Chapter 10

Binary choice and limited dependent variable models, and maximum likelihood estimation

10.1 Overview

The first part of this chapter describes the linear probability model, logit analysis, and probit analysis, three techniques for fitting regression models where the dependent variable is a qualitative characteristic. Next it discusses tobit analysis, a censored regression model fitted using a combination of linear regression analysis and probit analysis. This leads to sample selection models and heckman analysis. The second part of the chapter introduces maximum likelihood estimation, the method used to fit all of these models except the linear probability model.

10.2 Learning outcomes

After working through the corresponding chapter in the text, studying the corresponding slideshows, and doing the starred exercises in the text and the additional exercises in this subject guide, you should be able to:

- describe the linear probability model and explain its defects
- describe logit analysis, giving the mathematical specification
- describe probit analysis, including the mathematical specification
- calculate marginal effects in logit and probit analysis
- explain why OLS yields biased estimates when applied to a sample with censored observations, even when the censored observations are deleted
- explain the problem of sample selection bias and describe how the heckman procedure may provide a solution to it (in general terms, without mathematical detail)
- explain the principle underlying maximum likelihood estimation
- apply maximum likelihood estimation from first principles in simple models.

10.3 Further material

Limiting distributions and the properties of maximum likelihood estimators

Provided that weak regularity conditions involving the differentiability of the likelihood function are satisfied, maximum likelihood (ML) estimators have the following attractive properties in large samples:

(1) They are consistent.

(2) They are asymptotically normally distributed.

(3) They are asymptotically efficient.

The meaning of the first property is familiar. It implies that the probability density function of the estimator collapses to a spike at the true value. This being the case, what can the other assertions mean? If the distribution becomes degenerate as the sample size becomes very large, how can it be described as having a normal distribution? And how can it be described as being efficient, when its variance, and the variance of any other consistent estimator, tend to zero?

To discuss the last two properties, we consider what is known as the limiting distribution of an estimator. This is the distribution of the estimator when the divergence between it and its population mean is multiplied by $\sqrt{n}$. If we do this, the distribution of a typical estimator remains nondegenerate as n becomes large, and this enables us to say meaningful things about its shape and to make comparisons with the distributions of other estimators (also multiplied by $\sqrt{n}$).

To put this mathematically, suppose that there is one parameter of interest, θ, and that $\widehat{\theta}$ is its ML estimator. Then (2) says that:

$$\sqrt{n}\left(\widehat{\theta}-\theta\right) \sim N(0,\, \sigma^2)$$

for some variance σ^2. (3) says that, given any other consistent estimator $\tilde{\theta}$, $\sqrt{\tilde{\theta}-\theta}$ cannot have a smaller variance.

Test procedures for maximum likelihood estimation

This section on ML tests contains material that is a little advanced for an introductory econometrics course. It is provided because likelihood ratio tests are encountered in the sections on binary choice models and because a brief introduction may be of help to those who proceed to a more advanced course.

There are three main approaches to testing hypotheses in maximum likelihood estimation: likelihood ratio (LR) tests, Wald tests, and Lagrange multiplier (LM) tests. Since the theory behind Lagrange multiplier tests is relatively complex, the present discussion will be confined to the first two types. We will start by assuming that the probability density function of a random variable X is a known function of a single unknown parameter θ and that the likelihood function for θ given a sample of n observations on X, $L(\theta \,|\, X_1, \ldots, X_n)$, satisfies weak regularity conditions involving its differentiability. In particular, we assume that θ is determined by the first-order condition $\mathrm{d}L/\mathrm{d}\theta = 0$. (This rules out estimators such as that in Exercise A10.7) The null hypothesis is $H_0 : \theta = \theta_0$, the alternative hypothesis is $H_1 : \theta \neq \theta_0$, and the maximum likelihood estimate of θ is $\widehat{\theta}$.

Likelihood ratio tests

A likelihood ratio test compares the value of the likelihood function at $\theta = \widehat{\theta}$ with its value at $\theta = \theta_0$. In view of the definition of $\widehat{\theta}$, $L(\widehat{\theta}) \geq L(\theta_0)$ for all θ_0. However, if the null hypothesis is true, the ratio $L(\widehat{\theta})/L(\theta_0)$ should not be significantly greater than 1. As a consequence, the logarithm of the ratio:

$$\log\left(\frac{L(\widehat{\theta})}{L(\theta_0)}\right) = \log L(\widehat{\theta}) - \log L(\theta_0)$$

should not be significantly different from zero. In that it involves a comparison of the measures of goodness of fit for unrestricted and restricted versions of the model, the LR test is similar to an F test.

Under the null hypothesis, it can be shown that in large samples the test statistic:

$$LR = 2\left(\log L(\widehat{\theta}) - \log L(\theta_0)\right)$$

has a chi-squared distribution with one degree of freedom. If there are multiple parameters of interest, and multiple restrictions, the number of degrees of freedom is equal to the number of restrictions.

Examples

We will return to the example in Section 10.6 in the textbook, where we have a normally-distributed random variable X with unknown population mean μ and known standard deviation equal to 1. Given a sample of n observations, the likelihood function is:

$$L(\widehat{\mu} \,|\, X_1, \ldots, X_n) = \left(\frac{1}{\sqrt{2\pi}\mathrm{e}^{(X_1-\mu)^2/2}}\right) \times \cdots \times \left(\frac{1}{\sqrt{2\pi}\mathrm{e}^{(X_n-\mu)^2/2}}\right).$$

The log-likelihood is:

$$\log L(\widehat{\mu} \,|\, X_1, \ldots, X_n) = n \log\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}\sum (X_i - \widehat{\mu})^2$$

and the unrestricted ML estimate is $\widehat{\mu} = \overline{X}$. The LR statistic for the null hypothesis $H_0 : \mu = \mu_0$ is therefore:

$$\begin{aligned} LR &= 2\left(\left(n\log\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}\sum (X_i - \overline{X})^2\right) - \left(n\log\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}(X_i - \mu_0)^2\right)\right) \\ &= \sum (X_i - \mu_0)^2 - \sum (X_i - \overline{X})^2 = n(\overline{X} - \mu_0)^2. \end{aligned}$$

If we relaxed the assumption $\sigma = 1$, the unrestricted likelihood function is:

$$L(\widehat{\mu}, \widehat{\sigma} \,|\, X_1, \ldots, X_n) = \left(\frac{1}{\widehat{\sigma}\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}\left(\frac{X_1-\widehat{\mu}}{\widehat{\sigma}}\right)^2}\right) \times \cdots \times \left(\frac{1}{\widehat{\sigma}\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}\left(\frac{X_n-\widehat{\mu}}{\widehat{\sigma}}\right)^2}\right)$$

and the log-likelihood is:

$$\log L(\widehat{\mu}, \widehat{\sigma} \,|\, X_1, \ldots, X_n) = n \log \left(\frac{1}{\sqrt{2\pi}} \right) - n \log \widehat{\sigma} - \frac{1}{2\widehat{\sigma}^2} \sum (X_i - \widehat{\mu})^2.$$

The first-order condition obtained by differentiating by σ is:

$$\frac{\partial \log L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum (X_i - \mu)^2 = 0$$

from which we obtain:

$$\widehat{\sigma}^2 = \frac{1}{n} \sum (X_i - \widehat{\mu})^2.$$

Substituting back into the log-likelihood function, the latter now becomes a function of μ only (and is known as the concentrated log-likelihood function or, sometimes, the profile log-likelihood function):

$$\log L(\mu \,|\, X_1, \ldots, X_n) = n \log \left(\frac{1}{\sqrt{2\pi}} \right) - n \log \left(\frac{1}{n} \sum (X_i - \mu)^2 \right)^{1/2} - \frac{n}{2}.$$

As before, the ML estimator of μ is $\bar{X}$. Hence the LR statistic is:

$$\begin{aligned} LR &= 2 \left(\left(n \log \left(\frac{1}{\sqrt{2\pi}} \right) - n \log \left(\frac{1}{n} \sum (X_i - \overline{X})^2 \right)^{1/2} - \frac{n}{2} \right) \right. \\ &\quad \left. - \left(n \log \left(\frac{1}{\sqrt{2\pi}} \right) - n \log \left(\frac{1}{n} \sum (X_i - \mu_0)^2 \right)^{1/2} - \frac{n}{2} \right) \right) \\ &= n \left(\log \sum (X_i - \mu_0)^2 - \log \sum (X_i - \overline{X})^2 \right). \end{aligned}$$

It is worth noting that this is closely related to the F statistic obtained when one fits the least squares model:

$$X_i = \mu + u_i.$$

The least squares estimator of μ is $\overline{X}$ and $RSS = \sum (X_i - \overline{X})^2$.

If one imposes the restriction $\mu = \mu_0$, we have $RSS_{\mathrm{R}} = \sum (X_i - \mu_0)^2$ and the F statistic:

$$F(1, n-1) = \frac{\sum (X_i - \mu_0)^2 - \sum (X_i - \overline{X})^2}{\left(\sum (X_i - \overline{X})^2 \right) / (n-1)}.$$

Returning to the LR statistic, we have:

$$\begin{aligned} LR &= n \log \frac{\sum (X_i - \mu_0)^2}{\sum (X_i - \overline{X})^2} = n \log \left(1 + \frac{\sum (X_i - \mu_0)^2 - \sum (X_i - \overline{X})^2}{\sum (X_i - \overline{X})^2} \right) \\ &\cong n \frac{\sum (X_i - \mu_0)^2 - \sum (X_i - \overline{X})^2}{\sum (X_i - \overline{X})^2} = \frac{n}{n-1} F \cong F. \end{aligned}$$

Note that we have used the approximation $\log(1 + a) = a$ which is valid when a is small enough for higher powers to be neglected.

Wald tests

Wald tests are based on the same principle as t tests in that they evaluate whether the discrepancy between the maximum likelihood estimate θ and the hypothetical value θ_0 is significant, taking account of the variance in the estimate. The test statistic for the null hypothesis $H_0 : \widehat{\theta} - \theta_0 = 0$ is:

$$\frac{\left(\widehat{\theta} - \theta_0\right)^2}{\widehat{\sigma}^2_{\widehat{\theta}}}$$

where $\widehat{\sigma}^2_{\widehat{\theta}}$ is the estimate of the variance of θ evaluated at the maximum likelihood value. $\widehat{\sigma}^2_{\widehat{\theta}}$ can be estimated in various ways that are asymptotically equivalent if the likelihood function has been specified correctly. A common estimator is that obtained as minus the inverse of the second differential of the log-likelihood function evaluated at the maximum likelihood estimate. Under the null hypothesis that the restriction is valid, the test statistic has a chi-squared distribution with one degree of freedom. When there are multiple restrictions, the test statistic becomes more complex and the number of degrees of freedom is equal to the number of restrictions.

Examples

We will use the same examples as for the LR test, first, assuming that $\sigma = 1$ and then assuming that it has to be estimated along with μ. In the first case the log-likelihood function is:

$$\log L(\mu \,|\, X_1, \ldots, X_n) = n \log \left(\frac{1}{\sqrt{2\pi}} \right) - \frac{1}{2} \sum (X_i - \mu)^2.$$

The first differential is $\sum(X_i - \mu)$ and the second is $-n$, so the estimate of the variance is $1/n$. The Wald test statistic is therefore $n(\overline{X} - \mu_0)^2$.

In the second example, where σ was unknown, the concentrated log-likelihood function is:

$$\begin{aligned}\log L(\mu \,|\, X_1, \ldots, X_n) &= n \log \left(\frac{1}{\sqrt{2\pi}} \right) - n \log \left(\frac{1}{n} \sum (X_i - \mu)^2 \right)^{1/2} - \frac{n}{2} \\ &= n \log \left(\frac{1}{\sqrt{2\pi}} \right) - \frac{n}{2} \log \frac{1}{n} - \frac{n}{2} \log \left(\sum (X_i - \mu)^2 \right) - \frac{n}{2}.\end{aligned}$$

The first derivative with respect to μ is:

$$\frac{\mathrm{d} \log L}{\mathrm{d}\mu} = n \frac{\sum (X_i - \mu)}{\sum (X_i - \mu)^2}.$$

The second derivative is:

$$\frac{\mathrm{d}^2 \log L}{\mathrm{d}\mu^2} = n \frac{(-n) \left(\sum (X_i - \mu)^2\right) - \left(\sum (X_i - \mu)\right) \left(-2 \sum (X_i - \mu)\right)}{\left[\sum (X_i - \mu)^2\right]^2}.$$

Evaluated at the ML estimator $\widehat{\mu} = \overline{X}$, $\sum(X_i - \mu) = 0$ and hence:

$$\frac{\mathrm{d}^2 \log L}{\mathrm{d}\mu^2} = -\frac{n^2}{\sum (X_i - \mu)^2}$$

giving an estimated variance $\widehat{\sigma}^2/n$, given:

$$\widehat{\sigma}^2 = \frac{1}{n}\sum (X_i - \overline{X})^2.$$

Hence the Wald test statistic is:

$$\frac{(\overline{X} - \mu_0)^2}{\widehat{\sigma}^2/n}.$$

Under the null hypothesis, this is distributed as a chi-squared statistic with one degree of freedom.

When there is just one restriction, as in the present case, the Wald statistic is the square of the corresponding asymptotic t statistic (asymptotic because the variance has been estimated asymptotically). The chi-squared test and the t test are equivalent, given that, when there is one degree of freedom, the critical value of the chi-squared statistic for any significance level is the square of the critical value of the normal distribution.

LR test of restrictions in a regression model

Given the regression model:

$$Y_i = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ij} + u_i$$

with u assumed to be iid $N(0,\, \sigma^2)$, the log-likelihood function for the parameters is:

$$\log L(\beta_1, \ldots, \beta_k, \sigma \,|\, Y_i, X_i, i = 1, \ldots, n) = n \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2\sigma^2}\sum\left(Y_i - \beta_1 - \sum_{j=2}^{k}\beta_j X_{ij}\right)^2.$$

This is a straightforward generalisation of the expression for a simple regression derived in Section 10.6 in the textbook. Hence

$$\log L(\beta_1, \ldots, \beta_k, \sigma \,|\, Y_i, X_i, i = 1, \ldots, n) = -n\log\sigma - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}Z$$

where:

$$Z = \sum\left(Y_i - \beta_1 - \sum_{j=2}^{k}\beta_j X_{ij}\right)^2.$$

The estimates of the β parameters affect only Z. To maximise the log-likelihood, they should be chosen so as to minimise Z, and of course this is exactly what one is doing when one is fitting a least squares regression. Hence $Z = RSS$. It remains to determine the ML estimate of σ. Taking the partial differential with respect to σ, we obtain one of the first-order conditions for a maximum:

$$\frac{\partial \log L(\beta_1, \ldots, \beta_k, \sigma)}{\partial\sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}RSS = 0.$$

From this we obtain:

$$\widehat{\sigma}^2 = \frac{RSS}{n}.$$

Hence the ML estimator is the sum of the squares of the residuals divided by n. This is different from the least squares estimator, which is the sum of the squares of the residuals divided by $n-k$, but the difference disappears as the sample size becomes large. Substituting for $\widehat{\sigma}^2$ in the log-likelihood function, we obtain the concentrated likelihood function:

$$\begin{aligned}\log L(\beta_1,\ldots,\beta_k\,|\,Y_i,X_i,i=1,\ldots,n) &= -n\log\left(\frac{RSS}{n}\right)^{1/2}-\frac{n}{2}\log 2\pi-\frac{1}{2Z/n}RSS\\ &= -\frac{n}{2}\log\frac{RSS}{n}-\frac{n}{2}\log 2\pi-\frac{n}{2}\\ &= -\frac{n}{2}(\log RSS+1+\log 2\pi-\log n).\end{aligned}$$

We will re-write this as:

$$\log L_{\mathrm{U}}=-\frac{n}{2}(\log RSS_{\mathrm{U}}+1+\log 2\pi-\log n)$$

the subscript U emphasising that this is the unrestricted log-likelihood. If we now impose a restriction on the parameters and maximise the loglikelihood function subject to the restriction, it will be:

$$\log L_{\mathrm{R}}=-\frac{n}{2}(\log RSS_{\mathrm{R}}+1+\log 2\pi-\log n)$$

where $RSS_{\mathrm{R}}\geq RSS_{\mathrm{U}}$ and hence $\log L_{\mathrm{R}}\leq\log L_{\mathrm{U}}$. The LR statistic for a test of the restriction is therefore:

$$2(\log L_{\mathrm{U}}-L_{\mathrm{R}})=n(\log RSS_{\mathrm{R}}-\log RSS_{\mathrm{U}})=n\log\frac{RSS_{\mathrm{R}}}{RSS_{\mathrm{U}}}.$$

It is distributed as a chi-squared statistic with one degree of freedom under the null hypothesis that the restriction is valid. If there is more than one restriction, the test statistic is the same but the number of degrees of freedom under the null hypothesis that all the restrictions are valid is equal to the number of restrictions.

An example of its use is the common factor test in Section 12.3 in the text. As with all maximum likelihood tests, it is valid only for large samples. Thus for testing linear restrictions we should prefer the F test approach because it is valid for finite samples.

10.4 Additional exercises

A10.1 *What factors affect the decision to make a purchase of your category of expenditure in the CES data set?*

Define a new variable *CATBUY* that is equal to 1 if the household makes any purchase of your category and 0 if it makes no purchase at all. Regress *CATBUY* on *EXPPC*, *SIZE*, *REFAGE*, and *COLLEGE* (as defined in Exercise A5.6) using: (1) the linear probability model, (2) the logit model, and (3) the probit model. Calculate the marginal effects at the mean of *EXPPC*, *SIZE*, *REFAGE*, and *COLLEGE* for the logit and probit models and compare them with the coefficients of the linear probability model.

A10.2 Logit analysis was used to relate the event of a respondent working (*WORKING*, defined to be 1 if the respondent was working, and 0 otherwise) to the respondent's educational attainment (S, defined as the highest grade completed) using 1994 data from the National Longitudinal Survey of Youth 1979–. In this year the respondents were aged 29–36 and a substantial number of females had given up work to raise a family. The analysis was undertaken for females and males separately, with the output shown below (first females, then males, with iteration messages deleted):

```
. logit WORKING S if MALE==0

Logit Estimates                                       Number of obs =   2726
                                                      chi2(1)       =  70.42
                                                      Prob > chi2   = 0.0000
Log Likelihood = -1586.5519                           Pseudo R2     = 0.0217

------------------------------------------------------------------------------
 WORKING |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
       S |   .1511872   .0186177      8.121   0.000       .1146971    .1876773
   _cons |  -1.049543   .2448064     -4.287   0.000      -1.529355   -.5697314
------------------------------------------------------------------------------


. logit WORKING S if MALE==1

Logit Estimates                                       Number of obs =   2573
                                                      chi2(1)       =  75.03
                                                      Prob > chi2   = 0.0000
Log Likelihood = -802.65424                           Pseudo R2     = 0.0446

------------------------------------------------------------------------------
 WORKING |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
       S |   .2499295   .0306482      8.155   0.000       .1898601    .3099989
   _cons |  -.9670268   .3775658     -2.561   0.010      -1.707042   -.2270113
------------------------------------------------------------------------------
```

95 per cent of the respondents had S in the range 9–18 years and the mean value of S was 13.3 and 13.2 years for females and males, respectively.

From the logit analysis, the marginal effect of S on the probability of working at the mean was estimated to be 0.030 and 0.020 for females and males, respectively. Ordinary least squares regressions of *WORKING* on S yielded slope coefficients of 0.029 and 0.020 for females and males, respectively.

As can be seen from the second figure below, the marginal effect of educational attainment was lower for males than for females over most of the range $S \geq 9$. Discuss the plausibility of this finding.

As can also be seen from the second figure, the marginal effect of educational attainment decreases with educational attainment for both males and females over the range $S \geq 9$. Discuss the plausibility of this finding.

Compare the estimates of the marginal effect of educational attainment using logit analysis with those obtained using ordinary least squares.

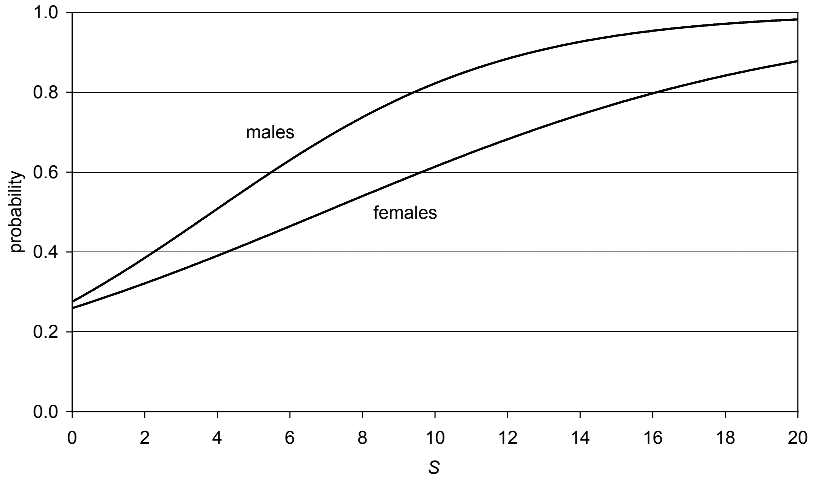

Figure 10.1: Probability of working, as a function of S.

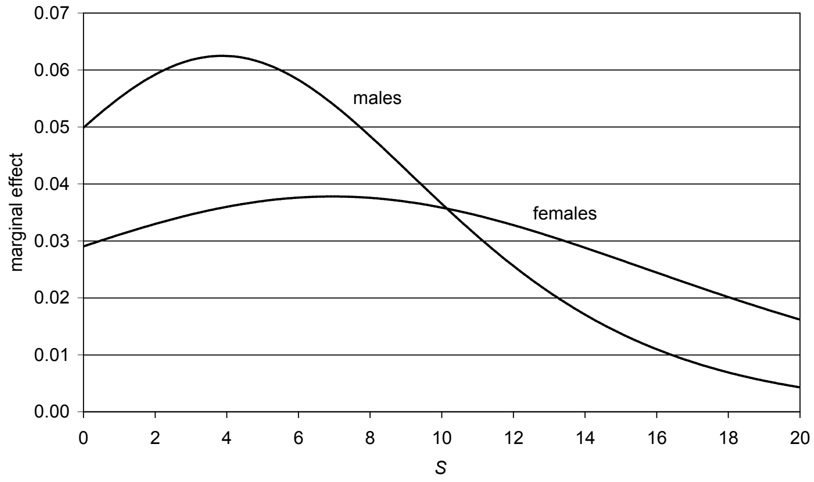

Figure 10.2: Marginal effect of S on the probability of working.

A10.3 A researcher has data on weight, height, and schooling for 540 respondents in the National Longitudinal Survey of Youth 1979– for the year 2002. Using the data on weight and height, he computes the body mass index for each individual. If the body mass index is 30 or greater, the individual is defined to be obese. He defines a binary variable, *OBESE*, that is equal to 1 for the 164 obese individuals and 0 for the other 376. He wishes to investigate whether obesity is related to schooling and fits an ordinary least squares (OLS) regression of *OBESE* on S, years of schooling, with the following result (t statistics in parentheses):

$$\begin{array}{rcll} \widehat{OBESE} & = & 0.595 - 0.021S & \qquad (1) \\ & & (5.30) \;\; (2.63) & \end{array}$$

This is described as the linear probability model (LPM). He also fits the logit model:

$$F(Z) = \frac{1}{1+\mathrm{e}^{-Z}}$$

where $F(Z)$ is the probability of being obese and $Z = \beta_1 + \beta_2 S$, with the following result (again, t statistics in parentheses):

$$\begin{aligned} \widehat{Z} &= \underset{(1.07)}{0.588} - \underset{(2.60)}{0.105}S \qquad (2) \end{aligned}$$

The figure below shows the probability of being obese and the marginal effect of schooling as a function of S, given the logit regression. Most (492 out of 540) of the individuals in the sample had 12 to 18 years of schooling.

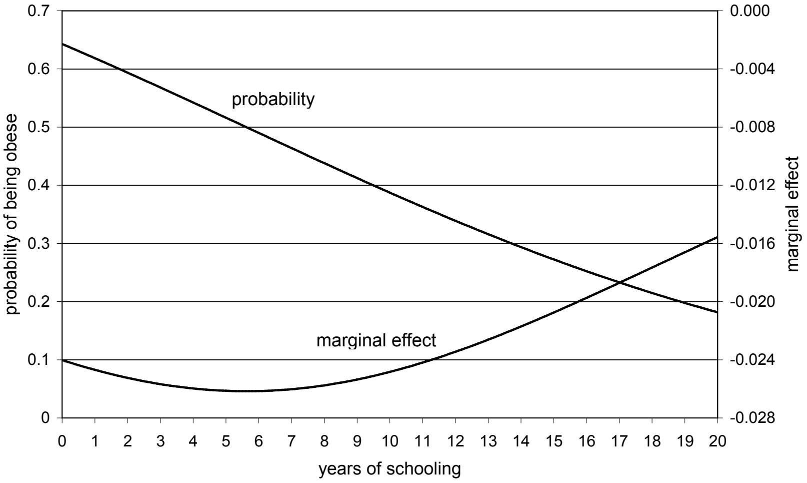

Figure 10.3: Scatter diagram of probability of being obese against years of schooling.

- Discuss whether the relationships indicated by the probability and marginal effect curves appear to be plausible.
- Add the probability function and the marginal effect function for the LPM to the diagram. Explain why you drew them the way you did.
- The logit model is considered to have several advantages over the LPM. Explain what these advantages are. Evaluate the importance of the advantages of the logit model in this particular case.
- The LPM is fitted using OLS. Explain how, instead, it might be fitted using maximum likelihood estimation:
  - ○ Write down the probability of being obese for any obese individual, given S_i for that individual, and write down the probability of not being obese for any non-obese individual, again given S_i for that individual.
  - ○ Write down the likelihood function for this sample of 164 obese individuals and 376 non-obese individuals.
  - ○ Explain how one would use this function to estimate the parameters. [Note: You are not expected to attempt to derive the estimators of the parameters.]

- ○ Explain whether your maximum likelihood estimators will be the same or different from those obtained using least squares.

A10.4 A researcher interested in the relationship between parenting, age and schooling has data for the year 2000 for a sample of 1,167 married males and 870 married females aged 35 to 42 in the National Longitudinal Survey of Youth 1979–. In particular, she is interested in how the presence of young children in the household is related to the age and education of the respondent. She defines *CHILDL6* to be 1 if there is a child less than 6 years old in the household and 0 otherwise and regresses it on AGE, age, and S, years of schooling, for males and females separately using probit analysis. Defining the probability of having a child less than 6 in the household to be $p = F(Z)$ where:

$$Z = \beta_1 + \beta_2 AGE + \beta_3 S$$

she obtains the results shown in the table below (asymptotic standard errors in parentheses).

	Males	Females
AGE	−0.137	−0.154
	(0.018)	(0.023)
S	0.132	0.094
	(0.015)	(0.020)
constant	0.194	0.547
	(0.358)	(0.492)
$\overline{Z}$	−0.399	−0.874
$f(\overline{Z})$	0.368	0.272

For males and females separately, she calculates:

$$\overline{Z} = \widehat{\beta}_1 + \widehat{\beta}_2 \overline{AGE} + \widehat{\beta}_3 \overline{S}$$

where $\overline{AGE}$ and $\overline{S}$ are the mean values of AGE and S and $\widehat{\beta}_1$, $\widehat{\beta}_2$, and $\widehat{\beta}_3$ are the probit coefficients in the corresponding regression, and she further calculates:

$$f(\overline{Z}) = \frac{1}{\sqrt{2\pi}} e^{-\bar{Z}^2/2}$$

where $f(Z) = \mathrm{d}F/\mathrm{d}Z$. The values of $\overline{Z}$ and $f(\overline{Z})$ are shown in the table.

- Explain how one may derive the marginal effects of the explanatory variables on the probability of having a child less than 6 in the household, and calculate for both males and females the marginal effects at the means of AGE and S.
- Explain whether the signs of the marginal effects are plausible. Explain whether you would expect the marginal effect of schooling to be higher for males or for females.

- At a seminar someone asks the researcher whether the marginal effect of S is significantly different for males and females. The researcher does not know how to test whether the difference is significant and asks you for advice. What would you say?

A10.5 A health economist investigating the relationship between smoking, schooling, and age, defines a dummy variable D to be equal to 1 for smokers and 0 for nonsmokers. She hypothesises that the effects of schooling and age are not independent of each other and defines an interactive term schooling*age. She includes this as an explanatory variable in the probit regression. Explain how this would affect the estimation of the marginal effects of schooling and age.

A10.6 A researcher has data on the following variables for 5,061 respondents in the National Longitudinal Survey of Youth 1979–:

- *MARRIED*, marital status in 1994, defined to be 1 if the respondent was married with spouse present and 0 otherwise;
- *MALE*, defined to be 1 if the respondent was male and 0 if female;
- *AGE* in 1994 (the range being 29–37);
- S, years of schooling, defined as highest grade completed, and
- *ASVABC*, score on a test of cognitive ability, scaled so as to have mean 50 and standard deviation 10.

She uses probit analysis to regress *MARRIED* on the other variables, with the output shown:

```
. probit MARRIED MALE AGE S ASVABC

Probit estimates                                  Number of obs   =       5061
                                                  LR chi2(4)      =     229.78
                                                  Prob > chi2     =     0.0000
Log likelihood = -3286.1289                       Pseudo R2       =     0.0338

------------------------------------------------------------------------------
     MARRIED |      Coef.   Std. Err.      z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        MALE |  -.1215281    .036332    -3.34   0.001    -.1927375   -.0503188
         AGE |    .028571   .0081632     3.50   0.000     .0125715    .0445705
           S |  -.0017465     .00919    -0.19   0.849    -.0197587    .0162656
      ASVABC |   .0252911   .0022895    11.05   0.000     .0208038    .0297784
       _cons |  -1.816455   .2798724    -6.49   0.000    -2.364995   -1.267916
------------------------------------------------------------------------------
```

Variable	Mean	Marginal effect
MALE	0.4841	−0.0467
AGE	32.52	0.0110
S	13.31	−0.0007
ASVABC	48.94	0.0097

The means of the explanatory variables, and their marginal effects evaluated at the means, are shown in the table.

- Discuss the conclusions one may reach, given the probit output and the table, commenting on their plausibility.
- The researcher considers including *CHILD*, a dummy variable defined to be 1 if the respondent had children, and 0 otherwise, as an explanatory variable. When she does this, its z-statistic is 33.65 and its marginal effect 0.5685. Discuss these findings.

10.7 Suppose that the time, t, required to complete a certain process has probability density function:

$$f(t) = \alpha \mathrm{e}^{-\alpha(t-\beta)} \quad \text{with } t > \beta > 0$$

and you have a sample of n observations with times $T_1, \ldots, T_n$.

Determine the maximum likelihood estimate of α, assuming that β is known.

A10.8 In Exercise 10.14 in the text, an event could occur with probability p. Given that the event occurred m times in a sample of n observations, the exercise required demonstrating that m/n was the ML estimator of p. Derive the LR statistic for the null hypothesis $p = p_0$. If $m = 40$ and $n = 100$, test the null hypothesis $p = 0.5$.

A10.9 For the variable in Exercise A10.8, derive the Wald statistic and test the null hypothesis $p = 0.5$.

10.5 Answers to the starred exercises in the textbook

10.1 [This exercise does not have a star in the text, but an answer to it is needed for comparison with the answer to Exercise 10.3.]

The output shows the result of an investigation of how the probability of a respondent obtaining a bachelor's degree from a four-year college is related to the score on *ASVABC*, using *EAWE* Data Set 21. *BACH* is a dummy variable equal to 1 for those with bachelor's degrees (years of schooling at least 16) and 0 otherwise. *ASVABC* is a measure of cognitive ability, scaled so that in the population it has mean 0 and standard deviation 1. Provide an interpretation of the coefficients. Explain why OLS is not a satisfactory estimation method for this kind of model.

```
. reg BACH ASVABC
----------------------------------------------------------------------------
    Source |       SS       df       MS              Number of obs =     500
-----------+------------------------------           F(  1,   498) =  123.14
     Model |  24.7674233     1  24.7674233           Prob > F      =  0.0000
  Residual |  100.160577   498  .201125656           R-squared     =  0.1983
-----------+------------------------------           Adj R-squared =  0.1966
     Total |     124.928   499  .250356713           Root MSE      =  .44847
----------------------------------------------------------------------------
      BACH |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+----------------------------------------------------------------
    ASVABC |   .2479312   .0223421    11.10   0.000     .2040348    .2918277
     _cons |   .4206845   .0209535    20.08   0.000     .3795163    .4618526
----------------------------------------------------------------------------
```

Answer:

The slope coefficient indicates that the probability of earning a bachelor's degree rises by 25 per cent for every additional unit of the *ASVABC* score. *ASVABC* is scaled so that one unit is one standard deviation and it has mean zero. While this may be realistic for a range of values of *ASVABC*, it is not for very low ones. Very few of those with scores in the low end of the spectrum earned bachelors degrees and variations in the *ASVABC* score would be unlikely to have an effect on the probability. The intercept literally indicates that an individual with average score would have a 42 per cent probability of earning a bachelor's degree.

However, the linear probability model predicts nonsense negative probabilities for all those with scores less of -1.70 or less. It also suffers from the problem that the standard errors and t and F tests are invalid because the disturbance term does not have a normal distribution. Its distribution is not even continuous, consisting of only two possible values for each value of *ASVABC*.

10.3 The output shows the results of fitting a logit regression for *BACH*, as defined in Exercise 10.1, with the iteration messages deleted. 48.8 per cent of the respondents earned bachelor's degrees.

```
. logit BACH ASVABC
------------------------------------------------------------------------------
Logistic regression                               Number of obs   =        500
                                                  LR chi2(1)      =     110.38
                                                  Prob > chi2     =     0.0000
Log likelihood = -291.23809                       Pseudo R2       =     0.1593
------------------------------------------------------------------------------
        BACH |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ASVABC |   1.240198   .1377998     9.00   0.000     .9701151     1.51028
       _cons |  -.4077999   .1088093    -3.75   0.000    -.6210623   -.1945375
------------------------------------------------------------------------------
```

The diagram shows the probability of earning a bachelor's degree as a function of *ASVABC*. It also shows the marginal effect function.

- With reference to the diagram, discuss the variation of the marginal effect of the *ASVABC* score implicit in the logit regression.
- Sketch the probability and marginal effect diagrams for the OLS regression in Exercise 10.1 and compare them with those for the logit regression.

Answer:

ASVABC is scaled so that it has a mean of zero. From the curve for the cumulative probability in the figure it can be seen that, for a respondent with mean score, the probability of graduating from college is about 40 per cent. For those one standard deviation above the mean, it is nearly 70 per cent. For those one standard deviation below, a little lower than 20 percent. Looking at the curve for the marginal probability, it can be seen that the marginal effect is greatest for those of average cognitive ability, and still quite high a standard deviation either way. For those two standard deviations above the mean, the marginal effect is low because most are going to college anyway. For those two standard deviations below, the effect is gain low, for the opposite reason.

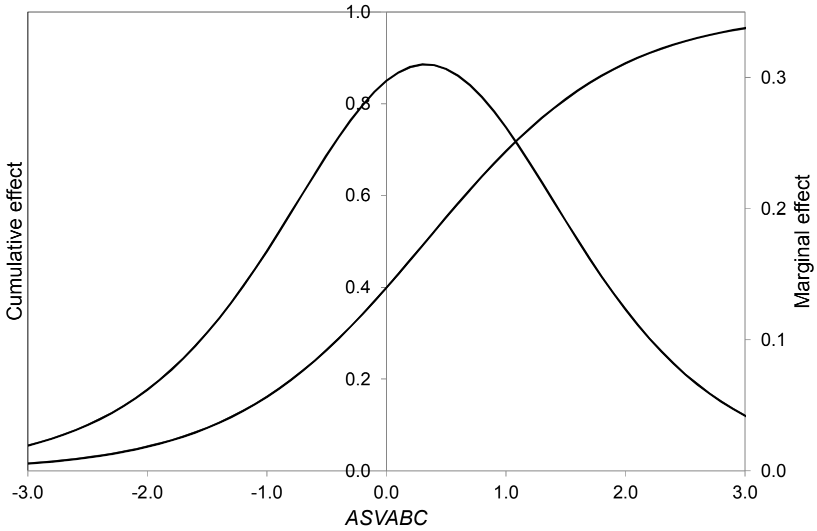

Figure 10.4: Scatter diagram of cumulative and marginal effects against *ASVABC*.

For the linear probability model in Exercise 10.1, the counterpart to the cumulative probability curve in the figure is a straight line using the regression result. In this example, the predictions of the linear probability model do not differ much from those of the logit model over the central range of the data. Its deficiencies become visible only at the extremes. The OLS counterpart to the marginal probability curve is a horizontal straight line at 0.25, showing that the marginal effect is somewhat underestimated in the central range and overestimated elsewhere.

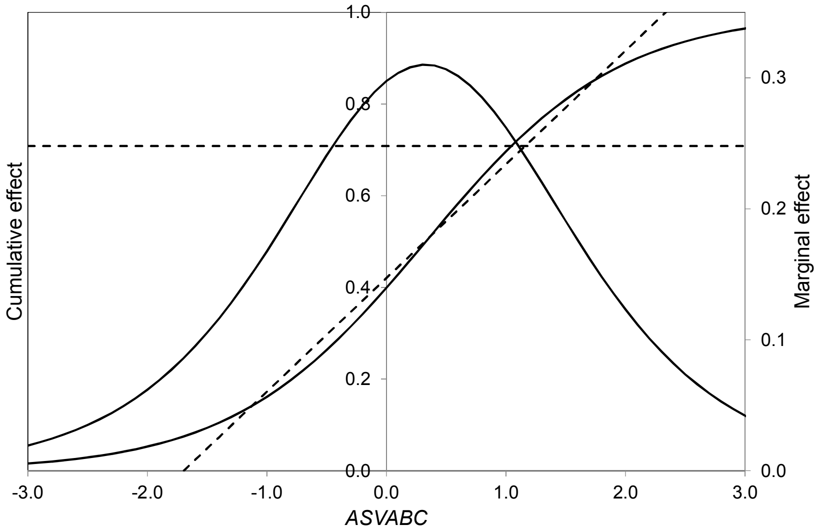

Figure 10.5: Scatter diagram of cumulative and marginal effects against *ASVABC*.

10.7 The following probit regression, with iteration messages deleted, was fitted using 2,108 observations on females in the National Longitudinal Survey of Youth using the *LFP2011* data set described in Exercise 10.2. The respondents were aged 27 to 31 and many of them were raising young families.

```
. probit WORKING S AGE CHILDL06 CHILDL16 MARRIED ETHBLACK ETHHISP if MALE==0
----------------------------------------------------------------------------
Probit regression                               Number of obs   =       2108
                                                LR chi2(7)      =     170.55
                                                Prob > chi2     =     0.0000
Log likelihood = -972.89229                     Pseudo R2       =     0.0806
----------------------------------------------------------------------------
   WORKING |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------+----------------------------------------------------------------
         S |   .1046085   .0127118     8.23   0.000     .0796939    .1295232
       AGE |  -.0029273   .0237761    -0.12   0.902    -.0495277     .043673
  CHILDL06 |  -.4490263     .08128    -5.52   0.000    -.6083322   -.2897204
  CHILDL16 |  -.3055774   .1060307    -2.88   0.004    -.5133938    -.097761
   MARRIED |  -.1286145   .0724189    -1.78   0.076    -.2705529    .0133239
  ETHBLACK |  -.1070784   .0861386    -1.24   0.214    -.2759069    .0617502
   ETHHISP |   .0364241   .0987625     0.37   0.712    -.1571468     .229995
     _cons |  -.1885982   .7046397    -0.27   0.789    -1.569667     1.19247
----------------------------------------------------------------------------
```

WORKING is a binary variable equal to 1 if the respondent was working in 2011, 0 otherwise. *CHILDL06* is a dummy variable equal to 1 if there was a child aged less than 6 in the household, 0 otherwise. *CHILDL16* is a dummy variable equal to 1 if there was a child aged less than 16, but no child less than 6, in the household, 0 otherwise. *MARRIED* is equal to 1 if the respondent was married with spouse present, 0 otherwise. The remaining variables are as described in Appendix B. The mean values of the variables are given in the output from the sum command:

```
. sum WORKING S AGE CHILDL06 CHILDL16 MARRIED ETHBLACK ETHHISP if MALE==0
-------------------------------------------------------------------
  Variable |       Obs        Mean    Std. Dev.       Min        Max
-----------+-------------------------------------------------------
   WORKING |      2108    .7988615    .4009465          0          1
         S |      2108    14.32922    2.882736          6         20
       AGE |      2108    28.99336    1.386405         27         31
  CHILDL06 |      2108    .4407021    .4965891          0          1
  CHILDL16 |      2108    .1465844    .3537751          0          1
   MARRIED |      2108     .420778    .4938011          0          1
  ETHBLACK |      2108    .1783681    .3829132          0          1
   ETHHISP |      2108    .1233397    .3289047          0          1
-------------------------------------------------------------------
```

Calculate the marginal effects and discuss whether they are plausible.

Answer:

The marginal effects are calculated in the table below. As might be expected, having a child aged less than 6 has a large adverse effect, very highly significant. Schooling also has a very significant effect, more educated mothers making use of their investment by tending to stay in the labour force. Age has a significant negative effect, the reason for which is not obvious (the respondents were aged 29–36 in 1994). Being black also has an adverse effect, the reason for which is likewise not obvious. (The *WORKING* variable is defined to be 1 if the individual has recorded hourly earnings of at least \$3. If the definition is tightened to including also the requirement that the employment status is employed, the latter effect is smaller, but still significant at the 5 per cent level.)

Variable	Mean	$\widehat{\beta}_2$	Mean×$\widehat{\beta}_2$	$f(Z)$	$\widehat{\beta}_2 \times f(Z)$
S	14.3292	0.1046	1.4990	0.2627	0.0275
AGE	28.9934	−0.0029	−0.0849	0.2627	−0.0008
$CHILD06$	0.4407	−0.4490	−0.1979	0.2627	−0.1180
$CHILDL16$	0.1466	−0.3056	−0.0448	0.2627	−0.0803
$MARRIED$	0.4208	−0.1286	−0.0541	0.2627	−0.0338
$ETHBLACK$	0.1784	−0.1071	−0.0191	0.2627	−0.0281
$ETHHISP$	0.1233	0.1233	0.0045	0.2627	0.0096
constant	1.0000	−0.1886	−0.1886		
Total			0.9141		

10.12 Show that the tobit model may be regarded as a special case of a selection bias model.

Answer:

The selection bias model may be written:

$$\begin{aligned} B_i^* &= \delta_1 + \sum_{j=2}^{m} \delta_j Q_{ji} + \varepsilon_i \\ Y_i^* &= \beta_1 \sum_{j=2}^{k} \beta_j X_{ji} + u_i \\ Y_i &= Y_i^* \qquad \text{for } B_i^* > 0 \\ & \quad Y_i \text{ is not observed for } B_i^* \le 0 \end{aligned}$$

where the Q variables determine selection. The tobit model is the special case where the Q variables are identical to the X variables and B^* is the same as Y^*.

10.14 An event is hypothesised to occur with probability p. In a sample of n observations, it occurred m times. Demonstrate that the maximum likelihood estimator of p is m/n.

Answer:

In each observation where the event did occur, the probability was p. In each observation where it did not occur, the probability was $(1-p)$. Since there were m of the former and $n-m$ of the latter, the joint probability was $p^m(1-p)^{n-m}$. Reinterpreting this as a function of p, given m and n, the log-likelihood function for p is:

$$\log L(p) - m \log p + (n-m)\log(1-p).$$

Differentiating with respect to p, we obtain the first-order condition for a minimum:

$$\frac{\mathrm{d} \log L(p)}{\mathrm{d}p} = \frac{m}{p} - \frac{n-m}{1-p} = 0.$$

This yields $p = m/n$. We should check that the second differential is negative and that we have therefore found a maximum. The second differential is:

$$\frac{\mathrm{d}^2 \log L(p)}{\mathrm{d}p^2} = -\frac{m}{p^2} - \frac{n-m}{(1-p)^2}.$$

Evaluated at $p = m/n$:

$$\frac{\mathrm{d}^2 \log L(p)}{\mathrm{d}p^2} = -\frac{n^2}{m} - \frac{n-m}{\left(1-\frac{m}{n}\right)^2} = -n^2\left(\frac{1}{m} + \frac{1}{n-m}\right).$$

This is negative, so we have indeed chosen the value of p that maximises the probability of the outcome.

10.18 Returning to the example of the random variable X with unknown mean μ and variance σ^2, the log-likelihood for a sample of n observations was given by equation (10.36):

$$\log L = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 + \frac{1}{\sigma^2}\left(-\frac{1}{2}(X_1-\mu)^2 - \cdots - \frac{1}{2}(X_n-\mu)^2\right).$$

The first-order condition forμ produced the ML estimator of μ and the first order condition for σ then yielded the ML estimator for σ. Often, the variance is treated as the primary dispersion parameter, rather than the standard deviation. Show that such a treatment yields the same results in the present case. Treat σ^2 as a parameter, differentiate $\log L$ with respect to it, and solve.

Answer:

$$\frac{\partial \log L}{\partial\sigma^2} = -\frac{n}{2\sigma^2} - \frac{1}{\sigma^4}\left(-\frac{1}{2}(X_1-\mu)^2 - \cdots - \frac{1}{2}(X_n-\mu)^2\right).$$

Hence:

$$\widehat{\sigma}^2 = \frac{1}{n}\left((X_1-\mu)^2 + \cdots + (X_n-\mu)^2\right)$$

as before. The ML estimator of μ is $\overline{X}$ as before.

10.19 In Exercise 10.7, $\log L_0$ is -1058.17. Compute the pseudo-R^2 and confirm that it is equal to that reported in the output.

Answer:

As defined in equation (10.48):

$$\text{pseudo-}R^2 = 1 - \frac{\log L}{\log L_0} = 1 - \frac{-972.8923}{-1058.17} = 0.0806$$

as appears in the output.

10.20 In Exercise 10.7, compute the likelihood ratio statistic $2(\log L - \log L_0)$, confirm that it is equal to that reported in the output, and perform the likelihood ratio test.

Answer:

The likelihood ratio statistic is $2(-972.89 + 1058.17) = 170.56$, which is that reported in the output, apart from rounding error in the last digit. Under the null hypothesis that the coefficients of the explanatory variables are all jointly equal to 0, this is distributed as a chi-squared statistic with degrees of freedom equal to the number of explanatory variables, in this case 7. The critical value of chi-squared at the 0.1 per cent significance level with 7 degrees of freedom is 24.32, and so we reject the null hypothesis at that level.

10.6 Answers to the additional exercises

A10.1 In the case of *FDHO* there were no non-purchasing households and so it was not possible to undertake the analysis.

The results for the logit analysis and the probit analysis were very similar. The linear probability model also yielded similar results for most of the commodities, the coefficients being similar to the logit and probit marginal effects and the t statistics being of the same order of magnitude as the z statistics for the logit and probit.

Most of the effects seem plausible with simple explanations. The total expenditure of the household and the size of the household were both highly significant factors in the decision to make a purchase for most categories of expenditure. The main exception, *TOB*. was instead influenced (negatively: survival bias?) by the age of the reference individual and, unsurprisingly, by his or her education.

Linear probability model, dependent variable *CATBUY*

		EXPPC$\times 10^{-4}$		*SIZE*$\times 10^{-2}$		*REFAGE*$\times 10^{-2}$		*COLLEGE*		Cases with probability	
	n	$\widehat{\beta}_2$	t	$\widehat{\beta}_3$	t	$\widehat{\beta}_4$	t	$\widehat{\beta}_5$	t	< 0	> 1
ADM	2,815	0.38	20.41	4.00	9.54	−0.34	−9.92	0.22	17.74	0	44
CLOT	4,500	0.33	18.74	5.38	13.61	−0.35	−10.72	0.05	4.12	0	144
DOM	1,661	0.30	17.37	4.18	10.78	0.16	5.08	0.09	7.99	0	181
EDUC	561	0.13	11.83	3.13	12.38	−0.12	−5.80	0.05	6.01	612	0
ELEC	5,828	0.08	7.33	2.71	11.09	0.16	7.76	0.02	2.07	0	254
FDAW	5,102	0.23	14.57	2.23	6.41	−0.27	−9.56	0.11	10.85	0	223
*FDHO**	6,334										
FOOT	1,827	0.28	15.83	5.93	14.81	−0.22	−6.65	0.01	1.01	0	4
FURN	487	0.14	13.47	1.65	6.87	−0.07	−3.74	0.01	1.66	149	0
GASO	5,710	0.09	7.70	3.23	12.07	−0.00	−0.14	0.07	8.61	0	331
HEAL	4,802	0.21	12.82	3.18	8.77	0.82	27.46	0.11	9.82	0	406
HOUS	6,223	0.03	5.24	0.52	4.36	0.04	4.44	0.01	2.30	0	484
LIFE	1,253	0.35	15.82	3.91	11.02	0.19	8.36	0.04	3.49	0	1
LOCT	692	0.04	3.42	−0.23	−0.80	−0.15	−6.38	0.00	0.42	0	0
MAPP	399	0.10	10.34	1.59	7.23	−0.00	−0.01	−0.01	−1.54	0	0
PERS	3,817	0.30	15.56	4.55	10.53	0.29	8.19	0.12	9.28	0	66
READ	2,287	0.25	13.48	2.52	5.98	0.37	10.76	0.16	13.03	0	10
SAPP	1,037	0.20	13.80	2.86	8.61	−0.03	−1.12	0.03	3.30	0	0
TELE	5,788	0.07	6.29	3.52	14.09	0.31	15.12	0.01	1.65	0	455
TEXT	992	0.19	13.25	2.45	7.50	−0.03	−1.22	0.04	3.84	0	0
TOB	1,155	−0.01	−0.54	0.24	0.69	−0.17	−5.90	−0.10	−9.16	0	0
TOYS	2,504	0.24	12.14	6.26	14.36	−0.13	−3.58	0.06	4.70	0	4
TRIP	516	0.23	21.63	0.93	3.88	−0.03	−1.39	0.03	4.58	415	0

**FDHO* had no observations with zero expenditure.

Logit model, dependent variable *CATBUY*

		$EXPPC \times 10^{-4}$		$SIZE \times 10^{-2}$		$REFAGE \times 10^{-2}$		$COLLEGE$	
	n	$\widehat{\beta}_2$	z	$\widehat{\beta}_3$	z	$\widehat{\beta}_4$	z	$\widehat{\beta}_5$	z
ADM	2,815	2.06	18.34	20.02	10.04	−1.69	10.02	1.00	16.52
CLOT	4,500	2.51	17.22	32.00	13.44	−1.72	−9.92	0.18	2.98
DOM	1,661	1.50	15.28	22.50	10.55	0.91	4.99	0.54	8.01
EDUC	561	1.38	11.60	35.93	12.32	−2.22	−7.14	0.81	6.99
ELEC	5,828	1.63	7.28	44.17	10.57	2.03	7.48	0.19	1.89
FDAW	5,102	2.71	14.40	17.42	6.78	−1.79	−8.99	0.63	9.16
FDHO	6,334								
FOOT	1,827	1.39	14.69	29.17	14.24	−1.25	−7.00	0.08	1.23
FURN	487	1.43	12.00	21.16	6.66	−1.28	−4.17	0.28	2.46
GASO	5,710	1.50	7.50	47.81	11.71	0.16	0.66	0.71	7.87
HEAL	4,802	2.29	13.58	21.11	9.12	5.22	24.36	0.59	8.61
HOUS	6,223	4.31	5.78	37.81	4.81	2.42	4.27	0.35	1.76
LIFE	1,253	1.38	13.94	24.61	10.71	1.28	6.33	0.27	3.71
LOCT	692	0.41	3.50	−1.75	−0.60	−1.57	−6.35	0.05	0.51
MAPP	399	1.21	9.65	23.27	5.89	−0.05	−0.16	−0.13	−1.11
PERS	3,817	1.78	15.07	21.91	10.92	1.30	8.11	0.48	8.46
READ	2,287	1.18	12.35	11.97	5.97	1.77	10.61	0.77	12.64
SAPP	1,037	1.24	12.47	19.99	8.37	−0.29	−1.37	0.29	3.71
TELE	5,788	1.24	6.20	51.87	12.34	3.82	13.66	0.18	1.78
TEXT	992	1.20	11.97	17.77	7.28	−0.31	−1.44	0.34	4.27
TOB	1,155	−0.07	−0.64	1.28	0.55	−1.17	−5.85	−0.62	−8.95
TOYS	2,504	1.04	11.53	27.08	13.84	−0.59	−3.69	0.27	4.70
TRIP	516	1.92	15.76	9.60	2.62	−0.42	−1.41	0.75	5.92

Probit model, dependent variable *CATBUY*

		$EXPPC \times 10^{-4}$		$SIZE \times 10^{-2}$		$REFAGE \times 10^{-2}$		$COLLEGE$	
	n	$\widehat{\beta}_2$	z	$\widehat{\beta}_3$	z	$\widehat{\beta}_4$	z	$\widehat{\beta}_5$	z
ADM	2,815	1.17	19.26	11.97	9.93	−1.01	−10.03	0.61	16.96
CLOT	4,500	1.34	18.00	18.37	13.62	−1.03	−10.00	0.12	3.31
DOM	1,661	0.89	15.77	13.35	10.52	0.53	5.00	0.31	7.95
EDUC	561	0.78	11.88	19.78	12.61	−1.15	−7.36	0.40	7.02
ELEC	5,828	0.71	7.18	19.93	10.53	0.96	7.17	0.10	2.03
FDAW	5,102	1.37	14.87	9.53	6.72	−1.03	−9.08	0.37	9.50
FDHO	6,334								
FOOT	1,827	0.82	15.39	17.60	14.43	−0.74	−6.98	0.05	1.29
FURN	487	0.80	12.45	11.37	6.83	−0.63	−4.15	0.12	2.24
GASO	5,710	0.61	7.37	21.79	11.79	0.08	0.60	0.40	8.43
HEAL	4,802	1.18	13.94	11.97	9.11	3.05	25.25	0.34	8.56
HOUS	6,223	1.33	5.76	14.17	4.56	0.98	4.22	0.19	2.26
LIFE	1,253	0.81	14.78	14.40	10.74	0.76	6.56	0.15	3.69
LOCT	692	0.21	3.30	−0.80	−0.54	−0.79	−6.26	0.02	0.50
MAPP	399	0.67	9.94	12.10	7.00	−0.03	−0.17	−0.07	−1.32
PERS	3,817	0.97	15.47	12.93	10.79	0.80	8.15	0.31	8.81
READ	2,287	0.70	12.74	7.14	5.86	1.07	10.63	0.47	12.87
SAPP	1,037	0.73	12.95	11.49	8.42	−0.15	−1.28	0.15	3.63
TELE	5,788	0.55	6.11	24.85	12.54	1.91	13.66	0.10	2.01
TEXT	992	0.71	12.53	10.21	7.33	−0.18	−1.46	0.18	4.16
TOB	1,155	−0.05	−0.79	0.84	0.63	−0.67	−5.86	−0.35	−8.89
TOYS	2,504	0.62	11.91	16.57	14.04	−0.37	−3.72	0.17	4.77
TRIP	516	1.06	16.91	4.84	2.66	−0.21	−1.42	0.35	5.93

Marginal effects, linear probability model, logit and probit						
	$EXPPC4 \times 10^{-4}$			$SIZE \times 10^{-2}$		
	LPM	logit	probit	LPM	logit	probit
ADM	0.38	0.51	0.46	4.00	4.93	4.72
CLOT	0.33	0.48	0.44	5.38	6.14	6.04
DOM	0.30	0.28	0.28	4.18	4.21	4.25
EDUC	0.13	0.09	0.10	3.13	2.24	2.57
ELEC	0.08	0.10	0.09	2.71	2.73	2.66
FDAW	0.23	0.36	0.34	2.23	2.32	2.37
FDHO						
FOOT	0.28	0.28	0.28	5.93	5.82	5.89
FURN	0.14	0.09	0.10	1.65	1.32	1.48
GASO	0.09	0.11	0.09	3.23	3.47	3.35
HEAL	0.21	0.35	0.33	3.18	3.23	3.34
HOUS	0.03	0.04	0.04	−0.23	−0.17	−0.15
LIFE	0.35	0.21	0.22	3.91	3.72	3.86
LOCT	0.04	0.04	0.04	−0.23	−0.17	−0.15
MAPP	0.10	0.07	0.08	1.59	1.27	1.39
PERS	0.30	0.42	0.37	4.55	5.18	4.96
READ	0.25	0.27	0.26	2.52	2.73	2.65
SAPP	0.20	0.16	0.17	2.86	2.60	2.74
TELE	0.07	0.08	0.07	3.52	3.14	3.29
TEXT	0.19	0.15	0.16	2.45	2.23	2.36
TOB	−0.01	−0.01	−0.01	0.24	0.19	0.22
TOYS	0.24	0.25	0.24	6.26	6.45	6.36
TRIP	0.23	0.11	0.13	0.93	0.58	0.61

Marginal effects, linear probability model, logit and probit

	$REFAGE\times 10^{-2}$			$COLLEGE$		
	LPM	logit	probit	LPM	logit	probit
ADM	−0.34	−0.42	−0.40	0.22	0.24	0.24
$CLOT$	−0.35	−0.33	−0.34	0.05	0.04	0.04
DOM	0.16	0.17	0.17	0.09	0.10	0.10
$EDUC$	−0.12	−0.14	−0.15	0.05	0.05	0.05
$ELEC$	0.16	0.13	0.13	0.02	0.01	0.01
$FDAW$	−0.27	−0.24	−0.26	0.11	0.08	0.09
$FDHO$						
$FOOT$	−0.22	−0.25	−0.25	0.01	0.02	0.02
$FURN$	−0.07	−0.08	−0.08	0.01	0.02	0.02
$GASO$	−0.00	0.01	0.01	0.07	0.05	0.06
$HEAL$	0.82	0.80	0.85	0.11	0.09	0.09
$HOUS$	0.04	0.02	0.03	0.01	0.00	0.01
$LIFE$	0.19	0.19	0.20	0.04	0.04	0.04
$LOCT$	−0.15	−0.15	−0.15	0.00	0.00	0.00
$MAPP$	−0.00	0.00	0.00	−0.01	−0.01	−0.01
$PERS$	0.29	0.31	0.31	0.12	0.11	0.12
$READ$	0.37	0.40	0.40	0.16	0.18	0.17
$SAPP$	−0.03	−0.04	−0.04	0.03	0.04	0.04
$TELE$	0.31	0.23	0.25	0.01	0.01	0.01
$TEXT$	−0.03	−0.04	−0.04	0.04	0.04	0.04
TOB	−0.17	−0.17	−0.17	−0.10	−0.09	−0.09
$TOYS$	−0.13	−0.14	−0.14	0.06	0.06	0.06
$TRIP$	−0.03	−0.03	−0.03	0.03	0.04	0.04

A10.2 The finding that the marginal effect of educational attainment was lower for males than for females over most of the range $S \geq 9$ is plausible because the probability of working is much closer to 1 for males than for females for $S \geq 9$, and hence the possible sensitivity of the participation rate to S is smaller.

The explanation of the finding that the marginal effect of educational attainment decreases with educational attainment for both males and females over the range $S \geq 9$ is similar. For both sexes, the greater is S, the greater is the participation rate, and hence the smaller is the scope for it being increased by further education.

The OLS estimates of the marginal effect of educational attainment are given by the slope coefficients and they are very similar to the logit estimates at the mean, the reason being that most of the observations on S are confined to the middle part of the sigmoid curve where it is relatively linear.

A10.3 *Discuss whether the relationships indicated by the probability and marginal effect curves appear to be plausible.*

The probability curve indicates an inverse relationship between schooling and the probability of being obese. This seems entirely plausible. The more educated tend to have healthier lifestyles, including eating habits. Over the relevant range, the marginal effect falls a little in absolute terms (is less negative) as schooling increases. This is in keeping with the idea that further schooling may have less effect on the highly educated than on the less educated (but the difference is not large).

Add the probability function and the marginal effect function for the LPM to the diagram. Explain why you drew them the way you did.

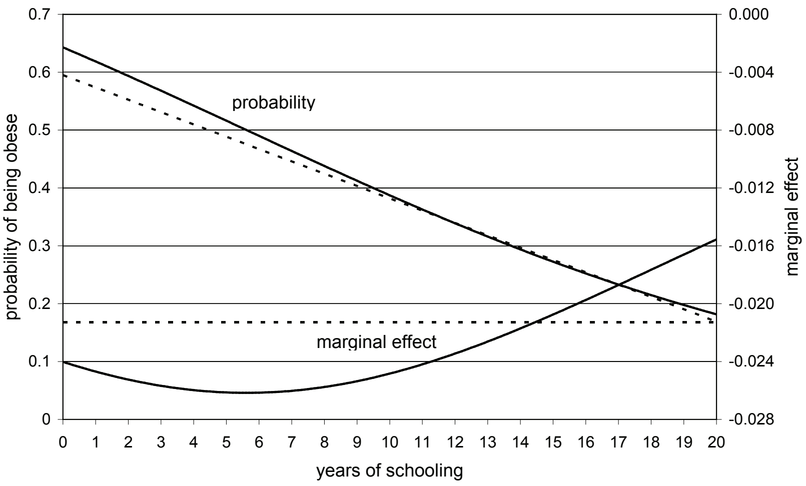

Figure 10.6: Scatter diagram of probability of being obese and marginal effect against years of schooling.

The estimated probability function for the LPM is just the regression equation and the marginal effect is the coefficient of S. They are shown as the dashed lines in the diagram.

The logit model is considered to have several advantages over the LPM. Explain what these advantages are. Evaluate the importance of the advantages of the logit model in this particular case.

The disadvantages of the LPM are (1) that it can give nonsense fitted values (predicted probabilities greater than 1 or less than 0); (2) the disturbance term in observation i must be equal to either $-1 - F(Z_i)$ (if the dependent variable is equal to 1) or $-F(Z_i)$ (if the dependent variable is equal to 0) and so it violates the usual assumption that the disturbance term is normally distributed, although this may not matter asymptotically; (3) the disturbance term will be heteroskedastic because Z_i is different for different observations; (4) the LPM implicitly assumes that the marginal effect of each explanatory variable is constant over its entire range, which is often intuitively unappealing.

In this case, nonsense predictions are clearly not an issue. The assumption of a constant marginal effect does not seem to be a problem either, given the approximate linearity of the logit $F(Z)$.

The LPM is fitted using OLS. Explain how, instead, it might be fitted using maximum likelihood estimation:

Write down the probability of being obese for any obese individual, given S_i for that individual, and write down the probability of not being obese for any non-obese *individual, again given S_i for that individual.*

Obese: $p_i^O = \beta_1 + \beta_2 S_i$; not obese: $p_i^{NO} = 1 - \beta_1 - \beta_2 S_i$.

Write down the likelihood function for this sample of 164 obese individuals and 376 non-obese individuals.

$$L(\beta_1, \beta_2 \,|\, \text{data}) = \prod_{\text{OBESE}} p_i^O \prod_{\text{NOT OBESE}} p_i^{NO} = \prod_{\text{OBESE}} (\beta_1 + \beta_2 S_i) \prod_{\text{NOT OBESE}} (1 - \beta_1 - \beta_2 S_i).$$

*Explain how one would use this function to estimate the parameters. [***Note:** *You are not expected to attempt to derive the estimators of the parameters.]*

You would use some algorithm to find the values of β_1 and β_2 that maximises the function.

Explain whether your maximum likelihood estimators will be the same or different from those obtained using least squares.

Least squares involves finding the extremum of a completely different expression and will therefore lead to different estimators.

10.4 *Explain how one may derive the marginal effects of the explanatory variables on the probability of having a child less than 6 in the household, and calculate for both males and females the marginal effects at the means of AGE and S.*

Since p is a function of Z, and Z is a linear function of the X variables, the marginal effect of X_j is:

$$\frac{\partial p}{\partial X_j} = \frac{\mathrm{d}p}{\mathrm{d}Z}\frac{\partial Z}{\partial X_j} = \frac{\mathrm{d}p}{\mathrm{d}Z}\beta_j$$

where β_j is the coefficient of X_j in the expression for Z. In the case of probit analysis, $p = F(Z)$ is the cumulative standardised normal distribution. Hence $\mathrm{d}p/\mathrm{d}Z$ is just the standardised normal distribution.

For males, this is 0.368 when evaluated at the means. Hence the marginal effect of AGE is $0.368 \times -0.137 = -0.050$ and that of S is $0.368 \times 0.132 = 0.049$. For females the corresponding figures are $0.272 \times -0.154 = -0.042$ and $0.272 \times 0.094 = 0.026$, respectively. So for every extra year of age, the probability is reduced by 5.0 per cent for males and 4.2 per cent for females. For every extra year of schooling, the probability increases by 4.9 per cent for males and 2.6 per cent for females.

Explain whether the signs of the marginal effects are plausible. Explain whether you would expect the marginal effect of schooling to be higher for males or for females.

Yes. Given that the cohort is aged 35–42, the respondents have passed the age at which most adults start families, and the older they are, the less likely they are to have small children in the household. At the same time, the more educated the respondent, the more likely he or she is to have started having a family relatively late, so the positive effect of schooling is also plausible. However, given the age of the cohort, it is likely to be weaker for females than for males, given that most females intending to have families will have started them by this time, irrespective of their education.

At a seminar someone asks the researcher whether the marginal effect of S is significantly different for males and females. The researcher does not know how to test whether the difference is significant and asks you for advice. What would you say?

Fit a probit regression for the combined sample, adding a male intercept dummy and male slope dummies for AGE and S. Test the coefficient of the slope dummy for S.

10.5 The Z function will be of the form:

$$Z = \beta_1 + \beta_2 A + \beta_3 S + \beta_4 AS$$

so the marginal effects are:

$$\frac{\partial p}{\partial A} = \frac{\mathrm{d}p}{\mathrm{d}Z}\frac{\partial Z}{\partial A} = f(Z)(\beta_2 + \beta_4 S)$$

and:

$$\frac{\partial p}{\partial S} = \frac{\mathrm{d}p}{\mathrm{d}Z}\frac{\partial Z}{\partial S} = f(Z)(\beta_3 + \beta_4 A).$$

Both factors depend on the values of A and/or S, but the marginal effects could be evaluated for a representative individual using the mean values of A and S in the sample.

A10.6 *Discuss the conclusions one may reach, given the probit output and the table, commenting on their plausibility.*

Being male has a small but highly significant negative effect. This is plausible because males tend to marry later than females and the cohort is still relatively young.

Age has a highly significant positive effect, again plausible because older people are more likely to have married than younger people.

Schooling has no apparent effect at all. It is not obvious whether this is plausible.

Cognitive ability has a highly significant positive effect. Again, it is not obvious whether this is plausible.

The researcher considers including CHILD, a dummy variable defined to be 1 if the respondent had children, and 0 otherwise, as an explanatory variable. When she does this, its z-statistic is 33.65 and its marginal effect 0.5685. Discuss these findings.

Obviously one would expect a high positive correlation between being married and having children and this would account for the huge and highly significant coefficient. However getting married and having children are often a joint decision, and accordingly it is simplistic to suppose that one characteristic is a determinant of the other. The finding should not be taken at face value.

A10.7 *Determine the maximum likelihood estimate of α, assuming that β is known.*

The log-likelihood function is:

$$\log L(\alpha \,|\, \beta, T_1, \ldots, T_n) = n \log \alpha - \alpha \sum (T_i - \beta).$$

Setting the first derivative with respect to α equal to zero, we have:

$$\frac{n}{\widehat{\alpha}} - \sum (T_i - \beta) = 0$$

and hence:

$$\widehat{\alpha} = \frac{1}{\overline{T} - \beta}.$$

The second derivative is $-n/\widehat{\alpha}^2$, which is negative, confirming we have maximised the loglikelihood function.

A10.8 From the solution to Exercise 10.14, the log-likelihood function for p is:

$$\log L(p) = m \log p + (n - m) \log(1 - p).$$

Thus the LR statistic is:

$$\begin{aligned} LR &= 2\left(\left(m \log \frac{m}{n} + (n-m) \log\left(1 - \frac{m}{n}\right)\right) - (m \log p_0 + (n-m)\log(1-p_0))\right) \\ &= 2\left(m \log\left(\frac{m/n}{p_0}\right) + (n-m)\log\left(\frac{1-m/n}{1-p_0}\right)\right). \end{aligned}$$

If $m = 40$ and $n = 100$, the LR statistic for $H_0 : p = 0.5$ is:

$$\mathrm{LR} = 2\left(40 \log\left(\frac{0.4}{0.5}\right) + 60 \log\left(\frac{0.6}{0.5}\right)\right) = 4.03.$$

We would reject the null hypothesis at the 5 per cent level (critical value of chi-squared with one degree of freedom 3.84) but not at the 1 per cent level (critical value 6.64).

A10.9 The first derivative of the log-likelihood function is:

$$\frac{\mathrm{d} \log L(p)}{\mathrm{d}p} = \frac{m}{p} - \frac{n-m}{1-p} = 0$$

and the second differential is:

$$\frac{\mathrm{d} \log L(p)}{\mathrm{d}p^2} = -\frac{m}{p^2} - \frac{n-m}{(1-p)^2}.$$

Evaluated at $p = m/n$:

$$\frac{\mathrm{d} \log L(p)}{\mathrm{d}p^2} = -\frac{n^2}{m} - \frac{n-m}{\left(1-\frac{m}{n}\right)^2} = -n^2\left(\frac{1}{m} + \frac{1}{n-m}\right) = -\frac{n^3}{m(n-m)}.$$

The variance of the ML estimate is given by:

$$\left(-\frac{\mathrm{d} \log L(p)}{\mathrm{d}p^2}\right)^{-1} = \left(\frac{n^3}{m(n-m)}\right)^{-1} = \frac{m(n-m)}{n^3}.$$

The Wald statistic is therefore:

$$\frac{\left(\frac{m}{n} - p_0\right)^2}{\frac{m(n-m)}{n^3}} = \frac{\left(\frac{m}{n} - p_0\right)^2}{\frac{1}{n}\frac{m}{n}\frac{n-m}{n}}.$$

Given the data, this is equal to:

$$\frac{(0.4 - 0.5)^2}{\frac{1}{100} \times 0.4 \times 0.6} = 4.17.$$

Under the null hypothesis this has a chi-squared distribution with one degree of freedom, and so the conclusion is the same as in Exercise A.8.