth the price sensitivity of demand does depend on the price level, then this iso-elasticity assumption will be a strong one imposed by the model whatever values we estimate its parameters a and b to take on. Empirically, given enough data, we can tell apart data generated by the linear demand model and the log-linear model since movements in supply at different price levels will provide us with information about the slope of demand and hence elasticities. Formally, we can use a "Box–Cox" test to distinguish the models (see, for example, Box and Cox 1964).

1.1.3.3 Discrete Choice Demand Models

Consumer choice situations can be sometimes best represented as zero–one "discrete" decisions between different alternative options. Consider, for example, buying a car. The choice is "which car" rather than "how-much car." In such situations,

a discrete choice demand model is typically used to capture consumer behavior. These models allow utility maximization to take place over existing options. One of the most popular discrete choice demand models is the multinomial logit (MNL) demand model, sometimes called "logit" for brevity (see McFadden 1973).

The MNL demand model assumes that the utility provided to a consumer who chooses to buy product j takes the form[10]

$$U_{ij} = \alpha x_j + \beta p_j + \varepsilon_{ij},$$

where $j = 1, \ldots, J$ indicates the product and i indicates a particular individual. The utility provided is determined by the good's characteristics x_j, the price p_j, and by an element of utility ε_{ij} which indicates the particular taste of individual i for good j. Product attributes provide utility to the consumer while higher prices reduce utility so β will typically be negative. As before, each individual is assumed to pick the option which provides her with the most utility, $\max_{j=1,\ldots,J} U_{ij}$. As before, aggregate demand in such situations is the sum of all individual demands. The MNL model simply makes a particularly convenient set of assumptions about the form of "consumer heterogeneity," i.e., the way in which one consumer is different from others in the population. In the MNL model, consumers are assumed to be identical except for the random additively separable terms ε_{ij}. A more detailed discussion of the logit model and other discrete choice models of demand is presented in chapter 9.

For now we note that we will see that in some cases estimation of MNL amounts to running a linear regression. Elasticities on the other hand generally need to be calculated as a second step once the parameters have been estimated. Discrete choice demand models are typically nonlinear and although some of them are mathematically intractable others are highly tractable.

1.1.4 Consumer Welfare

Many competition authorities around the world, at least in principle, use a "consumer welfare" standard to evaluate policy and firm behavior. Such a standard is not uncontroversial since some economists argue that there should be equal (or at least some) weight assigned to producer and consumer welfare with redistributions if desired achieved by other means such as taxation.[11] Whichever welfare standard

[10] More precisely, these are called "conditional indirect utilities." The reason is that it is the indirect utility obtained if product j is chosen, i.e., conditional on choosing product j. We will see in chapter 9 that these choice models can be motivated by using our familiar (utility maximization subject to a budget constraint) model by imposing constraints on the consumer's choice set. The "indirect" comes from the fact that the utility is specified as a function of price.

[11] We do not discuss the relative merits of arguments in this debate here, though it is certainly an important and interesting one. The proponents of consumer surplus standards usually cite a political economy reason: that consumers are large in number and have only very diffuse incentives to intervene individually in making markets work for them while large firms have far less diffuse incentives to extract surplus. The economics of Harbinger triangles suggests that pure static deadweight losses are sometimes "small." Putting deadweight losses to one side, standard monopoly pricing results in a transfer of surplus

is used, we must say what we mean by "consumer welfare" and generally, in practice, competition authorities often mean an approximation to consumer welfare, "aggregate consumer surplus," a term which we define below.[12] Generally, actions that permanently result in an increase of market output, a decrease in prices, or an increase in the customers' valuation of the product will increase "consumer surplus" and so are deemed beneficial for consumers. If firms provide tax revenue that is subsequently redistributed in part, or individuals invest in companies either directly or via pension funds, then the distinction between individual (rather than consumer) welfare and producer welfare is less clear cut than the consumer–producer distinction. Democratic governments that enact competition laws presumably ultimately care (at least) about all their citizens, which some argue means there should be at least some weight for shareholders via a weight on producer surplus. Such weight would probably lead to a less interventionist approach than a "pure" consumer welfare standard. Even within a consumer welfare standard, there are significant choices to be made. For example, to operationalize a "true" consumer welfare measure, an agency would need to decide how careful to be when weighting individuals' utilities by their respective marginal utilities of income. Doing so, or not, could lead to profoundly different practical outcomes in a competition agency. In particular, weighing consumers according to their marginal utilities of income may lead an agency to be involved in more intervention to protect poorer consumers, even potentially at the cost of richer consumers. Some in the competition policy world consider such income redistributions to be more in the realm of social policy than competition policy. Others disagree that an easy distinction is possible. For a concrete example where such issues might arise consider price discrimination for a good where inelastic demanders tend to be poorer. If so, price discrimination could involve poor customers paying high prices while rich consumers pay lower prices. A recent example, is electricity in the United Kingdom, where many poorer customers are, to an extent, "locked in" to prepay meters and hence are charged more per unit than their richer neighbors who pay monthly and can change provider. A competition agency acting to stop price discrimination would typically result in richer customers paying more and poorer customers paying less. Absent clear governmental instructions on the framework for analysis, an important question is whether a competition agency is in a suitable position to make such (distributional and hence political) judgments.

---

from consumers to producers. In addition, the evidence suggests that there are potentially important dynamic effects of competition on productivity, including cost reductions and also welfare gains resulting from increased variety and improved quality. Quantifying such effects is tremendously difficult but also potentially tremendously important. Efforts to do so include Nickell (1996) and more recently Aghion and Griffith (2008). The link between competition and productivity is important in competition policy but also in international trade and so much of the available evidence comes from that field. See, for example, the contributions and literature surveyed by Jensen et al. (2007).

[12] Many current authors attribute "consumer surplus" to Marshall (1890). However, Hotelling (1938) attributes "consumer surplus" to an engineer, Jules Dupuit, in his work of 1844. See the discussion in Hotelling (1938).

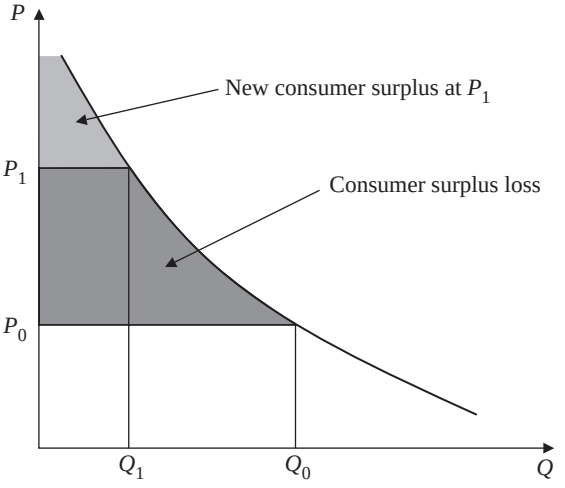

Figure 1.5. Reduction in consumer surplus following a price increase.

We note that some regulators do have legal obligations to protect consumers generally but also vulnerable consumer groups specifically (e.g., the water regulator in the United Kingdom, Ofwat).

1.1.4.1 Consumer Surplus

The consumer surplus derived from a unit of consumption is the difference between the price that a consumer would be willing to pay for it and what she actually pays, i.e., the market price. Since the demand curve describes the maximum that a consumer would have been willing to pay for each unit, the consumer surplus is simply the difference between the demand curve and the price actually paid. Every unit being consumed generates consumer surplus and so the total consumer surplus is the area below the demand curve that falls above the price paid for the good. Figure 1.5 represents the loss of consumer surplus after prices increase from P_0 to P_1, reducing demand.

1.1.4.2 Quantification of Consumer Surplus

If $P(Q)$ denotes the inverse demand curve, calculation of consumer surplus at price P_0 and quantity Q_0 involves the following calculation:

$$\mathrm{CS}_0 = \int_0^{Q_0} (P(Q) - P_0)\,\mathrm{d}Q = \int_0^{Q_0} P(Q)\,\mathrm{d}Q - P_0 Q_0.$$

Welfare measurements can be very sensitive to the demand specification chosen, so in practical circumstances one will sometimes need to examine several plausible specifications and describe the range of potential outcomes given assumptions about demand. In particular, the behavior of the inverse demand curve $P(Q)$ close to

$Q = 0$ can have a large impact on the value of consumer surplus obtained and it is something about which we will rarely have any data. Welfare estimates of changes within the realm of experience will tend to rely less heavily on our underlying assumptions about the demand curve (e.g., whether it is linear or log-linear). The welfare effect of a change in the market price from P_0 to P_1 is calculated by

$$\begin{aligned}\Delta \mathrm{CS} &= \mathrm{CS}_1 - \mathrm{CS}_0 \\ &= \int_0^{Q_1} P(Q)\,\mathrm{d}Q - P_1 Q_1 - \int_0^{Q_0} P(Q)\,\mathrm{d}Q + P_0 Q_0,\end{aligned}$$

where the subscripts "0" and "1" indicate the situation before and after the change. For some policy evaluations, the demand function in the two integrals will be different. For example, in chapter 10 we will examine the impact of a change in vertical ownership arrangements in the cable TV market on consumer welfare. Theory suggests that both the price and quality of the good provided may change as a result of the change in market structure.

One approach to estimating consumer surplus is to estimate the demand curve. However, there are also alternatives when evaluating welfare outcomes. For example, a simple technique for approximating deadweight loss in practice involves the method originally used by Harberger (1954) in his classic cross-industry study of the magnitude of deadweight loss. Deadweight loss is the surplus that is lost to consumers and not transferred to producers when prices increase, and is sometimes known as the Harberger triangle. Since consumers lose the surplus and producers do not gain it, it represents a fall in total welfare. In that study Harberger observed

(i) a measure of "excess" profits allowing for a 10.4% "normal" rate of return on capital in the calculation of total costs, $C(Q)$, $\Pi = P(Q)Q - C(Q)$, and

(ii) a measure of sales $R = P(Q)Q$ for each industry.

Our data tell us that

$$\frac{\Pi}{R} = \frac{P(Q)Q - C(Q)}{P(Q)Q} = \frac{P(Q) - AC(Q)}{P(Q)}$$

so that the "return on sales" ratio gives us the percentage monopolistic price markup (the Lerner index) under either the assumption that all industries neither benefit from economies of scale nor suffer from diseconomies of scale so that average and marginal costs were equal or alternatively if we measure monopoly distortions only relative to a "second best" welfare outcome where firms must price to make nonnegative returns because lump-sum transfers are not possible.

The elasticity of demand then tells us how much sales will fall following such a percentage price increase. The deadweight loss (Harberger triangle) is then estimated as one half of the price change times the predicted quantity change, each in levels rather than percentages, i.e.,

$$\text{Deadweight Loss} = \frac{(P - AC)\Delta Q}{2} = \frac{\Pi^2(-\eta)}{2(PQ)},$$

where the former is just the definition of deadweight loss from monopolistic pricing under our assumptions while the latter involves only our "data." The equality can be seen by expanding and canceling terms since[13]

$$\frac{(P - AC)\Delta Q}{2} = \frac{((P - AC)Q)^2}{2(PQ)} \frac{(-\Delta Q/Q)}{(P - AC)/P} = \frac{\Pi^2 \eta}{2(PQ)}.$$

Harberger assumed that all industry price elasticities of demand η were -1. One can also evaluate the transfer involved from consumers to firms as simply the "excess" profits being earned. Thus, for example, Harberger had an estimate of the excess profits (averaged over the period 1924–28) for the bakery products industry of \$17 million and an estimate of excess profits/sales, and therefore markups above average costs, of $100\Pi/R = 5.3\%$. Reverse engineering Harberger's calculation we learn that revenues were $R = \Pi/0.053 = \$320.8$ million and we can then calculate that

$$\text{Deadweight Loss} = \frac{\Pi^2(-\eta)}{2(PQ)} = \frac{17^2(1)}{2(320.8)} = 0.45 \text{ million},$$

about half a million dollars on sales of \$321 million. The transfer from consumers to firms of course involves all the \$17 million in "excess" profits, so that the order of magnitude of consumer surplus loss is greater than that of the deadweight loss. Notice finally that the more elastic demand is, for a given level of excess profits, the greater the expected deadweight loss.[14]

Such an exercise is not easy in a cross-industry study and, for example, it is striking that many of Harberger's estimates of excess profits (and prices) were in fact negative, suggesting that prices in many industries were "too low" rather than "too high." He derives the "normal" profit rate by allowing for a 10.4% return on capital employed, which he calculates by using the simple average of profit rates across industries in his study. In a modern application we would usually want to use a more sophisticated approach to the "cost of capital" which adjusts for risk across the various industries. (See, for example, the discussion on the weighted average cost of capital (WACC) in chapter 3.)

Consumer welfare calculations can be a useful tool for a rough approximation of an effect but given the crucial importance of assumptions, for which there is sometimes little factual evidence, the impact on consumer welfare is currently sometimes not actively quantified during investigations but rather qualitatively assessed in view of the conduct's expected impact on prices, output, and other variables relevant for consumer valuation.

---

[13] In this formula, η is measured as the percentage change in quantities that results from the percentage change in prices above cost, i.e., $(P - AC)/P$.

[14] In the U.K. Competition Commission's investigation into payment protection insurance, excess profits from PPI were estimated to be £1.4 billion on sales of £3.5 billion. If the price elasticity of demand were -1.5, then such a calculation suggests a deadweight loss of $(1{,}400)^2(1.5)/(2 \times 3{,}500) =$ £420 million. Harberger triangles need not always be small.

There are a number of related notions of consumer welfare in addition to consumer surplus and in fact consumer surplus is best considered an imperfect approximation for an "exact" welfare measure for a given individual. We may alternatively use equivalent variation (EV) or compensating variation (CV) to measure welfare "exactly" in a continuous choice demand context, while researchers also use expected maximum utility (EMU) in the discrete choice demand context. Compensating variation calculates the change in income that must be given to or taken from a consumer *after* a price change in order to bring her back to her previous utility level. The equivalent variation is the change in income (positive or negative) that one should give to or take from our consumer *before* a price change to give her the same utility level before and after a price change.[15] Marshall showed that consumer surplus will equal compensating variation if a consumer has a constant marginal utility of income.

In some cases, these objects are easy to calculate directly, for example, among other results Hausman (1981) provides analytic expressions for CV that arise from a single inside good (and one outside good) linear demand curve of the form we graphed in figure 1.3 (Hausman 1981; Hurwicz and Uzawa 1971).[16] This debate around approximating consumer welfare measures for a given individual is at one level only for the perfectionist; the consensus from the literature appears to be that measures of consumer surplus changes from price rises do not typically appear particularly sensitive to the approximation which motivates the use of consumer surplus. On the other hand, the approximation error can be a significant amount relative to a deadweight loss calculation. Of course, in interpreting such results it is important to keep in mind that authors will often assume that market demand curves can be rationalized as if they were a representative consumer's demand curve. As we have described, representative agent demand models require strong and probably unrealistic assumptions. In a more general model where aggregate demand depends on the distribution of income (and perhaps also on other elements of consumer heterogeneity), CV and EV measures can be calculated for each individual and then aggregated across individuals. One interpretation of this "result" is that authors must be very careful with deadweight loss calculations. Another far more controversial interpretation is that the classical deadweight losses are only of a similar order

[15] To illustrate the difference for the classic continuous choice demand case, readers may recall the difference between Marshall's demand curve, which is a function of price for a given level of income $d(p, y)$ and the Hicksian demand, which is described as a function of price for a given level of utility, $d(p, u)$ (Hicks 1956). See the discussion in, for example, Deaton and Muellbauer (1980b, chapter 7). For more on practical methods to compute exact welfare measures, see Vartia (1983). We follow practice rather than theory in this section, but also point the reader to Breslaw and Smith (1995), who very usefully provide computer code for approximating CV using Aptech's GAUSS matrix programming language using a method which avoids solving differential equations (à la the method suggested in Hausman (1981)).

[16] In looking at Hausman (1981) it is important to recall that a numerical error means he was far more negative about consumer surplus as an approximation than the actual results suggested (see Irvine and Sims 1998). See also Hausman and Newey (1995).

to our approximation error for welfare calculations and perhaps are therefore best considered typically small.[17]

This brings to a conclusion our brief review of the concepts from demand theory that are used daily in competition policy analysis. We will discuss each of these concepts in greater depth in future chapters, but next we turn to costs and production.

1.2 Technological Determinants of Market Structure

Firm decisions are an important driver of market structure, performance, and conduct and so, if we are to understand market outcomes, we must first understand firm decisions. In turn, if profits are an important driver of firm decisions, then we must understand the drivers of profits, namely revenues and costs. Demand analysis provides a toolbox for analyzing firm revenues. We now turn to the economists' toolbox for analyzing information on the cost side of the market.

Economists examining firms' cost structure, efficiency, and productivity have found three interrelated types of models particularly useful: production functions, cost functions, and input demand equations. We describe each below. We will see that each contains information about both technological possibilities for combining inputs into outputs and also about the cost of doing so. Along the way these tools facilitate an analysis of firm efficiency and productivity.

1.2.1 Production Functions

To produce output the firm must combine inputs according to a technological and/or managerial process. A production function describes the output that can be achieved by efficiently combining inputs.[18] It reflects technological reality and is expressed as $Q = f(K_1, \ldots, K_n; \alpha_1, \ldots, \alpha_m, u)$, where K_i are inputs, α_j are technological parameters, and u is a firm-specific (or plant- or occasionally process-specific) productivity indicator. The causes of the unknown (to the researcher) productivity indicator u are often of great interest as well as the differences in productivity across firms or plants. Whatever the causes, a firm with a higher u can for some reason combine inputs to produce more output than a firm of lower productivity. Reasons

[17] If competition authorities operated a total welfare standard, one might conclude that these short-run effects are small and antitrust intervention should therefore be highly limited. On the other hand, even if this were true, if competition authority interventions affect the incentives to reduce costs or to compete by introducing new or better products, then the relevant consumer (and total) surplus gains can be extremely large in the longer term. Moreover, there are examples where even the short-run measures of deadweight losses will be large.

[18] Recall that production possibility sets capture the ways in which inputs can feasibly be turned into outputs. In contrast, production frontiers capture the ways in which inputs can efficiently be turned into output, that is, the smallest levels of inputs required to produce a given level of output. Under technical assumptions, production functions capture the information in the production possibility frontier, that is, they describe the efficient ways of combining inputs to produce output.

might include the firms' respective levels of know-how and the managerial quality of their production processes.

When choosing a specification for a production function, it is important to be aware of the implications of a given functional form in terms of assumptions being made about the actual production process. Some functional forms are more flexible than others in that different values for the parameters can accommodate many different technological realities. Other functional forms, on the other hand, describe very specific production processes. Obviously, we are attempting to capture reality so our production function specification should be capable of doing so. To illustrate, in this section we first introduce some terminology and then we present two classic examples: the fixed-proportions technology and the Cobb–Douglas production function.

1.2.1.1 Terminology

Isoquants. The extent to which technology allows different inputs to substitute for one another is important for both the mix of inputs a firm will choose and also the amount of output a firm can produce. We call a contour describing the combinations of inputs that produce any given level of output an isoquant, where "iso" means "same" (so isoquant means same quantity). An example of an isoquant is provided in figure 1.6.

Marginal Product. The marginal product of an input is the increase in output due to an increase in that input alone. For example, the marginal product of input K_i is defined as $\text{MP}_{K_i} = \partial Q/\partial K_i$.

Marginal Rate of Technical Substitution. The slope of an isoquant tells us how much we need to increase one input to compensate for the decrease in another input if we want to maintain the same output level. This is called the marginal rate of technical substitution (MRTS):

$$\text{MRTS}_{jk} = \frac{\partial Q/\partial K_j}{\partial Q/\partial K_k}.$$

Returns to Scale. We sometimes consider what happens to the amount of output produced, $f(\lambda K_1, \ldots, \lambda K_n; \alpha_1, \ldots, \alpha_m, u)$, when all inputs are increased by a factor λ. For example, we might perhaps consider $\lambda = 2$ in which case we are considering what happens to output if we double all inputs. If output also increases by λ, then we say that the production function exhibits constant returns to scale (CRS). If output increases by more than λ, we say there are increasing returns to scale (IRS), and if output increases by less than λ, we say there are decreasing returns to scale (DRS).

There are increasing returns to scale in the transportation of oil and that is why supertankers exist. To see why, consider that an approximate formula for the volume

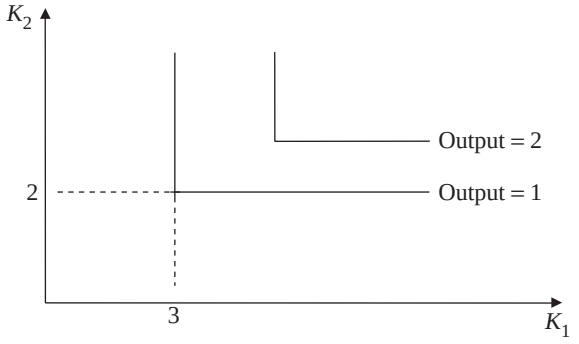

Figure 1.6. Isoquants for the fixed-proportions technology.

of oil that an oil tanker can carry is length × height × width. That means to double the volume of oil carried we need to double either the length, the height, or the width but definitely not all three. That in turn means we do not need to double the amount of steel used to build the oil tanker if we want to double the amount of oil that can be carried from one place to another. Similarly, we may not need to double the size of the crew.

Industries which will tend to exhibit CRS include those where we can build identical plants next to each other.

On the other hand, if it takes more and more inputs to produce a single extra unit of output, then we say there are DRS. Continuing our previous example, even if in principle there are CRS available from building identical plants next to one another, if management and coordination of all those plants become ever more complex as the firm grows, we may nonetheless suffer from DRS at the firm level.

Formally, assume a production function $Q = f(K_1, K_2; u)$.

$$\text{If } f(\lambda K_1, \lambda K_2; u) = \lambda f(K_1, K_2; u), \text{ there are CRS.}$$
$$\text{If } f(\lambda K_1, \lambda K_2; u) > \lambda f(K_1, K_2; u), \text{ there are IRS.}$$
$$\text{If } f(\lambda K_1, \lambda K_2; u) < \lambda f(K_1, K_2; u), \text{ there are DRS.}$$

The nature of returns to scale can differ at different levels of production. Indeed, one reason economies of scale can be important in competition policy is that returns to scale determine the minimum efficient scale of operation and so may help evaluate an "efficiency" defense in a merger. Alternatively, a monopoly may argue that it is a natural monopoly and therefore should not be broken up during a competition investigation.

1.2.1.2 Fixed-Proportions Technology

The fixed-proportions production technology provides an important if somewhat extreme example. It implies that to produce output we need to use inputs in fixed

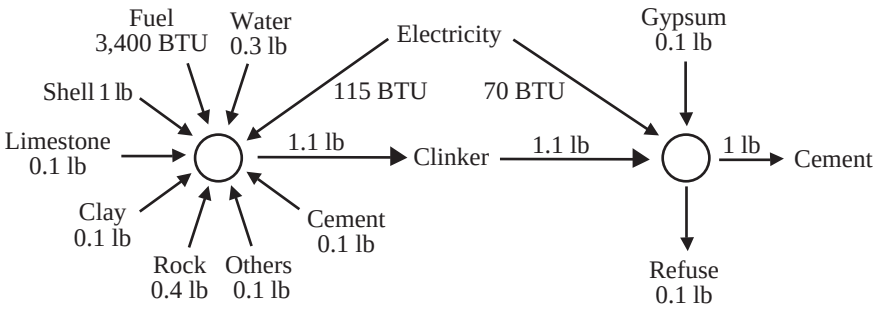

Figure 1.7. Zero substitution between inputs: an approximate recipe for portland cement. (Cement kilns are easy to spot—they are usually long cylindrical tubes which can be 750 feet long.) *Source*: Derived from a graph provided by Tom Stoker, MIT. Numbers amended to protect confidentiality.

proportions, that is, there is no way to substitute among the inputs to produce output. Suppose, for instance, a unit of output Q can only be produced with three units of K_1 and two units of K_2, where K_1 and K_2 are inputs. The production function is expressed by

$$Q = \min\{\tfrac{1}{3}K_1; \tfrac{1}{2}K_2\}.$$

The isoquants are shown in figure 1.6.

We see that in this example unless we have additional K_1 available we cannot increase production no matter how much available K_2 there is as there is no substitutability between the inputs. Such a production function could be that of the perfect martini, where gin and vermouth are combined in fixed proportions: with each martini requiring 75 ml of gin and 5 ml of vermouth.[19]

Another example of such a production function is provided by the recipe for portland cement (see figure 1.7). In this case, the mapping of isoquants is not possible on a two-dimensional scale but the characteristics of the production function are similar. Whenever a production process involves following a fixed "recipe" one must increase all inputs by a given factor to increase output.

Note that in this example of zero substitution among inputs, the marginal product of an extra unit of input holding fixed the amount of all other inputs is zero. When working with a fixed-coefficients production technology, to produce some more output we need more of each of the inputs.

1.2.1.3 The Cobb–Douglas Production Function

The Cobb–Douglas production function is frequently used for its flexibility and convenient properties. This function is named after C. W. Cobb and P. H. Douglas, who introduced it in 1928 in a study on the evolution of output, capital, and labor

[19] Winston Churchill is reputed to have had a slightly different fixed-proportions production function for the perfect dry martini, one which involved only a "glance" at the vermouth.

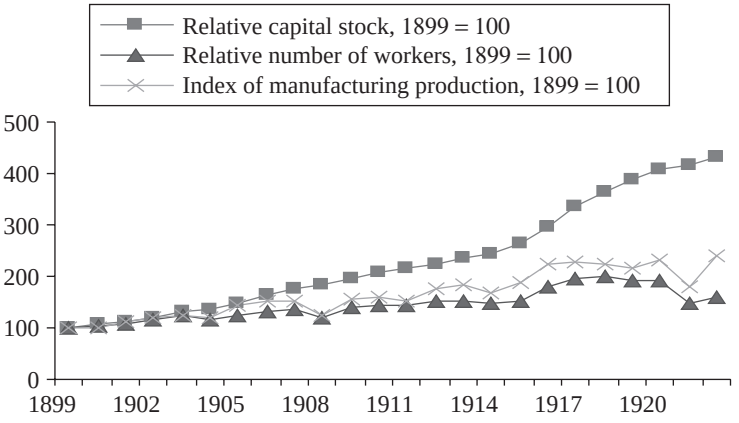

Figure 1.8. A plot of Cobb and Douglas's data.

in the United States between 1899 and 1924. Their time series evidence examines the relationship between aggregate inputs of labor and capital and national output during a period of fast growing U.S. labor and even faster growing capital stock. Their data are plotted in figure 1.8.[20]

Cobb and Douglas designed a function that could capture the relationship between output and inputs while allowing for substitution and which could be both empirically relevant and mathematically tractable. The Cobb–Douglas production function is defined as follows:

$$Q = a_0 L^{a_L} K^{a_K} u \quad \Longrightarrow \quad \ln Q = \beta_0 + a_L \ln L + a_K \ln K + v,$$

where $v = \ln u$, $\beta_0 = \ln a_0$, and where the parameters (a_0, a_L, a_K) can be easily estimated from the equation once it is log-linearized. As figure 1.9 shows, the isoquants in this function exhibit a convex shape indicating that there is a certain degree of substitution among the inputs.

Marginal products, the increase in production achieved by increasing one unit of an input holding other inputs constant, are defined as follows in a Cobb–Douglas function:

$$\mathrm{MP}_L \equiv \frac{\partial Q}{\partial L} = a_0 a_L L^{a_l - 1} K^{a_K} F^{a_F} u = a_L \frac{Q}{L},$$

$$\mathrm{MP}_K \equiv \frac{\partial Q}{\partial K} = a_0 L^{a_l} a_K K^{a_K - 1} F^{a_F} u = a_K \frac{Q}{K},$$

so that the marginal rate of technical substitution is

$$\mathrm{MRTS}_{LK} = \frac{\partial Q / \partial L}{\partial Q / \partial K} = \frac{a_L}{a_K} \frac{K}{L}.$$

[20] In their paper (Cobb and Douglas 1928), the authors report the full data set they used.

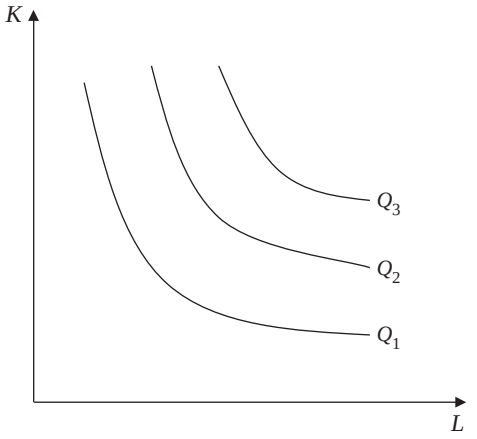

Figure 1.9. Example of isoquants for a Cobb–Douglas function.

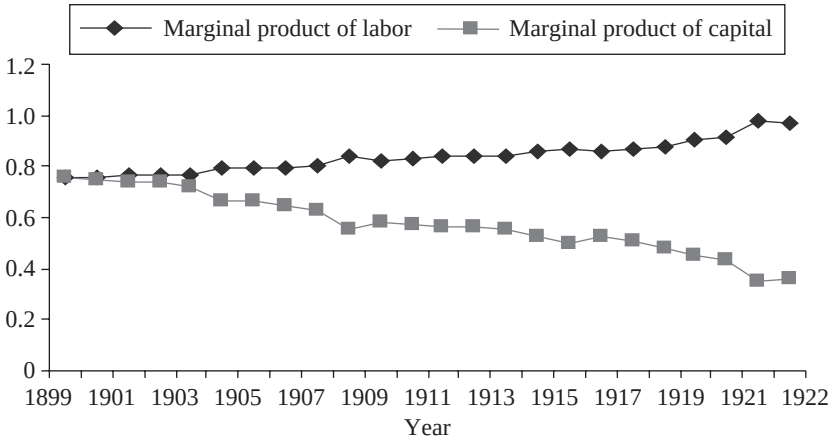

Figure 1.10. Cobb and Douglas's implied marginal products of labor and capital.

Cobb and Douglas's econometric evidence suggested that the increase in labor and particularly capital over time was increasing output, but not proportionately. In particular, as figure 1.10 shows their estimates suggested that the marginal product of capital was declining fast. Naturally, such a conclusion in 1928 would have profound implications for the likelihood of continued large capital flows into the United States.

1.2.2 Cost Functions

A production function describes how much output a firm gets if it uses given levels of inputs. We are directly interested in the cost of producing output, not least to decide how much to produce and as a result it is quite common to estimate cost functions.

Rather surprisingly, under sometimes plausible assumptions, cost functions contain exactly the same information as the production function about the technical possibilities for turning inputs into outputs but require substantially different data sets to estimate. Specifically, assuming that firms minimize costs allows us to exploit the "duality" between production and cost functions to retrieve basically the same information about the nature of technology in an industry.[21]

1.2.2.1 Cost Minimization and the Derivation of Cost Functions

In order to maximize profits, firms are commonly assumed to minimize costs for any given level of output given the constraint imposed by the production function with regards to the relation between inputs and output. Although the production function aims to capture the technological reality of an industry, profit-maximizing and cost-minimizing behaviors are explicit behavioral assumptions about the ways in which firms are going to take decisions. As such those behavioral assumptions must be examined in light of a firm's actual behavior.

Formally, cost minimization is expressed as

$$C(Q, p^L, p^K, p^F, u; \alpha) = \min_{L,K,F} p^L L + p^K K + p^F F \quad \text{subject to} \quad Q \leqslant f(L, K, F, u; a),$$

where p indicates prices of inputs L, K, and F, u is an unobserved cost efficiency parameter, and α and a are cost and technology parameters respectively. Given input prices and a production function, the model assumes that a firm chooses the quantities of inputs that minimize its total cost to produce each given level of output. Thus, the cost function presents the schedule of quantity levels and the minimum cost necessary to produce them.

An amazing result from microeconomic theory is that, if firms do indeed (i) minimize costs for any given level of output and (ii) take input prices as fixed so that these prices do not vary with the amount of output the firm produces, then the cost function can tell us everything we need to know about the nature of technology. As a result, instead of estimating a production function directly, we can entirely equivalently estimate a cost function. The reason this theoretical result is extremely useful is that it means one can retrieve all the useful information about the parameters of technology from available data on costs, output, and input prices. In contrast, if we were to learn about the production function directly, we would need data on output and input quantities.

This equivalency is sometimes described by saying that the cost function is the dual of the production function, in the sense that there is a one-to-one correspondence

[21] This result is known as a "duality" result and is often taught in university courses as a purely theoretical equivalence result. However, we will see that this duality result has potentially important practical implications precisely because it allows us to use very different data sets to get at the same underlying information.

between the two if we assume cost minimization. If we know the parameters of the production function, i.e., the input and output correspondence as well as input prices, we can retrieve the cost function expressing cost as a function of output and input prices.

For example, the cost function that corresponds to the Cobb–Douglas production function is (see, for example, Nerlove 1963)

$$C = kQ^{1/r} p_L^{\alpha_L/r} p_K^{\alpha_K/r} p_F^{\alpha_F/r} v,$$

where $v = u^{-1/r}$, $r = \alpha_L + \alpha_K + \alpha_F$, and $k = r(\alpha_0 \alpha_L^{\alpha_L} \alpha_K^{\alpha_K} \alpha_F^{\alpha_F})^{-1/r}$.

1.2.2.2 Cost Measurements

There are several important cost concepts derived from the cost function that are of practical use.

The *marginal cost* (MC) is the incremental cost of producing one additional unit of output. For instance, the marginal cost of producing a compact disc is the cost of the physical disc, the cost of recording the content on that disc, the cost of the extra payment on royalties for the copyrighted material recorded on the disc, and some element perhaps of the cost of promotion. Marginal costs are important because they play a key role in the firm's decision to produce an extra unit of output. A profit-maximizing firm will increase production by one unit whenever the MC of producing it is less than the marginal revenue (MR) obtained by selling it. The familiar equality MC = MR determines the optimal output of a profit-maximizing firm because firms expand output whenever MC < MR thereby increasing their total profits.

A *variable cost* (VC) is a cost that varies with the level of output Q, but we shall also use the term "variable cost" to mean the sum of all costs that vary with the level of output. Examples of variable costs are the cost of petrol in a transportation company, the cost of flour in a bakery, or the cost of labor in a construction company. *Average variable cost* (AVC) is defined as AVC = VC/Q. As long as MC < AVC, average variable costs are decreasing with output. Average variable costs are at a minimum at the level of output at which marginal cost intersects average variable cost from below. When MC > AVC, the average variable costs is increasing in output.

Fixed costs (FC) are the sum of the costs that need to be incurred irrespective of the level of output produced. For example, the cost of electricity masts in an electrical distribution company or the cost of a computer server in a consulting firm may be fixed—incurred even if (respectively) no electricity is actually distributed or no consulting work actually undertaken. Fixed costs are recoverable once the firm shuts down usually through the sale of the asset. In the long run, fixed costs are frequently variable costs since the firm can choose to change the amount it spends. That can make a decision about the relevant time-horizon in an investigation an important one.

Sunk costs are similar to fixed costs in that they need to be incurred and do not vary with the level of output but they differ from fixed costs in that they cannot be recovered if the firm shuts down. Irrecoverable expenditures on research and development provide an example of sunk costs. Once sunk costs are incurred they should not play a role in decision making since their opportunity cost is zero. In practice, many "fixed" investments are partially sunk as, for example, some equipment will have a low resale value because of asymmetric information problems or due to illiquid markets for used goods. Nonetheless, few investments are literally and completely "sunk," which means informed judgments must often be made about the extent to which investments are sunk.

In antitrust investigations, other cost concepts are sometimes used to determine cost benchmarks against which to measure prices. *Average avoidable costs* (AAC) are the average of the costs per unit that could have been avoided if a company had not produced a given discrete amount of output. It also takes into account any necessary fixed costs incurred in order to produce the output. *Long-run average incremental cost* (LRAIC) includes the variable and fixed costs necessary to produce a particular product. It differs from the average total costs because it is product specific and does not take into account costs that are common in the production of several products. For instance, if a product A is manufactured in a plant where product B is produced, the cost of the plant is not part of the LRAIC of producing A to the extent that it is not "incremental" to the production of product B.[22] Other more complex measures of costs are also used in the context of regulated industries, where prices for certain services are established in a way that guarantees a "fair price" to the buyer or a "fair return" to the seller.

In both managerial and financial accounts, variable costs are often computed and include the cost of materials used. Operating costs generally also include costs of sales and general administration that may be appropriately considered fixed. However, they may also include depreciation costs which may be approximating fixed costs or could even be more appropriately treated as sunk costs. If so, they would not be relevant for decision-making purposes. The variable costs or the operating costs without accounting depreciation are, in many cases, the most relevant costs for starting an economic analysis but ultimately judgments around cost data will need to be directly informed by the facts pertinent to a particular case.

[22] For LRAIC, see, for example, the discussion of the U.K. Competition Commission's inquiry in 2003 into phone-call termination charges in the United Kingdom and in particular the discussion of the approach in Office of Fair Trading (2003, chapter 10). In that case, the question was how high the price should be for a phone company to terminate a call on a rival's network. The commission decided it was appropriate that it should be evaluated on an "incremental cost" basis as it was found to be in a separate market from the downstream retail market, where phone operators were competing with each other for retail customers. In a regulated price setting, agencies sometimes decide it is appropriate for a "suitable" proportion of common costs to be recovered from regulated prices and, if so, some regulatory agencies may suggest using LRAIC "plus" pricing. Ofcom's (2007) mobile termination pricing decision provides an example of that approach.

1.2.2.3 Minimum Efficient Scale, Economies and Diseconomies of Scale

The minimum efficient scale (MES) of a firm or a plant is the level of output at which the long-run average cost (LRAC = AVC + FC/Q) reaches a minimum. The notion of long run for a given cost function deals with a time frame where the firm has (at least some) flexibility in changing its capital stock as well as its more flexible inputs such as labor and materials. In reality, cost functions can of course change substantially over time, which complicates the estimation and interpretation of long-run average costs. The dynamics of technological change and changing input prices are two reasons why the "long run" cannot in practice typically be taken to mean some point in time in the future when cost functions will settle down and henceforth remain the same.

We saw that average variable costs are minimized when they equal marginal costs. MES is the output level where the LRAC is minimized. At that point, it is important to note that MC = LRAC. For all plant sizes lower than the MES, the marginal cost of producing an extra unit is higher than it would be with a bigger plant size. The firm can lower its marginal and average costs by increasing scale. In some cases, plants bigger than the MES will suffer from diseconomies of scale as capital investments will increase average costs. In other cases average and marginal costs will become approximately constant above the MES and so all plants above the MES will achieve the same levels of these costs (and this case motivates the "minimum" in the MES). Figure 1.11 illustrates how much plant 1 would have to increase its plant size to achieve the MES. In that particular example, long-run costs increase beyond the MES. Even though MES is measured relative to a "long-run" cost measure, it is important to note that the "long run" in this construction refers to a firm's or plant's ability to change input levels holding all else equal. As a result, this intellectual construction is more helpful for an analyst when attempting to understand costs in a cross section of firms or plants at a given point in time than as an aid to understanding what will happen to costs in some distant time period. As we have already noted, over time both input prices and technology will typically change substantially.

We say a cost function demonstrates *economies of scale* if the long-run average cost decreases with output. A firm with a size lower than the MES will exhibit economies of scale and will have an incentive to grow. *Diseconomies of scale* occur when the long-run average variable cost increases with output.

In the short run, economies and diseconomies of scale will refer to the behavior of average and marginal costs as output is increased for a given capacity or plant size. Mathematically, define

$$S = \frac{\text{AC}}{\text{MC}} = \frac{C}{Q\partial C/\partial Q} = \frac{1}{\partial \ln C/\partial \ln Q}.$$

Thus we can derive a measure of the nature of economies of scale S directly from an estimated cost function by calculating the elasticity of costs with respect to

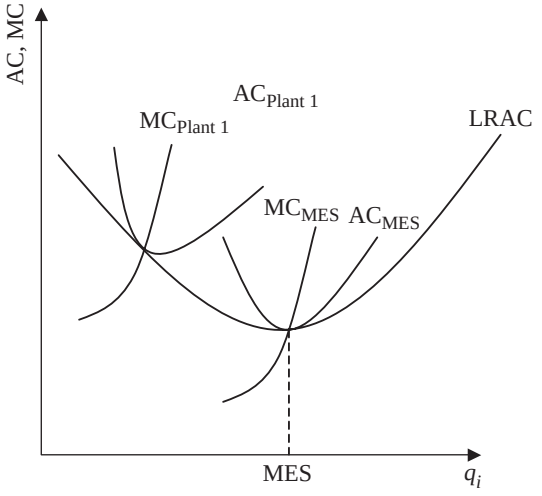

Figure 1.11. The minimum efficient scale of a plant.

output and computing its inverse. Alternatively, one can also use $S^* = 1 - \text{MC}/\text{AC}$ as a measure of economies of scale, which obviously captures exactly the same information about the cost function. If $S > 1$, we have economies of scale because AC is greater than MC. On the other hand, if $S < 1$, we have diseconomies of scale.

There are many potential sources of economies of scale. First, it could be that one of the inputs can only be acquired in large discrete quantities resulting in the firm having lower unit costs as it uses all of this input. An example would be the purchase of a passenger plane with several hundred available seats or the construction of an electricity grid. Also, as size increases, there may be scope for a more efficient allocation of resources within a firm resulting in cost savings. For example, small firms might hire generalists good at doing lots of things while a larger firm might hire more efficient, but indivisible, specialized personnel. Sources of economies of scale can be numerous and a good knowledge of the industry will help uncover the important ones.

If we have substantial economies of scale, the minimum efficient size of a firm may be big relative to the size of a market and as a result there will be few active firms in that market. In the most extreme case, to achieve efficiency a firm must be so large that only one firm will be able to operate at an efficient scale in a market. Such a situation is called a "natural" monopoly, because a benevolent social planner would choose to produce all market output using just one firm. Breaking up such a monopoly would have a negative effect on productive efficiency. Of course, since breaking up such a firm may remove pricing power, we may gain in allocative efficiency (lower prices) even though we may lose in productive efficiency (higher costs).

1.2.2.4 Scale Economies in Multiproduct Production

Determining whether there are economies of scale in a multiproduct firm can be a fairly similar exercise as for a single-product firm.[23] However, instead of looking at the evolution of costs as output of one good increases, we must look at the evolution of costs as the outputs of all goods increase. There are a variety of possible senses in which output can increase but we will often mean "increase in the same proportion." In that case, the term "economies of scale" will capture the evolution of costs as the scale of operation increases while maintaining a constant product mix.

Ray economies of scale (RES) occur when the average cost decreases with an increase in the scale of operation, or, equivalently, if the marginal cost of increasing the scale of operations lies below the average cost of total production.

In order to formalize our notion of economies of scale in a multiproduct environment, let us first define the multiproduct cost function, $C(q_1, q_2)$. Next fix two quantities q_1^0 and q_2^0 and define a new function

$$\tilde{C}(Q \mid q_1^0, q_2^0) \equiv C(Qq_1^0, Qq_2^0),$$

where Q is therefore a scalar measure of the scale of output which we will vary while holding the proportion of the two goods produced fixed. Total production can be expressed as

$$(q_1, q_2) = Q^*(q_1^0, q_2^0).$$

Graphically, if we trace a ray through all the points (Qq_1^0, Qq_2^0), $Q > 0$, our multiproduct measure of economies of scale will measure the economies of scale of the cost function above the ray (see figure 1.12).

The slope of the cost function along the ray is called the directional derivative by mathematicians, and provides the marginal cost of increasing the scale of operations:

$$\begin{aligned}\widetilde{\mathrm{MC}}(Q) &= \frac{\partial \tilde{C}(Q)}{\partial Q} = \frac{\partial C(Qq_1^0, Qq_2^0)}{\partial Q} \\ &= \frac{\partial C(q_1, q_2)}{\partial q_1}\frac{\partial q_1}{\partial Q} + \frac{\partial C(q_1, q_2)}{\partial q_2}\frac{\partial q_2}{\partial Q} \\ &= \sum_{i=1}^{2} \mathrm{MC}_i q_i^0.\end{aligned}$$

Given

$$\mathrm{RES} = \frac{\widetilde{\mathrm{AC}}}{\widetilde{\mathrm{MC}}} = \frac{\tilde{C}(Q)/Q}{\widetilde{\mathrm{MC}}(Q)} = \left(\frac{\partial \ln \tilde{C}(Q)}{\partial \ln Q}\right)^{-1},$$

$\mathrm{RES} > 1$ implies that we have ray economies of scale,
$\mathrm{RES} < 1$ implies that we have ray diseconomies of scale.

[23] For a very nice summary of cost concepts for multiproduct firms, see Bailey and Friedlander (1982).

1.2.2.5 Economies of Scope

Although economies of scale in multiproduct firms mirror the analysis of economies and diseconomies of scale in the single-output environment, important features of costs can also arise from the fact that several products are produced. The cost of producing one good may depend on the quantity produced of the other goods. Indeed, it may actually decrease because of the production of these other goods. For example, nickel and palladium are two metals sometimes found together in the ground. One option would be to build separate mines for extracting the nickel and palladium, but it would obviously be cheaper to build one and extract both from the ore.[24] Similarly, if a firm provides banking services, the cost of providing insurance services might be less for this firm than for a firm that only offers insurance. Such effects are referred to as economies of scope. Economies of scope can arise because certain fixed costs are common to both products and can be shared. For instance, once the reputation embodied in a brand name has been built, it can be cheaper for a firm to launch other successful products under that same brand.

Formally, *economies of scope* occur when it is cheaper to produce a given level of output of two products $(\tilde{q}_1, \tilde{q}_2)$ together compared with producing the two products separately by different firms (see Panzar and Willig 1981). To determine economies of scope we want to compare $C(\tilde{q}_1, \tilde{q}_2)$ and $C(\tilde{q}_1, 0) + C(0, \tilde{q}_2)$. If there are economies of scope, we want to understand the ranges over which they occur. For instance, we want to know the set of $(\tilde{q}_1, \tilde{q}_2)$ for which costs of joint production are lower than individual production:

$$\{(\tilde{q}_1, \tilde{q}_2) \mid C(\tilde{q}_1, \tilde{q}_2) < C(\tilde{q}_1, 0) + C(0, \tilde{q}_2)\}.$$

In addition, we will say *cost complementarities* arise when the marginal cost of production of good 1 is declining in the level of output of good 2:

$$\frac{\partial}{\partial q_2}\left(\frac{\partial C(q_1, q_2)}{\partial q_1}\right) = \frac{\partial^2 C(q_1, q_2)}{\partial q_2 \partial q_1} < 0.$$

An example of a cost function with economies of scope is the multiproduct function shown in figure 1.12. In the figure the cost of producing both goods is clearly lower than the sum of the costs of producing both goods separately. In fact, the figure shows there is actually a "dip" so that the cost of producing the two goods together is lower than the cost of producing them each individually. Clearly, this cost function demonstrates very strong form of economies of scope.[25]

[24] For example, the Norilsk mining center in the Russian high arctic produces nickel, palladium, and also copper. In that case, nickel mining began before the others at the surface, and underground mining began later.

[25] Note that it is sometimes important to be careful in distinguishing "economies of scope" from "subadditivity" where a single-product cost function satisfies $C(q_1 + q_2) < C(q_1 + 0) + C(0 + q_2)$.

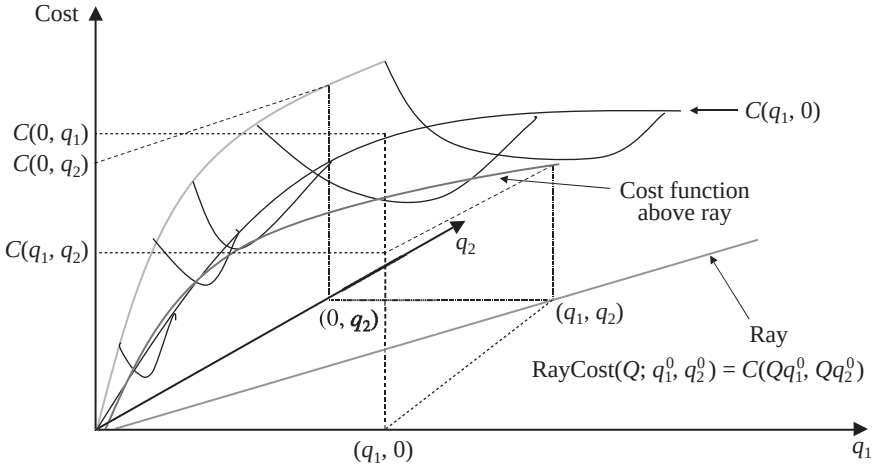

Figure 1.12. A multiproduct cost function. No unique notion of economies of scale in multiproduct environment, so we consider what happens to costs as expand production keeping output of each good in proportion. *Source*: Authors' rendition of a multiproduct cost function provided by Evans and Heckman (1984a,b) and Bailey and Friedlander (1982).

Economies of scope can have an effect on market structure because their existence will promote the creation of efficient multiproduct firms. When considering whether to break up or prohibit a multiproduct firm, it is in principle informative to examine the likely existence or relevance of economies of scope. In theory, it should be easy to evaluate economies of scope, but in practice when using estimated cost functions one must be extremely careful in assessing whether the cost estimates should be used. Very often one of the scenarios has never been observed in reality and therefore the hypothesis used in constructing the cost estimates can be speculative and with little possibility for empirical validation. A discussion of constructing cost data in a multiproduct context is provided in OFT (2003).[26]

In a multiproduct environment, conditional single-product cost functions tell us what happens to costs when the production of one product expands while maintaining constant the output of other products. Graphically, the cost function of product 1 conditional on the output of product 2 is represented as a slice of the cost function in figure 1.13 that, for example, is above the line between $(0, q_2)$ and (q_1, q_2).[27]

Conditional cost functions are useful when defining the *average incremental cost* (AIC) of increasing good 1 by an amount Δq_1, holding output of good 2 constant. This cost measure is commonly used to evaluate the cost of a firm's expansion in a particular line of products.

[26] See, in particular, chapter 6, "Cost and revenue allocation," as well as the case study examples in part 2.

[27] These objects are somewhat difficult to visualize in what is a complex graph. The central approach is to consider the univariate cost functions that result when the appropriate "slice" of the multivariate cost function is taken.

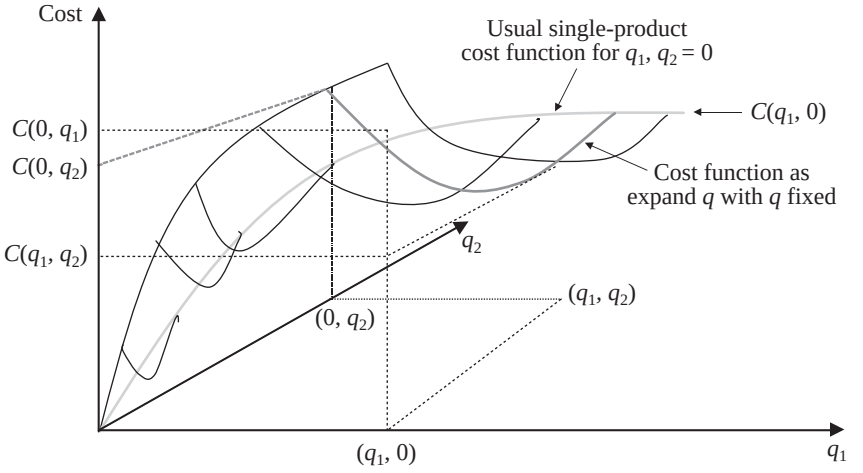

Figure 1.13. Conditional product cost function in multiproduct environment. We can still consider what happens to costs as the firm expands production of a single output at any fixed level of output of the other good.

Formally, the conditional average incremental cost function is defined as

$$\mathrm{AIC}_1(q_1 \mid q_2) = \frac{C(q_1 + \Delta q_1 \mid q_2) - C(q_1 \mid q_2)}{\Delta q_1}.$$

The conditional single-output marginal cost is defined as

$$\mathrm{MC}_1(q_1 \mid q_2) = \frac{\partial C(q_1, q_2)}{\partial q_1}.$$

Product-specific economies of scale can also be evaluated. Economies of scale in product 1, holding output of product 2 constant, are defined as

$$S_1(q_1 \mid q_2) = \frac{\mathrm{AIC}(q_1 \mid q_2)}{\mathrm{MC}(q_1 \mid q_2)}.$$

As usual, $S_1 > 1$ indicates the presence of economies of scale in the quantity produced of good 1 conditional on the level of output of good 2, while $S_1 < 1$ indicates the presence of diseconomies of scale.

1.2.2.6 Endogenous Economies of Scale

The discussion above has centered on economies of scale that are technologically determined. We discussed inputs that were necessary to production and that entered the production function in a way that was exogenously determined by the technology. However, firms may sometimes enhance their profits by investing in brands, advertising, and design or product innovation. The analysis of such effects involves

important demand-side elements but also has implications on the cost side. For example, if R&D or advertising expenditures involve large fixed outlays that are largely independent of the scale of production, they will result in economies of scale. Since firms will choose their level of R&D and advertising, these are often called "endogenous" fixed costs.[28] The decision to advertise or create a brand is not imposed exogenously by technology but rather is an endogenous decision of the firm in response to competitive conditions. The resulting economies of scale are also endogenous and, because the consumer welfare contribution of such expenditures may or may not be positive, it may or may not be appropriate to include them with the technologically determined economies of scale in the assessment of economies of scale and scope, depending on the context. For example, it would be somewhat odd for a regulator to uncritically allow a regulated monopoly to charge a price which covered any and all advertising expenditure, irrespective of whether such advertising expenditure was in fact socially desirable.

1.2.3 Input Demand Functions

Input demand functions provide a third potential source of information about the nature of technology in an industry. In this section we develop the relationship between profit maximization and cost minimization and describe the way in which knowledge of input demand equations can teach us about the nature of technology and more specifically provide information about the shape of cost functions and production functions.

1.2.3.1 The Profit-Maximization Problem

Generally, economists assume that firms maximize profits rather than minimize costs per se. Of course, minimizing the costs of producing a given level of output is a necessary but not generally a sufficient condition for profit maximization. A profit-maximizing firm which is a price-taker on both its output and input markets will choose inputs to solve

$$\max_{L,K,F} \Pi(L, K, F, p, p^L, p^k, p^F, u; \alpha) = \max_{L,K,F} pf(L, K, F, u; \alpha) - p^L L - p^K K - p^F F, \qquad (1.1)$$

where L denotes labor, K capital, F a third input, say, fuel, and $f(L, K, F, u; \alpha)$ the level of production; p denotes the price of the good produced and the other prices (p^L, p^K, p^F) are the prices of the inputs. The variable u denotes an unobserved efficiency component and α represents the parameters of the firm's production function.

[28] Sutton (1991) studies the case of endogenous sunk costs. In his analysis, he assumes that R&D and advertising expenditures are sunk by the time firms compete in prices although in other models they need not be.

If the firm is a price-taker on its output and input markets, then we can equivalently consider the firm as solving a two-step procedure. First, for any given level of output it chooses its cost-minimizing combination of inputs that can feasibly supply that output level. Second, it chooses how much output to supply to maximize profits.

Specifically,

$$C(Q, p^L, p^K, p^F, u; \alpha) = \min_{K,L,F} p^L L + p^K K + p^F F \quad \text{subject to} \quad Q \leqslant f(K, L, F, u; \alpha) \tag{1.2}$$

and then define

$$\max_Q \Pi(Q, p, p^L, p^K, p^F, u; \alpha) = \max_Q pQ - C(Q, p^L, p^K, p^F, u; \alpha). \tag{1.3}$$

With price-taking firms, the solution to (1.1) will be identical to the solution of the two-stage problem, solving (1.2) and then (1.3).

If the firm is not a price-taker on its output market, the price of the final good p will depend on the level of output Q and we will write it as a function of Q, $P(Q)$, in the profit-maximization problem. Nonetheless, we will still be able to consider the firm as solving a two-step problem provided once again that the firm is a price-taker on its input markets. Profit-maximizing decisions in environments where firms may be able to exercise market power will be considered when we discuss oligopolistic competition in section 1.3.[29]

1.2.3.2 Input Demand Functions

Solving the cost-minimization problem

$$C(Q, p^L, p^K, p^F, u; \alpha) = \min_{K,L,F} p^L L + p^K K + p^F F \quad \text{subject to} \quad Q \leqslant f(K, L, F, u; \alpha)$$

[29] If the firm is not a price-taker on its input markets, the price of the inputs may also depend on the level of inputs chosen and, while we can easily define the firm's cost function as

$$C(Q, u; \alpha, \vartheta_L, \vartheta_K, \vartheta_F) = \min_{K,L,F} p^L(L; \vartheta_L)L + p^K(K; \vartheta_K)K + p^F(F; \vartheta_F)F \quad \text{subject to} \quad Q \leqslant f(K, L, F, u; \alpha),$$

the resulting cost function should not, for example, depend on the realized values of the input prices but rather on the structure of the input pricing functions, $C(Q, u; \alpha, \vartheta_L, \vartheta_K, \vartheta_F)$. This observation suggests that estimation of cost functions in environments where firms can get volume discounts from their suppliers are certainly possible, but doing so requires both careful thought about the variables that should be included and also careful thought about interpretation of the results. In particular, in general the shape of the cost function will now capture a complex mixture of incentives generated by (i) substitution possibilities generated by the production function and (ii) of the pricing structures faced in input markets.

produces the conditional input demand equations, which express the inputs demanded as a function of input prices, conditional on output level Q:

$$L = L(Q, p^L, p^K, p^F, u; \alpha),$$
$$K = K(Q, p^L, p^K, p^F, u; \alpha),$$
$$F = F(Q, p^L, p^K, p^F, u; \alpha).$$

Conveniently, Shephard's lemma establishes that cost minimization implies that the inputs demanded are equal to the derivative of the cost function with respect to the price of the input:

$$L = L(Q, p^L, p^K, p^F, u; \alpha) = \frac{\partial C(Q, p^L, p^K, p^F, u; \alpha)}{\partial p^L},$$
$$K = K(Q, p^L, p^K, p^F, u; \alpha) = \frac{\partial C(Q, p^L, p^K, p^F, u; \alpha)}{\partial p^K},$$
$$F = F(Q, p^L, p^K, p^F, u; \alpha) = \frac{\partial C(Q, p^L, p^K, p^F, u; \alpha)}{\partial p^F}.$$

The practical relevance of Shephard's lemma is that it means that many of the parameters in the cost function can be retrieved from the input demand equations and vice versa. That means we have a third type of data set, data on input demands, that will potentially allow us to learn about technology parameters.[30]

Finally, if firms are price-takers on output markets, solving the profit-maximizing problem produces the unconditional input demand equations that express input demand as a function of the price of the final good and the prices of the inputs:

$$L = L(p, p^L, p^K, p^F, u; \alpha),$$
$$K = K(p, p^L, p^K, p^F, u; \alpha),$$
$$F = F(p, p^L, p^K, p^F, u; \alpha).$$

Note that both conditional (on Q) and unconditional factor demand functions depend on productivity, u. Firms with a higher productivity will tend to produce more but will use fewer inputs than other firms in order to produce any given level of output. That observation has a number of important implications for the econometric analysis of production functions since it can mean input demands will be correlated with the unobservable productivity, so that we need to address the endogeneity of input

[30] For a technical discussion of the result, see the section "Duality: a mathematical introduction" in Mas-Colell et al. (1995). In the terminology of duality theory, the cost function plays the role of the "support function" of a convex set. Specifically, let the convex set be $S = \{(K, L, F) \mid Q \leqslant f(K, L, F, u; \alpha)\}$ and define the "support function" $\mu(p_L, p_K, p_F) = \min_{(K,L,F)}\{p_L L + p_K K + p_F F \mid (L, K, F) \in S\}$, then roughly the duality theorem says that there is a unique set of inputs (L^*, K^*, F^*) so that $p_L L^* + p_K K^* + p_F F^* = \mu(p_L, p_K, p_F)$ if and only if $\mu(p_L, p_K, p_F)$ is differentiable at (p_L, p_K, p_F). Moreover, $L^* = \partial\mu(p_L, p_K, p_F)/\partial p_L$, $K^* = \partial\mu(p_L, p_K, p_F)/\partial p_K$, and $F^* = \partial\mu(p_L, p_K, p_F)/\partial p_F$.

demands in the estimation of production functions (see, for example, the discussion in Olley and Pakes 1996; Levinsohn and Petrin 2003; Ackerberg et al. 2005). The estimation of cost functions is discussed in more detail in chapter 3.

1.3 Competitive Environments: Perfect Competition, Oligopoly, and Monopoly

In a perfectly competitive environment, market prices and output are determined by the interaction of demand and supply curves, where the supply curve is determined by the firms' costs. In a perfectly competitive environment, there are no strategic decisions to make. Firms spend their time considering market conditions, but do not focus on analyzing how rivals will respond if they take particular decisions. In more general settings, firms will be sensitive to competitors' decisions regarding key strategic variables. Both the dimensions of strategic behavior and the nature of the strategic interaction will then be fundamental determinants of market outcomes. In other words, the strategic variables—perhaps advertising, prices, quantity, or product quality—and the specific way firms in the industry react to decisions made by rival firms in the industry will determine the market outcomes we observe. The primary lesson of game theory for firms is that they should spend as much time thinking about their rivals as they spend thinking about their own preferences and decisions. When firms do that, we say that they are interacting strategically. Evidence for strategic interaction is often quite easy to find in corporate strategy and pricing documents.

In this section, we describe the basic models of competition commonly used to model firm behavior in antitrust and merger analysis, where strategic interaction is the norm rather than the exception. Of course, since this is primarily a text on empirical methods, we certainly will not be able to present anything like a comprehensive treatment of oligopoly theory. Rather, we focus attention on the fundamental models of competitive interaction, the models which remain firmly at the core of most empirical analysis in industrial organization. Our ability to do so and yet cover much of the empirical work used in practical settings suggests the scope of work yet to be done in turning more advanced theoretical models into tools that can, as a practical matter, be used with real world data.

While some of the models studied in this section may to some eyes appear highly specialized, we will see that the general principles of building game theoretic economic (and subsequently econometric) models are entirely generic. In particular, we will always wish to (1) describe the primitives of the model, in this case the nature of demand and the firms' cost structures, (2) describe the strategic variables, (3) describe the behavioral assumptions we make about the agents playing the game, generally profit maximization, and then, finally, (4) describe the nature of equilibrium, generally Nash equilibrium whereby each player does the best they can given

the choice of their rival(s). We must describe the nature of equilibrium as each firm has its own objective and these often competing objectives must be reconciled if a model is to generate a prediction about the world.

1.3.1 Quantity-Setting Competition

The first class of models we review are those in which firms choose their optimal level of output while considering how their choices will affect the output decisions of their rivals. The strategic variable in this model is quantity, hence the name: quantity-setting competition. We will review the general model and then relate its predictions to the predicted outcomes under perfect competition and monopoly.

1.3.1.1 The Cournot Game

The modern models of quantity-setting competition are based on that developed by Antoine Augustin Cournot in 1838. The Cournot game assumes that the only strategic variable chosen by firms is their output level. The most standard analysis of the game considers the situation in which firms move simultaneously and the game has only one period. Also, it is assumed that the good produced is homogeneous, which means that consumers can perfectly substitute goods from the different firms and implies that there can only be one price for all the goods in the market. To aid exposition we first develop a simple numerical example and then provide a more general treatment.

For simplicity suppose there are only two firms and that total and marginal costs are zero. Suppose also that the inverse demand function is of the form

$$P(q_1 + q_2) = 1 - (q_1 + q_2),$$

where the fact that market price depends only on the sum of the output of the two firms captures the perfect substitutability of the two goods. As in all economic models, we must be explicit about the behavioral assumptions of the firms being considered. A probably reasonable, if sometimes approximate, assumption about most firms is that they attempt to maximize profits to the best of their abilities. We shall follow the profession in adopting profit maximization as a baseline behavioral assumption.[31] The assumptions on the nature of consumer demand, together with the assumption on costs, which here we shall assume for simplicity involve zero

[31] Economists quite rightly question the reality of this assumption on a regular basis. Most of the time we fairly quickly receive reassurance from firm behavior, company documents, and indeed stated objectives, at least those stated to shareholders or behind closed doors. Public reassurances and marketing messages are, of course, a different matter and moreover individual CEOs or other board members (and indeed investors) certainly can consider public image or other social impacts of economic activity. For these reasons and others there are always departures from at least a narrow definition of profit maximization and we certainly should not be dogmatic about any of our assumptions. And yet in terms of its predictive power, profit maximization appears to do rather well and it would be a very brave (and frankly irresponsible) merger authority which approved, say, a merger to monopoly because the merging parties told us that they did not maximize profits but rather consumer happiness.

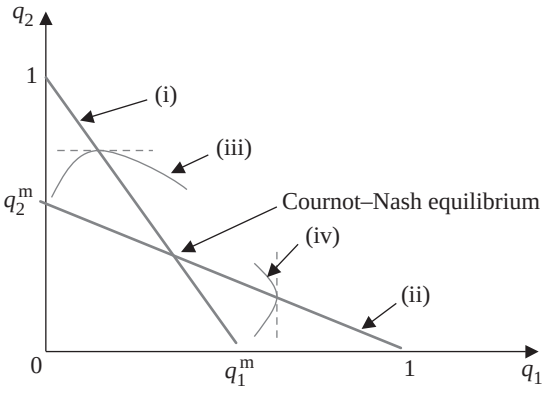

Figure 1.14. Reaction functions in the Cournot model. (i) $R_1(q_2)$: $q_1 = \frac{1}{2}(1-q_2)$; (ii) $R_2(q_1)$: $q_2 = \frac{1}{2}(1-q_1)$; (iii) $\bar{\pi}_1 = q_1(1-q_1-q_2)$ (isoprofit line for firm 1); (iv) $\bar{\pi}_2 = q_2(1-q_1-q_2)$ (isoprofit line for firm 2).

marginal costs, $c_1 = c_2 = 0$, allow us to describe the way in which each firm's profits depend on the two firms' quantity choices. In our example,

$$\pi_1(q_1,q_2) = (P(q_1+q_2)-c_1)q_1 = (1-q_1-q_2)q_1,$$
$$\pi_2(q_1,q_2) = (P(q_1+q_2)-c_2)q_2 = (1-q_1-q_2)q_2.$$

Given our behavioral assumption, we can define the reaction function, or best response function. This function describes the firm's optimal quantity decision for each value of the competitor's quantity choice. The reaction function can be easily calculated given our assumption of profit-maximizing behavior. The first-order condition from profit maximization by firm 1 is

$$\frac{\partial \pi_1(q_1,q_2)}{\partial q_1} = (1-q_2) - 2q_1 = 0.$$

Solving for the quantity of firm 1 produces firm 1's reaction function

$$q_1 = R_1(q_2) = \tfrac{1}{2}(1-q_2).$$

If both firms choose their quantity simultaneously, the outcome is a Nash equilibrium in which each firm chooses their optimal quantity in response to the other firm's choice. The reaction functions of firms 1 and 2 respectively are

$$R_1(q_2): \quad q_1 = \tfrac{1}{2}(1-q_2) \qquad \text{and} \qquad R_2(q_1): \quad q_2 = \tfrac{1}{2}(1-q_1).$$

Solving these two linear equations describes the Cournot–Nash equilibrium

$$q_1 = \tfrac{1}{2}(1-q_2) = \tfrac{1}{2}(1-\tfrac{1}{2}(1-q_1)) = \tfrac{1}{2}(\tfrac{1}{2}+\tfrac{1}{2}q_1) = \tfrac{1}{4}+\tfrac{1}{4}q_1,$$

so that the equilibrium output for firm 1 is

$$\tfrac{3}{4}q_1^{NE} = \tfrac{1}{4} \quad \Longrightarrow \quad q_1^{NE} = \tfrac{1}{3}.$$

The equilibrium output for firm 2 will then be

$$q_2^{\text{NE}} = \tfrac{1}{2}(1 - \tfrac{1}{3}) = \tfrac{1}{3}.$$

The resulting profits will be

$$\pi_1^{\text{NE}} = \pi_2^{\text{NE}} = \tfrac{1}{3}(1 - \tfrac{1}{3} - \tfrac{1}{3}) = \tfrac{1}{9}.$$

Graphically, the Cournot–Nash equilibrium is the intersection between the two firms' reaction curves as shown in figure 1.14.

The reaction function is the quantity choice that maximizes the firm's profits for each given quantity choice of its competitor. The profits for the different combinations of output choices in a Cournot duopoly are plotted in figure 1.15.

Isoprofit lines show all quantity pairs (q_1, q_2) that generate any given fixed level of profits for firm 1. These lines would be represented by horizontal slices of the surface in figure 1.15. We can define a given fixed level of profit $\bar{\pi}_1$ as

$$\bar{\pi}_1 = (1 - q_1 - q_2)q_1.$$

Note that given a level of profits and quantity chosen by firm 1, the output of firm 2 can be inferred as

$$q_2 = 1 - q_1 - \frac{\bar{\pi}_1}{q_1}.$$

Isoprofit lines can be drawn in a contour plot as shown in figure 1.16. Firm 1's best response to any given q_2 is where it reaches highest isoprofit contour. The figure reveals an important characteristic of the model: for a fixed output of firm 1, firm 1's profits increase as firm 2 lowers its output. If the competitor chooses not to produce, the profit-maximizing response is to produce the monopoly output and make monopoly profits. That is, if $q_2 = 0$, then $q_1 = \frac{1}{2}(1 - q_2) = 0.5$ and the profit will be

$$\bar{\pi}_1 = (1 - q_1 - q_2)q_1 = (1 - 0.5 - 0)0.5 = 0.25.$$

More generally, the first-order conditions in the Cournot game produce the familiar condition that marginal revenue is equated to marginal costs. Given the profit function

$$\pi_i(q_i, q_j) = P(q_1 + q_2)q_i - C_i(q_i),$$

the first-order conditions are

$$\frac{\partial \pi_i(q_1, q_2)}{\partial q_i} = \underbrace{P(q_1 + q_2) + q_i P'(q_1 + q_2)}_{\text{Marginal revenue}} - \underbrace{C_i'(q_i)}_{\text{Marginal cost}} = 0,$$

which in general defines an implicit function we shall call firm i's reaction curve, $q_i = R_i(q_{-i})$, where q_{-i} denotes the output level of the other firm(s).[32] In our

[32] That is, we can think of the first-order condition defining a value of q_i which, given the quantities chosen by other firms, will set the first-order condition to zero.

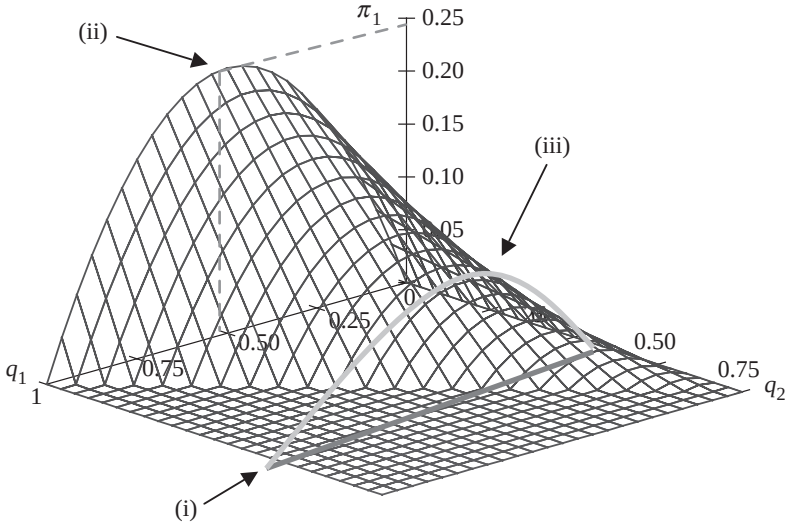

Figure 1.15. Profit function for a two-player Cournot game as a function of the strategic variables for each firm. (i) For each fixed q_2, firm 1 chooses q_1 to maximize her profits; (ii) the q_1 that generates the maximal level of profit for fixed value of q_2 is firm 1's best response to q_2; (iii) profits if firm 1 is a monopoly: $q_2 = 0$, $q_1 = 0.5$, $\Pi_1 = 0.25$.

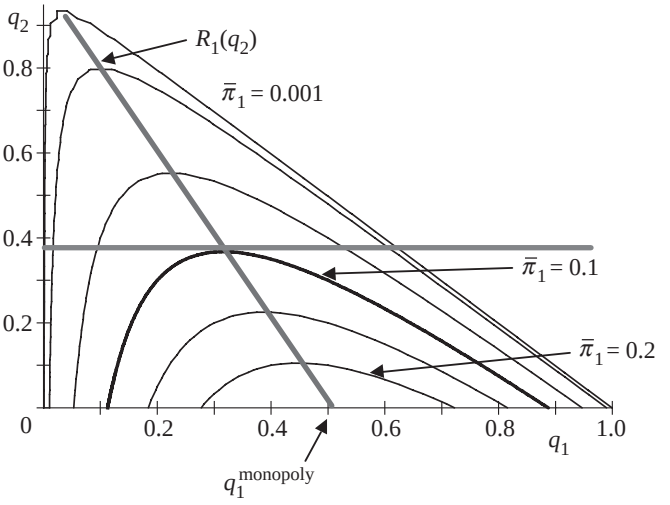

Figure 1.16. Isoprofit lines in simple Cournot model.

two-player case, we have two first-order conditions to solve, which can each in turn be used to define the reaction functions $q_1 = R_1(q_2)$ and $q_2 = R_2(q_1)$. In general, with N active firms we will have N first-order conditions to solve. Nash equilibrium is the intersection of the reaction functions so that solving the reaction functions can

involve solving N nonlinear equations. Our numerical example makes these equations linear (and hence easy to solve analytically) by assuming that inverse demand curves are linear and marginal costs constant. In general, however, computers can usually solve nonlinear systems of equations for us provided a solution exists.[33] Ideally, we would like a "unique" prediction about the world coming out of the model and we will get one only if there is a unique solution to the set of first-order conditions.[34]

Note that since profits are always revenues minus costs, marginal profitability can as always be described as marginal revenue minus marginal cost. At a maximum, the first-order condition will be zero and hence we have the familiar result that profit maximization requires that marginal revenue equals marginal cost.

To see the impact of strategic decision making, at this point it is worth taking a moment to relate the Cournot optimality conditions, with perhaps the more familiar results from perfect competition and monopoly.

1.3.1.2 Quantity Choices under Perfect Competition

In an environment with price-taking firms, the first-order condition from profit maximization leads to equating the marginal cost of the firm to the market price, provided, of course, that there are no fixed costs so that we can ignore the sometimes important constraint that profits must be nonnegative:

$$\pi_i(q_i) = pq_i - C_i(q_i) \implies \frac{\partial \pi_i(q_i)}{\partial q_i} = p - C_i'(q_i) = 0 \implies p = C_i'(q_i).$$

Evidently, if the price is €1 and the marginal cost of producing one more unit is €0.90, then my profits will increase if I expand production by that unit. Similarly, if the price is €1 while the marginal cost of production of the last unit is €1.01, my profits will increase if I do not produce that last unit. Repeating the calculation makes clear that quantity will adjust until marginal cost equals marginal revenue, which by assumption in this context is exactly equal to price.

Going further, since all firms face the same price, all firms will choose their quantities in order to help price equal marginal cost so that $C_i'(q_i) = C_j'(q_j) = p$. In particular, that means marginal costs are equalized across firms because all firms face the same selling price.

Note that joint cost minimization also implies that the marginal costs are equated across active firms. Consider what happens when we minimize the total cost of producing any given level of total output:

$$\min_{q_1, q_2} C_1(q_1) + C_2(q_2) \quad \text{subject to} \quad q_1 + q_2 = Q.$$

[33] For the conditions required for existence of a solution to these nonlinear equations and hence for Nash equilibrium, see Novshek (1985) and Amir (1996).

[34] In general, a system of N nonlinear equations may have no solution, one solution, or many solutions. In economic models the more commonly problematic situation arises when models have multiple equilibria. We discuss the issue of multiple equilibria further in chapter 5.

In particular, note that such a problem yields the following first-order optimality conditions,

$$C_1'(q_1) = C_2'(q_2) = \lambda,$$

where λ is the Lagrange multiplier in the constrained minimization exercise. Clearly, minimizing the total costs for any given level of production will involve equalizing marginal costs.

Intuitively, if we had firms producing at different marginal costs, the last unit of output produced at the firm with higher marginal costs could have been more efficiently produced by the firm with lower marginal costs. Perfect competition, and in particular the price mechanism, acts to ensure that output is distributed across firms in a way that ensures that all units in the market are as efficiently produced as possible given the existing firms' technologies. It is in this way that prices help ensure *productive efficiency*.

In perfectly competitive markets, prices also act to ensure that the marginal cost of output is also equal to its marginal benefit, so that we have *allocative efficiency*. To see why, recall that the market demand curve describes the marginal value of output to consumers at each level of quantity produced. At any given price, the last unit of the good purchased will have a marginal value equal to the price. The supply curve of the firm under perfect competition is the marginal cost for each level of quantity since firms adjust output until $p = \mathrm{MC}(q)$ in equilibrium. Therefore, when price adjusts to ensure that aggregate supply is equal to aggregate demand, it ensures that the marginal valuation of the last unit sold is equal to the marginal cost of its production. In other words, the market produces the quantity such that the last unit is valued by consumers as much as it costs to produce. It is this remarkable mechanism that ensures that the market outcome under perfect competition is socially efficient.

1.3.1.3 Quantity Setting under Monopoly

In a monopoly, there is only one firm producing and therefore the market price will be determined by this one firm when it chooses the total quantity to produce. As usual, the firm's profit function is

$$\pi_i(q_i) = P(q_i)q_i - C_i(q_i)$$

and the corresponding first-order condition is

$$\frac{\partial \pi_i(q_i)}{\partial q_i} = \underbrace{P(q_i) + P'(q_i)q_i}_{\text{Marginal revenue}} - \underbrace{C_i'(q_i)}_{\text{Marginal cost}} = 0.$$

Note that the first-order condition from monopoly profit maximization is a special case of the first-order condition under Cournot where the quantity of the other firms is set to zero. The monopolist, like any profit-maximizing firm in any of the scenarios analyzed, chooses its quantity in order to set marginal revenue equal to marginal cost.

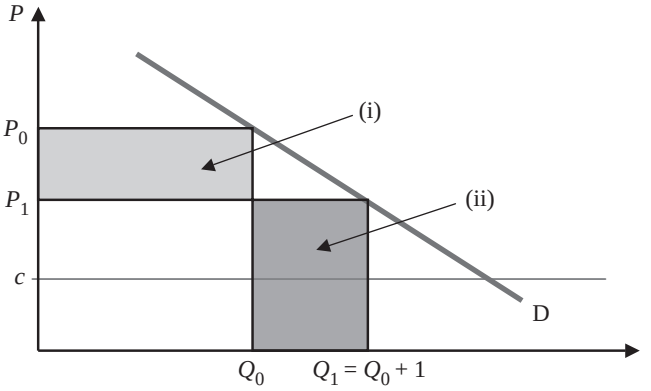

Figure 1.17. Demand, revenue, and marginal revenue.
(i) Loss of revenue $Q_0 \Delta P \to QP'(Q)$. (ii) Increase of revenue P_1.

Note that the slope of the inverse demand function $P'(q_i)$ is negative. That means that the marginal revenue generated by an extra unit sold is smaller than the marginal valuation by the consumers as represented by the inverse demand curve $P(q_i)$. Graphically, the marginal revenue curve is below the inverse demand curve for a monopolist. The reason for this is that the monopolist cannot generally lower the price of only the last unit. Rather she is typically forced to lower the price for all the units previously produced as well. Increasing the price therefore increases the revenue for each product which continues to be sold at the higher price, but reduces revenue to the extent that the number of units sold falls. Figure 1.17 illustrates the marginal revenue when the monopolist increases its sales by one unit from Q_0 to Q_1. To sell Q_1, the monopolist must reduce its selling price to P_1, down from P_0. The marginal revenue associated with selling that extra unit is therefore

$$\begin{aligned} \text{MR} &= P_1 Q_1 - P_0 Q_0 = P_1(Q_1 - Q_0) + Q_0(P_1 - P_0) \\ &= P_1 \times 1 + Q_0 \Delta P = P_1 + Q_0 \Delta P. \end{aligned}$$

Under a profit-maximizing monopoly, marginal revenue of the last unit sold is lower than the marginal valuation of consumers. As a result, the monopoly outcome is not socially efficient. At the level of quantity produced, there are consumers for whom the marginal value of an extra unit is greater than the marginal cost of supplying it. Unfortunately, even though some consumers are willing to pay more than the marginal cost of production, the monopolist prefers not to supply them to avoid suffering from lower revenues from the customers who remain. The welfare loss imposed by a monopoly market is illustrated in figure 1.18.

1.3.1.4 Comparing Monopoly and Perfect Competition to the Cournot Game

In all competition models, profit maximization implies that the firm will set marginal revenue equal to marginal cost: MR = MC. Whereas in perfect competition, firms'

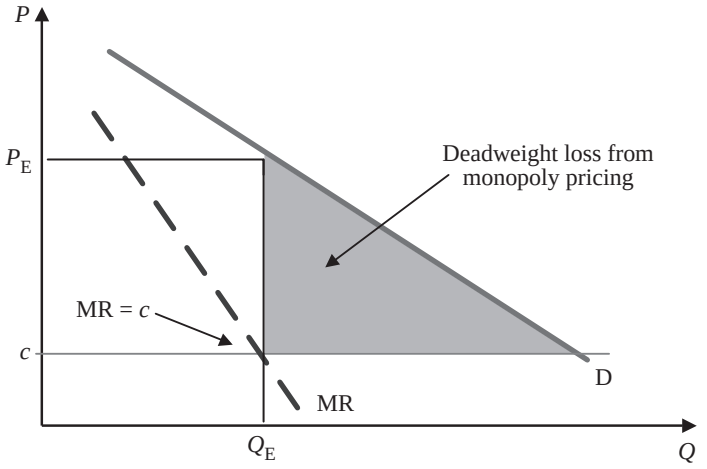

Figure 1.18. Welfare loss from monopoly pricing compared with perfect competition.

marginal revenue is the market price, in a monopoly market the marginal revenue will be determined by the monopolist's choice of quantity. In a Cournot game, the marginal revenue depends on the firm's output decision as well as on the rivals' output choices.

Specifically, in a Cournot game, we showed that the first-order condition from profit maximization,

$$\text{Max}_{q_i} \pi_i(q_i, q_j) = P(q_1 + q_2)q_i - C_i(q_i),$$

is

$$\frac{\partial \pi_1(q_1, q_2)}{\partial q_1} = P(q_1 + q_2) + q_1 P'(q_1 + q_2) - C_1'(q_1) = 0.$$

As always, the firm equates marginal revenue to marginal cost. As in the monopolist case, the marginal revenue is smaller than the marginal valuation by the consumer. In particular, because of the negative slope of the demand curve, we have

$$\text{MR}_1(q_1, q_2) = P(q_1 + q_2) + q_1 P'(q_1 + q_2) < P(q_1 + q_2).$$

Graphically, the marginal revenue curve is below the demand curve.

First notice that under Cournot, the effect of the decrease in price $P'(q_1 + q_2)$ is only counted for the q_1 units produced by firm 1, while under monopoly the effect is counted for the entire market output.

Second, under Cournot, the marginal revenue of each firm is affected by its output decision *and* by the output decisions of competing firms, outputs which do affect the equilibrium price. The result is a negative externality across firms. When firm 1 chooses its optimal quantity, it does not take into account the potential reduction in profits that other firms suffer with an increase in total output. This effect is called a "business stealing" effect. As a result Cournot firms will jointly produce and sell

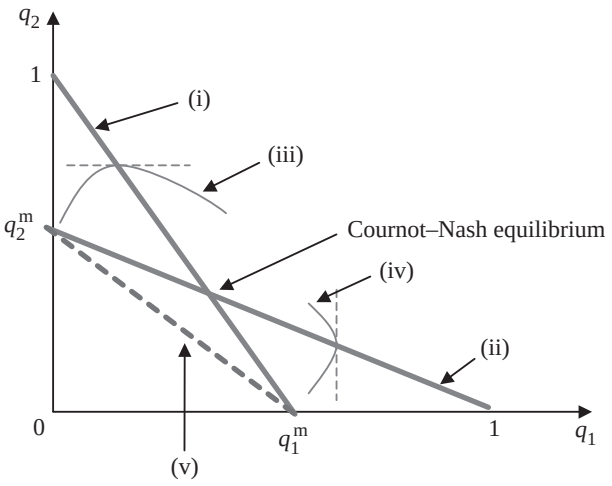

Figure 1.19. Cournot equilibrium versus monopoly: (i)–(iv) as in figure 1.14; (v) output combinations that maximize joint profits.

at a lower price than an equivalent (multiplant) monopolist. Figure 1.19 illustrates the joint industry profit-maximizing output combinations and the Cournot–Nash equilibrium. If firms have the same constant marginal cost, any output allocation among the two firms such that the sum of their output is the monopoly quantity, i.e., any combination fulfilling $q_1 + q_2 = Q^{\text{monopoly}}$, will maximize industry profits. The industry profit-maximizing output levels are represented by the dashed line in the figure. The Cournot–Nash equilibrium is reached by each firm maximizing its profits individually. It is represented by the intersection of the two firms' reaction function. The total output in the Cournot–Nash equilibrium is larger than under monopoly. At a very basic level, competition authorities which apply a consumer welfare standard are aiming to maintain competition so that the negative externalities across firms are preserved. In so doing they ensure that firms endow positive externalities on consumers, in the form of consumer surplus.

Under perfect competition, social welfare is maximized because the market equates the marginal valuations with the marginal cost of production. A monopolist firm will decide not to produce units that are valued more than their costs in order not to decrease total profits and therefore social welfare is not maximized. That said, production costs are still minimized.[35] Social welfare is not maximized with Cournot competition but the loss of welfare is less severe than in the monopoly game thanks to the extra output produced as a result of the Cournot externalities. Output and social welfare will be higher than in the monopolist case since the firm does not factor in the effect of lower prices on the other firms' revenues. When

[35] Experience suggests that monopolies will often, among other things, suffer from X-inefficiency as well as restricting output, so this result should probably not be taken too literally. (See the literature on X-inefficiency following Leibenstein (1966).)

a firm's output expansion only has a small effect on price, the Cournot outcome becomes close to the competitive outcome. This is the case when there are a large number of firms and each firm is small relative to total market output. In a Cournot equilibrium, marginal cost can vary across firms and so industry production costs are not necessarily minimized unless firms are symmetric and marginal costs are equal across firms.

In summary, Cournot equilibrium will be bad for the firms' profits but good for consumer welfare relative to the monopoly outcomes. On the other hand, Cournot will be good for the firms' profits but bad for consumer welfare relative to a market with price-taking firms.

The Cournot model has had a profound impact on competition analysis and it is sometimes described as the model that antitrust practitioner's have in mind when they first consider the economics of a given situation. As we discuss in chapter 6, the model is, among other things, the motivation for considering the commonly used Herfindahl–Hirschman index (HHI) of concentration.

1.3.2 Price-Setting Competition

Oligopoly theory was developed to explain what would happen in markets when there were small numbers of competing firms. Cournot's (1838) theory was based on a form of competition in which firms choose quantities of output and the construction appeared to fit with the empirical evidence that firms seemed to price above marginal cost, the price prediction of the perfect competition model. While Cournot was successful in predicting price above marginal cost, some unease remains about whether firms genuinely choose the level of output they produce or determine their selling price and sell whatever demand there is for the product at that price. This observation motivated the analysis of what became one of the most important theoretical results in oligopoly theory, Bertrand's paradox.

1.3.2.1 The Bertrand Paradox

Bertrand (1883) considered that Cournot's model embodied an unrealistic assumption about firm behavior. He suggested that a more realistic model of actual firm behavior was that firms choose prices and then supply the resulting demand for their product. If so, then price rather than quantity would be the relevant strategic variable for the firms. Bertrand's model does indeed seem highly intuitive since firms do frequently set prices for their products. Thus from the point of view of the description of actual firm behavior, it seems to fit reality better than Cournot's model. Nonetheless, we now treat Bertrand's model as important because it produces paradoxical, counterintuitive results.[36] Like many results in economics, Bertrand's results are

[36] A paradox is defined in the Oxford English Dictionary as a statement or tenet contrary to received opinion or belief; often with the implication that it is marvelous or incredible; sometimes with unfavorable connotation, as being discordant with what is held to be established truth, and hence absurd or fantastic.

usually considered important because they force us to ask carefully which of his assumptions are violated.[37]

Bertrand considers a duopoly in a homogeneous products market with a market demand curve $Q = D(p)$. If firm 1 prices above its competitors, customers will only buy from the cheaper firm and firm 1's demand will be 0. If firm 1 prices below its competitor, it will supply the whole market since no customer will want to buy from firm 2. If firm 1 and firm 2 offer the same price, then demand will be split between the two firms, we shall assume equally (the exact split is not crucial). The demand curve for firm 1 will be as follows:

$$q_1 = D_1(p_1, p_2) = \begin{cases} D(p_1) & \text{if } p_1 < p_2, \\ D(p_1)/2 & \text{if } p_1 = p_2, \\ 0 & \text{if } p_1 > p_2, \end{cases}$$

where demand is assumed to be split evenly if the two firms charge identical prices. Assuming constant marginal costs c for both firms, Bertrand showed that there is a unique Nash equilibrium: $p_1^* = p_2^* = c$.

The proof is based on the following arguments. If firm 2 prices above marginal costs, $p_2 > c$, then firm 1 can undercut slightly by setting $p_1 = p_2 - \varepsilon$, where ε is very small, and take the whole market. Provided that p_1 is above marginal cost, firm 1 will still make positive profits. However, firm 2 also has the incentive to undercut firm 1 by a slight amount and for as long as the prices are above marginal costs firms will have an incentive to undercut each other. No firm has an incentive to price below marginal costs because that would imply that they would make losses. Therefore, the only possible stable outcome is the Nash equilibrium, where both firms are pricing at marginal cost. In this situation, both firms make zero profits.

The Bertrand game has a very important, strong implication. Namely, Bertrand's result implies that as long as there is more than one player in the market, prices for all firms will be set to marginal cost and profits will be zero. In other words, as long as there are at least two firms in the market for a homogeneous product and no fixed costs, the market will produce the perfect competition equilibrium. Such a result occurs despite the fact that both firms would be better off if they both increased their prices! Bertrand's result is considered to be a paradox because intuitively, neither business people nor economists usually expect a duopoly to produce the same results as we would get from a perfectly competitive market. Moreover, the data substantiate such intuition: the vast majority of oligopolies have positive markups and the firms involved do not generally price at, or often even close to, marginal cost.

[37] Another example of such a result is the Modigliani and Miller theorem (1958). These authors showed that under certain—on the face of it highly plausible—assumptions, the capital structure of a firm does not matter for the value of the firm. Of course, most practitioners and academics believed and believe that the proportions of debt and equity do matter and so for fifty years corporate finance has studied violations of Modigliani and Miller's assumptions, which include the absence of taxes and bankruptcy costs as well as the presence of full information and efficient markets.

How do we react to Bertrand's paradox? Well, if you have a theory with well-defined assumptions which gives you implausible predictions, it is time to look at the assumptions. Following Bertrand's results, economists have examined a large variety of alternative assumptions in order to obtain predictions that conform better to reality.

In the next three sections we discuss three further important examples which, along with others we will discuss later in the book, have been found to modify Bertrand's model in a way that relaxes his strong conclusions. First, fixed costs can be introduced into the model. Second, product differentiation can be introduced. Product differentiation gives a certain degree of pricing power to each firm. Third, capacity constraints, which put a limit to the percentage of the market that any firm can supply, have been incorporated. We discuss each model in turn.

1.3.2.2 Bertrand Competition with Fixed Costs

First note that the Bertrand result that price equals marginal cost only applies in cases where fixed costs are zero. If fixed costs are nonzero, then firms maximize profits subject to the nonnegativity constraint that profits must be nonnegative while profits may well be negative if prices were set at marginal cost. Firm one's problem can be written as follows:

$$\max_{p_1}(p_1 - c_1)D_1(p_1, p_2) - F_1 \quad \text{subject to} \quad (p_1 - c_1)D_1(p_1, p_2) - F_1 \geq 0$$

and in a two-firm game, firm 2 will solve the analogous problem. If $F_1 = 0$, then the profit constraint is always there but under normal conditions does not constrain the profit-maximizing choice of price so that in informal analyses (e.g., in classrooms) it is usually ignored. However, if $F_1, F_2 > 0$, price undercutting will force the profit constraint to bind for at least one firm in equilibrium. Suppose firm 2's constraint binds first as prices are driven down by price undercutting. Firm 1 will then face a choice between (i) sharing the market (by setting its price equal to that charged by firm 2 when it makes zero profits at equal prices $D_1(p_1, p_2) = D(p_1)/2$ if $p_2 = p_1$) or (ii) slightly undercutting that price which will keep its rival out of the market so that $D_1(p_1, p_2) = D(p_1)$, where $p_1 = p_2 - \varepsilon$, with ε a small increment.[38] Generally, the latter will be more profitable and therefore this version of Bertrand competition with fixed costs results in the prediction that prices will be driven down to levels sufficient to keep the less efficient rival out of the market (see Chowdhury 2002). Slight changes to the game can, however, change this result. For example, a two-stage game with entry involving sinking a fixed cost and then price

[38] There is an easily overcome technical problem arising in this setting because firm 1 would want to be as close to firm 2's price as possible but still smaller than it, which can result in there being no solution to the firm's optimization problem. Technically, the optimization is over the open set $[0, p_2)$ and so need not have a solution. The problem is easily overcome by assuming that price increments occur in small discrete steps, perhaps pennies or cents.

competition will result in only one firm entering and that firm charging a monopoly price. The reason is that if two firms enter, thereby sinking their respective fixed costs, they would compete à la Bertrand at the second stage. That in turn means they would not recover their fixed costs and hence one of the firms will decide it is better not to enter the market. Finally, we note that such situations are also sometimes expected to experience "Edgeworth" cycles, where firms go through a process of undercutting each other until prices are so low that one firm prefers to jump back up to a high price, thereby beginning the cycle again (see Maskin and Tirole 1988b; Noel 2007; Castanias and Johnson 1993; Doyle et al. 2008).

1.3.2.3 Price Competition with Differentiated Products

Models with product differentiation assume that firms' products differ and so are imperfect substitutes for consumers. If so, then each product has a degree of uniqueness and certain consumers may be willing to pay a premium to get each particular product. The differentiation can come due to differences in concrete attributes such as product quality or location or in consumers' subjective perceptions such as those that may result from a brand's image.

Suppose we face a market with two differentiated goods and the following linear demand system:

$$\begin{aligned} &\text{Demand for good 1:} \qquad q_1 = a_1 - b_{11}p_1 + b_{12}p_2, \\ &\text{Demand for good 2:} \qquad q_2 = a_2 - b_{22}p_2 + b_{21}p_1. \end{aligned}$$

First note that good 1 is a substitute for good 2 if an increase in the price of good 2 increases the demand for good 1, which is equivalent to saying that $\partial q_1/\partial p_2 = b_{12} > 0$. Good 1 is a complement for good 2 if an increase in the price of good 2 decreases the demand for good 1 meaning that $\partial q_1/\partial p_2 = b_{12} < 0$.

Assuming firms choose prices, the profit-maximizing firm will choose its best response to the rivals' choices of price. Define the best response function of firm i as[39]

$$R_i(p_{-i}) = \underset{p_i}{\text{argmax}}\, \pi_i(p_i, p_{-i}).$$

If we assume constant marginal costs, the profit function can be expressed as

$$\pi_i(p_i, p_{-i}) = (p_i - c)D_i(p_i, p_{-i}).$$

Differentiating with respect to own price, the first-order condition for profit maximization will be

$$\begin{aligned} \frac{\partial \pi_i(p_i, p_{-i})}{\partial p_i} &= D_i(p_i, p_{-i}) + (p_i - c)\frac{\partial D_i(p_i, p_{-i})}{\partial p_i} = 0, \\ \iff \qquad & (a_i - b_{ii}p_i + b_{ij}p_j) + (p_i - c)(-b_{ii}) = 0, \end{aligned}$$

[39] The notation "argmax" may be new to some readers. It is shorthand for the "argument which maximizes" the function. Here, the price of firm i. The optimal price for firm i will depend on the prices charged by rivals and that dependence is captured in the statement of the reaction function as $p_i = R_i(p_{-i})$.

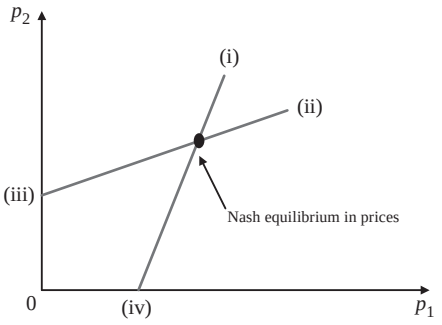

Figure 1.20. Best response curves in price competition with differentiated substitute products. (i) $R_1(p_2) = c/2 + (a_1 + b_{12}p_2)/2b_{11}$; (ii) $R_2(p_1) = c/2 + (a_2 + b_{21}p_1)/2b_{22}$; (iii) $c/2 + a_2/2b_{22}$; (iv) $c/2 + a_1/2b_{11}$.

or, more concisely,

$$a_i + b_{ij}p_j = (2p_i - c)b_{ii}.$$

Rearranging gives the best response function for the producer of product i to a given announcement of price p_j by i's rival firm j:

$$R_i(p_{-i}): \quad p_i = \frac{c}{2} + \frac{a_i + b_{ij}p_j}{2b_{ii}}.$$

Note that the slope of reaction function is $b_{ij}/2b_{ii}$, which, in particular, depends on b_{ij}. In fact, since b_{ii} will be positive, whether the reaction function slopes up or down depends only on the sign of b_{ij}. That in turn means that it depends directly on whether the goods are complements or substitutes.

In a differentiated product price game with demand substitutes ($b_{ij} > 0$), the reaction curves slope up. If firm i increases prices, the best response for firm j is also to increase prices. Graphically, our two-firm example can be represented with each firm pricing according to the best response curves pictured in figure 1.20.

Formally, a generic noncooperative game involves firm i choosing some strategic variable a_i to maximize its profits. The game produces a best reaction function: $a_i^* = R_i(a_{-i}) = \text{argmax}_{a_i}\, \pi_i(a_i, a_{-i})$, where "argmax" means the argument which maximizes the objective function, here the action a_i which maximizes firm i's profits.

Differentiating gives the following equality:

$$\left.\frac{\partial \pi_i(a_i^*, a_{-i})}{\partial a_i}\right|_{a_i^* = R_i(a_{-i})} = \pi_i^i(R_i(a_{-i}), a_{-i}) = 0$$

by definition of the best response function, where the notation "$|_{a_i^* = R_i(a_{-i})}$" denotes that the first-order condition is evaluated at the point where player i is playing a best response to its rival's strategies, a_{-i}. Intuitively, if I am at my optimal choice

of action, say, for example, output, then my marginal profit is zero as required by the optimization process.

Totally differentiating both sides of this equation with respect to another player j's action gives

$$\begin{aligned}\frac{\mathrm{d}\pi_i^i(R(a_{-i}),a_{-i})}{\mathrm{d}a_j} &= \left.\frac{\partial\pi_i^i(a_i,a_{-i})}{\partial a_i}\right|_{a_i^*=R_i(a_{-i})}\frac{\partial R_i(a_{-i})}{\partial a_j}+\left.\frac{\partial\pi_i^i(a_i,a_{-i})}{\partial a_j}\right|_{a_i^*=R_i(a_{-i})}\\ &= 0.\end{aligned}$$

Using double superscripts to indicate double derivatives, this equation can be expressed as

$$\frac{\mathrm{d}\pi_i^i(R_i(a_{-i}),a_{-i})}{\mathrm{d}a_j}=\pi_i^{ii}(r_i(a_{-i}),a_{-i})\frac{\partial R_i(a_{-i})}{\partial a_j}+\pi_i^{ij}(R_i(a_{-i}),a_{-i})=0,$$

which in turn can be rearranged to provide an expression for the slope of the reaction curve:

$$\frac{\partial R_i(a_{-i})}{\partial a_j}=\frac{-\pi_i^{ij}(R_i(a_{-i}),a_{-i})}{\pi_i^{ii}(R_i(a_{-i}),a_{-i})}.$$

(Alternatively, we could obtain this expression directly by applying the implicit function theorem to the first-order condition which implicitly defines firm i's reaction function. See your favorite mathematics or economics textbook, e.g., Mas-Colell et al. (1995, pp. 940–43).) The reaction curve describes the action that maximizes firm i's profits given its competitors' choices. Thus, the second-order condition requires that the second own derivative is negative at the profit-maximizing choice of action a_i, $R(a_{-i})$. That is, $\pi_i^{ii}(R_i(a_{-i}),a_{-i})<0$.

Thus this result says that the sign of the slope of the reaction function will then depend on the cross derivative of the firm profit function $\pi_i^{ij}(a_i,a_{-i})$ evaluated at the point $(R_i(a_{-i}),a_{-i})$. Intuitively, we said that at an optimum the marginal profitability of a firm given your action is zero. Now suppose a rival's action a_j goes up. We consider what happens to my optimal choice of action. Clearly, if $\pi_i^{ij}(R_i(a_{-i}),a_{-i})<0$ then my (firm i's) marginal profitability is falling in your action. That means, when you increased a_j, my marginal profitability fell below zero. The question of i's best response to the new a_j is the question of how to restore my marginal profitability back up to zero, i.e., how do I increase my marginal profitability. If $\pi_i^{ii}(a_i,a_{-i})<0$, then we know that decreasing my action a_i will increase my marginal profitability. In summary, when you increased a_j, then I optimally decreased my action a_i. Thus, if $\pi_i^{ij}(R_i(a_{-i}),a_{-i})<0$, my best response will be decreasing in your choice of action and my reaction function will be downward sloping. Analogously, if $\pi_i^{ij}(R_i(a_{-i}),a_{-i})>0$, then my best response will

be increasing in your choice of action and my reaction function will be upward sloping.

As an example, we showed that in the model the first-order conditions are

$$\pi_i^i(p_i, p_{-i}) = D_i(p_i, p_{-i}) + (p_i - c)D_i^i(p_i, p_{-i})$$

so that the cross derivative is

$$\pi_i^{ji}(p_i, p_{-i}) = D_i^j(p_i, p_{-i}) + (p_i - c)D_i^{ji}(p_i, p_{-i}).$$

With linear demands such as those described at the beginning of this section, the second term is zero $D_i^{ij}(p_i, p_{-i}) = 0$ and hence

$$\pi_i^{ji}(p_i, p_{-i}) = D_i^j(p_i, p_{-i}) = b_{ij}.$$

Whether the reaction functions are upward or downward sloping will depend on the sign of b_{ij}. If b_{ij} is positive so that the goods are substitutes, the reaction function will be upward sloping. If b_{ij} is negative and the goods are complements, the reaction functions will be downward sloping.

If reaction functions are downward sloping, then we will say the game is one of *strategic substitutes*. Returning to the material on the Cournot game, one can easily check that a Cournot game is a game of strategic substitutes, where we write the firm's action, or strategic variable, as quantity q. In Cournot games, competitors will react to a unilateral increase in quantity by decreasing their quantity. In price-setting games, if the goods are demand complements, then reaction curves will also slope downward and the game will also be one of strategic substitutes: firms will react to the increase in the price of a rival's complementary good by decreasing their own price. For this reason, price games among complementary goods will have many properties similar to Cournot-style quantity games.

If reaction functions are upward sloping, then we will say the game is one of *strategic complements*. This is the case in most pricing games such as differentiated products Bertrand pricing games, where the products are demand substitutes. In such cases, firms will react to a rival's unilateral increase in price by increasing their own price(s).

The introduction of product differentiation allows for a model of strategic interaction based on price-setting competition that allows for prices to be above marginal costs. Price competition in a market with differentiated products has become the most generally used model for differentiated product industries. It is, for example, used in particular to model competition in markets for branded consumer goods.

1.3.2.4 Price Competition with Capacity Constraints

One important attempt to reconcile Cournot and Bertrand while making apparently reasonable assumptions on behavior and maintaining consistency with empirically

observed outcomes was formulated in Kreps and Scheinkman (1983). They describe a two-stage game in which firms choose capacity in the first stage and then play a Bertrand competition game in the second stage, given their installed capacity. Kreps and Scheinkman show that, provided customers are allocated to the different producers according to the rule of "efficient rationing" in the second stage, the subgame perfect equilibrium of this two-stage game can be similar to the one-shot Cournot game.

When there are capacity constraints, the total supply can be less than the total demand for a given price. This means we must be concerned with "rationing rules." Rationing rules are assumptions about the way the good is assigned to consumers. It determines (i) who gets the good and who does not, and (ii) which firms supply to which customers. Common rationing assumptions are (i) efficient rationing, where the consumers who value the good most are served first by the lowest-price firm until the firm's capacity is exhausted, and (ii) proportional (random) rationing, where each consumer has an equal probability of being served by any of the existing firms.

With efficient rationing the residual demand of the lowest price firm looks as shown in figure 1.21 since the very highest valuation customers—those at the top-left of the market demand curve—are all served by the lowest price firm. Only when the lowest price firm's capacity is exhausted does the higher price firm begin to experience positive demand for its product.

Suppose firm 1 is the low-cost firm with capacity k_1. Under efficient rationing, the first k_1 units are always bought from firm 1. Firm 2's demand curve is then just a downward-sloping demand curve where at each price firm 2 faces the residual demand, that is, the market demand minus k_1. There is one more wrinkle, that firm 2 cannot sell more than its own capacity k_2. Kreps and Scheinkman show that when the total demand is larger than the sum of capacities in the market, the equilibrium of their two-stage (capacity then price competition) game will correspond to the solution of a one-stage Cournot game where the strategic variable is capacity instead of output produced.[40]

We follow Kreps and Scheinkman to solve for the equilibrium of the two-stage game, we proceed by backward induction, solving stage two first. At stage two, firms are playing a Bertrand price competition game with their capacities k_1 and k_2 for firm 1 and firm 2 respectively fixed. Sales for any firm will be

$$q_i(p_i, p_j; k_i, k_j) = \begin{cases} \min\{D(p_i), k_i\} & \text{if } p_i < p_j, \\ \min\{\max\{D(p_i) - k_j, 0\}, k_i\} & \text{if } p_i > p_j, \\ \min\{(k_i/(k_i + k_j))D(p), k_i\} & \text{if } p_i = p_j. \end{cases}$$

To see why, notice first that the firm gets all the market demand up to its full capacity when it prices below competitors. On the other hand, if a firm prices above

[40] As capacities increase, the nature of the equilibrium changes. In particular, one must use mixed strategies for medium capacities and with very large capacities the equilibrium is a Bertrand equilibrium.

its competitor, it will supply any positive residual demand up to its own capacity. The efficient rationing assumption is thus also embodied in the firm's assumed demand curve in the term $\max\{D(p_i) - k_j, 0\}$. If prices are the same across competitors, we assume that each firm will supply their share of the total capacity available at that price.

At the second stage of our game, the firms take capacities as given and choose their price to maximize their own profits, for each possible price of rivals. The best response function for each firm at stage 2 is therefore

$$\begin{aligned} R_i(p_j; k_i, k_j) &= \operatorname*{argmax}_{p_i} \pi_i(p_i, p_j; k_i, k_j) \\ &= \operatorname*{argmax}_{p_i} (p_i - c) q_i(p_i, p_j; k_i, k_j). \end{aligned}$$

There are two possible scenarios. If capacities are large, so large that capacities are not an effective constraint on sales, then the sales of each firm are

$$q_i(p_i, p_j; k_i, k_j) = \begin{cases} D(p_i) & \text{if } p_i < p_j, \\ 0 & \text{if } p_i > p_j, \\ (k_i/(k_i + k_j))D(p) & \text{if } p_i = p_j. \end{cases}$$

In this case, the firm's demand curve is exactly the one obtained in a Bertrand game with homogeneous products, except for the minor difference in the "splitting rule" when prices are equal. As a result, in this case the equilibrium of the subgame will involve setting price equal to marginal costs, $p^* = mc$. Since this case is less interesting than that of small capacities we focus on that case.

If capacities are small so that capacity constraints are binding, we have

$$0 \leqslant k_i \leqslant D(p_i) - k_j \leqslant D(p_i).$$

The first inequality follows since capacities are positive. The second illustrates that capacity constraints are binding and then capacity is smaller than residual demand at the current price while the latter inequality follows simply since k_j is positive. Rearranging the middle inequality gives that total capacity is no larger than demand:

$$k_i + k_j \leqslant D(p) \quad \Longleftrightarrow \quad k_i \leqslant \left(\frac{k_i}{k_i + k_j}\right) D(p)$$

and in that case sales will be

$$q_i(p_i, p_j; k_i, k_j) = \begin{cases} k_i & \text{if } p_i < p_j, \\ \min\{k_i, D(p_i) - k_j\} = k_i & \text{if } p_i > p_j, \\ k_i & \text{if } p_i = p_j. \end{cases}$$

Now assuming that equilibrium price adjusts to equate total industry capacity to market demand, we have that $k_i + k_j = D(p^*)$ or, inverting the demand equation, $p^* = P(k_i + k_j)$.

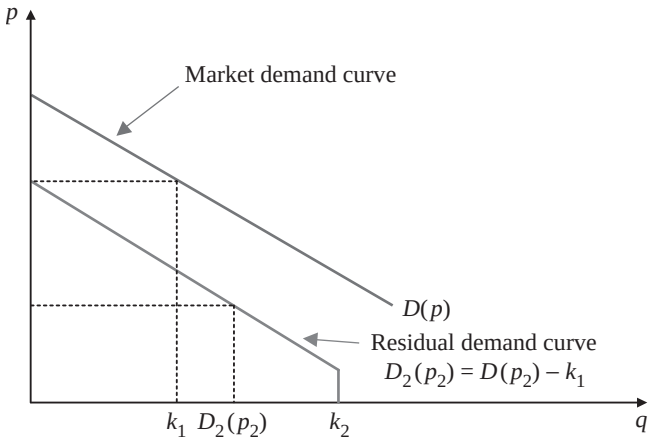

Figure 1.21. Residual demand curve with efficient rationing.

If so, then to solve for the equilibrium of the game in stage one, we need to substitute the optimal prices p^* into the reaction function of the capacity setting game. That is, each firm solves

$$\begin{aligned} R_i(k_j) &= \underset{k_i}{\operatorname{argmax}}\, \pi_i(p_i^*, p_j^*; k_i, k_j) \\ &= \underset{k_i}{\operatorname{argmax}}(p_i^* - c)q(p_i^*, p_j^*; k_i, k_j) \\ &= \underset{k_i}{\operatorname{argmax}}(P(k_i + k_j) - c)k_i. \end{aligned}$$

Clearly, since the objective function is the same as that used in the Cournot model, with qs replaced by ks, the reaction functions derived for the subgame perfect equilibrium of the two-stage game look exactly like the one-shot Cournot game profit function with the choice variable being capacity k instead of output q, and with the inverse demand function $P(k_i + k_j)$.

Deneckere and Davidson (1986) show that the Kreps and Scheinkman result is sensitive to the exact rationing rule used (see figure 1.21). They argue first that the "efficient rationing" is not very likely since under that rule the most highly valued units must be bought from the low-price firm. Second, they note that if, for example, consumers are randomly distributed between the two firms, then the Kreps and Scheinkman result disappears. On reflection, perhaps the fact that this result is sensitive is not really terribly surprising: Kreps and Scheinkman are trying to 'compress' a two-stage game into a simpler and yet equivalent one-stage game—clearly an endeavor which is, at least in general, only going to work under strong and restrictive assumptions.

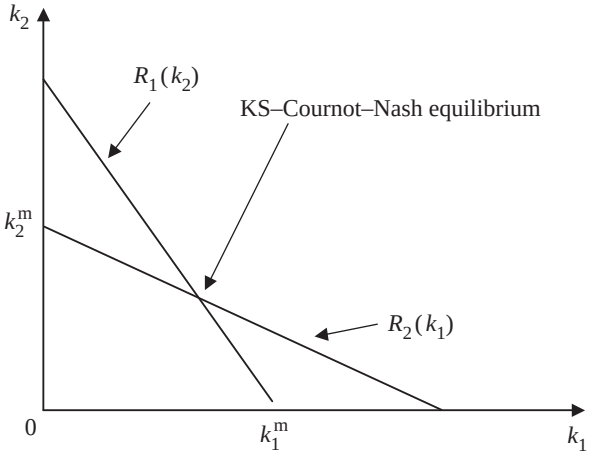

Figure 1.22. Reaction functions in Kreps and Scheinkman two-stage game.

1.3.3 The Monopoly and Dominant-Firm Models

In this section we first briefly revisit the monopoly model and then discuss a variant of that model in which a dominant firm faces a competitive fringe which acts nonstrategically.

1.3.3.1 Monopoly Models

The clearest "dominant" firm model is one in which a firm is a monopoly. Our baseline model of such a situation is that a monopolist will simply maximize profits in a way that is unconstrained by rivals. However, a monopolist may be a price-setting monopolist, a quantity-setting monopolist, a multiplant quantity-setting monopolist or a multiproduct quantity-setting monopolist or indeed a multiplant, multiproduct, price- or quantity-setting monopolist. Thus there is no single model of a monopoly. In order of complexity static monopoly models of the firm assume that they solve problems including:

1. Price-setting monopolist: $\max_p (p - c)D(p)$.
2. Quantity-setting monopolist: $\max_q (P(q) - c)q$.
3. Multiplant, quantity-setting monopolist:

$$\max_{q_1,\ldots,q_J} \sum_{j=1}^{J} (P(q_1 + q_2 + \cdots + q_J) - c_j(q_j))q_j.$$

4. Multiproduct, price-setting monopolist:

$$\max_{p_1,\ldots,p_J} \sum_{j=1}^{J} (p_j - c_j)D_j(p_1, \ldots, p_J).$$

5. Multiproduct, multiplant, price-setting monopolist:

$$\max_{p_1,\ldots,p_J} \sum_{j=1}^{J} (p_j - c_j(D_j(p_1,\ldots,p_J)))D_j(p_1,\ldots,p_J).$$

6. Multiproduct, multiplant, quantity-setting monopolist:

$$\max_{q_1,\ldots,q_J} \sum_{j=1}^{J} (P_j(q_1,\ldots,q_J) - c_j(q_j))q_j.$$

Single-product monopolists will act to set marginal revenue equal to marginal cost. In those cases, since the monopoly problem is a single-agent problem in a single product's price or quantity, our analysis can progress in a relatively straightforward manner. In particular, note that single-agent, single-product problems give us a single equation (first-order condition) to solve. In contrast, even a single agent's optimization problem in the more complex multiplant or multiproduct settings generates an optimization problem is multidimensional. In such single-agent problems, we will have as many equations to solve as we have choice variables. In simple cases we can solve these problems analytically, while, more generally, for any given demand and cost specification the monopoly problem is typically relatively straightforward to solve on a computer using optimization routines.

Naturally, in general, monopolies may choose strategic variables other than price and quantity. For example, if a single-product monopolist chooses both price and advertising levels, it solves the problem $\max_{p,a}(p-c)D(p,a)$, which yields the usual first-order condition with respect to prices,

$$\frac{p-c}{p} = -\left(\frac{\partial \ln D(p,a)}{\partial \ln p}\right)^{-1},$$

and a second one with respect to advertising,

$$(p-c)\frac{\partial D(p,a)}{\partial a} = 0.$$

A little algebra gives

$$\frac{p-c}{p}\,p\,\frac{D(p,a)}{a}\,\frac{\partial \ln D(p,a)}{\partial \ln a} = 0$$

and substituting in for $(p-c)/p$ using the first-order condition for prices gives the result:

$$\frac{a}{pD(p,a)} = \left(\frac{\partial \ln D(p,a)}{\partial \ln a}\right)\Big/\left(-\frac{\partial \ln D(p,a)}{\partial \ln p}\right),$$

which states the famous Dorfman and Steiner (1954) result that advertising–sales ratios should equal the ratios of the own-advertising elasticity of demand to the own-price elasticity of demand.[41]

[41] For an empirical application, see Ward (1975).

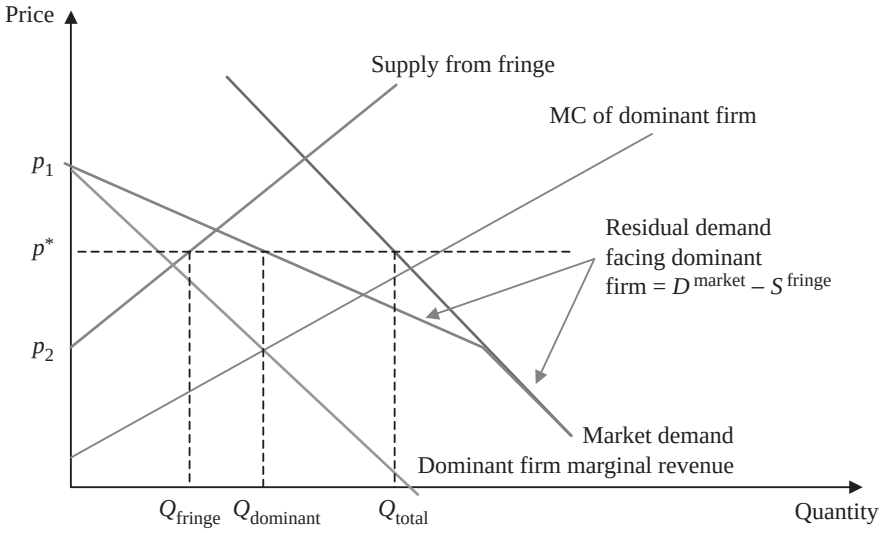

Figure 1.23. Deriving the residual demand curve.

1.3.3.2 The Dominant-Firm Model

The dominant-firm model supposes that there is a monopoly (or collection of firms acting as a cartel) which is nonetheless constrained to some extent by a competitive fringe. The central assumption of the model is that the fringe acts in a nonstrategic manner. We follow convention and develop the model within the context of a price-setting, single-product monopoly. Dominant-firm models analogous to each of the cases studied above are similarly easily developed.

If firms which are part of the competitive fringe act as price-takers, they will decide how much to supply at any given price p. We will denote the supply from the fringe at any given price p as $S^{\text{fringe}}(p)$. Because of the supply behavior of the fringe, if they are able to supply whomever they so desire at any given price p, the dominant firm will face the residual demand curve:

$$D^{\text{dominant}}(p) = D^{\text{market}}(p) - S^{\text{fringe}}(p).$$

Figure 1.23 illustrates the market demand, fringe supply, and resulting dominant-firm demand curve. We have drawn the figure under the assumption that (i) there is a sufficiently high price p_1 such that the fringe is willing to supply the whole market demand at that price leaving zero residual demand for the dominant firm and (ii) there is analogously a sufficiently low price p_2 below which the fringe is entirely unwilling to supply.

Given the dominant firm's residual demand curve, analysis of the dominant-firm model becomes entirely analogous to a monopoly model where the monopolist faces the residual demand curve, $D^{\text{dominant}}(p)$. Thus our dominant firm will set prices so

that the quantity supplied will equate the marginal revenue to its marginal cost of supply. That level of output is denoted Q_{dominant} in figure 1.23. The resulting price will be p^* and fringe supply at that price is $S^{\text{fringe}}(p^*) = Q_{\text{fringe}}$ so that total supply (and total demand) are

$$Q_{\text{total}} = Q_{\text{dominant}} + Q_{\text{fringe}} = S^{\text{fringe}}(p^*) + D^{\text{dominant}}(p^*) = D^{\text{market}}(p^*).$$

A little algebra gives us a useful expression for understanding the role of the fringe in this model. Specifically, the dominant firm's own-price elasticity of demand can be written as[42]

$$\begin{aligned}\eta^{\text{dominant}}_{\text{demand}} &\equiv \frac{\partial \ln D^{\text{dominant}}}{\partial \ln p} \\ &= \frac{\partial \ln(D^{\text{market}} - S^{\text{fringe}})}{\partial \ln p} \\ &= \frac{1}{D^{\text{market}} - S^{\text{fringe}}} \frac{\partial (D^{\text{market}} - S^{\text{fringe}})}{\partial \ln p}\end{aligned}$$

so that we can write

$$\eta^{\text{dominant}}_{\text{demand}} = \frac{1}{D^{\text{market}} - S^{\text{fringe}}} \left[\left(\frac{D^{\text{market}}}{D^{\text{market}}} \right) \frac{\partial D^{\text{market}}}{\partial \ln p} - \left(\frac{S^{\text{fringe}}}{S^{\text{fringe}}} \right) \frac{\partial S^{\text{fringe}}}{\partial \ln p} \right]$$

and hence after a little more algebra we have

$$\begin{aligned}\eta^{\text{dominant}}_{\text{demand}} &= \left(\frac{D^{\text{market}}}{D^{\text{market}} - S^{\text{fringe}}} \right) \frac{\partial \ln D^{\text{market}}}{\partial \ln p} \\ &\quad - \left(\frac{S^{\text{fringe}}/D^{\text{market}}}{(D^{\text{market}} - S^{\text{fringe}})/D^{\text{market}}} \right) \frac{\partial \ln S^{\text{fringe}}}{\partial \ln p} \\ &= \frac{1}{\text{Share}^{\text{dom}}} \eta^{\text{market}}_{\text{demand}} - \left(\frac{\text{Share}^{\text{fringe}}}{\text{Share}^{\text{dom}}} \right) \eta^{\text{fringe}}_{\text{supply}},\end{aligned}$$

where η indicates a price elasticity. That is, the dominant firm's demand curve—the residual demand curve—depends on (i) the market elasticity of demand, (ii) the fringe elasticity of supply, and also (iii) the market shares of the dominant firm and the fringe. Remembering that demand elasticities are negative and supply elasticities positive, this formula suggests intuitively that the dominant firm will therefore face a relatively elastic demand curve when market demand is elastic or when market demand is inelastic but the supply elasticity of the competitive fringe is large and the fringe is of significant size.

[42] Recall from your favorite mathematics textbook that for any suitably differentiable function $f(x)$ we can write

$$\frac{\partial \ln f(x)}{\partial \ln x} = \frac{1}{f(x)} \frac{\partial f(x)}{\partial \ln x}.$$

1.4 Conclusions

- Empirical analysis is best founded on economic theory. Doing so requires a good understanding of each of the determinants of market outcomes: the nature of demand, technological determinants of production and costs, regulations, and firm's objectives.
- Demand functions are important in empirical analysis in antitrust. The elasticity of demand will be an important determinant of the profitability of price increases and the implication of those price increases for both consumer and total welfare.
- The nature of technology in an industry, as embodied in production and cost functions, is a second driver of the structure of markets. For example, economies of scale can drive concentration in an industry while economies of scope can encourage firms to produce multiple goods within a single firm. Information about the nature of technology in an industry can be retrieved from input and output data (via production functions) but also from cost, output and input price data (via cost functions) or alternatively data on input choices and input prices (via input demand functions.)
- To model competitive interaction, one must make a behavioral assumption about firms and an assumption about the nature of equilibrium. Generally, we assume firms wish to maximize their own profits, and we assume Nash equilibrium. The equilibrium assumption resolves the tensions otherwise inherent in a collection of firms each pursuing their own objectives. One must also choose the dimension(s) of competition by which we mean defining the variables that firms choose and respond to. Those variables are generally prices or quantity but can also include, for example, quality, advertising, or investment in research and development.
- The two baseline models used in antitrust are quantity- and price-setting models otherwise known as Cournot and (differentiated product) Bertrand models respectively. Quantity-setting competition is normally used to describe industries where firms choose how much of a homogeneous product to produce. Competition where firms set prices in markets with differentiated or branded products is often modeled using the differentiated product Bertrand model. That said, these two models should not be considered as the only models available to fit the facts of an investigation; they are not.
- An environment of perfect competition with price-taking firms produces the most efficient outcome both in terms of consumer welfare and production efficiency. However, such models are typically at best a theoretical abstraction and therefore they should be treated cautiously and certainly should not systematically be used as a benchmark for the level of competition that can realistically be implemented in practice.