Decision trees; attitude towards risk; some applications; the expected value of perfect information (EVPI)

Chapter 1

Decision analysis

Texts

Varian, H.R. *Intermediate Microeconomics.* (New York: W.W. Norton and Co., 2006) seventh edition [ISBN 0393927024]. Chapter 12.

It seems appropriate to start a course on economics for management with decision analysis. Managers make decisions daily regarding selection of suppliers, budgets for research and development, whether to buy a certain component or produce it in-house and so on. Economics is in some sense the science of decision-making. It analyses consumers' decisions on which goods to consume, in which quantities and when, firms' decisions on the allocation of production over several plants, how much to produce and how to select the best technology to produce a given product. The bulk of economic analysis however considers these decision problems in an environment of certainty. That is, all necessary information is available to the agents making decisions. Although this is a justifiable simplification, in reality of course most decisions are made in a climate of (sometimes extreme) uncertainty. For example, a firm may know how many employees to hire to produce a given quantity of output but the decision of whether or how many employees to lay off during a recession involves some estimate of the length of the recession. Oil companies take enormous gambles when they decide to develop a new field. The cost of drilling for oil, especially in deep water, can be over US$1 billion and the payoffs in terms of future oil price are very uncertain. Investment decisions would definitely be very much easier if uncertainty could be eliminated. Imagine what would happen if you could forecast interest rates and exchange rates with 100 per cent accuracy.

Clearly we need to understand how decisions are made when at least some of the important factors influencing the decision are not known for sure. The field of decision analysis offers a framework for studying how these types of decisions are made or should be made. It also provides insight into the cost of uncertainty or, in other words, how much a decision-maker is or should be prepared to pay to reduce or eliminate the uncertainty. To illustrate the concept of value of information, consider the problem of a company bidding for a road maintenance contract. The costs of the project are unknown and the company does not know how low to bid to get the job. An important question to be answered in preparing the bid is whether to gather more information about the nature of the project and/or the competitors' bids. These efforts only pay off if, as a result, better decisions are taken.

Decision analysis is mainly used for situations in which there is one decision-maker whereas **game theory** deals with problems in which there are several decision-makers, each pursuing their own objectives. In decision analysis any form of uncertainty can be modelled including that arising from unknown features of competitors' behaviour (as in the bidding example). However, when decision analysis models are used to solve problems with several decision-makers, the competitors are not modelled as rational agents (i.e. it is not recognised that they are also trying to achieve certain objectives, taking the actions of other players into account). Instead, decision theory takes the view that, as long as probabilities can be attached to other decision-makers' actions, optimal decisions can be calculated. An obvious objection to this approach is that it is not clear how these probabilities become known to the decision-maker. Game theory avoids this problem as it takes a symmetric, simultaneous view. The reason decision analysis is used in these situations despite these shortcomings is that it is much simpler than game theory. For this reason and because some of the techniques of decision analysis (such as representing sequential decision problems on graphs or decision trees, and solving them backwards) can be used in game theory, we study decision analysis first.

Decision trees

A **decision tree** is a convenient representation of a decision problem. It contains all the ingredients of the problem:

- the decisions
- the sources of uncertainty
- the payoffs which are the results, in terms of the decision-maker's objective, for each possible combination of probabilistic outcomes and decisions.

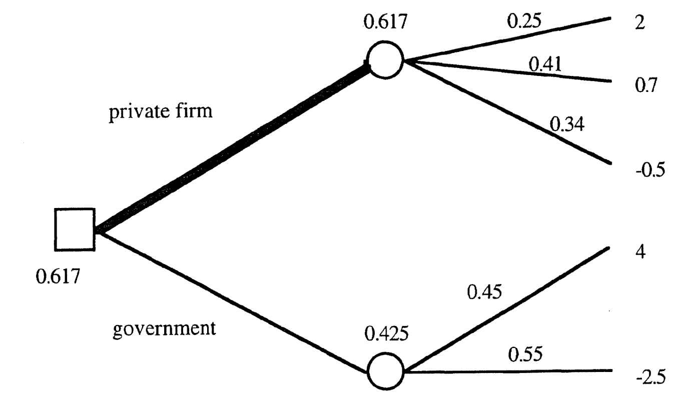
Drawing a decision tree forces the decision-maker to think through the structure of the problem s/he faces and often makes the process of determining optimal decisions easier. A decision tree consists of two kinds of nodes: decision or action nodes which are drawn as squares and probability or chance nodes drawn as circles. The arcs leading from a decision node represent the choices available to the decision-maker at this point whereas the arcs leading from a probability node correspond to the set of possible outcomes when some uncertainty is resolved. When the structure of the decision problem is captured in a decision tree, the payoffs are written at the end of the final branches and (conditional) probabilities are written next to each arc leading from a probability node. The algorithm for finding the optimal decisions is not difficult. Starting at the end of the tree, work backwards and label nodes as follows. At a probability node calculate the expected value of the labels of its successor nodes, using the probabilities given on the arcs leading from the node. This expected value becomes the label for the probability node. At a decision node x (assuming a maximisation problem), select the maximum value of the labels of successor nodes. This maximum becomes the label for the decision node. The decision which generates this maximum value is the optimal decision at this node. Repeat this procedure until you reach the starting node. The label you get at the starting node is the expected payoff obtained when the optimal decisions are taken. The construction and solution of a decision tree is most easily explained through examples.

Example 1.1

Cussoft Ltd., a firm which supplies customised software, must decide between two mutually exclusive contracts, one for the government and the other for a private firm. It is hard to estimate the costs Cussoft will incur under either contract but, from experience, it estimates that, if it contracts with a private firm, its profit will be £2 million, £0.7 million, or -£0.5 million with probabilities 0.25, 0.41 and 0.34 respectively. If it contracts with the government, its profit will be £4 million or -£2.5 million with respective probabilities 0.45 and 0.55. Which contract offers the greater expected profit?

In this very simple example, Cussoft has a choice of two decisions — to contract with the private firm or to contract with the government. In either case its payoff is uncertain. The decision tree with the payoffs and probabilities is drawn in Figure 1.1. The expected profit if the contract with the private firm is chosen equals (0.25)(2) + (0.41)(0.7) + (0.34)(-0.5) = 0.617 (£ million) whereas the contract with the government delivers an expected profit of (0.45)(4)+(0.55)(-2.5)=0.425 (£ million) so that the optimal decision is to go for the contract with the private firm. Optimal decisions are indicated by thick lines.

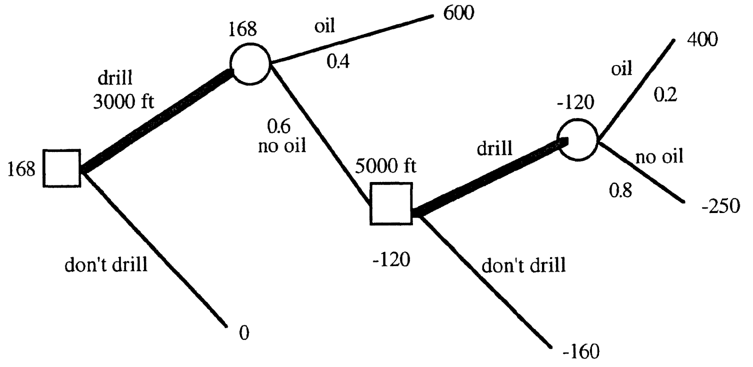
Figure 1.1: Decision tree for example 1.1



Example 1.2

Suppose the Chief Executive of an oil company must decide whether to drill a site and, if so, how deep. It costs £160,000 to drill the first 3,000 feet and there is a 0.4 chance of striking oil. If oil is struck, the profit (net of drilling expenses) is £600,000. If she doesn't strike oil, the executive can drill 2,000 feet deeper at an additional cost of £90,000. Her chance of finding oil between 3,000 and 5,000 feet is 0.2 and her net profit (after all drilling costs) from a strike at this depth is £400,000. What action should the executive take to maximise her expected profit? Try writing down and solving the decision tree yourself without peeking! You should get the following result.

Figure 1.2: Decision tree for example 1.2



Attitude towards risk

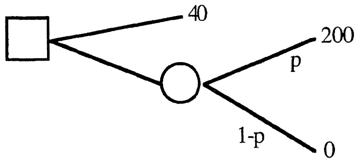
In the examples considered so far we have used the **expected monetary value (EMV) criterion** (i.e. we assumed that the decision-maker is interested in maximising the expected value of profits or minimising the expected value of costs). In many circumstances this is a reasonable assumption to make, especially if the decision-maker is a large company. To appreciate that it may not always be appropriate to use EMV consider the following story, known as the **St. Petersburg paradox.** I will toss a coin and, if it comes up heads, you will get £2. If it comes up tails, I will toss it again and, if it comes up heads this time, you will get £4; if it comes up tails, I will toss it again and, this time, you will get £8 if it comes up heads etc. How much would you be willing to pay for this gamble? I predict that you would not want to pay your week's pocket money or salary to play this game. However, if you calculate the EMV you find:

$$EMV=2(1/2)+4(1/4)+8(1/8)+...+2^n(1/2^n)+...=1+1+1+...= \infty!$$

Even when faced with potentially large gains, most people do not like to risk a substantial fraction of their financial resources. Although this implies that we cannot always use EMV, it is still possible to give a general analysis of how people make decisions even if they do not like taking risks. As a first step we have to find out the decision-maker's attitude towards risk. A useful concept here is the **certainty equivalent (CE)** of a risky prospect defined as the amount of money which makes the individual indifferent between it and the risky prospect. To clarify this, imagine you are offered a lottery ticket which has a 50-50 chance of winning £0 or £200. Would you prefer £100 for sure to the lottery ticket? What about £50 for sure? The amount £x so that you are indifferent between x and the lottery ticket is your certainty equivalent of the lottery ticket. If your x is less than £100, the EMV of the lottery, you are 'risk averse'. In general a decision-maker is **risk averse** if CE<EMV, **risk neutral** if CE=EMV and **risk loving** if CE>EMV. Suppose you have an opportunity to invest £1000 in a business venture which will gross £1100 or £1200 with equal probability next year. Alternatively you could deposit the £1000 in bank which will give you a riskless return. How large does the interest rate have to be for you to be indifferent between the business venture and the deposit account (i.e. what is your certainty equivalent? Are you a risk lover?)? Note that it is possible to be a risk lover for some lotteries and a risk hater for others.

By asking these types of questions, we can determine a decision-maker's degree of risk aversion summarised in his/her **utility of money function**. This enables us to still use expected value calculations but with monetary outcomes replaced by utility values (i.e we can use the **expected utility criterion**). It is possible to show that, if a decision-maker satisfies certain relatively plausible axioms, he can be predicted to behave as if he maximises expected utility. Furthermore, since a utility function $U^*(x) = aU(x) + b$, $a>0$, leads to the same choices as U(x) we can arbitrarily fix the utility of the worst outcome w at 0 (U(w)=0) and the utility of the best outcome b at 1(U(b)=1) for a given decision problem. To find the utility corresponding to an outcome x we ask the decision-maker for the value of p, the probability of winning b in a lottery with prizes b and w, which makes x the CE for the lottery. For example, if the worst outcome in a decision problem is £0(U(0)=0) and the best outcome is £200(U(200)=1), how do we determine U(40)? We offer the decision-maker the choice represented in Figure 1.3 and keep varying p until he is indifferent between 40 and the lottery.

Figure 1.3: Determining utility



When the decision-maker is indifferent, say for p=0.4, we have:

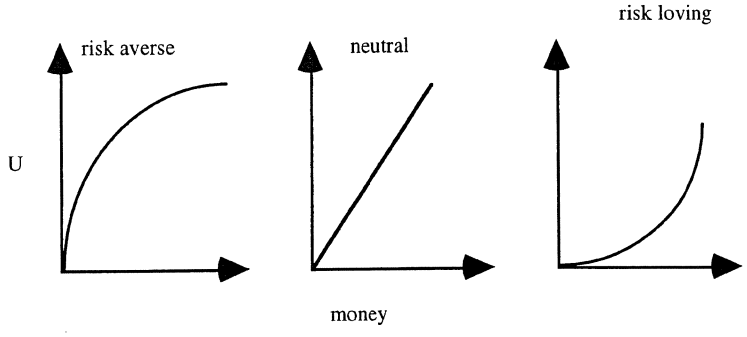
U(40)=U(lottery)=pU(200)+(1-p)U(0)=p=0.4.

Utility values can be obtained in a similar way for the other possible outcomes of the decision problem. Replacing the monetary values by the utility values and proceeding as before will lead to the expected utility maximising decisions.

The definition of risk aversion can be rephrased in terms of the utility function:

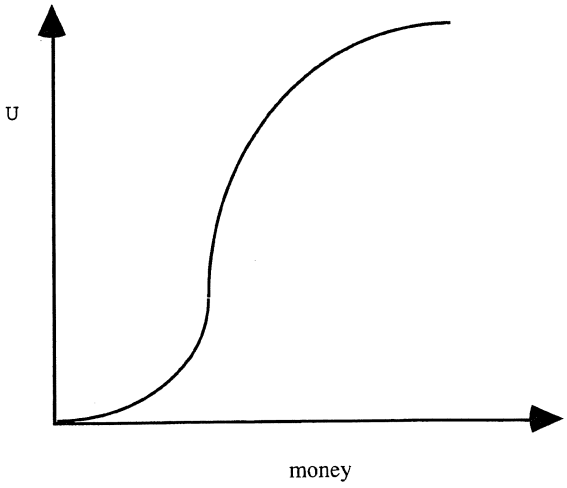
- a **risk averse** decision-maker has a concave utility function
- a **risk lover** has a convex utility function
- a **risk neutral** decision-maker has a linear utility function.

Figure 1.4: Attitudes towards risk



It is possible for a decision-maker to be risk averse over a range of outcomes and risk loving over another range. Indeed, this is how we can explain that the same people who take out home contents insurance buy a national lottery ticket every week. An example of a utility function corresponding to risk loving behaviour for small bets and risk averse behaviour for large bets is drawn in Figure 1.5.

Figure 1.5: Risk loving and risk averse behaviour



For continuous differentiable functions, there are two measures of risk aversion, the Pratt-Arrow **coefficient of absolute risk aversion** determined as $-U''(x)/U'(x)$ and the Pratt-Arrow **coefficient of relative risk aversion** determined as $-U''(x)x/U'(x)$. If U is concave $U''< 0$ and hence both these coefficients are positive for risk averse individuals.[1]

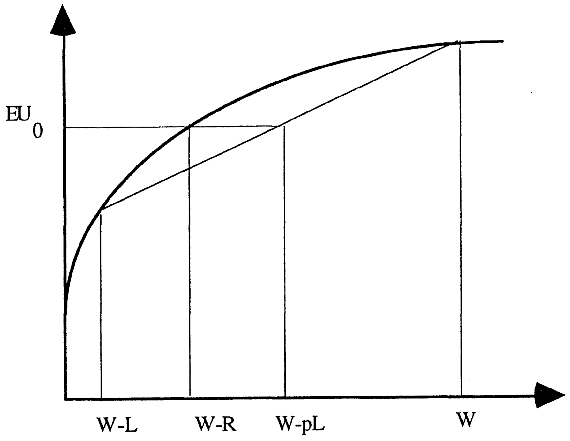
[1] Note that the coefficient of relative risk aversion is the negative of the elasticity of marginal utility of income and does not depend on the units in which income is measured

Some applications

The demand for insurance

People take out insurance policies because they do not like certain types of risk. Let us see how this fits into the expected utility model. Assume an individual has initial wealth W and will suffer a loss L with probability p. How much would she be willing to pay to insure against this loss? Clearly, the maximum premium R she will pay makes her just indifferent between taking out insurance and not taking out insurance. Without insurance she gets expected utility $EU_0 = p\ U(W\text{-}L) + (1\text{-}p)\ U(W)$ and, if she insures at premium R, her utility is U(W-R). Therefore the maximum premium satisfies U(W-R)=pU(W-L)+(1-p)U(W). This is illustrated for a risk averse individual in Figure 1.6. The expected utility without insurance is a convex combination of U(W) and U(W-L) and therefore lies on the straight line between U(W) and U(W-L); the exact position is determined by p so that EU_0 can be read off the graph just above W-pL. This utility level corresponds to a certain prospect W-R which, as can be seen from the figure, has to be less than W-pL, so that R>pL. This shows that, if a risk averse individual is offered actuarially fair insurance (premium R equals expected loss pL), he will insure.

Figure 1.6: Insurance



Example 1.3

Jamie studies at Cambridge University and uses a bicycle to get around. He is worried about having his bike stolen and considers taking out insurance against theft. If the bike gets stolen he would have to replace it which would cost him £200. He finds out that 10 per cent of bicycles in Cambridge are stolen every year. His total savings are £400 and his utility of money function is given by $U(x)=x^{1/2}$. Under what conditions would Jamie take out insurance for a year? What if he has utility of money U(x)=ln(x)?

If he takes out insurance he obtains utility U(400-R) where R is the premium. Without insurance he gets (0.1)U(200)+(0.9)U(400). Equalising these expected utilities and substituting $U(x)=x^{1/2}$, gives:

$$(0.1)\sqrt{200}+(0.9)\sqrt{400}=\sqrt{400-R}$$

or

$$R=23.09$$

which means that insuring is the best decision as long as the premium does not exceed £23.09. Similarly, if U(x)=ln(x), the maximum premium can be calculated (you should check this!) as £26.79.

The demand for financial assets

Consider the problem of an investor with initial wealth W who wants to decide on her investment plans for the coming year. For simplicity, let us assume that there are only two options: a riskless asset which delivers a gross return of R at the end of the year, and a risky asset which delivers a high return H with probability p and a low return L with probability 1-p. It is not difficult to allow for borrowing so that the investor can invest more than W but, to keep things simple, let us restrict the investor's budget to W. The decision problem then consists of finding the optimal amount of money A (<W) to be invested in the risky asset. Given A, the investor gets an expected return of:

$$\begin{aligned} EU(A) &= pU(R(W-A)+HA)+(1-p)U(R(W-A)+LA) \\ &= pU(RW+(H-R)A)+(1-p)U(RW+(L-R)A) \end{aligned}$$

Maximising EU(A) and assuming an interior solution (i.e. 0<A<W) leads to the following (first order) condition:

$$EU'(A) = pU'(RW+(H-R)A)(H-R)+(1-p)U'(RW+(L-R)A)(L-R) = 0$$

Note that, if the risky asset always yields a lower return than the riskless asset (H,L<R), there can be no solution to this condition since $U'>0$. In this scenario the investor would not invest in the risky asset (A=0). Similarly, if the risky asset always yields a higher return than the riskless asset (H,L>R) there can be no solution to this condition and under these circumstances the investor would invest all of her wealth in the risky asset (A=W). For the other scenarios (L<R<H) the first order condition above allows us, for a specific utility function, to calculate the optimal portfolio.

The expected value of perfect information

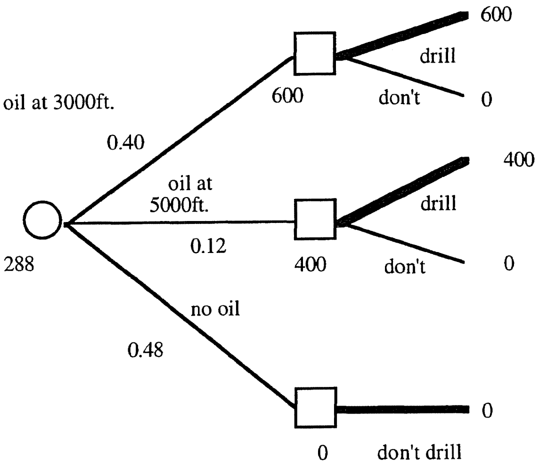
In most situations of uncertainty a decision-maker has the possibility of reducing, if not eliminating, the uncertainty regarding some relevant factor. Typically this process of finding more information is costly in financial terms (e.g. when a market research agency is contracted to provide details of the potential market for a new product) or in terms of effort (e.g. when you test drive several cars before deciding on a purchase). A crucial question in many decision-making contexts is therefore, 'How much money and/ or effort should the decision-maker allocate to reducing the uncertainty?' In this section, we study a procedure which gives an upper bound to this question. The upper bound which is the **expected value of perfect information** (EVPI) is derived as follows for a risk neutral decision-maker (we will go back to the expected utility model later).

Assuming you know the outcomes of all probabilistic events in a decision tree, determine the optimal decisions and corresponding payoff for each possible scenario (combination of outcomes at each probability node). Given that you know the probabilities of each scenario materialising, calculate the expected payoff under certainty using the optimal payoff under each scenario and the scenario's probability. This expected payoff is precisely the payoff you would expect if you were given exact information about what will happen at each probability node. The difference between this expected payoff under perfect information and the original optimal payoff is the EVPI. Since, in reality, it is almost never possible to get perfect information and eliminate the uncertainty completely, the EVPI is an upperbound on how much the decision-maker is willing to pay for **any** (imperfect) information. Generally better decisions are made when there is no uncertainty and therefore the EVPI is positive. However, it is possible that having more information does not change the optimal decisions and in those cases the EVPI is zero. While the concept of EVPI is extremely useful, it is really quite abstract and difficult to grasp. So let us look at an example.

Example 1.2 (continued)

The notion of perfect information in this problem translates into the existence of a perfect seismic test which could tell you with certainty whether there is oil at 3000ft., at 5000ft. or not at all. Assuming such a test exists, how much would the Chief Executive Officer (CEO) be willing to pay to know the test result? The tree in Figure 1.7 represents the (easy) decision problem if all uncertainty is resolved before any decisions are made. There are three scenarios: 'oil at 3000ft.', 'oil at 5000ft' and 'no oil' whose probabilities can be derived from the original tree as 0.40, 0.12, and 0.48 respectively. If the CEO is told which scenario will occur, her decision will be straightforward. Given the optimal decision corresponding to each scenario and the probabilities of the various scenarios, the optimal payoff with perfect information is £288,000. Recall that the original problem had an EMV of £168,000 and hence EVPI=£120,000. If a perfect seismic test were available, the CEO would be willing to pay up to £120,000 for it.

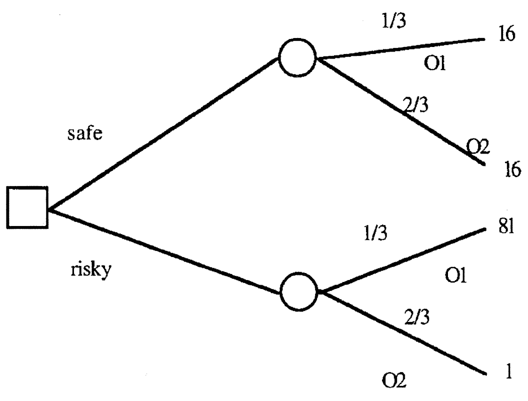
Figure 1.7: Calculating EVPI



If the decision-maker is not risk neutral, a similar method to the one we have just discussed can be used to evaluate the EVPI in utility terms (i.e. we calculate EU under the assumption of perfect information and compare this with the EU in the original problem). However, this does not tell us how much the decision-maker is willing to pay to face the riskless problem rather than the risky one! It is in fact quite tricky to determine the EVPI for an EU maximiser. Consider the simple decision tree in Figure 1.8 where an individual with money utility $U(x) = x^{1/2}$ chooses between a safe and a risky strategy, say investing in a particular stock or not. In either case there are two outcomes — O_1 and O_2 (e.g. the company is targeted for takeover or not) — resulting in the monetary payoffs indicated on the tree. The probabilities of the outcomes are independent of the decision taken. Using the EU criterion, the decision-maker chooses the safe strategy so that EU = 4.

Why?

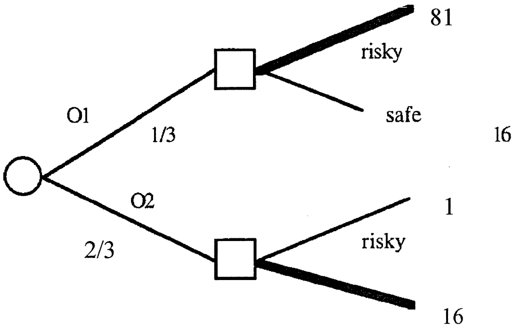
Figure 1.8: EU maximiser chooses safe



With perfect information however, the decision-maker chooses the 'risky' strategy if O_1 is predicted and the 'safe' strategy when 02 is predicted, as is indicated in Figure 1.9. This gives her:

$$EU = (1/3)\,(9) + (2/3)\,(4) = 17/3.$$

Figure 1.9: Optimal choice under perfect information



How much is the decision-maker willing to pay for this increase in EU from 4 to 17/3? Suppose she pays an amount R for the perfect information. She will be indifferent between getting the information and not getting it if the EU in both cases is equal, or $4=(1/3)(81\text{-}R)^{1/2} + (2/3)(16\text{-}R)^{1/2}$. Solving this for R (numerically) gives R approximately 12.5; hence, the EVPI for this problem is about 12.5. Note that this last equation can be solved algebraically.

Chapter summary

By the end of this chapter and the relevant reading, you should understand:

- the concept of **EVPI** and how it can be useful
- why we may want to use **expected utility** rather than expected value maximisation
- the concept of **certainty equivalent** and how it relates to expected value for a risk loving, risk neutral and risk hating decision-maker
- the application of decision analysis in insurance and finance.

You should be able to:

- structure simple decision problems in **decision tree** format and derive optimal decisions
- calculate **risk aversion coefficients**
- calculate **EVPI** for risk neutral and non risk neutral decision-makers.

Sample exercises

1. London Underground (LU) is facing a courtcase by legal firm Snook&Co, representing the family of Mr Addams who was killed in the Kings Cross fire. LU has estimated the damages it will have to pay if the case goes to court as follows: £1,000,000, £600,000 or £0 with probabilities 0.2, 0.5 and 0.3 respectively. Its legal expenses are estimated at £100,000 in addition to these awards. The alternative to allowing the case to go to court is for LU to enter into out-of-court settlement negotiations. It is uncertain about the amount of money Snook&Co. are prepared to settle for. They may only wish to settle for a high amount (£900,000) or they may be willing to settle for a reasonable amount (£400,000). Each scenario is equally likely. If they are willing to settle for £400,000 they will of course accept an offer of £900,000. On the other hand, if they will only settle for £900,000 they will reject an offer of £400,000. LU, if it decides to enter into negotiations, will offer £400,000 or £900,000 to Snook&Co. who will either accept (and waive any future right to sue) or reject and take the case to court. The legal cost of pursuing a settlement whether or not one is reached is £50,000. Determine the strategy which minimises LU's expected total cost.

2. Rickie is considering setting up a business in the field of entertainment at children's parties. He estimates that he would earn a gross revenue of £9,000 or £4,000 with a 50-50 chance. His initial wealth is zero. What is the largest value of the cost which would make him start this business:

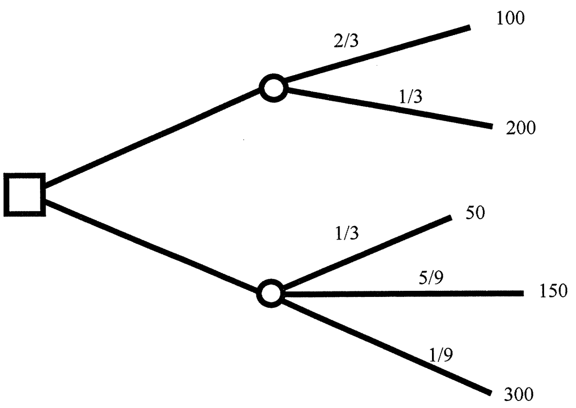
   a. if his utility of money function is $U(x) = ax+b$ where $a>0$

   b. if $U(x) = x^{1/2}$; for $x>0$ and $U(x)= -(-x)^{1/2}$ for $x<0$

   c. if $U(x) = x^2$, for $x>0$ and $U(x)= -x^2$ for $x<0$.

3. Find the coefficient of absolute risk aversion for $U(x) = a-b.\exp(-cx)$ and the coefficient of relative risk aversion for $U(x) = a+b.\ln(x)$.

4. Find a volunteer (preferably someone who doesn't know expected utility theory) and estimate their utility of money function to predict their choice between the two lotteries below. Payoffs are given in monetary value (£). Check your prediction.



5. An expected utility maximiser spends £10 on a lottery ticket, with a chance of 1 in 1 million of £1 million. He takes out home contents insurance at a premium of £100. His probability of an insurance claim of £1000 is 1%. Draw his utility of money function.

6. A decision-maker must choose between (1) a sure payment of £200; (2) a gamble with prizes £0, £200, £450, and £1000 with respective probabilities 0.5, 0.3, 0.1 and 0.1; (3) a gamble with prizes £0, £100, £200, and £520, each with probability 0.25.

   a. Which choice will be made if the decision-maker is risk neutral?
   b. Assume the decision-maker has a CARA (constant absolute risk aversion) utility of money function U(x)= -a exp(-cx) +b and her certainty equivalent for a gamble with prizes £1000 and £0 equally likely is £470. Which choice will be made?[2]

7. Henrika has utility function U= $M^{1/2}$ for $M \geq 0$ and U= - $(-M)^{1/2}$ for $M < 0$, over money payoffs M.
   a. Given a lottery with outcomes £0 and £36 with respective probabilities 2/3 and 1/3, how much is she willing to pay to replace the lottery with its expected value?
   b. Given the table of money payoffs below, which action maximises her expected utility?
   c. How much would Henrika be willing to pay for perfect information regarding the state of nature?

	S_1	S_2
A_1	4	4
A_2	9	1
	1/3	2/3

[2] ***Hint:*** *show that c has to equal 0.00024 approximately for U(0)= 0 and U(1000)=1*

Notes

Extensive form games; normal form games; Nash equilibrium; prisoners' dilemma; perfect equilibrium; perfect Bayesian equilibrium

Chapter 2

Game theory

Texts

Tirole, J. *The Theory of Industrial Organization.* (Cambridge, Mass.: The MIT Press, 1988) [ISBN 0262200716] Chapter 11.

Varian, H.R *Intermediate Microeconomics.* (New York: W.W. Norton and Co., 2006) seventh edition [ISBN 0393927024]. Chapters 28 and 29.

References cited

Axelrod, R. *The evolution of cooperation.* (New York: Basic Books, 1984 [ISBN 0465021212].

Kreps, D. and R. Wilson 'Reputation and imperfect information', *Journal of Economic Theory* (1982) 27: 253–79.

Milgrom, P. and J. Roberts 'Predation, reputation and entry deterrence', *Journal of Economic Theory* (1982) 27: 280–312.

'Mumbo jumbo, super-jumbo', *The Economist*, 12 June 1993, 98.

Nash, J. 'Noncooperative games', *Annals of Mathematics* (1951) 54: 289–95.

'Now for the really big one', *The Economist*, 9 January 1993, 61–62.

'Plane Wars', *The Economist*, I I June 1994, 83.

Selten, R. 'The chain store paradox', *Theory and Decision* (1978), 9: 127–59.

''The flying monopolists', *The Economist*, 19 June 1993, 18.

Game theory extends the theory of individual decision-making to situations of strategic interdependence: that is, situations where players (decision-makers) take other players' behavior into account when making their decisions. The pay-offs resulting from any decision (and possibly random events) are generally dependent on others' actions.

A distinction is made between **cooperative game theory** and **noncooperative game theory**. In cooperative games, coalitions or groups of players are analysed. Players can communicate and make binding agreements. The theory of noncooperative games assumes that no such agreements are possible. Each player in choosing his or her actions, subject to the rules of the game, is motivated by self-interest. Because of the larger scope for application of noncooperative games to managerial economics, we will limit our discussion to noncooperative games.

To model an economic situation as a game involves translating the essential characteristics of the situation into rules of a game. The following must be determined:

- the number of players
- their possible actions at every point in time
- the payoffs for all possible combinations of moves by the players
- the information structure (what do players know when they have to make their decisions?).

All this information can be presented in a game tree which is the game theory equivalent of the decision tree. This way of describing the game is called the **extensive form** representation.

It is often convenient to think of players' behavior in a game in terms of strategies. A strategy tells you what the player will do each time s/he has to make a decision. So, if you know the player's strategy, you can predict his behavior in all possible scenarios with respect to the other players' behavior. When you list or describe the strategies available to each player and attach payoffs to all possible combinations of strategies by the players, the resulting 'summary' of the game is called a **normal form** or **strategic form** representation.

In games of **complete information** all players know the rules of the game. In **incomplete information** games at least one player only has probabilistic information about some elements of the game (e.g. the other players' precise characteristics). An example of the latter category is a game involving an insurer — who only has probabilistic information about the carelessness of an individual who insures his car against theft — and the insured individual who knows how careless he is. A firm is also likely to know more about its own costs than about its competitors' costs. In games of **perfect information** all players know the earlier moves made by themselves and by the other players. In games of **perfect recall** players remember their own moves and do not forget any information which they obtained in the course of game play. They do not necessarily learn about other players' moves.

Game theory, as decision theory, assumes rational decision-makers. This means that players are assumed to make decisions or choose strategies which will give them the highest possible expected payoff (or utility). Each player also knows that other players are rational and that they know that he knows they are rational and so on. In a strategic situation the question arises whether it could not be in an individual player's interest to convince the other players that he is irrational. (This is a complicated issue which we will consider in the later sections of this chapter. All I want to say for now is that ultimately the creation of an impression of irrationality may be a rational decision.)

Before we start our study of game theory, a 'health warning' may be appropriate. It is not realistic to expect that you will be able to use game theory as a technique for solving real problems. Most realistic situations are too complex to analyse from a game theoretical perspective. Furthermore, game theory does not offer any optimal solutions or solution procedures for most practical problems. However, through a study of game theory, insights can be obtained which would be difficult to obtain in another way and game theoretic modelling helps decision-makers think through all aspects of the strategic problems they are facing. As is true of mathematical models in general it allows you to check intuitive answers for logical consistency.

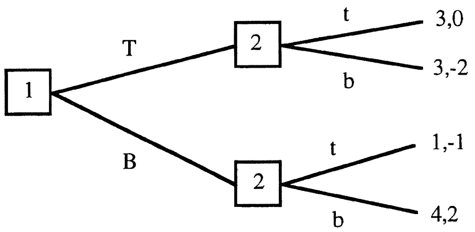
Extensive form games

As mentioned above, the extensive form representation of a game is similar to a decision tree. The order of play and the possible decisions at each decision point for each player are indicated as well as the information structure, the outcomes or payoffs and probabilities. As in decision analysis the payoffs are not always financial. They may reflect the player's utility of reaching a given outcome. A major difference with decision analysis is that in analysing games and in constructing the game tree, the notion of **information set** is important. When there is only one decision-maker, the decision-maker has perfect knowledge of her own earlier choices. In a game, the players often have to make choices not knowing which decisions have been taken or are taken at the same time by the other players. To indicate that a player does not know her position in the tree exactly, the possible locations are grouped or linked in an information set. Since a player should not be able to deduce from the nature or number of alternative choices available to her where she is in the information set, her set of possible actions has to be identical at every node in the information set. For the same reason, if two nodes are in the same information set, the same player has to make a decision at these nodes. In games of perfect information the players know all the moves made at any stage of the game and therefore all information sets consist of single nodes.

Example 2.1 presents the game tree for a dynamic game in which Player 2 can observe the action taken by Player 1. Example 2.2 presents the game tree for a static game in which players take decisions simultaneously.

Example 2.1

Figure 2.1: Game tree for a dynamic game



In this game tree Player 1 makes the first move, Player 2 observes the choice made by Player 1 (perfect information game) and then chooses from his two alternative actions. The payoff pairs are listed at the endpoints of the tree. For example, when Player 1 chooses B and Player 2 chooses t, they receive payoffs of 1 and -1 respectively. Games of perfect information are easy to analyse. As in decision analysis, we can just start at the end of the tree and work backwards (Kuhn's algorithm). When Player 2 is about to move and he is at the top node, he chooses t since this gives him a payoff of 0 rather than -2 corresponding to b. When he is at the bottom node, he gets a payoff of 2 by choosing b. Player 1 knows the game tree and can anticipate these choices of Player 2. He therefore anticipates a payoff of 3 if he chooses T and 4 if he chooses B. We can conclude that Player 1 will take action B and Player 2 will take action b.

Let us use this example to explain what is meant by a strategy. Player 1 has two strategies: T and B. (Remember that a strategy should state what the player will do in each eventuality.) For Player 2 therefore, each strategy consists of a pair of actions, one to take if he ends up at the top node and one to take if he ends up at the bottom node. Player 2 has four possible strategies, namely:

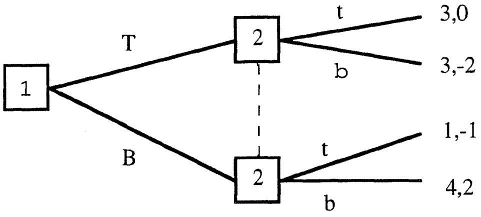
- (t if T, t if B)
- (t if T, b if B)
- (b if T, t if B)
- (b if T, b if B)

or {(t,t), (t,b), (b,t),(b,b)} for short.

Example 2.2

The game tree below is almost the same as in Example 2.1 but here Player 2 does not observe the action taken by Player 1. In other words, it is as if the players have to decide on their actions simultaneously. This can be seen on the game tree by the dashed line linking the two decision nodes of Player 2: Player 2 has an information set consisting of these two nodes. This game (of imperfect information) cannot be solved backwards in the same way as the game of Example 2.1.

Figure 2.2: Game tree for a simultaneous move game



Note that, although the game trees in the two examples are very similar, Player 2 has different strategy sets in the two games. In the second game his strategy set is just {t,b} whereas in the first game there are four possible strategies.

Normal form games

A two-person game in normal form with a finite number of strategies for each player is easy to analyse using a **payoff matrix**. The payoff matrix consists of r rows and c columns where r and c are the number of strategies for the row and the column players respectively. The matrix elements are pairs of payoffs (p_r,p_c) resulting from the row player's strategy r and the column player's strategy c, with the payoff to the row player listed first. The normal form representations of the games in Examples 2.1 and 2.2 are given below.

		Player 2			
		(t,t)	(t,b)	(b,t)	(b,b)
Player 1	T	3,0	3,0	3,-2	3,-2
	B	1,-1	4,2	1,-1	4,2

Normal form for example 2.1

		Player 2	
		t	b
Player 1	T	3,0	3,-2
	B	1,-1	4,2

Normal form for example 2.2

Example 2.3

Two competing firms are considering whether to buy television time to advertise their products during the Olympic Games. If only one of them advertises, the other one loses a significant fraction of its sales. The anticipated net revenues for all strategy combinations are given in the table below. We assume that the firms have to make their decisions simultaneously.

		Firm B	
		Advertise	Don't
Firm A	Advertise	10,5	13,2
	Don't	6,7	11,9

If firm A decides to advertise, it gets a payoff of 10 or 13 depending on whether B advertises or not. When it doesn't advertise it gets a payoff of 6 or 11 depending on whether B advertises or not. So, irrespective of B's decision, A is better off advertising. In other words, 'Don't advertise' is a dominated strategy for firm A. Given that we are assuming that players behave rationally, dominated strategies can be eliminated. Firm B can safely assume that A will advertise. Given this fact, B now only has to consider the top row of the matrix and hence it will also advertise. Note that both A and B are worse off than if they could sign a binding agreement not to advertise.

Example 2.4

In the payoff matrix below, only one payoff, the payoff to the row player, is given for each pair of strategies. This is the convention for zero-sum games (i.e. games for which the payoffs to the players sum to zero for all possible strategy combinations). Hence, the entry 10 in the (1,1) position is interpreted as a gain of 10 to the row player and a loss of 10 to the column player. An application of this type of game is where duopolists compete over market share. Then one firm's gain (increase in market share) is by definition the other's loss (decrease in market share). In zero sum games one player (the row player here) tries to maximise his payoff and the other player (the column player here) tries to minimise the payoff.

If we consider the row player first, we see that the middle row weakly dominates the bottom row. For a strategy A to strictly dominate a strategy B we need the payoffs of A to be strictly larger than those of B against all of the opponent's strategies. For weak dominance it is sufficient that the payoffs are at least as large as those of the weakly dominated strategy. In our case, the payoff of M is not always **strictly** larger than that of B (it is the same if Player 2 plays R). If we eliminate the dominated strategy B, we are left with a 2x3 game in which Player 1 has no dominated strategies. If we now consider Player 2, we see that C is weakly dominated by R (remembering that Player 2's payoffs are the negative of the values in the table!) and hence we can eliminate the second column. In the resulting 2x2 game, T is dominated by M and hence we can predict that (M,R) will be played.

		Player 2		
		L	C	R
Player 1	T	10	20	-30
	M	20	10	10
	B	0	-20	10

In general, we may delete dominated strategies from the payoff matrix and, in the process of deleting one player's dominated strategies, we generate a new payoff matrix which contains dominated strategies for the other player which in turn can be deleted and so on.This process is called **successive elimination of dominated strategies.**

Verify that, in the normal form of Example 2.1, this process leads to the outcome we predicted earlier, namely (B,(t,b)) but that, in the normal form of Example 2.2, there are no dominated strategies.

Sometimes, as in the example above, we will be left with one strategy pair, which would be the predicted outcome of the game but the usual scenario is that only a small fraction of the strategies can be eliminated.

Nash equilibrium

A Nash equilibrium is a combination of strategies, one for each player, with the property that no player would unilaterally want to change his strategy given that the other players play their Nash Equilibrium strategies. So a Nash equilibrium strategy is the best response to the strategies that a player assumes the other players are using.

Pure strategies are the strategies as they are listed in the normal form of a game. We have to distinguish these from **mixed strategies** (which will be referred to later). The game below has 1 Nash equilibrium in pure strategies, namely (T,L). This can be seen as follows. If the row player plays his strategy T, the best the column player can do is to play his strategy L which gives him a payoff of 6 (rather than 2 if he played R). Vice versa, if the column player plays L, the best response of the row player is T. (T,L) is the only pair of strategies which are best responses to each other.

	L	R
T	5,6	1,2
B	4,3	0,4

In some cases we can find Nash equilibria by successive elimination of **strictly** dominated strategies. If weakly dominated strategies are also eliminated we may not find all Nash equilibria. Another danger of successively eliminating **weakly** dominated strategies is that the final normal form (which may contain only one entry) after elimination may depend on the order in which dominated strategies are eliminated.

Example 2.5 'Battle of the sexes'

The story corresponding to this game is that of a husband and wife who enjoy the pleasure of each other's company but have different tastes in leisure activities. The husband likes to watch football whereas the wife prefers a night out on the town. On a given night the couple have to decide whether they will stay in and watch the football match or go out. The payoff matrix could look like this.

		wife	
		in	out
husband	in	10,5	2,4
	out	0,1	4,8

This game has two Nash equilibria: (in,in) and (out,out). Only when both players choose the same strategy is it in neither's interest to switch strategies. The battle of the sexes game is a paradigm for bargaining over common standards.

When electronics manufacturers choose incompatible technologies they are generally worse off than when they can agree on a standard. For example, Japanese, US and European firms were developing their own versions of high definition television whereas they would have received greater payoffs if they had coordinated. The computer industry, in particular in the area of operating system development, has had its share of battles over standards. This type of game clearly has a first mover advantage and, if firms succeed in making early announcements which commit them to a strategy, they will do better.

Example 2.6

This example is typical of market entry battles. Suppose two pharmaceutical companies are deciding on developing a drug for Alzheimer's disease or for osteoporosis. If they end up developing the same drug, they have to share the market and, since development is very costly, they will make a loss. If they develop different drugs they make monopoly profits which will more than cover the development cost. The payoff matrix could then look like this:

	A	O
A	-2,-2	20,10
O	10,20	-1,-1

There are two Nash equilibria in this game: (A,O) and (O,A). Note that there is a 'first mover advantage' in this game. If Firm 1 can announce that it will develop the drug for Alzheimer's it can gain 20 if the announcement is believed (and therefore Firm 2 chooses strategy O). Firms in this situation would find it in their interest to give up flexibility strategically by, for example, signing a contract which commits them to delivery of a certain product. In our scenario a firm could, with a lot of publicity, hire the services of a university research lab famous for research on Alzheimer's disease.

The type of first mover advantage illustrated in this example is prevalent in the development and marketing of new products with large development costs such as wordprocessing or spreadsheet software packages. The firm which can move fastest can design the most commercially viable product in terms of product attributes and the slower firms will then have to take this product definition as given. Other sources of first mover advantage in a new product introduction context include brand loyalty (first mover retains large market share), lower costs than the second mover due to economies of scale and learning curve effects.

Case: Airbus versus Boeing

Europe's Airbus Industrie consortium and Boeing are both capable of developing and manufacturing a large passenger aircraft. The rationale for pursuing such a project is clear. Airports are getting very crowded and, given the high volume of traffic forecast for the next decades, it will become increasingly difficult to find take-off and landing slots; an aircraft carrying, say, 700 or 800 passengers rather than the current maximum of 500 is likely to increase efficiency. The world market has room for only one entrant (predicted sales are about 500) and if both firms start development, they will incur severe losses (to bring a super-jumbo to market could cost US$15 billion). Assume the payoff matrix with strategies 'develop' (D) and 'don't develop' (DD) is similar to the one given below.

A\B	D	DD
D	-3,-3	10,-1
DD	-1,10	0,0

There are two Nash equilibria: one in which Airbus builds the aircraft (and Boeing doesn't) and one in which Boeing builds it (and Airbus doesn't). In this game there is a significant 'first-mover advantage' : if we allow Boeing to make a decision before its rival has a chance to make a decision it will develop the aircraft. (In reality the game is of course more complicated and there are more strategies available to the players. For example, Boeing could decide to make a larger version of its existing 450-seat 747, which would not be very big but could be developed relatively quickly and at lower cost. Or it could decide to collaborate with Airbus.)

The role of government regulation is not clear cut here. On the one hand, governments may want to prevent the inefficient outcome of both firms going ahead with development but, on the other hand, the prospect of monopoly is not attractive either. Particularly if Boeing and Airbus collaborate, the consequences of what would be an effective cartel would be disastrous for struggling airlines. Not only would they have to pay a high price for the super-jumbo but, if they want to buy a smaller aircraft, they would have to turn to Boeing or Airbus who might increase the prices of these smaller aircrafts to promote the super-jumbo.

One way to avoid the problem of both companies starting development is for the EU to announce that it will give a large subsidy to Airbus if it develops the aircraft, regardless of whether Boeing also develops it. Then 'develop' may become a dominant strategy for Airbus and one Nash equilibrium would be eliminated. (To see this add 5 to A's payoff if A chooses strategy D.)

The United States has persistently complained about Airbus subsidies and, in 1992, an agreement limiting further financial support was reached. As of July 1993 both firms had plans to go ahead independently with development of a large aircraft despite discussing a partnership to build a super-jumbo jet (the VLCT or very large civil transport project). In June 1994 Airbus unveiled a design of the A3xx-100, a double decker super-jumbo which would cost US$8 billion to develop.[1]

[1] *Case based on 'Now for the really big one'; 'Mumbo jumbo, super jumbo'; 'The flying monopolists'; 'Plane Wars'*

Example 2.7

In the payoff matrix below there is no Nash equilibrium 'in pure strategies' (i.e. none of the pairs (T,L),(T,R),(B,L) or (B,R) are stable outcomes). Consider, for example, (B,L). If the row player picks strategy B then the best response of the column player is L but, against L, the best response of the row player is T, not B. A similar analysis applies to the other strategy pairs.

	L	R
T	10,5	2,10
B	8,4	4,2

Nash (1951) showed that games in which each player has a finite number of strategies always have an equilibrium. However, players may have to use **mixed strategies** at the equilibrium. A mixed strategy is a rule which attaches a probability to each pure strategy. To see why it makes sense to use mixed strategies think of the game of poker. The strategies are whether to bluff or not. Clearly, players who always bluff and players who never bluff will do worse than a player who sometimes bluffs. Players using mixed strategies are less predictable and leaving your opponent guessing may pay off. To see how to find a Nash equilibrium in mixed strategies for a two-player game in which each of the players has two pure strategies, consider the payoff matrix of Example 2.7 again.

Example 2.7 (con'd)

Suppose the row player uses a mixed strategy (x,1-x) (i.e. he plays strategy T with probability x and B with probability 1-x) and the column player uses a mixed strategy (y,1-y) (i.e. he plays strategy L with probability y and R with probability 1-y). Then the expected payoffs to the row and column player are respectively:

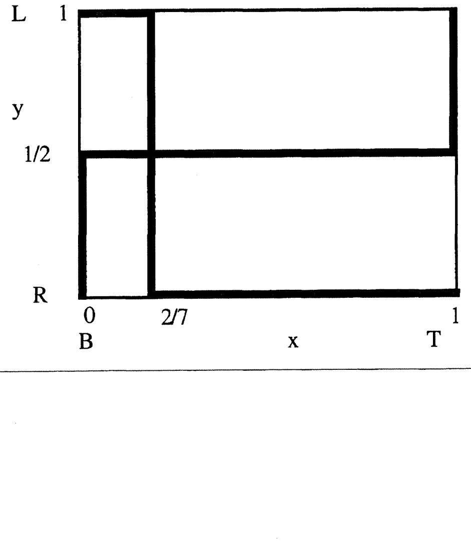
$$\pi_r = 10xy + 2x(1\text{-}y)+8(1\text{-}x)y+4(1\text{-}x)(1\text{-}y) \text{ and}$$

and

$$\pi_c = 5xy + 10x(1\text{-}y)+4(1\text{-}x)y+2(1\text{-}x)(1\text{-}y).$$

The row player chooses x so as to maximise her expected payoff and the column player chooses y so as to maximise his expected payoff. Given y, the expected payoff to the row player is increasing in x as long as y>1/2 and decreasing in x for y<1/2 (check this by differentiating π_r with respect to x) and therefore the best response to y is x = 0 for y<1/2, x=1 for y>1/2 and any x is optimal against y=1/2. Following a similar derivation for the column player we find that his best response to a given x is to set y=0 for x>2/7, y=1 for x<2/7 and any y is optimal against x=2/7. These best response functions are pictured in bold in Figure 2.3.

Figure 2.3: Mixed strategy Nash equilibrium



At a Nash equilibrium the strategies have to be best responses to each other. Therefore, the point (x=2/7, y=1/2), where the response functions intersect, forms a Nash equilibrium. We can then calculate the players' expected payoffs by substituting x and y in the expressions for π_r and π_c above.

The notion of Nash equilibrium is very helpful in a negative sense: any combination of strategies which does not form a Nash equilibrium is inherently unstable. However, when there is more than one Nash equilibrium, game theory does not offer a prediction as to which Nash equilibrium will be played. A topic of current research is to try to find justifications for selecting particular types of Nash equilibria, say the equilibria which are not Pareto dominated, over others. This type of research could help eliminate some Nash equilibria when there are multiple equilibria. Nevertheless game theorists have not succeeeded in agreeing on an algorithm which will select **the** equilibrium that is or should be played in reality.

Prisoners' dilemma

A class of two-person noncooperative games which has received much attention not only in economics but in social science in general, is the class of prisoners' dilemma games. The story the game is meant to model concerns two prisoners who are questioned separately, without the possibility of communicating, about their involvement in a crime. They are offered the following deal. If one prisoner confesses and the other does not, the confessor goes free and the other prisoner serves 10 years; if both confess, they each spend seven years in prison; if neither confesses, they each serve a two-year term. This is summarised in the payoff matrix below.

	confess	don't
confess	7,7	0,10
don't	10,0	2,2

This game is very easy to analyse: for both players the strategy 'confess' dominates. (Remember that you want to minimise the payoff here!) There is one Nash equilibrium in which both prisoners serve seven-year sentences. What is interesting about this game is that, if the prisoners could set up a binding agreement, they would agree not to confess and serve two years. (This type of model is used to explain difficulties encountered in arms control for example.)

The typical application of the prisoners' dilemma to managerial economics translates the prisoners' plight into the situation of duopolists deciding on their pricing policy. If both set a high price they achieve high profits; if both set a low price they achieve low profits; if one firm sets a low price and its rival sets a high price the discounter captures the whole market and makes very high profits whereas the expensive seller makes a loss. At the Nash equilibrium both firms set low prices.

Of course in reality firms do not interact only once but they interact in the market over many years and the question arises whether collusive behavior could be rationalised in a repeated prisoners' dilemma game. When the game is played over several periods rather than as a one-shot game, players might be able to cooperate (set a high price) as long as their rival is willing to cooperate and punish when the rival cheats (deviates from cooperation). This possibility of punishment should give players more of an incentive to cooperate in the long-term. Axelrod (1984) ran a contest in which he asked game theorists to submit a strategy for the repeated version of the prisoners' dilemma. He then paired the given strategies (some of which were very complicated and required significant computer programming) and ran a tournament. The 'tit-for-tat' strategy, which consistently outperformed most of the others, is very simple. This strategy prescribes cooperation as long as the other player cooperates but deviation as soon as and as long as the other player deviates from cooperation. It never initiates cheating and it is forgiving in that it only punishes for one period. If two players use the tit-for-tat strategy they will always cooperate.

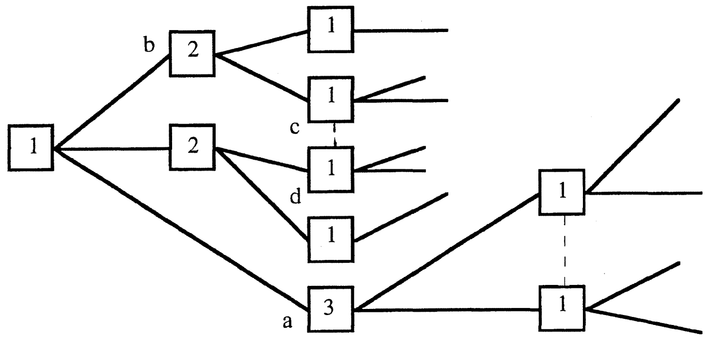
Let's think about what game theory can contribute to understanding players' behavior in the repeated prisoners' dilemma. If the game is repeated a finite number of times then collusive behavior cannot be rationalised. To see this, remember that the only reason to cooperate is to avoid retaliation in the future. This means that, in the last period, there is no incentive to cooperate. However, if both players are going to cheat in the last period, the next-to-last period can be analysed as if it were the last period and we can expect cheating then etc. so that we end up with the paradox that, even in the repeated prisoners' dilemma game, cheating is the unique equilibrium. (Of course we are assuming, as always, that players are intelligent, can analyse the game and come to this conclusion. If a player is known to be irrational, an optimal response could be to cooperate.) However, if the game is repeated over an infinite horizon or if there is uncertainty about the horizon (i.e. there is a positive probability (<1) of reaching the horizon), then cooperation can be generated. What is needed is that the strategies are such that the gain from cheating in one period is less than the expected gain from cooperation. For example both players could use trigger strategies (i.e. cooperate until the other player cheats and then cheat until the horizon is reached). This will be a Nash equilibrium if the gain from cheating for one period is smaller than the expected loss from a switch to both players cheating from then onwards.

Perfect equilibrium

So far, with the exception of the section 'Extensive form games', we have considered games in normal form. In this section we return to the extensive form representation. Consider Example 2.1 and its normal form representation at the beginning of the section on 'Normal form games'. From the payoff matrix it is clear that there are three Nash equilibria (T,(t,t)),(B,(t,b)) and (B,(b,b)). Two of these, the ones which have Player 2 playing b(t) regardless of what Player 1 plays, do not make much sense in this dynamic game. For example, (B,(b,b)) implies that Player 2 — **if** Player 1 plays T — would rather play b and get a payoff of -2 than t which gives payoff 0. The reason this strategy is a Nash equilibrium is that Player 1 will not play T.

The notion of perfect equilibrium was developed as a refinement of Nash equilibrium to weed out this type of unreasonable equilibria. Basically, the requirement for a perfect equilibrium is that the strategies of the players have to form an equilibrium in any subgame. A subgame is a game starting at any node (with the exception of nodes which belong to information sets containing two or more nodes) in the game tree such that no node which follows this starting node is in an information set with a node which does not follow the starting node. In the game tree in Figure 2.4, **a** is the starting node of a subgame but **b** is not since **c**, which follows **b**, is in an information set with **d** which does not follow **b**.

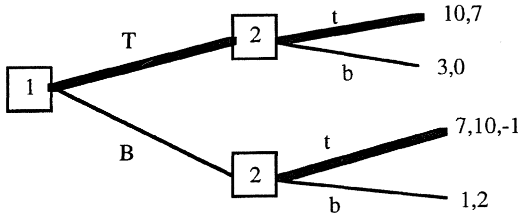
Figure 2.4: a starts a subgame, b does not



So (B,(b,b)) is not perfect since it is not an equilibrium in the (trivial) subgame starting at Player 2's decision node corresponding to Player 1's choice of T.

Example 2.8

Figure 2.5: Perfect equilibrium



Using backward induction it is easy to see that the perfect equilibrium is (T,(t,t)) as indicated on the game tree. If we analyse this game in the normal form, we find three Nash equilibria (marked with an asterisk in the payoff table). One of these, namely (B,(b,t)) can be interpreted as based on a threat by Player 2 to play b unless Player 1 plays B. Of course if such a threat were credible, Player 1 would play B. However, given the dynamic nature of the game, the threat by Player 2 is not credible since in executing it he would hurt not only his opponent but himself too (he would get a payoff of 0 rather than 7 which he could get from playing t). By restricting our attention to perfect equilibria we eliminate equilibria based on non-credible threats.

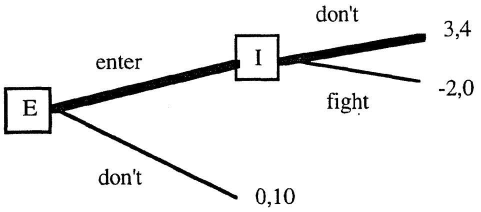
		Player 2			
		(t,t)	(t,b)	(b,t)	(b,b)
Player 1	T	10,7*	10,7*	3,0	3,0
	B	7,10	1,2	7,10*	1,2

Normal form

Example 2.9 ('entry deterrence')

The industrial organisation literature contains many game theoretic contributions to the analysis of entry deterrence. The simplest scenario is where the industry consists of a monopolist who has to determine his strategy vis-a-vis an entrant. The monopolist can decide to allow entry and share the market with the new entrant, or (threaten to) undercut the new entrant so he cannot make positive profits. The question arises whether the incumbent firm's threat to fight the entrant is credible and deters the entrant from entering. We can analyse this game using the notion of perfect equilibrium. On the game tree in Figure 2.6, E stands for entrant and I for incumbent. The entrant's payoff is listed first.

Figure 2.6: Entry deterrence



It is easy to see that, at the perfect equilibrium, the entrant enters and the incumbent does not fight. It is in the interest of the incumbent firm to accommodate the entrant. There are versions of the entry deterrence game which result in the incumbent fighting entry at equilibrium. In the 'deeper pockets' version, the incumbent has access to more funds since it is likely to be a better risk than the entrant, and can therefore outlast the entrant in a price war. In the incomplete information version, the entrant is not sure about the characteristics or payoffs of the incumbent (see Example 2.10). In the 'excess capacity' version, incumbent firms make large investments in equipment which affects their payoff if they decide to fight entry. If firms have a large capacity already in place, their cost of increasing output (to drive the price down) is relatively low and therefore their threat of a price war is credible. To see this in Example 2.9, find the perfect equilibrium if the incumbent's payoff of fighting increases to 6 when the entrant enters.

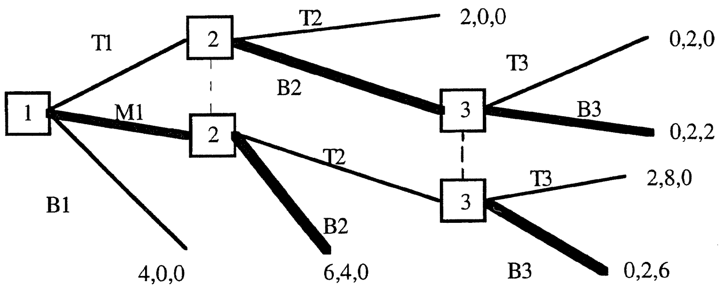
This example could be extended to allow for several potential entrants who move in sequence and can observe whether the incumbent (or incumbents if earlier entry was successful) allows entry or not. It would seem that, for an incumbent faced with repeated entry, it is rational to always undercut entrants in order to build a reputation for toughness which deters further entry. Unfortunately, as in the repeated prisoner's dilemma, this intuition fails. Selten (1978) coined the term 'chainstore paradox' to capture this phenomenon. The story is about a chain-store which has branches in several towns. In each of these towns there is a potential competitor. One after the other of these potential competitors must decide whether to set up business or not. The chain-store, if there is entry, decides whether to be cooperative or agressive. If we consider the last potential entrant, we have the one-shot game discussed above in which the entrant enters and the chain store cooperates. Now consider the next-to-last potential entrant. The chain-store knows that being aggressive will not deter the last competitor so the cooperative response is again best. We can go on in this way and conclude that all entrants should enter and the chain store should accommodate them all![2] We should remember that, as for the repeated prisoners' dilemma, the paradox arises because of the finite horizon (finite number of potential entrants). If we assume an infinite horizon, predatory behavior to establish a reputation can be an equilibrium strategy.

[2] *For an analysis of different versions of the chain store game in which the paradox is avoided see Kreps and Wilson (1982) and Milgrom and Roberts (1982).*

Perfect Bayesian equilibrium

In games of imperfect or incomplete information, the perfect equilibrium concept is not very helpful since there are often no subgames to analyse. Players have information sets containing several nodes. In these games an appropriate solution concept is perfect Bayesian equilibrium. Consider the imperfect information three-person game in extensive form represented in Figure 2.7. At the time they have to make a move, Players 2 and 3 do not know precisely which moves were made earlier in the game.

Figure 2.7: Bayesian equilibrium



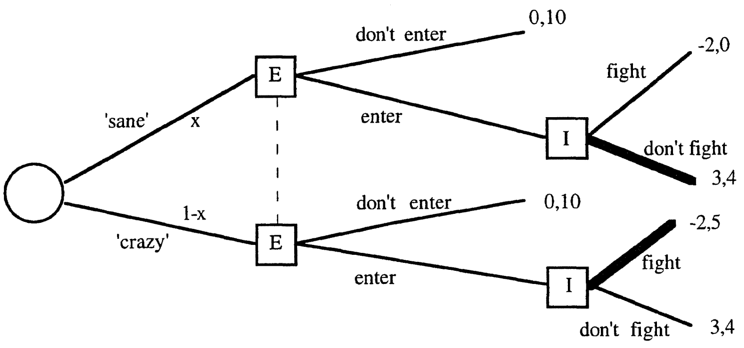
This game looks very complicated but is in fact easy to analyse. When Player 3 gets to move, she has a dominant strategy, B3: at each node in her information set, making this choice delivers her the highest payoff. Hence, whatever probability Player 3 attaches to being at the top or the bottom node in her information set, she should take action B3. Similarly, given Player 3's choice, Player 2 chooses B2. The choice of B2 leads to 2 or 4 depending on whether Player 2 is at the top or bottom node in his information set, whereas T2 leads to payoffs of 0 and 2 respectively. So again, independent of Player 2's probability assessment over the nodes in his information set, he chooses B2. Player 1, anticipating the other players' actions, chooses strategy M1.

In general for a perfect Bayesian equilibrium we require (a) that the equilibrium is perfect given players' assessment of the probabilities of being at the various nodes in their information sets and (b) that these probabilities should be updated using Bayes' rule and according to the equilibrium strategies. In other words, strategies should be optimal given players' beliefs and beliefs should be obtained from strategies. For the game represented in Figure 2.7, these requirements are satisfied if we set the probability of Player 2 being at his bottom node equal to 1 and the probability of Player 3 being at her bottom node equal to any number in [0,1].

Example 2.10

Let us return to the entry game of Example 2.9 and introduce incomplete information by assuming that the incumbent firm could be one of two types — 'crazy' or 'sane' — and that, while it knows its type, the entrant does not. The entrant subjectively estimates the probability that the incumbent is sane as x. This scenario is depicted in the game tree in Figure 2.8 with the payoffs to the entrant listed first. The important difference with the game of Example 2.9 is that here there is a possibility that the entrant faces a 'crazy' firm which always fights since its payoff of fighting (5) is higher than that of not fighting (4). The 'sane' firm always accommodates the entrant. The entrant's decision to enter or not will therefore depend on his belief about the incumbent's type. If the entrant believes that the incumbent firm is 'sane' with probability x then its expected payoff if it enters is 3x–2(1–x)=5x–2. Since the entrant has a payoff of zero if he doesn't enter, he will decide to enter as long as x>2/5.

Figure 2.8: Entry deterrence



Chapter summary

After this chapter and the relevant reading, you should understand:

- the concept of information set and why it is not needed in decision analysis
- why it is useful to have both extensive form and normal form representations of a game
- the importance of the prisoners' dilemma as a paradigm for many social interactions
- the concept of dominated strategies and the rationale for eliminating them in analysis of a game
- the concept of Nash equilibrium (This is absolutely essential!)
- the concept of non-credible threats and its application in entry deterrence.

You should be able to:

- represent a simple multi-person decision problem using a **game tree**
- translate from an extensive form representation to the **normal form** representation
- find **Nash equilibria in pure and mixed strategies**
- show why in a finitely **repeated prisoners' dilemma** game cheating is a Nash equilibrium
- explain the **chainstore paradox.**

Sample exercises

1. Consider the following matrix game.

	L	R
T	2,5	1,4
B	5,-1	3,1

Are there any dominated strategies? Draw the payoff region. Find the pure strategy Nash equilibrium and equilibrium payoffs. Is the Nash equilibrium Pareto efficient? Which strategies would be used if the players could make binding agreements?

2. Find the Nash equilibrium for the following zero sum game. The tabulated payoffs are the payoffs to Player I. Player II's payoffs are the negative of Player I's. How much would you be willing to pay to play this game?

	Player II			
Player I	-1	2	-3	0
	0	1	2	3
	-2	-3	4	-1

3. At price 50, quantity demanded is 1000 annually; at price 60 quantity demanded is 900 annually. There are two firms in the market. Both have constant average costs of 40. Construct a payoff matrix and find the Nash equilibrium. Assume that, if both firms charge the same price, they divide the market equally but, if one charges a lower price than the other, it captures the whole market. Suppose the two firms agree to collude in the first year and both offer a most favoured customer clause. What is the payoff matrix for the second year if they colluded the first year?

4. Find the pure and mixed strategy equilibria in the following payoff tables. How might the -100 payoff affect the players' actions?

	L	R
T	12,10	4,4
B	4,4	9,6

	L	R
T	12,10	4,4
B	4,-100	9,6

5. Students don't enjoy doing homework and teachers don't like grading it. However, it is considered to be in the students' long-term interest that they do their homework. One way to encourage students to do their homework is by continuous assessment (i.e. mark all homework), but this is very costly in terms of the teachers' time and the students do not like it either. Suppose the utility levels of students and teachers are as in the payoff matrix below.

		teacher	
		check	don't check
student	work	0,-3	0,0
	no work	-4,4	1,-2

a. What is the teacher's optimal strategy? Will the students do any work?

b. Suppose the teacher tells the students at the beginning of the year that all homework will be checked and the students believe her. Will they do the work? Is the teacher likely to stick to this policy?

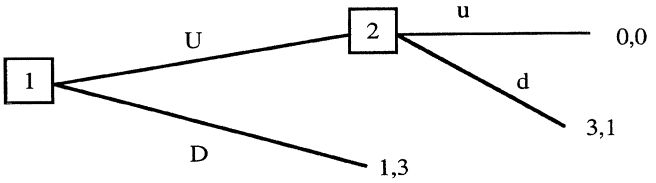
c. Suppose the teacher could commit to checking the homework part of the time but students will not know exactly when. What is the minimal degree of checking so that students are encouraged to do the work? (i.e. what percentage of homework should be checked?

6. Consider the two player simultaneous move game below where payoffs are in £. Find the pure strategy Nash equilibria. How would you play this game if you were the row player? How would you play this game if you were the column player?

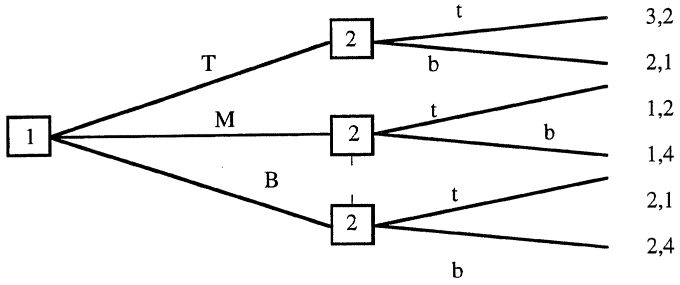
	a	b	c	d
u	100,3	2,2	2,1	0,-500
d	0,-5000	3,-500	3,2	1,10

7. Two neighbours had their house broken into on the same night and from each house an identical rare print went missing. The insurance company with whom both neighbours had taken out home contents insurance assures them of adequate compensation. However, the insurance company does not know the value of the prints and offers the following scheme. Each of the neighbours has to write down the cost of the print, which can be any (integer) value between £10 and £10000. Denote the value written down by Individual 1 as x_1 and the one written down by Individual 2 as x_2. If $x_1=x_2$ then the insurance company believes that it is likely that the individuals are telling the truth and so each will be paid x_1. If Individual 1 writes down a larger number than Individual 2, it is assumed that 1 is lying and the lower number is accepted to be the real cost. In this case Individual 1 gets x_2-2 (he is punished for lying) and Individual 2 gets x_2+2 (he is rewarded for being honest). What outcome do you expect? What is the Nash equilibrium?

8. Consider the extensive form game below. Find all Nash equilibria. Find the subgame perfect equilibrium.



9. Consider the extensive form game below. What are the players' information sets? Write this game in normal form and analyse it using the normal form and the game tree.



10. A firm may decide to illegally pollute or not. Polluting gives it an extra payoff of g>0. The Department of the Environment can decide to check for pollution or not. The cost of this inspection is c>0. If the firm has polluted and it is inspected, it has to pay a penalty p>g; in this case, the payoff to the department of the environment is s–c>0. If the firm has not polluted and it is checked, no penalty is paid. If the department of the environment does not inspect, its payoff is 0.

   a. Suppose that the Department of the Environment can observe whether the firm has polluted or not before it formally decides whether or not to check. Draw the game tree. What is the equilibrium?

   b. Now suppose the pollution cannot be observed before checking. Draw the game tree. Is there a pure strategy equilibrium? Compute the mixed strategy equilibrium. How does a change in the penalty affect the equilibrium?

Notes

The alternating offers bargaining game; incomplete information bargaining

Chapter 3

Bargaining

Texts

Tirole, J. *The Theory of Industrial Organization.* (New York: The MIT Press, 1988) [ISBN 0262200766] Section 11.5.2.

References cited

Fudenberg. D. and J. Tirole *Game Theory.* (Cambridge, Mass.: The MIT Press, 1991) [ISBN 0262061414].

Rubinstein, A. 'Perfect equilibrium in a bargaining model', *Econometrica* (1982) 50: 97–109.

The literature on bargaining has developed dramatically in the last decade or so due to advances in noncooperative game theory. Bargaining is an interesting topic of study because it has both **cooperative** and **conflictual** elements. For example, when a seller has a low reservation price for an object and a buyer has a high reservation price then, clearly, if the two parties can agree to trade, they will both be better off. On the other hand, conflict exists regarding the divisions of the gains of trade. The seller will naturally prefer a high price and the buyer will prefer a low price.

Game theory helps us to model bargaining situations carefully and allows us to check our intuition regarding, for example, how the outcome of the bargaining will depend on the parties' bargaining power and so on. Questions economists are interested in include:

- under which conditions will bargaining lead to an efficient outcome
- what are good bargaining strategies?

Bargaining problems arise whenever payoffs have to be shared among several players. When firms succeed in running a cartel,[1] for example, agreeing on how to divide cartel profits is a major problem. Managers are interested in bargaining models for their predictions in the context of management (e.g. for labour (union) negotiations and strikes). However, most game theoretic models of bargaining are either very simplistic (and that is certainly true for the ones I discuss in this chapter) or extremely complex and unrealistic in their assumptions about players' ability to reason and calculate.

In the hope that this does not discourage you too much, let us proceed.

[1] *See Chapter 11, 'Oligopoly'*

The alternating offers bargaining game

The alternating offers bargaining model was formulated and solved fairly recently by Rubinstein (1982). In this model two players have to decide on how to divide an amount of money between them. They alternate in making suggestions about this division. However, as time goes by, the amount of money available shrinks. It turns out that, at the perfect equilibrium of this game, the players agree immediately. To see how this conclusion is arrived at, suppose two players, Bert and Ernie, have a total of £100 to divide between them. Say Bert makes an offer first. He might decide, for example, to keep £70 to himself and offer £30 to Ernie. Ernie will then agree or make a counteroffer. If he agrees, he will get the £30; if he refuses the offer and makes a counteroffer, the £100 'cake' will shrink to £100δ. The discount factor $\delta(0 < \delta < 1)$ represents the cost of a delay in reaching an agreement (it represents the cost of a strike, for example, where potential output and profit is lost while parties disagree). If the game ends here (i.e. by Bert accepting or rejecting Ernie's counteroffer) and we assume that if no agreement has been reached both players get zero, it is not hard to see what will happen. (You may want to draw a game tree at this point.) Assuming Bert and Ernie are like the usual self-interested non-altruistic players then, if Ernie decides to make a counteroffer, he will offer Bert one penny which Bert will accept. Ernie in this case ends up with about £100δ. If Bert wants to avoid this predicament he will have to offer Ernie this payoff from the start so that he keeps £100 (1-δ) for himself. So if there is only one round of offers, at the perfect equilibrium, Bert will offer £100δ to Ernie and Ernie will accept.

Now consider the same game but Bert will get to make a second offer (i.e. the sequence of moves is Bert, Ernie, Bert). By the time Bert makes his second move, the cake will have shrunk to £100δ^2. Now Bert will have the 'last mover advantage' and will offer a penny to Ernie. Ernie will anticipate this and offer Bert £100δ^2 when he has the opportunity so that he keeps £100δ – £100δ^2 = £100δ(1-δ) for himself. However, Bert can improve on this by offering Ernie £100δ (1-δ) in the first move and by keeping £100 (1-δ+δ^2) for himself. You should be convinced by now that the subgame perfect equilibrium strategy for each player tells him to make an offer which leaves his opponent very close to indifferent between accepting the offer and continuing the game. As a consequence, **the first offer is always accepted.**

What happens if we don't give the players a deadline? Suppose they can keep making offers and counteroffers for an infinite number of periods but, as before, the cake shrinks in each period. Of course this eliminates the 'last mover advantage' and we can look for a symmetric equilibrium (players using the same strategy). Also, the equilibrium strategy must be stationary (i.e. it should give the same prescription in each period because, in the infinite version, when an offer has been rejected, the game is exactly as it was before the offer was made except for the shrinking). So, let us assume that Bert offers Ernie £100x and thus keeps £100(1–x) to himself. Ernie will consider accepting £100x or making an offer of £100δx to Bert and keeping £100$\delta(1-x)$. Since he should be indifferent between these two options we find $\delta(1-x)=x$ or $x=\delta/(1+\delta)$. At the equilibrium Bert gets £100/(1+δ) and Ernie gets £100δ/(1+δ). In the infinite version of this game it is an advantage to be able to make the first move. Again, as in the finite version, the first offer is always accepted.

As the time between offers and counteroffers shrinks, the discount factor approaches 1 and the asymmetry, caused by who moves first, disappears. It is not very difficult to extend this analysis to allow for different discount factors of the two bargaining parties. The conclusion of this modified bargaining game is that the more patient bargaining partner will get a larger slice of the cake.

Experimental work

Although the alternating offers bargaining game has a simple 'solution' and a stark prediction about the duration of the bargaining (one offer only), the game theoretic findings are not always replicated in experiments. About 10 years ago, when I was a student at LSE, I (and many others) participated in a series of bargaining experiments. Subjects were paired to an unknown bargaining partner and could only communicate via a formal computer program with their partner who was sitting in a different room. Experiments of this type generally show that 'real' players have a tendency to propose and accept what they consider a fair offer while rejecting what they consider a mean offer even if this rejection means they will be in a worse position. If you were offered one penny in the last round of the game in this section, would you accept?

Incomplete information bargaining

The alternating offers bargaining game does not look very appealing as a paradigm for real life bargaining situations in which disagreement is common and costly negotiations take place for several weeks or months. It turns out that to generate delayed agreement you have to assume that the players do not know all the information there is to know about their opponent (i.e. players have private information). In particular, players may have private information about their reservation price when bargaining over the sale of an item. Models of incomplete information bargaining can be extremely complex and I will not discuss them in general or even give an overview.[2] Instead, I will show you in a simple example that inefficiencies can occur. This means that, if it were possible to get both players to reveal their valuations truthfully, they could both be made better off. It is precisely because players are hiding their valuations that there are costly delays before an agreement is reached.

Two players, a seller and a buyer, are trying to come to an agreement about the price at which a good will be sold. The seller has a valuation (or reservation price) of 1 or 3, equally likely. The buyer has valuation 2 or 4, equally likely. I will refer to a player as being of type *i* if he has valuation *i*. The seller moves first and offers to sell for a price of 2 or 4. The buyer always accepts an offer of 2 but may reject a price of 4. If the buyer rejects, the seller can offer a price of 2 or 4 and the buyer has another chance to accept or reject. If there is delay any payoffs in the second period are discounted using a discount factor d_s for the seller and d_b for the buyer. Table 1 contains the possible strategies for each type of player and the payoffs corresponding to each possible strategy pair. The first payoff listed is the buyer's. Note that a seller of type 3 will never set the price equal to 2 and that a buyer of type 2 is never willing to buy at price 4. This restricts the number of strategies we have to consider.

	seller type 1			seller type 3
	ask 2	ask4, then 2	ask 4, then 4	always ask4
buyer type 2 always rejects 4	(0,1)	$(0,d_s)$	(0,0)	(0,0)
buyer type 4 reject 4 once	(2,1)	$(2d_b,d_s)$	$(0,3d_s)$	$(0,d_s)$
do not reject	(2,1)	(0,3)	(0,3)	(0,1)

For a buyer of type 4, rejecting a price of 4 in the first period weakly dominates accepting price 4 immediately. If we eliminate the last row in the table then, for a seller of type 1, asking a price of 2 dominates asking 4 first and then asking 2 so that we can eliminate the second column in the table. Given that the seller of type 1 thinks that the buyer he faces is equally likely to be of type 2 as of type 4, asking 2 gives him an expected payoff of 1 whereas asking 4 twice gives him an expected payoff of $0(1/2)+(3d_s)(1/2)$. Thus the seller of type 1 asks 2 if $1>(3d_s/2)$ or $d_s < 2/3$. It follows that, if $d_s > 2/3$, there is no trade, at the equilibrium, between a seller of type 1 and a buyer of type 2 which is clearly inefficient. If $d_s > 2/3$ there is an inefficient delay in the agreement between a seller of type 1 and a buyer of type 4 at the equilibrium. There is also an inefficient delay of the agreement between a seller of type 3 and a buyer of type 4.

[2] *For an advanced treatment of this material see, for example, Fudenberg and Tirole (1991) Chapter 10*

Chapter summary

After this chapter and the relevant reading, you should understand:

- the **cooperative and conflictual** aspects of bargaining
- the nature of the **inefficiencies** associated with incomplete information bargaining.

You should be able to:

- analyse **alternating offer bargaining games** with finite or infinite number of rounds.

Sample exercise

Consider the alternating offers bargaining game over £100, with Bert making the first offer and Ernie making a counteroffer if he wants to. Suppose Bert has an outside option of £50, that is, at any point during the game, Bert can stop bargaining and get £50 (discounted if he gets it after the first period). If Bert takes his outside option, Ernie gets zero. How does this affect the equilibrium strategies and payoffs?

Adverse selection; moral hazard; signalling and screening; principal-agent problems

Chapter 4

Asymmetric information

Texts

Varian, H.R. *Intermediate Microeconomics.* (New York: W.W. Norton and Co., 2006) seventh edition [ISBN 0393927024] Chapter 37.

References cited

Akerlof, G.A. 'The market for "lemons": quality uncertainty and the market mechanism', *Quarterly Journal of Economics* (1970) 85: 488–500.
'Flattened, sort of', *The Economist,* 5 November 1994, 100.
'Generation X-onomics', *The Economist,* 19 March 1994, 55–56.
'Keep taking the tablets', *The Economist,* 4 December 1993, 81.
Milgrom, P. and J. Roberts 'Price and advertising as signals of product quality', *Journal of Political Economy* (1986) 94(4): 796–821.
Rothschild, M. and J. Stiglitz, J. 'Equilibrium in competitive insurance markets: an essay on the economics of imperfect information', *Quarterly Journal of Economics* (1976) 91: 629–49.
Spence, M. *Market signalling.* (Cambridge: Harvard University Press, 1972).
'Still money in that franchise', International Banking Survey, *The Economist,*. 30 April 1994, 43–46.
'The consequences of kindness', The Nordic Countries Survey, *The Economist,* 5 November 1994, 13–16.

Situations of asymmetric information arise when one party involved in a transaction knows more than the other about relevant variables. For example, a seller may have more information about the quality of the product he is trying to sell than the buyer does. With asymmetric information decisions are taken which would be inefficient under symmetric information. In extreme cases, markets which would show vigorous trade under perfect information may collapse completely. It becomes very difficult under asymmetric information to draw inferences from people's behavior, as they may be trying to fool you. For example, when bargaining, a buyer may make a low bid in the hope that the seller will believe he has a low reservation price. By bluffing in this way he may increase his expected gain at the expense of an inefficient delay in reaching an agreement.

Adverse selection

Asymmetry of information may lead to **adverse selection** where less desirable agents are more likely to voluntarily participate or 'self-select' in trade. In adverse selection problems, the **type of agent is not observable** and has to be guessed. This is why adverse selection problems are also known as **hidden information** problems. For example, people who suffer from terminal diseases are more likely than others to want to buy medical insurance or life insurance. The individuals considering buying insurance have better information about their risks than an insurance company does.

As a consequence of this adverse selection problem, the insurance company cannot use risk estimates from the general population to set its premiums. The premiums will have to be higher than those based on the general population. Similarly, if a bank charges the same interest to low risk and high risk borrowers, the high risk ones will be more likely to borrow. The bank faces an adverse selection of borrowers. If you advertise a job vacancy at a given wage, you only get people applying who are willing to work at that wage. From a pool of potential applicants, only the less desirable workers might apply. Adverse selection can be responsible for market failure (i.e. there may be no market for a good whereas profitable transactions between buyers and sellers would be possible if everyone had full information).

The 'lemons' problem

The 'lemons' problem, first studied in a seminal paper by Akerlof (1970), provides an excellent example of an adverse selection problem. The terminology of this problem has its origins in American slang: used cars are called 'lemons' if they are of bad quality whereas 'plums', 'peaches' or 'cherries' are high-quality cars. The starting point of the analysis is a question you may have wondered about: 'Why are one-year old cars with low mileage so much cheaper than new cars?' The answer lies in the quality of one-year old cars which are offered for sale.

Example 4.1

Suppose one-year old cars of a particular make and model have a quality or value (to the original buyer) uniformly distributed on the interval from £10,000 to £20,000. The buyer of a new car only finds out the value the car will have at age 1 after he has bought the car. He is now considering selling the car on the used car market. Suppose a used car buyer (there are many of them) is willing to pay £500 more than the seller's valuation of the car. (The reason for this could be that new car buyers do not like driving cars over one-year old whereas used car buyers do not care much about how new the car is.) If buyers and sellers had perfect information about quality then all cars would be sold (for a price between the seller's valuation and the buyer's valuation) and everyone would be better off. Here it is assumed that the seller knows the quality but the buyer does not. What will the price be?

Because of buyer competition, the price in the market equals the (expected) value to the buyers. If the price is p, then all sellers who value their car at p and below will try to sell their car. So all cars offered for sale have values (to the seller) between £10,000 and p with an average quality of (£10,000 + p)/2. Buyers are willing to pay £500 more than the value to the seller i.e. p = (£10,000 + p)/2 +£500 which gives an equilibrium price p = £11,000. This means that only the lowest quality cars (those with value below £11,000) are offered for sale. Only 10 per cent of the cars are sold. You should verify that, if the buyer premium (the excess the buyer is willing to pay over the value of the car to the seller) is less than £500, even fewer cars get sold.

Although Example 4.1 was phrased in terms of the used car market, the 'lemons' problem can occur in any market where buyers and sellers have different information about the quality of the good being sold. In such markets there is a tendency for low quality to crowd out high quality. You might think that buyers and sellers should be able to write a contract stipulating the quality sold, with appropriate penalties if the delivery is later found to be of low quality. However, writing such contracts is costly and only makes sense if they can be enforced. It is often very difficult for a buyer, let alone a court, to assess quality accurately and to determine whether any defect existed at the time of the sale or was caused by negligent use. The adverse selection problem can be attenuated if reputation is important, as is the case when a seller aims to repeatedly sell to a given buyer, and by the use of warranties.[1]

[1] *Discussed in 'Signalling and screening'*

Example 4.2

Consider the example of home contents insurance. There is a large variance geographically in burglary rates, even within cities. For concreteness assume that a fraction p_h of potential buyers of insurance expect a loss (which becomes a claim if they are insured) of £400 per year and the remaining potential buyers (p_l) expect a loss of £100 per year. Since potential buyers of insurance are risk averse, they are willing to pay a premium R which is higher than their expected claims. Suppose the high claim individuals are willing to pay up to £500 per year and the low claim individuals' reservation price is £130 per year. The insurance company does not know who the high and low claim individuals are and we assume it has to set a uniform premium for all customers. Ignoring any administration costs and assuming everyone is insured, the company has to set the premium to cover its expected payout:

$$R \geq 400\ p_h + 100\ p_l.$$

As long as there are many low claim individuals this premium does not exceed the low claim individuals' reservation price:

$$400\ p_h + 100\ p_l = 400\ (1-p_l) + 100\ p_l < 130 \text{ for } p_l > 90\%.$$

However, when there are too few low claim individuals, the premium required by the insurance company to cover its cost would have to increase to above £130. This implies that only high claim individuals buy insurance at a high premium $R > 400$. Note that under perfect, symmetric information everyone would buy insurance at the appropriate rate for their risk category. However, because of the asymmetry of information, if the insurance company continues to set the premium based on the average claim, it will make losses since only the high risk properties will be insured. So it must set the premium based on the high risk clients and only these will be insured.

Problems of adverse selection can sometimes be overcome by compulsory insurance regulation requiring all homeowners to insure their homes or requiring everyone to take out medical insurance for example. Some employers insure all their employees as one package to avoid adverse selection problems. If everyone is insured, the high risk individuals will be better off since they pay a lower premium. The low risk individuals are not necessarily better off but they will get the insurance at a lower premium than when the premium was based on only high risk individuals. Whether low risk individuals are better off under this scheme depends on how risk averse they are as the insurance they are offered is not actuarially fair (premium > expected loss).

As another example of eliminating the adverse selection problem in this way, consider extended five year or 50,000 miles warranties for cars. These used to be purchased only by people who were likely to make extensive use of the warranty. Now some manufacturers give this kind of warranty as a standard feature and include its cost in the price of the car.

Example 4.3

The value of Target Ltd. under its current management is known only to the current management team. EFM Ltd. is interested in taking over Target and estimates that Target's current value is 50, 60 or 75 (£ million) — all equally likely — whereas, after takeover and under EFM's management, the value would be 50, 70 or 85 respectively. The procedure for the takeover bid is that EFM makes an offer which Target can accept or reject. EFM only gets one chance to make an offer. How much should they bid? Adverse selection occurs because, for any bid EFM makes, Target will only be interested if it has a low value (below the level of the bid). In fact we have an extreme case of adverse selection here in that it prohibits trade altogether:

- if EFM offers 50 it cannot make a profit since Target would only sell if its value was 50
- if EFM offers 60, Target sells if its value is 50 or 60 which gives an expected gain to EFM of (1/3)(–10)+(1/3)(10)=0
- if EFM offers 75 then Target will accept the offer for sure but EFM expects a loss: (1/3)(–25) + (1/3)(–5) + (1/3)(10) < 0.

In Example 4.3, a takeover bid is only accepted if the target's value under its current management is below it. If information were symmetric (i.e. if the bidder had equal access to information about the target's value as the target's current management, the inefficiency associated with the absence of a takeover could be avoided). A successful bid could be made in **any** case, whatever the target's value is, and trade would take place at some price between the target's value under its current management and its value after the takeover under a new management team.

Moral hazard

'Moral hazard' is about incentive problems where one agent **cannot observe the actions** of the other agent. This is why moral hazard problems are also known as problems of **hidden action.** In the typical example, being insured changes an individual's behavior. If someone has taken out insurance for theft or collision he is likely to be more careless than when he was not insured. Whereas the term 'adverse selection' generally applies to characteristics or qualities of one of the parties **before** a contract is entered into **(pre-contractual opportunism)**, 'moral hazard' describes situations in which one of the parties misbehaves **after** a contract is signed **(post-contractual opportunism)**. A restaurant which offers an all-you-can eat deal gets an adverse selection of big eaters as its clientele. If a group of people eat out and decide to share the bill equally, moral hazard implies that they are likely to overindulge. If a company rewards employees based on seniority rather than performance, it may get an adverse selection of low-achieving applicants. When compensation is based on team performance rather than individual performance, moral hazard leads to reduced efforts.

In the 1980s the American government paid out US$300 billion when hundreds of savings and loan associations failed. This debacle has been explained in terms of moral hazard on the part of the banks who (are required by law to) take out deposit insurance provided by the government. **With** the deposit insurance, banks do not have the proper incentive to avoid excessive risk taking.[2] For life insurance policies in the US, the insurance company pays the beneficiary after a suicide only if the suicide took place a year or two years from the time the policy was issued. Life insurance statistics show that the suicide rate is lowest in the 12th and 24th month and highest in the 13th and 25th month of the policy!

Of course, if the insurer can observe the behavior of individuals (e.g. where they park their cars or whether they lock them — in the case of motor insurance; whether they smoke or take drugs in the case of life insurance), it could group them in different risk classes and set a different premium for each class. The problem is that it is often impossible or very costly to monitor behavior. To counteract the moral hazard problem in insurance, companies often do not give complete insurance but use deductible provisions.

[2] *See 'Still money in that franchise'*

Example 4.4

Elmo wants to insure his car, of value V, against theft. Elmo who has wealth W (this includes the value of his car) is risk averse and has a strictly concave utility of money function U. Elmo knows he should really always lock his car and fix the anti-theft device he keeps in his trunk. However, everything else equal, he prefers to be lazy and leave the car unlocked. If he is careful his car gets stolen with probability p_f; if he is careless his probability of ending up carless (!) is $p_1(p_1 > p_f)$. If Elmo does not have insurance he will prefer to be careful since:

$$(1 - p_f)\ U(W) + p_f\ U(W - V) > (1-p_1)\ U(W) + p_1U(W-V).$$

The insurance company is risk neutral and the insurance industry is assumed to be perfectly competitive. This implies that profits are zero (the premium is determined as the expected claim) and the insurance company offers Elmo his utility maximising contract (R,D) where R is the premium and D is the payment Elmo receives if his car gets stolen. If the insurer could observe Elmo's behavior then it would maximise:

$$(1-p_i)U(W-R)+p_iU(W-V+D-R) \text{ subject to } p_1D = R,\ i = 1,\ f.$$

Substituting the constraint for R and setting the derivative with respect to D equal to zero leads to:

$$(1-p_1)\ U'(W-p_iD)(-p_i)+p_i\ U'(W-V+D-p_i\ D)(l-p_i) = 0$$

so that D = V or, in words, the insurer offers full insurance. (No surprises here – we've seen this in the chapter on decision analysis!) Elmo ends up with utility U(W-R) and will choose to be careful and pay a low premium.

Assume now that the insurer cannot observe its client's behaviour. If it offers full insurance, Elmo is going to be careless given any premium R he is asked to pay since, once he is insured, his expected utility is U(W-R) whether he is careless or not. The insurance company anticipates this behaviour and therefore has to set the premium accordingly: $R = p_1\ V$. Will Elmo accept this contract? It depends on whether his expected utility without insurance (and with carefulness) exceeds the expected utility of the insurance contract:

$$(1-p_f)U(W) + p_fU(W-V) >=< U(W-p_1V)$$

For example if U(x) = ln(x) then the equation above reduces to:

$$(1-p_f)\ \ln(W) + p_f\ln(W-V) > = < \ln(W-p_1V).$$

For W = 10,000; V = 7,000; p_f= 0.01 and p_1 = 0.015 Elmo would insure. If p_1 = 0.05, Elmo would not insure.

This type of market failure is however easily remedied through the use of a deductible. If the insurance company does not offer full insurance but sets D<V, then if Elmo buys insurance his expected utility:

$$(1-p_i)U(W-R)+p_iU(W-V+D-R)$$

is decreasing in p_i and Elmo therefore has an incentive to be careful. The insurance company can now afford to base its premium on anticipated careful behavior (i.e. $R = p_fD$). Note that the size of the deductible is unimportant; in fact, for utility maximisation, we want D as close to V as possible but not equal to it.

In this insurance example, inefficiency results from the policyholder's inability to commit to careful behavior after he has purchased a full cover policy. If the insurance company offers a policy with a deductible then the client is in some sense worse off because he is risk averse and would rather ensure 100 per cent. However, the inability to commit to careful behavior ceases to be a problem. By altering the insurance contract slightly from the full cover format, the client's incentives have changed dramatically and he is ultimately better off than under the full cover policy!

Signalling and screening

In situations of adverse selection or moral hazard, the **informed** party may have an incentive to reveal his information to the uninformed party. A seller of high quality used cars could certainly do better if he could tell potential buyers that her cars are indeed of high quality. However, just posting a sign saying 'I sell good cars' will not achieve very much. For the customers to believe the seller, the seller has to prove in some way that her cars are good. She could do this by offering a warranty. Clearly, if offering a warranty is **only** profitable when the cars are of good quality (i.e. a seller of 'lemons' would make a loss if he offered a warranty), then the warranty is a credible **signal** of good quality.

Similarly, firms which have low costs might want their potential competitors to know this for entry deterrence purposes. Entry may only be profitable if the incumbent has high costs. A low cost incumbent can signal through limit pricing, which would be unprofitable for a high cost incumbent. Advertising can have a pure signalling purpose. For some goods, advertisements do not tell customers anything other than that the good is for sale. The explanation might be that the seller feels so confident about his product's quality that he predicts that the customer who tries it will become a regular buyer. Indeed, if customers would not purchase repeatedly, the seller could not recuperate the advertising expenditure.[3] Generally, the cost of signalling should be lower for the higher quality parties. It should not pay for the lower quality parties to pretend to be high quality by imitating the behavior of high quality parties.

The **uninformed** party usually has an incentive to obtain information. There are certain measures the uninformed party can take to alleviate the asymmetric information problem. When it engages in activities to find out the other party's private information it is said to be **screening**. So the distinction between screening and signalling is that screening is done by the uninformed party and signalling is done by the informed party. Banks and insurance companies can try to identify risk classes. Car insurers, for example, look at a driver's past record, age and sex, whether the car owner has a garage etc. Business premises with fire sprinkler systems can get a fire insurance policy with a lower premium. Similarly, smokers pay more for medical insurance. For some policies, potential clients have to undergo a medical examination and insurance can be refused based on the outcome of this examination. Banks use sophisticated multivariate statistics to assess borrowers' credit histories and predict their probability of defaulting on a loan. Used car buyers in the UK get an AA (Automobile Association, not Alcoholics Anonymous!) inspector to check a car before purchase.

If characteristics of customers are unobservable, firms can use **self-selection constraints** as an aid in screening to reveal private information. For example, consider the phenomenon of rising wage profiles where workers get paid an increasing wage over their careers. An explanation may be that firms are interested in hiring workers who will stay for a long time. Especially if workers get training or experience which is valuable elsewhere, this is a valid concern. A firm will then pay workers below the market level initially so that only 'loyal' workers will self-select to work for the firm.

[3] *See, for example, Milgrom and Roberts (1986)*

Education as a signal

The classic example of **signalling**, first analysed by Spence (1974), is one in which high productivity individuals try to differentiate themselves from low productivity individuals by investing in education. The firm cannot distinguish between high and low quality workers but the workers themselves know their abilities. As I will show in the following example, good workers can use education as a signal of productivity since **only** the most productive workers invest in education. In other words, the signalling cost to the good workers is lower than to the low productivity workers and therefore the two types of workers can be identified because they make different choices.

Example 4.5

When workers enter their first job, it is difficult to estimate their productivity. Assume that workers know their own productivity but firms do not and that a high productivity worker is worth £30,000 per year and a low productivity worker is worth £15,000 per year to the firm. There are as many high productivity workers as there are low productivity workers. If the labour market is competitive, competition between firms drives up the wage to:

$$(0.5)\ £15{,}000 + (0.5)\ £30{,}000 = £22{,}500.$$

High productivity workers would like to convince the firm that they are indeed worth more than average. An opportunity to do exactly that may exist if workers can invest in education before they apply for a job. The employer, if he is convinced that workers who are educated are also highly productive, will pay £30,000 if the worker is educated and £15,000 otherwise. The cost of a university education to an individual consists of £15,000 tuition fees and living expenses (in excess of the government subsidy) plus an opportunity cost of £45,000 (i.e. three years of wages foregone while studying) plus a cost of effort. (This last cost varies for high and low productivity individuals.) Let's say that high productivity workers are intelligent and actually enjoy studying so that their cost of effort is zero whereas low productivity workers find studying extremely strenuous and they value the cost of effort involved in getting a university degree at £50,000. Hence the total cost of education is £60,000 for a high productivity worker and £110,000 for a low productivity worker.

For simplicity we assume that university education has no effect on productivity and we ignore what happens after a worker leaves his first job. If workers expect to stay in their first job for five years, signalling is effective. If firms pay university graduates £30,000 and others £15,000, all high productivity workers get a university degree and none of the low productivity workers go to university. To check this, we have to show that a high productivity worker gains from getting a degree (see (a) below) and a low productivity worker does not (see (b) below):

$$(30{,}000)\ (5) - 60{,}000 = 90{,}000 > (15{,}000)\ (5) = 75{,}000 \qquad \text{(a)}$$

$$(30{,}000)\ (5) - 110{,}000 = 40{,}000 < (15{,}000)\ (5) = 75{,}000 \qquad \text{(b).}$$

We have found a **separating equilibrium** in which the two types of workers are identified through signalling. In this example, education is beneficial to the high productivity workers but wasteful for society as a whole in the sense that, under perfect, symmetric information its cost could be avoided. It is useful only as a signal here.

You should be able to show that, if the expected tenure for the first job is less than four years, signalling does not pay and no worker gets a degree. In this case we find a **pooling equilibrium** in which the two types of workers use the same strategy and are paid £22,500. Similarly, if the expected tenure is long enough, say eight years or more, signalling will not work because everyone would prefer to get a degree and earn £30,000. Firms which want to break even cannot pay the high salary to graduates in this case and as a result nobody gets a degree.

Note that the high productivity workers are negatively affected by the existence of low productivity workers. Either they have to invest in signalling (in a separating equilibrium) or they get a wage below their productivity (in a pooling equilibrium). This type of externality is an important aspect of the adverse selection problem.

In the US, where labour markets are less regulated than in Europe and the minimum wage is only US$4.25 per hour, the wage differential between college graduates and others is large and increasing. In 1994, college graduates earned an average of 77 per cent more than high-school graduates.[4] Of course, people increase their productivity in college by learning computer skills and other skills valued by employers. At the same time, the sectors of the economy which employ unskilled workers have seen fiercer foreign competition which drives down wages. In addition it seems plausible that employers use college education as a proxy or a signal of ability and high productivity and that this explains part of the wage gap. Contrast this with Sweden's labour market where workers with a university education are paid 3.5 per cent more than workers who did not get a university education (few Swedes go to university).[5]

An insurance menu

Uninformed parties can screen by offering a menu of choices or possible contracts to prospective (informed) trading partners who 'self-select' one of these offerings. This type of screening was developed in an insurance context by Rothschild and Stiglitz (1976). They show that, if the insurer offers a menu of insurance policies with different premiums and amounts of cover, the high risk clients self-select into a policy with high cover. Example 4.6 illustrates how, in an insurance market, inefficiency due to adverse selection can be ameliorated through the insurance company offering clients two contracts: a low premium, partial cover policy and a high premium full cover policy.

Example 4.6

Assume there are no moral hazard problems (i.e. individuals do not have any control over their probability of a claim). There are low risk and high risk individuals (all risk averse) with claim probability p_l and p_h respectively. (The insurer knows p_l and p_h but he does not know who the low risk and high risk individuals are.) For simplicity I assume that low risk and high risk individuals are identical with repect to initial wealth W, size of the potential loss V and utility function U.

If the insurance company offers a policy (R,D) where R is the premium and D is the payment when the policyholder suffers a loss, then it is possible that only the high risk individuals are insured (see Example 4.2). Suppose now that the insurance company offers a menu of two policies, a low premium policy with a deductible (R_l, D_l) and a high premium policy with no deductible (R_h, V). The purpose of these two policies is to get clients to self select such that the low risk group chooses the first policy and the high risk group the second. Assuming, as before, that the insurance industry is competitive and therefore profit on any policy is zero, the premiums are set according to $R_l = p_l D_l$ and $R_h = p_h V$.

Clearly, the low risk group prefers the fair (premium equal to expected claim) insurance (R_l, D_l) to no insurance. Similarly, high risk individuals prefer (R_h, V) to no insurance. What is most important though is that the high risk group prefers (R_h, V) to (R_l,D_l). If it prefers the contract designed for the low risk group then the insurance company cannot break even. So we need:

$$U(W - R_h) > (1 - p_h)\, U(W - R_l) + p_h\, U(W - R_l + D_l - V). \quad (*)$$

[4] See 'Generation X-onomics'

[5] See 'The consequences of kindness'

The left hand side of this inequality is the expected utility to a high risk individual if he chooses policy (R_h, V) and the right hand side is his expected utility if he chooses policy (R_l, D_l). Given that $R_l = p_l D_l$ and $R_h = p_h V$, this puts a constraint on the design of the menu of policies in the form of an upper bound on D_l. The existence of the high risk group imposes a policy with a deductible on the low risk group. For the parameter values corresponding to the car theft insurance of Example 4.4 (W=10,000, V=7,000), U(x) = ln(x), p_l=0.01 and p_h = 0.1, the premiums are R_h = (0.1)(7000)=700 and R_l = (0.01)D_l. The inequality (*) for these values reduces to:

$$\ln(10{,}000 - 700) > (0.9)\ln(10{,}000 - 0.01\,D_l) + (0.1)\ln(10{,}000 - 0.01\,D_l + D_l - 7{,}000)$$

or

$$D_l < 1945.$$

The low risk group cannot be offered more than a 28 per cent (1,945/7,000) partial cover.

As we have seen in Example 4.6, insurance companies can overcome the situation in which the bad risks crowd out the good ones, by designing separate insurance policies for the different risk groups. Although the insurer cannot prevent a high risk client from buying a policy designed for a low risk client, it is possible to deter him by choosing the parameters of the policies carefully, and in particular by using a deductible in the policy designed for the low risks. Offering partial cover to the low risk group is only a partial (!) solution to the adverse selection problem since the low risk group would prefer a full cover policy. The low risk group still suffers from the existence of the high risk group but at least it is not priced out of the market anymore.

Principal-agent problems

Principal-agent problems are an important class of moral hazard problems in which one party, the principal, hires another party, the agent, to take certain actions. The agent generally has more information than the principal. In particular, the agent knows how much or how little effort he made in pursuing the principal's objectives. Since the principal and the agent have different objectives the agent has to be given appropriate incentives to act in the principal's interest in order to avoid or limit the moral hazard problem. In corporate governance, for example, the principals are the shareholders and the agents are the managers.

Designing appropriate incentive systems is of great practical importance. In the bank deposits insurance case, mentioned above, the government complements the provision of this insurance with regulation on minimum capital-adequacy standards. The European Union's Capital Adequacy Directive was passed in March 1993 and will be implemented at the end of 1995. This regulation gives the bank's owners incentives to avoid excessively risky investments since more of their own capital is at stake.[6] In Japan, doctors are known to over-prescribe drugs because their income is partly dependent on prescription. However, the health ministry, after several scandals involving deaths as a consequence of over-prescription, is rethinking the doctors' compensation system.[7]

In the standard context in which the principal-agent problem is studied, an employer hires a worker to do a particular job. Generally, the employer benefits from high effort levels but the worker dislikes providing much effort. A moral hazard problem arises: if the worker is not monitored or given appropriate incentives, he will shirk.

Effort can be observed

When the employer can observe the worker's efforts, or when there is a deterministic relationship between effort and performance, it is not difficult to find an incentive scheme which motivates the worker to provide the 'optimal' effort. Suppose the worker

[6] *See 'Still money in that franchise'; 'Flattened, sort of'*

[7] *See 'Keep taking the tablets'*

generates a profit P(e) for the employer if he works for e hours. The cost (monetary equivalent of his disutility) to the worker of working e hours is C(e) and, if the worker does not work for this employer, he can get an alternative job which gives him a utility (in money terms) of u. The employer chooses the effort he wants the worker to exert to maximise his profit minus the payment to the worker, taking into account the worker's reservation utility u and the fact that it should be in the worker's interest to make the effort chosen by the employer. The following simple payment schemes motivate the worker to provide the efficient amount of effort.

Payment based on effort

If the worker is paid based on his effort e according to $we+K$ (with w and K as constants), the employer's problem is: max $P(e)-(we+K)$. The employer has to take into account the **participation constraint** or the **individual rationality constraint** (i.e. the worker only works if he gets at least his reservation utility: $we+K-C(e)\geq u$). The employer has no reason to give the worker more than his reservation utility and so his objective becomes, after substituting the participation constraint, max $P(e)-(C(e)+u)$. The employer chooses the optimal effort e* such that the marginal profit of effort equals its marginal cost: $P'(e^*)= C'(e^*)$. The worker has to be encouraged to provide the optimal effort level which leads to the **incentive compatibility constraint**: the worker's net payoff $we+K-C(e)$ should be maximised at the optimal effort or $w=C'(e^*)$. We can conclude that the worker is paid a wage per hour equal to his marginal disutility of effort and a lump sum K which leaves him with his reservation utility.

Forcing contract

The employer could propose to pay the worker a lump sum L which gives him his reservation utility if he makes effort e^* i.e. $L = u + C(e^*)$ and zero otherwise. Clearly, the participation and incentive compatibility constraints are satisfied under this simple payment scheme. This arrangement is called a forcing contract because the employee is forced to make effort e^*. In 'payment based on effort' and 'franchise', the worker can choose his effort level.

Franchise

Suppose the worker keeps the profit of his efforts in return for a certain payment to the principal. This can be interpreted as a franchise structure. How should the 'employer' set the franchise fee F? The worker now maximises $P(e) - C(e) - F$ and therefore chooses the same optimal effort as before: e^* such that $P'(e^*) = C'(e^*)$. The principal can charge a franchise fee which leaves the worker with his reservation utility: $F=P(e^*) - C(e^*) - u$.

Effort cannot be observed

When the effort can be observed or when output or profit can be observed and there is a deterministic relationship between effort and output or profit, the moral hazard problem is easily solved as we have seen above. However, when the employer cannot observe effort the problem becomes more complicated. Suppose the employer can observe output but output is only stochastically related to effort. Payment based on effort is out of the question, so what about payment based on output? The employer could, for example, use the franchise solution. However, if the worker is risk averse, he will need to be compensated for taking on risk. Even when he makes a large effort, his profit could be low if he is a franchisee. The employer, on the other hand, is more likely to be risk neutral and willing to carry this risk. The franchise solution is inefficient here. I hope an example will clarify this.

Example 4.7

Assume the agent's (worker's) utility, as before, depends on his pay w and his effort level e. For simplicity let $U(w,e) = \surd w - e$. The principal (employer) is risk neutral. The agent can decide to shirk or not; shirking corresponds to $e = 0$ and not shirking corresponds to $e = 8$. The agent generates a revenue of £0 or £10,000 for the principal, with probabilities given in Table 4.1. You could think of the agent as a salesperson whose effort is important in getting a sale worth £10,000.

Table 4.1: Problems of generating low and high revenue

	revenue for principal	
	0	10,000
$e=0$ (shirk)	1/2	1/2
$e=8$ (work)	1/3	2/3

Of course, when the agent is not shirking, the principal's chance of earning the £10,000 revenue is higher. Suppose the agent's reservation utility is $u = 8$ then, if the principal could observe the agent's effort level, how much would he have to pay to elicit $e = 0$? For $e = 0$, the agent's utility in this job is $\surd w$ and for this to exceed $u = 8$ we need to pay him $w \geq 64$. If the principal wants to elicit effort level $e = 8$ which gives utility $\surd w\text{-}8$ to the agent, he needs to pay $w \geq 16^2 = 256$. You should be able to show that a risk neutral principal, who can observe effort, will elicit the high effort.

Suppose now that the principal cannot observe effort. An obvious method to reward the agent is to pay him based on his 'performance' (i.e. pay him x if the revenue to the principal is 0 and y if he secures a revenue of £10,000). Assume the principal cannot pay a negative amount ($x, y \geq 0$). How should the principal determine the appropriate payment scheme (x,y) if he wants to make the agent work? First of all, he wants to prevent the agent quitting his job or, in other words, the participation constraint should be satisfied:

$$(1/3)U(x,8) + (2/3)U(y,8) \geq 8, \text{ or}$$

$$(1/3)(\surd x - 8) + (2/3)(\surd y - 8) \geq 8. \qquad (1)$$

Also, given the compensation scheme (x,y) the agent should prefer working to shirking (i.e. the incentive compatibility constraint should be satisfied):

$$(1/3)U(x,8)(2/3)U(y,8) \geq (1/2)U(x,0)+(1/2)U(y,0), \text{ or}$$

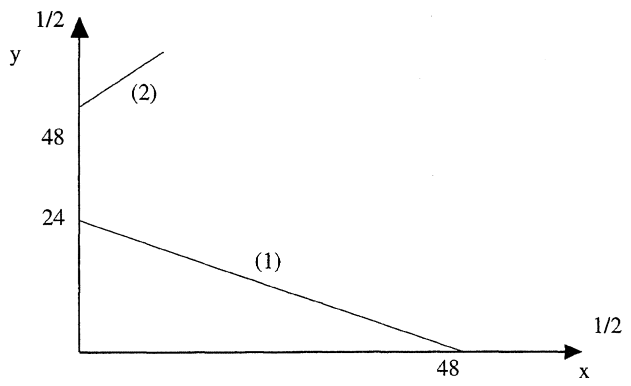
$$(1/3)(\surd x - 8) + (2/3)(\surd y - 8) \geq (1/2)\surd x + (1/2)\surd y. \qquad (2)$$

The constraints (1) and (2) can be rewritten as:

$$\surd x + 2\surd y \geq 48 \text{ and } -\surd x + \surd y - 48 \geq 0.$$

respectively. Therefore the area above the upwards sloping line in Figure 4.1 represents the set of feasible compensation schemes.

Figure 4.1: Feasible compensation schemes



The risk neutral principal will want to maximise his expected revenue which, given that the agent works, equals:

$$(1/3)(0-x)+(2/3)(10{,}000-y).$$

Maximising this is equivalent to minimising $2y+x$ which is achieved (over the region of feasible compensation schemes) at $\sqrt{x}=0$ and $\sqrt{y}=48$. This indicates that the principal should pay the agent only when he secures the £10,000 revenue and then the payment should be $48^2=2{,}304$. The principal thus expects a net revenue of:

$$(2/3)(10{,}000-2{,}304)=5{,}130.67$$

which exceeds his expected net revenue of inducing no effort:

$$(1/2)(0)+(1/2)(10{,}000)-64=4{,}936.$$

The principal will therefore motivate the agent to work.

The agent receives expected utility equal to his reservation utility of 8 when his effort can be observed. When effort is unobservable, his expected utility is $(1/3)(0-8)+(2/3)(48-8)=24$. His expected wage is also higher when effort cannot be observed: $(1/3)(0)+(2/3)(2304)=1536>256$. This higher expected wage is necessary to compensate the agent for carrying risk. Ideally the principal should carry all risk since he is risk neutral but, if the agent is paid a fixed wage, inefficiency due to moral hazard results.

Chapter summary

After this chapter and the relevant reading, you should understand:

- why asymmetric information represents a problem if we are concerned with efficiency the 'lemons' problem
- how using a 'deductible' can overcome moral hazard in the insurance context
- the difference between signalling and screening
- the relevance of the participation and incentive compatibility constraints in the **principal-agent** problem.

You should be able to:

- give examples (preferably your own) of **adverse selection** and **moral hazard**
- show how adverse selection in **insurance** can lead to low risk individuals being priced out of the market and explain the use of a **partial cover** insurance policy in this context

- explain Spence's **education as a signal** model and work out an example
- analyse a simple **principal-agent** problem.

Sample exercises

1. The value of a four year old Honda Accord to the owner is uniformly distributed between £3,800 and £5,800. Whatever the value to the owner is, the car is worth £200 more to a potential buyer. There are many potential buyers for a four year old Honda Accord. Find the equilibrium price.[8]

2. A travel agent sells holidays to the Cayman Islands. Consumers are willing to pay £1,400 if the holiday is excellent and £800 if the holiday is mediocre. The travel agent knows the quality of the holidays he sells but the consumer does not.

   a. If a fraction x of travel agents sell excellent holidays, how much is a risk neutral consumer willing to pay for a holiday?

   b. Suppose it costs a travel agent £1,000 to book a mediocre holiday and travel agents are perfect competitors. Is there a market equilibrium in which all holidays are mediocre?

   c. Suppose it costs a travel agent £1,000 to book a mediocre holiday and £1,100 to book an excellent holiday. Is there an equilibrium in which all holidays are excellent?

   d. Suppose it costs the travel agent the same amount, £1,000, to book a mediocre or an excellent holiday. Find the market equilibrium (equilibria). How much consumer surplus is generated at the equilibrium (equilibria)?

   e. For the scenarios of (c) and (d), if it were possible, would it be good competition policy to ban the sale of mediocre holidays?

   f. How can the inefficiencies resulting from asymmetric information be avoided in this context?

3. McClean & Co. and McDirty & Co. are in the entertainment business. McClean is interested in taking over McDirty. It does not know McDirty's value *v* under its current management but believes it to be between £3 million and £4 million, all values equally likely. The value *v* is known only to McDirty's management team who will decide on any takeover bid. Whatever the value *v* is, the company is worth *1.5v – £1.5 million* to McClean.

   a. Draw the value to McClean as a function of *v* and conclude that McDirty is always worth more in McClean's hands than under its current management.

   b. McClean has **one** chance to make a takeover bid *p* which McDirty will accept or reject. What is the expected gain to McClean of offering £3.5 million?[9]

   c. What is the optimal takeover bid?

4. Consider Spence's signalling model and allow workers to choose their number of years of education before they start working. Eighty per cent of the workforce is intelligent and highly productive whereas the remaining 20 per cent is not intelligent and unproductive. The intelligent workers are worth £50,000 per year to the employer and the others are worth £20,000 per year, which is what their salary would be if the employer could tell them apart. The cost of education is £10,000 per year for an intelligent student and £20,000 per year for a less intelligent student. You should add the opportunity cost of not working while studying to these numbers. Assume that the less intelligent workers decide to get zero education.

[8] ***Hint:*** *At any price p, which cars are offered for sale? What are they worth to a buyer?*

[9] ***Hint:*** *for which values of* v *will McDirty accept?*

a. Ignoring discounting, and assuming the expected length of the job is four years, how much education should the intelligent workers get for their education to be a credible signal of their productivity?

b. Show that everyone would be better off without signalling.

5. Martha has a disutility of effort function $C(e) = e^2/2$ (e is the number of hours she works) and reservation utility u=0. I am a risk neutral distributor of Finnish vodka and want Martha to work for me. For every hour she works, I make a profit of m on average. I can observe how many hours Martha works and thus pay her based on her effort.

   a. What is the optimal wage labour contract w(e) = we +K, where w is a per hour wage and K is a lump sum.

   b. Martha wants to buy the franchise from me. How much could I get her to pay for it?

   c. Next year Martha graduates and she will be able to get a better job. This will increase her reservation utility to u=2. How does that affect the answer to (a) and (b)?

6. A risk neutral principal hires a risk averse agent to do a job. The agent's utility function is $U(w,e) = \sqrt{w} - (e-1)$ where e is the agent's effort and w is his compensation and the agent has a reservation utility $u = 1$. The agent decides to work hard ($e = 2$) or shirk ($e = 1$). The principal's revenue depends on the agent's efforts but also on random factors outside the agent's control. The probabilities of obtaining revenue 10 or 30 are as indicated in the table below.

	Revenue	
	R = 10	R = 30
$e = 1$	2/3	1/3
$e = 2$	1/3	2/3

   a. Calculate the expected revenue if the agent works hard and if he shirks.

   b. If the principal could observe effort, how much would she have to pay to get the agent to work hard? to get the agent to make a low effort? What is her net revenue in either case? What is the optimal forcing contract?

   c. Suppose the principal can only observe revenue, not effort, and so has to pay the agent more for the high revenue than for the low revenue if she wants the agent to work hard. Write down the participation and incentive compatibility constraints. Draw these constraints and indicate feasible combinations of pay levels for high and low revenue.

   d. Write the principal's objective function as a function of the pay levels for high and low revenue. What is the optimal compensation scheme? Calculate the agent's expected utility and his expected wage and make a comparison with (b).

   e. Check that motivating the high effort is worthwhile for the principal.

Private and common value auctions; private value auctions and their 'optimal' bidding strategies; auction revenue; common value auctions; complications and concluding remarks

Chapter 5

Auctions and bidding

Texts

Varian, H.R. *Intermediate Microeconomics.* (New York:W.W. Norton and Co., 2006) seventh edition [ISBN 0393927024] Chapter 17.

References cited

Boyes W.J. and S.K. Happel 'Auctions as an allocation mechanism in academia: the case of faculty offices', *Journal of Economic Perspectives* (1989) 3(3): 37–40.

Capen, E., R. Clapp and W. Campbell 'Competitive bidding in high-risk situations', *Journal of Petroleum Technology* (1971) 23(1): 641–53.

Cassady, R., Jr. *Auctions and Auctioneering.* (University of California Press, 1967).

'Inside the stockade', *The Economist,* 2 April 1994, 69–70.

McAfee, R. and J. McMillan 'Auctions and bidding', *Journal of Economic Literature* (1987) 25(2): 699–754.

Milgrom, P 'Auction theory', in Bewley, T. (ed) *Advances in Economic Theory, Fifth World Congress.* (Cambridge: Cambridge University Press,1987). [ISBN 0521340446]

Milgrom, P. and R. Weber 'A theory of auctions and competitive bidding', *Econometrica* (1982) 50(5): 1089–1122.

'Speeding allowed', *The Economist,* 16 April 1994, 69.

'Teetering', *The Economist,* 27 August 1994, 7.

Wilson, R 'Auctions of shares', *Quarterly Journal of Economics* (1979) 93: 675–89.

'You say you want a revolution', *The Economist,* 8 January 1994, 27–28.

While the model of perfect competition involves many sellers competing to sell, auction theory analyses situations in which many buyers compete to buy from one seller. Auctions are a widely used alternative to the market mechanism for transferring goods from sellers to buyers as the following list of examples illustrates.

- Around 500 BC Babylonian women were auctioned off annually to prospective husbands. The most attractive ones were offered first and fetched high prices, while others were auctioned with large dowries made up from the sale of the attractive women.[1]
- The US Forest Service auctions timber rights.
- The US government regularly auctions off land which may contain oil.
- In sales of securities and bonds auction methods are used. The Treasury auctions off government debt issues to financial institutions such as banks, stock brokerages and insurance companies.
- Companies are sometimes sold through a formal bidding process which may involve an investment bank acting as the auctioneer.

[1] *Cassady (1967)*

- Boyes and Happel (1989) report that, in the department of economics at the College of Business, Arizona State University, office space was auctioned off to the faculty. The proceeds were used to set up a graduate scholarship fund. (The authors mention that the bidders didn't really care where the money went as long as the chairman didn't keep it!) Every interested faculty member submitted a sealed envelope with a bid before an announced deadline. The highest bidder then got first choice of offices, the second highest bidder could choose from the remaining offices, etc. Anyone who bid above $75 received a window office. The highest bid was $500, the next highest $250; a total of $3,200 was raised.
- In Singapore prospective buyers of new cars have to bid for a permit or Certificate of Entitlement (COE) before they make their purchases. Successful bidders pay the lowest accepted bid price. For example, if the quota for a particular category of cars (say 1,001 to 1,600 cc.) is 1,000 and the 1000th highest bid is S$600, then every successful bidder (with bid in top 1000) in this category pays S$600. The exceptions are company vehicles and heavy goods vehicles for which double the amount in their categories must be paid. There are exceptions for motorcycles as well. Their owners pay one-third of the 'open' category premium if they choose to enter this category (there is a special motorcycle category). Diplomatic vehicles are unaffected by the quota system.

In some auctions buyers bid to purchase items but sellers can also bid to sell a product or service. Many governments and large companies purchase almost exclusively by procurement through competitive bidding. Therefore you may have to quote a fee as a consultant for particular projects. The analysis of an auction is the same whether bidders bid to buy or to sell (except that in the latter case the lowest bidder wins). You should bear in mind that the purpose of bidding is not to win but to maximise your expected gain (i.e. to win only when you are better off by winning).

Given that auctions are so prevalent, we should consider the reasons for their use and the circumstances under which they are more appropriate than other methods of sale. Auctions are never used to sell standardised products for which there are competitive markets. The time and expense needed to organise an auction and gather all the interested parties in the same place would be wasted since the outcome would be the same as in the market. However, auctions are virtually the only mechanism used to sell special, unique items such as antiques, real estate, art, rare wine, oil drilling sites and mineral leases. It is very difficult to post a price for such items because of the uncertainty about demand for them; there are no historical data a seller can use to form an estimate about potential buyers' willingness to pay. Auctions are therefore used to determine prices for these special objects.

Private and common value auctions

In the theory of auctions a distinction is made between private and common value auctions. It is crucial that you understand this categorisation! The distinction rests on how the value of the auctioned object is modeled (i.e. which assumptions are made regarding how the potential buyers value the object). Take the (unrealistic) extreme case in which an item to be auctioned has a fixed value which is known to all bidders (e.g. a £50 note); the item will be sold to the highest bidder who pays his bid (we will see later that in some auctions the winning bidder does **not** pay his bid) and is allocated randomly if there is a tie. Then of course the Nash equilibrium[2] is for everyone to bid the known value (make sure you know why) — not very interesting! In all real-life auctions there is some uncertainty either about the value other bidders attach to the item or about the value of the item itself.

[2] *See Chapter 2, 'Game Theory'*

In a **private value** auction you know the value to yourself of winning the item (or the contract in a procurement situation) but you do not know its value to other competing bidders. Different bidders attach different values to the item. For example, you know that your reservation price for an antique clock is £12,000; another bidder may value it at £14,000. When you tender for a contract to build a tunnel you may know from past experience how much it would cost you to build it but you are unlikely to know exactly what your competitors' costs will be. In models of private value auctions it is often assumed that the valuations of bidders are drawn independently from the same distribution: the **IID** (independently identically distributed) **assumption.** In the procurement context this implies that costs for each company are taken as drawn from a common distribution. Auctions for artwork are considered private value auctions as long as bidders do not intend to buy for investment purposes. If they buy as an investment with the intention of resale, their private value is **not** independent of the other bidders' values.

In a **common value** auction you are uncertain about the value of the object to be auctioned but, whatever the value is, it is the same for all bidders. This would be the case for an auction for an unknown amount of money in a sealed envelope. Each bidder forms an estimate of the single uncertain value of the object. Typical examples of common value auctions are auctions in which a government sells the mineral rights to a plot of land. When oil field leases or gas drilling rights are auctioned, a bidder may not know how deep he will have to drill, how much oil or gas there is, what the future oil and gas prices are etc. but all these factors will affect the value of the oil field equally for all bidders. US Treasury bills are sold through common value auctions because the participating financial institutions resell the government debt. In a procurement context, the common value assumption is valid if all the bidders would incur the same costs to do the job but no bidder knows the precise cost. Some bidders may however have more information about the object than others: they may have more experience (for example, they have drilled in an adjacent site to the oil field on auction and can make assumptions about the site's characteristics). The important assumption in models of common value auctions is that the object has the same value to all bidders but each bidder forms his own estimate of this value.

In reality auctions can be a mixture of private value and common value auctions: the auctioned object will have a common value but it also has a buyer-specific value. For example, there may be several candidates interested in buying a particular company. The actual value of the company's assets is likely to be the same for all potential buyers. The company's activities, however, may offer varying degrees of synergy, depending on who the buyer is. Current auction theory studies this type of hybrid scenario with common and private values but the techniques used are quite advanced and beyond the scope of this subject guide. In discussing optimal bidding strategies, I will only refer to strictly private value and strictly common value auctions.

Private value auctions and their 'optimal' bidding strategies

In this section I present some types of auctions which are frequently used and discuss good ways of bidding in them. I will show what the profit-maximising bidding strategies are and how optimal bids depend on bidders' estimates of the valuations and on the number of bidders. Table 5.1 summarises the characteristics of the auction forms discussed in this section. In all auctions bidders try to choose the best strategy knowing that their competitors are also rationally trying to optimise. This of course leads to game theoretic analysis; auctions are very good examples of games of incomplete information: each bidder has private information about the value of the object being sold.

Table 5.1: Summary of private value auctions

Private value auctions	How bids are made?	What does the winner pay?	Optimal strategy
English	auctioneer starts with low bid, bidders make increasingly higher bids	the highest bid	stay in the auction until the bidding reaches your valuation
Standard (or first price) sealed bid	bids are submitted secretly and simultaneously	the highest bid	bid below your valuation
Dutch	auctioneer starts with high price and reduces it gradually until a bidder stops him	the price at which the auctioneer stopped	bid below your valuation
Second price sealed bid	bids are submitted secretly and simultaneously	the second highest bid	bid your true valuation

Private value English auctions

English auctions, or **oral ascending bid** auctions, are the most commonly used auction format, accounting for more than 75 per cent of auctions in the world.[3] Although the term English auction may conjure up images of Christie's or Sotheby's salerooms filled with art connoisseurs bidding millions of pounds for a Monet, use of this type of auctions is certainly not restricted to these glamorous settings. Many agricultural products, fish, repossessed houses and cars are also sold in oral auctions. These auctions follow a specific format: the auctioneer starts by setting a low price and asking for bids. As long as interested bidders remain, the price rises through competitive and increasingly higher (verbal) bids until there is only one bidder left. This highest bidder wins the object. It is important to realise that this highest bidder pays an amount which is only slightly higher than the last bid made by the second highest bidder. A Japanese version of this auction involves an electronic system which displays a rising price. All bidders who remain interested at the displayed price keep a button pressed. When the price gets too high for a bidder, he releases the button and, when there is only one bidder left, the price at which the second to last bidder drops out is displayed with the identity of the winner.

How should you bid in a private value English auction? Imagine you are actually in an auction room, bidding for a desired object. You have a reservation value v: this is your private valuation of the object. You are indifferent between acquiring the object at a price equal to your reservation value and not acquiring it. The bidding process starts at a bid below v. Will you make a bid? Suppose the bidding is quite competitive and the bid price rises to v. Will you make another bid? The optimal strategy, which you may have arrived at yourself, is given in the box below.

[3] *See Cassady (1967)*

The optimal strategy in a private value English auction is to stay in the auction until your valuation is reached.

In fact, this optimal strategy is **dominant**: whatever strategies the other bidders use, you cannot do better than follow the above strategy. The explanation is as follows. If you win the auction, your payoff is the difference between your valuation and your bid; if you do not win, your payoff is zero. Your only decision is when to drop out. If you drop out before the bid price reaches your valuation, you forego the opportunity of a positive payoff. If you stay in the auction after the bid price has exceeded your valuation, you risk obtaining a negative payoff (if you win the auction), and the highest payoff you can get is zero. **Therefore under no circumstances can you improve on dropping out at your valuation**. This points to a possible explanation for the popularity of English auctions: very little information gathering and preparation costs are involved since bidders' optimal strategies do not depend on how their competitors bid.

From the nature of the optimal strategy it follows that, in an English auction, the object goes to the person who values it most. He will pay a price approximately equal to the second highest reservation price since at this price his only remaining competitor drops out. The price is in reality slightly above the second highest valuation because bids are raised in discrete steps. In some real-life auctions the auctioneer determines the next possible bid on the basis of his perception of the competitiveness of the bidding. The bids may be raised by pennies or £100,000, depending on the context. (However, the issue of discrete bid raises is ignored in virtually all auction models.)

Private value standard sealed bid auctions

In a standard sealed bid auction, bids are invited from interested bidders and remain secret until a preannounced date on which they are revealed simultaneously. In the usual **first price** version of the sealed bid auction, the winner is the bidder who bids the highest amount and he pays his bid. Of course in the case of procurement or government contracts, where this type of auction is used almost exclusively, the winner is the one who offers the lowest price.

How should you bid in a first price sealed bid auction? Obviously, you shouldn't bid above your valuation (because you could get a negative payoff). If you bid your valuation, your payoff is zero. It is easy to see that you can do better if you bid below your valuation. When you bid below your valuation you are in fact trading off larger potential gains (valuation minus bid) against a reduced probability of winning the auction. This implies that there is no dominant strategy here because the reduction in the winning probability depends on your competitors' bids.Your equilibrium bid should take into account your estimate of your competitors' valuations and of their bids. The optimal bid is the one which maximises your expected gain given your probabilistic information about equilibrium competing bids. At the equilibrium each bidder shades his bid below his valuation. It turns out that the general (Nash equilibrium) rule for private value standard sealed bid auctions is as follows.

The optimal strategy in an IID private value first price sealed bid auction is to bid the expected value of the second highest valuation among the bidders assuming your valuation is the highest.

I will not give a proof of this result but I will show you how it works in the special case of uniformly distributed valuations. Suppose there are two risk-neutral bidders: you and an opponent. Each of you has a valuation which is known only to you and which is drawn independently from a uniform distribution on [0,1]. Your valuation is v_1 and you bid b_1. Suppose your opponent's strategy tells him to bid a fraction f of his valuation v_2

(i.e. $b_2 = f v_2$). Would you ever bid above f? If you bid above f you are sure to win since $b_2 < f$ for $v_2 < 1$. If you bid below f your probability of winning is the probability that the other bid is below yours:

$$P\,(\text{win}) = P(b_2 < b_1) = P(f\,v_2 < b_1) = P(v_2 < b_1 / f) = b_1 / f.$$

Hence your expected gain from a bid b_1 equals:

$$(v_1 - b_1)\,P(\text{win}) = (v_1 - b_1)\,(b_1/f).$$

Note that there is a tradeoff between the probability and the profitability of winning: if the bid increases, the profit $(v_1 - b_1)$ decreases but the probability (b_1/f) of winning increases. Optimising the expected gain with respect to b_1 gives $b_1 = v_1/2$. This optimal bid is independent of f: an optimal response to any linear bid is to bid half your valuation. It follows that the strategy pair in which bidders bid half their valuations forms a Nash equilibrium. How does this relate to the general rule above? Well, if you assume your value v_1 is the highest, then your expectation of the other bidder's value is equal to $v_1/2$.

Let us extend the analysis above to a sealed bid auction with n bidders, each with a valuation drawn from the uniform distribution on [0,1]. Assume you are Bidder 1 and all your opponents use the same linear bidding strategy: $b_i = f v_i$, $i = 2,\ldots,n$. You win the auction if all of your rivals' bids are below yours, that is, when $b_i = f v_i < b_1$ for all i:

$$P(\text{win}) = P(v_2 < b_1 / f \;\&\; v_3 < b_1 / f \;\&\ldots v_n < b_1 / f) = (b_1 / f)^{n-1}.$$

Hence your expected gain from a bid b_1 equals:

$$(v_1 - b_1)P(\text{win}) = (v_1 - b_1)\,(b_1 / f)^{n-1}.$$

Optimising the expected gain with respect to b_1 leads to $b_1 = ((n-1)/n)v_1$. The equilibrium strategies are therefore to bid (n-1)/n of your value. You should be able to derive the equilibrium bidding strategy for valuations from a uniform distribution on [L,U]. The answer is $b_i = (1/n)\,L + ((n-1)/n)\,v_i$. We have of course assumed that each bidder knows how many bidders there are. This is an unrealistic assumption if the number of bidders is very large but it is precisely in this case that uncertainty about the number of bidders is not a serious problem: the optimal bidding strategy is to bid close to your valuation.

It is very important that you realise what is meant by an 'equilibrium strategy'. As we have seen in the game theory chapter, an equilibrium strategy is only optimal against equilibrium strategies. It is the best you can do **if** the other bidders play their equilibrium strategies. If you have reason to believe that your rivals are not using an equilibrium strategy, your optimal response is likely to differ from the equilibrium strategy. In general, to find an equilibrium bidding strategy, look for best responses (i.e. strategies which tell the players what to bid given their valuations and the other players' bidding strategies). With IID valuations there is always a symmetric equilibrium, in which each player bids an increasing function of his value. As a consequence, the bidder with the highest valuation will make the highest bid and win the object.

As you can see from the results above, as the number of bidders increases, each bids a higher fraction of his value and the final price goes up. It follows from this analysis that it is in the seller's interest to get as many bidders as possible since the expected maximum value is increasing in n and each bidder bids close to his value if n is large. The effect of the number of bidders on the expected revenue has become a relevant consideration in determining appropriate antitrust policy with respect to takeovers and restructuring in the defense industry. The government cannot expect to fulfill its military requirements at a low cost if there are only one or two potential suppliers for any type of weapon.[4] However, this lack of competition among suppliers could be counteracted by the government's monopsony power.

[4] *See 'Inside the stockade'*

Private value Dutch auctions

In a Dutch auction the auctioneer starts at a very high price and reduces it gradually until one of the bidders shouts 'mine!'. The winning bidder then pays that price. This is called a Dutch auction because it is the form of auctioning used in the Netherlands for the wholesale of fresh flowers. A mechanical version of the Dutch auction involves a giant clock and a switch for each bidder. As the clock ticks, the price goes down continuously until one of the bidders uses his switch to stop it. The price and the identity of the winning bidder are then displayed. This version of the Dutch auction is used in Ontario tobacco auctions.[5]

> The optimal strategy in a private value Dutch auction is same as for a first price sealed bid auction.

In essence, the Dutch auction and the first price sealed bid auction describe the same game. Each player's strategy consists of a function of his valuation. The only (immaterial) difference is that, in a sealed bid auction, the bid is submitted in writing and a Dutch auction is an oral auction. In a Dutch auction, the bidding strategy can be interpreted as the plan to claim the item if the bid level goes down to the number predetermined in the bidding strategy. A bidder in a private value Dutch auction does not gain anything from his presence; he could send an agent to do the bidding.

Although Dutch and English auctions are both oral auctions, they are very different from the bidder's perspective. As mentioned before, in an English auction, as the price increases, the bidder can assess at any time whether he can gain by increasing the bid level. In particular, when the bid level reaches the bidder's valuation it is clear that he cannot make a gain from the auction. In a Dutch auction, the bidder is not certain at any price (below his valuation) whether he can increase his gain by waiting or bidding. He knows that he will win the auction if he bids but he could get a larger surplus (valuation minus bid) if he waits.

Private value second price sealed bid auctions

The second price sealed bid auction is the same as the standard sealed bid auction where the winner is the highest bidder but now the winning (highest) bidder will pay the second highest price bid. For example, if A bids £4,000, B bids £5,000 and C bids £6,000, C would be the winner of both a first price auction and a second price auction. In the first price auction C would have to pay £6,000 whereas in the second price auction he would have to pay £5,000. Surprisingly, the second price sealed bid auction is easier to analyse than the first price sealed bid auction.

> The optimal strategy in a private value second price sealed bid auction is to bid your true valuation.

Note that we found before that this is also the optimal strategy in an English auction. Bidding your valuation is a dominant strategy, as it is for the English auction. Why? Suppose you bid **below** your valuation. Then you will gain zero if you don't win. If you win, then you would have won bidding your true valuation and paid the same amount (the second highest bid). So, bidding below your valuation is dominated by bidding your valuation since bidding lower only decreases your chances of winning and obtaining a positive payoff. Now suppose you bid **above** your valuation. If you lose, you would certainly have lost bidding your valuation (your bid would have been lower). Hence you did not do better by bidding higher than your valuation in this case. If you win and the second bid is above your valuation you will get a negative payoff. You do worse than by bidding your valuation. If you win and the second bid is below your valuation, you would have won bidding your valuation as well. Bidding above your value is dominated by bidding your value since bidding above only increases your probability of winning the auction when you don't want to win it.

[5] *See Cassady (1967)*

Note that the optimality or dominance of the strategy 'bid your valuation' does not depend on knowing other bidders' valuations or their distributions. In other words, whatever the other bidders' strategies are, you cannot do better than bidding your valuation. As was mentioned for the English auction, the dominance of the optimal bidding strategy makes second price sealed bidding an attractive auction form. Bidders do not have to gather any information; they do not in fact have to think strategically. They just have to be rational enough to realise that they should bid their valuation.

Auction revenue

An important issue for anyone considering the sale of an object at an auction is how much revenue can be expected from such a sale and which type of auction generates the highest revenue. Let us first look at the revenue to the seller from first and second price sealed bid auctions using the example of independent, uniformly distributed valuations. We will use the following property of the uniform distribution. For a uniform distribution on [0,1], the expected value is 1/2; if you make two independent draws from the distribution, the expected value of the largest is 2/3 and of the smallest 1/3; in general, if you make n independent draws, the expected value of the largest is n/(n+1), the second largest has expected value (n-1)/(n+1) and so on; the minimum has expected value 1/(n+1).

First price sealed bid revenue

Recall that, in the two-bidder uniform valuation case, the expected winning bid is half the expectation of the highest value (i.e. (1/2)(2/3)=1/3). In the n bidder case, the expected revenue is a fraction (n-1)/n of the winner's valuation or (n-1)/n of the expected value of the maximum: that is, ((n-1)/n)(n/(n+1))=(n-1)/(n+1).

Second price sealed bid revenue

What is the expected revenue to the seller if the two bidders bid their valuation (which they are supposed to do as their optimal strategy) in a second price sealed bid auction? The seller gets the lowest bid, so if there are two bidders with valuations drawn independently from a uniform distribution on [0,1] he will expect a revenue of 1/3, the same as in a first price sealed bid auction. For n bidders, he will expect to get (n-1)/(n+1), the expected second highest valuation. So the expected revenue is again the same as for the first price sealed bid auction. In fact, this conclusion holds quite generally:

> The revenue equivalence result: Irrespective of the distribution of values, in a private values IID auction for the sale of one item and risk neutral bidders, a first price sealed bid auction (or a Dutch auction) and a second price sealed bid auction (or an English auction) yield the same expected revenue. The expected revenue in these auctions is the expected second highest valuation. Each of these auctions is an optimal auction (i.e. it produces the maximum revenue of all possible auction methods).

Common value auctions

Common value English auctions

Private value auctions are relatively easy to analyse because the bidders have all the necessary information at the start. They know their valuation of the object for sale and there is nothing to learn. In a **common value** auction, in contrast, where bidders are uncertain about the value of the item for sale, and where the item has the same value independent of who acquires it, bidders in English auctions can learn by observing each other's behaviour.

In particular, a bidder can observe how many active bidders there are at any price and when they drop out. This gives some indication of their estimates. Sometimes the identity of the bidders conveys useful information. This is the reason why an expert buyer, whose presence indicates that a particular item is desirable, may want to hire an agent to do the bidding for him. If he were to bid himself, other bidders might revise their value estimates upwards and bid more competitively.

There is no dominant strategy for a common value English auction and the equilibrium bidding strategy depends on the structure of the information among the bidders (who knows what?) and the possibilities for obtaining additional information about the value during the auction.

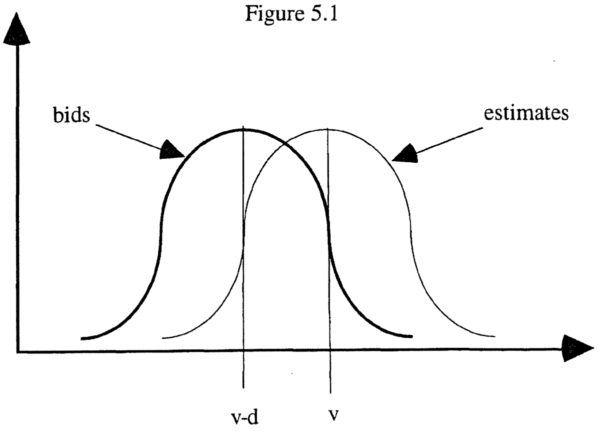
Common value first price sealed bid auctions

In a **common value** first price sealed bid auction, as in all common value auctions, each bidder has his own estimate of the value. If bidders bid according to their estimates, then the most optimistic bidder (the one with the highest estimate) wins the object. However, he is very likely to have overestimated the value of the object. This is the **winner's curse:** 'If you lose, you lose and if you win you also lose!'

To avoid the winner's curse you should bid below your estimate.

The winner's curse phenomenon is illustrated in the Figure 5.1 below. Assume that the probability density function of bidders' estimates has the actual value of the object as its expected value. This means that bidders' valuations are on average correct but half of the bidders overestimate the value. Given this large probability of overestimation, bidders will be cautious and reduce their estimates when determining the amount to bid. In the figure they all reduce their estimate by an amount d. This gives the distribution of bids on the left in the figure. Depending on the discount d the most optimistic bidder may still suffer from the winner's curse. If the discount is too small the highest bid still exceeds v, as it does in the figure. As d increases, the bid curve moves further to the left and the chances that the winner has bid too much become smaller.

Figure 5.1



In competitive procurements there are frequent occurrences of the winner's curse. If you get the consultancy job you tendered for, you are likely to have underestimated the costs. In common value competitive bidding for, say, an offshore oil lease, you have to correct for the possibility of overestimation. Capen, Clapp & Campbell (1971) analyse bids for offshore oil tracts, auctioned by the US government in the late 1960s. They present evidence that the winners may have been cursed! For example, the winning bid for offshore Texas Tract 506 was US$43.5 million, nearly three times the second highest bid of US$15.5 million. Failure to properly discount their estimates may partly expain why some independent television firms are in trouble, having paid too much for their franchises which were sold by the UK government in a bidding process in 1991.[6]

[6] *See 'You say you want a revolution'*

If there are many bidders, you have to shade your bid even more in order to account for the winner's curse phenomenon. To see this, suppose each bidder estimates the value as its real value plus some (positive or negative) error and bids his estimate minus some discount. Then the probability of the bid of the winner exceeding the value is larger when there is a large number of bidders since the probability that the maximum of n IID variables exceeds any number increases in n. This is an interesting observation because there is a natural tendency to bid more agressively when there are many bidders.

Note that I haven't told you how to bid in a common value first price sealed bid auction other than that you should discount your valuation. This is because no simple optimal bidding strategy exists. How you should bid depends on the nature of uncertainty about the value and the information available to the bidders. As for the private value versions, the optimal bidding strategy in a common value Dutch auction is the same as in a common value first price sealed bid auction.

Common value second price sealed bid auctions

For the **common value** version, the optimal strategies in the second price sealed bid auction are different from the English auction. In the English auction there is an opportunity to learn from observing other bidders which does not arise in a sealed bid auction. You should not bid your expected value based on your estimate alone (for winner's curse reasons) and you should take into account that you win if your estimate is highest. You can expect higher bids in a second price sealed bid auction because the price which is paid is lower than the winning bid.

Complications and concluding remarks

The theory of auctions is a fascinating field which has grown dramatically over the last two decades, in parallel with the advance of game theoretic modelling in economic theory. I have tried to give you an introduction to the more accessible material. In this section I briefly mention some additional issues which are of relevance to real-life auctions.

Phantom bids

As mentioned above, bidders in common value English auctions obtain information by observing who drops out and when. If the number of active bidders decreases rapidly, bidders are likely to think they have overestimated the value and revise their estimate downwards. This explains the temptation by sellers to use agents who keep bidding against the last bidder. The disadvantages of using agents in this way are that (a) the seller incurs a risk of the high bidder dropping out and one of his agents 'winning' the auction and (b) bidders may account for the possibility that they are bidding against an agent and shade their bids accordingly. Clearly, using phantom bids skillfully is an art!

Can the auctioneer use phantom bids profitably in a private value English auction? Yes, he can. Recall that the winner pays the second highest valuation. If the auctioneer takes a phantom bid after all but one 'real' bidder have dropped out, he increases the seller's revenue. The larger the difference between the highest and the second highest valuation, the more profitable the use of phantom bids will be.

In a second price (private or common value) sealed bid auction, an unscrupulous seller could 'cheat' by inserting a phantom bid just below the winning bid, thus increasing the price paid by the winner. This may partly explain the unpopularity of the second price sealed bid auction. Of course, the seller could only profitably cheat in this way if the bidders are naive. If the bidders account for the possibility of a dishonest seller, they will shade their bids and they will tend to bid as in a first price sealed bid auction.

Do you know why? Think about what the winner pays if the seller uses a high phantom bid.

Collusion

In some auctions it is possible for bidders to collude. For example, a 'ring' of bidders in an English auction could agree not to bid against each other (i.e. not to bid when the current bid was made by one of the ring members). In this way the ring can buy the object for less than would be possible without the ring. The object is then sold in the ring to the highest bidder and the profit is divided among the ring members. Rings are more likely when bidders know each other. In 1994, property developers in Hong Kong were accused of collusion in government land auctions.[7]

Although bidders could also collude in sealed bid auctions, the anonymity of the bidding process makes cheating (i.e. breaking the collusive agreement) more likely. This may be the reason why sealed bid auctions are preferred to English auctions in procurement. Especially in auctions which are not one-off events but are repeated very regularly with more or less the same bidder population (for example Treasury auctions), collusion is likely. To counteract collusion it is advisable for the seller to withhold information such as the identity of the winner and the winning bid. This eliminates the possibility of colluders punishing the cheaters which in turn increases the probability of cheating.

Risk aversion

In deriving optimal strategies we have implicitly assumed that bidders are risk-neutral. Would we find different results if **bidders** are assumed to be **risk averse**? The answer to this question depends on the type of auction under consideration. In a private value English or second price sealed bid auction the dominance of bidding your true valuation still applies. The argument given for this dominance does not assume anything about the bidders' attitude to risk.

However, risk averse and risk neutral bidders **do** bid differently in a private value first price sealed bid auction and a Dutch auction. When the clock in a Dutch auction is ticking away, indicating lower and lower prices, a risk averse bidder will become increasingly nervous as the price moves below his value. He is likely to stop the clock earlier than a risk neutral person with the same value. A risk averse person bids **higher** because he is willing to pay more to avoid the zero payoff associated with losing. If this does not make sense to you intuitively, think of the auction as a lottery in which there are two possible payoffs: a surplus equal to the value of the item minus the bid if you win and zero if you lose. The probability of winning depends on your bid. If you are risk averse you may prefer a low positive payoff with a high probability to a high positive payoff with a low probability. It follows that a (risk neutral) seller can make more revenue in a first price sealed bid auction than in a second price sealed bid auction if the bidders are risk averse, hence the qualification referring to risk neutrality in the revenue equivalence result.

Example 5.1

Consider a private value first price sealed bid auction with two risk averse bidders who have identical utility functions $U(x) = x^{1/2}$ and values drawn independently from the uniform distribution on [0,1]. Recall that the Nash equilibrium strategy for risk neutral bidders is to bid half their valuation. As before, assume Bidder 2 uses a linear bidding strategy: $b_2 = fv_2$. The probability that Bidder 1 wins with a bid b_1 remains b_1/f. Bidder 1 maximises his expected utility $U = (b_1 / f)\ (b_1 - v_1)^{1/2}$ which results in $b_1 = 2/3\ v_1$. The best response against any linear bid function is $b_1 = 2/3\ v_1$. We have therefore shown that, for risk averse bidders with a square root utility of money function, bidding 2/3 of their valuation is a Nash equilibrium strategy. The seller gets 2/3 of the expected maximum valuation (2/3), that is 4/9, which is higher than the expected revenue of 1/3 in the risk neutral version.

[7] *'See 'Speeding allowed'*

A **risk averse seller** will also prefer first price over second price sealed bid auctions. The reason is that the variance of the winning bid is larger in the latter auction form. Example 5.2 (which can be skipped if you are frightened of integrals!) illustrates this.

Example 5.2

Consider an auction with uniform valuations and risk-neutral bidders. We know that in a **first price** auction the winning bid is half the maximum value. To determine the variance of the winning bid we therefore need to look at the distribution of the highest value. For any number x in [0,1]:

$$\text{P(highest value} < x) = \text{P(both values} < x) = x^2.$$

This implies that the distribution function of the maximum is $F(x) = x^2$ and its density function is $f(x) = 2x$. You should check that the expectation of half the highest value is:

$$\int_0^1 (x/2)f(x)dx = 1/3$$

To find the variance calculate the second moment:

$$E(x^2) = \int_0^1 x^2 f(x)dx = 1/2$$

The variance of the maximum value equals $E(x^2)-(E(x))^2 = 1/18$ (since $E(x) = 2/3$) and so the variance of the winning bid is $\text{Var}(x/2)=\text{Var}(x)/4 = 1/72$.

In a **second price** auction with two bidders, the winning bid is the minimum (second highest) value. We therefore have to find the distribution of the minimum valuation:

$$\text{P(lowest valuation} > x) = \text{P(both values} > x) = (1-x)^2$$

$$\text{so that } F(x) = 1-(1-x)^2 = -x^2 + 2x.$$

The density of the minimum valuation is then $f(x)=2 - 2x$ and you can check that the expected minimum valuation is 1/3. This confirms that the seller expects the same revenue in the private value second price sealed bid auction as in the private value first price sealed bid auction. The variance of the minimum value can be calculated as

$$E(x^2)-(E(x))^2 = \int_0^1 x^2 f(x)dx - (1/3)^2 = 1/6 - 1/9 = 1/18$$

The variance of the selling price in the second price auction is four times higher than in the first price auction.

Sequential auctions

In most auction models it assumed that the sale of an item can be considered in isolation. In reality however, many auctions are of a sequential or repeated nature (e.g. the government may have an annual round of procurement tenders for stationery; several art objects may be auctioned on the same day). In some procurement auctions the quantity auctioned is not fixed but can be determined by the buyer (i.e. the buyer solicits bids and his quantity demanded is given by a demand function q(p) where p is fixed by the auction). In some share auctions a bidder bids a price and a quantity he wants to buy. The winner will be the bidder who offered the highest price who will get the quantity he bid. If there are any shares left, the bidder who offered the second highest price will get the quantity he bid at the price he offered and so on. Bidders in share auctions may be allowed to submit several price-quantity bids. Conceptually this is equivalent to submitting a demand curve.

I have already mentioned that repeated auctions are more likely to lead to bidder collusion but there are other complications. The value of a good to a potential buyer may depend on whether he has been able to acquire complementary goods in an earlier auction. Imagine a property developer interested in purchasing a substantial chunk of land to build a new entertainment complex. Suppose the land is sold in relatively small land to build a new entertainment complex. Suppose the land is sold in relatively small plots. If the bidder is not successful in the early bidding rounds, his valuations for plots coming up for sale later on may decrease dramatically. In particular, the value of any plot may depend on whether he is able to acquire the adjoining plots. Similarly, a construction company, with limited capacity, values winning a road maintenance contract according to its available capacity which depends on the number of contracts it has won recently.

Another consideration in sequential auctions is the effect of winning in an early round on your ability to bid later on. If you have a fixed budget or are not able to borrow easily, winning the auction for the modern painting may leave you resourceless and hence without a chance to win the silverware auction later on. Other bidders may realise this and may be able to use this fact to their advantage. By bidding against you to ensure you pay a large amount of money for the painting, they hope to ensure themselves a good deal in the silverware auction.

Other factors

Procurement contracts are not always awarded on the basis of price alone. In many instances a buyer will consider quality issues as well. He could do this either when the bids are opened (if he has asked bidders to specify quality variables) or by preselecting the suppliers who are allowed to tender. For reasons of public accountability, governments and other public institutions are more likely to use the latter option, especially if quality is not easily assessed objectively.

Time may be an important consideration in building contracts. When the Santa Monica freeway was damaged in the 1994 Los Angeles earthquake, the California Department of Transportation (Caltrans) scrapped its normal bidding procedures and construction firms were required to bid on both the cost and the time needed to get the repairs done.[8]

Conclusion

I have only discussed a few types of auctions but many more exist. For the more advanced and interested reader I recommend Wilson (1979), Milgrom and Weber (1982), Milgrom (1987) and McAfee and McMillan (1987) as further reading. Even for the auctions described in this chapter there are many variants and the assumptions which are made to enable us to analyse these auctions would have to be modified to deal with each variant. For example, it is quite common for the seller to set a reserve price below which he is not willing to sell. This reserve could be several millions of pounds for an Impressionist painting for example. An astute auctioneer or seller may be able to generate a higher expected revenue by carefully selecting a reserve price. In the models considered here, the number of buyers is fixed but in reality it is endogenous because potential suppliers trade off the cost of preparing a bid (which may include an 'entry fee') against their expected payoff from participating in the tender. It is fair to say that game theory has offered significant insights into the mechanisms of auctions and strategies for rational bidders. At the same time, many practical issues are ignored in current auction theory which limits its value for real life bidders.

[8] *See 'Speeding allowed'*

Chapter summary

After reading this chapter and the relevant reading, you should understand:

- the difference between **private** and **common value** auctions
- the structure of the auctions discussed and the optimal bidding strategies
- the **IID** assumption on buyers' valuations
- the effect of the **number of bidders** on the auction revenue and the bids
- how a **bidders' ring** works
- the **revenue equivalence result**
- the **winner's curse**
- the effect of **buyer risk aversion** in a first price sealed bid auction
- the effect of **seller risk aversion** on his preference between first and second price sealed bid auctions.

You should be able to:

- show why bidding your true valuation is a **dominant strategy** in English and second price sealed bid auctions
- derive the optimal bidding strategy in first price sealed bid auctions with uniformly distributed private values
- analyse simple auctions of the type discussed.

Sample exercises

1. A house which is currently rented is going to be sold. There are two potential buyers: the current tenant and an outsider. The owner solicits a price from the outside buyer and allows the tenant to buy the house if he matches this offer. The buyers and the owner know that the two buyers' values for the house are independent of each other and that for each buyer the value is drawn from a uniform distribution on [£100,000; £160,000].

   a. How should the tenant decide whether or not to match the outside offer?

   b. What offer should the outside buyer make if her value is £120,000? What offer should she make as a function of her value v?

   c. (Only attempt this part if you feel brave!) What is the seller's expected revenue? Assuming there are two bidders, what would the seller's expected revenue be under an English auction?

2. In a private value first price sealed bid auction there are two bidders with values v_1 and v_2 drawn independently from a uniform distribution on [0,1]. Assume bidder 2's strategy is to bid $b_2 = v_2^{1/2}$. Show that the optimal response of bidder 1 is to bid $b_1 = 2/3\ v_1$. Do these strategies form a Nash equilibrium?

3. I have a beautiful picture of the famous ballerina Belle Taper which I will auction in class. I charge a £2.50 auction entry fee. You cannot bid unless you pay this fee. You can only see the picture after you have decided to participate in the auction (and paid!) You should not assume that you will derive any pleasure from seeing the picture, only from possessing it. After you have seen the picture you will know your valuation of it. Everyone's prior distribution of their and any other bidder's valuation is uniform on [£0, £15]. The auction will be of the first price sealed bid type. Would you participate in the auction if one other bidder participates? What if two other bidders participate? What if the auction is of the second price sealed bid type?

4. a. 'A second price sealed bid auction is better for the seller than a first price auction because in a second price auction the buyer pays less than his bid and will therefore bid higher.' True or false?

   b. 'A first price sealed bid auction is better for the seller than a second price sealed auction because in a first price auction your optimal bid depends on what your opponent bids. This makes the bidding more competitive.' True or false?

5. Suppose there are four bidders, each with a private value uniformly distributed on [0,1]. What is the expected price in an English auction? in a first price sealed bid auction?

6. Steven and Robert are 2 risk neutral potential buyers of a dingelhopper. Each buyer values the dingelhopper at an amount v_i known only to him. This valuation is considered by everyone, including the seller, to have been drawn from a given distribution, uniform on [0,10] for Steven and uniform on [10,30] for Robert. The draws for the two individuals are independent.

   a. Who will win in an English auction? What is the expected selling price?

   b. Suppose the seller can impose a minimum bid level of 10. What is his expected revenue in this case? What if he sets the minimum bid level at 15? What is the optimal minimum bid level?

Notes

Reviewing consumer choice; state-contingent commodities model; intertemporal choice; labour supply; risk and return

Chapter 6

Topics in consumer theory

Texts

Varian, H.R. *Intermediate Microeconomics.* (New York: W.W Norton and Co., 2006) seventh edition [ISBN 0393927024] Chapters 2, 3, 4, 5, 6, 8, 9, 10, 12, 13, 14 and 15.

References cited

Battalio, R.C., L. Green and J.H. Kagel 'Income-leisure tradeoffs of animal workers', *American Economic Review* (1981) 71(4): 621–32.

Biddle, J.E. and D.S.Hamermesh 'Sleep and the allocation of time', *Journal of Political Economy* (1990) 98(5): 922–43.

Gravelle, H. and R. Rees *Microeconomics.* (Harlow: FT Prentice Hall, 2004) third edition [ISBN 0582404878] Chapter 3 and 19: you may find this treatment too theoretical and I do not expect you to study the material there.

Solberg, E.J. *Microeconomics for Business Decisions.* (Lexington: D.C. Heath and Co., 1992) [ISBN 0669167053] Chapters 4, 5, 6 and 7.

Markowitz, H. *Portfolio selection.* (New York: John Wiley, 1959).

'Buyers and builders', Asia Survey. *The Economist,* 30 October 1993a, 17–20.

'Giving the economy a fix', *The Economist,* 10 April 1993b, 90.

'Hard Labour', *The Economist,* 11 September 1993c, 30.

'Sharing the burden', *The Economist,* 13 November 1993d, 18–20.

'The home front', *The Economist,* 11 December 1993e, 31–32.

'New-found plans', *The Economist,* 7 January 1994a, 79.

'Risk and return', *The Economist,* 19 February 1994b, 137.

In this chapter my aim is to guide you through some fundamental concepts in the theory of the consumer. You have studied many of these concepts in your introductory course so I will leave it up to you to review some of the material there.

Consumer theory deals with the individual consumer's choice of how to spend his income. The standard model of consumer choice assumes that consumers maximise a utility function subject to a budget constraint. In addition to a brief review of the theory of the consumer, I will show you some of its important applications to finance, welfare economics, uncertainty, intertemporal decision-making and labour supply.

Reviewing consumer choice

Utility, budget constraint and demand[1]

Make sure you know what is meant by the following terms:

- perfect complements (L-shaped indifference curves)
- perfect substitutes (linear indifference curves)
- bads.

[1] *Read Solberg (1992) Chapter 5; Varian (2006) Chapters 2, 3, 4, 5, 6*

You should know how the budget constraint is affected by taxes, quantity discounts and vouchers. A voucher is an income transfer which can be spent on a particular good.[2]

You should remember how demand for a good can be derived from the utility maximisation problem subject to a budget constraint. For the two goods case, at an **interior solution,** the optimality condition for consumer choice is that the marginal rate of substitution equals the price ratio. In a **corner solution** one of the goods is not consumed. If the goods are perfect substitutes you will always find a corner solution unless the price ratio equals the marginal rate of substitution. You should be able to derive consumer demand algebraically. Study the derivation of demand for Cobb-Douglas utility $U(x_1, x_2) = x_1^c x_2^d$ in Varian (2006) p.93. The resulting demand functions

$$x_1 = \frac{cm}{(c+d)p_1} \text{ and } x_2 = \frac{dm}{(c+d)p_2} \text{ where } m \text{ is income.}$$

Note that for Cobb-Douglas utility the expenditure shares of each good are constant (i.e. $p_1 x_1/m = c/(c+d)$ and $p_2 x_2/m = d/(c+d)$ do not vary with prices or income). Empirically this implies that we may be justified in assuming Cobb-Douglas utility if the fraction of their income which consumers spend on various products is relatively stable. Once we know a consumer's utility function we can assess how he is affected by price and income changes; this is useful in evaluating policy decisions regarding taxation for example.

Once you know how to derive individual demand curves, it is relatively easy to determine market demand. For any given price, market demand is the sum of the individual demands. In other words, to find market demand add individual demands **horizontally.**

Slutsky and Hicks[3]

Make sure you know what is meant by:

- normal good (note that a good can be normal at some income levels and inferior at others)
- inferior good
- Giffen good.

When the price of a good changes relative to prices of other goods, there are two effects. The first one is the **substitution effect** which represents the fact that the rate at which you can trade off the good under study with another good has changed. The second effect is the **income effect** which represents the idea that your purchasing power increases after a price decrease and decreases after a price increase. There are two versions of the substitution effect. In the **Slutsky version** the budget line rotates through the original consumption bundle until it has the same slope as the new budgetline. The tangency of this rotated budgetline and the highest possible indifference curve determines the change in consumption due to the Slutsky substitution effect. Since the rotated budgetline passes through the original consumption bundle, the latter is still affordable and in that sense purchasing power has remained constant. In the **Hicks version** the budgetline rolls under the original optimal indifference curve until it has the same slope as the new budgetline and the new tangency (with the same indifference curve so that utility is now held constant) determines the change in consumption due to the Hicks substitution effect.

The **Slutsky demand** for a good depends on the initial consumption bundle since we are holding the consumer's purchasing power constant in the sense that he can still (just) afford to buy his original bundle. Figure 6.1 illustrates the Slutsky substitution effect and demand.

[2] *I recommend Varian (2006) pp. 29–31 on the Food Stamp programme as an example.*

[3] *Read Varian (2006) Chapter 8*

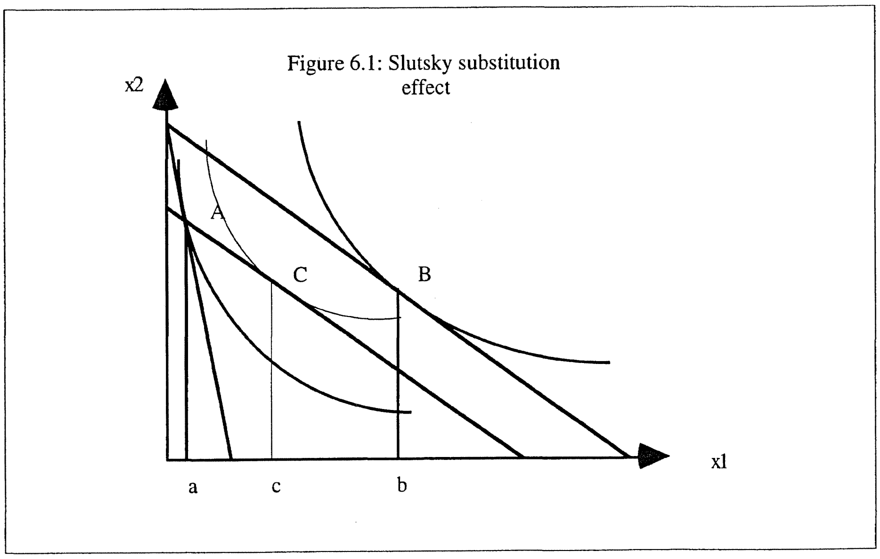


Figure 6.1: Slutsky substitution effect

The original budget line leads to a consumer optimum at A. After the decrease in price of good 1, the optimum shifts to B. The increase in Slutsky demand due to the price decrease is given by c-a. Point C is the tangency of a new (higher) indifference and a budgetline corresponding to the new relative prices drawn through the original bundle A. We will have to introduce some notation now to derive the Slutsky equation. The original prices for which A is the optimum are p_1^0 and p_2^0 and the quantities consumed at A are x_1^0 and x_2^0. From the figure it is easy to see that the Slutsky demand (x_1^s) which is obtained by rotating the budgetline through A as the price of good 1 decreases and drawing the relationship between price and quantity, is in fact ordinary demand (x_1) where income is held such that A is just affordable or:

$$x_1^s(p_1, p_2, x_1^0, x_2^0) = x_1(p_1, p_2, m = p_1 x_1^0 + p_2 x_2^0). \qquad (1)$$

To derive the Slutsky equation, we take the derivative with respect to p_1:

$$\partial x_1^s(p_1, p_2, x_1^0, x_2^0)/\partial p_1 = \partial x_1(p_1, p_2, m = p_1 x_1^0 + p_2 x_2^0)/\partial p_1$$
$$+ (\partial x_1(p_1, p_2, m = p_1 x_1^0 + p_2 x_2^0)/\partial m)\, x_1^0$$

which can be rearranged to give:

$$\partial x_1(p_1, p_2, m = p_1 x_1^0 + p_2 x_2^0)/\partial p_1 = \partial x_1^s(p_1, p_2, x_1^0, x_2^0)/\partial p_1$$
$$- (\partial x_1(p_1, p_2, m = p_1 x_1^0 + p_2 x_2^0)/\partial m)\, x_1^0. \qquad (2)$$

This last equation decomposes the price effect into the Slutsky substitution effect and the income effect.

Example 6.1

Assume Cobb-Douglas utility $U(x_1, x_2) = (x_1 x_2)^{1/2}$ and original prices $p_1^0 = p_2^0 = 1$ and income $m = 100$. From the section on 'Utility, budget constraint and demand', we know that the ordinary demands are given by $x_1 = m/(2p_1)$ and $x_2 = m/(2p_2)$ and hence $x_1^0 = x_2^0 = 50$. From (1) we know that Slutsky demand is given by:

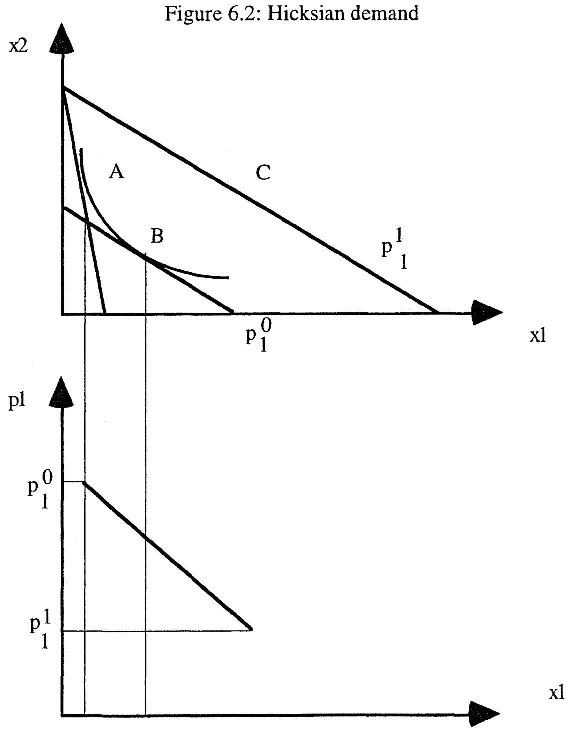
$$x_1^s(p_1, p_2, x_1^0 = 50, x_2^0 = 50) = x_1(p_1, p_2, m = p_1 x_1^0 + p_2 x_2^0 = 50(p_1+p_2)) =$$
$$25(p_1 + p_2)/p_1.$$

This implies that, when the price of good 1 decreases from 1 to 1/2, ordinary demand for good 1 increases from 50 to 100 whereas Slutsky demand increases from 50 to 75. The difference between 75 and 100 is due to the income effect. You should check the Slutsky equation (2) for this example:

$$-75/(2p_1^2) = -25\,p_2/p_1^2 - 1/(2p_1)50 = -150 \text{ for } p_1 = 1/2 \text{ and } p_2 = 1.$$

You should know how to derive **Hicksian or compensated demand** for the two goods case. Remember that the compensated demand function shows the response of quantity demanded to price changes if income is varied so as to keep the consumer on the same indifference curve (keeping utility constant). In Figure 6.2 the derivation of compensated demand is illustrated. For a high initial price of good x_1 the optimal consumption bundle is at A. If the price of good 1 decreases, so that budget line C applies, the consumer will choose a point on C. However, we are not interested in the ordinary demand now; we want to find compensated demand. We therefore find the optimal consumption bundle if the price ratio is as after the price decrease but the consumer stays on the original indifference curve. In the bottom figure Hicksian demand can be drawn simply by plotting the optimal quantity of x_1 as the price ratio changes and drawing the relationship between price and quantity. You can visualise the budget line rolling around the bottom of the indifference curve. As you can see on the graph, the optimal consumption bundle is found at the tangency of the budgetline and the indifference curve, as for ordinary demand. What we are doing mathematically when we derive the Hicksian demand is we minimise expenditure (push the budget line down) subject to the consumer achieving a given utility level. The Hicksian demand function is thus a function of prices and utility.

Figure 6.2: Hicksian demand



Example 6.2

Suppose utility is given by $U(x_1, x_2) = (x_1\, x_2)^{1/2}$. What do the compensated demands look like? The optimisation problem is:

$min\, p_1\, x_1 + p_2 x_2 \;\; s.t. \;\; U(x_1\,, x_2)= u_0.$.

From the tangency condition we know that the price ratio has to equal the marginal rate of substitution or:

$$\frac{p_1}{p_2} = \frac{\partial U/\partial x_1}{\partial U/\partial x_2} = \frac{1/2\sqrt{x_2/x_1}}{1/2\sqrt{x_1}/x_2} = \frac{x_2}{x_1} \text{ or } x_1 = \frac{p_2 x_2}{p_1}$$

Substitution in the constraint results in the Hicksian demand functions:

$$x_2 = \sqrt{\frac{p_1}{p_2} u_0} \text{ and } x_1 = \sqrt{\frac{p_2}{p_1} u_0}$$

You should understand why the substitution effect of a price change (Slutsky and Hicks) is always negative so that Slutsky and Hicks demand curves are always downward sloping. The income effect can be positive or negative so that ordinary demand is not necessarily although almost always downward sloping. In practice, the income effect is small unless consumers spend a significant fraction of their income on the good.

Most students find the material in this seetion quite abstract and hard to digest. I hope that the applications of substitution and income effects to intertemporal choice and labour supply will convince you that studying these concepts is worthwhile.

Consumer welfare effects of a price change[4]

[4] *Read Varian (2006) Chapter 14*

Public policy frequently has an effect on relative prices of goods. This is most obvious in the case of tax policy. If tax on particular commodities increases or if there is a shift towards excise tax to allow a reduction in income tax, consumer welfare is clearly affected. One way to measure this change in the consumer's well-being would be to compare his utility level before and after the change. This is problematic because the actual utility number has no significance. This may not matter if we only want to know whether the consumer is better or worse off. However, if we want a more precise measurement or if we want to aggregate the effect of a policy change over all consumers we need something more precise. For instance, we could determine how much income the consumer would be willing to give up to avoid a price increase or how much money the government would need to give the consumer to compensate him for a policy change which leads to a price increase (e.g. raising VAT on fuel). Using consumers' own monetary valuations of price changes allows aggregation.

There are several measures which could be used to assess a change in consumer welfare. **Compensating variation** (CV) is the amount of money which the individual requires to compensate him for a price increase or the amount of money which could be taken away from the individual after a price decrease to make him **as well off as before** the price decrease. **Equivalent variation** (EV) is the amount of money the consumer is willing to pay to avoid a price increase or the amount of money the individual should get to achieve **the same utility as after** a price decrease. You should remember that CV refers to a change in income which brings you back to the original indifference curve whereas EV refers to a change in income which brings you to the new indifference curve. CV indicates the change in income after the price change has taken place whereas EV indicates the change in income before the price change takes place.

You will probably need to think about this for a while. Making some graphs might help. Remember that you can read CV and EV on the vertical axis if x_2 has unit price.[5]

[5] *You could read the analysis in Gravelle and Rees (2004) 58–65 where the relationship between EV, CV and Hicksian demand is explained but I do not expect this level of mathematical sophistication from you.*

The three measures of changes in consumer welfare due to price changes (EV, CV and the change in consumer surplus CS) are equal when the income effect of a price change is zero. This is the case for **quasilinear utility** or $U(x_1, x_2) = v(x_1) + x_2$. Note that, on the indifference map corresponding to this utility function, the vertical distance between two indifference curves is constant (i.e. it does not vary with x_1). CV, EV and the change in CS are all different except for quasilinear utility where all three coincide. (It would be a good exercise for you to check this.) The change in CS is a good approximation when the income effect is small which in turn is likely when expenditures on the commodity as a share of total income is small.

Example 6.3

Using the same utility function $U(x_1 , x_2) = (x_1, x_2)^{1/2}$ as above and initial prices and income $p_1 = p_2 = 1$, $m = 100$, we can calculate the values of the three measures of changes in consumer welfare when p_1 decreases to 1/2. CV is the amount of money we can take away from the consumer after the price decrease to bring him back to his original utility level. To calculate CV we need to know how optimal utility is affected by a change in price. If we substitute the demand functions $x_1 = m/(2p_1)$ and $x_2 = m/(2p_2)$ into U we find the **indirect utility function:**

$$V = m/(2(p_1p_2)^{1/2})$$

so that utility after the price change is:

$$V_1 = m/(2(1/2.1)^{1/2}) = m/2^{1/2}$$

and utility before the price change is:

$$V_0 = m/(2(1.1)^{1/2}) = m/2$$

CV is the amount of money we can take away from the consumer to leave him as well off as initially. Hence $(m\text{-}CV)/2^{1/2} = m/2$ or $(100\text{-}CV)/2^{1/2} = 50$ which gives $CV=29.3$. EV is the amount of money you have to give the consumer to make him as well off as after the price decrease. This means that EV is such that:

$$(m+EV)/2 = m/2^{1/2} \text{ or } (100+EV)/2 = 100/2^{1/2}$$

which gives $EV=41.4$. The change in consumer surplus due to the price decrease is the area to the left of demand between the original and the new price, hence:

$$CS = \int_{1/2}^{1} \frac{m}{2p_1} dp_1 = 50(\log(1) - \log(1/2)) = 34.7$$

Elasticity[6]

Elasticity measures tell us how responsive demand is to changes in price, income, prices of substitutes or complements, etc. Whether we look at price elasticity or income elasticity or cross price elasticity, the measure is always defined in proportional terms. It tells us what percentage change in demand we can expect if the other variable changes by one per cent

You should be aware that elasticities are time-dependent. For most products it is unrealistic to assume that the price and income elasticities will be unchanged over a long period of time. The simplicity, at least in theory, of calculating values for, say, income elasticity is misleading. For marketing purposes this figure has to be interpreted carefully.Let us take the example of Malaysian car sales. A 40 per cent rise in incomes between 1987 and 1991 was accompanied by a 290 per cent rise in car sales, giving a rough estimate of income elasticity of 7. Should the car industry attach much importance to this number? A closer look at the car market and the market for many goods in fast growing economies reveals that households do not consume a good when

[6] *Read Solberg (1992) Chapter 4; Varian (2006) Chapter 15*

their income is below a threshold level but as soon as their income reaches the threshold level they do buy. This means that, even with small increases in per capita income, large increases in purchases could occur. At the same time, a large increase in per capita income may not have a significant effect on sales if the households whose income increased were already above the threshold level. Clearly, it is more useful to have a good forecast of the growth in percentage of households above the threshold than of growth in per capita income.[7]

Whether demand for a good is price elastic depends on several factors including:

- **the number and closeness of substitutes**. The elasticity of demand for a particular brand of cigarettes is large whereas the demand for cigarettes as a whole is inelastic. The demand for petrol and salt is inelastic
- **the period of time over which the response to a price change is assessed**. For many goods if you study the response after a longer time period the elasticity will be larger because consumers need some time to look for substitutes after a price increase. For consumer durables such as electrical appliances or cars however, the opposite is true. It appears that if the price of these goods increases people postpone their purchases and do not buy in the short-term; hence there is a high short-term response. In the longer term consumers cannot postpone replacement of their appliance any further which leads to a lower long-term response to a price increase
- **luxuries versus necessities.** The demand for luxury items (income elasticity > 1) is more elastic than the demand for necessities (income elasticity <1) because of their large income effect.

The notion of price elasticity is crucial in pricing decisions. Let us ignore costs for the time being and just determine how price *p* should be set if we want to maximise revenue. Revenue *R(p)* is the product of price and quantity demanded at that price: *R(p)=pq(p)*. To see how revenue varies with price, take the derivative:

$$\frac{\partial R(p)}{\partial p}=p\frac{\partial q(p)}{\partial p}+q(p)=\left(\frac{p}{q(p)}\frac{\partial q(p)}{\partial p}+1\right)q(p)=(1-\eta)q(p)$$

where η is the price elasticity. Therefore revenue is increasing in *p* (positive derivative) as long as demand is inelastic (η<1) and it decreases in price when demand is elastic (η>1). A revenue maximiser sets price so that demand has unit elasticity. An important consequence is that a **profit**-maximising firm will never set price such that demand is inelastic. Why? If demand is inelastic, an increase in price generates more revenue. At the same time, because of the increase in price, less is sold and therefore costs decrease. Revenue increases and costs decrease after a price increase so profit can be increased by increasing price as long as demand is inelastic.

Price elasticity determines the optimal markup over costs. As you would expect intuitively, if demand is inelastic the markup is higher than when demand is elastic. To see why this is true, consider the case of constant marginal cost c and express revenue as a function of quantity rather than price: R(*q*) = p(*q*)*q*. For profit maximisation we set marginal revenue equal to marginal cost. Marginal revenue is simply the derivative of revenue with respect to quantity and so:

$$\frac{\partial R}{\partial q}=p(q)+q\frac{\partial p}{\partial q}=p\left(1+\frac{q}{p}\frac{\partial p}{\partial q}\right)=c \text{ or } p\left(1-\tfrac{1}{\eta}\right)=c$$

which can be rewritten as:

$$p=c\left(\frac{\eta}{\eta-1}\right)$$

[7] *See 'Buyers and builders'*

The optimal markup over cost $\frac{\eta}{\eta-1}(\eta>1)$ is decreasing in the price elasticity (you should check this) as mentioned above.

In practice **constant elasticity demand** is often assumed when estimating demand functions:

$$q = ap^{-\eta}m^{\gamma}t^{\beta} \quad (3)$$

where q is the quantity demanded, p is the price, m is income and t is the price of a substitute or a complement. Other factors such as demographic variables are often included. The parameters η, γ and β are the price elasticity, income elasticity and cross-price elasticity respectively.[8] The multiplicative functional form (3) assumes that the elasticities are constant (i.e. they do not vary with values of price, income, etc). The popularity of this model can be explained by its ease of use in econometric work (linear regression can be used after taking logarithms on both sides).[9]

[8] *You should check this using the formula for elasticity*

[9] *If you are curious about how demand functions are estimated you should read the examples in Chapter 8 and Chapter 9 of Solberg (1992)*

Regression analysis and time series analysis are not the only methods you can use to get information about demand for your product. In fact in some cases regression analysis based on historical data is out of the question because such data are not available (e.g. for a new product). Marketing managers then resort to consumer labs or market experiments. In a consumer lab or clinic, simulations of purchasing decisions take place. The 'consumers' are given a budget which they can spend on the product under study as well as substitute products (other brands) and other products. The experiments then consist of varying the consumers' budget and the relative prices. The resulting information is analysed and elasticities are calculated. In the case of test marketing, the experiment goes on in the real world. The market for the product is divided into geographical areas and a different marketing mix (price, advertising, etc) is used in each area. Consumers' responses are measured and demand information results from analysis of these responses.

State-contingent commodities model[10]

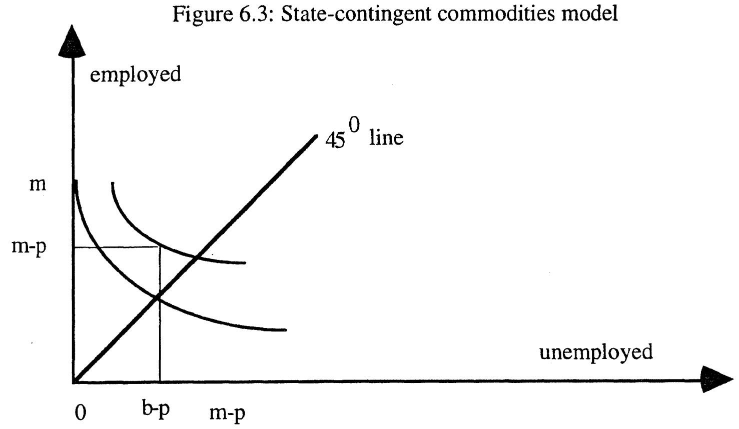
[10] *Read Varian (2006) Chapter 12*

When we analysed decision-making under uncertainty, we used the expected utility model. There is an alternative approach to studying uncertainty: the **state-contingent commodities model.** This approach uses standard consumer theory terminology which is why it is discussed here. Whereas under the expected utility hypothesis any number of outcomes can be considered, the state-contingent commodities model, as it relies heavily on graphical analysis, is most useful when there are only two outcomes.

The 'products' a consumer derives utility from are income or wealth in two states of the world. The states of the world could be 'car is stolen' and 'car is not stolen' for example. Of these states of the world one and only one will actually occur. This points to a major departure from standard consumer theory. In the state-contingent commodities model, only one of the goods is consumed **by definition** whereas in the standard consumer theory model, both goods can be (and usually are) consumed in positive quantities.

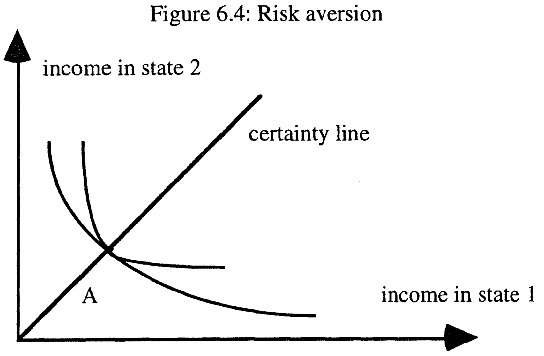
Individuals derive utility from income bundles consisting of one income level for each possible state of the world. If I do not have unemployment insurance, my income bundle could be (*0, m*) indicating a zero income level in the state 'I lose my job' and income *m* in the state 'I do not lose my job'. If I have unemployment insurance, my income bundle could be (*b-p, m-p*), where *b* is the unemployment benefit the insurance company pays me when I lose my job and *p* is the insurance premium. Depending on the values of *b* and *p* and my chance of becoming unemployed, I may prefer the second bundle to the first. Figure 6.3 shows an indifference map corresponding to this situation.

Figure 6.3: State-contingent commodities model



The 45 degree line or the **certainty line** contains income bundles for which income in both states is equal and hence there is no uncertainty. The indifference curves are convex for a risk averse individual because, loosely speaking, such an individual prefers the same income in both states. Consider an individual at point A on the 45 degree line in Figure 6.4. If this individual were more risk averse she would be willing to trade her secure position A for fewer income bundles. The income bundles she would trade for are of course the ones above her indifference curve through A. So higher risk aversion corresponds to more convex indifference curves.

Figure 6.4: Risk aversion



In general, utility (and thus the position of the indifference curves) depends on income in both states and the probability (q_i) of each state of the world occurring:

$$U(m_1, m_2, q_1, q_2).$$

An example of such a utility function is the expected utility function we studied before:

$$U(m_1, m_2, q_1, q_2) = q_1 u(m_1) + q_2 u(m_2) \qquad (4)$$

but the state-contingent income approach allows for more general utility functions. Using the expected utility formulation (4) let's see how we can determine the certainty equivalent and the risk premium graphically. On an indifference curve $dU(m_1, m_2, q_1, q_2) = 0$ or, from (4),

$$q_1 u'(m_1)\, dm_1 + q_2 u'(m_2)\, dm_2 = 0.$$

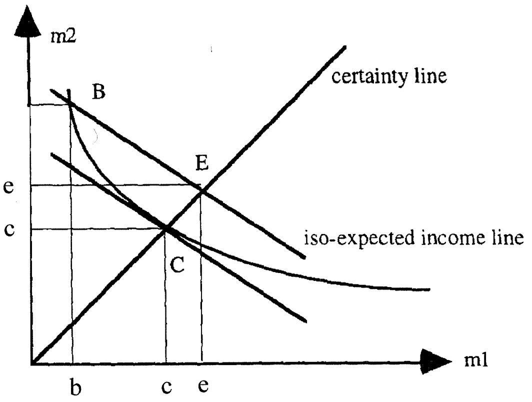
The slope of the indifference curve is thus:

$$dm_2 / dm_1 = - q_1 u'(m_1) / q_2 u'(m_2) \qquad (5)$$

which is the marginal rate of substitution between income in the two states, adjusted by the probabilities of the states occurring. At the intersection of an indifference curve with the certainty line (where $m_1 = m_2$), the slope is $dm_2 / dm_1 = - q_1 / q_2$.

The **iso-expected income line** is the locus of income bundles *(m_1, m_2)* which have the same expected value *E(m)*. The equation for the iso-expected income line for given probabilities (q_1, q_2) is $q_1 m_1 + q_2 m_2 = E(m)$. The slope of this line is also $-q_1/q_2$ which implies that the tangency of an indifference curve and an iso-expected income line occurs at the intersection of the indifference curve with the certainty line. This is illustrated in Figure 6.5.

Figure 6.5: Certainty equivalent and risk premium



We now want to situate the certainty equivalent of income bundle B graphically. By definition the individual is indifferent between all points on his indifference curve and, in particular, he is indifferent between B and C. At point C the individual gets equal income *c* in both states. The certainty equivalent of B is therefore *c*: the individual is indifferent between getting *c* for sure or getting state-contingent income B.

It is straightforward to determine the risk premium graphically. The risk premium is the difference between the expected value and the certainty equivalent. Where can we read off the expected value of B? Income bundle E is on the iso-expected income line through B and therefore has the same expected value as B, namely $q_1 e + q_2 e = e$. The risk premium therefore equals *e-c*.

The state-contingent commodities model has applications in the analysis of insurance demand.[11]

Intertemporal choice[12]

I strongly recommend that you read Varian's excellent treatment of intertemporal choice for revision purposes or if you need some clarification. You should know how to set up the intertemporal consumption choice problem and derive how saving and borrowing respond to changes in the interest rate *r*.

Suppose the consumer has income m_1 in period 1 and income m_2 in period 2 (the endowment point E in the figure 6.6 below). He can borrow and lend at interest rate *r*. If he does not consume in period 1 then the amount available for consumption in period 2 equals $m_2 + (1 + r)m_1$ .Similarly, if he plans not to consume in period 2, he can borrow

[11] *See Gravelle and Rees (2004) 507–14 if you want to study this but I do not expect you to.*

[12] *Read Solberg (1992) Chapter 6; Varian (2006) Chapter 9.*

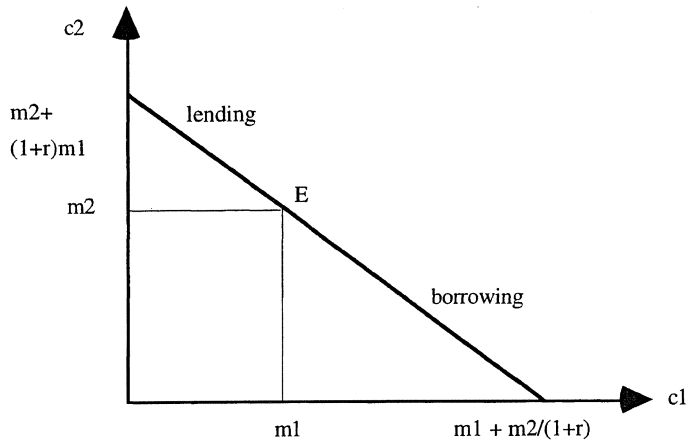
an amount $m_2/(1+r)$ in period 1 and hence pay back m_2 in period 2. In general, consumption in period 2 will equal income in period 2 plus any savings from period 1 or minus any repayments of loans from period 1:

$$c_2 = m_2 + (1+r)\,(m_1 - c_1). \qquad (6)$$

Equation (6) is the equivalent of the budget constraint in the standard consumer choice problem. Note that, as the interest rate r changes, the budget line rotates through the endowment point E, becoming steeper with an increase in r and flatter with a decrease in r.

If the individual can borrow but not lend, he has to choose a point on the line segment to the right of E and, if he can lend but not borrow, he has to choose a point on the segment to the left of E. If the borrowing and lending interest rates are different, the budget constraint is piecewise linear (and concave if the borrowing rate is higher than the lending rate) with a kink at E.

Figure 6.6: Intertemporal choice



The consumer's preferences over combinations of current and future consumption can be represented by a utility function $U(c_1, c_2)$. At the optimum bundle of current and future consumption, the budget line is tangent to the highest possible indifference curve. The MRS in this context is defined as $-(1+\rho\,)$ where $\rho\ (>0)$ measures subjective impatience.

Why?

At the optimum the MRS equals the slope of the 'budget line' $-(1+r)$ which implies that each individual saves or borrows such that the rate r at which he can trade current for future consumption in the market coincides with the rate at which he wishes to make these trades.

Using this model, we can derive the amount of saving or borrowing every individual in the market plans to do, given the interest rate. The market interest rate can be determined by plotting the total amount of saving and the total amount of borrowing against the interest rate. You read off the equilibrium interest rate where the two curves intersect.

Example 6.4

Suppose the consumer has utility function $U(c_1, c_2) = \ln(c_1) + a \ln(c_2)$, where a (<1) is a constant. The return on loans to and from the bank is r and income is m_1 in period 1 and m_2 in period 2. The consumer's optimisation problem is:

$$Max\ \ln(c_1) + a \ln(c_2)\ s.t.\ c_2 = m_2 + (m_1 - c_1)(1+r).$$

If we substitute the constraint into the objective function we obtain:

$$\ln(c_1) + a \ln(m_2 + (m_1 - c_1)(1+r)).$$

Setting the derivative with respect to c_1 equal to zero, we find:

$$1/c_1 + (a/(m_2 + (m_1 - c_1)(1+r))(-1)(1+r) = 0$$

which can be solved for c_1:

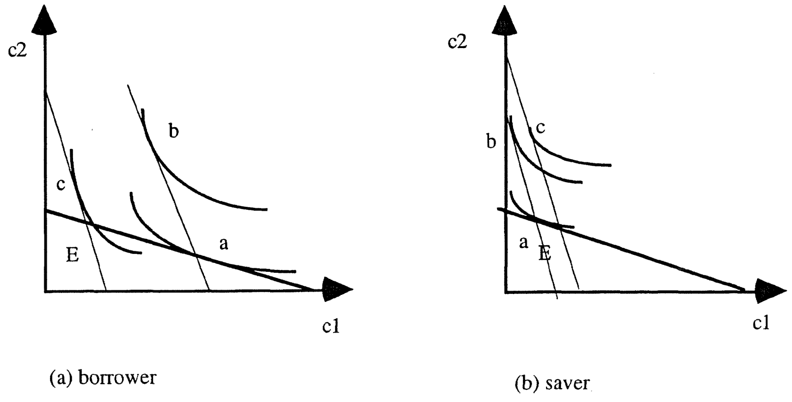
$$c_1{}^* = (m_1 + m_2/(1+r))/(1+a).$$

If the consumer can lend but not borrow then the solution remains unchanged as long as $c_1{}^* < m_1$. If $c_1{}^* > m_1$ *i.e.* $m_2 > a\, m_1\, (1+r)$ then the consumer would like to borrow but, if he is prevented from doing so, he will consume his endowment:

$$(c_1, c_2) = (m_1, m_2).$$

The effect of an interest rate change can be decomposed into a substitution and an income effect. Using this decomposition it is possible to show, for example, that an increase in r causes a decrease in c_1 for a borrower whereas it could cause a increase in c_1 for a lender.

Figure 6.7 : Effect of an interest rate increase



(a) borrower

(b) saver

Figure 6.7(a) shows what happens to a borrower's consumption plans when the interest rate increases. Initially he is planning to consume at point a, to the right of his endowment E. After the interest rate increase he modifies his plans to point c. The dashed line through a has the same slope as the new budget line (I am using Slutsky substitution here). The move from a to b is caused by the substitution effect, which always works in the direction of less c_1 if r increases. The move from b to c is due to the income effect which leads to less c_1 if c_1 is a normal good. Figure 7(b) shows the effect of the same interest rate increase on the saver's consumption plans. The substitution effect works in the same direction as for the borrower but the income effect now works in the opposite direction. If the income effect is large enough, c_1 could increase. Therefore, it is possible to get a **backward bending supply of savings** (i.e. as the interest rate increases to a high level, savers may reduce their savings!).

As an exercise, show that a saver can be made better or worse off by an interest rate decrease. Note that a saver's initial consumption plan is not feasible after the interest rate decrease.

Case: Britain benefits from lower interest rates

When you analyse the effect of interest rate changes on the economy in different countries you have to take into account the amount of debt at floating rates. Fixed rate debt is by definition not affected by a change in the interest rate. In an economy with mainly fixed rate debt and variable rate savings, interest payments on loans will not change much if the interest rate falls but interest receipts from savings could go down dramatically. American households paid $6 billion less in interest payments between 1992 and 1992 but their interest receipts fell by $73 billion.

Britain is an exception among the big economies in the world in that it is a net debtor if we consider assets and debts held at floating rates by firms and households. A large majority of British homeowners hold variable rate mortgages (90 per cent versus less than 20 per cent in the US and 10 per cent or less in Germany, France and Japan). Also, less than half of British company debt is at fixed rates whereas in the U.S., France and Japan it makes up 60 per cent or more.[13]

Since Britain is a debtor, it is better off when the interest rate decreases whereas the rest of the world could be worse off. Why is that? For a borrower, when the interest rate decreases, the substitution and income effects reinforce each other and cause him to borrow more. (Make a graph to see this!) He ends up on the new budget line to the right of his original position. The original consumption bundle is still feasible which means that the borrower is at least as well off as before the fall in interest rate.

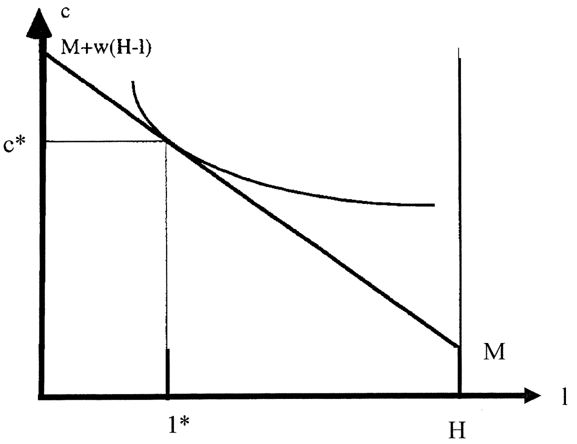
[13] *See 'Giving the economy a fix'.*

Labour supply[14]

[14] *Read Solberg (1992) Chapter 5; Varian (2006) Chapter 9*

The standard consumer choice problem is easily rephrased in the context of the labour supply decision. The 'goods' are leisure l and a composite good c (with price =1) which can be interpreted as 'income available to spend on consumer goods'. The budget constraint indicates how much consumption is possible for each choice of leisure and its implied choice of work time given a wage rate w. The worker may have non-labour income M which is taken into account in the formulation of the budget constraint. In addition to the monetary constraint, the consumer also faces a time constraint: work

Figure 6.8: Labour supply

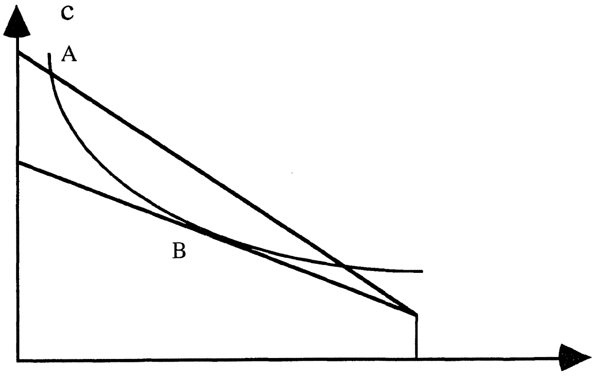


time and leisure time have to add up to 24 hours per day or, more generally, *H* hours per time period. The worker's optimisation problem, represented graphically in Figure 6.8, is: *Max U(c, l) s.t. c = M + w(H–l)* and at the optimum the MRS equals the price ratio-*w*.

You should know how to modify the model to allow for progressive income tax. The budget constraint becomes piecewise linear and concave. Investment in education can be modeled as a decrease in *M* (fees, opportunity cost, etc) and an increase in *w* (higher wages for the more educated) so that the budget constraint becomes steeper.

Of course the basic labour supply model assumes that the worker has a free choice of working hours or length of working week. In reality the employer will want to place some restrictions on that choice, often in the form of a lower bound on the number of hours worked. We can nevertheless still use the model to study the employee's reaction to these restrictions. Statistics show a large variance of number of hours worked per week among workers. In Britain more than five million people, or 20 per cent of the workforce, work more than 48 hours per week.[15] There is some evidence suggesting that people are working harder than they want to in the sense that they would rather earn less and work less than they currently do. Job satisfaction is higher for women part-timers than for full-time workers.[16] The Economist[17] quotes a survey in which 80 per cent of British part-timers say they prefer working part-time to working full-time and a third of full-timers in the EU said they would rather work fewer hours than have a pay rise. If workers have a strong preference for working part-time, then firms which are willing to be flexible rather than insisting on, say, a 40-hour work week can obtain labour at a lower cost. We can show this by drawing the indifference curve through *H-l* (hours worked) = 40 and the corresponding *c* (point A in Figure 6.9). We can now determine how much lower the wage rate could be (while keeping the worker as happy as before) if the worker is allowed to choose the length of the working week. Rotate the budget line down until you find a tangency (point B) with the indifference curve. At B the worker is as well off as at A although his income has decreased compared to at A. An interesting statistic in this context is that women part-timers earned only 72 per cent of the rate per hour of full-time women workers in 1992.[18]

Figure 6.9: Flexible versus fixed hours



As the wage increases, different choices between leisure and work time are made. This relationship between wage and hours worked is the **labour supply curve**. The effect of an increase in wage can be decomposed in a substitution and income effect. The substitution effect pushes in the direction of more work whereas the income effect pushes for more leisure if you assume that leisure is a normal good. If the income effect is large enough, we get a **backward bending** labour supply curve where wage increases result in fewer hours of work.

[15] *See 'Hard Labour'*

[16] *See 'The home front'*

[17] *See 'Sharing the burden'*

[18] *See 'The home front*

A tax on labour earnings has the same effect as a decrease in wage: the income and substitution effect counteract each other so that the worker could end up working more or less whereas a lump sum tax only has an income effect making the worker work more. As an exercise you could make a graph illustrating the effect of a fixed rate income tax when the labour supply curve is backward bending and show that introduction of the tax reduces labour supply for low wage rates and increases labour supply for high wage rates.

In line with the mainstream empirical work in this area, Battalio, Green and Kagel (1981) find evidence of a backward bending labour supply curve at high 'wages' for pigeons and rats! In their experimental work they discovered that laboratory animals trade off income for leisure and that leisure is a normal good. 'Work' for pigeons is typically key pecking or treadle running whereas rats are conditioned to do 'work' tasks such as lever pressing and wheel running. The animals have to perform the task a certain number of times to get a reward. The authors claim that rats and pigeons behave as if they were maximising a utility function *U(c,l)* and they conclude, 'Substitutability of income and leisure appears to be a fundamental, biologically-based, law of behavior.'

Example 6.5

A consumer has a Cobb-Douglas utility function, $U(c,l) = c^{\alpha} l^{1-\alpha}$ and budget constraint $c = M + w(H-l)$. To derive the individual's labour supply curve, set the ratio of marginal utilities equal to the slope of the budget constraint:

$$\frac{\partial U/\partial c}{\partial U/\partial l} = \frac{\alpha c^{\alpha-1} l^{1-\alpha}}{c^{\alpha}(1-\alpha)l^{-\alpha}} = \frac{1}{w} \text{ or } \frac{\alpha l}{(1-\alpha)c} = \frac{1}{w}$$

Solving this for *l* gives

$$l = \frac{(1-\alpha)c}{\alpha w} \qquad (*)$$

and substituting this expression into the budget constraint gives:

$$c = M + w(H - \frac{(1-\alpha)c}{\alpha w}) \text{ or } c = \alpha(M + wH)$$

Substituting this expression for *c* into (*) leads to:

$$l = (1-\alpha)\left(H + \frac{M}{w}\right)$$

and the labour supply curve is therefore:

$$H - l = \alpha H - (1-\alpha)\frac{M}{w}.$$

The number of hours worked, *H-l,* is increasing in *w* and hence, for this particular utility function, the labour supply curve is not backward bending.

The simple model discussed above can be used to analyse many interesting questions. Suppose the government pays a fixed amount in welfare benefit for unemployed workers (i.e. if they work zero hours they get the benefit but if they work at all, they do not).

You should be able to show that this provides an incentive to be idle especially if the wage rate is low.

In many countries, welfare or unemployment benefit is withdrawn completely as soon as the recipient enters the labour market. This 'welfare trap' has detrimental effects on work incentives. Contrast this with the Canadian province of Newfoundland which has recently introduced a plan to avoid the problem of sudden cut-off of benefit. Every individual is guaranteed a basic income if he does not work. Rather than losing this welfare payment when people start working they in fact receive a **higher** payment in the form of a work subsidy. The income supplement then levels off and eventually decreases to zero at an income of C$42,500 for a family of four[19]. You should be able to

[19] *See 'New-found plans'*

show graphically that this plan reduces voluntary unemployment and that it may be possible to reduce total welfare payments (by substituting wage subsidies for unemployment benefit) without making the recipients worse off.

In a recent paper, Biddle and Hamermesh (1990) tackle a subject which has long been ignored not only by economists but by social scientists in general: the demand for sleep. Given that most survey respondents when asked state that they sleep more hours than they work, understanding the demand for sleep seems crucial for developing a model of allocation of time by consumers. Biddle and Hamermesh suggest that there are three approaches, or hypotheses, to incorporating the demand for sleep in the standard leisure-consumption tradeoff model we have worked with in this chapter.

1. The minimum amount of sleep an individual needs is assumed to be biologically determined and consumers do not derive utility from sleep. Under this hypothesis, an individual sleeps according to his biological minimum and we can set the amount of time an individual has available for work and leisure, *H*, equal to 24 hours per day minus the amount of sleep.
2. A more plausible hypothesis is that the individual derives utility from hours spent sleeping as he derives utility from leisure hours in general. In terms of the labour supply model, the question arises whether we make a mistake by bunching sleeping with other leisure activities.
3. There is some evidence that sleep affects productivity and hence wage. Clearly, if we are going to take this hypothesis seriously, our labour supply model has to be revised rather dramatically. The wage rate becomes endogenous as it is affected by the consumer's choice of sleeping time: $w = w_e + w_s t_s$ where w_e is the exogenous part of the wage rate and w_s measures the sensitivity of productivity or wage to time spent asleep t_s. Total time H is made up of working, waking leisure and sleeping: $H = t_w + t_l + t_s$ .The consumer's problem can thus be written as:

$$Max\ U(c\ ,t_s\ ,\ t_l) s.t. c=M+(w_e+w_s t_s)(H-t_s-t_l).$$

Biddle and Hamermesh come up with the following empirical findings:

- people with higher wages sleep less
- if the wage rate increases, the number of hours men work does not increase but sleeping time is reduced in favour of leisure time
- consistent with prior research, men's supply of work hours is much less sensitive to changes in the wage rate than women's. The labour supply elasticity is measured as -0.021 for men and 0.191 for women
- sleeping time seems to be a normal good: in a sample of economies, a higher GNP is associated with significantly longer sleep duration.

Risk and return[20]

Further reading

If, after reading this section, you are interested in learning more about finance (and if you have a lot of spare time!), I recommend:

Brealey, R.A. and Myers, S.C. *Principles of corporate finance.* (Singapore: McGraw-Hill, 2005) eighth edition [ISBN 0073130826]: this is a very accessible but thorough text.

In the chapter on decision analysis, I touched on the subject of investment choice between risky assets. Suppose you are offered two alternative investment opportunities each with a given set of uncertain outcomes and a probability distribution over these outcomes. A straightforward approach would be to use the expected utility model

[20] *Read Solberg (1992) Chapter 7; Varian (2006) Chapter 13*

(i.e. calculate a weighted sum of utilities of each outcome, with the weights equal to the probability of the outcome materialising) and choose the asset which delivers the highest expected utility. Although this approach is valid, major progress in financial economics was made by using an assumption to simplify this process. Harry Markowitz (1959) who, in 1990, was awarded the Nobel Prize in Economics jointly with William Sharpe and Merton Miller, proposed a model of choice among risky assets where utility depends on expected value and standard deviation of income rather than on the entire probability distribution.

When risk averse investors decide on their investment in a financial asset they consider not only expected return (dividend or interest receipts plus capital gain or loss) but also the risk involved. When given a choice between investment A which gives a return of 10 per cent for sure and investment B which gives five per cent or 15 per cent with equal probability, A is chosen. In general, an investor is only willing to take on higher risk when there is a compensating increase in expected return (i.e. his indifference curves in the risk-return plane are upward sloping: expected return is a 'good' whereas risk is a 'bad'). The level of risk associated with an investment is usually measured as the standard deviation of the return. (This may be a good time to look for your statistics notes!) Investments with larger standard deviations have to be rewarded by greater expected returns. For example, average annual returns for venture capital from 1945 to 1993 were 15.9 per cent and the standard deviation in the returns was 25.5 per cent; stocks included in the S&P 500 averaged a return of 11.7 per cent over the same period and a correspondingly lower standard deviation of 16.3 per cent. The exceptions to the rule were gold and silver which had low average returns of about five per cent and very high standard deviations of 25.8 per cent and 55.8 per cent respectively.[21]

Given an investor's preferences we could predict which of a list of assets, with given risk and return, he would invest in. In reality, however, the investor is not restricted to investing his entire budget in one asset. He could invest in several assets or a **portfolio.** Investing in several assets reduces risk when the returns on the assets are not perfectly positively correlated. I will come back to this later. I want to turn to a simple example illustrating the role of correlation between returns first. Suppose I can invest in a business selling umbrellas and/or a business selling ice cream. The profitability of these businesses depends on the weather. Assume I have an investment budget of £10,000 and my net payoff from investing £1 in either business is according to the table below.

weather	umbrellas	ice cream	probability
good	-1	1	1/2
bad	10	-1	1/2

If I invest my £10,000 only in umbrellas, my **expected** payoff is £45,000 and if I invest only in ice cream, my **expected** payoff is £45,000. However, suppose I invest £5,000 in both then, when the weather is good, I gain £45,000 **for sure** and, when the weather is bad, I gain £45,000 **for sure.** In other words, rather than running considerable risk by investing in one business, I can eliminate all risk without reducing my expected return by diversifying my investment.

Before you get too excited about this I have to warn you that in practice it is impossible to find two stocks which have perfect negative correlation and so you have to settle for some uncertainty. By choosing a portfolio carefully this uncertainty can be reduced. Let us see how that works if we have two assets A and B with expected returns r_A and r_B and standard deviation of the returns σ_A and σ_B. Empirically it turns out that the normal distribution works well to describe the volatility of returns but we do not have to make any assumption about the distribution here. Assume $r_A < r_B$ and $\sigma_A < \sigma_B$

[21] *See 'Risk and return'*

otherwise we could do no better than invest in the asset with the highest return and lowest risk. If of every £1 in my budget I invest a share α in asset A and the remaining share $(1-\alpha)$ in asset B, the expected return and risk of my portfolio are given by:

$$r_p = \alpha r_A + (1-\alpha) r_B \qquad (7)$$

and

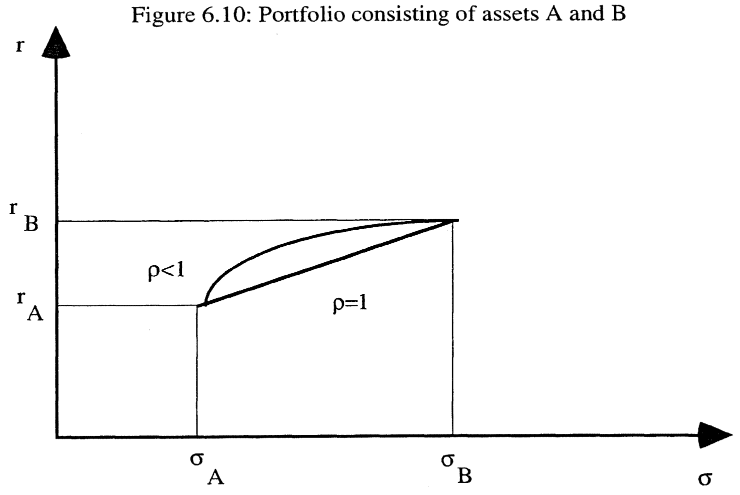
$$Var_p = \alpha^2 \sigma_A^2 + (1-\alpha)^2 \sigma_B^2 + 2\alpha(1-\alpha)\rho\sigma_A \sigma_B \qquad (8)$$

where ρ is the correlation between the returns of assets A and B. This means that, by varying the share of my budget I invest in A, it is possible to attain the combinations of expected return and risk given by (7) and (8). Consider the (easier) case of $\rho = 1$ first. For $\rho = 1$,

$$Var_p = (\alpha\sigma_A + (1-\alpha)\sigma_B)^2 \text{ or } \sigma_p = \alpha\sigma_A + (1-\alpha)\sigma_B \qquad (9)$$

which combined with (7) implies that (σ_p, r_p) lies on a straight line between (σ_A, r_A) and (σ_B, r_B) as is shown in Figure 6.10.

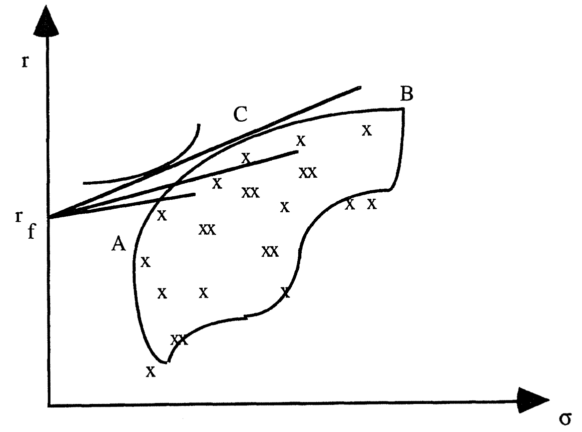
Figure 6.10: Portfolio consisting of assets A and B



For $\rho < 1$, $\sigma_p < \alpha\sigma_A + (1-\alpha)\sigma_B$ and hence the combinations of risk and expected return which are feasible will be to the left of the upward sloping line segment in Figure 6.10. The lower the correlation, the lower the risk of the portfolio. Ideally the two assets in the portfolio should have negative correlation but, as mentioned before, in practice finding stocks which are negatively correlated is difficult.

We could extend our exercise to more than two assets and determine the possible risk — expected return combinations which are possible by varying the shares of the budget allocated to each asset. The set of these feasible combinations could look like the figure below with the x's representing expected return — risk combinations for individual assets in the portfolio. Given this feasible set which portfolio would you choose? Of course it will depend on how risk averse you are but whatever your utility function is, we can slim down the choice set using a dominance argument. Given two portfolios with the same σ, you would never choose the one with the lower *r* because it is **dominated** by the portfolio which has the same risk and a higher return. So, for each possible risk level σ, we need only consider the portfolio which gives the highest expected return. Similarly, given two portfolios with the same *r*, you would always prefer the one with the lower risk. Hence, for portfolios with a given level of *r*, the portfolio with the lowest *s* dominates the others. The portfolios which are not dominated are called **efficient** and are on the curve between A and B in Figure 6.11. There are computer programs available which calculate the set of efficient portfolios given the risk and expected return of each available asset.

Figure 6.11: Efficient portfolios



Now suppose we have a risk-free asset (e.g. a Treasury bill) available in addition to our collection of risky assets. The risk-free asset gives a return r_f and (by definition) risk $\sigma_f = 0$. If we invest a share α in the risk-free asset and the remaining in a portfolio (σ_p , r_p) then we know from (7) and (8) that our investment will have an expected return of $r = \alpha\, r_f + (1-\alpha)r_p$ and a variance of $Var = (1-\alpha)^2\, \sigma_p^{\,2}$. This means that, by varying the share α invested in the risk free asset, we can attain all points on the line between $(0 , r_f)$ and (σ_p, r_p). We could do this for any portfolio in the feasible set but of course the portfolios on the efficiency line are more desirable and, of those, the portfolio at which the highest line from $(0 , r_f)$ is tangent to the efficiency line is the optimal one to combine with investment in the risk-free asset (point C in Figure 6.11).

In fact not only points between $(0 , r_f)$ and C are feasible but points to the right of C as well. By borrowing money at interest rate r_f and investing it in the optimal portfolio we can extend our 'budget line' to the right of C. Clearly, investors should choose a point on this budget line and lend or borrow money at fixed interest according to whether they pick to the left or the right of C. Exactly which combination of risk-free and risky investment is chosen depends on the investor's utility function, as represented by the indifference curve in the figure above. It is not difficult to derive the equation for the budget line through $(0, r_f)$ and (σ_p, r_p): $r = r_f + ((r_p - r_f)/ \sigma_p)\, \sigma$. The slope of this line $(r_p - r_f)/'\sigma_p$ is the **price of risk** which at the investor's optimum equals his MRS.

Chapter summary

After this chapter and the relevant reading, you should understand:

- the decomposition of the price effect into income and substitution effects (Hicks and **Slutsky** method)
- the concepts EV and CV
- the relationship between price elasticity and the effect of a price change on revenue
- the relationship between price elasticity and the optimal markup
- the role of the income effect in generating **backward bending labour supply**
- the tradeoff between **risk and return** in investment
- the importance of correlation between asset returns for risk reduction
- why everyone chooses the same portfolio of risky assets when a risk-free asset is available

You should be able to:

- derive demand functions and labour supply functions given a utility function
- derive the **Slutsky equation**
- derive Slutsky and compensated demand functions
- calculate **elasticities**
- explain the **state-contingent commodities model**
- set up the **intertemporal choice problem** and analyse the effect of changes in the interest rate
- set up the **labour supply** model
- derive the shape of the feasible set of risk-return combinations for portfolios of two risky assets.

Sample exercises

1. Goods x and y are perfect substitutes and consumers have utility functions of the form $U(x,y) = x + y$. Consumer i has income m_i. Derive and draw his demand for good x. Derive and draw the market demand for good x.
2. True/False. If the income elasticity is positive, the compensated (Hicks) demand curve is steeper than the ordinary demand.
3. Joanna likes lipstick (x_1) and earrings (x_2) which give her utility $U(x_1, x_2) = x_1x_2$. Her pocket money is M per week and lipsticks and earrings cost $p_1 = 4$ and $p_2 = 1$ respectively. Suppose p_1 falls to 1. By how much would her pocket money have had to increase to make her as well off as after this price change? By how much could her pocket money be reduced while keeping her as well off as before the price decrease? Which is the compensating variation? Which is the equivalent variation? Calculate her increase in consumer surplus due to the price decrease.
4. The price elasticity of demand for wixies is 1.5. The wixies producers form a cartel and agree wixies quotas which will have the effect of reducing output such that price rises by 10%. What is the effect on total revenue? How does your answer change if the price elasticity of demand is 0.75?
5. Using the state-contingent commodities model, show graphically that, for full insurance, a risk averse individual is willing to pay an insurance premium which is higher than his expected loss.

6. A consumer lives for two periods and receives income 100 in each period. In the first period he can invest some of his income in a stock which has a return r of 10% or 20%, each equally likely. His utility function $U(c_1, c_2)$ is a function of his consumption levels in both periods: $U(c_1, c_2) = c_1^2 + 0.7\ c_2^2$ for $c_1, c_2 \geq 0$. Find the optimal consumption plan.[22]

7. A worker can choose the fraction of a day, y, that he is not working (so he works a fraction 1-y of a day). He derives utility from consumption goods c (assumed to have price=1) and leisure y according to $U(c,y) = c^4\ y^2$. He has a fixed nonlabour income of M per day and the wage rate is w per day. Write his budget constraint and derive his labour supply curve. Is it backward bending at high wages?

8. Empirical studies show that (a) for male employment small changes in the wage rate do not affect the number of hours worked and (b) the number of hours women work increases when taxes fall because of their increased labour market participation. Explain these findings graphically.

9. Portfolio theory assumes that investors care about expected return and risk. Draw indifference curves for a risk neutral investor.

10. The return on stock A will be -10% or +20%, each with probability 0.5. The return on stock B will be -5% or +15%, with probabilities 0.3 and 0.7 respectively. The correlation coefficient between the returns of the two stocks is 0.6. Calculate the expected return and the variance of the return for each stock. Calculate the covariance between the returns of the stocks. Can you decrease your risk by investing in a combination of these two stocks rather than in either of the stocks?

11. Suppose the risk-free rate of return is 5% and you consider investing in a risky asset with r = 10% and $\sigma = 4\%$. Show all combinations of risk and expected return which are feasible if you invest in a portfolio consisting of the risk-free asset and the risky asset. If you are willing to carry a risk of 3% but not more what is the best portfolio?

12. During the past six years the returns on Air Angleterre and French Airways have moved as follows:

Year	1	2	3	4	5	6
AA	25	-16	18	5	13	2
FA	10	-3	15	3	20	8

Calculate all the parameters you need to derive the locus of feasible expected return-risk combinations if you invest a share of your budget in AA and the remainder in FA.

[22] ***Hint**: maximise expected utility*

Notes

Production functions and isoquants; firm demand for inputs; industry demand for inputs; from production function to cost function; division of output among plants; estimation of cost functions

Chapter 7

Production, factor demands and costs

Texts

Varian, H.R. *Intermediate Microeconomics.* (New York: W.W. Norton and Co., 2006) seventh edition [ISBN 0393927024]. Chapters 18, 19 and 26.

References cited

'Doleful', *The Economist,* 9 October 1993, 17.
'The trade unions scent a victory', *The Economist,* 3 September 1994, 27–28.
'Tuning in to the future', *The Economist,* 5 September 1992, 27–28.
Moroney, J. 'Cobb-Douglas production functions and returns to scale in US manufacturing industry', *Western Economic Journal* (1967), 39–51.

What I want to do in this chapter goes under the heading of 'theory of the firm' in most microeconomics textbooks. However, over the last 15 years or so, there has been a growing interest in what determines the boundaries of the firm, how firms are organised internally, how their organisational design can be explained in terms of efficiency and so on. All of these topics are elements of a 'theory of the firm'. The neoclassical view of the firm as a black box, hiring or buying inputs, transforming inputs into outputs and selling these outputs so as to maximise profits has ceased to be the **only** theory of the firm worth studying. This is not to say that traditional neoclassical economics has nothing to contribute to our understanding of how firms operate. It in fact complements the new organisational theories in that it emphasises the technical relationships within the firm and its market interactions. Furthermore, just as the theory of the consumer is used to derive market demand, the neoclassical model of production provides a building block in the determination of market supply.

Production functions and isoquants

The theory of production is very similar to the theory of the consumer. The same type of reasoning is used, for example, to find conditional input demands (amount of input used to produce a given level of output) as to determine the consumer's optimal consumption bundle. Isoquants are contour lines of the production function whereas indifference curves are contour lines of the utility function. Given this similarity and that you have studied this material in your introductory economics course, I will only give a brief review here. Although the models in this chapter are presented in terms of the firm producing **one** output using one or two inputs, most of the analysis can be generalised to multiproduct firms using many inputs.

It is important to realise that a production function $f(x_1, x_2)$ indicates the **maximum** amount of output which can be produced for the given quantities of inputs x_1 and x_2. Similarly, any input bundle (x_1,x_2) lies on the isoquant corresponding to the highest level of output which can be achieved using these levels of inputs. To derive input demands we will need to know the slope of the isoquants. Since output is constant on the isoquant, we have:

$$d\,f(x_1, x_2) = \frac{\partial f(x_1,x_2)}{\partial x_1} dx_1 + \frac{\partial f(x_1,x_2)}{\partial x_2} dx_2 = 0$$

so that the slope of the isoquant or the **marginal rate of technical substitution** (MRTS) equals the ratio of marginal products:

$$dx_2 / dx_1 = - \frac{\partial f(x_1,x_2)}{\partial x_1} / \frac{\partial f(x_1,x_2)}{\partial x_2} \equiv - MP_1 / MP_2$$

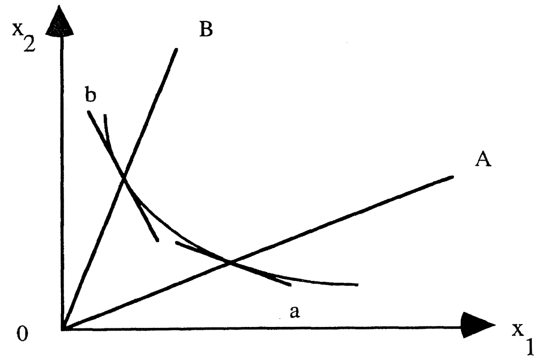
The MRTS is the counterpart of the MRS in consumer theory. It shows the rate at which one input can be substituted for another while maintaining the same output level. For convex isoquants, the MRTS decreases in x_1 which means that if a large quantity of x_1 (and hence a relatively small quantity of x_2) is used, we can reduce the amount of x_1 significantly while increasing the level of x_2 slightly and still produce the same amount. The MRTS is sensitive to the scale of measurement which is why the **elasticity of substitution** is sometimes used to measure the curvature of the isoquant. It is defined as:

$$\sigma = \frac{\text{\% change in the input ratio } x_2 \ / \ x_1}{\text{\% change in MRTS}}. \qquad (1)$$

In Figure 7.1, the change in the input ratio corresponds to the difference in slope of rays OA and OB and the change in the MRTS corresponds to the difference in slope of the lines a and b tangent to the isoquant. Clearly, if for the input ratios determined by A and B, the change in MRTS is larger, then the isoquant is more convex and σ is smaller. The extreme values for s are 0 for perfect complements (L shaped isoquants) and ∞ for perfect substitutes (linear isoquants).

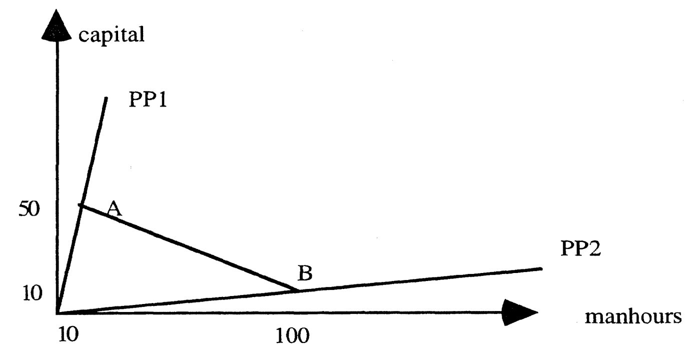
Why? Convince yourself by drawing a graph.

Figure 7.1: Elasticity of substitution



To justify the assumption of smooth continuous isoquants it is useful to think of production processes (or production activities) represented by a ray from the origin. The angle of the ray indicates in which fixed proportions the inputs have to be combined in that particular production process. Suppose that using production process 1 (PP1) you can produce 100 units of output using 10 manhours and 50 units of capital; using production process 2 (PP2) you can produce 100 units of output using 100 manhours and 10 units of capital. For both production processes output increases and decreases proportionally with increases and decreases in inputs. If you want to produce 100 units you have a choice between using exclusively PP1 or PP2 or a combination of the production processes. For example, you could produce 50 units using PP1 (employing five manhours and 25 units of capital) and 50 units using PP2 (employing 50 manhours and five units of capital). If you combine the production processes, the combination of the total amount of inputs used will be on the straight line AB in Figure 7.2.

Figure 7.2: The linear model



Although the linear model of production discussed above is appealing, it is not convenient to work with mathematically. The assumption of smooth convex isoquants, which are convenient, can be seen as a reasonable approximation to the linear model when there are many possible production processes.

Students often get confused about the difference between 'diminishing returns' and 'decreasing returns to scale'. Whereas the first expression refers to decreasing marginal product of **one** input, the latter refers to what happens to the level of output if **all** input levels increase by the same proportion (e.g. for a production function f of two inputs x_1 and x_2, $f(\alpha x_1 , \alpha x_2) < \alpha f(x_1 , x_2)$ for $\alpha > 1$ indicates decreasing returns to scale). Note that it is possible for a production function to have increasing returns to scale for some output levels and decreasing returns to scale for others. You should be able to show that the Cobb-Douglas production function $f(x_1 , x_2) = ax_1^b x_2^c$ has decreasing, constant or increasing returns to scale depending on whether b+c < = > 1.

The notion of decreasing returns to scale is difficult to understand in a realistic setting. If we can produce 100 units with a given set of inputs why can't we just duplicate the set of inputs to produce 200 units? It must be that we are not really able to duplicate **all** inputs and we are therefore keeping some inputs constant. Maybe it is impossible to duplicate the entrepreneurial input or the R&D effort without loss of quality. Increasing returns to scale can arise for example when an increase in man hours allows for specialisation. Also inventories of spare parts and back-up equipment may not have to increase with an increase in output. Sometimes a basic engineering relationship is responsible for generating increasing returns to scale. Think of the output of an oil pipeline as the amount of oil which flows through per time unit. This output is proportional to the square of the radius r of the pipeline (the cross section has area πr^2) whereas the input in terms of materials is proportional to the radius (the circumference is $2\pi r$). Ignoring other inputs such as horsepower needed to create a flow, output is a quadratic function of input. In real world industries, production characterised by constant returns to scale or close to constant returns seems the norm.[1]

The production function reflects the current state of technology; if there is technological progress, the production function changes. If you consider a production function with one variable input, say labour, then the effect of technological progress is to shift the production function up: with the same amount of labour, a higher level of output can be achieved. For two variable inputs, the effect of technological progress can be visualised on the isoquant map: isoquants shift inwards towards the origin. For example, in the steel industry, American and Japanese 'mini-mills' — which are a quarter of the size of giant European firms (and involve a quarter of their capital investment) — use cheap scrap rather than iron ore and coke. To make a tonne of steel they use a quarter of the labour needed in the European firms.

[1] *See Moroney (1967)*

Empirical estimation of production functions is not easy. One popular approach is to use statistical techniques such as regression analysis on time series or cross section data (data on inputs and outputs for several firms or plants in the industry) to estimate a specific functional form. The Cobb-Douglas specification plays a role which is similar to the constant elasticity demand function in consumer theory because of its mathematical convenience: the parameters are estimated directly in linear regression. When cross section data are used to estimate production functions the implicit assumption is made that the same technological possibilities are available to all firms in the industry and hence the same production function applies to all of the firms. Similarly, when using time series data, it is assumed that the state of technology is stable over the period under study. Apart from measurement problems (it is difficult to determine precisely the amount of all inputs used) there is the problem of whether the observations reflect efficient production, which is what the production function should describe. An alternative to using existing data is of course to create new data through experiments.

Firm demand for inputs

It is important to make a distinction between the following two questions:

- given that the firm wants to produce a given level of output, what is its optimal choice of inputs?
- what are the firm's profit maximising input demands and hence what is the firm's optimal level of output?

The answer to the first question gives the conditional input demands (i.e. the amount of inputs the firm demands **conditional** on the level of output it produces). The conditional input demands are used to determine the cost function as we will see later on. The answer to the second question gives the **unconditional** input demands (i.e. the amount of inputs the firm hires or buys to produce the optimal level of output). Clearly the optimal level of output is determined by the structure of the market in which the firm operates. As we will see, there are two avenues to determining the firm's unconditional demand for inputs. We can find the optimal output level first and then set demands for input equal to the conditional demands for this optimal output level. Alternatively, we calculate unconditional demands directly and the optimal output is determined as the maximum output level the firm can produce given these levels of inputs.

Let us start by deriving the conditional input demands. The easiest scenario to analyse is the case of a firm which is a **price taker** on the two input markets. For such a firm the locus of points with equal expenditure E on inputs is a straight line, an **isocost line** ($p_1 x_1 + p_2 x_2 = E$), the analogue of the budget constraint for the consumer. Given the input prices and the output level it wishes to produce with corresponding isoquant I (see Figure 7.3) the firm wants to minimise expenditure by finding the lowest (furthest to the south-west) isocost line tangent to the isoquant I. The graph looks the same as for the consumer problem.

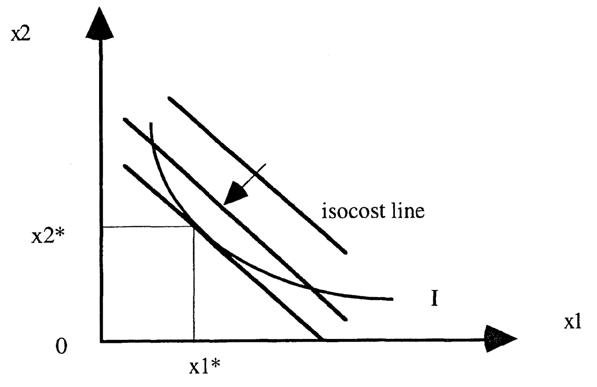
At the optimum (minimum cost) solution (if it is not a corner solution), the slopes of isoquant and isocost line are equal, hence:

$$p_1/p_2 = MRTS = MP_1/MP_2.$$

This implies that, at the optimum, the elasticity of substitution, defined in (1) equals:

$$\sigma = \frac{\%\text{ change in the input ratio } x_2\ /\ x_1}{\%\text{ change in } p_1\ /\ p_2}.$$

Figure 7.3: Optimal input mix



Hence, at the optimum, the input ratio changes significantly when the input price ratio changes if the inputs are substitutes (σ large) whereas, for complementary inputs (with σ small), the input price ratio has to change dramatically to have an effect on the combination of inputs used.

For more than two inputs the optimality condition for cost minimisation subject to an output constraint generalises to 'the ratio of MP to price is equal for all inputs which are actually used'. The intuition behind this result is easy to grasp. If it were the case that the MP-price ratio of input 1 was higher than that of input 2 then, by diverting some money from the budget for input 2 to that for input 1, you could increase output or you could keep output constant and reduce total expenditure. It is important to remember that the input levels which satisfy the tangency conditions are the input demands for the **given level of output**. We can determine what happens to these input levels when the output level q changes. The relationships we find then are precisely the conditional input demands $x_1(q)$, $x_2(q)$. Clearly these conditional input demands depend on the prices of the inputs and the substitutability of the inputs.

Using this simple model of conditional input demands, we can make some interesting predictions. When the interest rate (cost of capital) falls, firms tend to invest in labour-saving equipment. If labour costs differ between economies, economies with relatively cheaper labour will tend to use labour intensive production processes, even when they have the same technological knowledge.

Example 7.1

The production function is given by $q = f(x_1, x_2) = a\, x_1^{1/2} x_2^{1/2}$. To determine the optimal use of inputs, set the MP-price ratio's equal:

$$(a/2)\,(x_2 / x_1)^{1/2} / p_1 = (a/2)\,(x_1 / x_2)^{1/2} / p_2 .$$

This can be simplified to:

$$x_1 / x_2 = p_2 / p_1 \text{ or } x_1.= x_2\,(p_2 / p_1).$$

Using the production function we can substitute $x_2 = (q/a)^2 / x_1$:

$$x_1 =(q/a)^2 (p_2 / p_1)\, /x_1 .$$

The conditional input demands are:

$$x_1 =(q/a)\,(p_2 / p_1)^{1/2} \text{ and } x_2 = (q/a)(p_1 / p_2)^{1/2}.$$

We are now in a position to discuss the abuse of labour productivity measures to construct arguments about technological change and to make international comparisons. To view an increase in labour productivity as technological progress is clearly wrong when the increase could be due to a rise in wages which affects labour's MP-price ratio so that less labour will be used. If less labour is used then, because of the law of diminishing return, its MP has to increase. When international comparisons regarding productivity are made, what is often ignored are the differences in capital intensity between production techniques used in different countries. If Germany uses relatively capital intensive production methods (maybe because of its high wage costs), then its labour productivity will be higher than elsewhere.

Suppose we want to determine how much of each input to use while at the same time deciding how much output *q* to produce. Assume there is only one variable input labour L with unit cost *w*. (The model generalises to any number of inputs.) We will first consider the relatively simple case where the firm is a price taker in the input and output markets. The firm's problem is to use the amount of input necessary to maximise profit:

$$\max p\, f(L) - w\, L$$

where p is the price per unit of output. Let us assume that the second order condition is satisfied. The solution to the maximisation problem is obtained by setting L such that:

$$p\, f'(L) = w, \quad \text{where } f' = df/dL.$$

The marginal return from labour should equal the marginal expenditure on it. This implicitly gives the **unconditional demand** for labour:

$$L = f'^{-1}(w/p), \quad \text{where } f'^{-1} \text{ is the inverse function of } f'.$$

Of course we still have to check that positive profits are made at this solution. Substitute p f' (L) = w in the profit function to get:

$$p\, f(L) - p\, f'(L)\, L$$

which is positive when:

$$f(L)/L > f'(L)$$

(i.e. when average product of labour (AP) exceeds marginal product of labour (MP)). We therefore have to qualify our result by saying that the unconditional demand for labour is:

$$L = f'^{-1}(w/p) \text{ when } AP > MP \text{ and zero otherwise.}$$

Example 7.2

A competitive firm sells its output at price p and has production function $q = L^{1/2}$. To calculate the unconditional labour demand set p x MP equal to w:

$$p\,(1/2\, L^{-1/2}) = w \text{ or } L = (p/2w)^2.$$

Labour demand is increasing in price of output and decreasing in the wage rate.
We have to check that no losses are incurred which is the case since:

$$AP>MP\ (1/L^{1/2} > (1/2)\, L^{-1/2}).$$

Now let us consider the more complicated case of a firm which is not a price taker. In the output market this means that the market price of the firm's output depends on how much it decides to sell. In the input market similarly it means that the price the firm pays per unit of input depends on the quantity it buys. As an example consider a university which is the only employer of cleaners in a small town. If the university wants to hire 10 cleaners it can offer a wage of £7 per hour. However, if 11 cleaners are needed, the wage has to rise to £7.50 per hour (the supply of cleaners is upward sloping). If the university cannot discriminate between the workers, the marginal cost of the 11th worker is not just his salary of £7.50 but also the increase in the other workers' salary: £7.50 + £0.50*10 = £12.50. Intuitively we can deduce that, if a firm is not a price taker, it will hire fewer people because it takes into account the effect of its demand on the wage level whereas a price taking firm ignores this externality.

A **monopsony** is a firm which is the only buyer in a market with many sellers who take the market price as given. A monopsony could be a monopoly or it could sell its output in a competitive market. In some countries farmers have to sell their produce to a national marketing board which (if the domestic market is isolated) operates as a monopoly or sells the produce on the international market at the world price (in this case price taking is likely if the country does not represent a large share of the world market). Let us now consider the most general case of the unconditional input demand model, the profit maximisation problem for the monopsonist — monopolist:

$$\max p(f(L))\, f(L) - w(L)\, L.$$

The monopsonist takes into account the supply of labour function w(L) (i.e. he knows that if he hires more workers he has to offer a higher wage). The first order condition for the optimisation problem is

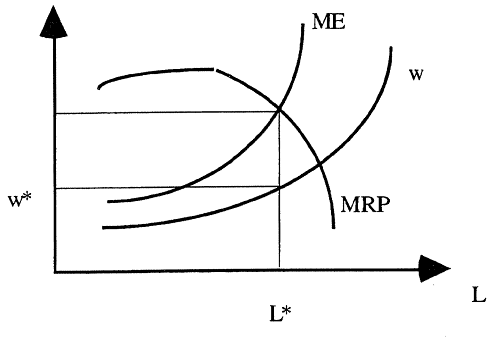
$$p'\ f'\, f(L) + p\, f'(L) - w(L) - L\ w'(L) = 0$$

or

$$(p' f(L) + p)\, f'(L) = w(L) + L\ w'(L)\,.$$

The lefthand side is the marginal return from L or its **marginal revenue product (MRP)**: the product of marginal revenue and MP and the righthand side is the **marginal expenditure (ME)** on L. ME is larger than the wage if we assume supply of labour to be upward sloping ($w' > 0$). The rule MRP = ME captures the demand for inputs in the most general setting; it is applicable to any combination of market power or price taking behaviour in the input and output market. The model of course reduces to the simple case we analysed above: for a price taker, marginal revenue equals price and marginal expenditure equals wage. Figure 7.4 illustrates how the monopsonist's demand for labour is determined. Since ME equals MRP and exceeds the wage, we are justified in saying that a monopsonist pays workers below the value of their marginal product.

Figure 7.4: The MRP rule - monopsony



Many textbooks make comparative statements of the type 'since marginal revenue is less than the price for a monopolist, a monopolist uses less input than a competitive firm'. The implicit assumption here is that a competitive firm operates with the same production function as the monopoly. Also it seems necessary to assume that it is the only firm in the industry because, otherwise, would it not be more interesting to compare the **industry** demand with the monopoly demand? However, if the firm is the only firm in the industry, assuming price taking behavior is not very realistic.

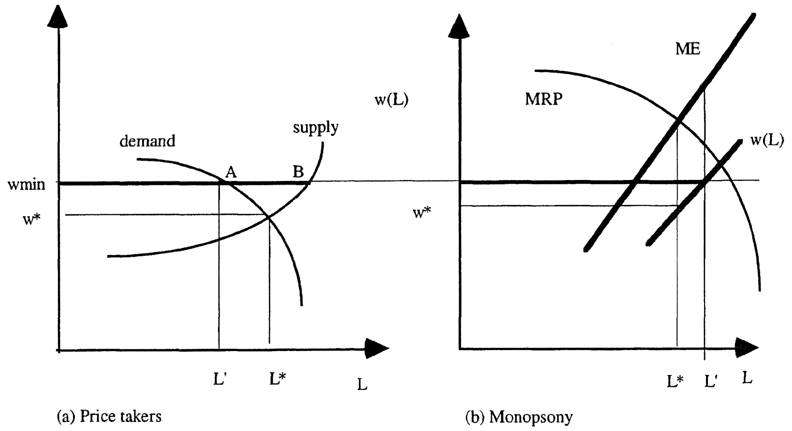
Example 7.3

The demand in an industry is given by p = 100 - 2Q, where Q is industry output. The production function for each firm in the industry is q = f(L) = 2 L and the wage rate is fixed at w = 4. If the industry consists of one monopolistic firm, we determine its labour demand from MRP = ME. Marginal revenue equals MR = 100 - 4 q = 100 - 8L and MP = 2. The optimality condition thus reads (100-8L)2=4 or L* = 98/8 which results in output of q* = 98/4. If the industry consists of a number of competitive firms, then each firm will set MRP = p MP equal to w, or 2p= 4. At the optimum, p has to equal 2 so that industry output equals Q = (100-2)/2 = 98/2 (from the demand function). This requires L*=98/4 units of labour. (Since each firm uses q/2 units of labour, the industry demand for labour is Q/2.)

Case: monopsony and minimum wages

An important objection to minimum wage policy is that it tends to create rather than solve problems for the people it is trying to help because of its effect on unemployment. This argument is very compelling: increase the wage rate and you will reduce the demand for labour, creating unemployment; at the same time a higher wage induces more people to enter the labour market, thereby further increasing unemployment. Surprisingly, the unemployment effect of introducing a minimum wage in a sector differs according to whether the sector is a monopsony. Figure 7.5a represents a labour market with firms acting as price takers. The equilibrium wage w* and employment L* are determined by the intersection of supply and demand. Imposition of a minimum wage w_{min} is equivalent to making the supply curve horizontal at w_{min} up to where the horizontal line intersects the original supply curve. No labour is available at wages below w_{min} and, at wages above w_{min}, labour supply is unchanged. The result is unemployment, corresponding to the distance AB. Employment has decreased from L* to L'.

Figure 7.5: Minimum wages



(a) Price takers

(b) Monopsony

Figure 7.5b represents the situation in monopsony. The effect of the minimum wage on the labour supply curve is as for the price taking sector. However for the monopsonist this has a dramatic effect on ME. He can now hire as many people as are willing to work for the minimum wage without driving wages up. His ME curve is now horizontal at w_{min} up to the intersection with the supply curve at L' from where it coincides with the original ME curve since the expenses on labour are $E=w_{min}$ L for L< L' and w(L)L otherwise. The monopsonist ends up employing L' workers so that employment has in fact **increased**!

Empirical studies suggest that minimum wages do not automatically reduce employment. Consider the recent rise in employment in New Jersey's fast-food industry. In 1992 New Jersey had raised its minimum wage from \$4.25 per hour to \$5.05 per hour whereas neighbouring Pennsylvania stuck to \$4.25. Employment in New Jersey rose by 13 per cent more than in it did in Pennsylvania. A recent forecast suggests that in Britain, a minimum wage equal to half male median earnings (£4 per hour) would have no significant impact on jobs if it did not apply to young workers under 21 years of age.[2] This is not to say that all the evidence points towards harmlessness of minimum wage legislation. In France, where the minimum wage is more than 50 per cent of average earnings, a quarter of people under 25 years old are unemployed. Contrast this with the United States where the minimum wage is approximately 30 per cent of average earnings and there is far less youth unemployment.[3]

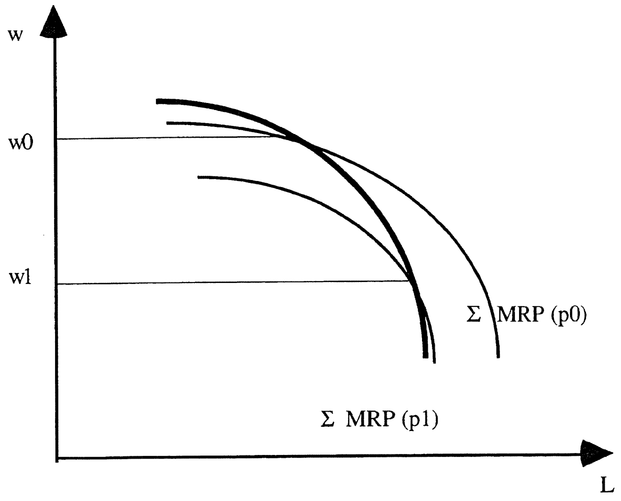
The model of unconditional input demand is interesting because it provides a link between input and output markets. It allows us to predict the effect of changes in the output market, such as moves towards more or less concentration, on the input market. Price wars in the PC market have the effect of reducing the MRP of microprocessors putting pressure on microprocessor prices. Similarly, fare wars in the airline industry have an effect on the aircraft market. As satellite and cable TV gain ground and more TV channels compete for writers, directors and producers, it is likely that wages in these professions will increase.[4]

The model of unconditional input demand summarises the firm's problem. Once we have determined the unconditional demand for inputs for a firm we have solved the problem of how much to produce (i.e. the maximum amount which can be produced given the inputs available) and what price to charge (this can be read off the demand curve given the output level).

Industry demand for inputs

In the consumer theory chapter we concluded that the market demand curve is simply the horizontal sum of individual demand curves. Since I am devoting a separate section to the topic of industry demand for inputs you may already suspect that the situation is not quite as simple here. Of course, if the industry is a monopoly we have no aggregation problem: the industry demand is the firm demand. So let us consider the case of demand for an input, say labour, by a competitive industry. Each individual firm

Figure 7.6: Industry demand for inputs



[2] *See 'The trade unions scent a victory'*

[3] *See 'Doleful'*

[4] *See 'Tuning in to the future'*

in the industry takes the input price or wage w as given and determines its requirement from MRP = p MP = w where p is the output price. Now suppose the wage decreases from w_0 to w_1. Each firm, **given** *p*, wants to hire more workers. However, *p* **is not fixed** in the sense that when all firms in the industry start hiring more workers and producing more output, the output price falls (if the demand curve does not shift). When the output price falls, each individual firm's demand for labour, and hence the horizontal sum of the demands ΣMRP, shifts to the left, as is indicated on Figure 7.6 for a drop in price from p_0 to p_1. We conclude that the industry demand (in bold on the figure) is steeper or less elastic than the sum of firm demands.

There are alternative possible scenarios however. In the analysis above we have ignored the effect of entry and exit in the industry. If the lower wages lead to positive profits, new firms will enter the industry with two counteracting consequences:

- these new firms help to increase total industry output and hence depress output price but
- there are now more firm demands to sum horizontally.

Depending on which of these effects is strongest, the industry demand could be more-or-less elastic than ΣMRP. This analysis has a mirror image in the determination of market supply.[5]

Example 7.4

In Example 7.2 we determined the demand for labour by a competitive firm with production function $q=L^{1/2}$ as $L=(p/2w)^2$. Suppose there are 100 identical firms in this industry and industry demand is given by p = 50 - Q/50 where Q is total industry output (Q=100q). Applying p MP = w for each firm but now taking the effect of industry output on price (via the demand function) into account, we find:

$$(50\text{-}2q)\ (1/2)\ L^{-1/2} = w$$

and after substituting the production function:

$$(50\text{-}2L^{1/2}) \,/\, (2L^{1/2}) = w \text{ or } L = (25 \,/\, (w+1))^2.$$

The industry demand for labour is thus $100(25 / (w+1))^2$ rather than the sum of firm demands $100(p/2w)^2$.

Case: A worked out example of the effect of industry concentration on input price

It is often argued that competition among buyers should be encouraged if the objective is to get a good deal (a high price) for input producers. Concentrated industries are said to be mean buyers of inputs. This argument has been used to abolish state marketing boards for agricultural products for example, and in the same sector, to promote competition among food processing plants. The reasoning is intuitively appealing. When firms act as price takers, they will demand more inputs since they do not worry about the effect of their demand on the input price. Also, a competitive industry is likely to produce more output than a monopoly which again should increase demand for inputs and hence their price. What I want to show you in this example is that, although there may be very good arguments to promote competition in an industry, it is not necessarily always the case that more competition leads to higher input prices.

For simplicity we look at an industry with one input, labour. The supply of labour is given by w = 1 + L. The production function is $q = L^{1/2}$ for each firm in the industry and market demand for the output is given by Q = 2 – p, where Q is total industry output and p is output price.

[5] *See the section on 'Perfect competition' in chapter 9*

Monopoly

Let us consider first what happens if the industry consists of one firm, a monopoly - monopsony. Its total expenditure on labour equals Lw = L (1+L) so that marginal expenditure equals ME = 1+2L. Marginal revenue for the monopolist is MR = 2 – 2q = 2 – $2L^{1/2}$ and MP=(1/2) $L^{-1/2}$. Applying the rule MRP = ME, (and doing some numerical work because this equation cannot be solved analytically), gives an unconditional input demand of L* = 0.18. The corresponding wage can be found from the labour supply function as w = 1.18. Given its demand for labour, we know from the production function how much output the monopolist produces: q=0.42, and hence final product price (from the demand function) is p=1.58.

Perfect competition

Now let us see what happens if a number of (identical) firms in this industry behave as price takers in input and output markets. Each firm sets MRP = p MP equal to ME = w or p /($2L^{1/2}$) = w. To determine the industry demand for labour we cannot just add these individual demand curves as we have seen above. We need to take the output price effect of increased input demand into account. The optimality condition for the firm becomes (2-nq)/($2L^{1/2}$) = w, where n is the number of firms in the industry or, after substituting for q, (2-$nL^{1/2}$)/($2L^{1/2}$) = w. We can solve this latter equation for L and find L = $(w+n/2)^{-2}$ as the amount of labour used by each firm. The total demand for labour in the industry is n times this amount: L^i = n $(w+n/2)^{-2}$ = $4n(2w+n)^{-2}$. The equilibrium wage is such that demand and supply of labour are equal: $4n(2w+n)^{-2}$= w-1. This equation cannot be solved analytically. The table below gives the equilibrium wage (rounded to two decimal places) for some specific values of n. It is not difficult to show that the limiting value of w, as the number of firms n increases, equals 1.

number of firms	equilibrium wage w
5	1.34
10	1.26
20	1.16
100	1.04
1000	1

At first sight we have a counterintuitive result. The more competitive the industry (the larger n) the lower the equilibrium wage. Also, when the number of firms is large, the competitive industry pays lower wages than a monopsonist! To see what is happening here look at the table below which gives the values of the key variables for a monopsonist, a competitive industry with n=10 firms and a competitive industry with n=20 firms. L* is total labour demand in the industry and P is profit per firm.

	w	L*	q	Q	p	Π
monopsony	1.18	0.18	0.42	0.42	1.58	0.45
n=10	1.26	0.26	0.16	1.6	0.4	0.03
n=20	1.16	0.16	0.09	1.79	0.21	0.01

If you compare the monopsony with the competitive industry with n=10 firms, you notice that total output Q is larger in the competitive industry and the demand for labour is higher, hence the higher wage rate. However, as the number of firms in the industry increases from 10 to 20, more output is produced with significantly **less** labour and, as a consequence, the wage rate drops. The explanation for this phenomenon is the decreasing returns to scale production function. An industry consisting of many small plants can produce the output more efficiently than one consisting of a few large plants. I admit that this makes my example rather artificial since if there are decreasing returns, the monopsonist could operate several small plants rather than one large operation. However the analysis above was carried out for zero fixed or sunk costs. If a sunk cost is incurred when a plant is put into operation, there is a tradeoff between efficiency of production and setup costs. (It is a good exercise for you to check how the assumption of positive sunk cost would affect the analysis.) The purpose of my example was to show you that you should analyse an industry and its related industries in detail before you can make general statements.

From production function to cost function

Whereas the production function gives the maximum amount of output which can be produced as a function of the level of inputs, the cost function gives the least cost at which a given level of output can be produced. If there is only one variable input, it is very easy to derive the cost function from the production function. For a production function q=f(L), the input requirement for production of q units is $L = f^{-1}(q)$ and the corresponding cost function $C(q) = w f^{-1}(q)$. If there are several variable inputs, the problem is a bit more involved. By definition of the cost function, we have to determine, for each output level, the cheapest way to produce it. For each output level q we have to find the best combination of inputs. But this is exactly the problem we solved when we determined conditional input demands! The procedure to find the cost function corresponding to some production function is thus to find the conditional input demands and then determine the budget necessary to cover these requirements.

Example 7.1 cont'd

In Example 7.1 we showed that the conditional demands for the production function:

$q = f(x_1,x_2) = a\, x_1^{1/2} x_2^{1/2}$ are $x_1 = (q/a)(p_2 / p_1)^{1/2}$ and $x_2 = (q/a)(p_1 / p_2)^{1/2}$.

The cost function is therefore:

$$C(q) = p_1 x_1 + p_2 x_2 = (q/a)((p_2 / p_1)^{1/2} p_1 + (p_1 / p_2)^{1/2} p_2) = (2q/a)(p_1 p_2)^{1/2}.$$

In the example above we had a constant returns to scale production function and a linear cost function. In fact, this observation holds more generally whenever the firm behaves as a price taker in its input markets. Similarly, increasing (decreasing) returns to scale production functions lead to cost functions which are less than (more than) linearly increasing. For example, the cost function corresponding to the Cobb-Douglas production function:

$$f(x_1, x_2) = a\, x_1^{\,b} x_2^{\,c}$$

is

$$C(q) = k\, p_1^{\,b/(b+c)} p_2^{\,c/(b+c)} q^{1/(b+c)}$$

where k is a constant depending on a, b and c.

Division of output among plants

Suppose a firm operates or is considering operating several plants or establishments from which it serves the same market. A natural planning question which arises is: 'How much output should each plant produce?' Intuitively, we could say that, if several plants are in operation, the marginal cost (MC) in each plant should be equal otherwise the firm could produce the same amount of output at lower cost by transferring output from a plant with high MC to a plant with lower MC. Also, the MC in each plant should equal marginal revenue (MR). To show this mathematically, assume I produce a total output q, q_1 in plant 1, q_2 in plant 2 ($q_1 + q_2 = q$) and the demand curve is given by p(q). My profit maximisation problem is:

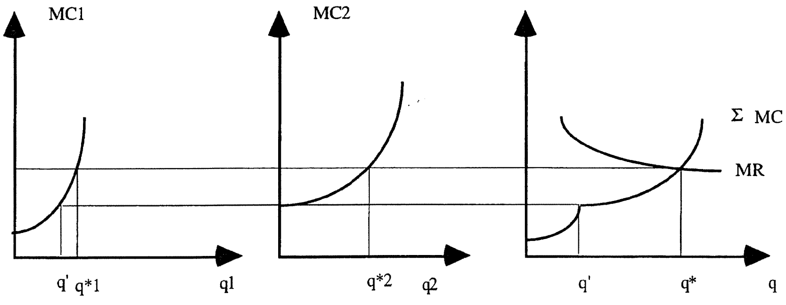
$$\max_{q_1, q_2} \Pi = p(q_1 + q_2)(q_1 + q_2) - C_1(q_1) - C_2(q_2).$$

The first order conditions for an **interior** solution $\partial\pi / \partial q_1 = \partial\pi / \partial q_2 = 0$ result in:

$$MR(q_1 + q_2) = MC_1(q_1) = MC_2(q_2)$$

which confirms our intuition. We can represent the multi-plant problem, and the optimality condition that MR equals MC at each plant, graphically as is done in Figure 7.7. Suppose there are two plants with increasing MC and $MC_1(0) < MC_2(0)$. Clearly, if the firm's total output is less than q' where $MC_1(q') = MC_2(0)$, then only Firm 1 should be used. As output increases beyond q' the firm should start to use Plant 2, adding units of output to the plant with minimal MC. In other words, the firm's aggregate marginal cost curve is the horizontal sum of the plants' MC curves. The firm's optimal output level q is determined by the intersection of this aggregate MC curve with the MR curve. We can trace back from this intersection to the individual plants' MC curves to find the optimal output level for each plant.

Figure 7.7: Division of output among plants



The optimality condition above is only valid for an **interior solution** (i.e. a production plan in which the firm actually produces positive output in both plants). In general we have to compare the profit obtained with such a plan with profit obtained by operating one plant only. If, for example, the fixed costs of Plant 1 are high, it may be better to produce only in Plant 2 even if Plant 2 is less efficient in the sense that its MC is higher. Example 7.5 shows how fixed costs determine where production should take place.

Example 7.5

A firm which faces a demand curve $p(q) = 50–10\ q$ is considering whether it should operate from two establishments and, if so, how much to produce in each. The cost functions associated with the two plants are: $C_1\ (q_1\ ) = F_1 + 10\ q_1 + 10\ q_1^2$ and $C_2\ (q_2\ ) = F_2 + 5\ q_2^2$ respectively, where F_i is fixed cost in Plant i, so that $MC_1(q_1\ ) = 10 + 20\ q_1$ and $MC_2(q_2)=10q_2$.

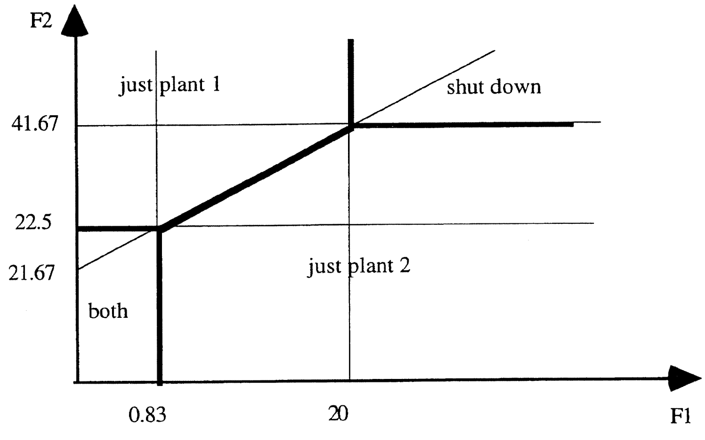
Let us first check the interior optimum by setting $MR(q_1 + q_2) = MC_1(q_1\ ) = MC_2(q_2)$. The marginal revenue corresponding to the demand curve is $MR(q_1 + q_2) = 50 - 20\ (q_1 + q_2)$. Setting this equal to $MC_1(q_1\ ) = 10 + 20\ q_1$ and solving for q_1 gives $q_1 = 1\text{-}q_2\ /2$ (*). Setting $MR(q_1+ q_2)$ equal to $MC_2(q_2)=10q_2$ and substituting (*) for q_1 results in $q_2^0= 3/2$ and, using (*) again, $q_1^0= 1/4$. You should check this result by calculating the marginal revenue and marginal cost functions at q_1^0 and q_2^0. You should also check that if we had $MC_2(q_2) = 60+10q_2$ instead of the original MC function at Plant 2, there is no interior solution i.e. under these conditions it does not pay to produce at Plant 2.

Even when we obtain an interior solution, we have not necessarily identified a profit maximising production plan. The profit level corresponding to $(q_1^0, q_2^0\ )$ is calculated as:

$$\Pi = (50\text{-}10(7/4))(7/4) - F_1 - 10\ (1/4) - 10\ (1/4)^2\ \text{-}F_2\ \text{-}5(3/2)^2 = 42.5 - F_1 - F_2.$$

We compare this to the maximum profit attainable if we operate a single plant. If only Plant 1 is in operation ($q_2 = 0$), its optimal output level is such that $MR = MC_1$ or $50\ \text{-}20\ q_1 = 10 + 20\ q_1$ such that $q_1 = 1$. The corresponding profit level is $\Pi_1 = (40)(1) - F_1 - 10\text{-}10\text{=}20\text{-}F_1$. Similarly, operating just Plant 2 ($q_1 = 0$) leads to $50–20q_2 = 10q_2$ or $q_2 = 5/3$ and corresponding profit level $\Pi_2\text{=}(50–10\,(5/3))(5/3)–F_2–5(5/3)^2 = 41.67–F_2$. Depending on whether Π, Π_1 or Π_2 is largest, we should produce at both plants, at Plant 1 only or at Plant 2 only. Clearly, which is optimal depends on the fixed costs as is summarised in the figure below. For example, Π is largest if $42.5 - F_1 - F_2 > 20 - F_1$ and $42.5 - F_1 - F_2 > 41.67 - F_2$ which implies that operating both plants (the interior solution) is profit maximising if $F_1 < 0.83$ and $F_2 < 22.5$. You should check that Π_1 is largest when $F_2 > \max(21.67 + F_1\ , 22.5)$ and Π_2 is largest when $F_1 > 0.83$ and $F_2 < 21.67 + F_1$. The north-east corner of Figure 7.8 corresponds to the scenario where fixed costs are too large at both plants so that it is impossible for the firm to make positive profits.

Figure 7.8: Fixed costs and multiplant firms



The example above shows how firms can decide whether to operate one or two manufacturing plants. Using this type of reasoning we can explain why real-life firms shift production from one plant to another due to exogenous cost shocks. Hoover, the domestic appliance maker, shut down its plant in Dijon thereby ending 600 jobs in France in January 1993. It decided to concentrate vacuum-cleaner production in Scotland, creating 400 new jobs at its plant near Glasgow. By shifting its production, Hoover slashed its costs by a quarter, partly due to economies of scale (all production in one plant) and the rest from lower wages and non-wage costs such as health insurance which are a lower percentage of overall wages in Britain than in France. The French have accused Britain of 'social dumping'.

Estimation of cost functions

Our analysis of how firms decide on optimal output levels and how they allocate production between plants rests heavily on the firm's knowledge of its cost function. How do firms obtain this knowledge? One approach is to use econometric analysis, using time series and/or cross section data (for several firms or plants in the same industry). Using regression analysis the relationship between cost and output level, input prices, plant size, and so on can be calculated.

As in other applications, we have to decide which functional form to use. Is the cost function linear, quadratic or cubic in output or does a Cobb-Douglas function give a reasonable fit to the data? As we have seen before, the Cobb-Douglas specification, which is very popular, takes the form:

$$C(q) = k\, p_L^{\alpha}\, p_K^{\beta}\, p_E^{\gamma}\, q^{\delta},$$

where p_L, p_K, and p_E are labour, capital and energy (or other input) prices and α, β, γ, δ are the parameters to be estimated.

In practice however, estimating cost functions is problematic. To start with, there are serious problems with the availability of suitable data. Expenditures on capital are difficult to measure. Most firms are multiproduct firms and it is difficult to isolate cost-related data for a single product. Accounting data reflect various ways of allocating overheads over products and may hide important factors such as economies of scope (i.e. the scenario in which it is cheaper to produce several products together rather than separately).

Using time series data leads to additional problems. To get reliable results, we need a reasonable sample size of, say, 40 observations. Ideally we should use monthly observations; it is unreasonable to assume that firms are using the same production technology (and hence operate according to the same cost function) over a period of 40 years! In agriculture for example mechanisation has dramatically changed the labour intensity of many production activities over timespans of 10 or 20 years.

Remember that in the definition of the cost function it is assumed that firms are using the most efficient combination of inputs. If the real life firms in the sample are not operating efficiently, the estimated cost function will be biased. In fact, we cannot expect firms to be operating 'efficiently' at all times; this would require perfect foresight **and** immediate adaptation of scale of inputs to changed market environments.

An alternative method which is used to estimate the cost function is the engineering approach. Engineers are asked to estimate the quantity of inputs such as plant size and machinery required for given levels of output. To estimate the cost function, these requirements are simply multiplied by the price of the inputs. Because of the technical nature of this exercise, the costs of controlling and managing are often overlooked.

Chapter summary

After this chapter and the relevant reading, you should understand:

- how the smooth isoquants continuous production model relates to the linear production model
- the difference between conditional and unconditional input demands
- the problems involved in estimating production and cost functions
- the MRP = ME rule for determining unconditional input demands
- why imposition of a minimum wage may increase employment in a monopsony
- why the industry demand for an input is not necessarily the horizontal sum of firm demands
- why MR should equal MC in each producing plant of a multiplant firm.

You should be able to:

- derive the **cost function** from a given production function *i.e.* derive conditional input demands (input as a function of output) and write the expenditure on these inputs as a function of output
- derive **unconditional input demands** for monopsony, monopoly and price-taking firms
- solve the problem of which plant(s) to operate and at what level for a **two plant firm** given the plants' cost functions and the demand function.

Sample exercises

1. Draw an isoquant map representing increasing returns to scale up to some output level and then decreasing returns to scale.
2. You can cook meals from 'scratch' using fresh ingredients or you can buy frozen dinners. Draw an isoquant assuming the only inputs in meal production are time and money.
3. For the linear version of the production model where every production process is characterised by a ray through the origin in the isoquant plane, show (graphically) the effect of technological progress which enables the same production levels as before with less labour input.
4. In many LDCs, production is constrained by the availability of capital. Show graphically (on an isoquant map) how the cost function reflects this constraint on the amount of capital available. How would you find the cost function analytically?
5. A firm's production function is given by $q(K,L) = K^{1/3} L^{2/3}$ and it faces fixed input prices r (for capital) and w (for labour).
   a. Derive the conditional input demands and the cost function.
   b. If the demand function is given by $q = (21 - 4p)/8$, $w = 1$ and $r = 1/2$, what is the firm's optimal output level? At this output level how much capital and how much labour is used?
   c. Suppose the wage increases from $w =1$ to $w = 8$. If the firm continues to produce the ouput level in (b) what is its new demand for labour i.e. what is the **substitution effect of the change in input price?**
   d. Given the wage increase in (c), what is the new optimal output level and its associated demand for labour? What is the **output effect of the change in input price?**
6. Find the cost function corresponding to the following production functions:
   a. the inputs are perfect complements: $q(x_1 , x_2) = \min (x_1 , x_2)$
   b. the inputs are perfect substitutes: $q(x_1 , x_2) = x_1 + x_2$.
7. **True/false**
   a. The industry demand for an input is less elastic than the horizontal sum of the marginal revenue product curves if the industry is perfectly competitive
   b. same question if the industry consists of local monopolists with independent markets.

Notes

Efficiency wages; internal labour markets

Chapter 8

Topics in labour economics

References cited

‘A bit rich’, *The Economist,* 26 November 1994, 89.

‘A racket in need of reform’, *The Economist, 27* August 1994, 21–28.

‘Bosses on the run’, *The Economist,* 28 January 1995, 26.

Frank, R.H. ‘Are workers paid their marginal products?’ *American Economic Review* (1984) 74(4): 549–71.

Lazear, E.P. ‘Agency, earnings profiles, productivity, and hours restrictions’, *American Economic Review* (1981) 71(4): 606–20.

‘Low risks, high rewards’, *The Economist,* 11 December 1993, 123–25.

Main, B.G., A. Bruce and T. Buck ‘Total hoard remuneration and company performance’, (Discussion Paper, Department of Economics, University of Edinburgh, 1994).

Malcomson, J.M. ‘Work incentives, hierarchy, and internal labor markets’, *Journal of Political Economy* (1984) 92(3): 486–507.

‘Nicely does it’, *The Economist,* 19 March 1994, 94.

‘Ordinary deaths’. *The Economist,* 5 November 1994, 74.

Ransom, M.R. ‘Seniority and monopsony in the academic labor market’, *American Economic Review* (1993) 83(1): 221–33.

Shapiro, C. and J.E. Stiglitz. ‘Equilibrium unemployment as a worker discipline device’, *American Economic Review* (1984) 74(3): 433–44.

‘The high price of freeing markets’, *The Economist,* 19 February 1994, 43–44.

‘The sour taste of gravy’, *The Economist, 5* November 1994, 50.

Yellen, J. ‘Efficiency wage models of unemployment’, *American Economic Review, Papers and Proceedings* (1984) 74(2): 200–08.

The previous chapters have provided all the ingredients of the traditional neoclassical model of the labour market. The supply of labour is derived from individual utility maximising decisions regarding consumption of leisure and other goods. The demand for labour by a firm is derived from the marginal revenue product rule which says that firms hire an amount of labour such that the marginal return (the marginal revenue product or MRP) equals the marginal expenditure. The total demand for labour (as for other inputs) is obtained by summing the firms’ demands taking into account any effects of this aggregation on the marginal revenue of output (see the section on industry demand for inputs). The equilibrium employment and wage levels are determined by the intersection of supply and demand.

Although the traditional neoclassical framework summarised above is undoubtedly an elegant construction which provides a wealth of testable predictions, as it stands, it is neither realistic nor a great help to managers. For a start, the model cannot explain involuntary unemployment. If there is unemployment (demand less than supply), wages simply adjust downwards until an equilibrium is reached. This ignores, among other things, the effect of minimum wage regulations, welfare payments and union power. Furthermore, downward pressure on wages can only exist if the pool of unemployed workers is available where they are needed; if they are in a geographically different location they may face prohibitive moving costs. Also, the theory predicts that wages should be continually adjusted to reflect a worker's productivity whereas in practice wages are fairly stable and are almost never adjusted downwards. Similarly, firms are assumed to adjust the size of their workforce continually in response to changes in demand conditions whereas in practice labour turnover, at least in Europe, is limited.

The MRP rule is hardly a practical guideline for managers making hiring decisions and setting wage levels. At the time a worker is hired, there is uncertainty about his productivity. This problem is likely to be worse for skilled or professional workers for whom, even after they have been hired, it is difficult to measure performance or productivity. Managers do not treat workers like other inputs. In addition to making decisions about how many workers to hire and how much to pay them, there are the important problems of **whom** to hire, whether and how to train them and how to motivate them. To explain certain features of the internal labour market such as promotions, life-time employment with the same company, increasing wages over a career path etc. different models or at least variations on the conventional framework are needed.

Efficiency wages

Standard labour economics models assume that labour is of uniform quality and that this quality can be observed by the employer. Indeed, the employer is assumed to base the worker's reward on his quality or productivity. In reality of course employers do not have perfect information about workers' quality or (lack of) effort. If the employer cannot detect 'cheating' such as entertaining friends instead of customers on the business expense account, selling company secrets, taking bribes, working less than contracted for, etc. then workers will cheat. Even when an employer can detect these activities, the possibilities for punishment are limited. According to recent reports in China's *Legal Daily*, a worker in an American joint venture company was forced to lick clean a glass product which the supervisor found dirty and a woman who stole two pairs of shoes from a Taiwanese factory was made to sit in a dog cage.[1] These types of punishments are not ordinarily available to employers.

In many cases, the only means of disciplining a worker is to fire him or her. However, if the terminated worker can immediately find a new job at the same wage (there is no unemployment), even termination is not an effective punishment and hence threat of termination will not induce good behaviour. Only when workers fear that they will not be able to find a job or at least not as good a job, are they deterred from cheating by the threat of termination. The question then arises whether by paying workers a wage premium in the form of an **efficiency wage**, the employer can induce good quality work. Note that 'good quality work' can have many interpretations ranging from higher productivity and better attitudes towards customers to reduced absenteeism and staff turnover rate. If workers are paid a wage higher than the market wage, they incur a real penalty when they are fired. This question was examined by Shapiro and Stiglitz (1984) in their seminal paper on efficiency wages.[2] Shapiro and Stiglitz focused on the consequences of payment of efficiency wages on unemployment. If firms find it profitable to pay higher wages, the argument goes, then the demand for labour decreases which creates unemployment. Furthermore, unemployment is not only a by-product of efficiency wages, it is a necessary ingredient in the efficiency wage model. If there is full employment, as mentioned above, the threat of termination is not effective. Unemployment is a worker discipline device.

[1] *See 'Ordinary deaths'*

[2] *Yellen (1984) provides a good overview of efficiency wage models*

A simple efficiency wage model

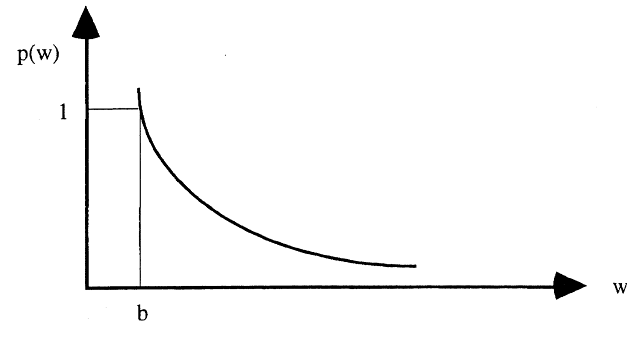
Let us look at a simple version of the efficiency wage model. Suppose the employee gains g from cheating if undetected. If the market wage is w_0 and there is no unemployment, firing is not a threat because a fired worker can get another job at the same wage w_0. It is important in this model that there is no stigma attached to being fired. All workers are assumed identical and equally likely to misbehave given the same circumstances. This implies that it is not rational for an employer to discriminate against workers fired from a previous job. Given that termination is not a punishment, all workers cheat and firms do not spend any resources stopping them. Workers get a payment $w_0 + g$ per period; w_0 from the employer and g as a bribe for example.

Now consider the problem of an employer who is willing to pay a higher wage w to induce workers to be honest. How much does he have to pay? The employer will monitor workers and catch cheaters with probability p. Let's just look at one period and assume that cheaters are caught (and fired) at the beginning of the period. If a worker is fired, he receives unemployment benefit b. We assume here that the higher wages lead to unemployment. A risk neutral worker cheats if the expected benefit from doing so exceeds his wage:

$$(w+g)(1-p) + bp > w \text{ or } w < b + g(1-p)/p \equiv w_1. \qquad (1)$$

The efficiency wage is w_1; it is the lowest wage the employer can pay if he wants the workers to be honest. The efficiency wage is increasing in the gain the worker gets from cheating. If the temptations to cheat are large so has to be the compensation for refraining from cheating. A rise in unemployment benefit also drives up the efficiency wage because it lowers the punishment value of getting fired. If workers are not likely to be caught (p low) the efficiency wage has to be relatively higher than when the firm can detect cheating easily. This last relationship illustrates the tradeoff a firm faces if it can determine the probability p, through appropriate choice of monitoring intensity, as well as the wage w_1. Either the firm pays high wages and does little monitoring or it monitors seriously and can get honesty for lower wages. This is illustrated in figure 8.1 below. All combinations to the north-east of the curve ensure that workers are honest. Generally, monitoring activities designed to increase the detection probability p are costly. This implies that a cost minimising firm chooses a combination of p and w on the curve.

Figure 8.1: Monitoring-wage tradeoff



Suppose the firm can fix p by allocating resources to monitoring and incurring a monitoring cost M(p) per worker. If the firm minimises total wage and monitoring expenditure per worker subject to inducing honest behaviour, its problem, which is known as the **minimum cost implementation problem** is:

$$\min w_1 (p) + M(p) = b + g(1-p)/p + M(p).$$

Assuming an interior solution and assuming the second order condition is satisfied, we find:

$$p^2 \, M'(p) = g. \qquad (2)$$

If $M'(p) > 0$ and not decreasing in p, (2) implies that, if the gain to the worker is high, it is in the firm's interest to devote a significant resource to monitoring.

For the firm, the move from paying w_0 and suffering dishonesty to the efficiency wage scenario is profitable if:

$$w_0 + C > w_1 + M(p) \qquad (3)$$

where C is the cost to the firm caused by the worker's cheating. Hence, the model does not predict that efficiency wages are paid by all firms in all circumstances. It is quite possible for an employer to be better off permitting bad behaviour than punishing it if the necessary rise in wage and monitoring cost is high. For example, restaurant owners often allow staff to 'steal' meals or snacks. Many business employees 'cheat' by inflating their business travel expenses.

We have determined the wage w_1 the firm would have to pay to keep a worker **who joins the firm** from cheating. We still have to check that the firm can actually attract or retain workers at this wage. If the alternative for the worker is to stay at wage level w_0, it has to be the case that:

$$w_1 > w_0 + g. \qquad (4)$$

Clearly, C has to be larger than g; the cost to the firm due to the worker shirking has to be larger than the gain to the worker for both condition (3) and condition (4) to be satisfied. If these conditions **are** satisfied, the move to an efficiency wage situation constitutes a Pareto improvement: both the firm and the workers are better off. Efficiency is increased hence the terminology 'efficiency wages'! The efficiency wage is higher than the original market wage w_0 and, if the MRP is unaffected, the demand for labour decreases and unemployment is generated. In this simple model, a rise in unemployment benefit is likely to increase the efficiency wage and hence increase unemployment.

Real-life firms and some enlightened management consultants know that being nice to workers may pay off. In the efficiency wage model, it is shown that it may be possible to increase the quality of work, reduce staff turnover or eliminate unproductive behavior by paying higher wages. Nordstrom, an American department store chain, pays its sales staff twice the industry average in the hope of attracting and retaining good workers. Clearly similar arguments could be made for non-pecuniary rewards for workers. It may be more profitable (less costly) for a firm to increase job satisfaction than to increase wages. Employees may appreciate non-monetary aspects of their jobs such as company cars, training programmes, pride in their work, more responsibility and autonomy. Chaparral Steel, a Texan steel producer, allows workers to travel around the world to select their own equipment. Every employee of Motorola spends at least a week every year in training. Levi Strauss 'empowers' workers by allowing them to redesign parts of the production process. In half of America's large companies, workers are organised in self-managing teams.[3] When firms create a worker-friendly environment, the effect on monitoring costs is equivalent to that of an increase in wage. If firms are nice to workers, their tradeoff between wage and detection probability (see Figure 8.1) can be shifted to the left so that, for the same amount of monitoring, firms can pay a lower efficiency wage or, for the same wage, they can do less monitoring. Within organisations, the latter possibility creates conflict as the reliance on middle managers to do the monitoring is diminished.

[3] *See 'Nicely does it'*

Case: Politicians, sleaze and efficiency wages

Politicians regularly exploit an efficiency wage argument to award themselves salary increases. In theory at least, paying a cabinet minister or any other government official a high wage should make him think twice about accepting bribes and getting involved in corruption and other scandals. In 1994 government ministers in Singapore saw their salaries increase by about 25 per cent. The Singapore government argues that high wages for officials have ensured that Singapore, as one of very few Asian countries, does not have a corruption problem. The starting pay for a cabinet minister in Singapore is \$419,285 a year. The annual salary of the prime minister is \$780,000, a very nice reward indeed compared to the \$200,000 Bill Clinton has to get by on or the measly \$122,000 for John Major. We can offer at least a partial explanation for these differences in pay in terms of our simple model of tradeoff between monitoring and wage. In most western democracies, politicians are monitored by the media. The more intrusive the media, the lower the efficiency wage! Singapore, for example, would not tolerate a press as aggressive as the British press.[4]

In the United States scandals involving politicians are relatively rare compared to Europe. This is even more surprising when one considers the enormous efforts of the American lobbying industry to influence politics. The US offers its elected representatives a salary of about \$140,000, nearly three times the amount a British Member of Parliament is paid. The secretarial and research allowance for an American politician is more than eight times the British allowance (\$577,000 versus \$68,000).[5]

In Russia, public officials and soldiers are demoralised and underpaid. Corruption, bribery and even sale of government property such as weapons is the order of the day, fuelling crime. The police are paid so badly that in 1992 more than 2000 crimes were blamed on police officers.[6]

[4] *See 'A bit rich'*

[5] *See 'The sour taste of gravy'*

[6] *See 'The high price offreeing markets'*

Efficiency wages and minimum wages

It is possible to discuss minimum wage legislation in the simple efficiency wage framework outlined above. Suppose the initial situation is as above with all firms paying low wages w_0, all workers cheating and no unemployment. The workers receive a payoff of $w_0 + g$ and the cost per worker to the firm is $w_0 + C$. Assume this is an equilibrium (i.e. it is not in any firm's interest to deviate by paying a higher (efficiency) wage). If a firm considers paying an efficiency wage w_1 then to induce honest behaviour w_1 has to satisfy:

$$(w_1 + g)(1-p) + (w_0 + g)\,p < w_1 \text{ or } w_1 > w_0 + g/p$$

where the first term in the first inequality represents the payoff corresponding to cheating and not being caught and the second term represents the payoff corresponding to cheating, being caught and getting a job elsewhere at the low wage w_0. Suppose the solution to the firm's minimum cost implementation problem is the wage w_1 and detection probability $p(w_1)$. Hence the firm would incur a total cost of $w_1 + M(p(w_1))$ per worker. It is not profitable for the firm to implement this solution when this cost exceeds its original cost (i.e. when $w_1 + M(p(w_1)) > w_0 + C$). Only an employer who incurs a high cost of worker misbehaviour would find it profitable to pay the high wage w_1.

However, if all firms paid a higher wage and this leads to unemployment then the worker's fall back position is lowered to the unemployment benefit. The resulting efficiency wage (see (1)) w_1 is now lower:

$$w_1 > g(1-p)/p + b.$$

Suppose the solution of the minimum cost implementation problem is w' and p(w') and the cost per worker is lower than the original cost: $w' + M(p(w')) < w_0 + C$. This means that, if all firms could coordinate their employment policies and offer the efficiency wage w', they would all be better off. They are in fact stuck in a low pay-low worker quality equilibrium from which unilateral deviations are not profitable. The government can facilitate firms' coordination by setting a minimum wage level at w'. If it becomes illegal to offer wages below w', all firms will pay w' and no firm has difficulties attracting workers. Do any welfare gains result from the introduction of the minimum wage legislation? It turns out that, as long as the monitoring costs are not too high, there may be welfare improvement. Total welfare, is initially g – C and, after the minimum wage, ignoring unemployment, it is $-M(p(w'))$ which represents an increase if $M(p(w')) - C < -g$. The minimum wage legislation could improve welfare if the monitoring costs are low relative to the cost of shirking and if the worker's gain from shirking is not too large. Of course, unemployment increases and so the net welfare effect of introducing a minimum wage is ambiguous.

Are the workers better off when the government introduces minimum efficiency wages? The answer is ambiguous. The workers who keep their jobs are now receiving $w' = g(1-p)/p + b$ compared to $w_0 + g$. Given $b<w_0$, workers are likely to be worse off unless p is low which it would be if firms face high costs of monitoring. Of course the workers who lose their jobs are worse off. The move to efficiency wages cannot be a Pareto improvement.

Firm demand for labour

Let us return now to efficiency wage theory and see if we can model the effect of the introduction of an efficiency wage on the firm's demand for labour. We already know that, for the efficiency wage theory to work, there should be some unemployment. However, this is not to say that each firm should reduce its labour demand. Suppose workers choose a low or high effort level e_L or e_H which gives them a disutility of c_L and c_H respectively and delivers an **expected** output of q_L or q_H respectively. It is important that an individual worker's output cannot costlessly be observed by the firm. However, if the firm monitors the worker, it can gain information about whether the worker is working hard (e_H) or not (e_L). In particular, if the worker is shirking (e_L), he has a chance p of being caught. The firm decides on p and incurs a monitoring cost M(p) per worker. All L workers are identical and, if they all work hard, the expected output is Lq_H whereas, if they shirk, total expected output is Lq_L. Both the workers and the firm are risk neutral.

If it is impossible for the firm to monitor workers, workers have no incentive to work hard since their shirking is not detected. The firm thus determines its demand for labour from:

$$\max_{L} Lq_L\, P(Lq_L) - w\, L$$

where P(Q) is the firm's demand function (i.e. the price the firm can charge if it sells Q units of output) and $w > c_L$ is the wage rate. The result of this optimisation problem is the familiar MRP = ME condition:

$$\left(P(Lq_L) + \tfrac{\partial P}{\partial Q} Lq_L\right) q_L = w. \qquad (5)$$

Now suppose it becomes possible for firms to monitor. Consider the problem of a firm wishing to pay an efficiency wage. Assume the market wage is w i.e. a worker has a choice between working for the market wage w and shirking or working for the efficiency wage w' at a high effort level. The firm, when it detects shirking, fires the worker and pays him zero. A shirking worker who is not caught gets the same wage w' as a non-shirker. Workers are therefore attracted to the firm (anticipating that they will have to work hard there) if:

$$w' - c_H > w - c_L. \qquad (6)$$

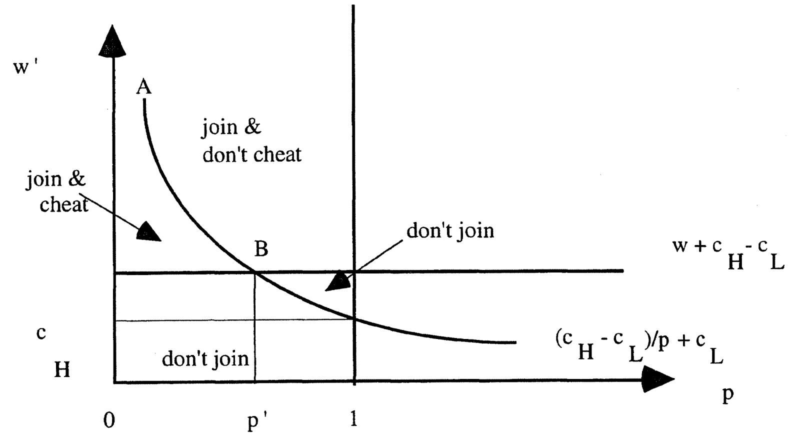
Once they join the firm, they work hard if the payoff of working hard exceeds the expected payoff of shirking:

$$w' - c_H > (1-p)\,(w' - c_L). \qquad (7)$$

It is assumed in (7) that firms detect worker shirking before workers have suffered the disutility and that a worker who is caught shirking cannot start working elsewhere at wage w immediately. The two conditions (6) and (7) which are very similar to the participation and incentive compatibility constraints in principal- agent problems, are represented graphically in Figure 8.2. Note that (7) can be rewritten as:

$$w' > (c_H - (1\text{-}p)\, c_L) / p = (c_H - c_L)/p + c_L\,.$$

Figure 8.2: Conditions on the efficiency wage



Now the firm maximises total profit subject to conditions (6) and (7):

$$\max_{w',\,p,\,L} \quad Lq_H\, P(Lq_H) - w'L - M(p)\,L \qquad (8)$$

$$\text{s.t. } w' > w + c_H - c_L \text{ and } w' > (c_H - c_L) / p + c_L.$$

The feasible set of combinations of w' and p is the 'join & don't cheat' area in the figure above.

Define p' as the minimum detection probability necessary to induce honest behaviour if the firm pays the minimum wage $w + c_H - c_L$ which attracts workers.

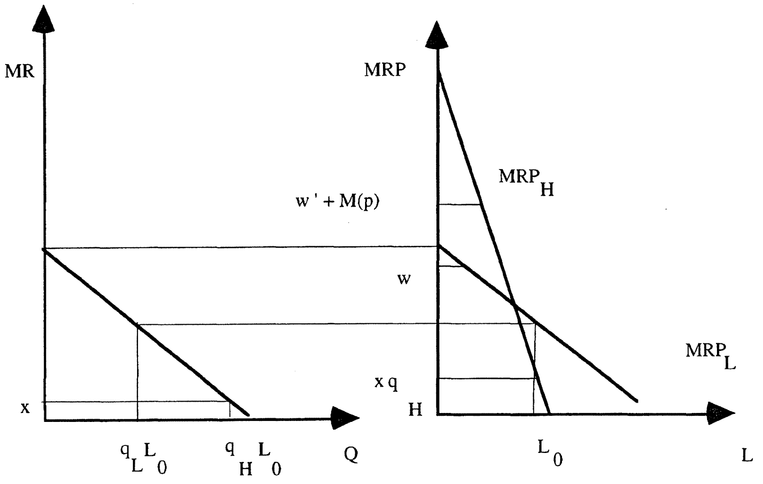
We can find p' at the intersection of the curves corresponding to conditions (6) and (7) as $p' = (c_H\text{-}c_L)/(w + c_H\text{-}2c_L)$. Clearly it can never be optimal for the firm to set p higher than p' because it can guarantee that the worker does not shirk at any wage above w + c_H - c_L with a lower detection probability p. The solution to the firm's problem therefore has to be a point on the curve AB. In addition, at the optimal solution of (8), the MRP = ME rule has to hold:

$$MRP_H = MR(q_H\,L)\, q_H = w' + M(p) \qquad (9)$$

where MR(Q) is the marginal revenue function. This is of course very similar to the optimality condition (5) before the introduction of an efficiency wage which can be rewritten as:

$$MRP_L = MR(q_L\,L)\, q_L = w. \qquad (10)$$

Figure 8.3: Effect of efficiency wage on firm demand for labour



(a) Marginal revenue

(b) Marginal revenue product

To keep things simple, let us assume the MR function is linear and set $q_L = 1$ and $q_H > 1$. Figure 8.3a represents the MR function. Figure 8.3b gives the corresponding MRP functions before (MRP_L) and after (MRP_H) the introduction of the efficiency wage. These curves are derived from the MR curve as follows. Consider MRP_L first. From (10) we know that:

$$MRP_L\ (L) = MR(L) \text{ for } q_L = 1.$$

Hence for L=0, MR(L) is given by the intercept in Figure 8.3a. For $L=L_0$, we can read $MR(L_0)$ on Figure 8.3a. These two points are sufficient to draw MRP_L in Figure 8.3b. Now consider MRP_H. For L = 0, $MRP_H = MR(0)q_H$ which gives us the intercept for MRP_H in Figure 8.3b. For $L = L_0$, we see that $MR(q_H\ L_0) = x$ in Figure 8.3a which gives $MRP_H\ (L_0) = q_H\ x$ in Figure 8.3b.

Conditions (9) and (10) tell us that the firm demand for labour is determined by the intersection of MRP_L and w and the intersection of MRP_H and $w' + M(p)$ respectively. If the initial wage w is relatively high as on the figure and if the efficiency wage and monitoring costs are not too high, the firm's demand for labour could increase when it starts paying efficiency wages!

Internal labour markets

Real-life companies often have employment and compensation policies which seem to bear no resemblance to the neoclassical result that labour should be hired and paid according to the MRP rule. Workers are not laid off whenever there is a temporary shift in marginal revenue; wages are often determined according to a fixed scale and the generally low variance in wages does not seem consistent with workers being paid the value of their marginal product in each period. Many employees are in an internal labour market and move up the hierarchy inside an organisation. Mobility in and out of the internal labour market is limited so that conditions in the external market have only a limited effect on the internal market. This is especially true for white collar jobs, professionals and managers or primary sector workers. The internal labour market is insulated to some extent from the external labour market except at a few entry points such as the hiring of graduates where firms compete with other employers. Workers may have outside options during their careers and firms may hire from outside for higher level jobs. Only in these situations is a firm forced to pay market wages. In this section we look at some of these features of employment policies in organisations and discuss alternative explanations economists have put forward.

Why do wages rise over a career path?

In most jobs employees are paid a relatively low wage when they start work and the wage gradually increases with years in the job or **seniority.** We could explain this phenomenon by referring to the higher value of more senior workers because of their acquired (firm-specific) **human capital** which means the knowledge and skills which determine productivity. More senior workers have had more formal training as well as on-the-job experience which is likely to make them more productive relative to job entrants. Certainly, in some situations this explanation is valid but overall it does not seem reasonable to maintain that the most senior workers are the most productive. Wages generally rise faster than productivity. It is interesting to note that university professors (in the USA) are an exception to the rule of salaries increasing with seniority. In a recent study, Ransom (1993) found that for his samples of university professors, after controlling for experience and some other variables, there was actually a negative correlation between seniority and salary.

In some circumstances, human capital theory offers an explanation for increases in wages. Firms investing heavily in **training** which increases the workers' value not only in their current job but elsewhere as well, tend to pay new trainee employees a low wage. This wage is increased after the training is completed. Examples of this are found in accountancy, law and architecture. The reason for the jump in wage is that, while the employee is developing marketable skills at the firm's expense, he pays for this training by working at a low wage. As soon as the training phase ends and outside offers would be forthcoming, the firm is in a position to offer a higher wage and retain the trained workers. What is harder to explain is that older workers are paid more than younger and sometimes more productive workers with the same training and qualifications.

Lazear (1981) uses an **efficiency** argument to explain such rising wage profiles. Workers are paid less than their MRP early in their careers and more than their MRP later in their careers so that **over their lifetime** they are getting the value of their marginal product. Given this wage profile it is obviously painful for a worker to lose his job after he has worked with a firm for a few years: he loses his 'investment' in the form of larger wages later on. Therefore workers with rising wages are more likely to work hard and avoid being fired whereas workers whose wages do not increase do not have the same incentive to put in high effort. By offering increasing wages the employer can succeed in making both the worker and himself better off because higher efforts by the worker lead to higher productivity. The increased cost of shirking leads to a Pareto improvement.

Because workers are paid very high wages when they are older, mandatory retirement is necessary otherwise workers would prefer to continue working after they had been paid their lifetime value to the firm. Also, there has to be some mechanism to prevent the firm from firing workers (without cause) before it has paid them the higher wages to which they are entitled. If firms care about their reputations, they will not renege on the (implicit) agreement to pay the workers their lifetime productivity value by firing them early because then they will not be able to attract workers in the future.

The efficiency explanation above rests on the idea of using rising wages as an incentive mechanism. In jobs where the employer can easily monitor whether the worker is performing adequately or where alternative incentive mechanisms such as piece-rates or commissions can be used, we would not expect to see wages rise much with seniority. Conversely, in large organisations where monitoring costs are high, rising wage profiles are likely.

Efficiency wage theory can also explain why wages rise over a career path up to a point where they exceed marginal revenue product. Initially employees are usually offered low responsibility jobs which are easier to monitor. In terms of the minimum cost implementation problem, this implies that the firm can set a high detection probability p at low cost M(p) which in turn leads to a relatively low wage. Later on, when the employee has more responsibility, the monitoring cost increases so less monitoring will be done and a higher wage results.

Firms also offer seniority-based pay for jobs where it is important to retain people. Paying a low salary initially and compensating later **screens for loyalty** in the sense that only workers who intend to stay with the firm over a longer period find this an attractive proposition. Firms also use 'golden handcuffs' such as deferred compensation and unvested pensions as screening devices. For example, Bell South, an American regional telephone company, encourages employees to put up to 25 per cent of their pay in a special account. The company augments this contribution and, when they retire, the employees get more than twice the normal market interest whereas, if they quit, they just get the market interest.

When an employer hires a new worker neither she nor the worker knows for sure how productive the worker will be. Workers are typically **risk averse** and would prefer a contract which offers them a secure job at a wage which corresponds to average productivity to a contract offering them a wage corresponding to their actual productivity. However, if a firm pays a wage equal to the value of the average productivity it will make a loss: the low productivity workers will stay and the highly productive workers will leave for better outside offers. To guard against this, the firm has to start everyone at a low wage and increase wages of workers who turn out to be productive. The difference between the low initial wage and MRP for the productive workers is like an insurance premium they pay at the time when they do not know their productivity.

Finally, in cross section data, wages can show positive correlation with seniority even if **individual** workers do not experience rising wage profiles. There are several explanations for this. Firstly, workers who are productive in their jobs are highly paid and are likely to stay whereas workers who are not productive and get low pay tend to leave. As a consequence, we find that the highly paid workers have higher seniority. Efficiency wage theory offers an alternative explanation: firms which pay efficiency wages have low turnover which means that workers in these higher paying firms have higher seniority again generating the positive correlation.

Why is there long-term employment?

In contrast to the assumption of perfect mobility of labour underlying the neoclassical model, firms do not lay off workers when there is a temporary reduction in demand and they do not start hiring immediately when demand for their products increases. Workers similarly do not usually change jobs when slightly higher wage offers are made. Apart from the obvious explanations of moving costs for workers and legal impediments and costs of firing workers, are there other reasons for the firm-worker relationship to be permanent?

One reason for long-term employment is the existence of **firm-specific human capital.** Employees may acquire knowledge and skills which increase their productivity **as long as** they stay in the same job but which are not useful for other jobs. From an efficiency point of view then, long-term commitment is necessary for the firm and/or the worker to have an incentive to invest in firm-specific human capital.

Sometimes it is necessary to offer a long-term contract to ensure that the employee's incentives are aligned with the **firm's long-term objectives**. If the employee's horizon is one year away, he may not care about the firm's profitability several years into the future. It may also be difficult to judge an employee's performance in the short-term if the effect of his work is not felt until much later. This is likely to be the case for managerial jobs.

Some versions of **efficiency wage theory** can also explain long-term employment. If workers are not offered job security they are more likely to shirk because the punishment of losing their job is not as severe.

What is the role of promotions?

Generally promotions help to assign people to jobs where they are most productive. However, promotions are also used as rewards to provide incentives for good performance.

A problem arises when these two goals are in conflict. Good performers at one level of the hierarchy may not be very productive when promoted to a higher level. This is especially likely when the nature of the job changes dramatically after promotion. To avoid the problem of good technical employees (engineers and researchers) becoming mediocre supervisors and managers, companies such as IBM and 3M use separate career ladders for scientists and managers.

Promotion is a reward not only because it brings with it a higher salary but it also gives the worker a chance to compete in the next round to get promoted further. Since the 1980s, firms have eliminated levels of management and, as a consequence, promotion is less of an incentive: at any level there are more candidates and hence a lower probability of getting promoted. Using promotion as a reward is not entirely harmless. When workers are involved in team work, rewards based on relative performance may lead to reduced cooperation. Another problem which can arise when relative performance assessment is used is that of worker collusion.

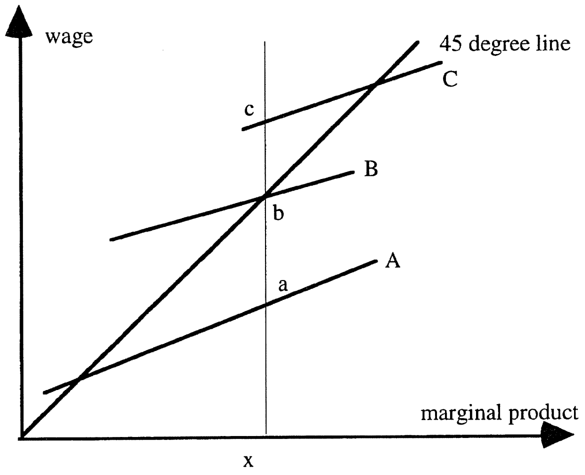
Malcomson (1984) proposes an interesting moral hazard explanation for promotions. The starting point is that workers are paid more after promotion. In many situations, especially when a substantial amount of team work is involved, the employer cannot assess absolute individual performance accurately. Relative performance is easier to judge. If the employer would promise to pay according to individual performance, workers may doubt his honesty. There is a moral hazard problem on the part of the employer in that he could claim that no workers performed up to the standard. The workers anticipate that the employer will misbehave in this way and are not encouraged to work hard. If instead the employer promises to reward relative performance by promoting the top 10 performers, workers can easily check that the promise was kept. Of course it is then in the employer's interest to select the best performers to encourage all workers to work hard.

Usually workers at higher levels of a firm's hierarchy have been promoted from within the firm. There are only a few ports of entry where outsiders may be recruited. This is consistent with firms ignorant of individual abilities paying low wages for young workers of mixed productivity and screening them to learn which are best. This points to a reason for the firm to delay promotion whenever it can. Promotion makes an employee more attractive to outside firms because it acts as a signal of worker quality. Higher wages have to be offered after promotion to prevent the worker taking up attractive outside offers.

Wage compression

Another feature of internal labour markets is that within firms there seems to be less variability in wages than in productivity. An explanation for this fact could be that it is just too difficult to measure individual productivity especially where team work is involved. A more interesting explanation is given by Frank (1984) in terms of status. The starting point is the idea that people care about their position in the income hierarchy of the firm and they care more the more they interact with their co-workers. Employees are willing to sacrifice some income if they can be the highest paid in the firm and conversely they need to be compensated for being the being the one with the lowest pay. This is illustrated in Figure 8.4. The three lines crossing the 45 degree line represent wage schedules for three firms. An individual with marginal productivity x could choose to work in Firm A where he is near the top of the salary scale and enjoy relatively high status. He would then pay a price ab in terms of the difference between his wage and his marginal productivity. If he does not care at all about relative status the individual with marginal productivity *x* could earn a wage premium *bc* by working for Firm C near the bottom of its salary scale.

Figure 8.4: Status and wage compression



Managerial and executive pay

It may be difficult to measure performance or productivity at the lower echelons of an organisation; it is very hard to gauge performance of managers and executives. This is no doubt why many firms have fixed salary scales attaching salaries to jobs rather than people in these jobs. It is unrealistic to assume that everyone in a particular job would have the same productivity but, since productivity is so hard to measure, this may be the best that can be done. Fixed salary scales show increased compensation for jobs which carry higher responsibility and higher diversity of tasks. When measures such as number of people reporting to a manager are used to determine compensation, perverse incentives may be given to encourage bureaucratic expension.

When a manager can reasonably be expected to have a significant influence on the profitability of the firm, performance-related pay in the form of bonuses of a percentage share in profits can be used. However, this may interfere with managers being risk neutral with respect to investment decisions in the firm. When pay is related to performance, managers may take decisions which maximise their own risk averse utility functions. This could even be a problem when pay is not performance-related but managers consider that their future employment possibilities depend on their current company's results. In Japan, where job mobility is very low, these considerations are less important and this, in addition to the concept of management by consensus where no single manager carries responsibility for failure of projects, should make Japanese managers less risk averse.

In 1990, total CEO (Chief Executive Officer) compensation for the 200 largest American companies averaged almost US$3 million per year.[7] The figures for Europe and Japan are much less although there has recently been public outrage in Britain over the 'excessive' salary increases top executives of newly privatised industries award themselves. In January 1995, Cedric Brown, the CEO of British Gas was asked to justify to Parliament his 75 per cent salary increase to £475,000.[8] One of the reasons for these huge remuneration packages seems to be the use of incentive schemes for high-level executives. The fixed salary part of the compensation package may be relatively modest but CEOs typically also receive stock options which are rights to buy company shares at or above the current share price. Clearly, there is no downside risk involved here. If the share price is low during the specified time in which the options can be exercised, the manager does not buy any shares. Nevertheless share options can be a useful device to lengthen the CEO's horizon and give him an incentive to maximise the long-term value of the firm. The remuneration package may also contain straightforward stock awards which consist of shares given or sold at a discount. Sometimes there are restrictions on the sale of these shares, for example, they may only be sold after retirement. This is another example of the use of **golden handcuffs** to align an employee's interest to the long-term future of the company.

Another reason for what seems like excessive compensation at the top of large firms is that, once an employee reaches this highest level, promotion can no longer be used as a reward. There is also a 'political' explanation for high executive compensation. Non-executive directors who determine executive pay clearly have an interest in exaggerating these pay awards as they are often executives elsewhere and can only benefit if the 'going rate' for top jobs rises. A similar story can be told to explain the inactivity of institutional shareholders. Top executives of banks and pension funds hardly have an interest in calling attention to high pay.[9] The absolute size of CEO pay should not be a major concern of shareholders as long as there is a link with performance. The structure of pay and the provision of incentives is what matters. It seems that too often executives are paid large amounts of money when their companies are not doing so well. It is all the more surprising therefore that large British institutional investors call for ceilings on the amount of share options that can be issued to executives. The Association of British Insurers (ABI) and the National Association of Pension Funds (NAPF) recommend a limit of four times the fixed salary component for stock options.[10]

[7] *See 'Low risks, high rewards'*

[8] *See 'Bosses on the run'*

[9] *See 'A racket in need of reform'*

[10] *Main et al. (1994)*

Chapter summary

After this chapter and the relevant reading, you should understand:

- the role of unemployment both as a consequence and a necessary condition in **efficiency wage theory**
- the dependence of the **efficiency wage** on the size of a potential gain from cheating, the probability of being caught and unemployment benefit
- why an individual **firm's demand for labour** could increase when it introduces efficiency wages
- why we observe **long-term employment**
- the role of **promotions** in the internal labour market.

You should be able to:

- discuss the difficulties in attempting to apply **standard neoclassical models** of labour supply and demand to internal labour markets
- analyse a simple **efficiency wage model**
- set up and solve the **minimum cost implementation problem**
- give explanations for **increasing wage profiles**
- discuss **Frank's theory of wage compression**
- discuss **managerial and executive compensation** and the role of performance related pay in this context.

Sample exercises

Review the minimum cost implementation problem for efficiency wages and use this framework to argue whether efficiency wages will be higher or lower than in the standard model if:

a. workers dislike monitoring

b. workers incur a large psychological cost when they are unemployed

c. workers are risk averse

d. monitoring is very costly

e. workers have high switching costs (moving expenses, loss of firm specific human capital) when they lose their jobs.

Determinants of market structure; measures of market structure; perfect competition; monopoly; monopolistic competition

Chapter 9

Market structure

Texts

Varian, H.R. *Intermediate Microeconomics.* (New York: W.W. Norton and Co., 2006) seventh edition [ISBN 0393927024] Chapters 22, 23 and 24.

References cited

'Ahead for now', The computer industry survey, *The Economist,* 17 September 1994, 12–16.

Dorfman, R. and P.O. Steiner. 'Optimal advertising and optimal quality', *American Economic Review* (1954) 44: 826–36.

Gourvish, T.R. and R.G. Wilson. *The British Brewing Industry, 1830–1980.* (Cambridge: Cambridge University Press, 1994 [ISBN 0521452325]

'Insurers get that sinking feeling', *The Economist,* 20 August 1994, 65–66.

Martin, S. *Advanced industrial economics.* (Oxford: Blackwell, 1993) [ISBN 063117852X].

'National lottery', *The Economist,* 28 May 1994, 28–33.

'Squeezing into Hong Kong', *The Economist,* 4 December 1993, 88,

Sutton, J. *Sunk costs and market structure,* (Cambridge, Mass.: The MIT Press, 1992) [ISBN 0262193051].

'The Post Office'. *The Economist,* 8 January 1994, 29.

'Unsure about insurance', *The Economist,* 31 October 1992, 101.

Industries differ in the way they are structured. An important defining characteristic of market structure is the number of firms serving the market. However, it is certainly not the case that we can predict how firms will behave just on the basis of how many firms there are in the industry. The presence of barriers to entry also plays a crucial role in maintaining market structure. It is generally believed that, when entry is easy, firms are disciplined in their pricing policies even when the industry is concentrated. Empirically, high entry barriers coincide with high profit margins. Economists classify industries according to the degree of market power firms have (i.e. do they behave as price takers or price makers?), whether firms behave strategically and whether the product is homogenous or differentiated. For example, in a competitive industry all firms are assumed to be price takers; in an oligopoly there is strategic interaction between firms.

Determinants of market structure

The question of why real-life industries are structured the way they are is very difficult to answer. There are a multitude of factors which contribute to the likelihood of one or the other market structure evolving. The following list is not exhaustive.

Economies of scale

The world market for disposable plastic syringes is dominated by three firms: Becton Dickinson, Sherwood/Brunswick and the Japanese firm Turumo, with worldwide market shares in 1992 of 31 per cent, 16 per cent and 18 per cent respectively. There are significant economies of scale in this industry: the production technology is such that, as output increases, average costs fall dramatically. The source of economies of scale may be large setup or fixed costs. In extreme cases the presence of significant economies of scale leads to a 'natural monopoly' which refers to the situation where average costs decrease beyond the output level corresponding to market demand. In such cases it is obviously cheaper for one firm to satisfy market demand. The natural monopoly argument has been used as an argument against privatisation of the Royal Mail in Britain. Local newspapers are also characterised by significant economies of scale so that often there will be just one local newspaper in a town.

The size of the market relative to the degree of economies of scale seems a relevant predictor of market structure. A useful concept in this context is the **minimal efficient scale** (MES) of an establishment in the industry. It is usually defined as the output level at which long-run average cost (LRAC) is minimised. An alternative definition of MES is 'the output level beyond which further increases in output would lead to a reduction in LRAC of no more than 10 per cent'. In practical terms this means it is the output level where the LRAC curve begins to flatten out. If we know the MES for a given production technology and we know the total consumption in the market, we can calculate how many firms can be accommodated in the industry. For example, if the MES for commercial aircraft manufacturing corresponds to 10 per cent of the US market, then there is room for 10 firms to serve this market. If the MES is large relative to the size of the market, the industry is likely to be concentrated.

The importance of economies of scale as a determinant of market structure is made evident by the changes in structure after technological innovation. The beer industry, for example, used to be fragmented and consisted of thousands of local breweries. Beer was unpasteurised and had to be kept cold so it could not be transported far. Since the introduction of bottling, however (a process which has significant economies of scale), the number of beer producers has decreased dramatically, mainly through takeovers, and now a handful of large companies dominate the industry.[1]

Strategy of incumbents

Market structure and the presence of entry barriers undoubtedly affect the behavior of established or incumbent firms in an industry. Conversely, incumbents may engage in activities designed to maintain the market structure or alter it in their favour. An example of the former is **limit pricing** where a monopolist, rather than maximising short-run profits, prices such that it is not possible for an entrant to make a profit at the current price. An example of the latter is **predatory pricing**, where a firm sets a low price (often below cost) in order to drive a rival out of business. Firms have instruments other than prices to their disposal when trying to drive competitors out of business. For example, after the deregulation of buses in some British cities, some bus companies tried to organise their timetables so that they would pick up passengers a few minutes before their rivals. Note that limit pricing and predatory pricing are only possible if the firm has a cost advantage. In some cases a monopolist does not have to set a low price to deter entry. The threat to retaliate and flood the market when a competitor enters may be sufficient. Such a threat can be made credible if the monopolist has excess capacity so that increasing output (after entry) is not very costly.

Incumbent firms manipulate **dealer relationships** to shut out competition. An extreme case is that of **exclusive dealership** where a manufacturer or wholesaler agrees to supply retailers only if they do not sell any substitute products. Supermarkets typically carry a fairly limited number of brands of consumer goods. Manufacturers negotiate deals which involve the supermarkets allocating them a certain amount of shelf space. Some PC makers sell their machines with a copy of MS-DOS or Windows already installed. Some airlines have developed computer reservation systems for use by travel agents. Of course flights by the airline which developed the system will be more prominently displayed and easier to book than others and, once travel agents have invested in learning how to use such a system, they are 'locked in'.

[1] *Gourvish and Wilson (1994)*

Incumbent firms often try to erect entry barriers by heavily **advertising** and promoting their products. This is effective in industries where customer loyalty is important or where buyers prefer established brands so that it becomes very difficult for an entrant to convince consumers to try his product. Large advertising budgets also increase an entrant's capital requirements.[2] Companies can increase customer loyalty by offering loyalty rebates. Airlines' frequent flyer programmes are a good example of this.

Capital requirements

Large capital requirements can be a source of a cost disadvantage for a potential entrant. Entrants are usually perceived as riskier prospects by banks and other money-lenders. This is not without reason as in many industries entrants are much more likely than established firms to go out of business. They therefore pay a higher risk premium. Large **sunk costs** (investments such as R&D which cannot be recuperated upon exit from the market) make the entry decision particularly risky. They are sources of barriers to entry in for example car manufacturing, defence industries, oil refining, deep sea drilling, chemicals, electronics, aerospace, pharmaceuticals, etc. Empirically it is found that, the larger an industry (in sales), the larger the number of firms, but this is not true for markets with high advertising and R&D costs.[3] Large fixed costs even when they are not entirely sunk make entry more difficult. Capital requirements are obviously larger if the industry is vertically integrated and it is impossible to enter unless you enter all layers of the industry. The computer industry used to be vertically integrated with computer manufacturers making most of the parts as well as the software in-house. They used to do most of their own marketing, distribution, sales and service as well. There were few independent suppliers of parts. Currently, PCs are assembled from readily available parts used to supply the consumer electronics industry. Most PC makers were never vertically integrated. Barriers to entry in the computer industry are much lower than they used to be and there is more competition in every layer of the industry.

Patents and other legal controls

Governments clearly influence the structure of some industries by setting up legal barriers to entry. Taxi drivers in many cities have to obtain a license before they are allowed to establish themselves. In New York the powerful taxi industry lobby has been able to keep the number of licenses ('medallions') unchanged at about 12,000 for almost 50 years. Governments decide which airlines are allowed to fly a certain route. By taking protectionist measures, governments can make or keep domestic industries concentrated. South Africa uses very high import tariffs and most South African industries contain a few major players. Patent law, which grants monopoly rights for a specified period of time, is used to encourage innovation.

Control of resource

An obvious instance of market structure determined by entry barriers is when incumbents control a special non-replicable resource such as mineral deposits. Some well-known examples are De Beers in diamonds, OPEC and French Champagne. Take-off and landing slots at airports are also very important resources without which entry into the airline industry is impossible. In the same industry, the practice at most airports of long-term leasing of gates represents an entry barrier. There are however limits to the market power which derives from owning a special resource. The Organisation of Petroleum Exporting Countries (OPEC) found that, when they raised the price of oil too much, other countries started searching for oil. The oil crisis of the early 1970s was largely responsible for the exploration of oil fields in the North Sea.

[2] *See next section*

[3] *Sutton (1992)*

Cost advantage and first mover advantage

In many industries the firms which entered first have a significant advantage in terms of entrepreneurial skills and/or technological know-how. Late entrants tend to have a cost disadvantage especially when **learning curve** effects are important (i.e. average production cost decreases with cumulative output). Sunk costs are also a source of cost advantage as they are only incurred on entry.

Measures of market structure

Given the complexity of real industries and the many factors which determine how firms behave in these industries we cannot expect that a simple measure of market structure paints the entire picture. Nevertheless, a summary measure of how many firms there are and their market share can convey useful information. The **concentration curve** represents a complete statistical summary of the number of firms and their market shares. On the X-axis the firms are listed starting with the largest in terms of market share. Market share can mean percentage of sales or assets or employment in the industry. On the Y-axis cumulative market share is measured so that by definition the concentration curve is concave.

Concentration ratios (CR_m) give the total market share accounted for by the largest *m* firms in the industry. UK official sources use CR_3 and CR_5 whereas, in the US, CR_4, CR_8 and CR_{12} are used. If there is only one firm in the industry the concentration ratio is 1. When there are *n* firms of equal size the concentration ratio is *m/n* which is close to 0 if *n* is large. Empirically it is found that concentration ratios are similar across industries from one country to another (e.g. consumer durables industries — electrical appliances, washing machines, refrigerators — tend to be concentrated CR_5=95%; for salt, CR_4=99.5% in UK, 93% in Germany, 98% in France). Most manufacturing industries have CR_4<50%.[4] Roughly speaking, an industry is classified as a monopoly if $CR_1 > 90\%$ and as perfectly competitive when $CR_4 < 40\%$. Monopolistic competition and oligopoly correspond to $CR_4 > 40\%$.

[4] *Sutton (1992)*

The **Herfindahl-Hirschman-Index** (HHI), like the concentration curve, uses information about all the firms in the industry. It is defined as the sum of the squares of all market shares:

$$\text{HHI} = \Sigma_{i=1}^{n} s_i^2 .$$

If there is only one firm in the industry HHI=1 and if there are many firms with equal market share, HHI is close to 0. The HHI has the interesting property that its inverse corresponds to the number of equally sized firms which would give this value of HHI. For example, if the HHI is 0.25, the industry is as concentrated (according to HHI) as an industry consisting of four firms each with market share 25% since $4(0.25)^2 = 0.25$. The advantage of the HHI is that it uses information on all firms in the industry. This is important when analysing the evolution of industries through merger activity for example. Whereas the CR does not change after a merger unless the very largest firms in the industry are involved, HHI always increases by $2s_is_j$ where s_i and s_j are the market shares of the merged firms. (Make sure you know why!) The HHI is used in the US Department of Justice Merger Guidelines to indicate when a merger is likely to be opposed. For example, a guideline might be formulated as 'a merger which increases HHI by x when HHI is currently over y will be opposed.'

As I mentioned before, these measures only give a very crude idea of what an industry is like and they have to be interpreted with care. For a start, CRs are usually measured at national level. This leads to two important distortions. Firstly, it ignores imports and this may lead to overestimation of concentration in the domestic industry. Secondly, the relevant market may be regional due to high transportation costs or other factors.

Although there may be many cement and sugar producers nationally, within a region in which these goods are normally distributed there may be only two or three firms. Similarly, there are many local newspapers in Britain but in each town only one or two are usually distributed. National CRs are therefore meaningless in this market. The definition of the product or the market is also very important. In antitrust cases the defendants often claim that the market should be defined more widely to include substitute products. If we consider the computer industry, for example, we find different concentration levels depending on whether we just consider laptop manufacturers or we include mainframes, desktops and workstations. In the pharmaceutical industry concentration may be relatively low if we consider the market for 'drugs' whereas, if we define the market as 'drugs for gout' it may be highly concentrated. Given the importance of barriers to entry for firms' behaviour in an industry, the simple measures of concentration have to be supplemented by further analysis. In particular the evolution of concentration measures may give an indication of whether firms in an industry compete fiercely. If the shares are relatively stable, this may indicate the absence of competition.

There are some measures which, unlike CRs, are not officially published but can be constructed from industry data. The **Lerner monopoly index** measures the wedge between price and marginal cost:

$$L= (p-MC)/p.$$

It is meant to give an indication of market power and entry barriers. When an industry is concentrated and has low entry barriers, the concentration ratio is high but the Lerner index would be expected to show evidence of limit pricing. Alternative measures of market power are **cross-elasticities**. Firms producing a good which has many close substitutes (large positive cross elasticity) have less market power than those producing goods with few distant substitutes.

Perfect competition

In a perfectly competitive market there are many buyers and sellers who act independently (non-strategically) and have perfect knowledge of the market and in particular of the price charged by all sellers. The product is homogenous so that firms do not engage in non-price competition. There is no point in advertising for example if consumers are perfectly informed and all products are identical. This also means that there is no price dispersion (i.e. there is a unique market price at which everyone buys and sells). No firm has control over this market price. It is often also assumed that all firms have access to the same technology and this assumption, combined with the fact that firms face the same input prices, implies that the firms will have identical LRAC curves. Inputs are assumed to be perfectly mobile and responsive to monetary incentives and there are no barriers to entry in the long-run. Entry then occurs until economic profit (of the marginal firm if they are not all identical) is zero. Zero economic profit of course allows for a normal rate of return on equity.

As a result of these assumptions a firm behaves as if the demand for its product is horizontal. Clearly the firm is in a sense making a mistake here: when it decides to increase its supply the price of output is affected because the industry demand curve is downward sloping. The same comment applies to consumers. If any consumer increases his consumption of the good he has a (very small) effect on price since industry supply is upward sloping. The idea that competitive firms and consumers ignore their effect on output price should be seen as a convenient simplification. The assumptions of the perfect competition model may seem unrealistic but there are several real world industries which conform relatively closely to the competitive framework. Markets for some agricultural products operate in a competitive fashion as do financial markets for savings and stocks. The packaged tour business is also very competitive with operators regularly going out of business. There are virtually no barriers to entry in this business. Chartering aircraft or setting up a travel agency does not require major capital investment.

From cost function to supply function

Competitive firms are assumed to maximise profits which, given their price taking behavior, means:

$$\max_{q} pq\text{-}C(q) \qquad (1)$$

where q is the firm's output level, p is the market price and C is the firm's cost function. The first order condition for (1) is:

$$p = MC(q) \qquad (2)$$

which means that the firm chooses its output level so that its marginal cost equals the market price. The firm's supply curve therefore coincides with its MC curve. We need to qualify this to ensure that the firm is not making negative profits. If the best output level according to (2) leads to losses, the firm should shut down (set its supply q=0). The shutdown conditions are as follows:

1. in the short-term, if revenue exceeds variable costs, the firm should continue to produce since revenue contributes to fixed cost
2. in the long-term, if revenue does not cover all costs, the firm should shut down.

These conditions imply that the firm's supply function is only a section of the MC curve, the section above average variable cost for short-run supply and the section above average cost for long-run supply. Short-run supply coincides with the short-run marginal cost curve (some inputs are held fixed) and long-run supply with the long-run marginal cost curve.

Case: Competition in the insurance industry

In many countries the insurance industry is fiercely competitive. The 'insurance cycle' is a well-documented phenomenon. During such a cycle the industry goes through a phase of overcapacity and price wars. This leads to losses which cause the marginal (usually smaller) firms to exit. Historically when the industry has been free of price wars for a short period of time a natural catastrophe such as Hurricane Andrew has struck eliminating the need for a price war to rid the industry of the weaker players. In any case a new equilibrium settles in with fewer firms and higher premiums and profits start to appear, encouraging new entry and buildup of excess capacity. A typical cycle lasts three to five years.[5]

From firm supply to industry supply

As a first approximation for industry supply we could sum the individual firm supply curves horizontally (i.e. for each price we determine (using the MC curve) how much each firm is willing to supply and we add these output levels to find industry supply at that price).

[5] *See 'Unsure about insurance'; 'Insurers get that sinking feeling'*

Example 9.1

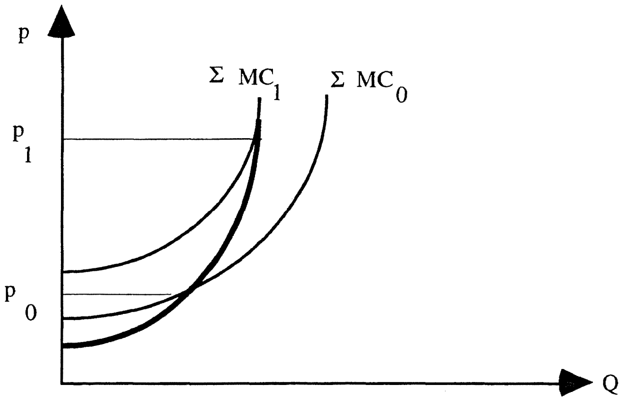
Suppose the annual market demand for wild pasta is given by D(p)=100-5p and there are 10 firms in the industry, each with production function $q=(LK)^{1/2}$ where L is water and K is wheat. In the short-run the amount of wheat is fixed at 1 unit. The input prices are r=2 for wheat and w=1 for water. Since $q=L^{1/2}$ in the short-run (K=1) each firm's requirement of water is $L=q^2$ so that the cost function is given by $C(q)=q^2+2$ (since w=1, K=1 and r=2). Each firm's supply function is thus p=MC=2q or q=p/2. We do not have to worry about shutdown here since MC always exceeds average variable cost. The industry supply curve or the sum of the individual supply curves is then Q=10(p/2)=5p and we find the equilibrium in this industry by equating demand and supply: 100-5p=5p which results in a market price of p=10 and quantity Q=50. At the market price of 10 each firm wants to produce five units and makes a profit of (10)(5)-C(5)=50-25-2=23.

As an exercise you should check how many firms could enter this industry and still make a profit.[6]

The distinction between short- and long-term is also important for industry supply. Long-run supply is more elastic than short-run supply because all inputs are variable and there is entry and exit. Sometimes the notion of intermediate run is used meaning the period of time in which existing firms can adjust most or all of their inputs whereas in the long-run market entry and exit can take place. If all firms are identical the entry and exit process ensures that each firm produces the output level at which its LRAC are minimised (i.e. at its efficient scale) and price equals LRAC. Why is this so? Clearly price cannot go below minimum LRAC or nothing would be produced. If it is above LRAC then firms can, by choosing their output level carefully, make positive profits which would induce entry. Note that, if firms are not identical, long-run supply is no longer constant but upward sloping. At the equilibrium in such an industry only the 'marginal' firms make zero profits; the other firms make positive profits.

Recall the derivation of industry demand for inputs where we concluded that the output price effect of a change in input price meant that we could not simply sum individual firm demands for the input to obtain industry demand. A very similar story can be told here. As the output price increases the demand for inputs will shift out which will cause input prices to rise (see Figure 9.1). At the changed input prices each firm's marginal cost or supply curve will shift inwards which makes industry supply less elastic than the horizontal sum of firm supply curves. This input price effect of a change in output price is likely to be significant when the industry is a large user of any of its inputs.

Figure 9.1. Input price effect on industry supply



[6] *Hint: Let the total number of firms be n and calculate the equilibrium market output and price as a function of n. Then find the maximum n for which profit is positive*

It is very difficult to assess the size or the significance of the input price effect. In some cases it may be counteracted by new entry into the industry as higher product prices lead to (temporary) larger profits. When industry output increases there may be some positive externality effects as well such as technological improvements through learning by doing for example or improved support services. These factors would tend to make industry supply **more** elastic than the horizontal sum of MC curves.

Monopoly

A monopoly is an industry consisting of one firm. Apart from state owned or heavily regulated industries such as electricity there are not many markets which conform to this extreme market structure. In practice however, we assume that a firm will behave as a monopolist when it is has a dominant market share of, say, 60-70 per cent and there is no major rival. Microsoft and Intel for example have market shares of about 80 per cent in operating software and microprocessor chips for PCs respectively.[7] Monopolies are relatively common in emerging economies and former communist countries because of protectionist measures and the lack of antitrust legislation. For example, in Mexico, Vitra has a market share of 90 per cent in flat glass. Philip Morris took over Tabak (a Czech tobacco company) which gives it control over 80 per cent of the market in the Czech Republic. A few other examples of monopolies are:

Monopolies based on patents

When firms invent a new product or process they are entitled by patent law to have temporary exclusive rights to market or license this innovation. Companies such as Polaroid and Xerox were in the fortunate position of building up a significant first mover advantage due to patenting. In the pharmaceutical industry each new drug is patent protected. When such a patent expires, competition from generic drugs cuts price dramatically. In addition to the patent laws, governments can create monopolies by granting a government franchise or 'sole rights' to operate a certain business. In 1994 the British government awarded Camelot, a consortium of several companies, the contract to run the national lottery as a monopoly for a seven year period. Other firms are prohibited from running a lottery with the exception of clubs, charities and local councils and then only if the turnover is small (less than £5 million a year). This setup mirrors what goes on in other European countries.[8] In Sweden alcohol stores are state-owned and have a monopoly on the sale of strong beer, wines and spirits. This should change now that Sweden has entered the EU as an alcohol monopoly is against EU competition law.

Utilities

Whereas in Britain, New Zealand and the US you can choose your telephone company, in many countries the telecom industry is a monopoly. Hongkong Telecom will face competition from three new competitors from 1995 but is allowed to keep its monopoly on international calls (which generate most of its profits) until 2006.[9] In Britain the Post Office keeps its monopoly on letter deliveries below £1.[10]

The standard monopoly model is very simple. Industry demand equals firm demand and, although we could make a distinction between short- and long-term with respect to the monopolist's cost function, there is no entry or exit. The profit maximisation problem can be stated in terms of price or quantity. Both formulations give the same result of course. Using the quantity formulation leads to the familiar 'marginal revenue equals marginal cost' result. If we think of the monopolist's problem as one of choosing an optimal price, we have:

$$\max_{p} pq(p)-C(q(p))$$

[7] *See 'Ahead for now'*

[8] *See 'National lottery'*

[9] *See 'Squeezing into Hong Kong'*

[10] *See 'The Post Office'*

The first order condition is:

$$p\,q'(p) + q(p) - MC(q(p))\;q'(p) = 0.$$

With some algebraic manipulation this can be rewritten in terms of the price elasticity as:

$$(p-MC)/p = 1/\eta \text{ or } p = (\eta/(\eta-1))\,MC.$$

The **markup** $\eta/(\eta-1)$ clearly depends on how price elastic demand is. For example, for the constant elasticity demand curve $q=ap^{-2}$, the markup equals 2 so that the optimal price is double the marginal cost: p=2 MC. This result also shows that, if the firm is profit-maximising, the Lerner index is the inverse of price elasticity.

The markup formula bears a strong resemblance to the popular practice of **cost-plus pricing**. Many pricing decisions are made using the rule of thumb that you should take an estimate of average variable cost, add a charge for overhead and a profit margin. If the profit margin reflects market conditions (price elasticity) and if economic costs (i.e. opportunity costs rather than accounting costs) are used, there is not too much wrong with this. The main problem is the overhead charge. For profit maximisation only marginal cost is relevant, not average cost.

The simple monopoly model is not very interesting and there are several reasons, such as regulation, why we do not observe the behavior it predicts. The really interesting questions arise when we consider how monopoly arises as a market structure and how a firm can stay a monopoly. Unless the barriers to entry are very high or the monopolist has a significant cost advantage, potential entrants are attracted by juicy monopoly profits. If the monopolist is not prepared or forced to share the market, how can he discourage entry? If the monopolist engages in limit pricing, we would not observe MR = MC. If he has a cost advantage, he could threaten to flood the market when entry occurs. If this threat is credible the monopolist may use the MR = MC rule. It is likely, however, that the threat is not credible (subgame perfect) as it may be more profitable to accommodate the entrant.[11] If the incumbent could commit to a 'fighting' strategy by announcing or advertising that he is willing to match or undercut any price offered elsewhere, entry can be deterred. Similarly, if the monopolist can obtain some large long-term contracts from major buyers he is more likely to be in a position to discourage entry. We will return to these issues later when we have looked at oligopoly. At this point these questions are hard to tackle because we have to model what happens after entry to decide whether it is worthwhile to try to deter entry for example.

Another reason why we may not observe the MR=MC rule in real-life monopolies is that some of these monopolies operate in **contestable markets**. Such markets have very low entry and exit barriers and the reason they are monopolies is usually due to economies of scale. Average cost decreases over the relevant output range so that only one firm can survive but many firms may compete to be the single supplier. Contestable markets are markets in which sunk costs are low. A frequently cited example is that of an airline route between two small cities. Entry is easy for anyone who has planes available (on other less lucrative routes) and exit is easy because planes can be sold or leased or switched to other routes. In these circumstances, a monopolist cannot exercise his market power because, if he did, he would be replaced by a willing and able entrant. In fact monopolists may even behave as if they were in a perfectly competitive industry if the market is contestable.

[11] Recall the discussion of the chainstore paradox in Chapter 2, 'Game theory'

Monopolistic competition

Monopolistic competition is the market structure which emerges if we relax the assumption of homogenous goods in the perfect competition model. There are many sellers producing a differentiated product. For example, in the beer industry, each brewery markets a unique product (although it has many close substitutes). The main consequence of relaxing the assumption of homogeneity of products is that some forms of non-price competition such as advertising and providing different types of service can be profitable. Monopolistic competition is associated with industries of firms with large advertising budgets. Many 'fast moving consumer goods' (FMCG) sectors fall into this category. Firms make enormous efforts to convince their consumers that the washing powder they market, for example, is not identical to the one marketed by their rivals. You could say that advertising is necessary in monopolistic competition to ensure that buyers and sellers have perfect information about the market which is what we continue to assume. There are, apart from advertising, many methods firms use to create differentiation. The following list contains a few examples:

- **packaging**: two chemically identical dishwasher detergents are perceived as different by consumers if one is packaged in a plastic bottle and the other in a cardboard box
- **product design**: shampoo and conditioner are sold separately and combined
- **type of service**: in the PC industry it is becoming increasingly difficult to produce a differentiated product. The same components are used in the assembly of most PCs. Almost all PCs use Intel's microprocessors. Dell however has been very successful in this industry because it invented a new way to sell PCs. Dell computers are sold by mail order and there is a telephone hot-line which offers technical advice and after-sale service
- **location**: supermarkets and restaurants, even when they offer the same range of products, are differentiated by their location. It is unrealistic to assume that, if a supermarket charges a penny more for butter than the supermarket at the other end of the high street, it will lose all its business, or conversely that, when it undercuts its rival by a small amount, it captures the whole market.

Short-run and long-run

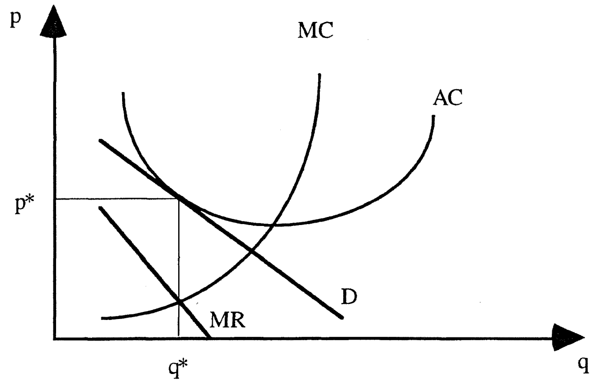
A crucial assumption in the monopolistic competition model is that, as in perfect competition, firms do not act strategically when formulating price and output decisions. They do not take potential rivals' responses into account and they do not collude. This distinguishes monopolistic competition from differentiated oligopoly.

Whereas perfectly competitive firms are forced to sell at the market price, in the monopolistic competition model, a firm does not lose all of its sales if it prices slightly above its competitors. This implies that each firm faces a downward sloping demand curve which is highly but not completely elastic. Each firm has some monopoly power. Indeed one of the reasons to engage in non-price competition such as improving service is that a firm which produces a differentiated product does not have to cut its price whenever a rival cuts his price.

The short-run monopolistic competition model is in fact exactly the same as the monopoly model with firms setting output levels such that MR = MC. As in perfect competition there are no or low barriers to entry so that in the long-run entry and exit ensure that profits are zero. The long-run equilibrium conditions for a monopolistically competitive industry are as follows:

- at the optimal output level q* for each firm, price should equal average cost (AC) so that profit is zero
- at its optimal output level each firm maximises profit and hence MR=MC at q*.

Figure 9.2: Monopolistic Competition - Long Run Equilibrium



From the long-run equilibrium conditions we know that, for the optimal output level q*, MR intersects MC. The price corresponding to q* can be read off the demand curve and has to equal AC as is illustrated in Figure 9.2 . This leaves the possibility of demand intersecting AC rather than being tangential to it at q*. However, this would mean that there are output levels for which the firm could make a positive profit contradicting the fact that q* is profit maximising. The figure shows that, at any output level other than q*, the firm makes a loss. Because demand has to be tangent to AC at the long-run equilibrium, firms operate at output levels to the left of the efficient AC minimising level. If they could increase their output, their AC would decrease. The cost of this 'excess capacity' is often referred to as the price of variety. Consumers are willing to pay a premium, so the argument goes, to have a choice between differentiated products. When, during your weekly food shopping, you have to find your way through the enormous variety of breakfast cereals for sale on the supermarket shelves you may be forgiven for being sceptical about this. In the US in 1990 alone, 48 new cold or cough remedies were introduced. It is hard to believe that consumers want this much variety.

Many industries satisfy the assumptions of the monopolistic competition model although inevitably there are entry barriers because of large advertising budgets for example. In such cases the analysis is still valid but we cannot insist on zero profits at the long-run equilibrium.

Because of product differentiation it is generally not easy to define a monopolistically competitive industry. For example, does the market for soap include liquid soap or can we restrict it to soap bars? Whereas a perfectly competitive industry consists of firms producing perfect substitutes, in monopolistic competition products are **close** but not perfect substitutes. What 'close' means in practice is difficult to say and always entails some arbitrary decisions. In empirical work a judgement has to be made when specifying demand functions as to which cross price effects to estimate and which to ignore.

Advertising and the Dorfman-Steiner model

Advertising plays a major role in monopolistically competitive industries. This gives us an opportunity to look at a model of advertising expenditures here. Even if two shampoos are chemically identical their demand curves are not perfectly elastic as in the perfect competition model if one is advertised by Naomi Campbell and the other by...let's just say someone like me. In general, advertising a product has the effect that consumers consider fewer products as good substitutes for the product. Advertising shifts the demand curve and changes price elasticity. Analytically this implies that the decisions of how much to advertise and which price to set cannot be taken in isolation.

Deriving the optimal advertising budget and price is not hard if we know how demand responds to changes in price and advertising. Let q(p,a) be the demand function. Then the firm's profit function is:

$$\pi = pq(p,a) - C(q(p,a)) - a.$$

The first order conditions are:

$$\frac{\partial \pi}{\partial p} = p\frac{\partial q(p,a)}{\partial p} + q(p,a) - \frac{\partial C}{\partial q}\frac{\partial q(p,a)}{\partial p} = 0 \qquad (3)$$

and

$$\frac{\partial \pi}{\partial a} = p\frac{\partial q(p,a)}{\partial a} - \frac{\partial C}{\partial q}\frac{\partial q(p,a)}{\partial a} - 1 = 0. \qquad (4)$$

After some algebraic manipulation, we can rewrite these conditions in terms of the price and advertising elasticity of demand:

$$\eta = p / (p - MC) \qquad (5)$$

and

$$\alpha \equiv \frac{\partial q(p,a)}{\partial a}\frac{a}{q} = \frac{a}{q(p-MC)}. \qquad (6)$$

Equation (5) is the familiar markup formula for monopoly pricing. From equation (6) we deduce that advertising increases with the responsiveness of demand to advertising, the output level and the profit margin. At the optimum, the ratio of advertising expenditures to sales revenue equals the ratio of advertising and price elasticities (divide (6) by (5)):

$$a / (pq) = \alpha / \eta \qquad (7)$$

This result was first shown by Dorfman and Steiner (1954) and has generated a huge literature in the field of marketing models. In empirical studies it is found that the advertising-sales ratio is positively related to price-cost margins[12] which confirms the Dorfman-Steiner result (substitute (5) into (7)) since price elasticity is negatively correlated with the price cost margin. Of course this simple model does not capture all of the interesting aspects of the advertising decision. For example, it ignores dynamic effects of advertising. A company's current advertising budget is unlikely to just affect current sales. By advertising now a stock of goodwill is built up which leads to higher sales in the future as well.

[12] *Martin (1993) 159*

Chapter summary

After this chapter and the relevant reading, you should understand:

- the importance of **entry barriers** as determinants of market structure and firm behaviour
- the concepts **concentration curve, CR_m, HHI and Lerner index**
- why in a perfectly competitive industry the **supply curve** is not necessarily the horizontal sum of individual firm supply curves
- why we may not observe MR=MC in a **monopoly**
- the **Dorfman-Steiner model.**

You should be able to:

- describe the important factors in the determination of **market structure** giving (preferably your own) examples
- calculate the **market structure measures**
- derive perfectly competitive firm and industry **supply**
- find profit maximising price and output for a **monopoly**
- explain short run and long run equilibrium in a **monopolistically competitive industry.**

Sample exercises

1. The market demand for a product is as indicated in the table below. The industry is perfectly competitive and the second table gives each firm's long-run total cost. Each firm can produce only integer numbers of units of output. How many firms will be in the industry in the long-run?

price (£)	quantity demanded
30	200
20	300
10	400
5	600
3	800
output	**total cost (£)**
1	10
2	12
3	15
4	30

2. The demand for a textbook is given by p=20-0.0002q. The publisher's marginal cost function is MC=6+0.00168q. The author of the textbook gets a royalty of 20% of total revenue. Suppose the publisher decides on the price. What is his preferred price-quantity pair? Suppose the author could decide on the price. What is her preferred price-quantity pair? In the first scenario (the publisher setting the price), how much would the publisher offer the author as a lump sum payment in return for the author giving up the royalty. Should the author accept?

3. Mickey Mouse Publishing Ltd. is a price-taking firm in the market for microeconomics textbooks. It produces books with total cost function $C(q) = 3q^2+6q+3$. What is its short-run supply function? What is its long-run supply function?

4. The Seow Food Corporation has bought exclusive rights to sell chocolate bars at Wembley arena. The fee it paid for the concession was £1000 per event. The cost (excluding this fee) of obtaining and marketing each candy bar is 10 pence. The demand schedule for candy bars in the arena is as indicated in the table below. Prices must be in multiples of five pence. What price should Seow charge for a candy bar and what is the maximum amount that it should pay for a concession for a single event?

Price per candy bar (pence)	Thousands of candy bars sold per game
20	10
25	9
30	8
35	7
40	6
45	5
50	4

5. Demand for a monopolist's goods is given by $q=S/p$ if $p \leq p_o$ and $q=0$ if $p>p_o$. S is a constant. What is the monopoly price?

Notes

Price discrimination; commodity bundling; multiproduct firms; transfer pricing

Chapter 10

Monopolistic pricing practices

Texts

Tirole, J. *The Theory of Industrial Organization.* (Cambridge: Mass.: The MIT Press, 1988) [ISBN 0262200716] Sections 1.1.2, 1.5.2, Chapter 3.

Varian, H.R. *Intermediate Microeconomics.* (New York: W.W. Norton and Co., 2006) seventh edition [ISBN 0393927024] Chapter 25.

References cited

Adams, W. and J. Yellen ‘Commodity bundling and the burden of monopoly’, *Quarterly Journal of Economics* (1976) 90: 475–98.

‘Ad nauseam’, *The Economist,* 5 February 1994, 71.

‘A hell of an operating system’, *The Economist,* 18 January 1995, 17–18.

‘Better than price-fixing?’, *The Economist,* 3 October 1992, 100,

‘By the seat of their pants’, Airlines Survey, *The Economist,* 12 June 1993, 26–28.

‘Europe’s car market. Carved up’, *The Economist,* 31 October 1992, 94.

‘Death of the brand manager’, *The Economist, 9* April 1994, 79–80.

Fisher, K. ‘Return to sender’, *The Economist,* 4 February 1995, 8.

“Hard sell’, *The Economist,* 4 March 1995, 89–92.

Hirshleifer, J. ‘On the economics of transfer pricing’, *Journal of Business* (1956) 29: 172–84.

‘IOU all over again?’, *The Economist, 2* July 1994, 40–41.

‘Labouring in obscurity’, *The Economist,* 17 September 1994, 94.

Lan, L. and A. Kanafani ‘Economics of park-and-shop discounts: a case of bundled pricing strategy’, *Journal of Transport Economics and Policy,* 3/27, 291–303.

‘Learning to fly all over again’, Airlines Survey, *The Economist,* 12 June 1993, 13.

‘Life after Lenin’, *The Economist,* 29 January 1994, 74.

‘Managing the future’, *The Economist,* 19 December 1992, 70–75.

‘Movie mystery’. *The Economist,* 19 November 1994, 36,

‘Not dead yet’, *The Economist,* 28 January 1995, 76–80.

Oi, W. ‘A Disneyland dilemma: two-part tariffs for a Mickey Mouse monopoly’, *Quarterly Journal of Economics* (1971) 85:77–90.

Schmalensee, R. ‘Gaussian demand and commodity bundling’, *Journal of Business* (1984) 57:1, 2, S21 1–S230.

‘Tax deficient’, *The Economist, 22* May 1993, 20.

Stigler, G.J. ‘*United States v. Loew’s Inc:* A note on block booking’, *Supreme Court Review* (1963) 152–57. Reprinted in Stigler, G.J. *The organization of industry.* (Homewood, I11: Irwin, 1968).

‘Taxing questions’, *The Economist,* 22 May 1993, 83.

‘The high-tech war’, *The Economist,* 26 December–8 January 1993, 83–84.

‘The richest islands in the world, maybe’, *The Economist.* 6 November 1993, 80.

‘Unhappy returns’, *The Economist,* 2 April 1994, 88.

‘Unhappy returns for Barclays’, *The Economist,* 25 June 1994, 97.

van Ackere, A. and Reyniers, D.J. 'A rationale for trade-ins', *Journal of Economics and Business* (1993) 45:1-16.

van Ackere, A. and Reyniers, D.J. 'Trade-ins and introductory offers in a monopoly', *RAND Journal of Economics* (1995) 26:58-74.

Webster, T. 'Rooms with a view to saving', *Evening Standard* 30 November 1993, 38.

'What went wrong at IBM', *The Economist*, 16th January 1993, 23-25.

In our discussion of monopoly pricing (Chapter 9) we found that the monopolist sets price such that marginal revenue equals marginal cost. We implicitly assumed that the monopolist had no option but to set a single price, the same for all buyers, independent of the quantity they buy, whether they buy any other products from the monopolist, etc. In reality, sellers often have the opportunity to market their products at different prices (price discrimination) or sell them as part of a package (commodity bundling). So far, we have also assumed that the firm markets one product or that we could analyse production and marketing of individual goods in isolation. However, most firms produce a range of products and they find it profitable to keep the pricing of the entire product line in mind when making pricing decisions for each product.

Many firms are structured as multidivisional organizations with one division buying goods from another division within the firm. The transfer price at which such goods are sold between divisions is of crucial importance for the profitability of the firm. If selling divisions are rewarded on the basis of their divisional profit, they will tend to set the transfer price too high which is detrimental to overall profitability. Clearly, transfer prices have to be determined centrally in such situations. In this chapter we turn our attention to these types of pricing problems.

Price discrimination

Of all pricing practices, monopolistic or otherwise, price discrimination in its many forms is certainly the most visible. Price discrimination occurs whenever the same good or service (i.e. a good or a service produced at identical cost) is sold at different prices. More generally, we can say that, when price differences between consumers are larger than is warranted by cost differences, there is price discrimination. For example, British Gas offers a reduction of your gas bill if you pay by direct debit. To the extent that collecting revenue by direct debit is less costly than by alternative methods, the discount is not an indication of price discrimination. It is more difficult to argue that price differences between first and second class on trains, or business and economy on planes, are justifiable on the basis of cost differences. There is something else going on here.

For a firm in a perfectly competitive industry it is not possible to indulge in price discrimination. If such a firm sets a price above the market price, it loses all of its customers. Some degree of market power is necessary to enable a firm to price discriminate. It is easiest to analyse price discrimination by a monopolist although in reality it occurs mostly in oligopolies. (Models of price discrimination by oligopolists are the subject of recent and current research in industrial organisation.) It is easy to understand the attraction of charging different prices for the same good. Clearly, profits cannot decrease as a consequence of price discrimination since one option is to set the same price in all markets. In general, profits increase if price discrimination is possible.

Why is price discrimination more common in the sale of services than manufactured goods? For services, resale is generally not possible. I cannot get a haircut or a massage and sell it to you, for example. Utilities such as telephone and electricity companies often successfully use price discrimination. Customers cannot resell their services.

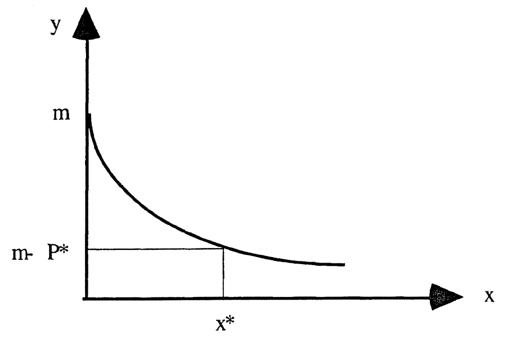
When resale is possible, price discrimination may not succeed because of arbitrage. The term arbitrage is used when it is possible for low-price buyers to sell in a high price market. If a wholesaler offers quantity discounts to retailers, he has to ensure that it is impossible for a retailer to buy a very large quantity taking advantage of the discount and then resell the product to the wholesaler's other customers. Similarly, when a Japanese car manufacturer wants to charge different prices in Europe and in the US, the potential price differential is limited by transportation costs. If the price difference is too large (e.g. when Japanese cars in the US are much cheaper than in Europe), they will be shipped from the US to Europe (in the absence of quotas etc.)

Economic theory classifies price discrimination practices according to how much information the monopolist has, whether consumers can choose from a **price menu** (e.g. when quantity discounts are offered) and whether markets are isolated (no arbitrage possibility) so that consumers in each market or market segment are quoted a segment specific price.

First degree price discrimination

Suppose a consumer with income *m* derives utility from widgets *x* and income remaining *y* after paying for the widgets. A seller of widgets who knows the utility function can offer the consumer a take-it-or-leave-it deal. She can determine the maximum amount the consumer is willing to pay by comparing the utility of consuming widgets and paying a certain amount *P*, to the utility without widgets. You can think of the seller choosing a point on the consumer's indifference curve through *(0,m)* as is illustrated in figure 10.1. If the seller asks for a payment P^* in return for x^* widgets, the consumer will just about accept this offer. If the seller asks for more than P^* or offers less than x^* in return, the consumer rejects the offer.

Figure 10.1: First degree price discrimination



Example 10.1

A consumer with income *m* derives utility from a good *x* produced by a monopolist and from his remaining income *y* according to utility function $U(x,y) = x^{1/2}y + y$. Hence his utility is $U(x,m-P) = x^{1/2}(m-P) + (m-P)$ when he buys *x* units and pays the monopolist *P* and utility $U(0,m) = m$ if he goes without product *x*. The monopolist who knows the consumer's utility function is able to make a take-it-or-leave-it offer, that is, he says 'I will give you *x* units if you give me *£P;* if you refuse, you will get nothing'. The consumer accepts such an offer if $U(x,m\text{-}P) > U(0,m)$:

$$x^{1/2}(m-P)+(m-P)\geq m \text{ or } x\geq\left(\frac{P}{m-P}\right)^2. \qquad (1)$$

Assume the monopolist has constant marginal costs c. To determine his offer P optimally he maximises his profit P – cx subject to (1). Clearly he is not going to give the consumer more than is necessary to make him accept the offer. This implies that he sets x = (P/(m–P))^2. His problem then becomes one of finding the optimal payment P:

$$Max \quad \pi = P - cx = P - c\left(\frac{P}{m-P}\right)^2 . \qquad (2)$$

The first order condition corresponding to (2) is:

$$1 - 2c\left(\frac{P}{m-P}\right)\left(\frac{m}{(m-P)^2}\right) = 0$$

which can be rewritten as:

$$\frac{1}{2c} = \frac{Pm}{(m-P)^3} . \qquad (3)$$

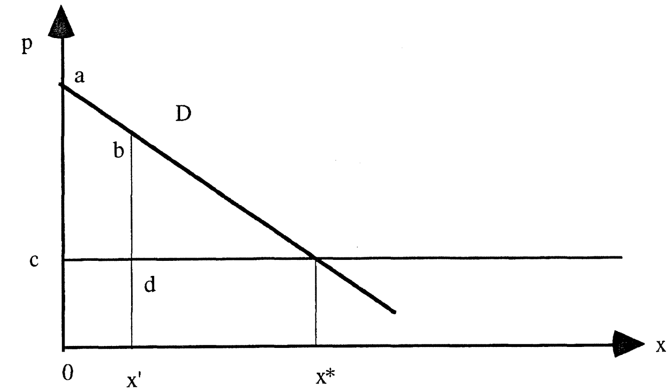
You should check that the second order condition is satisfied for P<m. The optimality condition (3) does not have an analytical solution but given any values of m and c we can solve for P. If marginal cost is c=4/3 and income m=3, then the monopolist should set P=1 and x=(P/(m-P))^2=1/4 , i.e. he should make an offer of 1/4 unit at a payment of 1. We can check in (1) that the consumer will just accept this offer: (1/2)(3–1) + (3–1) = 3. The monopolist, who is in the most enviable position of knowing the consumer's utility function, gets profit π =1–(4/3)(1/4) = 2/3. You should check that, when the monopolist is restricted to setting a fixed unit price, his optimal profit would be less than 2/3.[1]

There is an alternative way to analyse the problem of first degree price discrimination. It starts out with a consumer's demand function rather than the utility function and uses the idea that the area under the demand curve (consumer surplus) represents 'willingness to pay' or a monetary equivalent of the utility a consumer derives from consuming the good. However, this interpretation is only valid for special classes of utility functions such as **quasi-linear utility.** A utility function of type *U(x,y) = v(x)+y,* where *x* is the quantity of widgets produced by the monopolist and *y* is remaining income, is quasi-linear. For such a utility function, the indifference curves in the *x-y* plane are vertically parallel and hence the income elasticity of widgets is zero. When we **can** use consumer surplus as a measure of how much consumption of a good is worth to a consumer, the monopolist's problem can be illustrated as in Figure 10.2. For each choice of *x,* the monopolist knows how much the consumer is willing to pay e.g. for x= x', the seller can demand the consumer surplus *ab* x' *0.* His costs are *cd* x' *0* so that his profits are *abdc.* Profits are maximised when the monopolist offers to sell *x** units and demands payment of the entire consumer surplus. In other words, the monopolist is selling each unit at the highest possible price the consumer is willing to pay. Although we have concentrated on the idea of the monopolist offering take-it-or-leave-it deals to individual consumers, we can think of *D* as the market demand curve. The same conclusion holds in that the monopolist will sell the quantity which maximises consumer surplus minus costs and demand payment of the entire consumer surplus which is the sum of individuals' consumer surplus.

Under first degree price discrimination, assuming constant marginal cost, the seller deals with each consumer and determines individual take-it-or-leave-it-offers. The offer *(P,x)* varies from person to person. Of course this is all quite unrealistic. In practice you might come reasonably close to the situation described above when there are few buyers and the seller has individual contact with each of them. For example, a car dealer may have some idea of how much a potential customer is willing to pay. In any case, the perfect price discrimination model provides a useful benchmark.

[1] ***Hint****: determine the consumer's demand function and set MR=MC*

Figure 10.2: Extraction of consumer surplus



Second degree price discrimination

Second degree price discrimination is also referred to as **nonlinear pricing** or **block pricing**. Under this scheme, the seller offers consumers a menu consisting of prices corresponding to different quantities. Usually price is lower if a higher quantity is bought (quantity or bulk discount). Foods and drinks are often sold at lower per unit prices in 'family' packs. This is price discrimination unless the price difference with single item sales can be justified by differences in distribution or packaging costs. Utilities (gas, electricity, water, telephone) offer price menus on which price decreases with the number of units bought. In theory a seller could offer a price menu with prices increasing when larger quantities are bought. In practice consumers would buy small quantities several times rather than a big quantity unless of course the seller can keep track of all this by making sure that sales are not anonymous. Indeed, for bank loans where price (interest) increases with quantity, buyers (borrowers) are not anonymous.

In contrast to first degree price discrimination, under second degree price discrimination every individual who buys the same quantity pays the same price. In a sense consumers **self-select** to pay different average prices through their choice of consumption quantity.

A frequently used special case of nonlinear pricing is **two-part pricing.** Consumers pay a lump sum 'entry fee' which gives them the right to purchase the good plus a regular price per unit. Of course, this type of pricing is only possible if customers cannot resell (to customers who did not pay the fixed fee). One of the first papers[2] on two-part pricing quotes Disneyland amusement park which used to charge per ride in addition to the entrance fee.

The following can all be interpreted as two-part pricing:

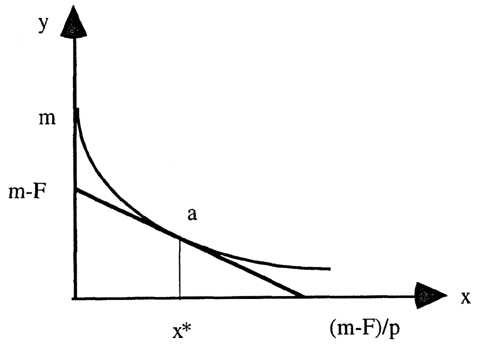
- telephone, electricity and gas companies charge a fixed rental fee plus a price per unit
- nightclubs use entry fees in addition to a fixed price per drink
- for products which need complementary inputs or replacement parts such as cameras and shavers, the price of the product is like an entrance fee and in addition the consumer pays a fixed price per roll of film or per blade
- taxis often charge a fixed fee plus a price per distance travelled.

In some cases, there is a cost based justification for two-part pricing, for example, for telephones there is a connection cost and printing jobs involve setup cost so that larger batch sizes are less costly per unit.

[2] *Oi (1971)*

As an illustration of two-part pricing in the same framework as first degree price discrimination, consider the problem of the monopolist who knows an individual's utility function. He offers the individual a combination of a fixed fee *F* and a unit price *p*. Figure 10.3 shows that this problem is equivalent to the one we encountered in first degree price discrimination. The seller can in fact choose any point on the indifference curve through *(0,m)* by an appropriate choice of *F* and *p*. The entry fee *F* determines the intercept of the consumer's budget line whereas *p* determines its slope.

Figure 10.3: Two-part pricing



If we analyse the problem in terms of the demand function (which we are allowed to do only if the income effect is negligible), then for any price *p* the seller charges, the maximum fee is the consumer surplus at price *p*. Again the best the monopolist can do leads to the same result as in the take-it-or-leave-it scenario, where the entire consumer surplus was captured. Here, price is set equal to marginal cost and *F* is the consumer surplus corresponding to *p=c*. This result is not dependent on the assumption of constant marginal costs. The seller's optimisation problem is:

$$\max\, p(q)q - C(q) + CS(q) = \max\, p(q)q - C(q) + \int_0^q p(t)dt - p(q)q$$

where *CS(q)* is the consumer surplus of consuming *q* units. The solution to this maximisation problem is to set *p* equal to marginal cost.

The two-part pricing problem becomes more realistic and more interesting but also much more complicated when the seller does not know individual utility functions or demand curves. Suppose the seller knows there are two types of consumers out there with demand functions $q_1(p)$ and $q_2(p)$ respectively but he cannot distinguish between them. He also knows the fraction of consumers belonging to each of these groups. The seller has to offer the same two-part pricing offer *(F, p)* to all consumers. How is he going to determine the best *(F, p)* offer? A first consideration is whether the seller should aim to sell to both groups. It may be that one of the groups has very low willingness to pay for the product and is not worth trying to sell to. If we assume that both groups are served, then we could determine the optimal *F* for any price *p*. It will be the minimum of the consumer surplus at *p* for a Type 1 and a Type 2 consumer. (Remember that we want both groups to buy.) This implies that the seller does not succeed in extracting all consumer surplus. One group will enjoy positive consumer surplus whereas the other's will be zero. Average profit per consumer consists of this fee *F* plus the profit margin times the average demand (at price *p*), which is dependent on the relative size of the groups. Although all consumers will be offered the same deal, the two groups end up paying different average prices *(F+pq)/q = F/q+p* because their quantity choice differs.

Third degree price discrimination

Third degree price discrimination is 'group' price discrimination. Different people pay different prices according to which group they belong. Typically they cannot change their membership status. For example, my hairdresser gives a 50 per cent discount to elderly ladies (No, I do not qualify!). It is important that it is impossible or unprofitable for low-price buyers to resell to potentially high-price buyers. This is why third degree price discrimination is most frequently used when selling services. Usually in third degree price discrimination schemes you pay the same price for each unit you buy (i.e. there are no quantity discounts). (In theory you could combine second and third degree price discrimination.) The following list gives a few examples of third degree price discrimination (I am sure you can think of others):

- senior citizen discounts for movies and buses; student discounts for the theatre; children's discounts for the zoo
- academic journals which can be three times more expensive for libraries than for individuals
- seasonal pricing at resorts
- some restaurants which offer the same menu for lunch and dinner but dinner prices could be double the lunch prices
- afternoon and evening performances at the theatre may differ in price
- trade-in discounts are offered for many appliances
- fare classes on trains (first and second class) and planes (first, business and economy class)
- companies with large travel budgets are given corporate discounts by airlines and hotel chains.

The analysis of third degree price discrimination is similar to that of division of output between plants. Indeed, you could analyse the (realistic) scenario of a company manufacturing a product in several plants and selling in several markets in this same framework. For simplicity we will focus on price discrimination here and assume that everything is produced in one plant with cost function $C(q)$. The seller can charge different prices, p_1 and p_2, in his two markets with demand functions $p_1(q_1)$ and $p_2(q_2)$. The seller's profit maximisation problem is:

$$\max_{q_1, q_2} \Pi = p_1(q_1)\, q_1 + p_2(q_2)\, q_2 - C(q_1 + q_2)$$

with first order conditions:

$$\begin{aligned} MR_1(q_1) &= MC(q_1 + q_2) \\ MR_2(q_2) &= MC(q_1 + q_2). \qquad (4) \end{aligned}$$

The intuition behind this result is straightforward: you should always allocate a unit of output to the market where it will have the highest revenue. **If** you supply both markets, their marginal revenue should be equal. Furthermore you should only sell another unit if its MR exceeds its MC. You could decide not to sell at all in one market. This means that the MR in the market which is not supplied is always below the MR in the market which is supplied and is below MC **at the optimal output level**. It is very important to realise that the marginal cost should be measured at the total output $q_1 + q_2$. When MC is constant it obviously does not matter where it is measured but in the general case of increasing MC it does. Figure 10.4, which illustrates the solution we have just discussed, is similar to the 'division of output' figure. The intersection of the horizontal sum of the marginal revenue curves with the marginal cost curve determines optimal total output. We can then trace back to the individual MR curves to determine how much is sold in each market. The optimal prices are read off the demand curves. If marginal costs are sufficiently high (MC'), only the 'high willingness to pay market' (Market 1) will be supplied.

Figure 10.4 Third degree price discrimination

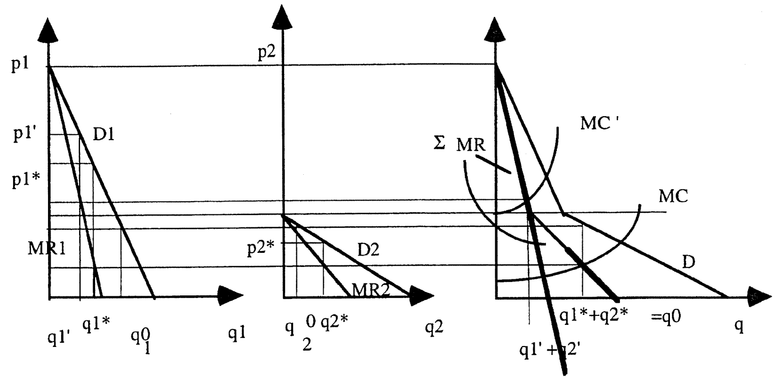


Figure 10.4 also illustrates the monopolist's pricing problem when he is prevented from price discriminating between the two markets. In this case, his demand curve is the horizontal summation of the market demands $D = D_1 + D_2$ and optimal output $q^0 = q^0_1 + q^0_2$ is determined where MC intersects the MR corresponding to D. Since D is kinked, the MR (the bold lines) is not continuous. In the figure, because the demand functions are linear and it continues to be optimal for the monopolist to serve both markets, the optimal output is unchanged. Of course the prices in the two markets are different: the low price is increased and the high price is reduced. In general, when price discrimination is not allowed, the new common price should be between the two previous prices so that consumers in the previously high price market benefit at the expense of consumers in the previously low price market. For example, Distillers had a dual pricing system in the seventies. Whisky exported to the continent was more expensive than that sold in the domestic market. When this was ruled to be illegal by the EU Commission, prices in the UK rose considerably for brands such as Red Label.

When price discrimination is not allowed, a monopolist is more likely to not to serve the low willingness to pay market.

Can you draw a MC curve on Figure 10.4 for which the monopolist only sells in both markets when he is allowed to charge different prices?

Generally output is lower when price discrimination is not allowed. If the monopolist is constrained to charge the same price in all markets, he may not want to expand production since it would lower the price of all output but if he can price discriminate he can expand sales in one market without hurting revenue in another. This is why the welfare effects of price discrimination are ambiguous. Generally, some consumers gain and others lose; the seller of course always gains.

Example 10.2

The domestic and foreign annual demands for a textbook entitled *Pricing for monopolists* are $p_d = 100 - 15q_d$ and $p_f = 60 - 2.5q_f$ respectively where p is the price in dollars and q is the number of copies sold (in thousands). Any copy of the textbook can only be sold in the country to which the publisher has consigned it (i.e. re-export is illegal). The publisher's annual production costs for this textbook are:

$$C(q) = 10.8 + 20\,q + 0.1\,q^2.$$

The MR curves corresponding to the two demand functions are:

$$MR_d = 100\text{-}30q_d$$

and

$$MR_f = 60\text{-}5q_f.$$

The optimality condition (4) implies that these two expressions are set equal to:

$$MC(q_d + q_f) = 20 + 0.2\ (q_d + q_f).$$

Solving for q_d and q_f results in:

$$q_d = 2.6 \text{ and } q_f = 7.6$$

which suggests (use the demand functions) prices:

$$p_d = 61 \text{ and } p_f = 41.$$

At output $q_d + q_f = 10.2$, MC equals 22 which is also the marginal revenue in both markets for these sales levels. Profit equals:

$$\Pi = (2.6)(61) + (7.6)(41) - 225.2 = 245 \text{ or \$245,000 per year.}$$

Now suppose the publisher is accused of **dumping** (i.e. selling at a lower price abroad than domestically) and is forced to sell the textbook at the same price domestically and abroad. Assuming the publisher still finds it profitable to sell in both markets, he faces a total demand:

$$q = q_d + q_f = (100\text{-}p)/15\ +\ (60\text{-}p)/2.5.$$

The publisher's profit maximisation problem can be expressed with price as the decision variable:

$$\max_{p}\ \Pi = pq(p) - C(q(p)).$$

The optimality condition is:

$$\left(p - \frac{\partial C}{\partial q}\right)\frac{\partial q}{\partial p} = -q. \qquad (5)$$

If we substitute the total demand q in (5), and solve for p we find $p = 43.83$. At this price 3,740 copies $(q_d = 3.74)$ are sold on the domestic market and 6,470 $(q_f = 6.47)$ are sold on the foreign market. The no dumping restriction hurts the publisher's profit which is now

$$\Pi = (43.83)(10.2) - (10.8 + 20(10.2) + 0.1(10.2)^2) = 221.9$$

so that profit has fallen to \$221,900 annually.

Case: The European car market

Although price discrimination is illegal within the European Union, it is a fact that, on 13 January 1993, the Opel Corsa pre-tax price was 9,276 ecus ($12,000) in Britain but only 6,137 ecus in Belgium; a Mercedes 200 cost 30,351 ecus in Ireland but 21,545 ecus in Germany; and a Peugeot 205 cost 9,339 ecus in Germany but 7,205 ecus in Portugal. These large price differentials are sustained by the bureaucratic difficulties involved in importing a car in an EU country. Documents have to be filled out, special registration procedures followed and rules on technical standards protect domestic manufacturers. The Greek government does not allow its citizens to buy foreign currency for the purpose of importing a car. Dealers may not honour manufacturer guarantees when cars bought overseas break down. Unlike the situation in the US, dealers in Europe only sell one manufacturer's cars. EU competition rules forbid exclusive dealerships but, due to pressure from the car manufacturers lobby, a 1985 regulation exempted the car industry from these rules for 10 years. This means that car makers have influence over dealers to keep prices up and to discourage cross-border sales (e.g. Renault ordered its dealers in Belgium to charge much more for right-hand-drive models destined for British roads).[3]

[3] *See 'Europe's car market. Carved up'*

Rewriting the optimality conditions (4) in terms of elasticity of demand allows us to explain some features of third degree price discrimination in practice. At the optimum, we find:

$$p_1 = \left(\frac{\eta_1}{\eta_1 - 1}\right) MC(q_1 + q_2) \text{ and}$$

$$p_2 = \left(\frac{\eta_2}{\eta_2 - 1}\right) MC(q_1 + q_2).$$

Since $\frac{\eta_1}{\eta_1 - 1} > \frac{\eta_2}{\eta_2 - 1}$ when $\eta_1 < \eta_2$, higher price will be set in the market with the least elastic demand. Every day, you can see applications of this principle. People are charged more when they want to have things done in a hurry (their demands are likely to be inelastic): for example, one-hour versus 24-hour photo processing. This type of price discrimination based on urgency of service also applies to same day versus same week dry-cleaning etc.

Japanese car manufacturers charge more for their cars in Japan than abroad because the overseas car market is more competitive (and hence its price elasticity higher) than the protected Japanese market. Dumping is illegal for goods imported in the US. To offset the price differential, extra duties are imposed on importers found dumping. The situation in the European Union is similar. However, in an interesting case of double standards, the EU persistently dumps surplus agricultural commodities, arising from farm support programs, in poor countries.

Because the income effect of a price change is likely to be larger for poor customers, their demands are generally more elastic. In many instances we do observe that poorer customers are charged lower prices. Now you know that this may have nothing to do with charity as it can be explained by pure profit maximising behavior on the part of the seller. It is rational for doctors and dentists to charge poor people a lower fee. Often geographical proxies are used in the sense that everyone in a generally poorer area benefits from lower prices and conversely everyone in a relatively richer area is asked to pay more. For example, Warner Brothers' cinema in Leicester Square in London charges £7 for a ticket about double of what is charged (£3.80) in its cinema in Bury near Manchester.[4] Apparently, Russian newspapers have caught on to this trick. In January 1994, Izvestia charged 20m roubles ($13,300) for a full page advertisement if the customer was a Russian company and $30,000 if it was a foreign company.[5] In China, outdoor advertising rates for foreign products are up to five times those for domestic products.[6] Keith Fisher (1995), the director of Royal Mail International, points out in a letter to *The Economist* that the Royal Mail charges a developing country one penny to deliver an incoming letter whereas an advanced country pays 16 pence. Measures have to be taken to avoid abuse by 'remailers' who essentially arbitrage by shipping bulk mailings to developing countries.

Marketing managers use various devices to discriminate between segments of the market. Although grocery **coupons** have other uses such as gathering information on demand by experimenting with the price, it seems that sellers may be trying to discriminate between people who collect and redeem coupons and those who do not bother. The assumption is that the redeemers have more elastic demand. Coupons are used for many FMCG such as breakfast cereals. The importance of this type of promotion is illustrated by the fact that grocery-coupon redemption amounted to $.6 billion in the US in 1992.[7] Similar reasoning applies to **price matching** which refers to the practice of offering to match any lower price the consumer can find. Usually an advertisement of the lower price has to be presented. Again there are other explanations for this marketing practice but, since very few customers collect on price guarantees, price matching is an easy way to discriminate between customers with low search costs (for lower prices) and those with high search costs. It can be argued that the latter have more inelastic demands.

[4] *See 'Movie mystery'*

[5] *See 'Life after Lenin'*

[6] *See 'Not dead yet'*

[7] *See 'Death of the brand manager'*

Skimming or intertemporal price discrimination

A monopolist wants to market a new durable good such as a CD player, the type of product that, within a horizon comparable to its lifecycle, a consumer buys at most once. A frequently observed pattern for these types of goods is that they are initially sold at extremely high prices. After a few years these innovative goods become more affordable. This pricing pattern of decreasing prices over time is referred to as intertemporal price discrimination. It is a form of third degree price discrimination in the sense that consumers who buy later pay a lower price than those who buy early on. At the same time you could argue that rational consumers can predict that prices will decrease and can therefore 'self-select' when they will buy, which makes intertemporal price discrimination more like second degree price discrimination.

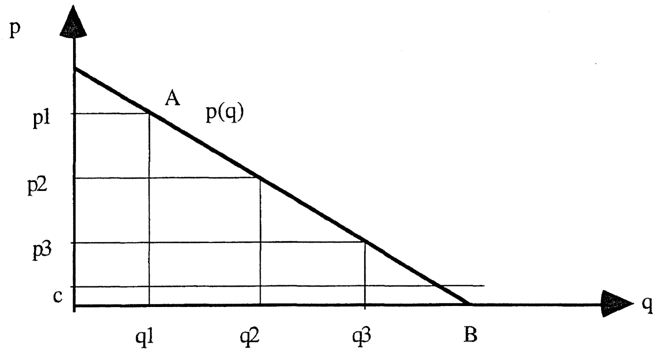
Price decreases over time are the norm in publishing. Publishers are in fact monopolists since each book is sold by only one publisher. Although there are some cost differences between hardbacks and paperbacks, it is difficult to justify the price differences on a cost basis. Furthermore, paperbacks are always sold when the demand for the hardback version has virtually dried up. Between the hardback and the paperback stage, books are sold through bookclubs as well. The Softback Preview bookclub is an English company which specialises in high-quality paperback versions of original hardback editions at around half the publisher's hardback price. An example of market skimming, where price cuts are not disguised by quality differences is provided by Intel. Until Intel lost its monopoly, (when AMD (Advanced Micro Devices) entered the market), its pricing strategy was to keep new microprocessor prices very high and ease them down slowly over time.

Sometimes intertemporal and third degree price discrimination are used simultaneously. Tickets for the Kasparov-Short chess championship in London, sponsored by the Times, were initially offered at up to £150. When they did not prove popular and very few seats were sold, prices were reduced down to £20 (**intertemporal price discrimination**) and readers of the *Sun* (a London daily newspaper) were offered tickets for £10 (**third degree price discrimination**). The frequently used marketing practice of offering a 'trade-in discount' on goods such as vacuum cleaners and other domestic appliances can be explained as an attempt to price discriminate both intertemporally and between customers who have bought before and those who have not.[8]

If we assume that consumers are **myopic**, which means that they do not think ahead and hence do not anticipate price changes, price skimming is easy to model. The population of buyers and their willingness to pay can be represented by a demand curve $p(q)$, as in Figure 10.5. If the monopolist can costlessly change the price he ends up with the entire consumer surplus (above marginal cost c). He 'skims' the market by selling at a high price to the most eager (or rich) consumers with very high willingness to pay first and then gradually moves down the demand curve selling to consumers with lower and lower reservation prices. In reality of course durable goods prices do not change every hour. The monopolist may have to keep the price constant for a given period of time, for example, because the price has appeared in a printed advertisement or a catalogue. In that case we would expect the price to jump from period to period. As in Figure 10.5, during the first year, the price may bep_1 and q_1 units are sold. At the beginning of the second year, since all consumers with reservation price above p_1 have left the market, the monopolist faces a new **residual** demand curve representing consumers still in the market (line AB in Figure 10.5). He then picks a new (lower) price p_2 and sells $q_2 - q_1$ units at this price and so on. Clearly, a large fraction of the consumer surplus can be captured in this way.

[8] *van Ackere and Reyniers (1993, 1995)*

Figure 10.5: Market Skimming



Example 10.3

A monopolist seller of microprocessors has a horizon of two years. Any microprocessor not sold after two years is considered obsolete and is scrapped. The demand for this durable product is given by $p(q) = 1-q$. Consumers are myopic and buy a microprocessor at most once. The production costs are negligible. Assuming the seller can set different prices in year 1 and in year 2, what is his optimal pricing strategy?

As we know from Figure 10.5, the price set in the first period determines second period demand. The easiest way to deal with this analytically is to work out what the seller should do in period 2 when he has set a price p_1 in year 1. When we have worked out this problem, we can solve the problem of setting p_1 optimally. In year 2, the (residual) demand is $p_2(q)=p_1-q$ or $q=p_1-p_2$ so that period 2 price is determined by

$$\max \Pi_2 = p_2\,(p_1 - p_2).$$

The optimal second period price is half the first period price:

$$p_2 = p_1/2 \text{ and } \Pi_2 = p_1^2/4.$$

In year 1, the monopolist who foresees what will happen in year 2, sets the price to maximise total discounted profits:

$$\max \Pi_1 + d\,\Pi_2 = p_1\,(1 - p_1) + d\,p_1^2/4,$$

where δ is the discount factor. The optimal first period price is thus $p_1 = 1/(2-\delta/2)$. If the seller is patient and does not discount the second year ($\delta = 1$), the price path is $p_1 = 2/3$ and $p_2 = 1/3$. A more impatient ($\delta < 1$) seller sets higher prices.

So far we have discussed the price skimming problem assuming consumers do not think ahead. If consumers are rational and forward-looking, the problem becomes more complicated. A consumer, anticipating a lower price in the future, may decide to postpone his purchase unless he is very impatient. This has a moderating effect on the monopolist's skimming plans. However, as long as prices cannot change instantaneously, there will be some intertemporal price discrimination with eager consumers buying earlier and paying more than others. When prices can change instantaneously, consumers anticipate a very rapid decline in price which means that they will refuse to buy until the price is very low (equal to marginal cost). This is the essence of the **Coase conjecture**: a durable goods monopolist who can change price instantly loses all his monopoly power when faced with rational consumers.

When consumers are rational, a seller who can commit to keeping the price up may be better off than a seller who cannot make such a (credible) commitment. It is not easy to make such a commitment and convince consumers you will not renege. Suppose the monopolist charges the one period monopoly price in the first period. Given that not everyone buys at this price, there is a very tempting residual demand to be satisfied in period 2. Generally, consumers will not believe that the seller will be able to resist the temptation to lower the price later on. If, however, a most-favoured customer clause is used so that the monopolist has to compensate earlier buyers if he ever lowers the price, the commitment to keep prices up may become credible. (Models of these phenomena involve consumers forming rational expectations about future price movements and are outside the scope of this syllabus.[9])

Case: Yield management

Airlines are champions of intertemporal and third degree price discrimination. It is relatively rare to find more than a handful of passengers on a plane who have actually paid the same fare. The number of tickets differentiated by a myriad of restrictions is overwhelming. The Saturday night stopover requirement is an attempt by airlines to steer business travellers away from cheap fares offered to tourists. The price difference is usually so large that it makes sense to buy two restricted tickets with a Saturday night stopover, one starting when you want to leave and the other returning when you want to return and just use the outgoing part of the first one and the return part of the second one. In trying to squeeze as much revenue as possible out of certain classes of passengers, airlines sometimes create what seem absurd pricing patterns. It is not uncommon for a return ticket to be cheaper than a one-way ticket, for example. Holiday packages including a flight and self-catering accommodation are sometimes cheaper than the flight only.

Due to computer reservation systems (CRS) it has become feasible to change fares instantaneously in response to rivals' price changes or demand predictions. Travel agents use a CRS to check which flights are available, at which fare and to make bookings. A CRS may be owned by one airline (e.g. the Sabre network belongs to American Airlines) but most are operated by groups of airlines. Amadeus — which is operated jointly by Lufthansa, Air France and Iberia — also has links with American and Asian airlines.

In addition to the CRS which is publicly accessible, airlines run their own sophisticated programs aimed at maximising revenue from each flight. A revenue management system (RMS) has as its objective to sell each seat and to obtain the maximum fare possible. A RMS checks several times a day on how forthcoming flights are filling up and updates its predictions accordingly. If bookings are coming in quickly, the number of low fares on offer is reduced; if it looks like the plane may not be full, more cheap seats are offered. Some full fare seats are always held back until close to take-off for late booking business passengers or passengers who connect to other (long-haul and more profitable) flights. Airlines now even have flexibility with respect to the proportion of seats allocated to first, business and economy classes. The new Boeing 777 airliner uses quick changing bulkheads so that these proportions can be changed rapidly.[10]

Although yield management is mainly associated with the airline industry, hotels face very similar problems and use similar revenue maximising techniques. Discounts for guests booking in advance, differential pricing between weekdays and weekend nights, loyalty bonuses and quantity discounts (e.g. 50 per cent off for a two-night stay) are just a few of the marketing ploys characterising this business. One of the nicest examples of third degree price discrimination must be the deal offered by The Royal Orchid Sheraton Hotel in Bangkok. To celebrate its 10th birthday in 1993, it offered a percentage discount on the room rate rising according to the guest's age.[11]

[9] *If you are interested you could read more in Tirole (1990) 79-87*

[10] *See 'Learning to fly all over again'; 'By the seat of their pants'; 'The high-tech war'*

[11] *Webster (1993)*

Peak-load pricing

For many goods and services, demand fluctuates with the seasons or time of day so that the seller faces several demand curves. The demand at these different times is satisfied by a common facility but typically different prices are charged for peak and off-peak use. The demand for electricity peaks in the morning and the evening and there is seasonal variation because of heating and air conditioning. Telephone companies have lower rates for long-distance calls in the evenings and at weekends. Health clubs sometimes offer off-peak membership at a lower fee. Hotels are cheaper off-season when they have spare capacity.

Charging different prices at different times of day is not really price discrimination if costs vary with output because of overtime or use of less efficient production methods when demand is high. Often marginal cost is constant until a critical capacity is reached. Electricity generating companies use technologically efficient methods for lower levels of output but when these are exhausted they have to use older technologies which could be ten times as expensive per unit of output (Kilowatt hour). The same is true for airlines which charge high fares in times of peak demand when older and less fuel-efficient planes have to be brought into operation. It is not always easy to establish whether a price difference can be justified by a cost difference. In many instances, demand in the peak period is less elastic and therefore, even if costs were the same, a monopolist would charge a higher price.

In its most general form, peak-load pricing is not easy to model. We should really start out with a model of consumers deciding whether they want to consume in the peak or off-peak period depending on the relative prices. However, we will skip this step here and take the demand functions which would result from such a process as given. Let us look at a simple model of peak-load pricing in which the peak $(D_p(p))$ and off-peak $(D_o(p))$ demands have the same elasticity so that the standard third degree price discrimination argument would give identical prices. An example of such a situation is when, for every possible price, peak demand is a multiple of off-peak demand: $D_p(p)= mD_o(p)$, where m is a constant. Assume further that the cost of providing the service, ignoring capacity costs, is $C(q)$, the same for both periods. The same capacity K is used in peak and off-peak periods. The opportunity or investment cost per unit of capacity is constant at c per unit. The profit maximisation problem is:

$$\max \Pi = p_p(q_p)\, q_p + p_o(q_o)\, q_o - C(q_p) - C(q_o) - cK \qquad (6)$$

subject to q_p , $q_o \leq K$, where q_p and q_o are the units sold in peak and off-peak periods respectively. There are two cases we need to consider. The first case is when capacity is fully used in the peak period only $(q_o < q_p = K)$. Optimal prices are derived by rewriting (6) as:

$$\max \Pi = p_p(q_p)\, q_p + p_o(q_o)\, q_o - C(q_p) - C(q_o) - cq_p$$

for which the first order conditions are:

$$MR_p\,(q_p) = MC(q_p) + c \qquad (7)$$

and

$$MR_o\,(q_o) = MC(q_o).$$

The other case we need to consider is when capacity is fully used in both periods $(q_o=q_p=q=K)$. This leads to rewriting (6) as:

$$\max \Pi = p_p(q)\, q + p_o(q)\, q - C(q) - C(q) - cq$$

which leads to the following result:

$$MR_p(q) + MR_o(q) = 2MC(q) + c. \qquad (8)$$

In either case, the optimality conditions (7) and (8) determine quantity sold in each period, which by substitution in the demand functions gives the optimal prices. To determine whether it is best to set prices so that capacity is only binding in the peak period or whether the seller should attempt to equalise demand in both periods, profit has to be calculated for both cases.

Commodity bundling

Sellers sometimes offer special deals when goods or services are bought in packages. For example, it is often cheaper to buy a holiday package which includes flight, accommodation and meals than to buy these items separately. Some computer manufacturers (e.g. Gateway 2000 and Zeos) sell bundled software as part of a package with a computer. PCs are often sold in a package with peripherals. Many durable goods sellers offer financing with their products. Luxury cars may be sold with leather seats, sunroof, wooden dashboard, CD player and driver and passenger airbags included at no extra cost. Again this marketing practice can be explained in terms of price discrimination but, as it is rather special, it deserves a special section.

Commodity bundling is sometimes used when the seller has to offer all buyers the same deal because explicit price discrimination is impossible or illegal. As for most forms of price discrimination, there may be cost-based justifications for offering discounts when a bundle of products is bought. For example, when a stereo system consisting of a specific tape deck, tuner and CD player is sold as a bundle, the manufacturer and the retailer may benefit from this standardisation and resulting economies of scale and lower handling costs. Similarly, a restaurant which offers set meals reduces waste of perishable items.

Commodity bundling, also called **joint purchase discounts**, was first analysed by Adams and Yellen (1976) although the idea that bundling can be used to extract consumer surplus first appeared in a paper by Stigler (1963) who discusses the case of 'block booking' by the American film industry. Monopoly producers used to (before it was ruled illegal) bundle films when offering them to theaters. Theaters could not buy the film they were interested in if they were only interested in one film; they had to buy the bundle or nothing. This is an example of **pure** rather than **mixed** bundling. **Pure bundling** refers to the situation in which the goods are only available as part of a package or bundle; they cannot be bought separately. A newspaper may sell advertising space in the morning paper only if an advertisement is also placed in the evening edition. Mail order film developers include free films when prints are returned. Most tape decks and cassette-radios have built-in speakers. Many professional societies and charities sell journals or magazines to their members. Often membership 'includes' subscription to these publications and members are not allowed to opt out and just pay a membership fee. **Mixed bundling** on the other hand allows for goods to be sold separately as well as in a bundle. Theater groups and concert organisers usually sell separate tickets for individual performances as well as season tickets. Airlines sell one-way as well as round-trip tickets. A restaurant may offer dishes which feature in the day menu on the à la carte menu as well. An interesting marketing practice which is equivalent to mixed bundling consists of offering customers discount coupons for another product in the seller's range.

To illustrate how bundling can be profitable, let us consider the pricing problem of a cafeteria manager who sells apple pie and cappucino. He figures that 50 per cent of his customers are mainly thirsty and the other 50 per cent are mainly hungry. A thirsty customer is willing to pay up to £2 for a cappucino and up to £2 for a slice of apple pie. A hungry customer is willing to pay up to £1 for a cappucino and up to £3 for the apple pie. For simplicity we will ignore costs. (You should, after reading this, be able to repeat the analysis taking costs into account.) If the manager aims to sell apple pie to both groups of customers, he cannot charge more than £2 per slice. This is of course better than just aiming to sell to hungry customers since that would result in expected revenue of £1.50 per customer. Similarly, if the manager wants everyone to buy cappucino, he should charge £1, which gives him the same revenue as selling to thirsty customers only at a price of £2. Using this pricing strategy, the manager gets an (expected) revenue of £3 per customer. Now let's see how bundling can make him better off. For each customer, the sum of reservation prices for cappucino and apple pie is £4. This means that, if the two items are offered in a bundle at £4, revenue per customer increases from £3 to £4.

The groups of customers in this example have negatively correlated reservation prices for the two items. This feature makes the bundling very profitable since the alternatives of either selling to the high willingness to pay group only, for each item, or charging low prices so everyone buys, are not attractive.

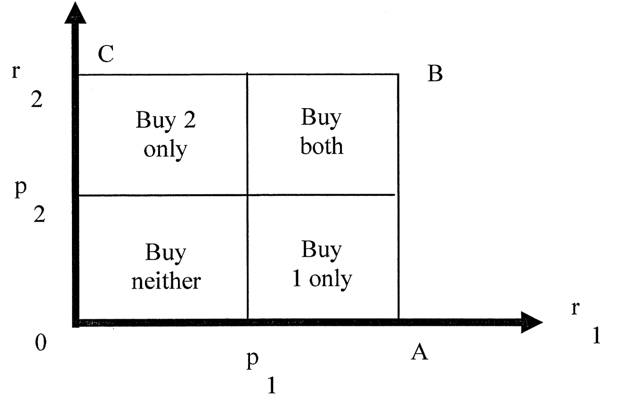
It is however not necessary for reservation prices to be negatively correlated for bundling to be profitable. Schmalensee (1984) points out that pure bundling reduces buyer diversity in the sense that the standard deviation of reservation prices for the bundle is always smaller than the sum of the standard deviations of reservation prices for the goods which compose the bundle. Let σ_1 and σ_2 be the standard deviations of reservation prices r_1 and r_2 for potential consumers of goods 1 and 2. The correlation between the reservation prices is ρ. If consumers' reservation price for the bundle is the sum of the reservation prices for the product (i.e. $r_b = r_1 + r_2$) then

$$Var\,(r_b) = \sigma_1^2 + \sigma_2^2 + 2\,Cov\,(r_1, r_2) = \sigma_1^2 + \sigma_2^2 + 2\,\rho\,\sigma_1\sigma_2 < (\sigma_1 + \sigma_2)^2.$$

The lower the correlation between reservation prices, the greater the reduction in heterogeneity achieved by bundling. Reduced diversity allows for more efficient extraction of consumer surplus and hence increased profits.

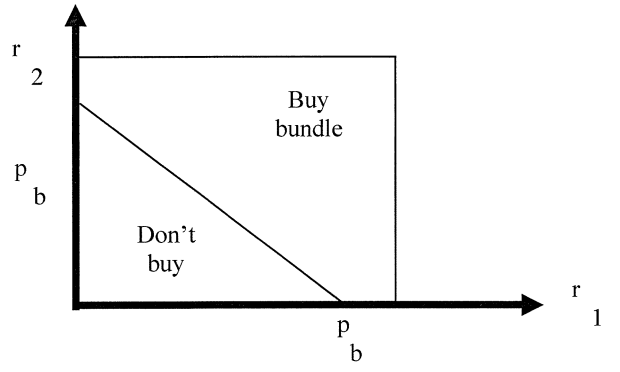
Figure 10.6 illustrates pure and mixed bundling for consumers with independent demands for two goods. Each consumer is represented by a point in the square OABC, marking the pair of reservation prices for both goods. Consumers buy at most one unit of each product and a consumer's reservation price for the bundle equals the sum of his reservation prices for the two goods. Consumers cannot resell goods. If goods are sold separately (unbundled sales) then all consumers with reservation prices above the set prices p_1 and p_2 buy Product 1 and Product 2 respectively (see Figure 10.6a).

Figure 10.6a: Unbundled sales



If the seller only offers the goods as a bundle, then only consumers whose reservation price for the bundle exceeds the bundle price ($r_1 + r_2 \geq p_b$) will buy (see Figure 10.6b).

Figure 10.6b: Pure bundling



In the case of mixed bundling, the seller offers the individual components of the bundle as well as the bundle. Obviously the individual products are priced such that their prices add up to more than the bundle price, otherwise nobody would ever buy the bundle.

Now consumers have four options: (a) buy nothing, (b) buy Product 1 only, (c) buy Product 2 only and (d) buy the bundle. They will make their choice according to which option gives them the highest consumer surplus, that is, according to max ($0, r_1 - p_1, r_2 - p_2, r_1 + r_2 - p_b$). On Figure 10.6c, the consumers who buy nothing are in area OABCD and the consumers who prefer option (b) are in area DCEF since there:

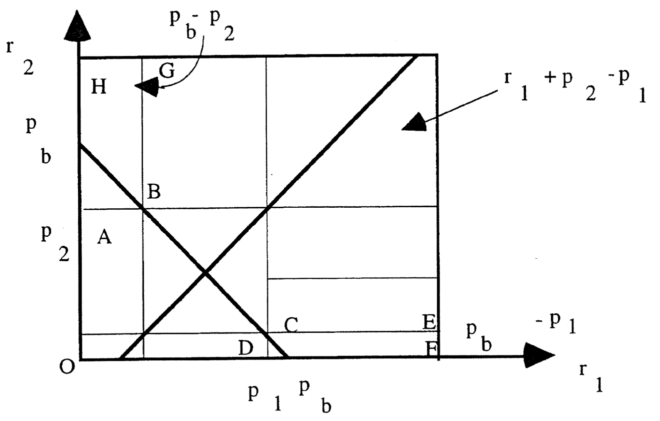
$$r_1 > p_1,\ r_2 < r_1 + p_2 - p_1 \text{ and } r_2 < p_b - p_1$$

Similarly, consumers in area AHGB prefer option (c) since there:

$$r_2 > p_2,\ r_2 > r_1 + p_2 - p_1 \text{ and } r_1 < p_b - p_2 .$$

This leaves consumers in area GBCE to buy the bundle.

Figure 10.6c: Mixed bundling



Commodity bundling is a current research topic. Different assumptions regarding consumers' reservation prices can be made. It may not always be reasonable to assume, as we have done here, that a consumer's reservation price for the bundle equals the sum of the reservation prices for individual items. Also, the effect of different market structures has not been conclusively studied. If the seller is an oligopolist for example he may bundle products to gain an advantage over his competitors. When a good for which the seller is a monopolist is bundled with goods which are sold competitively the monopolist may be trying to extend his monopoly power. In a current lawsuit Microsoft is accused of doing just that. By bundling software packages with the operating system it makes PC buyers who get this software installed 'for free' reluctant to buy a package from one of Microsoft's rivals. Several firms which used to sell programs to do tasks now performed by Windows, Microsoft's operating system, have gone out of business.[12]

Sometimes different sellers offer components of the bundle. In an interesting application of commodity bundling, Lan and Kanafani (1993) discuss park-and-shop discounts. Parking facilities in city centres and commodities people shop for in these city centres, are owned by different firms. Nevertheless the authors quote examples of successful bundling strategies through which shoppers get a discount on car parking. The case of pure bundling in this context is also quite common. It occurs whenever a shop or supermarket parking lot is only accessible to customers (who may have to show a receipt). Schemes which give discounts to shoppers who use public transport are an alternative form of commodity bundling which may be preferable if congestion and pollution in the city centre is a problem. In this spirit, London Underground's one day Travelcards have been used to offer discounts on museum entrance fees and films.

The decision whether to market a product on its own or in combination with another product can be of crucial importance. Commercial Brake is the name of a new gadget which, when used with a video recorder, automatically skips over the commercial breaks. This invention is for sale for $199 and one million units are expected to be sold during its first year. Total yearly VCR sales however are about 50 million. Arthur D. Little Enterprises (ADLE), the consultancy marketing the new technology, not surprisingly, tries to convince VCR manufacturers to sell their products with Commercial Brake fitted.[13]

[12] *See 'Hard sell'; 'A hell of an operating system'*

[13] *See 'Ad nauseum'*

The term **tied sales** is often used to describe a marketing practice which is related to commodity bundling. Whereas in commodity bundling the proportions of the items in the bundle are fixed (e.g. one bar of soap and one towel), under tied sales, the consumer can decide on the quantities in the bundle. Under tied sales, a manufacturer refuses to supply a product for which he has a monopoly unless the consumer agrees to also buy some (complementary) product from the manufacturer. When this complementary product is available in a competitive market, tying sales can be seen as an attempt to extend monopoly power. Usually there is another reason to tie sales, namely to discriminate between types of customers. As part of their standard leasing arrangement for photocopiers, Xerox used to insist that customers bought Xerox paper. This allowed it to price discriminate between high intensity and low intensity users by charging a high paper price. If higher use of paper indicates a higher willingness to pay for the photocopier lease then customers reveal their type through the amount of paper they use. Tied sales is a mechanism to extract more consumer surplus from the keener users. Of course Xerox could have negotiated different rentals with low and high intensity users but this leaves the problem of identifying which class any user belongs to.

Manufacturers who supply replacement complementary goods to the durable good they sell may be tempted to design their products so that only products in their range are compatible. This constitutes a **technological tie**. For example, a printer manufacturer could design his products so that only cartridges from the same make can be used. Camera manufacturers design lenses and camera bodies so that they have limited compatibility — usually only within the brand name — and often there is incompatibility between successive generations of products. In the US there have been several lawsuits concerning technological ties and designed incompatibility.

Multiproduct firms

In mainstream microeconomic models the firm is usually modelled as a single product manufacturer. In our discussion of production and pricing we have focused on the case of a firm producing a single output. This is done for convenience of course as we know that practically all firms are multiproduct firms. In fact some of the most interesting pricing questions arise when firms have to take into account that their pricing policy for one of their products has significant implications on the demand and profitability of another. Generally, when pricing and marketing decisions for individual products are not coordinated the firm does not perform as well as it could. Particularly, a company structure in which separate divisions are given responsibility for the profitability of an individual product can be undesirable.

Example 10.4

Noindent, a company which markets game computers and videogames, has set up two divisions. The hardware division has responsibility for pricing the game machines whereas the software division is in charge of pricing the game cartridges. The demand functions for hardware and software respectively are as follows:

$$p_h = 1000 - 30\,q_h + 2.5\,q_s$$

$$p_s = 2.5\,q_h + 500 - 20\,q_s$$

where p_h and p_s are the prices and q_h and q_s the quantities (in thousands) demanded per week at these prices. Clearly, the two products are complements and increases in the demand for one have a positive effect on the demand for the other. Game computers and video game cartridges are produced at constant marginal cost of £100 and £50 respectively.

The hardware division of Noindent maximises its profits, taking the software division's pricing as given. In game theory terms, we will be looking for a Nash equilibrium. Profit from selling game computers is:

$$\Pi_h = (900 - 30\, q_h + 2.5\, q_s)\, q_h$$

which is maximised (set the derivative with respect to q_h zero) for:

$$q_h = 15 + q_s/24. \tag{9}$$

Substituting this expression for q_h in the demand function results in:

$$p_h = 550 + 1.25\, q_s \tag{10}$$

We can do a similar exercise for the software division with profits:

$$\Pi_s = (2.5\, q_h + 450 - 20\, q_s)\, q_s$$

maximised at:

$$q_s = 11.25 + q_h/16. \tag{11}$$

Substitution in the demand function gives:

$$p_s = 275 + 1.25\, q_h \tag{12}$$

Solving (9) and (11) for q_h and q_s gives $q_h = 15.509$ and $q_s = 12.219$ so that the prices set by the divisions are determined by (10) and (12) as $p_h = 565.27$ and $p_s = 294.39$. The maximum profit when divisions act separately can be calculated for these results as $\Pi_{sep} = 10{,}202$.

Suppose Noindent restructures so that hardware and software pricing decisions are made within one division or profit centre. Such a division has incentives to take demand interactions into account by maximising overall profit:

$$\Pi = (900 - 30\, q_h + 2.5\, q_s)\, q_h + (2.5\, q_h + 450 - 20\, q_s)\, q_s$$

Setting derivatives with respect to q_h and q_s zero and solving for q_h and q_s gives $q_h = 16.105$ and $q_s = 13.263$ so that the profit maximising prices are determined by (10) and (12) as $p_h = 550$ and $p_s = 275$. The maximum profit when the software and hardware pricing decisions are made together is $\Pi_{tog} = 10{,}232$. Prices are set lower when the demand complementarity between products is taken into account. When divisions only care about their own profitability, they tend to set prices too high because they ignore the externality effect of their pricing decision on the other division's profitability.

Firms selling complementary goods often price one product at or below marginal cost in order to stimulate demand (at high prices) for the other products. The cheap products are **loss leaders** aimed at attracting customers for more profitable goods or services. For example, engine manufacturers may sell engines rather cheaply and charge a lot for service and spare parts. Amstrad wordprocessors were priced low to fuel the market for peripherals and software. Airlines sometimes take a loss on certain routes because some passengers on these routes connect to profitable flights. In extreme cases, companies offer gifts of free products to build market share for other products.

An important issue for many multiproduct firms is that of **cannibalisation.** When firms add new products or new varieties of existing products to their range, they may hurt their own sales. Whenever Intel introduces a new generation of microprocessors, it competes against its own previous generation. Firms are sometimes multiproduct firms not by choice but because of technological reasons. This is the case when they create valuable byproducts in the course of manufacturing some other product. There are many examples of such **joint products** in the agriculture and chemical industries. When raising cattle for beef, the farmer also produces leather. Shell creates propylene, the basic ingredient in polypropylene, as a byproduct when it makes ethylene.[14] In contrast

[14] *See 'Better than price-fixing?'*

to the other situations described in this section, joint products are related in production rather than consumption. The joint product scenario forms an extreme case of **economies of scope** which refers to a production technology whereby it is cheaper to produce two products together than separately. For example, it is relatively cheap for a company laying cables for cable television to add a telephone connection.

Let us focus on a monopolist manufacturing and marketing joint products. Each 'unit' consists of a unit of Product A and a unit of Product B. If you find this too abstract, think of the 'unit' as an egg: Product A is eggyolks (which could be used by a mayonnaise manufacturer) and Product B is eggwhites (sold to a manufacturer of low cholesterol egg substitutes). If we know the cost function of the 'composite good' (eggs) and the demand functions of Product A and Product B, how much should be produced? Note that the quantity produced of A always equals that of B. Intuitively we can deduce that production should be increased until the marginal cost exceeds the **sum** of the marginal revenues for A and B. To derive this rule mathematically, write the profit function as:

$$\Pi = p_A(q_A)\, q_A + p_B(q_B)\, q_B - C\,(q), \quad q = max\,(q_A, q_B) \qquad (13)$$

where q_A and q_B are the quantities of A and B sold and $C(q)$ is the cost of producing q units of A and of B. If we assume that everything which is produced will be sold (see below for reasons why this is not always optimal), then we can set $q{=}q_A{=}q_B$. Optimising (13) then gives the following result:

$$MR_A\,(q) + MR_B\,(q) = MC(q). \qquad (14)$$

When the seller is a monopolist an interesting problem can arise. A monopolist's marginal revenue decreases in output and may tend to zero for large enough output levels. It is thus possible that, while MR for Product A is still positive and above marginal cost, MR for Product B is negative. Of course a firm never sells at negative MR since it can avoid this by selling up to zero MR and wasting the excess production. We will see exactly how this works in an example but, before tackling the example, let's think about why this type of waste does not occur for joint products in a perfectly competitive market. In such markets, MR equals price which is positive (as long as the good is a 'good' and not a 'bad'!) and a competitive firm does not need to worry about selling so much that it drives marginal revenue down. Of course there may be waste in a competitive market if the price **net of transportation costs** is not high enough. Gas which is released when drilling for oil is often burnt because its transportation cost from remote oil fields is too high. In the Falkland Islands where sheep are farmed for wool, the meat goes to waste presumably because of the high transportation costs. This state of affairs is not helped by British regulations requiring that all meat for the British forces is imported including 'mutton granules'.[15]

Example 10.5

A pineapple grower in the Bahamas produces cans of sliced pineapples and cans of pineapple juice. The demand functions for cans of pineapples and cans of juice are respectively:

$$q_p = 80 - 5\,p_p$$

and

$$q_j = 50 - 5\,p_j.$$

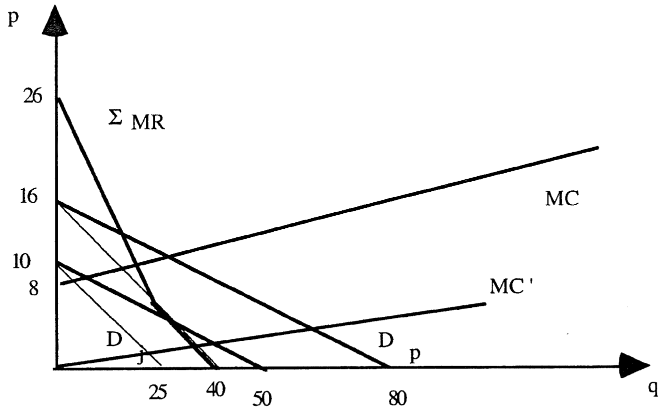
At a cost of $C(q) = 25 + 8q + 0.05q^2$, the production process produces q cans of pineapple slices and q cans of pineapple juice. Applying rule (14) to this problem gives:

$$(80\text{-}2q)/5 + (50\text{-}2q)/5 = 8 + 0.1q$$

[15] *See 'The richest islands in the world, maybe'*

so that $q=20$ pineapple cans and juice cans should be produced. Figure 10.7 shows how to find the optimal quantity graphically by adding the marginal revenue curves *vertically* and identifying the intersection of this sum with marginal cost. Note that we assume that the seller does not sell at negative marginal revenue and hence for $q> 25$ the sum of marginal revenues coincides with the marginal revenue curve for pineapple cans. This intersection is at 20 units and the prices can be read off the demand curves for $q=20$ as $p_p = 12$ and $p_j = 6$. At 20 units of output, marginal revenue is positive for both products.

Figure 10.7: Joint products



Now suppose the cost function is $C(q) = 25 + q^2/35$ and there is free disposal (i.e. it is costless to discard excess pineapple juice). Note in Figure 10.8 that marginal cost (MC') now intersects the sum of the marginal revenue curves where marginal revenue of juice would be negative if all of the juice produced is sold. The profit maximising producer should set $MR_p(q_p)=MC'(q_p)$ or $(80 - 2q)/5 = 2q/35$ which implies production of $q_A= 35$ cans of pineapple. Rather than selling 35 cans of pineapple juice, the seller should restrict sale of juice to the level where marginal revenue is zero (i.e. $q_j = 25$ cans).

Transfer pricing

A transfer price is the price which is charged when one division of a vertically integrated firm (say the production division) sells an intermediate good or service to another division (say the marketing division). When such internal sales take place, maybe in addition to sales of the intermediate good on the external market, the determination of the transfer price is important for the firm's profitability. At first sight this statement may puzzle you: why should it matter what price GM's parts division charges its assembly plants? By definition, the payment for parts is a transfer; it is like putting money from the left pocket into the right pocket. The reason transfer prices are important is that they are used to provide proper incentives for individual divisions to take actions which maximise **total** firm profit. Many large-scale enterprises have a decentralised structure consisting of semi-autonomous profit centers. Such stand-alone divisions are often established to avoid excessive communication and coordination costs in large firms. Since in this scenario divisions are rewarded on the basis of their performance, the selling and buying divisions have opposing interests in the determination of the transfer price. The selling division has an interest in setting the transfer price high, whereas the buying division prefers a low transfer price. If transfer prices are not set properly or if the transfer price is left to negotiation between divisions, total firm profit suffers.

Following Hirshleifer's (1956) seminal paper on transfer pricing we will develop the analysis for three scenarios distinguished by whether there is an outside or external market for the intermediate good and if there is, whether this market is competitive or not. To simplify the analysis we assume that units are defined so that one unit of the final good requires one unit of intermediate good. We also assume that there are no technological synergies between the production and the marketing division (i.e. the cost function for either division is not affected by the output of the other).

Scenario 1: no external market for the intermediate good

The simplest scenario is that of a vertically integrated firm which does not sell or buy the intermediate good on the external market. There may be several reasons for this arrangement. An important consideration is that of transaction costs the firm incurs when it deals with the external market. Hirshleifer gives the example of a steel mill consisting of two divisions exchanging molten iron. Buying and selling of molten iron on the external market is prohibitively expensive. It could be an important strategic decision not to have a market for the intermediate good. American Airlines, for example, has blacklisted some of its information systems such as flight-scheduling and yield-management programmes for sale to its domestic competitors whereas other services such as data punching into computers are sold on the external market.[16] IBM's bad fortune in the early nineties has been blamed on its decision to buy the essential components for its personal computers from the outside market. Instead of using its own microprocessor chips and software, it bought chips from Intel and operating system software from Microsoft. This allowed standardisation in the PC market and made IBM machines very easy to clone.[17]

In what follows we use subscripts *f*, *i* and *e* to denote variables related to the *f*inal good, the *i*ntermediate good and the *e*xternal market for the intermediate good. Since, in this first scenario, the firm does not buy or sell the intermediate good externally, the final good quantity q_f equals the quantity of the intermediate good produced. Given the demand function $p_f(q_f)$ this quantity q_f is the only decision variable. The integrated firm's problem is simply to maximise revenue on the final good market minus the costs incurred by the two divisions:

$$max\ p_f(q_f)\, q_f - C_i(q_f) - C_f(q_f).$$

For profit maximisation the firm should produce a quantity q_f such that the marginal revenue equals **total** marginal cost:

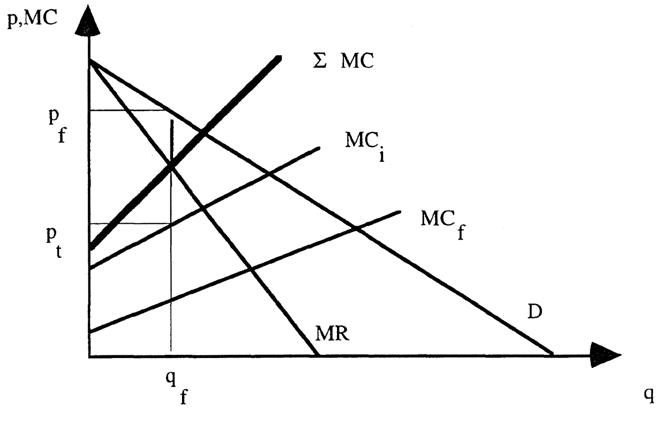
$$MR_f(q_f) = MC_i(q_f) + MC_f(q_f).$$

This result is illustrated in Figure 10.8 where ΣMC is the vertical summation of the marginal cost curves. The intersection of ΣMC and MR, the marginal revenue for the final good, determines optimal output. Of course this analysis also applies when the firm sells the final good in a perfectly competitive market. The only difference there is that marginal revenue is horizontal at the market price.

[16] *See 'Managing the future'*

[17] *See 'What went wrong at IBM'*

Figure 10.8: No external market



Given this determination of the optimal output level, the firm has to find a way of implementing a suitable transfer pricing scheme. If the marginal cost curves are known, the transfer price should be set at p_t in figure 10.8, such that $p_t = MC_i\,(q_f)$. With this transfer price rule in place, the firm can then instruct both divisions to maximise profit subject to this restriction and this will lead to the desired result: the manufacturing division produces the output q_f for which marginal cost equals 'marginal revenue' p_t and the marketing division wants to use exactly what is supplied by the manufacturing division since at output q_f its marginal cost, $p_t + MC_f(q_f)$, equals marginal revenue.

However, we have really only solved the problem for an ideal world where the cost functions are known and where managers do not try to negotiate 'a better deal' for their division. In practice, when the headquarters does not know divisions' marginal cost functions, the divisions may not find it in their interest to volunteer this information or to give accurate information about costs. Especially when conditions are such that one of the divisions is making very low profits, managers will try to bargain to increase or decrease the transfer price. For example, when the manufacturing division has constant marginal costs, its profits are zero under the optimal transfer price rule. If costs are known it may be better in terms of providing incentives to make such a division a cost centre with rewards tied to cost reductions rather than reward it on the basis of profitability.

Scenario 2: perfectly competitive external market for the intermediate good

In this scenario the output of the production division need not be equal to the output of the final product. If there is excess demand the firm buys on the external market and if there is excess supply it sells on the external market. In addition to the assumptions of scenario 1, we assume **demand independence** between the production and marketing division (i.e. sale of an extra unit by either division does not affect demand for the other). This is a strong assumption since generally when the firm sells more of the intermediate good we would expect that less of the final good is demanded and vice versa. This is obviously the case when the firm sells the intermediate product to its competitors in the final goods market. Therefore demand independence is only realistic if the intermediate good is sold to an industry which produces an imperfect substitute or better, a product unrelated to the final good. Hirschleifer quotes the example of a copper concern selling copper to its wire manufacturing division and on the external market to firms which use copper as an input in the production of products other than wire, such as pots and pans.

It is useful to consider the situation in which the firm sells to the external market first. The firm now has to make two decisions: how much of the intermediate good and of the final good to manufacture. It maximises revenue from the final good and from the intermediate good minus costs:

$$max\, p_f(q_f) + p_e q_e - C_i(q_f + q_e) - C_f(q_{f)}.$$

Since the intermediate good market is perfectly competitive, the firm takes the price on that market p_e as given. For profit maximisation we thus have:

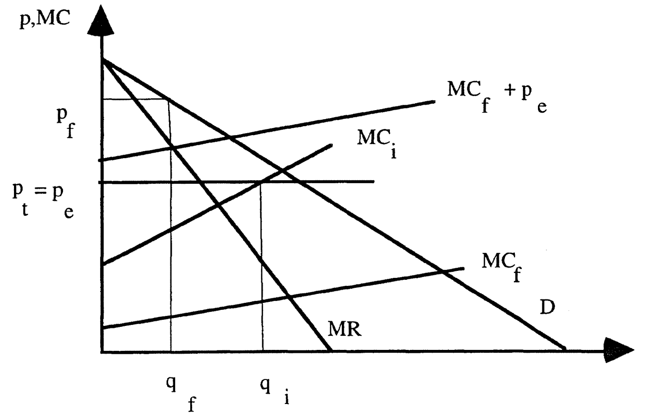
$$MR_f(q_f) = MC_i(q_f + q_e) + MC_f(q_f)$$

and

$$p_e = MC_i(q_f + q_e).$$

Figure 10.9 illustrates these results graphically. To determine the amount of final good to produce, the firm sets marginal revenue equal to the sum of marginal costs **at the optimum output levels** for final and intermediate goods. The optimal output level for the intermediate good $q_i = q_f + q_e$ is determined where its marginal cost MC_i equals the market price p_e. The quantity of the final good produced is then found at the intersection of $MC_f + p_e$ and MR. The external market price represents the opportunity cost of using a unit of the intermediate good internally. Of course the same analysis applies when the final good is sold in a competitive market so that its MR curve is horizontal at the market price.

Figure 10.9: Competitive external market



If the firm is a net buyer of the intermediate good, its optimisation problem becomes:

$$max\, p_f(q_f)\, q_f - p_e q_e - C_i(q_f - q_e) - C_f(q_f)$$

which results in

$$MR_f(q_f) = MC_i(q_f - q_e) + MC_f(q_f)$$

and

$$p_e = MC_i(q_f - q_e).$$

This situation could be illustrated graphically as in Figure 10.9 but with MC_i and p_e intersecting to the left of the optimal output level for the final good.

When there is an external market, profit maximisation generally calls for the vertically integrated firm to use it. The optimal quantities of intermediate and final good produced are not equal. As a consequence, when the divisions are prevented from using the external market, profitability suffers. For example, a cement producer cannot increase profitability when he buys a ready-mix-concrete firm with the sole intention of using it as a captive customer.

The proper transfer price when there is a competitive external market is the external market price. This will induce both divisions to produce the optimal quantities if they are instructed to maximise profit. This transfer price rule is also intuitively appealing to managers of the separate divisions. The manufacturing division would not want to sell below the market price and forego a more profitable outside sale and the marketing division does not want to pay more than the market price. Headquarters, when setting the transfer price in this scenario, does not need any cost information. Because of its simplicity the rule of equating transfer price to external market price is used almost without exception in practice.

Note that the marketing division does not benefit from being vertically integrated with the manufacturing division. It could buy the intermediate good under the same conditions on the external market. There is no rationale for the vertically integrated structure under this scenario. We obtain this rather surprising result because of the assumption of demand independence. In the more realistic setting, when an increase in the quantity of the intermediate good sold on the outside market has a depressing effect on the demand for the final good, we would expect that the firm sets the transfer price below the external market price. On the other hand, when demand dependence and technological dependence are negligible, there is a case to be made for 'outsourcing' or spinning off a manufacturing division which operates in a competitive market. This process has been taking place in the computer industry which used to have a multi-layered vertical structure. Most PC makers in fact restrict themselves to little more than contracting with electronics suppliers, assembling and marketing their machines. Several of these electronics manufacturers arose from restructuring in computer companies which spun off some of their production capacity.[18]

The issues of make-or-buy decisions and demerger are however very complex and we cannot hope that a simple transfer price model like the one outlined in this section will give us all the answers. Arguments in favour of outsourcing have to be balanced with the firm's product differentiation strategy for example. Consumer perceptions and hence valuations of a product may be determined to a larger or lesser extent by what is outsourced. Consumers may not care whether a part in a domestic appliance engine is made in-house. They may, however, due to Intel's advertising campaigns, care about which microprocessor is built into their PC.

Scenario 3: imperfectly competitive external market for the intermediate good

The last scenario we consider is the one in which the firm has monopoly power in both of its markets. It sells the intermediate good internally at a transfer price and externally according to a downward sloping demand curve $p_e(q_e)$. Hence its profit maximisation problem is:

$$max \quad p_f\,(q_f)\,q_f + p_e\,(q_e)\,q_e - C_i\,(q_f + q_e) - C_f(q_f),$$

which results in the following optimality conditions:

$$MR_f(q_f) = MC_i\,(q_f + q_e) + MC_f(q_f) \tag{18}$$

and

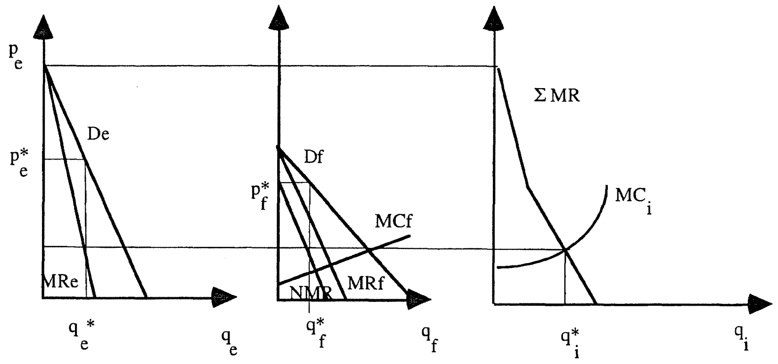
$$MR_e(q_e) = MC_i\,(q_f + q_e). \tag{19}$$

These conditions are similar to the conditions for optimal third degree price

[18] *See 'Labouring in obscurity'*

discrimination. The manufacturing division sets its marginal cost $MC_i(q_f + q_e)$ equal to marginal revenue in the external market $MR_e(q_e)$ and to *net* marginal revenue in the final good market $MR_f(q_f) - MC_f\ (q_f)$. Figure 10.10 shows graphically what is going on. The left panel represents conditions in the external market for the intermediate good. The middle panel shows the derivation of the net marginal revenue curve *NMR* as the vertical distance between MR_f and MC_f. In the panel on the right the manufacturing division sets its output level where marginal cost intersects ΣMR, the **horizontal** sum of the *NMR* and MR_e curves. Again, the analysis for a firm which sells its final product in a competitive market is very similar.

Figure 10.10: Imperfectly competitive external market



The optimal transfer price rule consists of selling internally at a price equal to marginal revenue on the outside market. Because price on the outside market exceeds marginal revenue, the internal division gets a discount (i.e. the intermediate good is sold for less internally). The intuitive reason for this is the double marginalisation problem[19] which would occur if the manufacturing division used its monopoly power internally.

[19] see chapter 12, section on successive monopoly

Example 10.6

Siemips, a major producer of microprocessor chips for telecoms equipment, sells these chips on the outside market as well as internally. It has monopoly power in the chips market as is reflected in the downward sloping external demand $p_e = 1200 - 0.8\ q_e$. The demand for Siemips' final good is $p_f = 1200 - 0.625\ q_f$. Each unit of the final good requires one microprocessor chip. Since Siemips has a policy of only supplying chips to manufacturers it is not in competition with, there is demand independence between the chips and equipment divisions. The equipment division has marginal cost $MC_f(q_f) = 20 + 0.025\ q_f$ and the chips division has marginal cost $MC_i(q_i) = 200 + 0.375\ q_i = 200 + 0.375\ (q_f + q_e)$.

To determine the optimal output level on the final good market, we use (18) and set marginal revenue equal to the sum of marginal costs:

$$1200 - 1.25\ q_f = 200 + 0.375\ (q_f + q_e) + 20 + 0.025\ q_f. \qquad (20)$$

Siemips acts as a monopolist on the external market and sets marginal revenue equal to marginal cost as in (19):

$$1200 - 1.6\ q_e = 200 + 0.375\ (q_f + q_e). \qquad (21)$$

Rewriting (20) and (21) gives

$$q_f = 593.93 - 0.227\ q_e \text{ and} \qquad (22)$$

$$q_f = 2666.67 - 5.267\ q_e. \qquad (23)$$

Equating these two and solving for q_e results in $q_e = 411.27$ and hence from the demand function $p_e = 871$. Substituting into (22) or (23) gives us the optimal output level for equipment $q_f = 500.6$. The transfer price is set equal to marginal revenue in the external market $1200 - 1.6\, q_e = 541.95$ which means that the internal division gets a discount of about 330 on a chip.

In all three scenarios we have considered, the transfer price should be set equal to marginal cost. In scenario 1, marginal cost corresponding to the optimal output level of the final good is relevant; in scenario 2, the market price is used but this is equal to marginal cost when the manufacturing division maximises profits; and in scenario 3, marginal cost corresponding to total production (for internal and external use) is the valid price. In practice, because it is convenient to set transfer price equal to market price, mistakes are made when market conditions are as in scenario 3. More complicated scenarios can be modeled for firms in different market structures and where the assumptions of demand independence and technological independence do not hold. This is however beyond the scope of this subject guide.

Case: Transfer pricing and taxation

Multinational enterprises manipulate their transfer pricing systems in order to redistribute profits between countries so as to minimise their overall tax liability. If country A has lower profit tax than country B then the transfer price will be set low in B so that profits are realised in A. This is illegal but it is very difficult to monitor especially if there is no external market. If there is an external market, the price on the external market can in some cases be used as an estimate of the proper transfer price. In the early nineties the US International Revenue Service investigated American subsidiaries of foreign firms, especially Japanese ones, on the suspicion that they had underpaid US corporate income taxes by as much as $12 billion. Of the nearly 37,000 foreign-owned companies filing returns in 1986, more than half reported no taxable income.

Most countries tax multinationals based on the profits that the firm earned within their borders as if it had made those profits as a stand-alone business at arm's length from the parent company (the **arm's length system**) . Firms are expected to use proper transfer prices which, in many cases, means the prices they would pay if they had to buy the internally transferred goods and services from outside. This is of course problematic when there is no outside market. For many intangibles there is no comparable external market. The advantage of the arm's length system is that profits are normally taxed only once.

California has operated a **unitary tax** where the tax base is simply a slice of the worldwide profits of the multinational determined by the proportion of the firm's worldwide workforce, property and sales sited within the taxing country (or state in this case). The state of California which is strapped for cash can guarantee in this way that foreign companies with Californian subsidiaries pay tax irrespective of whether the subsidiaries (claim to) make a profit. The advantages of a unitary tax system are that the rules for taxing are clear, simple and not easily manipulated but double taxation of profits is likely. Barclays Bank was involved in a 17-year legal battle with California claiming that its profits were subject to double taxation and challenging the state's right to use the unitary tax system. Since Barclays first took the state to court, California has given up unitary taxation. From 1988 firms which are willling to pay a fee and go through the paperwork required to get an exemption can do so and in October 1993 the arm's length system was adopted. Barclays did not drop the case and demanded a refund of taxes paid under the unitary system. On 20 June 1994, the 'tax war' ended with the Supreme Court ruling in favor of California. If California had lost, it would have had to refund $1.5 billion in taxes to Barclays and other multinational firms.[20]

[20] *See 'Taxing questions'; 'Tax deficient'; 'Unhappy returns for Barclays'; 'Unhappy returns'; 'IOU all over again?'*

Chapter summary

After reading this chapter and the relevant reading, you should understand:

- the differences between first, second and third degree price discrimination
- why under first degree price discrimination all consumer surplus can be extracted
- why under two part pricing, at least for identical consumers, unit price should be set equal to marginal cost
- the relationship between price in a market and its price elasticity when third degree price discrimination is practised
- intertemporal price discrimination and the Coase conjecture
- the peak-load pricing model
- the difference between pure and mixed commodity bundling
- the concepts of loss leaders, cannibalisation, joint products and economies of scope
- the joint products model and why a monopolist might waste rather than sell some of his output
- the three transfer pricing scenarios
- the relevance of taxation systems on the practice of transfer pricing.

You should be able to:

- give (your own) examples of price discrimination and commodity bundling
- derive the optimal **take-it-or-leave-it-offer** in first degree price discrimination
- derive the optimal **two-part pricing** scheme
- derive optimal prices and quantities in third degree price discrimination (analytically and graphically) and determine the optimal price and quantity when the monopolist faces two demand curves and is not allowed to price discriminate
- solve a simple two period **skimming** model with myopic consumers
- solve a simple **peak-load pricing** problem
- show graphically how consumers with different reservation prices distribute themselves over the various options in **commodity bundling**
- for a **multiproduct firm** derive prices and quantities for two products when

  a. each product is handled by a separate division and divisions act noncooperatively and

  b. when the decisions are taken to maximise firm profit
- solve a **joint products** problem
- derive the optimal **transfer price** (analytically, graphically and numerically).

Sample exercises

1. This is an extract from 'Serving your needs. A guide to telephone services in your home 1993-94.' published by British Telecom:

   'We work out the cost of each call individually, and measure the call in whole units. Each unit buys a period of time. (The basic unit currently costs 4.935 p including VAT.) The length of time given for each unit depends on when and where you are phoning from, and where your call is to. (The unit rate may vary depending on how many calls you make...We have introduced three 'Customer Options' which give discounts on the basic unit rate depending on the number of calls you make:

   Standard Personal: A 5% discount is automatically given on all direct dialled calls you make over £58.75 in one quarter. This discount increases to 8% on calls made over £293.75 in a quarter. (A quarter is an average of 91 days.)

   Option 15: For a £4 quarterly charge you can apply for our new high-value scheme which offers a 10% discount on all direct-dialled calls at the basic unit rate. You will benefit from this if your call charges are consistently more than £40 per quarter including VAT.

   Supportline: If you make very few calls (less than 125 units per quarter) you can get: half-price standard rental and the first 30 units free. On the next 120 units per quarter, you pay a rate which is higher than the standard unit rate. After you have used 150 units the price falls to the standard unit rate.'

   a. Draw budgetlines corresponding to each of the three 'Customer Options'. The x-axis is the number of units per quarter and the y-axis is remaining income. Remember the rental.

   b. If my current phone bill is less than £40 per quarter, it does not make sense for me to go for Option 15. True or false?

   c. Why does BT offer these options?

2. In 1979 the Belgian government forced BMW to charge a low price for cars sold in Belgium (price ceiling). BMW attempted to ban its dealers from exporting the cheaper cars, offered for sale in Belgium, to other countries. The European Commission condemned these bans. Show the possible effect of the Commission's decision by graphing BMW's price-discriminating strategy (selling at price ceiling in Belgium and at higher price elsewhere) and graphing the non-discriminating scenario. Show that BMW might decide to quit the Belgian market altogether or alternatively, charge the Belgium controlled price throughout the EEC.

3. Off-peak demand for a service is given by $q_1 = 100 - 2p_1$ and peak demand is $q_2=300-p_2$. Marginal cost per unit is $c = 10$ and capacity cost $g=90$ per unit. Find the optimal peak and off-peak price.

4. Suppose Gillette knows it has two types of customers: about 50% are of Type 1 and 50% are of Type 2. A Type 1 customer is willing to pay up to £1 for a razor and up to £1.50 for a package of 10 razorblades. A Type 2 customer is willing to pay up to £0.80 for a razor and up to £1.70 for a package of 10 razorblades. The marginal cost of razors is c_1 per unit and a package of 10 razorblades costs c_2 to produce.

   a. Given that Gillette has to charge the same prices to all customers, what is the best pricing policy? How does it depend on c_1 and c_2 ?

   b. What is the best pricing policy if Gillette sells its products as a bundle consisting of a razor and 10 blades, and the individual products are not on sale?

   c. If it uses mixed bundling how should it price the bundle and the individual products?

5. Due to increased awareness of the dangers associated with a high-fat diet, the demand for skimmed milk has increased. Assuming milk products are sold through a monopoly distributor (a milk marketing board), what is the effect of this shift in demand on the price of cream?

6. The Croucher Corporation which is in the woolly hat industry is composed of a marketing division and a production division. The marginal cost of producing a woolly hat is £10 per unit and the marginal cost of marketing it is £4 per unit. The seasonal demand for woolly hats is p=100-0.0lq. There is no external market for the woolly hats Croucher produces (they look rather odd and are basically worthless unless properly marketed which requires special skills only Croucher's own marketing division can provide).

   a. How many woolly hats are sold each season?
   b. What is the price of a woolly hat?
   c. How much should the production division charge the marketing division for each hat?

7. Harry enjoys conversations about economics with Sally and wants her to be his tutor. Sally does not enjoy tutoring but she enjoys getting paid. She knows that Harry has utility function $U(x,y) = x^2 + y$, where x is number of hours of tutoring per week from her and y is income remaining after paying for tutoring. She also knows that he has an income m per week. Sally's disutility associated with tutoring (or, to be precise, the monetary equivalent of her disutility) is given by $C(x) = x^3/3$.

   a. Because of her monopoly power, Sally can make Harry a take-it-or-leave-it offer (P, x), which means she offers x hours per week for a fee of P. Determine the optimal offer.
   b. Harry's friends get to hear about Sally. Now Larry, Barry and Terry are also interested in economics tutoring. If each has the same utility function as Harry, what offer or offers should Sally now make? [Hint: she does not have to tutor all of them.]

8. Hummer & Co has two divisions. Division 1 produces cotton which is used by division 2 in the manufacture of bandages. The market for cotton is competitive and the price is p_e = \$350 per unit. One unit of bandages can be produced from one unit of cotton. The marginal costs for the two divisions are $MC_1 = 200 + 0.375q$ and $MC_2 = 100 + 0.5q$. Demand for the company's bandages is given by: $p = 1000 - 0.625q$.

   a. Determine the optimal quantity and price for bandages and the transfer price for cotton. How much cotton is bought or sold on the external market?
   b. Suppose now that Hummer & Co has market power in the external market for cotton. It cannot buy cotton from the external market, but it can sell cotton according to an external demand function $p_e = 358 - 0.09q_e$. Determine the optimal quantity and price for bandages, and how much cotton (if any) is sold on the external market and, if any is sold, the transfer price.

Notes

Strategic asymmetry; symmetric models; collusion; dynamic interaction; conclusion and extensions

Chapter 11

Oligopoly

Texts

Tirole, J. *The Theory of Industrial Organization.* (Cambridge, Mass.:The MIT Press, 1988). [ISBN 0262200716] Chapters 5 and 6.

Varian, H.R. *Intermediate Microeconomics.* (New York: W.W. Norton and Co., 2006) seventh edition [ISBN 0393927024] Chapter 27.

References cited

'A bonus for Saddam'. *The Economist,* 11 March 1995, 105–06.

Abreu, D., D. Pearce and E. Stacchetti 'Optimal cartel equilibria with imperfect Monitoring', *Journal of Economic Theory* (1986) 39(1): 191–225.

Allen, B. and J.-F. Thisse 'Price equilibria in pure strategies for homogenous oligopoly', *Journal of Economics and .Management Strategy* (1992) 1(1): 63–82.

Anderson, S.P. and M. Engers 'Stackelberg versus Cournot oligopoly equilibrium', *International Journal of Industrial Organization* (1992) 10: 127–35.

Bertrand, J. Book review of 'Theorie Mathematique de la Richesse Sociale' and of 'Recherches sur les Principes Mathematiques de la Theorie des Richesses', *Journal des Savants* (1883) 68: 499–508.

Boyer, M. and M. Moreaux 'Being a leader or a follower', *International Journal of Industrial Organization* (1987) 5: 175–192.

Brander, J.A. and A. Zhang 'Dynamic oligopoly behaviour in the airline industry', *International Journal of Industrial Organization* (1993) 11: 407–35.

Brannman. L.E. and J.D. Klein 'The effectiveness and stability of highway bid-rigging' in Audretsch, D.B. and Siegfried, J.J. *(eds) Empirical Studies in Industrial Organization: Essays in Honor of Leonard W Weiss.* (Dordrecht: Kluwer, 1992) [ISBN 0792318064] 61–75

'Business and Finance', *The Economist,* 3 December 1994, 7.

'Clean streets', *The Economist,* 12 March 1994, 53–56.

Cooper, T.E. 'Most-favored-customer pricing and tacit collusion', *Rand Journal of Economics* (1986) 17(3): 377–88.

Cournot, A.A. *Recherches sur les Principes Mathematiques de la Theorie des Richesses.* (1838) English edition, Bacon, N. (ed) *Researches into the Mathematical Principles of the Theory of Wealth.* (New York: Macmillan, 1897).

'Disputes are forever', *The Economist,* 17 September 1994, 93–94.

Edgeworth, F. 'La Teoria Pura del Monopolio', *Giornale degli Economisti* (1897) 40: 13–31.

Encaoua, D and Jacquemin, A 'Degree of monopoly, indices of concentration and threat of entry', *International Economic Review* (1980) 21(1): 87–105.

Geroski, PA.. Phlips, L. and Ulph, A *Oligopoly, competition and welfare.*(Oxford: Blackwell, 1985) [ISBN 063114479X].

Gravelle, H.S.E and R. Rees *Microeconomics* (Harlow: FT Prentice Hall, 2004) third edition [ISBN 0582404878] Chapter 16.

Green, E.J. and Porter, R.H. 'Non-cooperative collusion under imperfect price information', *Econometrica* (1984) 52(1):87-100.

Hall, R.L. and Hitch, C.J. 'Price theory and business behavior', *Oxford Economic Papers* (1939) 2:12-45.

Kreps, D.M. and Scheinkman, J.A. 'Quantity precommitment and Bertrand competition yield Cournot outcomes', *Bell Journal of Economics* (1983) 14(2):326-337.

Levinstein, M. 'Price wars and the stability of collusion: a study of the pre-World War I bromine industry', *NBER Working Paper Series on Historical Factors in Long-Run Growth* (August 1993) 50.

Neilson, W.S. and Winter, H. 'Bilateral most-favored-customer pricing and collusion', *Rand Journal of Economics* (1993) 24(1):147-155.

'Oil rolls over', *The Economist*, 2 April 1994, 75.

Ono, Y. 'The equilibrium of duopoly in a market of homogenous goods', *Economica* (1978) 45:287-295.

'Scored', *The Economist*, 20 November 1993, 5.

Spulber, D.F. 'Bertrand competition when rivals' costs are unknown', *The Journal of Industrial Economics* (1995) XLIII(1):1-11.

'Smelt a rat', *The Economist*, 23 July 1994, 72-73.

'Steel woes', *The Economist*, 19 February 1994, 18.

Stigler, G. 'A theory of oligopoly', *Journal of Political Economy* (1964) 72(1):44-61.

'Still smokin', *The Economist*, 11 March 1995, 93-94.

Sweezy, P.M. 'Demand under conditions of oligopoly', *Journal of Political Economy* (1939) 47(4):568-573.

'Then there were seven', *The Economist*, 5 February 1994, 19-24.

von Stackelberg, H. *Marktform und Gleichgewicht.* (Berlin: Springer, 1994).

In terms of number of firms and consumer price, **oligopoly** is an intermediate market structure between monopoly and perfect competition. Firms may price at marginal cost as in perfect competition, they may set a monopoly price or a price between these two extremes. An oligopolistic industry consists of a few firms who recognise their strategic interdependence. This strategic interaction, rather than the number of firms, defines oligopoly. Whereas monopoly, monopolistic competition and perfect competition could be analysed from an optimisation perspective, oligopoly requires a game theory framework. Actions taken by each firm (choice of price or output level) have implications for the payoffs of all firms in the market and therefore individual firms cannot ignore the effect of their actions on their rivals' behavior. In contrast with monopoly and perfect competition, there is no unified theory of how oligopolies behave. Furthermore, the predictions of the various models diverge enormously and depend on exactly which assumptions are made regarding decision variables and strategic symmetry (firms making decisions simultaneously) or asymmetry (leader-follower models). The diversity of models and their predictions correspond to the diversity of real-life behaviour patterns of oligopolistic industries. To decide which model is appropriate, the particular industry under study has to be analysed carefully. There is an abundance of industries which could be described as oligopolies. The US domestic car market contains three major players (the Big Three), namely: General Motors, Ford and Chrysler. In the airline industry, most routes are served by a few airlines.

The oldest and best-known models are models of pure oligopoly (i.e. firms producing a homogeneous good such as zinc or salt). In this chapter we restrict ourselves to homogenous good industries. Oligopoly models can be classified on the basis of their assumptions regarding the toughness of price competition in the industry; on the basis of whether sequential (strategic symmetry) or simultaneous moves (strategic asymmetry) are assumed and according to which variable(s) are used as the decision variable(s). The models we discuss in this chapter are listed in Table 11.1. In addition to these static (one-period) models we analyse repeated versions of some of these standard models.[1]

Table 11.1: Oligopoly models

	tough price competition	sequential/ simultaneous	decision variable
Stackelberg (section 2.1)	no	seq.	quantity
Dominant firm (section 2.2)	no/yes	seq.	price
Cournot (section 3.1)	intermediate	sim.	quantity
Bertrand (section 3.2)	yes	sim.	price
Joint profit max. (section 4)	no	sim.	quantity

Given that in pure oligopoly the product is homogenous, there is an industry demand curve, relating price and total market demand. If firms choose output levels then the price they obtain is determined from this market demand function.When quantity is the decision variable, firms are not making any choices regarding price. Price is determined by the demand curve. This seems unrealistic in that it is hard to imagine how real-life firms can operate without posting prices or publishing price lists. But remember that in the study of monopoly the same problem is encountered and the monopolist's optimal quantity decision generally has to be 'translated', using the demand curve, into a pricing policy to become operational. There is no reason why the same method of analysis cannot also be applied to oligopoly. The oligopolists could determine their and their rivals' 'optimal' or equilibrium production quantities and infer the price from the demand function. You may wonder how price could be a decision variable for firms selling a homogenous product. Surely they will all have to charge the same price? Without running too far ahead let's think of an industry consisting of two firms, A and B, and assume A charges a lower price than B ($p_A < p_B$). If A has sufficient capacity, it can serve the entire market (i.e. the quantity demanded at p_A according to the demand function). If A has limited capacity however, A sells up to its capacity and B faces the residual demand. When A is capacity-constrained it is not *a priori* impossible for B to charge a different (higher) price. In all of these scenarios we can model firms as taking decisions with respect to the quantity **or** the price variable, the other one being determined through market demand. Firms choose price in the Bertrand model and quantity in the other one-period models we will discuss.

In practice, there are several factors determining whether price or quantity is used as the decision variable in a particular industry. For **strictly** homogenous goods it is generally accepted that there are no price choices to be made by individual firms. The price is determined by the market (demand function) and quantity is therefore the strategic variable. For heterogenous goods, price is more likely to be the decision variable but firms could also compete on advertising, capacity investments, service, etc. If the firms

[1] *See the section on 'Dynamic interaction' in this chapter*

in a particular industry customarily print catalogues which list prices, then price is likely to be the strategic variable. Production is adjusted to meet short-term variations in demand when the production process allows for quick changes in the rate of production. Alternatively, inventories are used to deal with variable demand. If the firm has high inventory costs and/or has to plan production ahead, it is likely to consider quantity as its strategic variable and deal with any situations of demand shocks by adjusting its pricing policy. An easy way to adjust pricing in the short-term is through offering discounts from a book price which is set as the maximum the firm would consider charging. In the car industry, quantity is the strategic variable since dealer discounts are adjusted more easily than production schedules. Also when firms are operating close to capacity and changing capacity is costly, a model with quantity as the decision variable is likely to be a good approximation. In this case capacity choice is really the strategic decision. In more sophisticated dynamic models, firms can choose price and quantity and if the chosen price output combination does not lie on the demand curve, inventory is used to absorb the difference between supply and demand.

For convenience we will mainly concentrate on the two firm or duopoly case. Duopoly was the prevalent market structure in the European airline industry before deregulation. On the route between Country A and B an airline from Country A and an airline from Country B (usually the state subsidised airlines) shared the market. Countries had bilateral agreements which determined how many passengers each airline could carry and at what fare. Some well-known duopolies in the US market are Pepsi and Coke (soft drinks) and Procter & Gamble and Kimberley-Clark (disposable diapers). In the UK market for white salt, there are effectively two producers: British Salt (a subsidiary of Stavely Industries) and ICI Weston Point (part of the Mond division of ICI). In the models we consider here it is also assumed that every firm knows the demand function and the cost parameters of all firms in the industry. Each firm has complete information.

You may be familiar with the 'kinked demand curve model' studied by Sweezy (1939) and Hall and Hitch (1939). In this model, oligopolists assume that their rivals will follow any downward change in price but not an upward change in price. This implies that each firm's demand curve is elastic above the status quo price and inelastic below it (a price increase leads to significant reduction in quantity whereas a price decrease is followed by a relatively small increase in sales): hence, the kink in the demand curve. Because of the kink, the marginal revenue curve is discontinuous and, as a consequence, small changes in marginal cost do not call for changes in the optimal price. Because of its relative unpopularity in modern industrial organisation, I will not discuss the kinked demand curve model in detail. The kinked demand curve theory is incomplete in the sense that it does not explain how the initial price (corresponding to the kink) is established. At first sight the model is appealing in that it seems to give an explanation for price rigidities in oligopoly. However, empirically it has been found that prices in oligopoly are not more rigid than monopoly prices.

Strategic asymmetry

In this section we study models in which firms for whatever reason do not make their decisions simultaneously. Simultaneity does not necessarily have a chronological meaning. In game theory 'simultaneous decisions' refers to the situations in which players make decisions ignorant of the decisions taken by the other players. In sequential models there is a 'leader' who makes a first move, after which the 'followers' make their decisions. This type of analysis is obviously only appropriate when there is an industry leader and it is clear which firm assumes the leadership role.

Stackelberg

In the Stackelberg model, the 'leader' (Firm 1) makes a quantity decision, which can be observed by the 'follower' (Firm 2) who in turn decides on quantity. Price depends on total output and the exact relationship between output and price is given by the demand curve $p(q_1+q_2)$. Let's consider the follower's problem first. The follower has to determine an optimal output level q_2 given that the leader has chosen q_1. Hence, his optimisation problem is:

$$\max_{q_2} \quad p(q_1+q_2)\, q_2 - c_2(q_2)\,.$$

The result (i.e. the optimal choice of q_2) will generally depend on q_1: $q_2 = q_2(q_1)$. The leader can anticipate the follower's choice of quantity and its dependence on his own quantity decision and therefore his problem is:

$$\max_{q_1} \quad p(q_1+q_2(q_1))q_1 - c_1(q_1)\,.$$

Example 11.1

Demand in an industry consisting of two firms is given by $p=24 - q_1 - q_2$. Firm 1 is the leader and has constant marginal costs of 8 per unit; Firm 2 is the follower with constant marginal costs of 4 per unit. Given q_1, Firm 2 determines its optimal quantity by maximising:

$$(24 - q_1 - q_2)\, q_2 - 4\, q_2.$$

From the first order condition we find $q_2=10 - q_1/2$. Firm 1 can anticipate this dependence of q_2 on q_1 and maximises $(24 - q_1 - 10+q_1/2)q_1 - 8q_1$ which leads to a quantity choice of $q_1=6$. So the Stackelberg model predicts $q_1=6$, $q_2=7$ and $p=11$ for this industry. The profits are 18 and 49 for Firm 1 and Firm 2 respectively.

Because of the strategic asymmetry assumed in the Stackelberg model (i.e. the leader can influence the follower's quantity choice but not vice versa), the leader generally has an advantage. This first mover advantage is an essential ingredient of the Stackelberg model and makes the model applicable only to industries where such an advantage clearly exists. If we interpret quantity choice as capacity choice, it may be easier to find situations in which the Stackelberg model is empirically plausible. It would mean that the firm which invests first acts as the leader. With this interpretation, one could argue that IBM had a leader's role in the mainframe computer market.

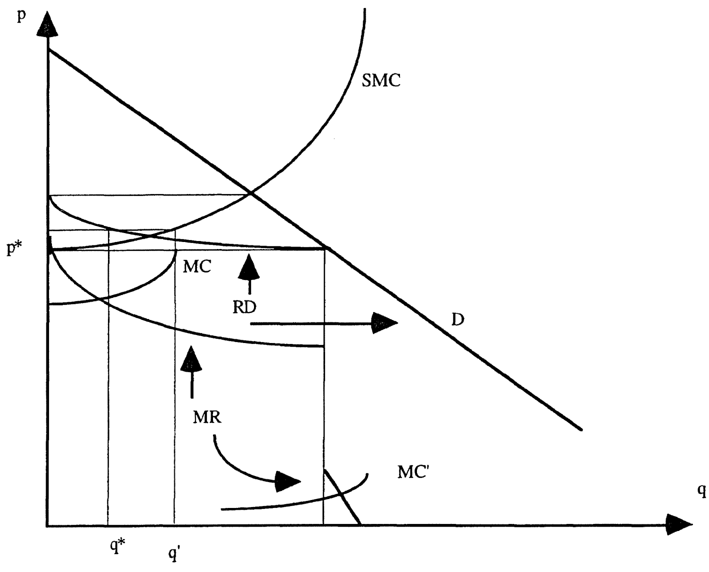
Dominant firm

The dominant firm model applies to industries in which there is one large firm (or a cartel) acting as a price leader and several small firms. The small firms in the industry (the fringe) act as competitive firms in that they take the price set by the dominant firm as given and they supply all they want at this price. In other words, they maximise their profits by determining their supply quantity as the quantity for which marginal cost equals the given price. In contrast to the perfectly competitive model there is no condition of zero profits. The dominant firm supplies the residual demand. As in the Stackelberg model, the dominant firm as the leader has the advantage of influencing the fringe firms' or followers' quantities, in this case by imposing the price they have to sell at.

Figure 11.1 illustrates how the price decision is made. For any price, the dominant firm can determine how much will be sold by the fringe firms by horizontally adding their MC curves. This is indicated by the SMC (sum of marginal cost) curve. The horizontal difference between the industry demand curve D and the SMC is the residual demand (RD) faced by the dominant firm. Profit maximisation by the price leader requires setting marginal revenue (MR) corresponding to the residual demand equal to the firm's marginal cost (MC). This leads to an output of q^* by the dominant firm. The optimal price can be read from the residual demand curve as p^*. Fringe firms supply q' at price p^*.

Note that when the dominant firm has a significant cost advantage as indicated by MC' in Figure 11.1, it finds it optimal to set the price below the fringe firms' marginal cost SMC. The fringe firms therefore do not produce and the dominant firm is the monopoly producer. In practice, a dominant firm may decide not to undercut the fringe firms even when that would deliver the highest profit. The reason is that refusing to accommodate the fringe firms or taking over some of them may provoke legal action.

Figure 11.1: Dominant firm model



If there are no barriers to entry and the small firms are making positive profits, entry will occur, shifting the SMC curve to the right and lowering the dominant firm's market share. For example, US Steel's market share decreased from 75 per cent in 1903 to less than 25 per cent in the 1960s. American Can which used to supply 90 per cent of the tin market in 1901, saw its market share decrease to 40 per cent by 1960. Such market entry may alter the way in which the industry operates. The dominant firm may not be willing to tolerate a very large fringe sector.

Theoretically, for the dominant firm model to apply, the dominant firm should be powerful so that it can credibly threaten to punish fringe firms if they do not accept its price leadership. This punishment would take the form of driving the fringe firms out of business through a price war. Therefore the dominant firm should have low costs, substantial market share and large production capacity to force the other firms in the industry to set the same price. In the UK white salt market, however, there is evidence of British Salt following ICI in its pricing whereas British Salt is the low-cost producer. Examples of industries for which the dominant firm model may be plausible include (dominant firms in brackets):

- the UK chemicals industry (ICI)
- the European soft drinks industry (Coca-Cola)
- the US aluminium industry (ALCOA)
- the US airline industry (American Airlines)
- the European steel industry (British Steel)
- photocopiers (XEROX)
- cars (GM)
- cameras and film (KODAK).

The reasons why particular firms get to be leaders may be related to size, management style or they may be historical. For example, Coca-Cola was exempt from sugar rationing during the war in return for supplying cheap Coke to American troops in Europe. This gave Coca-Cola a significant advantage over Pepsi Co at least in the European market. Federal Express, probably because of first mover advantages, is seen as the price leader in overnight delivery.

The leadership role can switch between large players in an industry. In the US candy bar market Hershey was dominant in the 1960s. This dominance was later challenged by Mars in the early 1970s. The term **barometric price leadership** refers to firms deliberately alternating the dominant firm role to avoid attracting the authorities' attention. The firm which takes the leader role is the first to announce a price change when a change in cost or demand conditions warrants it.

The dominant firm model can be modified to allow for **several** large firms dominating the market.[2] These market leaders can be modeled as joint profit maximisers (a cartel) in which case our analysis above is basically unaltered or they may be assumed to behave in Cournot fashion with respect to residual demand.[3] Industries with a few (rather than one) major players are empirically important. In the US tobacco industry in 1995 for example, three giant companies (BAT Industries, Philip Morris and RJR Nabisco) jointly have a 90 per cent market share.[4]

Example 11.2

Assume industry demand is given by $p=a-bQ$, where Q is total industry output. There are n identical small firms in the industry with cost function $C(q_s)=c_s q_s^2$. The dominant firm has constant marginal costs c_d and produces an output q_d so that $Q=nq_s+q_d$. Suppose the dominant firm sets a price p. The small firms set marginal cost equal to p so that $p=2c_s q_s$ or $q_s=p/(2c_s)$. Hence, residual demand for the dominant firm is given by:

$$q_d=Q-nq_s=(a-p)/b \;-\; np/(2c_s).$$

This can be rewritten as:

$$p=(a-bq_d)/(1+bn/(2c_s)).$$

The dominant firm sets marginal revenue corresponding to this demand equal to its marginal cost:

$$c_d=(a-2bq_d)/(1+bn/(2c_s)).$$

which results in:

$$q_d=(a-(1+bn/(2c_s))\,c_d)/(2b).$$

The optimal price set by the dominant firm can be found by substituting q_d in the residual demand curve, which gives:

$$p=c_d/2+c_s a/(2c_s+bn).$$

For $n=0$ this reduces to $p=(c_d+a)/2$, the monopoly price. The optimal price is decreasing in the number of fringe firms.

[2] See, for example, Encaoua and Jacquemin (1980)

[3] See the section on 'Cournot' in this chapter

[4] See 'Still smokin'

Symmetric models

The theoretical difficulty of justifying asymmetric models lies in the fact that the asymmetry is not explained **within** the model. There have been some recent attempts[5] to endogenise the strategic (a)symmetry in oligopoly models. Unless we know the particular industry we are studying well, and can appeal to, say, historical reasons for the asymmetry, it is perhaps more sensible to assume that no firm has a strategic advantage. This assumption is the starting point for simultaneous moves oligopoly models.

[5] *See, for example, Ono (1978); Boyer and Moreaux (1987); Anderson and Engers (1992)*

Cournot

In the Cournot (1883) model firms make simultaneous quantity decisions. In game theory terminology, the duopolists are engaged in a non-cooperative game and we are interested in the Nash equilibrium of this game (i.e. a pair of quantity choices such that it is in neither firm's interest to alter its choice unilaterally). These strategies are consistent in the sense that no firm has *ex post* regret when it observes its competitor's quantity decision. The good produced by the Cournot duopolists is homogenous and therefore the price is determined by the total quantity supplied: $p(q_1 + q_2)$. To find the Nash equilibrium, we find best response functions (i.e. we find the best quantity choice of Firm 1(2) as a function of Firm 2(1)'s quantity choice). The intersection of these best response functions gives the Nash equilibrium. Mathematically this means that for Firm 1 we solve:

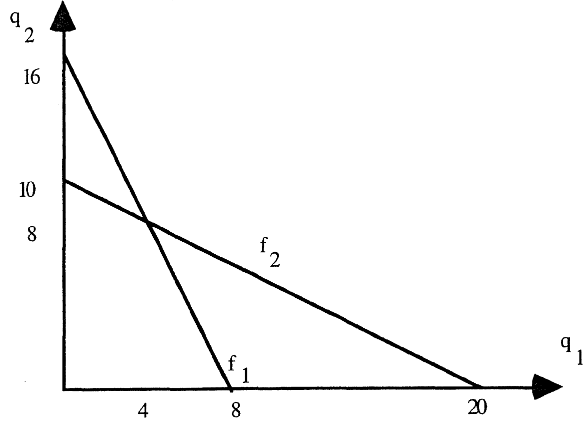
$$\max_{q_1} \qquad p(q_1 + q_2)\, q_1 \; - c_1(q_1)$$

and find $q_1 = f_1(q_2)$ and for Firm 2 we find $q_2 = f_2(q_1)$ similarly. The Nash equilibrium is then found by solving $q_1 = f_1(f_2(q_1))$.

Example 11.3

Assume the same demand function and marginal costs as in example 1. The firms' profits can be written as $\pi_1 = (24 – q_1 – q_2 – 8)q_1$ and $\pi_2 = (24 – q_1 – q_2 – 4)q_2$. The first order conditions lead to the response functions $q_1 = f_1(q_2) = (16 – q_2)/2$ and $q_2 = f_2(q_1) = (20 – q_1)/2$. These response functions are drawn in Figure 11.2. The unique Nash equilibrium is at the intersection where $q_1 = 4$ and $q_2 = 8$. Hence, the Cournot model predicts a price of 12. Compared to the Stackelberg model, the Cournot model predicts a lower output for Firm 1 (the Stackelberg leader) and more for Firm 2 (the follower). Firm 1 has lower profits and Firm 2 has higher profits than in the Stackelberg model. These conclusions hold generally, not just for this example. In the Stackelberg scenario the leader has the opportunity to select a high output to induce the follower to cut back his production. In the Cournot model this is not possible because firms take output decisions simultaneously.

Figure 11.2: Cournot response functions



Example 11.4

Consider an industry consisting of n identical firms with constant marginal cost c and demand function $p(Q)$ where Q is total industry output. Firm i's profit function is given by:

$$\pi_i = p(Q)q_i - cq_i$$

Its optimal quantity decision given the other firms' quantity choices is determined by the first order condition:

$$p(Q) + q_i \frac{\partial p}{\partial Q} - c = 0.$$

Since all firms are identical we look for a symmetric Nash equilibrium in which all output decisions are identical i.e. $q_{i=}$ Q/n so that the condition above can be rewritten as:

$$p + \left(\frac{Q}{n}\right)\frac{\partial p}{\partial Q} = c,$$

which, using the definition of demand elasticity, reduces to:

$$p - \frac{p}{n\eta} = c$$

Or

$$\frac{p-c}{p} = \frac{1}{n\eta}.$$

We conclude that, at the Cournot solution, each firm's monopoly power, measured as its markup-price ratio, varies inversely with the demand elasticity and the number of firms in the industry. If the number of firms n is very large, the Cournot solution approximates the equilibrium of a perfectly competitive industry.

Bertrand

Bertrand (1883) wrote a book review of Cournot's work and disagreed with Cournot about his assumptions on the behavior of oligopolists. Bertrand thought that oligopolists would collude rather than compete and to demonstrate how unrealistic Cournot' s description was, he developed a model in which the decision variable is price. He showed how Cournot's way of thinking leads to the implausible result that in a duopoly the perfect competition price is charged.

We can think of the Bertrand model as a non-cooperative game in which each firm decides on price, taking the other prices as given. We then look for a Nash equilibrium as in the Cournot model. The assumptions of the Bertrand model are that the product is homogenous, consumers have complete information and there are no search or transport costs so that they buy from the firm charging the lowest price. If firms charge the same price they are assumed to share the market and have identical market shares. Assume firms have identical and constant marginal costs and unlimited capacity. Under these assumptions we find the very striking result that there is a unique Nash equilibrium in which both firms set price equal to marginal cost. Why is this the Nash equilibrium? Clearly, price could not be below marginal cost since then firms would be making losses. Price cannot be above marginal cost since then one of the firms could reduce its price by a very small amount, capture the whole market and make a positive profit. If firms differ in (constant) marginal costs, then the low cost firm prices just below the second lowest marginal cost at equilibrium unless that price is above the monopoly price.

Example 11.5

Assume the same data as in Example 11.1. We look for a Nash equilibrium in prices. For any price at or above Firm 1's MC, Firm 2, which has the lower MC, can undercut Firm 1 and capture the whole market. At the Nash equilibrium Firm 1 charges $p_1=8$ and Firm 2 charges $p_2=8 - \varepsilon$ so that Firm 2 is the sole supplier. Note that Firm 2 is not charging the monopoly price however (check that a monopolist would charge $p = 14$) since that would induce Firm 1 to enter the market. The Bertrand model thus predicts a price of $8 - \varepsilon$ and a quantity of $16 + \varepsilon$. Compare this to the predictions of the Cournot model in which the high cost producer, because of the lower intensity of the price competition, obtains a 1/3 market share.

The assumption of consumers buying from the firm charging the lowest price, even if the price difference is very small , is quite restrictive and justifiable only when the goods are very close substitutes. In fact, oligopolists strategically differentiate their products to reduce tough price competition. When products are differentiated, a firm can charge a higher price than its rivals without losing all its business. Even when the good is homogenous, it is not realistic to assume that all demand goes to the lower price firm. Allen and Thisse (1992) have developed a **pure** oligopoly model in which some customers do not care about small price differences. This implies that firms do not lose all of their sales if they are undercut slightly by a rival. The equilibrium of their pricing game allows for some market power by the firms. Spulber (1995) shows that the assumption of firms knowing rivals' costs is a crucial ingredient in the Bertrand model. If firms know only their own costs and have probabilistic information about rivals' costs, they set price above marginal cost at the equilibrium of the pricing game.

It is difficult to think of situations in which the Bertrand model is a reasonable approximation of reality. The scenario which fits the original Bertrand model best is maybe that of firms submitting sealed bids for a procurement contract where the contract is awarded to the firm quoting the lowest price. However, even in these circumstances, firms find ways to avoid tough price competition. In the 1950s, General Electric and three other firms rotated sealed-bid business for circuit breakers. They held secret meetings to decide market shares in advance. More recently, two of America's biggest bakeries, Continental Baking and Campbell Taggert, have been investigated for alleged bidrigging in sales of bread to schools and hospitals. In Japan in 1993, the Fair Trade Commission conducted raids on some big electronics companies, including Sony and Toshiba, who allegedly colluded on bids for electronic billboards in sports stadiums.[6]

Whereas Bertrand had assumed that firms have sufficient capacity to serve the entire market, Edgeworth (1897) modified Bertrand's analysis by restricting firms' output levels to exogenously determined capacity levels. As an example, think of airlines offering scheduled flights between London Heathrow and New York JFK. In the short-run at least, an airline cannot change the number of seats available on the route (unless the flight frequency is increased which is problematic because of a shortage of take-off and landing slots). If price is set equal to marginal cost, there may be excess demand. When this is the case, both firms pricing at marginal cost and making zero profit is not an equilibrium. One firm could increase price slightly and make positive profits. Of course, customers would prefer to buy from its rival (pricing at marginal cost) but they are rationed there. The analysis of the Edgeworth model is quite complicated and I will not discuss it here other than to state the conclusion: with capacity constraints duopolists competing on price set price above marginal cost unless they have excess capacity to the extent that one firm could serve the entire market demand for price equal to marginal cost.

[6] *See 'Scored'*

Since the predictions of Bertrand and Cournot are very different, it is important to decide whether price or quantity is the strategic variable. In a much quoted paper, Kreps and Scheinkman (1983) allow duopolists to use both variables. In a first stage, 'capacity' is chosen which limits each firm's output in the second stage. In this second stage, firms set prices. The model is in fact an Edgeworth model with capacities endogenised. The firm which sets the lowest price can produce and sell up to its capacity whereas the high price firm is left with any residual demand. Kreps and Scheinkman show that, at the unique subgame perfect equilibrium of this game, the capacities chosen in the first stage are the Cournot equilibrium quantities and in the second stage firms choose (identical) prices such that demand equals joint capacity.

Collusion

Chamberlin was one of the first economists (after Adam Smith) to suggest that oligopolistic sellers would form a cartel and cooperate to set joint profit maximising price and output levels. Cartels were common before the antitrust laws were passed. Levinstein (1993) vividly describes how the US and Europe bromine producers colluded in the 30 years before World War I. Currently, in most of the western world (Switzerland is a notable exception), the norm is that collusion is illegal although some cartels are sanctioned by governments. For example, the International Air Transport Association, consisting of American and European airlines flying transatlantic routes, used to set uniform fares for transatlantic flights. The US allows agricultural committees to determine prices and production quotas for some products. Sometimes collusion between firms in the domestic market is explicitly forbidden but governments allow firms to participate in international cartels. The European steel industry operated as a price fixing cartel until 1988 and was fined in 1994 by the European Commission.[7] In the same year, the Commission fined 42 cement firms and trade groups for running a cartel.[8]

[7] *See 'Steel woes'*

[8] *See 'Business and Finance'*

Because price fixing is illegal, there are not many **open** cartels but firms may nevertheless find ways to cooperate at the expense of customers. Powerful firms use the media to announce price changes which are followed by the other firms in the industry. All firms in the industry use the leader's price as a focus point and the risk of 'misunderstandings' is minimal. This is a form of implicit or **tacit collusion.** Trade associations or professional associations are often vehicles for collusion. This certainly seems to be the case for the Associations of Trade Waste Removers and their central council which hold a tight grip on the rubbish-collection business for commercial customers in New York City. The rate charged by New York carters is fixed at three or four times the rate in Los Angeles or Chicago. The business is notorious for its connections with the Mafia.[9]

[9] *See 'Clean streets'*

Technically, the problem of a cartel is at first sight identical to that of a monopoly operating several plants. The optimal output is found by horizontally summing marginal cost curves and determining the intersection of this cartel marginal cost curve with the marginal revenue curve (or the horizontal sum of the marginal revenue curves if there are several markets). The output is allocated to individual plants such that marginal cost is equal in all plants and equal to marginal revenue. For joint profit maximisation, it is necessary to allocate high output quotas to low cost firms and low quotas to high cost firms which may even shut down if they are too inefficient.

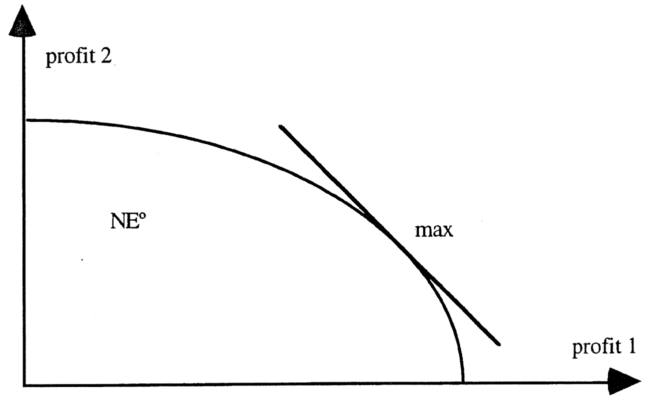
In an oligopoly, finding the optimal output division between cartel members is only a first step. An important but often ignored problem is how to divide the cartel profit. If firms can bargain over sidepayments, it is not as difficult to agree on optimal quotas as when no sidepayments are allowed. By definition of the collusive solution (i.e. that it is the solution which maximises joint profits) it is always possible to make all cartel members better off than they would be in a non-cooperative setting such as Cournot. This is not to say that the division of profits is trivial when side payments are possible but it seems that the bargaining should not be too difficult. Because of the problems associated with divisions of profits, as well as uncertainties about market demand, output allocations are often clearly suboptimal in real life cartels.When there are no sidepayments, each cartel member wants a large share of the output. Production allocations may be based on past sales, capacity or a geographic segmentation of the market ('market sharing cartels'). Having the most skilled negotiator may be more important than relative production efficiency.

Usually cartel members do not make contractual agreements about sidepayments or redistribution of profits since this would be evidence of (illegal) collusion. In this case (if firms are not identical), some firms may be better off at the Cournot solution than at the joint profit maximising solution. This is illustrated for a duopoly in Figure 11.3 where a **profit possibility frontier** is drawn. By definition, all combinations of profits under this curve can be attained by a suitable choice of q_1 and q_2. The curve itself is derived by solving:

$$\max_{q_1, q_2} \lambda\pi_1 + (1-\lambda)\pi_2$$

for all values of the parameter λ between 0 and 1. The point which maximises joint profits corresponds to the solution to the optimisation problem for λ = 1/2. NE in Figure 11.3 indicates the Cournot-Nash equilibrium payoff pair. By moving from NE to the point (max) where joint profits are maximised, firm 2's profit decreases.

Figure 11.3: Profit possibility frontier



Mathematically, a cartel maximises joint profit:

$$\max_{q_1, q_2} \pi = p(Q)(q_1 + q_2) - c_1(q_1) - c_2(q_2)$$

where $Q = q_1 + q_2$. The first order conditions are:

$$\frac{\partial \pi}{\partial q_1} = \frac{\partial p}{\partial Q} \times (q_1 + q_2) + p(q_1 + q_2) - MC_1(q_1) = 0 \qquad (1)$$

and

$$\frac{\partial \pi}{\partial q_2} = \frac{\partial p}{\partial Q} \times (q_1 + q_2) + p(q_1 + q_2) - MC_2(q_2) = 0.$$

which gives the result mentioned above: marginal revenue equals each firm's marginal cost at its optimal output level.

We can use this analysis to show that, for cartel members, since they are playing a prisoners' dilemma game, there is always a temptation to cheat by overproducing. Given that firms, even when they have formally agreed to collude, have no recourse to a court of law (because collusion is illegal), this temptation is very real. Consider firm 1's profit given that firm 2 produces the joint profit maximising quantity q_2*:

$$\pi_1(q_1, q_2^*) = p(q_1 + q_2^*)\, q_1 - c_1(q_1).$$

If firm 1 assumes that firm 2 is going to stick to q_2* it finds its optimal output level q_1* *by* differentiating this function with respect to q_1 which gives:

$$p(q_1 + q_2^*) + q_1 \frac{\partial p}{\partial Q} - \frac{\partial c_1}{\partial q_1}. \qquad (2)$$

Comparing this to the first order condition for q_1 above in (1) reveals that (2) is equal to:

$$-p(q_2)\frac{\partial p}{\partial Q}$$

which is positive. This indicates that Firm 1 can increase its profits by increasing output if Firm 2 keeps output constant at the 'collusive' amount. The intuition for this is that marginal revenue for the cheater is close to the cartel price whereas his marginal cost equals marginal revenue and is thus lower than the cartel price.

Even when cartels can monitor and enforce prices effectively, cartel members may engage in non-price competition by offering free extras such as after-sale service or free delivery and installation. Airlines offer better meals, free in-flight entertainment etc., in order to lure customers away from other cartel members. This type of non-price competition can be interpreted as cheating and it is not as easily detected as price cutting.

Case: OPEC

The standard example of a cartel is OPEC which restricts oil output by its members so as not to spoil the market. OPEC's history has repeatedly illustrated the difficulties of maintaining a cartel and avoiding cheating by cartel members. Within OPEC there is tension between the rich countries with large oil reserves and a small population (e.g. Kuwait and Saudi-Arabia, which alone accounts for a third of OPEC's output) and poor countries with small oil reserves and a large population (e.g. Libya and Indonesia). The countries with large reserves want a low price since they are concerned about the long-term effect of high oil prices: customers may start switching to other energy sources and there may be market entry in the form of non-OPEC exploration and production such as in the North Sea. Countries with low reserves do not worry about the long-term effects and, since short-term demand elasticity for oil is low, they argue for a high price. The poor OPEC countries have consistently overproduced their quotas and even the United Arab Emirates decided to increase its production to 1.5 million barrels a day from its allocation of one million barrels in 1988. The lifting of sanctions against Iraq, imposed by the UN after the invasion of Kuwait, may put more pressure on the cartel.[10]

[10] *See 'Oil rolls over'; 'A bonus for Saddam'*

Case: Diamonds

When industries succeed in filtering sales through a common distributor, collusion is not an unlikely mode of behaviour. The Central Selling Organisation (CSO), a subsidiary of De Beers, aims to control the world wholesale market in rough diamonds. In 1993, De Beers itself accounted for 50 per cent of CSO sales and Russian sales amounted to 26 per cent. Even this seemingly cosy cartel arrangement has had problems. Russia agreed in 1990 to sell 95 per cent of its output through the CSO for a period of five years but it is unlikely that this agreement was honored. In particular, Russia has exploited loopholes such as classifying diamonds as 'technical' and hence not covered by the agreement. De Beers has had to buy this leaked output to keep control of the flow of diamonds on to the market and hence their price.[11]

Given the incentive to overproduce, cartels use several mechanisms to detect and prevent cheating e.g. forcing firms to publicise prices. Firms may adopt 'facilitating practices' which eliminate the incentive to cheat. For example, cartel members may agree to offer a 'meet or release clause'. Such a clause applies when a customer finds a lower price offered by another supplier, in which case the low price is met by the original seller or the customer is released from the obligation to buy. This way cartel members are informed if any undercutting takes place. Another 'trick' used by colluders is the 'most favored customer clause' (MFC). Under MFC, firms promise their customers that if they ever lower the price, they will offer a rebate, equal to the difference between the price customers pay now and the new price. In this scenario, if firms can agree to collude for a small number of periods while offering the MFC, then their payoffs are such that deviating from the collusive outcome by lowering price is no longer profitable because of the penalties which would have to be paid to previous customers. Both General Electric and Westinghouse, manufacturers of turbine generators, used MFC, effective for six months after a sale, in the 1960s and 1970s until they agreed to end the practice to avoid antitrust prosecution.[12]

Dynamic interaction

So far we have discussed models which describe oligopolistic behavior when the firms play a one-shot game. More realistic models assume a dynamic setting and in particular the **repeated** prisoner's dilemma provides a natural framework. As we have seen before, collusion can be an equilibrium if the time horizon is infinite or if there is uncertainty about the length of the horizon. Such a collusive equilibrium is sustained by players' threats to retaliate (i.e. to revert to the non-cooperative Nash outcome as soon as one player deviates). To illustrate this idea consider a simple infinitely repeated Bertrand game. All n firms are identical and the price which maximises joint profit is p^m, the monopoly price. If all firms charge p^m each makes a profit equal to a share $1/n$ of the monopoly profit Π^m. If one of the firms undercuts however, it captures the whole market. Although repetition of the one-period Bertrand solution with all firms pricing at marginal cost in each period is an equilibrium in this game, there may be a Nash equilibrium in which firms collude. Suppose firms use **trigger strategies** of the following type: collude (i.e. set price p^m) until at least one firm deviates; when one or more firms deviate in a given period, set price equal to marginal cost from the next period onwards. For this combination of trigger strategies to form an equilibrium it should be the case that it does not pay for any firm to do something else assuming all the other firms are sticking to their trigger strategies. If a firm decides to deviate, it gains:

$$\Pi^m - \Pi^m/n = \Pi^m \ (n-1)/n$$

[11] *See 'Disputes are forever"*

[12] *See Cooper (1986) and Neilson and Winter (1993)*

immediately as it becomes a monopoly. Its loss, compared to the collusive outcome, due to retaliation from the next period onwards is:

$$(\Pi^m/n)(\delta + \delta^2 + \ldots) = (\Pi^m/n)\,(\delta/(1-\delta))$$

where δ is a discount factor. The firm cheats if the gain exceeds the loss or when:

$$\delta < 1 - 1/n.$$

This proves that, at least in this simple pricing model, cheating is more likely when the number of firms in a cartel is large. When $n = 50$, a discount factor above 0.98 is needed to sustain collusion whereas for an industry with $n=2$ firms, a discount factor above 0.5 is sufficient. The dependence on the discount factor is clear: when firms are patient (high δ), they evaluate their future losses more highly and this deters them from cheating.

We have analysed repetition of the Bertrand pricing game here but a similar story can be told about repeating the static Cournot game. The trigger strategies there involve playing Cournot equilibrium output forever as soon as cheating occurs. In contrast to the Bertrand situation where you don't sell at all when someone cheats, firms in a Cournot industry may not be able to observe cheating immediately but maybe two or three periods after it has taken place. Cheaters could get away with earning profit above their share of the collusive profit for several periods. Detection lags therefore make cheating more tempting.

Stigler (1964) has pointed out that, when demand fluctuates and there is a lot of uncertainty, there is larger scope for misunderstanding moves which merely reflect changed demand conditions as attempts to cheat. This observation has formed the basis for recent game theoretic work on dynamic oligopoly.[13] Recall that the repeated Bertrand and Cournot models predict that, when collusion is sustained, prices and outputs are stable. Firms choose the joint profit maximising price or output and there are no price wars. In the recent models which allow for uncertainty, price wars do occur. To illustrate the main ideas here, let's assume a homogenous oligopoly with identical firms. The time horizon is infinite or there is a chance, in each period, that the game ends. Firms decide on output levels and they observe only their own output level and the market price (which, as always, depends on total industry output). Each firm uses these observables to deduce whether a rival has defected from the (tacitly) collusive arrangement (i.e. when the market price is low, overproduction, compared to the joint profit maximising output, has taken place). So far, we are describing a repeated Cournot game. The ingredient which has to be added to generate equilibrium price wars is uncertainty. The relationship between industry output and price is stochastic so that a low price can be caused by a rival flooding the market or an exogenous fall in demand.

Green and Porter (1984) show that, when there is uncertainty, there is an equilibrium in which firms periodically revert from the collusive solution to static Cournot output levels for a few periods, after which they return to the collusive output levels. At this equilibrium, firms are using trigger strategies which tell them to start behaving non-cooperatively when price falls below a given level. The non-cooperative punishment phases are called price wars. What is interesting in this and other **regime-switching** models is that no firm ever defects at this type of equilibrium yet price wars do and have to occur to sustain the equilibrium. In other words, if there is no trigger in terms of a low price level setting off a price war, it would not be in any individual player's interest to continue to collude. Paradoxically, price wars can be seen as evidence of collusion. If firms act non-cooperatively (play static Cournot in each period), there is no reason for them to revise prices periodically and consequently price and output is stable. However,

[13] See for example Green and Porter (1984); Abreu et al. (1986)

when they try to collude, punishment periods **are** necessary. In the Green-Porter model, the 'collusive' output is not the output a monopolist would set. When determining output, firms have to take into account that when a low output level is set, the temptation to cheat is higher and in order to sustain the collusive equilibrium the punishment periods would have to be longer.

In terms of empirical predictions from this type of model, we should expect to observe price wars in recessionary periods when there is surplus capacity. There is some casual evidence for price wars triggered by an exogenous fall in demand. Green and Porter offer the American rail freight industry in the 1880s as an example of an industry which behaved in a manner consistent with their model. The European car market went through price wars when it declined by 16 per cent in volume in 1993. It has overcapacity of at least three million cars per year.[14] In an attempt to test Green-Porter type models, Brander and Zhang (1993) have analysed data for the 1984-1988 period on economy and discount fares on routes to or from Chicago for which American Airlines and United Airlines are duopolists. They find some evidence in favour of regime switching behaviour with quantity as the decision variable and reversion to Cournot. Punishments appear to have taken place during the first three quarters of 1985 and the first quarter of 1987. The hypotheses that American and United were playing one shot Bertrand or Cournot or joint profit maximising, were rejected.

Case: Aluminium

Aluminium producers certainly seem to be aware of the dangers of excess capacity. They agreed to cut back capacity in 1994 and prices and profits have increased since this agreement. However, Russia — which is a large producer and exporter — is believed to be cheating by cutting down its smelting capacity by much less than the agreed 500,000 tonnes. Furthermore, it is quite possible that, without the agreement, Russia would have had to close more capacity than it has done now. Russian costs have increased and inefficient Russian smelters are kept in business only because prices have climbed. Also, as firms become more profitable, the temptation to cheat and bring capacity back into production may be hard to resist even for 'honest' western producers.[15]

There are models which reach exactly the opposite conclusion to Green-Porter, namely: that cartels are less stable during expansionary periods because firms then have less capacity or inventories to punish the cheaters. Also the gain from cheating is larger when there is a boom in demand. Studies of construction industry auctions concur with this latter prediction.[16]

Conclusion and extensions

In this chapter we have discussed Stackelberg, dominant firm, Cournot, Bertrand and collusion models as the basic models of oligopoly. These models vary with respect to the decision variable firms are assumed to manipulate and the toughness of price competition between sellers. The Bertrand model assumes an extreme degree of price competition; the joint profit maximisation model assumes no price competition; the Cournot model is somewhere in between. The Bertrand model predicts a low price and high output whereas the joint profit maximisation or collusion model predicts a high price and low output. Whereas Bertrand, Cournot and joint profit maximisation models assume simultaneous decision-making, in the Stackelberg model and the dominant firm model, it is assumed that one firm (the leader) makes a decision which can be observed by the follower(s) who in turn make their decisions taking the leader's decision into account. From a theoretical point of view, the simple leader-follower models are not complete because they do not explain how firms adopt leader or follower roles. More recent versions of the Stackelberg model endogenise the role choice and are therefore more sound.

[14] *See 'Then there were seven'*

[15] *See 'Smelt a rat'*

[16] *See, for example, Brannman and Klein (1992)*

The predictions of the standard models are not always very convincing and this is the reason why industrial organisation economists have proposed alternative and often more complex models. In developing these models, they have dropped or revised some of the assumptions in the older models, for example the **one-shot assumption** which leads one to ignore the fact that real-life oligopolistic firms interact with each other over time. We already know from the discussion of the repeated prisoners' dilemma, that repetition **can** make a difference. In the infinitely repeated Cournot game, firms could play trigger strategies (or other strategies which allow for punishment if one player cheats) as equilibrium strategies so that they both set low output rates (and hence a high price is sustained). This could ensure that a point on the profit possibility frontier is reached but the Nash equilibrium of the one-shot game, as well as a large number of other combinations of strategies, is also an equilibrium in the repeated version and so one cannot claim that collusion must necessarily be achieved. When it is achieved, prices and output levels are stable in the repeated versions of the Cournot and the Bertrand games. This is not the case when demand is uncertain and firms cannot be sure whether cheating has occurred. In regime switching models firms alternate between collusive periods and punishment periods. Price wars start when (stochastic) demand falls below a trigger level.

We have only considered pure or homogenous good oligopolies but real life products are often differentiated. Fortunately, differentiated oligopolies can be analysed in essentially the same way as pure oligopolies.[17] The main difference is that, when products are differentiated, each products has its own demand curve whereas in pure oligopoly it is assumed that price depends on total industry output. As a result, prices charged by firms in a differentiated oligopoly are not identical. Demand for a good produced by one oligopolist depends on the prices of the other products as well as its own price. It can be argued that collusion is less likely in differentiated oligopoly because of the larger informational requirements and the difficulty of monitoring. When a manufacturer of a substitute product lowers his price it is not known whether this is cheating or whether it reflects changed cost or demand conditions. Given that in a differentiated oligopoly technologies are less likely to be identical, and even at the collusive outcome prices differ, monitoring is an awesome task.

In the models we have discussed, firms have complete information. In practice it is unlikely that firms have exact information about each other's costs. When firms have private information they can strategically reveal this information.[18] In experiments when information about competitors is limited and no communication is allowed, prices tend to the Bertrand solution and there is some evidence in favour of the Cournot model when subjects are asked to decide on quantity. The collusive outcome becomes likely when there are few experienced players who have perfect information about each other's costs and actions. [19]

[17] *See for example Gravelle and Rees (2004) Chapter 16*

[18] *See for example Tirole (1988) Chapter 6*

[19] *See Geroski, Phlips and Ulph (1985)*

Chapter summary

After reading this chapter and the relevant reading, you should understand:

- the differences in **assumptions** with respect to decision variables and strategic (a)symmetry between the various oligopoly models discussed in this chapter
- why there is a **temptation to cheat** in a cartel
- the nature of the Nash equilibrium in the Bertrand model
- the concept of a **profit possibility frontier**
- why the **division of profit** can be problematic in a cartel
- how **MFC** clauses can facilitate collusion
- the nature of **regime-switching models**, the role of uncertainty in these models and their predictions about price wars.

You should be able to:

- solve Stackelberg, Cournot, Bertrand and collusion models
- solve a **dominant firm model** analytically and graphically
- show why in a **repeated Bertrand model** cheating is more likely when there are many firms and firms are impatient.

Sample exercises

1. Let the inverse demand curve for a homogenous product be $p = 70 - Q$. Suppose there are two firms in the industry, each with constant marginal costs of 10. Assuming they behave as Cournot duopolists, what will be the price and output? What price and output levels does the Stackelberg model predict? What price and output levels does the Bertrand model predict? What is the collusive price and output level? Now assume the firms interact each period and the time horizon is infinite. For which values of the discount factor is there a collusive equilibrium sustained by trigger strategies if firms play Cournot forever after cheating has occurred? What if they play Bertrand forever after cheating?
2. An industry consists of two firms, each of which produce output at a constant unit cost of 10 per unit. The demand function for the industry is $Q = 1{,}000{,}000/p$. Find the Cournot and Stackelberg equilibrium price and quantity.
3. PowerGen and National Power are the only two electricity generating companies in the UK. The demand for electricity is $p = 580 - 3q$. The total cost function of PowerGen is $TC_P = 410q_P$ and the total cost function of National Power is $TC_N = 460q_N$.
    a. If these two firms collude to maximise their combined profits, how much will each firm produce?
    b. How will they divide profits?
4. An industry consists of one dominant firm and five small (fringe) firms which behave competitively, taking the price set by the dominant firm as given. Industry demand is $Q = 10 - p$. The fringe firms have identical cost functions $C(q) = q^2$ and the dominant firm has constant marginal cost of 1 per unit. Derive the dominant firm's residual demand and corresponding marginal revenue. Illustrate all your derivations on a graph. What price will the dominant firm set? At this price, how much does it supply and how much is supplied by the fringe firms?
5. Demand is given by $p = 1 - Q$ and production costs are zero. Find Cournot, Bertrand and Stackelberg price, output and profit levels. Suppose firm 2 (which is the follower in the Stackelberg game) has a fixed cost F. How does this change your answers?