5 : CHAPTER

Models with Limited Dependent Variables

In previous chapters, we have dealt with models where the range of the dependent variable is unbounded. Now we shall consider imposing both an upper and a lower bound on the variable. We shall concentrate on a case where the dependent variable is a binary or dichotomous variable which can take only the values of zero and one.

A model whose dependent variable has an upper and a lower bound can be derived from an ordinary regression model by mapping the dependent variable y through a sigmoid or S-shaped function. As $y \to -\infty$, the sigmoid tends to its lower asymptote whereas, as $y \to \infty$, the sigmoid tends to the upper asymptote.

A model with a binary dependent variable may be obtained from an ordinary regression model by mapping the dependent variable y through a step function representing a threshold mechanism. When y falls short of the threshold value, the response of the mechanism is to generate a zero. When y exceeds the threshold, an unit is delivered.

The models of this chapter are of a sophisticated sort which comprise both a sigmoid function and a threshold mechanism. For conceptual purposes, these models may be broken into two parts.

The first part is a probability model. Here a systematic value, which is derived from a set of explanatory variables, is mapped through a sigmoid function to generate a probability value π which is bounded above by unity and below by zero.

The second part of the model is a sampling process which generates the observations by means of the threshold mechanism based on the probability generated by the first part of the model. In the simplest case, the sampling process is a point binomial which is akin to the tossing of a biased coin when heads—the unit outcome—has the probability π and tails—the zero outcome—has the probability $1 - \pi$. In other cases, the sampling process is described by a binomial distribution, and the outcome is akin to the number of heads resulting from n tosses of a coin.

Until recent years, the models of this chapter had been used more commonly in the life sciences and in experimental psychology than in economics. Their belated discovery by economists in the 1970's led to a veritable flood of applications in economics throughout the 1980's.

Probability Models

Imagine that one wishes to explain the occurrence or the non-occurrence of an event which can affect each of n individuals in a given sample. The binary or boolean variable $y_i \in \{0, 1\}$ serves to indicate whether or not the event has affected the ith individual.

The object is to find a function of a set of observable variables with which to express the probability of the occurrence of the event in the case of any individual. If all the members of a group or a population were affected by the same global values of the variables, then such a function should enable us to predict the proportion who experience the event and to predict how this might change with changing values of the variables.

If we were to record, for each individual, the values of k variables which influence the probability of the event, then we could express this probability, in the ith instance, by

$$P(y_i = 1) = \pi(x_{i.}, \beta), \tag{1}$$

where $x_{i.} = [x_{i1}, \ldots, x_{ik}]$ are the variables and β is a vector of parameters. If we are able to specify the form of the function π, and if the values $x_{i.}; i = 1, \ldots, n$ are sufficiently heterogeneous, then we may hope to derive an estimate of β from the sample of n individuals. When the same values of x are experienced by groups of individuals, then we have to observe several groups in varying circumstances before we can hope to make make such inferences.

It is helpful to regard π as the composition of two mappings:

$$\pi = \pi\big\{h(x)\big\}. \tag{2}$$

The function $h = h(x)$ is often a linear function of the observations which takes the form of $h_i = x_{i.}\beta$. The function $\pi = \pi(h)$ is a distribution function which fulfils the condition

$$0 \leq \pi(h) \leq 1 \quad \text{with} \quad \pi(-\infty) = 0 \quad \text{and} \quad \pi(\infty) = 1. \tag{3}$$

There are three common choices for $\pi(h)$:

(4) (i) *The uniform distribution*

$$\pi(h) = \begin{cases} 0, & \text{if } h \leq 0; \\ h, & \text{if } 0 \leq h \leq 1; \\ 1, & \text{if } 1 \leq h. \end{cases}$$

(ii) *The logistic distribution*

$$\pi(h) = \frac{e^h}{1 + e^h}.$$

(iii) *The normal distribution*

$$\pi(h) = \int_{-\infty}^{h} \frac{1}{\sqrt{2\pi}} e^{-\zeta^2/2} d\zeta.$$

In the first case, we have a linear probability model, in the second case, we have a logistic probability model or logit model and, in the third case, we have a probit model.

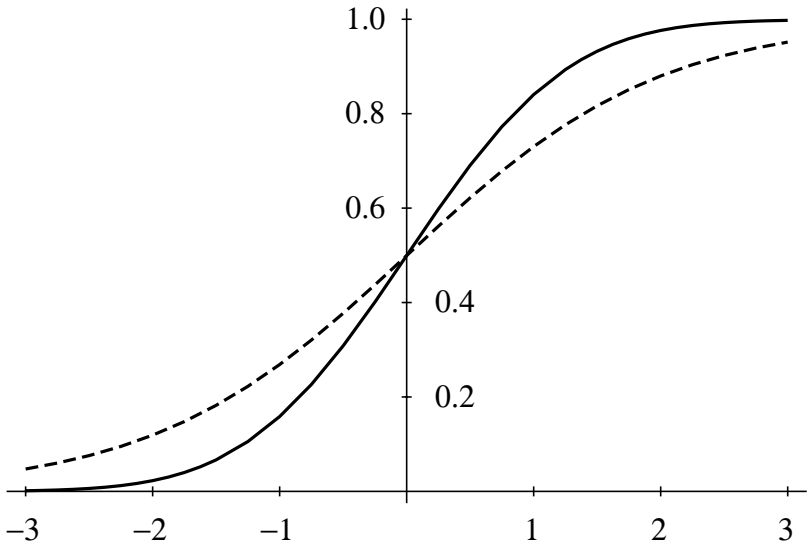

Figure 1. The cumulative standard normal distribution, the plain line, and the logistic function, the broken line, plotted to the same scale.

The attraction of the probit model is that it is based on a distribution—the normal distribution—for which there is often a clear statistical interpretation. A disadvantage of the normal distribution is that there is no closed-form expression for its integral, which can be a hindrance when it comes to computing the estimates of the parameters of the function $h(x, \beta)$. This factor explains the attraction of the alternative logistic distribution which is far more tractable from the point of view of computation.

In the figure above, the cumulative standard normal distribution and the logistic function are plotted to the same scale. By applying a scale factor to the logistic curve it can be made to approximate the normal curve quite closely.

The choice of the scale factor is not an unequivocal matter. One way of rescaling the logistic function is to take account of the fact that the corresponding density function has a variance of $\pi^2/3$ which is to be compared with the unit variance of the standard normal density function. Therefore one might apply a scale factor of $\sqrt{3}/\pi = 0.551$ to the logistic function in order to standardise the dispersion of the distribution.

However, if we consider that the logistic and the normal curves are intended to provide probability values, then it appears that, instead of equating the dispersion of the two density functions, we should compare the cumulative distribution functions. On the basis of such reasoning, Amemiya (1981) has suggested that it would be more appropriate to apply a scale factor of 0.625 to the logistic function. This improves the approximation of the distribution functions in the vicinity of $z = 0$ at the cost of sacrificing the accuracy of their approximation at extreme values of z.

The Classical Probit Model

The classical example of a probit model concerns the effects of a pesticide upon a sample of insects. For the ith insect, the lethal dosage is the quantity δ_i which is the realised value of a random variable; and it is assumed that, in the population of these insects, the values $\xi = \log(\delta)$ are distributed normally with a mean of μ and a variance of σ^2. If an insect is selected at random and is subjected to the dosage d_i, then the probability that it will die is $P(\xi_i < x_i)$, where $x_i = \log d_i$. This is given by

$$\pi(x_i) = \int_{-\infty}^{x_i} N(\xi; \mu, \sigma) d\xi. \tag{5}$$

The function $\pi(x)$ with $x = \log(d)$ also indicates the fraction of a sample of insects which could be expected to die if all the individuals were subjected to the same global dosage d.

Let $y_i = 1$ if the ith insect dies and $y_i = 0$ if it survives. Then the situation of the insect is summarised by writing

$$y_i = \begin{cases} 0, & \text{if } x_i \leq \xi_i \quad \text{or, equivalently,} \quad d_i \leq \delta_i; \\ 1, & \text{if } \xi_i < x_i \quad \text{or, equivalently,} \quad \delta_i < d_i. \end{cases} \tag{6}$$

By making the assumption that it is the log of the lethal dosage which follows a normal distribution, rather than the lethal dosage itself, we avoid the unwitting implication that insects can die from negative dosages. The lethal dosages are said to have a log-normal distribution.

The log-normal distribution has an upper tail which converges rather slowly to zero. Therefore the corresponding tail of the cumulative distribution converges slowly to the upper asymptote of unity, which implies that some

individuals are virtually immune to the effects of the pesticide. In a laboratory experiment, one would expect to find, to the contrary, that there is a moderate dosage which is certain to kill all the insects. In the field, however, there is always the chance that some insects will be sheltered from the pesticide.

The integral of (5) may be expressed in terms of a standard normal density function $N(\zeta; 0, 1)$ so as to accord with the formulation under (4, iii). Thus

$$\begin{aligned} &P(\xi_i < x_i) \quad \text{with} \quad \xi_i \sim N(\mu, \sigma^2) \\ &\text{is equal to} \\ &P\left(\frac{\xi_i - \mu}{\sigma} = \zeta_i < h_i = \frac{x_i - \mu}{\sigma}\right) \quad \text{with} \quad \zeta_i \sim N(0, 1). \end{aligned} \tag{7}$$

Moreover, the standardised variable h_i, which corresponds to the dose received by the ith insect, can be written as

$$\begin{aligned} &h_i = \frac{x_i - \mu}{\sigma} = \beta_0 + \beta_1 x_i, \\ &\text{where} \quad \beta_0 = -\frac{\mu}{\sigma} \quad \text{and} \quad \beta_1 = \frac{1}{\sigma}. \end{aligned} \tag{8}$$

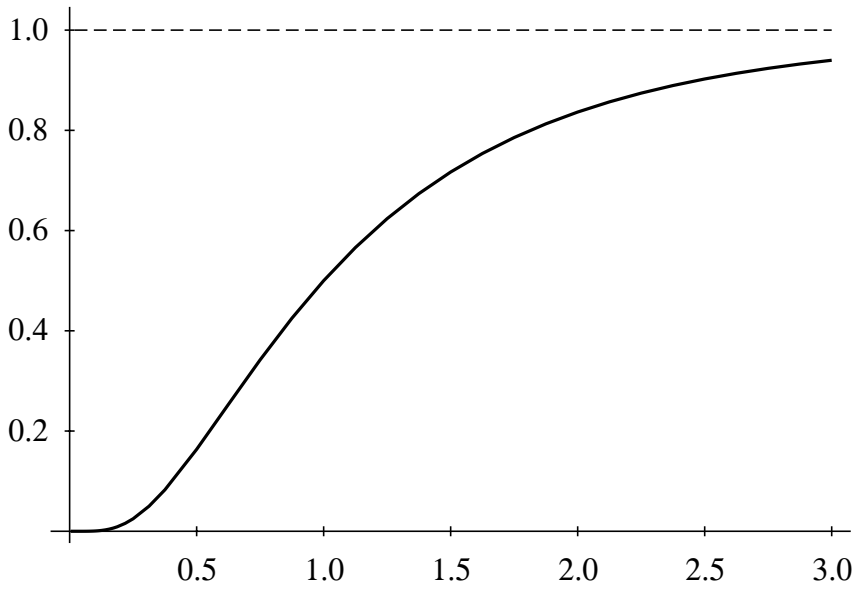

Figure 2. The cumulative log-normal distribution. The logarithm of the log-normal variate is a standard normal variate.

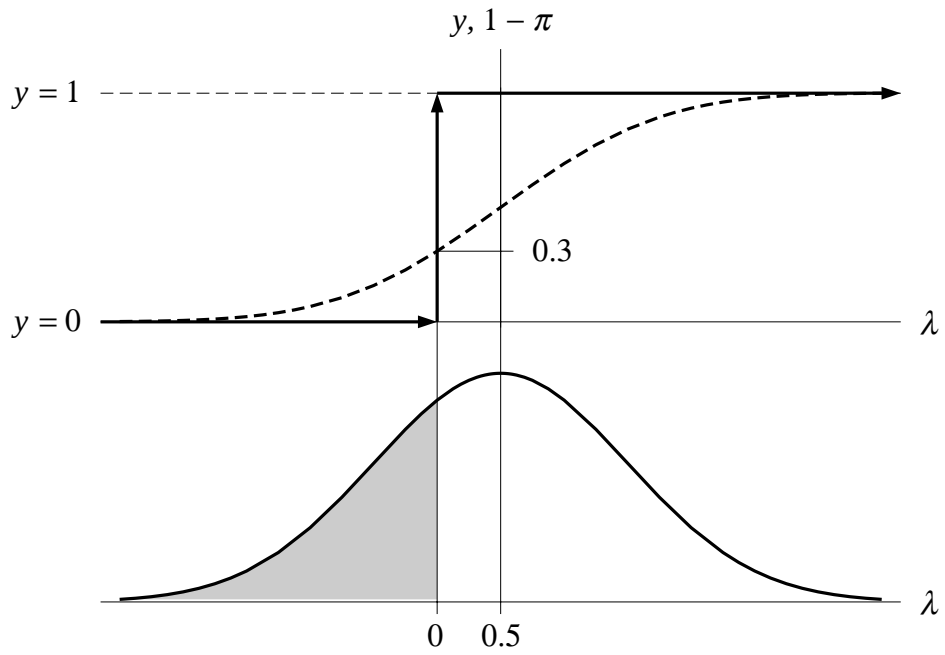

Figure 3. If $\lambda = h + \varepsilon$ exceeds the threshold value of zero, then the step function, indicated by the arrows in the upper digram, delivers $y = 1$. When $\varepsilon \sim N(0, 1)$ and $h = 0.5$, the probability that λ will fall short of the threshold is 0.3, which is the area of the shaded region in the lower figure.

The Latent-Variable Formulation

In the example above, an individual threshold ξ_i, whose value is essentially unobservable, has been attributed to each of the sample elements. In many applications, it is helpful to think in terms of a universal threshold and to attribute to each of the sample elements an unobservable latent variable which either exceeds or falls short of that threshold. Let λ_i be the value of the latent variable for the ith individual, and let the threshold be located at zero. Then we have

$$y_i = \begin{cases} 0, & \text{if } \lambda_i \leq 0; \\ 1, & \text{if } 0 < \lambda_i. \end{cases} \tag{9}$$

An example is provided by a consumer who weighs the costs and benefits of an economic decision which might be whether or not to purchase a durable item via a credit agreement. If the discounted future benefits exceed the discounted costs, then the net value λ_i of the prospective purchase is positive, and the agreement is signed giving $y_i = 1$. Otherwise, with $\lambda_i \leq 0$, we get $y_i = 0$.

In the case of the probit model, we can set

$$\begin{aligned} &\lambda_i = h(x_i; \beta) - \zeta_i \\ &\text{where} \quad \zeta_i \sim N(0, 1). \end{aligned} \tag{10}$$

It follows that

$$\begin{aligned} P(y_i = 0) &= P(\lambda_i \leq 0) \\ &= P(h_i \leq \zeta_i) = 1 - \pi(h_i), \end{aligned} \tag{11}$$

and that

$$\begin{aligned} P(y_i = 1) &= P(0 < \lambda_i) \\ &= P(\zeta_i < h_i) = \pi(h_i). \end{aligned} \tag{12}$$

The logistic model can be represented in the same way.

If $h = \beta_0 + \beta_1 x$ is a simple linear function and if we set $-\zeta_i = \varepsilon_i$, then we get $\lambda_i = \beta_0 + \beta_1 x + \varepsilon_i$, which suggests a comparison with the RHS of an ordinary regression model.

Estimation with Individual Data

Imagine that we have a sample of observations $(y_i, x_{i.}); i = 1, \ldots, n$ where $y_i \in \{0, 1\}$ for all i. Then, assuming that the events affecting the individuals are statistically independent and taking $\pi_i = \pi(x_{i.}, \beta)$ to represent the probability that the event will affect the ith individual, we can write represent the likelihood function for the sample as

$$L(\beta) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1 - y_i} = \prod_{i=1}^{n} \left(\frac{\pi_i}{1 - \pi_i} \right)^{y_i} (1 - \pi_i). \tag{13}$$

This is the product of n point binomials. The log of the likelihood function is given by

$$\log L = \sum_{i=1}^{n} y_i \log \left(\frac{\pi_i}{1 - \pi_i} \right) + \sum_{i=1}^{n} \log(1 - \pi_i). \tag{14}$$

Differentiating $\log L$ with respect to β_j, which is the jth element of the parameter vector β, yields

$$\begin{aligned} \frac{\partial \log L}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{y_i}{\pi_i (1 - \pi_i)} \frac{\partial \pi_i}{\partial \beta_j} - \sum_{i=1}^{n} \frac{1}{1 - \pi_i} \frac{\partial \pi_i}{\partial \beta_j} \\ &= \sum_{i=1}^{n} \frac{y_i - \pi_i}{\pi_i (1 - \pi_i)} \frac{\partial \pi_i}{\partial \beta}. \end{aligned} \tag{15}$$

To obtain the second-order derivatives which are also needed, it is helpful to write the final expression of (15) as

$$\frac{\partial \log L}{\partial \beta_j} = \sum \left\{ \frac{y_i}{\pi_i} - \frac{1 - y_i}{1 - \pi_i} \right\} \frac{\partial \pi_i}{\partial \beta_j}. \tag{16}$$

Then it can be seen more easily that

$$\frac{\partial^2 \log L}{\partial\beta_j\beta_k} = \sum_i \left\{ \frac{y_i}{\pi_i} - \frac{1-y_i}{1-\pi_i} \right\} \frac{\partial^2\pi_i}{\partial\beta_j\beta_k} - \sum_i \left\{ \frac{y_i}{\pi_i^2} + \frac{1-y_i}{(1-\pi_i)^2} \right\} \frac{\partial\pi_i}{\partial\beta_j} \frac{\partial\pi_i}{\partial\beta_k}. \tag{17}$$

The negative of the expected value of the matrix of second derivatives is the information matrix whose inverse provides the asymptotic dispersion matrix of the maximum-likelihood estimates. The expected value of the expression above is found by taking $E(y_i) = \pi_i$. On taking expectations, the first term of the RHS of (17) vanishes and the second term is simplified, with the result that

$$E\left(\frac{\partial^2 \log L}{\partial\beta_j\beta_k} \right) = \sum_i \frac{1}{\pi_i(1-\pi_i)} \frac{\partial\pi_i}{\partial\beta_j} \frac{\partial\pi_i}{\partial\beta_k}. \tag{18}$$

The maximum-likelihood estimates are the values which satisfy the conditions

$$\frac{\partial \log L(\beta)}{\partial\beta} = 0. \tag{19}$$

To solve this equation requires an iterative procedure. The Newton–Raphson procedure serves the purpose.

The Newton–Raphson Procedure

A common procedure for finding the solution or root of a nonlinear equation $\alpha(x) = 0$ is the Newton–Raphson procedure which depends upon approximating the curve $y = \alpha(x)$ by its tangent at a point near the root. Let this point be $[x_0, \alpha(x_0)]$. Then the equation of the tangent is

$$y = \alpha(x_0) + \frac{\partial\alpha(x_0)}{\partial x}(x - x_0) \tag{20}$$

and, on setting $y = 0$, we find that this line intersects the x-axis at

$$x_1 = x_0 - \left[\frac{\partial\alpha(x_0)}{\partial x} \right]^{-1} \alpha(x_0). \tag{21}$$

If x_0 is close to the root λ of the equation $\alpha(x) = 0$, then we can expect x_1 to be closer still. To find an accurate approximation to λ, we generate a sequence of approximations $\{x_0, x_1, \ldots, x_r, x_{r+1}, \ldots\}$ according to the algorithm

$$x_{r+1} = x_r - \left[\frac{\partial\alpha(x_r)}{\partial x} \right]^{-1} \alpha(x_r). \tag{22}$$

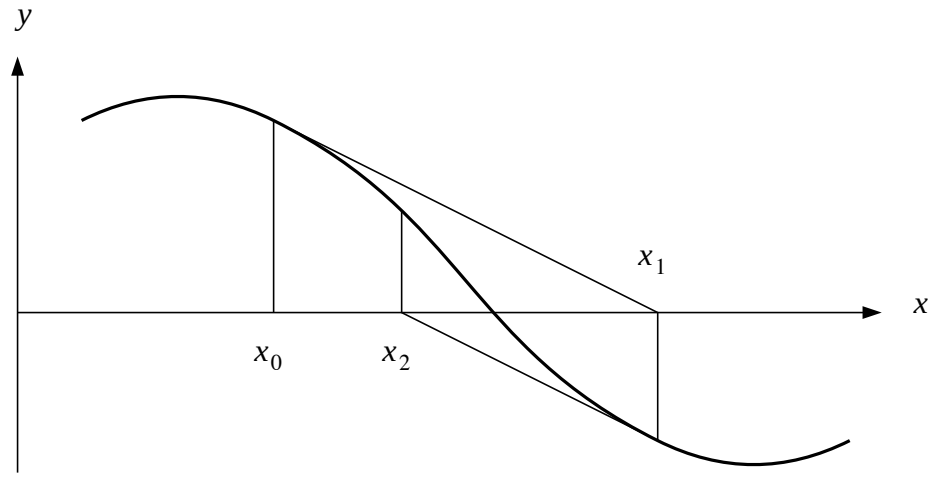

Figure 4. If x_0 is close to the root of the equation $\alpha(x) = 0$, then we can expect x_1 to be closer still.

The Newton–Raphson procedure is readily adapted to the problem of finding the value of the vector β which satisfies the equation $\partial \log L(\beta)/\partial\beta = 0$ which is the first-order condition for the maximisation of the log-likelihood function. Let β consist of two elements β_0 and β_1. Then the algorithm by which the $(r+1)$th approximation to the solution is obtained from the rth approximation is specified by

$$\begin{bmatrix} \beta_0 \\ \\ \beta_1 \end{bmatrix}_{(r+1)} = \begin{bmatrix} \beta_0 \\ \\ \beta_1 \end{bmatrix}_{(r)} - \begin{bmatrix} \dfrac{\partial^2 \log L}{\partial \beta_0^2} & \dfrac{\partial^2 \log L}{\partial \beta_0 \beta_1} \\ \dfrac{\partial^2 \log L}{\partial \beta_1 \beta_0} & \dfrac{\partial^2 \log L}{\partial \beta_1^2} \end{bmatrix}_{(r)}^{-1} \begin{bmatrix} \dfrac{\partial \log L}{\partial \beta_0} \\ \dfrac{\partial \log L}{\partial \beta_1} \end{bmatrix}. \tag{23}$$

It is common to replace the matrix of second-order partial derivatives in this algorithm by its expected value which is the negative of information matrix. The modified procedure is known as Fishers's method of scoring. The algebra is often simplified by replacing the derivatives by their expectations, whereas the properties of the algorithm are hardly affected.

In the case of the simple probit model, where there is no closed-form expression for the likelihood function, the probability values, together with the various derivatives and expected derivatives to be found under (15) to (18), which are needed in order to implement one or other of these estimation procedures, may be evaluated with the help of tables which can be read into the computer.

Recall that the probability values π are specified by the cumulative normal

distribution

$$\pi(h) = \int_{-\infty}^{h} \frac{1}{\sqrt{2\pi}} e^{-\zeta^2/2} d\zeta. \tag{24}$$

We may assume, for the sake of a simple illustration, that the function $h(x)$ is linear:

$$h(x) = \beta_0 + \beta_1 x. \tag{25}$$

Then the derivatives $\partial\pi_i/\partial\beta_j$ become

$$\frac{\partial \pi_i}{\partial \beta_0} = \frac{\partial \pi_i}{\partial h} . \frac{\partial h}{\partial \beta_0} = N\{h(x_i)\} \quad \text{and} \quad \frac{\partial \pi_i}{\partial \beta_1} = \frac{\partial \pi_i}{\partial h} . \frac{\partial h}{\partial \beta_1} = N\{h(x_i)\} x_i, \tag{26}$$

where N denotes the normal density function which is the derivative of π.

Estimation with Grouped Data

In the classical applications of probit analysis, the data was usually in the form of grouped observations. Thus, to assess the effectiveness of an insecticide, various levels of dosage $d_j; j = 1, \ldots, J$ would be administered to batches of n_j insects. The numbers $m_j = \sum_i y_{ij}$ killed in each batch would be recorded and their proportions $p_j = m_j/n_j$ would be calculated.

If a sufficiently wide range of dosages are investigated, and if the numbers n_j in the groups are large enough to allow the sample proportions p_j accurately to reflect the underlying probabilities π_j, then the plot of p_j against $x_j = \log d_j$ should give a clear impression of the underlying distribution function $\pi = \pi\{h(x)\}$.

In the case of a single experimental variable x, it would be a simple matter to infer the parameters of the function $h = \beta_0 + \beta_1 x$ from the plot. According to the model, we have

$$\pi(h) = \pi(\beta_0 + \beta_1 x). \tag{27}$$

From the inverse $h = \pi^{-1}(\pi)$ of the function $\pi = \pi(h)$, one may obtain the values $h_j = \pi^{-1}(p_j)$. In the case of the probit model, this is a matter of referring to the table of the standard normal distribution. The values of π or p are found in the body of the table whilst the corresponding values of h are the entries in the margin. Given the points (h_j, x_j) for $j = 1, \ldots J$, it is a simple matter to fit a regression equation in the form of

$$h_j = b_0 + b_1 x_j + e_j. \tag{28}$$

In the early days of probit analysis, before the advent of the electronic computer, such fitting was often performed by eye with the help of a ruler.

To derive a more sophisticated and efficient method of estimating the parameters of the model, we may pursue a method of maximum-likelihood. This method is a straightforward generalisation of the one which we have applied to individual data.

Consider a group of n individuals which are subject to the same probability $P(y = 1) = \pi$ for the event in question. The probability that the event will occur in m out of n cases is given by the binomial formula:

$$B(m, n, \pi) = \binom{n}{m}\pi^m(1-\pi)^{n-m} = \frac{n!}{m!(n-m)!}\pi^m(1-\pi)^{n-m}. \tag{29}$$

If there are J independent groups, then the joint probability of their outcomes $m_1, \ldots, m_j$ is the product

$$L = \prod_{j=1}^{J}\binom{n_j}{m_j}\pi_j^{m_j}(1-\pi_j)^{n_j-m_j} = \prod_{j=1}^{J}\binom{n_j}{m_j}\left(\frac{\pi_j}{1-\pi_j}\right)^{m_j}(1-\pi_j)^{n_j}. \tag{30}$$

Therefore the log of the likelihood function is

$$\log L = \sum_{j=1}^{J}\left\{m_j \log\left(\frac{\pi_j}{1-\pi_j}\right) + n_j \log(1-\pi_j) + \log\binom{n_j}{m_j}\right\}. \tag{31}$$

Given that $\pi_j = \pi(x_{j.}, \beta)$, the problem is to estimate β by finding the value which satisfies the first-order condition for maximising the likelihood function which is

$$\frac{\partial \log L(\beta)}{\partial \beta} = 0. \tag{32}$$

To provide a simple example, let us take the linear logistic model

$$\pi = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}. \tag{33}$$

The so-called log-odds ratio is

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x. \tag{34}$$

Therefore the log-likelihood function of (31) becomes

$$\log L = \sum_{j=1}^{J}\left\{m_j(\beta_0+\beta_1 x_j) - n_j \log(1-e^{\beta_0+\beta_1 x_j}) + \log\binom{n_j}{m_j}\right\}, \tag{35}$$

and its derivatives in respect of β_0 and β_1 are

$$\begin{aligned}\frac{\partial \log L}{\partial \beta_0} &= \sum_j \left\{ m_j - n_j \left(\frac{e^{\beta_0+\beta_1 x_j}}{1+e^{\beta_0+\beta_1 x_j}} \right) \right\} = \sum_j (m_j - n_j \pi_j), \\ \frac{\partial \log L}{\partial \beta_1} &= \sum_j \left\{ m_j x_j - n_j x_j \left(\frac{e^{\beta_0+\beta_1 x_j}}{1+e^{\beta_0+\beta_1 x_j}} \right) \right\} = \sum_j x_j (m_j - n_j \pi_j).\end{aligned} \tag{36}$$

The information matrix, which, together with the above derivatives, is used in estimating the parameters by Fishers's method of scoring, is provided by

$$\begin{bmatrix} \sum_j m_j \pi_j (1-\pi_j) & \sum_j m_j x_j \pi_j (1-\pi_j) \\ \sum_j m_j x_j \pi_j (1-\pi_j) & \sum_j m_j x_j^2 \pi_j (1-\pi_j) \end{bmatrix}. \tag{37}$$