11 Limited dependent variable models

Learning Outcomes

In this chapter, you will learn how to

- Compare between different types of limited dependent variables and select the appropriate model
- Interpret and evaluate logit and probit models
- Distinguish between the binomial and multinomial cases
- Deal appropriately with censored and truncated dependent variables
- Estimate limited dependent variable models using maximum likelihood in EViews

11.1 Introduction and motivation

Chapters 4 and 9 have shown various uses of dummy variables to numerically capture the information qualitative variables – for example, day-of-the-week effects, gender, credit ratings, etc. When a dummy is used as an explanatory variable in a regression model, this usually does not give rise to any particular problems (so long as one is careful to avoid the *dummy variable trap* – see chapter 9). However, there are many situations in financial research where it is the explained variable, rather than one or more of the explanatory variables, that is qualitative. The qualitative information would then be coded as a dummy variable and the situation would be referred to as a *limited dependent variable* and needs to be treated differently. The term refers to any problem where the values that the dependent variables may take are limited to certain integers (e.g. 0, 1, 2, 3, 4) or even where it is a binary number (only 0 or 1). There are numerous examples of instances where this may arise, for example where we want to model:

- Why firms choose to list their shares on the NASDAQ rather than the NYSE
- Why some stocks pay dividends while others do not
- What factors affect whether countries default on their sovereign debt
- Why some firms choose to issue new stock to finance an expansion while others issue bonds
- Why some firms choose to engage in stock splits while others do not.

It is fairly easy to see in all these cases that the appropriate form for the dependent variable would be a 0–1 dummy variable since there are only two possible outcomes. There are, of course, also situations where it would be more useful to allow the dependent variable to take on other values, but these will be considered later in section 11.9. We will first examine a simple and obvious, but unfortunately flawed, method for dealing with binary dependent variables, known as the *linear probability model*.

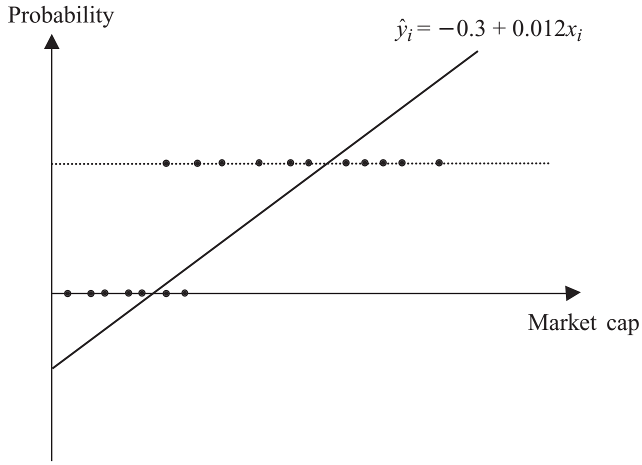
11.2 The linear probability model

The linear probability model (LPM) is by far the simplest way of dealing with binary dependent variables, and it is based on an assumption that the probability of an event occurring, P_i, is linearly related to a set of explanatory variables $x_{2i}, x_{3i}, \ldots, x_{ki}$

$$P_i = p(y_i = 1) = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki} + u_i, \quad i = 1, \ldots, N \tag{11.1}$$

The actual probabilities cannot be observed, so we would estimate a model where the outcomes, y_i (the series of zeros and ones), would be the dependent variable. This is then a linear regression model and would be estimated by OLS. The set of explanatory variables could include either quantitative variables or dummies or both. The fitted values from this regression are the estimated probabilities for $y_i = 1$ for each observation i. The slope estimates for the linear probability model can be interpreted as the change in the probability that the dependent variable will equal 1 for a one-unit change in a given explanatory variable, holding the effect of all other explanatory variables fixed. Suppose, for example, that we wanted to model the probability that a firm i will pay a dividend ($y_i = 1$) as a function of its market capitalisation (x_{2i}, measured in millions of US

Figure 11.1
The fatal flaw of the linear probability model



dollars), and we fit the following line:

$$\hat{P}_i = -0.3 + 0.012x_{2i} \tag{11.2}$$

where $\hat{P}_i$ denotes the fitted or estimated probability for firm i. This model suggests that for every \$1m increase in size, the probability that the firm will pay a dividend increases by 0.012 (or 1.2%). A firm whose stock is valued at \$50m will have a $-0.3 + 0.012 \times 50 = 0.3$ (or 30%) probability of making a dividend payment. Graphically, this situation may be represented as in figure 11.1.

While the linear probability model is simple to estimate and intuitive to interpret, the diagram should immediately signal a problem with this setup. For any firm whose value is less than \$25m, the model-predicted probability of dividend payment is negative, while for any firm worth more than \$88m, the probability is greater than one. Clearly, such predictions cannot be allowed to stand, since the probabilities should lie within the range (0,1). An obvious solution is to truncate the probabilities at 0 or 1, so that a probability of -0.3, say, would be set to zero, and a probability of, say, 1.2 would be set to 1. However, there are at least two reasons why this is still not adequate:

(1) The process of truncation will result in too many observations for which the estimated probabilities are exactly zero or one.
(2) More importantly, it is simply not plausible to suggest that the firm's probability of paying a dividend is either exactly zero or exactly one. Are we really certain that very small firms will definitely never pay a dividend and that large firms will always make a payout? Probably not, so a different kind of model is usually used for binary dependent

variables – either a *logit* or a *probit* specification. These approaches will be discussed in the following sections. But before moving on, it is worth noting that the LPM also suffers from a couple of more standard econometric problems that we have examined in previous chapters. First, since the dependent variable takes only one or two values, for given (fixed in repeated samples) values of the explanatory variables, the disturbance term[1] will also take on only one of two values. Consider again equation (11.1). If $y_i = 1$, then by definition

$$u_i = 1 - \beta_1 - \beta_2 x_{2i} - \beta_3 x_{3i} - \cdots - \beta_k x_{ki};$$

but if $y_i = 0$, then

$$u_i = -\beta_1 - \beta_2 x_{2i} - \beta_3 x_{3i} - \cdots - \beta_k x_{ki}.$$

Hence the error term cannot plausibly be assumed to be normally distributed. Since u_i changes systematically with the explanatory variables, the disturbances will also be heteroscedastic. It is therefore essential that heteroscedasticity-robust standard errors are always used in the context of limited dependent variable models.

11.3 The logit model

Both the logit and probit model approaches are able to overcome the limitation of the LPM that it can produce estimated probabilities that are negative or greater than one. They do this by using a function that effectively transforms the regression model so that the fitted values are bounded within the (0,1) interval. Visually, the fitted regression model will appear as an S-shape rather than a straight line, as was the case for the LPM. This is shown in figure 11.2.

The logistic function F, which is a function of any random variable, z, would be

$$F(z_i) = \frac{e^{z_i}}{1 + e^{z_i}} = \frac{1}{1 + e^{-z_i}} \tag{11.3}$$

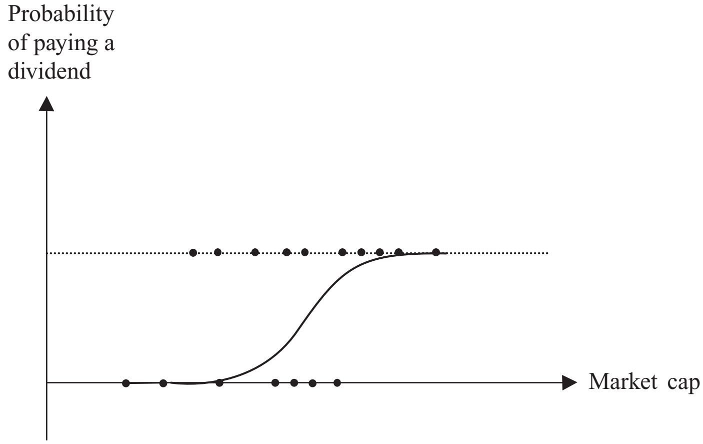
where e is the exponential under the logit approach. The model is so called because the function F is in fact the cumulative logistic distribution. So the logistic model estimated would be

$$P_i = \frac{1}{1 + e^{-(\beta_1 + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i)}} \tag{11.4}$$

where again P_i is the probability that $y_i = 1$.

[1] N.B. The discussion refers to the disturbance, u_i, rather than the residual, $\hat{u}_i$.

Figure 11.2
The logit model



With the logistic model, 0 and 1 are asymptotes to the function and thus the probabilities will never actually fall to exactly zero or rise to one, although they may come infinitesimally close. In equation (11.3), as z_i tends to infinity, e^{-z_i} tends to zero and $1/(1+e^{-z_i})$ tends to 1; as z_i tends to minus infinity, e^{-z_i} tends to infinity and $1/(1+e^{-z_i})$ tends to 0.

Clearly, this model is not linear (and cannot be made linear by a transformation) and thus is not estimable using OLS. Instead, maximum likelihood is usually used – this is discussed in section 11.7 and in more detail in the appendix to this chapter.

11.4 Using a logit to test the pecking order hypothesis

This section examines a study of the pecking order hypothesis due to Helwege and Liang (1996). The theory of firm financing suggests that corporations should use the cheapest methods of financing their activities first (i.e. the sources of funds that require payment of the lowest rates of return to investors) and switch to more expensive methods only when the cheaper sources have been exhausted. This is known as the 'pecking order hypothesis', initially proposed by Myers (1984). Differences in the relative cost of the various sources of funds are argued to arise largely from information asymmetries since the firm's senior managers will know the true riskiness of the business, whereas potential outside investors will not.[2] Hence, all else equal, firms will prefer internal finance and then, if

[2] 'Managers have private information regarding the value of assets in place and investment opportunities that cannot credibly be conveyed to the market. Consequently, any risky security offered by the firm will not be priced fairly from the manager's point of view' (Helwege and Liang, p. 438).

further (external) funding is necessary, the firm's riskiness will determine the type of funding sought. The more risky the firm is perceived to be, the less accurate will be the pricing of its securities.

Helwege and Liang (1996) examine the pecking order hypothesis in the context of a set of US firms that had been newly listed on the stock market in 1983, with their additional funding decisions being tracked over the 1984–1992 period. Such newly listed firms are argued to experience higher rates of growth, and are more likely to require additional external funding than firms which have been stock market listed for many years. They are also more likely to exhibit information asymmetries due to their lack of a track record. The list of initial public offerings (IPOs) came from the Securities Data Corporation and the Securities and Exchange Commission with data obtained from Compustat.

A core objective of the paper is to determine the factors that affect the probability of raising external financing. As such, the dependent variable will be binary – that is, a column of 1s (firm raises funds externally) and 0s (firm does not raise any external funds). Thus OLS would not be appropriate and hence a logit model is used. The explanatory variables are a set that aims to capture the relative degree of information asymmetry and degree of riskiness of the firm. If the pecking order hypothesis is supported by the data, then firms should be more likely to raise external funding the less internal cash they hold. Hence variable 'deficit' measures (capital expenditures + acquisitions + dividends − earnings). 'Positive deficit' is a variable identical to deficit but with any negative deficits (i.e. surpluses) set to zero; 'surplus' is equal to the negative of deficit for firms where deficit is negative; 'positive deficit × operating income' is an interaction term where the two variables are multiplied together to capture cases where firms have strong investment opportunities but limited access to internal funds; 'assets' is used as a measure of firm size; 'industry asset growth' is the average rate of growth of assets in that firm's industry over the 1983–1992 period; 'firm's growth of sales' is the growth rate of sales averaged over the previous 5 years; 'previous financing' is a dummy variable equal to 1 for firms that obtained external financing in the previous year. The results from the logit regression are presented in table 11.1.

The key variable, 'deficit,' has a parameter that is not statistically significant and hence the probability of obtaining external financing does not depend on the size of a firm's cash deficit.[3] The parameter on the 'surplus'

[3] Or an alternative explanation, as with a similar result in the context of a standard regression model, is that the probability varies widely across firms with the size of the cash deficit so that the standard errors are large relative to the point estimate.

Table 11.1 Logit estimation of the probability of external financing

Variable	(1)	(2)	(3)
Intercept	−0.29	−0.72	−0.15
	(−3.42)	(−7.05)	(−1.58)
Deficit	0.04	0.02	
	(0.34)	(0.18)	
Positive deficit			−0.24
			(−1.19)
Surplus			−2.06
			(−3.23)
Positive deficit × operating income			−0.03
			(−0.59)
Assets	0.0004	0.0003	0.0004
	(1.99)	(1.36)	(1.99)
Industry asset growth	−0.002	−0.002	−0.002
	(−1.70)	(−1.35)	(−1.69)
Previous financing		0.79	
		(8.48)	

Note: a blank cell implies that the particular variable was not included in that regression; t-ratios in parentheses; only figures for all years in the sample are presented.
Source: Helwege and Liang (1996). Reprinted with the permission of Elsevier Science.

variable has the correct negative sign, indicating that the larger a firm's surplus, the less likely it is to seek external financing, which provides some limited support for the pecking order hypothesis. Larger firms (with larger total assets) are more likely to use the capital markets, as are firms that have already obtained external financing during the previous year.

11.5 The probit model

Instead of using the cumulative logistic function to transform the model, the cumulative normal distribution is sometimes used instead. This gives rise to the probit model. The function F in equation (11.3) is replaced by:

$$F(z_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{z_i^2}{\sigma}\right)} \tag{11.5}$$

This function is the cumulative distribution function for a standard normally distributed random variable. As for the logistic approach, this function provides a transformation to ensure that the fitted probabilities will lie between zero and one. Also as for the logit model, the marginal impact of a unit change in an explanatory variable, x_{4i} say, will be given by $\beta_4 F(z_i)$, where β_4 is the parameter attached to x_{4i} and $z_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + u_i$.

11.6 Choosing between the logit and probit models

For the majority of the applications, the logit and probit models will give very similar characterisations of the data because the densities are very similar. That is, the fitted regression plots (such as figure 11.2) will be virtually indistinguishable and the implied relationships between the explanatory variables and the probability that $y_i = 1$ will also be very similar. Both approaches are much preferred to the linear probability model. The only instance where the models may give non-negligibility different results occurs when the split of the y_i between 0 and 1 is very unbalanced – for example, when $y_i = 1$ occurs only 10% of the time.

Stock and Watson (2006) suggest that the logistic approach was traditionally preferred since the function does not require the evaluation of an integral and thus the model parameters could be estimated faster. However, this argument is no longer relevant given the computational speeds now achievable and the choice of one specification rather than the other is now usually arbitrary.

11.7 Estimation of limited dependent variable models

Given that both logit and probit are non-linear models, they cannot be estimated by OLS. While the parameters could, in principle, be estimated using non-linear least squares (NLS), maximum likelihood (ML) is simpler and is invariably used in practice. As discussed in chapter 8, the principle is that the parameters are chosen to jointly maximise a log-likelihood function (LLF). The form of this LLF will depend upon whether the logit or probit model is used, but the general principles for parameter estimation described in chapter 8 will still apply. That is, we form the appropriate log-likelihood function and then the software package will find the values of the parameters that jointly maximise it using an iterative search procedure. A derivation of the ML estimator for logit and probit models is given in the appendix to this chapter. Box 11.1 shows how to interpret the estimated parameters from probit and logit models.

Once the model parameters have been estimated, standard errors can be calculated and hypothesis tests conducted. While t-test statistics are constructed in the usual way, the standard error formulae used following the ML estimation are valid asymptotically only. Consequently, it is common to use the critical values from a normal distribution rather than a t

Box 11.1 Parameter interpretation for probit and logit models

Standard errors and t-ratios will automatically be calculated by the econometric software package used, and hypothesis tests can be conducted in the usual fashion. However, interpretation of the coefficients needs slight care. It is tempting, but incorrect, to state that a 1-unit increase in x_{2i}, for example, causes a β_2% increase in the probability that the outcome corresponding to $y_i = 1$ will be realised. This would have been the correct interpretation for the linear probability model.

However, for logit models, this interpretation would be incorrect because the form of the function is not $P_i = \beta_i + \beta_2 x_i + u_i$, for example, but rather $P_i = F(x_{2i})$, where F represents the (non-linear) logistic function. To obtain the required relationship between changes in x_{2i} and P_i, we would need to differentiate F with respect to x_{2i} and it turns out that this derivative is $\beta_2 F(x_{2i})$. So in fact, a 1-unit increase in x_{2i} will cause a $\beta_2 F(x_{2i})$ increase in probability. Usually, these impacts of incremental changes in an explanatory variable are evaluated by setting each of them to their mean values. For example, suppose we have estimated the following logit model with 3 explanatory variables using maximum likelihood

$$\hat{P}_i = \frac{1}{1 + e^{-(0.1+0.3x_{2i}-0.6x_{3i}+0.9x_{4i})}} \tag{11.6}$$

Thus we have $\hat{\beta}_1 = 0.1$, $\hat{\beta}_2 = 0.3$, $\hat{\beta}_3 = -0.6$, $\hat{\beta}_4 = 0.9$. We now need to calculate $F(z_i)$, for which we need the means of the explanatory variables, where z_i is defined as before. Suppose that these are $\bar{x}_2 = 1.6$, $\bar{x}_3 = 0.2$, $\bar{x}_4 = 0.1$, then the estimate of $F(z_i)$ will be given by

$$\hat{P}_i = \frac{1}{1 + e^{-(0.1+0.3\times1.6-0.6\times0.2+0.9\times0.1)}} = \frac{1}{1 + e^{-0.55}} = 0.63 \tag{11.7}$$

Thus a 1-unit increase in x_2 will cause an increase in the probability that the outcome corresponding to $y_i = 1$ will occur by $0.3 \times 0.63 = 0.19$. The corresponding changes in probability for variables x_3 and x_4 are $-0.6 \times 0.63 = -0.38$ and $0.9 \times 0.63 = 0.57$, respectively. These estimates are sometimes known as the *marginal effects*.

There is also another way of interpreting discrete choice models, known as the random utility model. The idea is that we can view the value of y that is chosen by individual i (either 0 or 1) as giving that person a particular level of utility, and the choice that is made will obviously be the one that generates the highest level of utility. This interpretation is particularly useful in the situation where the person faces a choice between more than 2 possibilities as in section 11.9 below.

distribution with the implicit assumption that the sample size is sufficiently large.

11.8 Goodness of fit measures for linear dependent variable models

While it would be possible to calculate the values of the standard goodness of fit measures such as RSS, R^2 or $\bar{R}^2$ for linear dependent variable models, these cease to have any real meaning. The objective of ML is to maximise the value of the LLF, not to minimise the RSS. Moreover, R^2 and adjusted

R^2, if calculated in the usual fashion, will be misleading because the fitted values from the model can take on any value but the actual values will be only either 0 and 1. To illustrate, suppose that we are considering a situation where a bank either grants a loan ($y_i = 1$) or it refuses ($y_i = 0$). Does, say, $\hat{P}_i = 0.8$ mean the loan is offered or not? In order to answer this question, sometimes, any value of $\hat{P}_i > 0.5$ would be rounded up to one and any value <0.5 rounded down to zero. However, this approach is unlikely to work well when most of the observations on the dependent variable are one or when most are zero. In such cases, it makes more sense to use the unconditional probability that $y = 1$ (call this $\bar{y}$) as the threshold rather than 0.5. So if, for example, only 20% of the observations have $y = 1$ (so $\bar{y} = 0.2$), then we would deem the model to have correctly predicted the outcome concerning whether the bank would grant the loan to the customer where $\hat{P}_i > 0.2$ and $y_i = 1$ and where $\hat{P}_i < 0.2$ and $y_i = 0$.

Thus if $y_i = 1$ and $\hat{P}_i = 0.8$, the model has effectively made the correct prediction (either the loan is granted or refused – we cannot have any outcome in between), whereas R^2 and $\bar{R}^2$ will not give it full credit for this. Two goodness of fit measures that are commonly reported for limited dependent variable models are as follows.

(1) The percentage of y_i values correctly predicted, defined as 100 $\times$ the number of observations predicted correctly divided by the total number of observations:

$$\text{Percent correct predictions} = \frac{100}{N} \sum_{i=1}^{N} y_i I(\hat{P}_i) + (1 - y_i)(1 - I(\hat{P}_i)) \tag{11.8}$$

where $I(\hat{y}_i) = 1$ if $\hat{y}_i > \bar{y}$ and 0 otherwise.

Obviously, the higher this number, the better the fit of the model. Although this measure is intuitive and easy to calculate, Kennedy (2003) suggests that it is not ideal, since it is possible that a 'naïve predictor' could do better than any model if the sample is unbalanced between 0 and 1. For example, suppose that $y_i = 1$ for 80% of the observations. A simple rule that the prediction is always 1 is likely to outperform any more complex model on this measure but is unlikely to be very useful. Kennedy (2003, p. 267) suggests measuring goodness of fit as the percentage of $y_i = 1$ correctly predicted plus the percentage of $y_i = 0$ correctly predicted. Algebraically, this can be calculated as

$$\text{Percent correct predictions} = 100 \times \left[\frac{\sum y_i I(\hat{P}_i)}{\sum y_i} + \frac{\sum (1 - y_i)(1 - I(\hat{P}_i))}{N - \sum y_i} \right] \tag{11.9}$$

Again, the higher the value of the measure, the better the fit of the model.

(2) A measure known as 'pseudo-R^2', defined as

$$\text{pseudo} - R^2 = 1 - \frac{LLF}{LLF_0} \tag{11.10}$$

where LLF is the maximised value of the log-likelihood function for the logit and probit model and LLF_0 is the value of the log-likelihood function for a restricted model where all of the slope parameters are set to zero (i.e. the model contains only an intercept). Pseudo-R^2 will have a value of zero for the restricted model, as with the traditional R^2, but this is where the similarity ends. Since the likelihood is essentially a joint probability, its value must be between zero and one, and therefore taking its logarithm to form the LLF must result in a negative number. Thus, as the model fit improves, LLF will become less negative and therefore pseudo-R^2 will rise. The maximum value of one could be reached only if the model fitted perfectly (i.e. all the $\hat{P}_i$ were either exactly zero or one corresponding to the actual values). This could never occur in reality and therefore pseudo-R^2 has a maximum value less than one. We also lose the simple interpretation of the standard R^2 that it measures the proportion of variation in the dependent variable that is explained by the model. Indeed, pseudo-R^2 does not have any intuitive interpretation.

This definition of pseudo-R^2 is also known as McFadden's R^2, but it is also possible to specify the metric in other ways. For example, we could define pseudo-R^2 as $[1 - (RSS/TSS)]$ where RSS is the residual sum of squares from the fitted model and TSS is the total sum of squares of y_i.

11.9 Multinomial linear dependent variables

All of the examples that have been considered so far in this chapter have concerned situations where the dependent variable is modelled as a binary (0,1) choice. But there are also many instances where investors or financial agents are faced with more alternatives. For example, a company may be considering listing on the NYSE, the NASDAQ or the AMEX markets; a firm that is intending to take over another may choose to pay by cash, with shares, or with a mixture of both; a retail investor may be choosing between five different mutual funds; a credit ratings agency could assign 1 of 16 (AAA to B3/B−) different ratings classifications to a firm's debt.

Notice that the first three of these examples are different from the last one. In the first three cases, there is no natural ordering of the alternatives: the choice is simply made between them. In the final case, there is an obvious ordering, because a score of 1, denoting a AAA-rated bond, is better than a score of 2, denoting a AA1/AA+-rated bond, and so on (see section 4.14 in chapter 4). These two situations need to be distinguished and a different approach used in each case. In the first (when there is no natural ordering), a multinomial logit or probit would be used, while in the second (where there is an ordering), an ordered logit or probit would be used. This latter situation will be discussed in the next section, while multinomial models will be considered now.

When the alternatives are unordered, this is sometimes called a *discrete choice* or *multiple choice* problem. The models used are derived from the principles of utility maximisation – that is, the agent chooses the alternative that maximises his utility relative to the others. Econometrically, this is captured using a simple generalisation of the binary setup discussed earlier. When there were only 2 choices (0,1), we required just one equation to capture the probability that one or the other would be chosen. If there are now three alternatives, we would need two equations; for four alternatives, we would need three equations. In general, if there are m possible alternative choices, we need $m - 1$ equations.

The situation is best illustrated by first examining a multinomial linear probability model. This still, of course, suffers from the same limitations as it did in the binary case (i.e. the same problems as the LPM), but it nonetheless serves as a simple example by way of introduction.[4] The multiple choice example most commonly used is that of the selection of the mode of transport for travel to work.[5] Suppose that the journey may be made by car, bus, or bicycle (3 alternatives), and suppose that the explanatory variables are the person's income (I), total hours worked (H), their gender (G) and the distance travelled (D).[6] We could set up 2 equations

$$BUS_i = \alpha_1 + \alpha_2 I_i + \alpha_3 H_i + \alpha_4 G_i + \alpha_5 D_i + u_i \tag{11.11}$$
$$CAR_i = \beta_1 + \beta_2 I_i + \beta_3 H_i + \beta_4 G_i + \beta_5 D_i + v_i \tag{11.12}$$

where $BUS_i = 1$ if person i travels by bus and 0 otherwise; $CAR_i = 1$ if person i travels by car and 0 otherwise.

[4] Multinomial models are clearly explained with intuitive examples in Halcoussis (2005, chapter 12).
[5] This illustration is used in Greene (2002) and Kennedy (2003), for example.
[6] Note that the same variables must be used for all equations for this approach to be valid.

There is no equation for travel by bicycle and this becomes a sort of reference point, since if the dependent variables in the two equations are both zero, the person must be travelling by bicycle.[7] In fact, we do not need to estimate the third equation (for travel by bicycle) since any quantity of interest can be inferred from the other two. The fitted values from the equations can be interpreted as probabilities and so, together with the third possibility, they must sum to unity. Thus, if, for a particular individual i, the probability of travelling by car is 0.4 and by bus is 0.3, then the possibility that she will travel by bicycle must be 0.3 ($1-0.4-0.3$). Also, the intercepts for the three equations (the two estimated equations plus the missing one) must sum to zero across the three modes of transport.

While the fitted probabilities will always sum to unity by construction, as with the binomial case, there is no guarantee that they will all lie between 0 and 1 – it is possible that one or more will be greater than 1 and one or more will be negative. In order to make a prediction about which mode of transport a particular individual will use, given that the parameters in equations (11.11) and (11.12) have been estimated and given the values of the explanatory variables for that individual, the largest fitted probability would be set to 1 and the others set to 0. So, for example, if the estimated probabilities of a particular individual travelling by car, bus and bicycle are 1.1, 0.2 and -0.3, these probabilities would be rounded to 1, 0, and 0. So the model would predict that this person would travel to work by car.

Exactly as the LPM has some important limitations that make logit and probit the preferred models, in the multiple choice context multinomial logit and probit models should be used. These are direct generalisations of the binary cases, and as with the multinomial LPM, $m-1$ equations must be estimated where there are m possible outcomes or choices. The outcome for which an equation is not estimated then becomes the reference choice, and thus the parameter estimates must be interpreted slightly differently. Suppose that travel by bus (B) or by car (C) have utilities for person i that depend on the characteristics described above (I_i, H_i, G_i, D_i), then the car will be chosen if

$$\begin{aligned}&(\beta_1+\beta_2 I_i+\beta_3 H_i+\beta_4 G_i+\beta_5 D_i+v_i)\\&\quad>(\alpha_1+\alpha_2 I_i+\alpha_3 H_i+\alpha_4 G_i+\alpha_5 D_i+u_i)\end{aligned}\tag{11.13}$$

That is, the probability that the car will be chosen will be greater than that of the bus being chosen if the utility from going by car is greater.

[7] We are assuming that the choices are exhaustive and mutually exclusive – that is, one and only one method of transport can be chosen!

Equation (11.13) can be rewritten as

$$\begin{aligned}(\beta_1-\alpha_1)+(\beta_2-\alpha_2)\,I_i+(\beta_3-\alpha_3)\,H_i\\+(\beta_4-\alpha_4)\,G_i+(\beta_5-\alpha_5)\,D_i>(u_i-v_i)\end{aligned} \tag{11.14}$$

If it is assumed that u_i and v_i independently follow a particular distribution,[8] then the difference between them will follow a logistic distribution. Thus we can write

$$P(C_i/B_i)=\frac{1}{1+e^{-z_i}} \tag{11.15}$$

where z_i is the function on the left hand side of (11.14), i.e. $(\beta_1-\alpha_1)+(\beta_2-\alpha_2)\,I_i+\cdots$ and travel by bus becomes the reference category. $P(C_i/B_i)$ denotes the probability that individual i would choose to travel by car rather than by bus.

Equation (11.15) implies that the probability of the car being chosen in preference to the bus depends upon the logistic function of the differences in the parameters describing the relationship between the utilities from travelling by each mode of transport. Of course, we cannot recover both β_2 and α_2 for example, but only the difference between them (call this $\gamma_2=\beta_2-\alpha_2$). These parameters measure the impact of marginal changes in the explanatory variables on the probability of travelling by car relative to the probability of travelling by bus. Note that a unit increase in I_i will lead to a $\gamma_2 F(I_i)$ increase in the probability and not a γ_2 increase – see equations (11.5) and (11.6) above. For this trinomial problem, there would need to be another equation – for example, based on the difference in utilities between travelling by bike and by bus. These two equations would be estimated simultaneously using maximum likelihood.

For the multinomial logit model, the error terms in the equations (u_i and v_i in the example above) must be assumed to be independent. However, this creates a problem whenever two or more of the choices are very similar to one another. This problem is known as the 'independence of irrelevant alternatives'. To illustrate how this works, Kennedy (2003, p. 270) uses an example where another choice to travel by bus is introduced and the only thing that differs is the colour of the bus. Suppose that the original probabilities for the car, bus and bicycle were 0.4, 0.3 and 0.3. If a new green bus were introduced in addition to the existing red bus, we would expect that the overall probability of travelling by bus should stay at 0.3 and that bus passengers should split between the two (say, with half using each coloured bus). This result arises since the new colour of the bus is

[8] In fact, they must follow independent log Weibull distributions.

irrelevant to those who have already chosen to travel by car or bicycle. Unfortunately, the logit model will not be able to capture this and will seek to preserve the relative probabilities of the old choices (which could be expressed as $\frac{4}{10}$, $\frac{3}{10}$ and $\frac{3}{10}$ respectively). These will become $\frac{4}{13}$, $\frac{3}{13}$, $\frac{3}{13}$ and $\frac{3}{13}$ for car, green bus, red bus and bicycle respectively – a long way from what intuition would lead us to expect.

Fortunately, the multinomial probit model, which is the multiple choice generalisation of the probit model discussed in section 11.5 above, can handle this. The multinomial probit model would be set up in exactly the same fashion as the multinomial logit model, except that the cumulative normal distribution is used for $(u_i - v_i)$ instead of a cumulative logistic distribution. This is based on an assumption that u_i and v_i are multivariate normally distributed but unlike the logit model, they can be correlated. A positive correlation between the error terms can be employed to reflect a similarity in the characteristics of two or more choices. However, such a correlation between the error terms makes estimation of the multinomial probit model using maximum likelihood difficult because multiple integrals must be evaluated. Kennedy (2003, p. 271) suggests that this has resulted in continued use of the multinomial logit approach despite the independence of irrelevant alternatives problem.

11.10 The pecking order hypothesis revisited – the choice between financing methods

In section 11.4, a logit model was used to evaluate whether there was empirical support for the pecking order hypothesis where the hypothesis boiled down to a consideration of the probability that a firm would seek external financing or not. But suppose that we wish to examine not only whether a firm decides to issue external funds but also which method of funding it chooses when there are a number of alternatives available. As discussed above, the pecking order hypothesis suggests that the least costly methods, which, everything else equal, will arise where there is least information asymmetry, will be used first, and the method used will also depend on the riskiness of the firm. Returning to Helwege and Liang's study, they argue that if the pecking order is followed, low-risk firms will issue public debt first, while moderately risky firms will issue private debt and the most risky companies will issue equity. Since there is more than one possible choice, this is a multiple choice problem and consequently, a binary logit model is inappropriate and instead, a multinomial logit is used. There are three possible choices here: bond issue, equity issue and private debt issue. As is always the case for multinomial models, we estimate equations for one fewer than the number of possibilities, and so equations are estimated for equities and bonds, but not for private debt. This choice then becomes the reference point, so that the coefficients measure the probability of issuing equity or bonds rather than private debt, and a positive parameter estimate in, say, the equities equation implies that an increase in the value of the variable leads to an increase in the probability that the firm will choose to issue equity rather than private debt.

The set of explanatory variables is slightly different now given the different nature of the problem at hand. The key variable measuring risk is now the 'unlevered Z score', which is Altman's Z score constructed as a weighted average of operating earnings before interest and taxes, sales, retained earnings and working capital. All other variable names are largely self-explanatory and so are not discussed in detail, but they are divided into two categories – those measuring the firm's level of risk (unlevered Z-score, debt, interest expense and variance of earnings) and those measuring the degree of information asymmetry (R&D expenditure, venture-backed, age, age over 50, plant property and equipment, industry growth, non-financial equity issuance, and assets). Firms with heavy R&D expenditure, those receiving venture capital financing, younger firms, firms with less property, plant and equipment, and smaller firms are argued to suffer from greater information asymmetry. The parameter estimates for the multinomial logit are presented in table 11.2, with equity issuance as a (0,1) dependent variable in the second column and bond issuance as a (0,1) dependent variable in the third column.

Overall, the results paint a very mixed picture about whether the pecking order hypothesis is validated or not. The positive (significant) and negative (insignificant) estimates on the unlevered Z-score and interest expense variables respectively suggest that firms in good financial health (i.e. less risky firms) are more likely to issue equities or bonds rather than private debt. Yet the positive sign of the parameter on the debt variable is suggestive that riskier firms are more likely to issue equities or bonds; the variance of earnings variable has the wrong sign but is not statistically significant. Almost all of the asymmetric information variables have statistically insignificant parameters. The only exceptions are that firms having venture backing are more likely to seek capital market financing of either type, as are non-financial firms. Finally, larger firms are more likely to issue bonds (but not equity). Thus the authors conclude that the results 'do not indicate that firms strongly avoid external financing as the pecking order predicts' and 'equity is not the least desirable source of financing since it appears to dominate bank loans' (Helwege and Liang (1996), p. 458).

Table 11.2 Multinomial logit estimation of the type of external financing

Variable	Equity equation	Bonds equation
Intercept	−4.67	−4.68
	(−6.17)	(−5.48)
Unlevered *Z*-score	0.14	0.26
	(1.84)	(2.86)
Debt	1.72	3.28
	(1.60)	(2.88)
Interest expense	−9.41	−4.54
	(−0.93)	(−0.42)
Variance of earnings	−0.04	−0.14
	(−0.55)	(−1.56)
R&D	0.61	0.89
	(1.28)	(1.59)
Venture-backed	0.70	0.86
	(2.32)	(2.50)
Age	−0.01	−0.03
	(−1.10)	(−1.85)
Age over 50	1.58	1.93
	(1.44)	(1.70)
Plant, property and equipment	(0.62)	0.34
	(0.94)	(0.50)
Industry growth	0.005	0.003
	(1.14)	(0.70)
Non-financial equity issuance	0.008	0.005
	(3.89)	(2.65)
Assets	−0.001	0.002
	(−0.59)	(4.11)

Notes: *t*-ratios in parentheses; only figures for all years in the sample are presented.
Source: Helwege and Liang (1996). Reprinted with the permission of Elsevier Science.

11.11 Ordered response linear dependent variables models

Some limited dependent variables can be assigned numerical values that have a natural ordering. The most common example in finance is that of credit ratings, as discussed previously, but a further application is to modelling a security's bid–ask spread (see, for example, ap Gwilym *et al.*, 1998). In such cases, it would not be appropriate to use multinomial logit or probit since these techniques cannot take into account any ordering in the dependent variables. Notice that ordinal variables are still distinct from the usual type of data that were employed in the early chapters in this book, such as stock returns, GDP, interest rates, etc. These are examples of cardinal numbers, since additional information can be inferred from their actual values relative to one another. To illustrate, an increase in house prices of 20% represents twice as much growth as a 10% rise. The same is not true of ordinal numbers, where (returning to the credit ratings example) a rating of AAA, assigned a numerical score of 16, is not ‘twice as good’ as a rating of Baa2/BBB, assigned a numerical score of 8. Similarly, for ordinal data, the difference between a score of, say, 15 and of 16 cannot be assumed to be equivalent to the difference between the scores of 8 and 9. All we can say is that as the score increases, there is a monotonic increase in the credit quality. Since only the ordering can be interpreted with such data and not the actual numerical values, OLS cannot be employed and a technique based on ML is used instead. The models used are generalisations of logit and probit, known as *ordered logit* and *ordered probit*.

Using the credit rating example again, the model is set up so that a particular bond falls in the AA+ category (using Standard and Poor’s terminology) if its unobserved (latent) creditworthiness falls within a certain range that is too low to classify it as AAA and too high to classify it as AA. The boundary values between each rating are then estimated along with the model parameters.

11.12 Are unsolicited credit ratings biased downwards? An ordered probit analysis

Modelling the determinants of credit ratings is one of the most important uses of ordered probit and logit models in finance. The main credit ratings agencies construct what may be termed *solicited* ratings, which are those where the issuer of the debt contacts the agency and pays them a fee for producing the rating. Many firms globally do not seek a rating (because, for example, the firm believes that the ratings agencies are not well placed to evaluate the riskiness of debt in their country or because they do not plan to issue any debt or because they believe that they would be awarded a low rating), but the agency may produce a rating anyway. Such ‘unwarranted and unwelcome’ ratings are known as *unsolicited* ratings. All of the major ratings agencies produce unsolicited ratings as well as solicited ones, and they argue that there is a market demand for this information even if the issuer would prefer not to be rated.

Companies in receipt of unsolicited ratings argue that these are biased downwards relative to solicited ratings and that they cannot be justified without the level of detail of information that can be provided only by the rated company itself. A study by Poon (2003) seeks to test the conjecture that unsolicited ratings are biased after controlling for the rated company's characteristics that pertain to its risk.

The data employed comprise a pooled sample of all companies that appeared on the annual 'issuer list' of S&P during the years 1998–2000. This list contains both solicited and unsolicited ratings covering 295 firms over 15 countries and totalling 595 observations. In a preliminary exploratory analysis of the data, Poon finds that around half of the sample ratings were unsolicited, and indeed the unsolicited ratings in the sample are on average significantly lower than the solicited ratings.[9] As expected, the financial characteristics of the firms with unsolicited ratings are significantly weaker than those for firms that requested ratings. The core methodology employs an ordered probit model with explanatory variables comprising firm characteristics and a dummy variable for whether the firm's credit rating was solicited or not

$$R_i^* = X_i\beta + \epsilon_i \tag{11.16}$$

with

$$R_i = \begin{cases} 1 & \text{if } R_i^* \le \mu_0 \\ 2 & \text{if } \mu_0 < R_i^* \le \mu_1 \\ 3 & \text{if } \mu_1 < R_i^* \le \mu_2 \\ 4 & \text{if } \mu_2 < R_i^* \le \mu_3 \\ 5 & \text{if } R_i^* > \mu_3 \end{cases}$$

where R_i are the observed ratings scores that are given numerical values as follows: AA or above = 6, A = 5, BBB = 4, BB = 3, B = 2 and CCC or below = 1; R_i^* is the unobservable 'true rating' (or 'an unobserved continuous variable representing S&P's assessment of the creditworthiness of issuer i'), X_i is a vector of variables that explains the variation in ratings; β is a vector of coefficients; μ_i are the threshold parameters to be estimated along with β; and ϵ_i is a disturbance term that is assumed normally distributed.

The explanatory variables attempt to capture the creditworthiness using publicly available information. Two specifications are estimated: the first includes the variables listed below, while the second additionally

[9] We are assuming here that the broader credit rating categories, of which there are 6, (AAA, AA, A, BBB, BB, B) are being used rather than the finer categories used by Cantor and Packer (1996).

incorporates an interaction of the main financial variables with a dummy variable for whether the firm's rating was solicited (SOL) and separately with a dummy for whether the firm is based in Japan.[10] The financial variables are ICOV – interest coverage (i.e. earnings interest), ROA – return on assets, DTC – total debt to capital, and SDTD – short-term debt to total debt. Three variables – SOVAA, SOVA and SOVBBB – are dummy variables that capture the debt issuer's sovereign credit rating.[11] Table 11.3 presents the results from the ordered probit estimation.

The key finding is that the SOL variable is positive and statistically significant in Model 1 (and it is positive but insignificant in Model 2), indicating that even after accounting for the financial characteristics of the firms, unsolicited firms receive ratings on average 0.359 units lower than an otherwise identical firm that had requested a rating. The parameter estimate for the interaction term between the solicitation and Japanese dummies (SOL$*$JP) is positive and significant in both specifications, indicating strong evidence that Japanese firms soliciting ratings receive higher scores. On average, firms with stronger financial characteristics (higher interest coverage, higher return on assets, lower debt to total capital, or a lower ratio of short-term debt to long-term debt) have higher ratings.

A major flaw that potentially exists within the above analysis is the *self-selection bias* or *sample selection bias* that may have arisen if firms that would have received lower credit ratings (because they have weak financials) elect not to solicit a rating. If the probit equation for the determinants of ratings is estimated ignoring this potential problem and it exists, the coefficients will be inconsistent. To get around this problem and to control for the sample selection bias, Heckman (1979) proposed a two-step procedure that in this case would involve first estimating a 0–1 probit model for whether the firm chooses to solicit a rating and second estimating the ordered probit model for the determinants of the rating. The first-stage probit model is

$$Y_i^* = Z_i \gamma + \xi_i \tag{11.17}$$

where $Y_i = 1$ if the firm has solicited a rating and 0 otherwise, and Y_i^* denotes the latent propensity of issuer i to solicit a rating, Z_i are the

[10] The Japanese dummy is used since a disproportionate number of firms in the sample are from this country.

[11] So SOVAA $= 1$ if the sovereign (i.e. the government of that country) has debt with a rating of AA or above and 0 otherwise; SOVA has a value 1 if the sovereign has a rating of A; and SOVBBB has a value 1 if the sovereign has a rating of BBB; any firm in a country with a sovereign whose rating is below BBB is assigned a zero value for all three sovereign rating dummies.

Table 11.3 Ordered probit model results for the determinants of credit ratings

Explanatory variables	Model 1		Model 2	
	Coefficient	Test statistic	Coefficient	Test statistic
Intercept	2.324	8.960***	1.492	3.155***
SOL	0.359	2.105**	0.391	0.647
JP	−0.548	−2.949***	1.296	2.441**
JP∗SOL	1.614	7.027***	1.487	5.183***
SOVAA	2.135	8.768***	2.470	8.975***
SOVA	0.554	2.552**	0.925	3.968***
SOVBBB	−0.416	−1.480	−0.181	−0.601
ICOV	0.023	3.466***	−0.005	−0.172
ROA	0.104	10.306***	0.194	2.503**
DTC	−1.393	−5.736***	−0.522	−1.130
SDTD	−1.212	−5.228***	0.111	0.171
SOL∗ICOV	–	–	0.005	0.163
SOL∗ROA	–	–	−0.116	−1.476
SOL∗DTC	–	–	0.756	1.136
SOL∗SDTD	–	–	−0.887	−1.290
JP∗ICOV	–	–	0.009	0.275
JP∗ROA	–	–	0.183	2.200**
JP∗DTC	–	–	−1.865	−3.214***
JP∗SDTD	–	–	−2.443	−3.437***
AA or above	>5.095		>5.578	
A	>3.788 and ≤5.095	25.278***	>4.147 and ≤5.578	23.294***
BBB	>2.550 and ≤3.788	19.671***	>2.803 and ≤4.147	19.204***
BB	>1.287 and ≤2.550	14.342***	>1.432 and ≤2.803	14.324***
B	>0 and ≤1.287	7.927***	>0 and ≤1.432	7.910***
CCC or below	≤0		≤0	

Note: *, ** and *** denote significance at the 10%, 5% and 1% levels respectively.
Source: Poon (2003). Reprinted with the permission of Elsevier Science.

variables that explain the choice to be rated or not, and γ are the parameters to be estimated. When this equation has been estimated, the rating R_i as defined above in equation (11.16) will be observed only if $Y_i = 1$. The error terms from the two equations, ϵ_i and ξ_i, follow a bivariate standard normal distribution with correlation $\rho_{\epsilon\xi}$. Table 11.4 shows the results from the two-step estimation procedure, with the estimates from the binary probit model for the decision concerning whether to solicit a rating in panel A and the determinants of ratings for rated firms in panel B.

A positive parameter value in panel A indicates that higher values of the associated variable increases the probability that a firm will elect to

Table 11.4 Two-step ordered probit model allowing for selectivity bias in the determinants of credit ratings

Explanatory variable	Coefficient	Test statistic
Panel A: Decision to be rated		
Intercept	1.624	3.935***
JP	−0.776	−4.951***
SOVAA	−0.959	−2.706***
SOVA	−0.614	−1.794*
SOVBBB	−1.130	−2.899***
ICOV	−0.005	−0.922
ROA	0.051	6.537***
DTC	0.272	1.019
SDTD	−1.651	−5.320***
Panel B: Rating determinant equation		
Intercept	1.368	2.890***
JP	2.456	3.141***
SOVAA	2.315	6.121***
SOVA	0.875	2.755***
SOVBBB	0.306	0.768
ICOV	0.002	0.118
ROA	0.038	2.408**
DTC	−0.330	−0.512
SDTD	0.105	0.303
JP∗ICOV	0.038	1.129
JP∗ROA	0.188	2.104**
JP∗DTC	−0.808	−0.924
JP∗SDTD	−2.823	−2.430**
Estimated correlation	−0.836	−5.723***
AA or above	>4.275	
A	>2.841 and ≤ 4.275	8.235***
BBB	>1.748 and ≤ 2.841	9.164***
BB	>0.704 and ≤ 1.748	6.788***
B	>0 and ≤ 0.704	3.316***
CCC or below	≤ 0	

Note: *, ** and *** denote significance at the 10%, 5% and 1% levels respectively.
Source: Poon (2003). Reprinted with the permission of Elsevier Science.

be rated. Of the four financial variables, only the return on assets and the short-term debt as a proportion of total debt have correctly signed and significant (positive and negative respectively) impacts on the decision to be rated. The parameters on the sovereign credit rating dummy variables (SOVAA, SOVA and SOVB) are all significant and negative in sign, indicating that any debt issuer in a country with a high sovereign rating is less likely to solicit its own rating from S&P, other things equal.

These sovereign rating dummy variables have the opposite sign in the ratings determinant equation (panel B) as expected, so that firms in countries where government debt is highly rated are themselves more likely to receive a higher rating. Of the four financial variables, only ROA has a significant (and positive) effect on the rating awarded. The dummy for Japanese firms is also positive and significant, and so are three of the four financial variables when interacted with the Japan dummy, indicating that S&P appears to attach different weights to the financial variables when assigning ratings to Japanese firms compared with comparable firms in other countries.

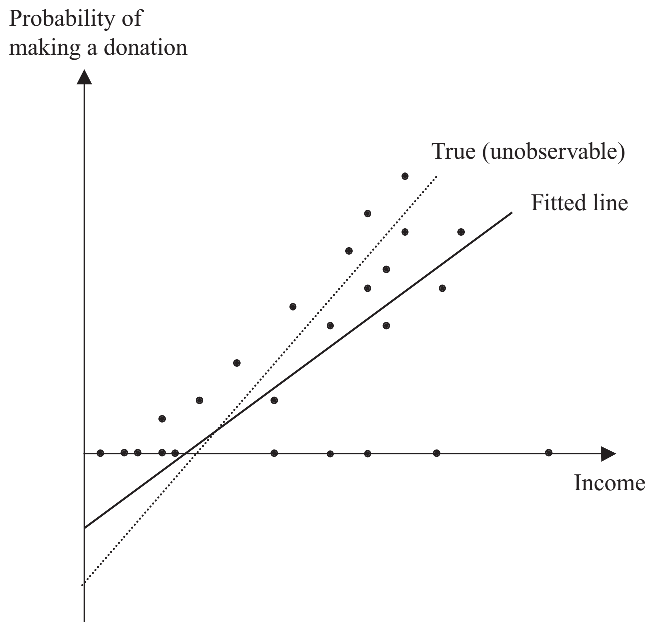
Finally, the estimated correlation between the error terms in the decision to be rated equation and the ratings determinant equation, $\rho_{\epsilon\xi}$, is significant and negative (-0.836), indicating that the results in table 11.3 above would have been subject to self-selection bias and hence the results of the two-stage model are to be preferred. The only disadvantage of this approach, however, is that by construction it cannot answer the core question of whether unsolicited ratings are on average lower after allowing for the debt issuer's financial characteristics, because only firms with solicited ratings are included in the sample at the second stage!

11.13 Censored and truncated dependent variables

Censored or truncated variables occur when the range of values observable for the dependent variables is limited for some reason. Unlike the types of limited dependent variables examined so far in this chapter, censored or truncated variables may not necessarily be dummies. A standard example is that of charitable donations by individuals. It is likely that some people would actually prefer to make negative donations (that is, to receive from the charity rather than to donate to it), but since this is not possible, there will be many observations at exactly zero. So suppose, for example, that we wished to model the relationship between donations to charity and people's annual income, in pounds. The situation we might face is illustrated in figure 11.3.

Given the observed data, with many observations on the dependent variable stuck at zero, OLS would yield biased and inconsistent parameter estimates. An obvious but flawed way to get around this would be just to remove all of the zero observations altogether, since we do not know whether they should be truly zero or negative. However, as well as being

Figure 11.3
Modelling charitable donations as a function of income



inefficient (since information would be discarded), this would still yield biased and inconsistent estimates. This arises because the error term, u_i, in such a regression would not have an expected value of zero, and it would also be correlated with the explanatory variable(s), violating the assumption that $Cov(u_i, x_{ki}) = 0\ \forall k$.

The key differences between censored and truncated data are highlighted in box 11.2. For both censored and truncated data, OLS will not be appropriate, and an approach based on maximum likelihood must be used, although the model in each case would be slightly different. In both cases, we can work out the marginal effects given the estimated parameters, but these are now more complex than in the logit or probit cases.

11.13.1 Censored dependent variable models

The approach usually used to estimate models with censored dependent variables is known as tobit analysis, named after Tobin (1958). To illustrate, suppose that we wanted to model the demand for privatisation IPO shares, as discussed above, as a function of income (x_{2i}), age (x_{3i}), education (x_{4i}) and region of residence (x_{5i}). The model would be

$$\begin{aligned} y_i^* &= \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + u_i \\ y_i &= y_i^* \quad \text{for } y_i^* < 250 \\ y_i &= 250 \quad \text{for } y_i^* \geq 250 \end{aligned} \tag{11.18}$$

Box 11.2 The differences between censored and truncated dependent variables

Although at first sight the two words might appear interchangeable, when the terms are used in econometrics, censored and truncated data are different.

- Censored data occur when the dependent variable has been '*censored*' at a certain point so that values above (or below) this cannot be observed. Even though the dependent variable is censored, the corresponding values of the independent variables are still observable.
- As an example, suppose that a privatisation IPO is heavily oversubscribed, and you were trying to model the demand for the shares using household income, age, education and region of residence as explanatory variables. The number of shares allocated to each investor may have been capped at, say, 250, resulting in a truncated distribution.
- In this example, even though we are likely to have many share allocations at 250 and none above this figure, all of the observations on the independent variables are present and hence the dependent variable is censored, not truncated.
- A truncated dependent variable, meanwhile, occurs when the observations for both the dependent and the independent variables are missing when the dependent variable is above (or below) a certain threshold. Thus the key difference from censored data is that we cannot observe the x_is either, and so some observations are completely cut out or *truncated* from the sample. For example, suppose that a bank were interested in determining the factors (such as age, occupation and income) that affected a customer's decision as to whether to undertake a transaction in a branch or online. Suppose also that the bank tried to achieve this by encouraging clients to fill in an online questionnaire when they log on. There would be no data at all for those who opted to transact in person since they probably would not have even logged on to the bank's web-based system and so would not have the opportunity to complete the questionnaire. Thus, dealing with truncated data is really a sample selection problem because the sample of data that can be observed is not representative of the population of interest – the sample is biased, very likely resulting in biased and inconsistent parameter estimates. This is a common problem, which will result whenever data for buyers or users only can be observed while data for non-buyers or non-users cannot. Of course, it is possible, although unlikely, that the population of interest is focused only on those who use the internet for banking transactions, in which case there would be no problem.

y_i^* represents the true demand for shares (i.e. the number of shares requested) and this will be observable only for demand less than 250. It is important to note in this model that β_2, β_3, etc. represent the impact on the number of shares demanded (of a unit change in x_{2i}, x_{3i}, etc.) and not the impact on the actual number of shares that will be bought (allocated).

An interesting financial application of the tobit approach is due to Haushalter (2000), who employs it to model the determinants of the extent of hedging by oil and gas producers using futures or options over the

1992–1994 period. The dependent variable used in the regression models, the proportion of production hedged, is clearly censored because around half of all of the observations are exactly zero (i.e. the firm does not hedge at all).[12] The censoring of the proportion of production hedged may arise because of high fixed costs that prevent many firms from being able to hedge even if they wished to. Moreover, if companies expect the price of oil or gas to rise in the future, they may wish to increase rather than reduce their exposure to price changes (i.e. 'negative hedging'), but this would not be observable given the way that the data are constructed in the study.

The main results from the study are that the proportion of exposure hedged is negatively related to creditworthiness, positively related to indebtedness, to the firm's marginal tax rate, and to the location of the firm's production facility. The extent of hedging is not, however, affected by the size of the firm as measured by its total assets.

Before moving on, two important limitations of tobit modelling should be noted. First, such models are much more seriously affected by non-normality and heteroscedasticity than are standard regression models (see Amemiya, 1984), and biased and inconsistent estimation will result. Second, as Kennedy (2003, p. 283) argues, the tobit model requires it to be plausible that the dependent variable can have values close to the limit. There is no problem with the privatisation IPO example discussed above since the demand could be for 249 shares. However, it would not be appropriate to use the tobit model in situations where this is not the case, such as the number of shares issued by each firm in a particular month. For most companies, this figure will be exactly zero, but for those where it is not, the number will be much higher and thus it would not be feasible to issue, say, 1 or 3 or 15 shares. In this case, an alternative approach should be used.

11.13.2 *Truncated dependent variable models*

For truncated data, a more general model is employed that contains two equations – one for whether a particular data point will fall into the observed or constrained categories and another for modelling the resulting variable. The second equation is equivalent to the tobit approach. This two-equation methodology allows for a different set of factors to affect the sample selection (for example, the decision to set up internet access to a

[12] Note that this is an example of a *censored* rather than a *truncated* dependent variable because the values of all of the explanatory variables are still available from the annual accounts even if a firm does not hedge at all.

bank account) from the equation to be estimated (for example, to model the factors that affect whether a particular transaction will be conducted online or in a branch). If it is thought that the two sets of factors will be the same, then a single equation can be used and the tobit approach is sufficient. In many cases, however, the researcher may believe that the variables in the sample selection and estimation equations should be different. Thus the equations could be

$$a_i^* = \alpha_1 + \alpha_2 z_{2i} + \alpha_3 z_{3i} + \cdots + \alpha_m z_{mi} + \varepsilon_i \tag{11.19}$$
$$y_i^* = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki} + u_i \tag{11.20}$$

where $y_i = y_i^*$ for $a_i^* > 0$ and, y_i is unobserved for $a_i^* \leq 0$. a_i^* denotes the relative 'advantage' of being in the observed sample relative to the unobserved sample.

The first equation determines whether the particular data point i will be observed or not, by regressing a proxy for the latent (unobserved) variable a_i^* on a set of factors, z_i. The second equation is similar to the tobit model. Ideally, the two equations (11.19) and (11.20) will be fitted jointly by maximum likelihood. This is usually based on the assumption that the error terms, ε_i and u_i, are multivariate normally distributed and allowing for any possible correlations between them. However, while joint estimation of the equations is more efficient, it is computationally more complex and hence a two-stage procedure popularised by Heckman (1976) is often used. The Heckman procedure allows for possible correlations between ε_i and u_i while estimating the equations separately in a clever way – see Maddala (1983).

11.14 Limited dependent variable models in EViews

Estimating limited dependent variable models in EViews is very simple. The example that will be considered here concerns whether it is possible to determine the factors that affect the likelihood that a student will fail his/her MSc. The data comprise a sample from the actual records of failure rates for five years of MSc students in finance at the ICMA Centre, University of Reading contained in the spreadsheet 'MSc_fail.xls'. While the values in the spreadsheet are all genuine, only a sample of 100 students is included for each of five years who completed (or not as the case may be!) their degrees in the years 2003 to 2007 inclusive. Therefore, the data should not be used to infer actual failure rates on these programmes. The idea for this example is taken from a study by Heslop and Varotto (2007) which seeks to propose an approach to preventing systematic biases in admissions decisions.[13]

The objective here is to analyse the factors that affect the probability of failure of the MSc. The dependent variable ('fail') is binary and takes the value 1 if that particular candidate failed at first attempt in terms of his/her overall grade and 0 elsewhere. Therefore, a model that is suitable for limited dependent variables is required, such as a logit or probit.

The other information in the spreadsheet that will be used includes the age of the student, a dummy variable taking the value 1 if the student is female, a dummy variable taking the value 1 if the student has work experience, a dummy variable taking the value 1 if the student's first language is English, a country code variable that takes values from 1 to 10,[14] a dummy variable that takes the value 1 if the student already has a postgraduate degree, a dummy variable that takes the value 1 if the student achieved an A-grade at the undergraduate level (i.e. a first-class honours degree or equivalent), and a dummy variable that takes the value 1 if the undergraduate grade was less than a B-grade (i.e. the student received the equivalent of a lower second-class degree). The B-grade (or upper second-class degree) is the omitted dummy variable and this will then become the reference point against which the other grades are compared – see chapter 9. The reason why these variables ought to be useful predictors of the probability of failure should be fairly obvious and is therefore not discussed. To allow for differences in examination rules and in average student quality across the five-year period, year dummies for 2004, 2005, 2006 and 2007 are created and thus the year 2003 dummy will be omitted from the regression model.

First, **open a new workfile** that can accept 'unstructured/undated' series of length 500 observations and then **import the 13 variables**. The data are **organised by observation** and start in cell **A2**. The country code variable will require further processing before it can be used but the others are already in the appropriate format, so to begin, suppose that we estimate a linear probability model (LPM) of fail on a constant, age, English, female and work experience. This would be achieved simply by running a linear regression in the usual way. While this model has a number of very undesirable features as discussed above, it would nonetheless provide a

[13] Note that since this book uses only a sub-set of their sample and variables in the analysis, the results presented below may differ from theirs. Since the number of fails is relatively small, I deliberately retained as many fail observations in the sample as possible, which will bias the estimated failure rate upwards relative to the true rate.

[14] The exact identities of the countries involved are not revealed in order to avoid any embarrassment for students from countries with high relative failure rates, except that Country 8 is the UK!

useful benchmark with which to compare the more appropriate models estimated below.

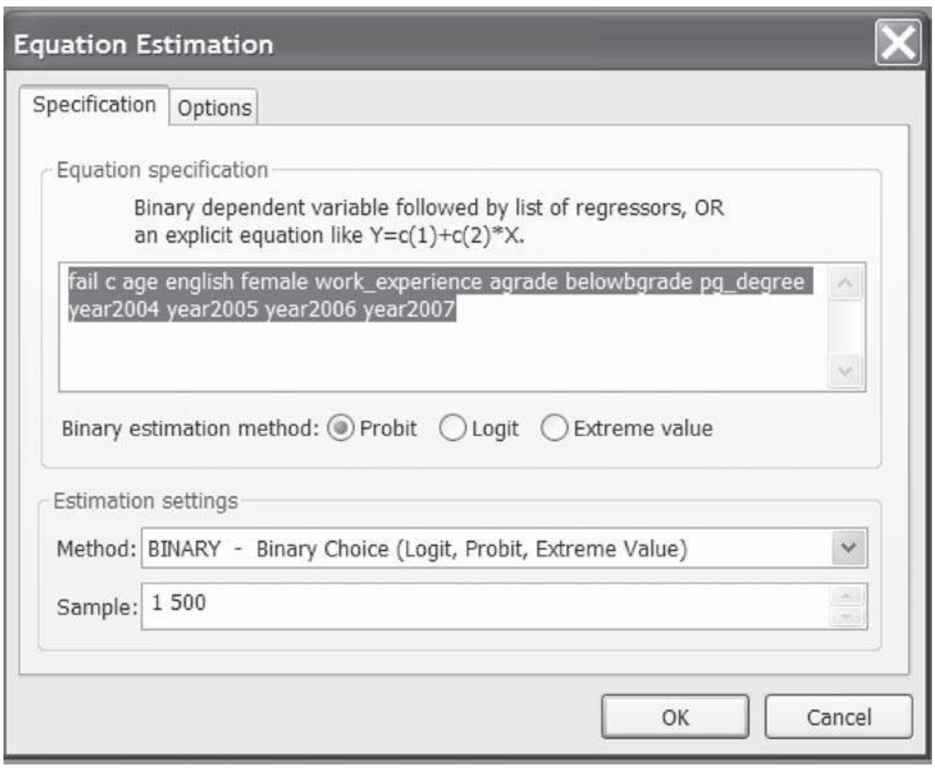
Next, estimate a probit model and a logit model using the same dependent and independent variables as above. Choose **Quick** and then **Equation Estimation**. Then type the dependent variable followed by the explanatory variables

FAIL C AGE ENGLISH FEMALE WORK_EXPERIENCE AGRADE BELOWBGRADE PG_DEGREE YEAR2004 YEAR2005 YEAR2006 YEAR2007

and then in the second window, marked 'Estimation settings', select **BINARY – Binary Choice (Logit, Probit, Extreme Value)** with the whole sample 1 500. The screen will appear as in screenshot 11.1.

Screenshot 11.1

'Equation Estimation' window for limited dependent variables



You can then choose either the probit or logit approach. Note that EViews also provides support for truncated and censored variable models and for multiple choice models, and these can be selected from the drop-down menu by choosing the appropriate method under 'estimation settings'. Suppose that here we wish to choose a probit model (the default). Click on the **Options** tab at the top of the window and this enables you to select **Robust Covariances** and **Huber/White**. This option will ensure that the standard error estimates are robust to heteroscedasticity (see screenshot 11.2).

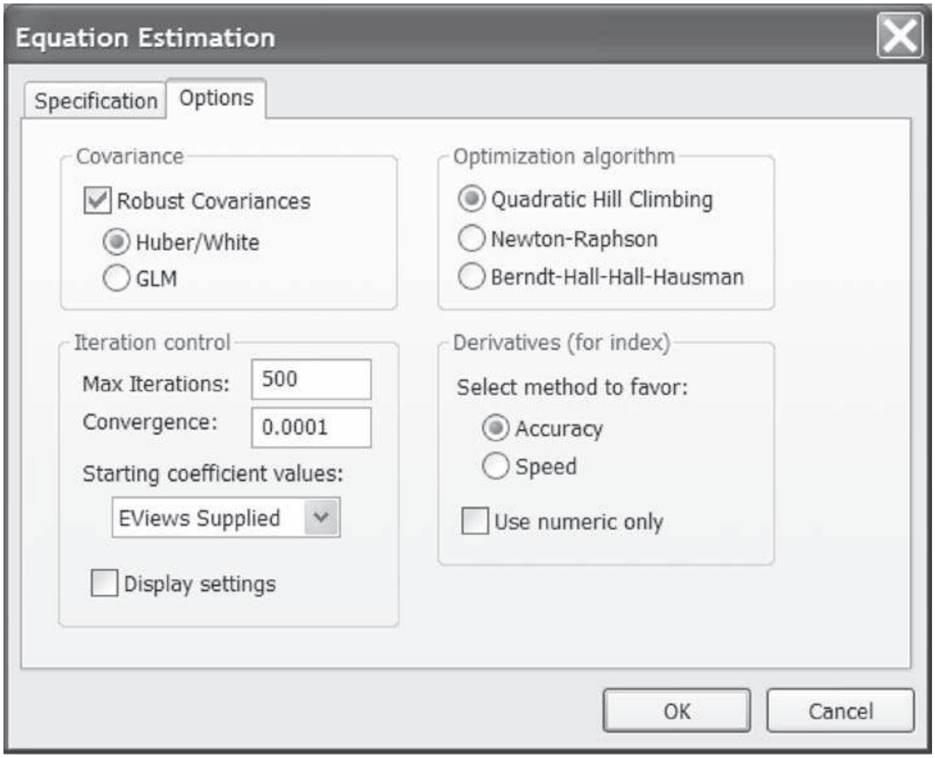
There are other options to change the optimisation method and convergence criterion, as discussed in chapter 8. We do not need to make any modifications from the default here, so click **OK** and the results will appear. **Freeze and name this table** and then, for completeness, **estimate a logit model**. The results that you should obtain for the probit model are as follows:

Dependent Variable: FAIL
Method: ML – Binary Probit (Quadratic hill climbing)
Date: 08/04/07 Time: 19:10
Sample: 1 500
Included observations: 500
Convergence achieved after 5 iterations
QML (Huber/White) standard errors & covariance

	Coefficient	Std. Error	z-Statistic	Prob.
C	−1.287210	0.609503	−2.111901	0.0347
AGE	0.005677	0.022559	0.251648	0.8013
ENGLISH	−0.093792	0.156226	−0.600362	0.5483
FEMALE	−0.194107	0.186201	−1.042460	0.2972
WORK_EXPERIENCE	−0.318247	0.151333	−2.102956	0.0355
AGRADE	−0.538814	0.231148	−2.331038	0.0198
BELOWBGRADE	0.341803	0.219301	1.558601	0.1191
PG_DEGREE	0.132957	0.225925	0.588502	0.5562
YEAR2004	0.349663	0.241450	1.448181	0.1476
YEAR2005	−0.108330	0.268527	−0.403422	0.6866
YEAR2006	0.673612	0.238536	2.823944	0.0047
YEAR2007	0.433785	0.24793	1.749630	0.0802

McFadden R-squared	0.088870	Mean dependent var	0.134000
S.D. dependent var	0.340993	S.E. of regression	0.333221
Akaike info criterion	0.765825	Sum squared resid	54.18582
Schwarz criterion	0.866976	Log likelihood	−179.4563
Hannan-Quinn criter.	0.805517	Restr. log likelihood	−196.9602
LR statistic	35.00773	Avg. log likelihood	−0.358913
Prob(LR statistic)	0.000247		
Obs with Dep=0	433	Total obs	500
Obs with Dep=1	67		

As can be seen, the pseudo-R^2 values are quite small at just below 9%, although this is often the case for limited dependent variable models. Only the work experience and A-grade variables and two of the year



Screenshot 11.2 'Equation Estimation' options for limited dependent variables

dummies have parameters that are statistically significant, and the Below B-grade dummy is almost significant at the 10% level in the probit specification (although less so in the logit). As the final two rows of the tables note, the proportion of fails in this sample is quite small, which makes it harder to fit a good model than if the proportions of passes and fails had been more evenly balanced. Various goodness of fit statistics can be examined by (from the logit or probit estimation output window) clicking **View/Goodness-of-fit Test**. . . . A further check on model adequacy is to produce a set of 'in-sample forecasts' – in other words, to construct the fitted values. To do this, click on the **Forecast** tab after estimating the probit model and then **uncheck the forecast evaluation box** in the 'Output' window as the evaluation is not relevant in this case. All other options can be left as the default settings and then the plot of the fitted values shown on figure 11.4 results.

The unconditional probability of failure for the sample of students we have is only 13.4% (i.e. only 67 out of 500 failed), so an observation should be classified as correctly fitted if either $y_i = 1$ and $\hat{y}_i > 0.134$ or $y_i = 0$ and $\hat{y}_i < 0.134$. The easiest way to evaluate the model in EViews is to click **View/Actual,Fitted,Residual Table** from the logit or probit output screen.

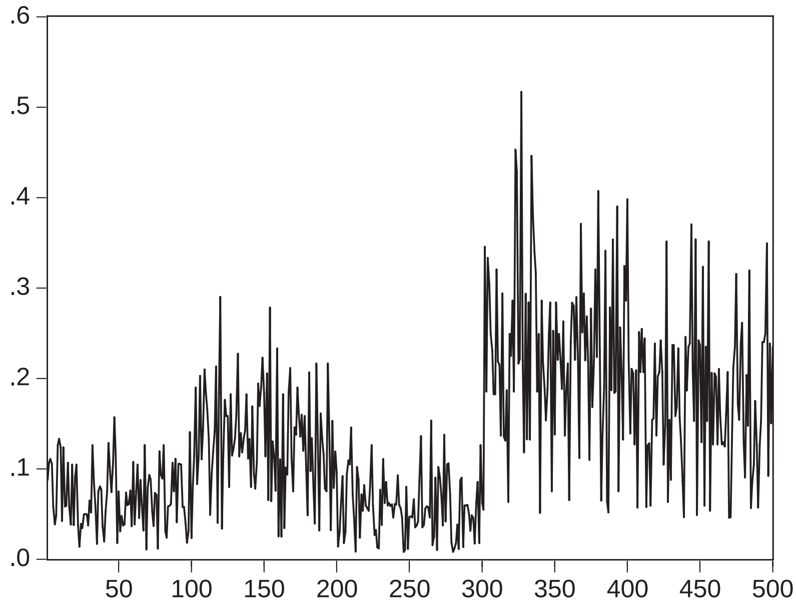


Figure 11.4
Fitted values from the failure probit regression

Then from this information we can identify that of the 67 students that failed, the model correctly predicted 46 of them to fail (and it also incorrectly predicted that 21 would pass). Of the 433 students who passed, the model incorrectly predicted 155 to fail and correctly predicted the remaining 278 to pass. Eviews can construct an 'expectation-prediction classification table' automatically by clicking on **View/Expectation-Prediction Table** and then entering the unconditional probability of failure as the cutoff when prompted (**0.134**). Overall, we could consider this a reasonable set of (in sample) predictions.

It is important to note that, as discussed above, we cannot interpret the parameter estimates in the usual way. In order to be able to do this, we need to calculate the marginal effects. Unfortunately, EViews does not do this automatically, so the procedure is probably best achieved in a spreadsheet using the approach described in box 11.1 for the logit model and analogously for the probit model. If we did this, we would end up with the statistics displayed in table 11.5, which are interestingly quite similar in value to those obtained from the linear probability model.

This table presents us with values that can be intuitively interpreted in terms of how the variables affect the probability of failure. For example, an age parameter value of 0.0012 implies that an increase in the age of the student by 1 year would increase the probability of failure by 0.12%, holding everything else equal, while a female student is around 2.5–3%

Table 11.5 Marginal effects for logit and probit models for probability of MSc failure

Parameter	logit	probit
C	−0.2433	−0.1646
AGE	0.0012	0.0007
ENGLISH	−0.0178	−0.0120
FEMALE	−0.0360	−0.0248
WORK_EXPERIENCE	−0.0613	−0.0407
AGRADE	−0.1170	−0.0689
BELOWBGRADE	0.0606	0.0437
PG_DEGREE	0.0229	0.0170
YEAR2004	0.0704	0.0447
YEAR2005	−0.0198	−0.0139
YEAR2006	0.1344	0.0862
YEAR2007	0.0917	0.0555

(depending on the model) less likely than a male student with otherwise identical characteristics to fail. Having an A-grade (first class) in the bachelors degree makes a candidate either 6.89% or 11.7% (depending on the model) less likely to fail than an otherwise identical student with a B-grade (upper second-class degree). Finally, since the year 2003 dummy has been omitted from the equations, this becomes the reference point. So students were more likely in 2004, 2006 and 2007, but less likely in 2005, to fail the MSc than in 2003.

Key concepts

The key terms to be able to define and explain from this chapter are

- limited dependent variables
- logit
- probit
- censored variables
- truncated variables
- ordered response
- multinomial logit
- marginal effects
- pseudo-R^2

Review questions

1. Explain why the linear probability model is inadequate as a specification for limited dependent variable estimation.
2. Compare and contrast the probit and logit specifications for binary choice variables.

3. (a) Describe the intuition behind the maximum likelihood estimation technique used for limited dependent variable models.
   (b) Why do we need to exercise caution when interpreting the coefficients of a probit or logit model?
   (c) How can we measure whether a logit model that we have estimated fits the data well or not?
   (d) What is the difference, in terms of the model setup, in binary choice versus multiple choice problems?
4. (a) Explain the difference between a censored variable and a truncated variable as the terms are used in econometrics.
   (b) Give examples from finance (other than those already described in this book) of situations where you might meet each of the types of variable described in part (a) of this question.
   (c) With reference to your examples in part (b), how would you go about specifying such models and estimating them?
5. Re-open the 'fail_xls' spreadsheet for modelling the probability of MSc failure and do the following:
   (a) Take the country code series and construct separate dummy variables for each country. Re-run the probit and logit regression above with all of the other variables plus the country dummy variables. Set up the regression so that the UK becomes the reference point against which the effect on failure rate in other countries is measured. Is there evidence that any countries have significantly higher or lower probabilities of failure than the UK, holding all other factors in the model constant? In the case of the logit model, use the approach given in box 11.1 to evaluate the differences in failure rates between the UK and each other country.
   (b) Suppose that a fellow researcher suggests that there may be a non-linear relationship between the probability of failure and the age of the student. Estimate a probit model with all of the same variables as above plus an additional one to test this. Is there indeed any evidence of such a nonlinear relationship?

Appendix: The maximum likelihood estimator for logit and probit models

Recall that under the logit formulation, the estimate of the probability that $y_i = 1$ will be given from equation (11.4), which was

$$P_i = \frac{1}{1 + e^{-(\beta_1 + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + u_i)}} \tag{11A.1}$$

Set the error term, u_i, to its expected value for simplicity and again, let $z_i = \beta_1 + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}$, so that we have

$$P_i = \frac{1}{1 + e^{-z_i}} \tag{11A.2}$$

We will also need the probability that $y_i \neq 1$ or equivalently the probability that $y_i = 0$. This will be given by 1 minus the probability in (11A.2).[15] Given that we can have actual zeros and ones only for y_i rather than probabilities, the likelihood function for each observation y_i will be

$$L_i = \left(\frac{1}{1 + e^{-z_i}}\right)^{y_i} \times \left(\frac{1}{1 + e^{z_i}}\right)^{(1-y_i)} \tag{11A.3}$$

The likelihood function that we need will be based on the joint probability for all N observations rather than an individual observation i. Assuming that each observation on y_i is independent, the joint likelihood will be the product of all N marginal likelihoods. Let $L\,(\theta\,|x_{2i}, x_{3i}, \ldots, x_{ki}; i = 1, N\,)$ denote the likelihood function of the set of parameters $(\beta_1, \beta_2, \ldots, \beta_k)$ given the data. Then the likelihood function will be given by

$$L\,(\theta) = \prod_{i=1}^{N} \left(\frac{1}{1 + e^{-z_i}}\right)^{y_i} \times \left(\frac{1}{1 + e^{z_i}}\right)^{(1-y_i)} \tag{11A.4}$$

As for maximum likelihood estimator of GARCH models, it is computationally much simpler to maximise an additive function of a set of variables than a multiplicative function, so long as we can ensure that the parameters required to achieve this will be the same. We thus take the natural logarithm of equation (11A.4) and this log-likelihood function is maximised

$$LLF = -\sum_{i=1}^{N} [y_i \ln(1 + e^{-z_i}) + (1 - y_i) \ln(1 + e^{z_i})] \tag{11A.5}$$

Estimation for the probit model will proceed in exactly the same way, except that the form for the likelihood function in (11A.4) will be slightly different. It will instead be based on the familiar normal distribution function described in the appendix to chapter 8.

[15] We can use the rule that

$$1 - \frac{1}{1 + e^{-z_i}} = \frac{1 + e^{-z_i} - 1}{1 + e^{-z_i}} = \frac{e^{-z_i}}{1 + e^{-z_i}} = \frac{e^{-z_i}}{1 + \frac{1}{e^{z_i}}} = \frac{e^{-z_i} \times e^{z_i}}{1 + e^{z_i}} = \frac{1}{1 + e^{z_i}}.$$