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

Economists are often interested in the factors behind the decision-making of individuals or enterprises. Examples are:

- Why do some people go to college while others do not?
- Why do some women enter the labor force while others do not?
- Why do some people buy houses while others rent?
- Why do some people migrate while others stay put?

The models that have been developed are known as binary choice or qualitative response models with the outcome, which we will denote Y, being assigned a value of 1 if the event occurs and 0 otherwise. Models with more than two possible outcomes have been developed, but we will restrict our attention to binary choice. The linear probability model apart, binary choice models are fitted using maximum likelihood estimation. The chapter ends with an introduction to this topic.

11.1 The linear probability model

The simplest binary choice model is the linear probability model where, as the name implies, the probability of the event occurring, p, is assumed to be a linear function of a set of explanatory variable(s):

$$p_i = p(Y_i = 1) = \beta_1 + \beta_2 X_i. \tag{11.1}$$

Graphically, the relationship is as shown in Figure 11.1, if there is just one explanatory variable. Of course p is unobservable. One has data only on the outcome, Y. In the linear probability model this is used as a dummy variable for the dependent variable.

As an illustration, we investigate the factors influencing graduating from high school. We will define a variable *GRAD* that is equal to 1 for those individuals who graduated, and 0 for those who dropped out, and we will regress it on *ASVABC*, the composite cognitive ability test score. The regression output

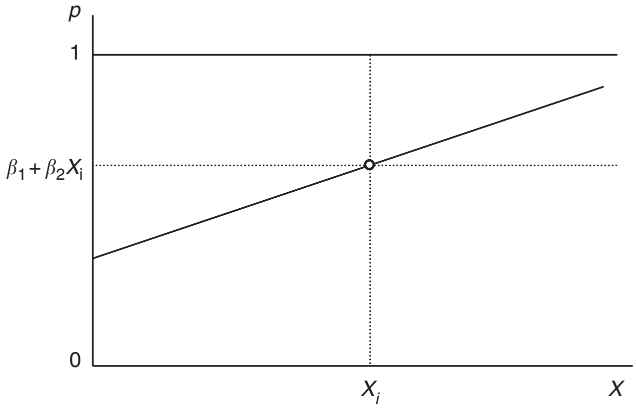


Figure 11.1 Linear probability model

Table 11.1

```
. reg GRAD ASVABC

  Source |       SS       df       MS              Number of obs =     570
---------+------------------------------           F(  1,   568) =  112.59
   Model |  7.13422753     1  7.13422753           Prob > F      =  0.0000
Residual |  35.9903339   568  .063363264           R-squared     =  0.1654
---------+------------------------------           Adj R-squared =  0.1640
   Total |  43.1245614   569  .07579009            Root MSE      =  .25172

------------------------------------------------------------------------------
    GRAD |      Coef.   Std. Err.       t     P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
  ASVABC |   .0121518   .0011452     10.611   0.000     .0099024    .0144012
   _cons |   .3081194   .0583932      5.277   0.000     .1934264    .4228124
```

shows the result of fitting this linear probability model, using *EAEF* Data Set 21 (Table 11.1).

The regression result suggests that the probability of graduating from high school increases by a proportion 0.012, that is, 1.2 percent, for every point increase in the *ASVABC* score. *ASVABC* is scaled so that it has mean 50 and standard deviation 10, so a one-standard deviation increase in the score would increase the probability of graduating by 12 percent. The intercept implies that if *ASVABC* were 0, the probability of graduating would be 31 percent. However the *ASVABC* score is scaled in such a way as to make its minimum about 20, and accordingly it is doubtful whether the interpretation should be taken at face value.

Unfortunately, the linear probability model has some serious defects. First, there are problems with the disturbance term. As usual, the value of the dependent variable Y_i in observation i has a nonstochastic component and a random component. The nonstochastic component depends on X_i and the parameters and is the expected value of Y_i given $X_i, E(Y_i|X_i)$. The random component is the disturbance term.

$$Y_i = E(Y_i|X_i) + u_i. \tag{11.2}$$

It is simple to compute the nonstochastic component in observation i because Y can take only two values. It is 1 with probability p_i and 0 with probability $(1 - p_i)$:

$$E(Y_i) = 1 \times p_i + 0 \times (1 - p_i) = p_i = \beta_1 + \beta_2 X_i. \tag{11.3}$$

The expected value in observation i is therefore $\beta_1 + \beta_2 X_i$. This means that we can rewrite the model as

$$Y_i = \beta_1 + \beta_2 X_i + u_i. \tag{11.4}$$

The probability function is thus also the nonstochastic component of the relationship between Y and X. It follows that, for the outcome variable Y_i to be equal to 1, as represented by the point A in Figure 11.2, the disturbance term must be equal to $(1 - \beta_1 - \beta_2 X_i)$. For the outcome to be 0, as represented by the point B, the disturbance term must be $(-\beta_1 - \beta_2 X_i)$. Thus the distribution of the disturbance term consists of just two specific values. It is not even continuous, never mind normal. This means that the standard errors and the usual test statistics are invalidated. For good measure, the two possible values of the disturbance term change with X, so the distribution is heteroscedastic as well. It can be shown that the population variance of u_i is $(\beta_1 + \beta_2 X_i)(1 - \beta_1 - \beta_2 X_i)$, and this varies with X_i.

The other problem is that the predicted probability may be greater than 1 or less than 0 for extreme values of X. In the example of graduating from high school, the regression equation predicts a probability greater than 1 for the 176 respondents with *ASVABC* scores greater than 56.

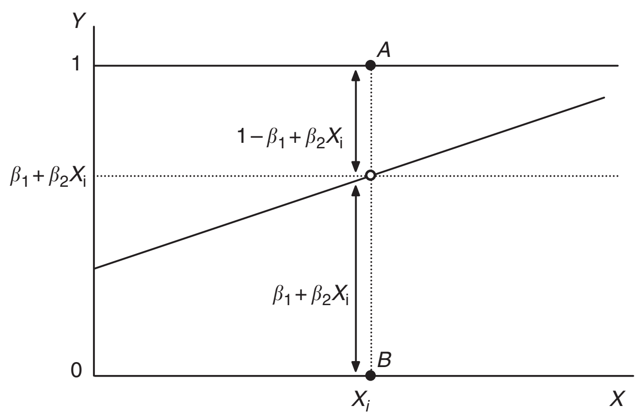


Figure 11.2 Disturbance term in the linear probability model

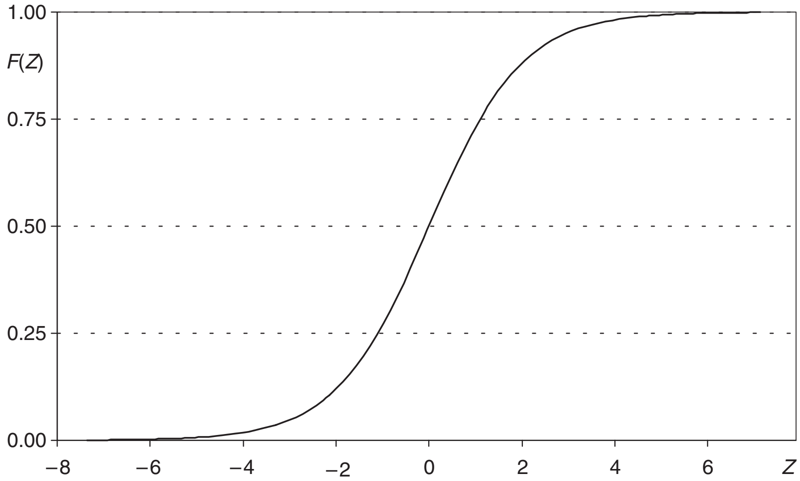


Figure 11.3 Logistic function

The first problem is dealt with by fitting the model with a technique known as maximum likelihood estimation, described in Section 11.6, instead of least squares. The second problem involves elaborating the model as follows. Define a variable Z that is a linear function of the explanatory variables. In the present case, since we have only one explanatory variable, this function is

$$Z_i = \beta_1 + \beta_2 X_i. \tag{11.5}$$

Next, suppose that p is a sigmoid (S-shaped) function of Z, for example as shown in Figure 11.3. Below a certain value of Z, there is very little chance of the individual graduating from high school. Above a certain value, the individual is almost certain to graduate. In between, the probability is sensitive to the value of Z.

This deals with the problem of nonsense probability estimates, but then there is the question of what should be the precise mathematical form of this function. There is no definitive answer to this. The two most popular forms are the logistic function, which is used in logit estimation, and the cumulative normal distribution, which is used in probit estimation. According to one of the leading authorities on the subject, Amemiya (1981), both give satisfactory results most of the time and neither has any particular advantage. We will start with the former.

11.2 Logit analysis

In logit estimation one hypothesizes that the probability of the occurrence of the event is determined by the function

$$p_i = F(Z_i) = \frac{1}{1 + e^{-Z_i}}. \tag{11.6}$$

This is the function shown in Figure 11.3. As Z tends to infinity, e^{-Z} tends to 0 and p has a limiting upper bound of 1. As Z tends to minus infinity, e^{-Z} tends to infinity and p has a limiting lower bound of 0. Hence there is no possibility of getting predictions of the probability being greater than 1 or less than 0.

The marginal effect of Z on the probability, which will be denoted $f(Z)$, is given by the derivative of this function with respect to Z:

$$f(Z) = \frac{dp}{dZ} = \frac{e^{-Z}}{(1+e^{-Z})^2}. \tag{11.7}$$

The function is shown in Figure 11.4. You can see that the effect of changes in Z on the probability is very small for large positive or large negative values of Z, and that the sensitivity of the probability to changes in Z is greatest at the midpoint value of 0.

In the case of the example of graduating from high school, the function is

$$p_i = \frac{1}{1+e^{-\beta_1-\beta_2 ASVABC_i}}. \tag{11.8}$$

If we fit the model, we get the output shown in Table 11.2.

The model is fitted by maximum likelihood estimation and, as the output indicates, this uses an iterative process to estimate the parameters.

The z statistics in the Stata output are approximations to t statistics and have nothing to do with the Z variable discussed in the text. (Some regression applications describe them as t statistics.) The z statistic for *ASVABC* is highly significant. How should one interpret the coefficients? To calculate the marginal effect of *ASVABC* on p we need to calculate $dp/dASVABC$. You could calculate the differential directly, but the best way to do this, especially if Z is a function

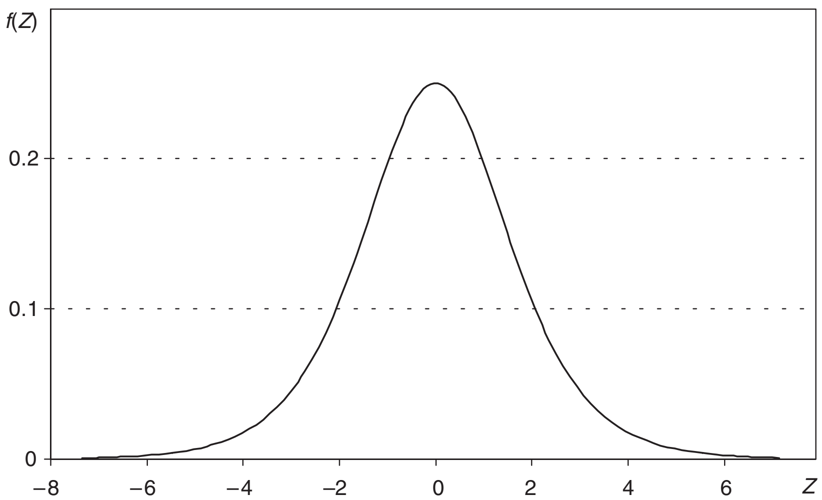


Figure 11.4 Marginal effect of Z on the probability

Table 11.2

```
. logit GRAD ASVABC

Iteration 0:  Log Likelihood =-162.29468
Iteration 1:  Log Likelihood =-132.97646
Iteration 2:  Log Likelihood =-117.99291
Iteration 3:  Log Likelihood =-117.36084
Iteration 4:  Log Likelihood =-117.35136
Iteration 5:  Log Likelihood =-117.35135

Logit Estimates                                   Number of obs =    570
                                                  chi2(1)       =  89.89
                                                  Prob > chi2   = 0.0000
Log Likelihood = -117.35135                       Pseudo R2     = 0.2769

------------------------------------------------------------------------------
    GRAD |      Coef.   Std. Err.      z     P>|z|      [95% Conf. Interval]
---------+--------------------------------------------------------------------
  ASVABC |   .1666022   .0211265     7.886   0.000      .1251951    .2080094
   _cons |  -5.003779   .8649213    -5.785   0.000     -6.698993   -3.308564
```

of more than one variable, is to break it up into two stages. p is a function of Z, and Z is a function of $ASVABC$, so

$$\frac{dp}{dASVABC} = \frac{dp}{dZ}.\frac{dZ}{dASVABC} = f(Z).\beta, \tag{11.9}$$

where $f(Z)$ is as defined above. The probability of graduating from high school, and the marginal effect, are plotted as functions of $ASVABC$ in Figure 11.5.

How can you summarize the effect of the $ASVABC$ score on the probability of graduating? The usual method is to calculate the marginal effect at the mean value of the explanatory variables. In this sample the mean value of $ASVABC$ was 50.15. For this value, Z is equal to 3.3514, and e^{-Z} is equal to 0.0350. Using this, $f(Z)$ is 0.0327 and the marginal effect is 0.0054:

$$f(Z)\beta_2 = \frac{e^{-Z}}{(1+e^{-Z})^2}\beta_2 = \frac{0.0350}{(1.0350)^2} \times 0.1666 = 0.0054. \tag{11.10}$$

In other words, at the sample mean, a one-point increase in $ASVABC$ increases the probability of going to college by 0.5 percent. This is a very small amount and the reason is that, for those with the mean $ASVABC$, the estimated probability of graduating is very high:

$$p = \frac{1}{1+e^{-Z}} = \frac{1}{1+0.0350} = 0.9661. \tag{11.11}$$

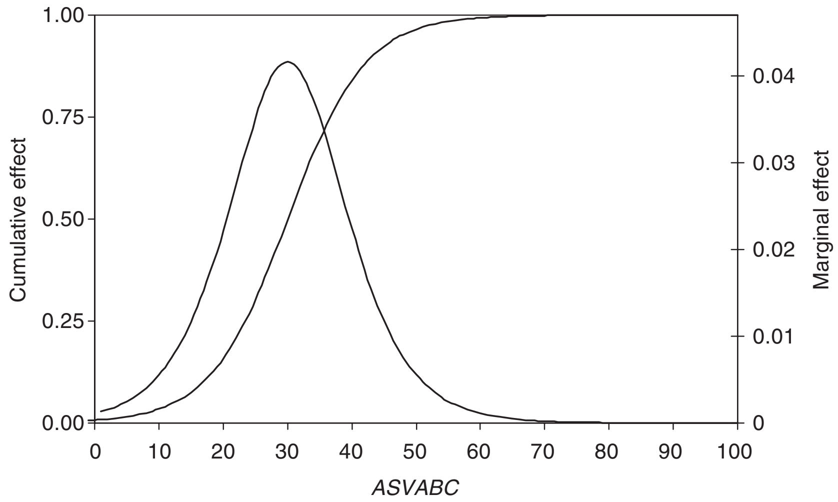


Figure 11.5 Cumulative and marginal effects of *ASVABC*

See also Figure 11.5. Of course we could calculate the marginal effect for other values of *ASVABC* if we wished and in this particular case it may be of interest to evaluate it for low *ASVABC*, where individuals are at greater risk of not graduating. For example, when *ASVABC* is 30, Z is $-0.0058, e^{-Z}$ is 1.0058, $f(Z)$ is 0.2500, and the marginal effect is 0.0417, or 4.2 percent. It is much higher because an individual with such a low score has only a 50 percent chance of graduating and an increase in *ASVABC* can make a substantial difference.

Generalization to more than one explanatory variable

Logit analysis is easily extended to the case where there is more than one explanatory variable. Suppose that we decide to relate graduating from high school to *ASVABC*, *SM*, the number of years of schooling of the mother, *SF*, the number of years of schooling of the father, and a dummy variable *MALE* that is equal to 1 for males, 0 for females. The Z variable becomes

$$Z = \beta_1 + \beta_2 ASVABC + \beta_3 SM + \beta_4 SF + \beta_5 MALE. \tag{11.12}$$

The corresponding regression output (with iteration messages deleted) is shown is Table 11.3.

The mean values of *ASVABC*, *SM*, *SF*, and *MALE* were as shown in Table 11.4, and hence the value of Z at the mean was 3.3380. From this one obtains 0.0355 for e^{-Z} and 0.0331 for $f(Z)$. The table shows the marginal effects, calculated by multiplying $f(Z)$ by the estimates of the coefficients of the logit regression.

According to the computations, a one-point increase in the *ASVABC* score increases the probability of going to college by 0.5 percent, every additional

Table 11.3

```
. logit GRAD ASVABC SM SF MALE
Logit Estimates                                     Number of obs =    570
                                                    chi2(4)       =  91.59
                                                    Prob > chi2   = 0.0000
Log Likelihood = -116.49968                         Pseudo R2     = 0.2822

------------------------------------------------------------------------------
    GRAD |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
  ASVABC |   .1563271   .0224382     6.967   0.000     .1123491    .2003051
      SM |   .0645542   .0773804     0.834   0.404    -.0871086     .216217
      SF |   .0054552   .0616822     0.088   0.930    -.1154397      .12635
    MALE |  -.2790915   .3601689    -0.775   0.438    -.9850095    .4268265
   _cons |   -5.15931    .994783    -5.186   0.000    -7.109049   -3.209571
```

Table 11.4 Logit estimation. Dependent variable: *GRAD*

Variable	Mean	b	Mean × b	$f(Z)$	$bf(Z)$
ASVABC	50.151	0.1563	7.8386	0.0331	0.0052
SM	11.653	0.0646	0.7528	0.0331	0.0021
SF	11.818	0.0055	0.0650	0.0331	0.0002
MALE	0.570	−0.2791	−0.1591	0.0331	−0.0092
Constant	1.000	−5.1593	−5.1593		
Total			3.3380		

year of schooling of the mother increases the probability by 0.2 percent, every additional year of schooling of the father increases the probability by a negligible amount, and being male reduces the probability by 0.9 percent. From the regression output it can be seen that the effect of *ASVABC* was significant at the 0.1 percent level but the effects of the parental education variables and the male dummy were insignificant.

Goodness of fit and statistical tests

There is no measure of goodness of fit equivalent to R^2 in maximum likelihood estimation. In default, numerous measures have been proposed for comparing alternative model specifications. Denoting the actual outcome in observation i as Y_i, with $Y_i = 1$ if the event occurs and 0 if it does not, and denoting the predicted probability of the event occurring $\hat{p}_i$, the measures include the following:

- the number of outcomes correctly predicted, taking the prediction in observation i as 1 if $\hat{p}_i$ is greater than 0.5 and 0 if it is less;

- the sum of the squared residuals $\sum_{i=1}^{n}(Y_i - \hat{p}_i)^2$;
- the correlation between the outcomes and predicted probabilities, $r_{Y_i \hat{p}_i}$.
- the pseudo-R^2 in the logit output, explained in Section 11.6.

Each of these measures has its shortcomings and Amemiya (1981) recommends considering more than one and comparing the results.

Nevertheless, the standard significance tests are similar to those for the standard regression model. The significance of an individual coefficient can be evaluated via its t statistic. However, since the standard error is valid only asymptotically (in large samples), the same goes for the t statistic, and since the t distribution converges on the normal distribution in large samples, the critical values of the latter should be used. The counterpart of the F test of the explanatory power of the model (H_0: all the slope coefficients are 0, H_1: at least one is nonzero) is a chi-squared test with the chi-squared statistic in the logit output distributed under H_0 with degrees of freedom equal to the number of explanatory variables. Details are provided in Section 11.6.

Exercises

11.1 Investigate the factors affecting going to college using your *EAEF* data set. Define a binary variable *COLLEGE* to be equal to 1 if $S > 12$ and 0 otherwise. Regress *COLLEGE* on *ASVABC*, *SM*, *SF*, and *MALE* (1) using ordinary least squares, and (2) using logit analysis. Calculate the marginal effects in the logit analysis and compare them with those obtained using OLS.

11.2* A researcher, using a sample of 2,868 individuals from the NLSY, is investigating how the probability of a respondent obtaining a bachelor's degree from a four-year college is related to the respondent's score on *ASVABC*. 26.7 percent of the respondents earned bachelor's degrees. *ASVABC* ranged from 22 to 65, with mean value 50.2, and most scores were in the range 40 to 60. Defining a variable *BACH* to be equal to 1 if the respondent has a bachelor's degree (or higher degree) and 0 otherwise, the researcher fitted the OLS regression (standard errors in parentheses):

$$BACH = -0.864 + 0.023 ASVABC. \quad R^2 = 0.21$$
$$(0.042) \quad (0.001)$$

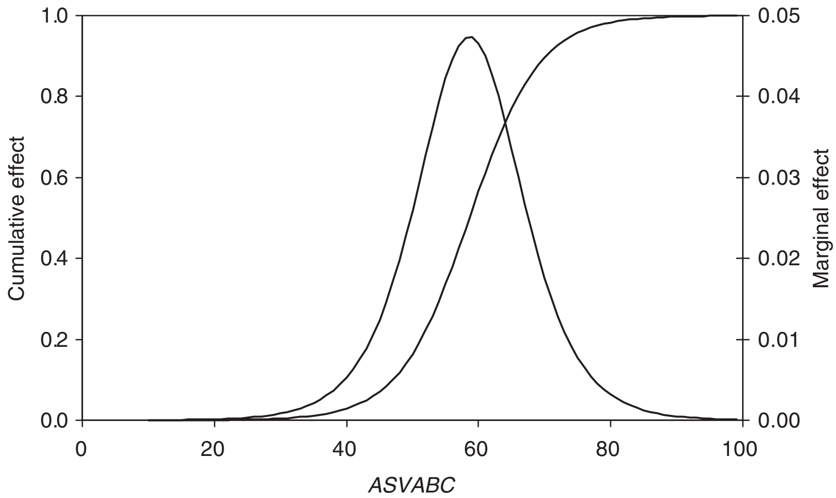
She also fitted the following logit regression:

$$Z = -11.103 + 0.189\, ASVABC,$$
$$(0.487) \quad (0.009)$$

where Z is the variable in the logit function. Using this regression, she plotted the probability and marginal effect functions shown in the diagram:

(a) Give an interpretation of the OLS regression and explain why OLS is not a satisfactory estimation method for this kind of model.

(b) With reference to the figure, discuss the variation of the marginal effect of the *ASVABC* score implicit in the logit regression and compare it with that in the OLS regression.



(c) Sketch the probability and marginal effect diagrams for the OLS regression and compare them with those for the logit regression. (In your discussion, make use of the information in the first paragraph of this question.)

11.3 Probit analysis

An alternative approach to the binary choice model is to use the cumulative standardized normal distribution to model the sigmoid relationship $F(Z)$. (A standardized normal distribution is one with mean 0 and unit variance.) As with logit analysis, you start by defining a variable Z that is a linear function of the variables that determine the probability:

$$Z = \beta_1 + \beta_2 X_2 + \cdots + \beta_k X_k. \tag{11.13}$$

$F(Z)$, the standardized cumulative normal distribution, gives the probability of the event occurring for any value of Z:

$$p_i = F(Z_i). \tag{11.14}$$

Maximum likelihood analysis is used to obtain estimates of the parameters. The marginal effect of X_i is $\partial p/\partial X_i$ which, as in the case of logit analysis, is best computed as

$$\frac{\partial p}{\partial X_i} = \frac{dp}{dZ} \cdot \frac{\partial Z}{\partial X_i} = f(Z).\beta_i. \tag{11.15}$$

Now since $F(Z)$ is the cumulative standardized normal distribution, $f(Z)$, its derivative, is just the standardized normal distribution itself:

$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z^2}. \tag{11.16}$$

Figure 11.6 plots $F(Z)$ and $f(Z)$ for probit analysis. As with logit analysis, the marginal effect of any variable is not constant. It depends on the value of $f(Z)$, which in turn depends on the values of each of the explanatory variables. To obtain a summary statistic for the marginal effect, the usual procedure is parallel to that used in logit analysis. You calculate Z for the mean values of the explanatory variables. Next you calculate $f(Z)$, as in (11.16). Then you calculate $f(Z)\beta_i$ to obtain the marginal effect of X_i.

This will be illustrated with the example of graduating from high school, using the same specification as in the logit regression. The regression output, with iteration messages deleted, is shown in Table 11.5.

The computation of the marginal effects at the sample means is shown in Table 11.6. Z is 1.8418 when evaluated at the mean values of the variables and $f(Z)$ is 0.0732. The estimates indicate that a one-point increase in the *ASVABC* score increases the probability of going to college by 0.6 percent, every additional year of schooling of the mother increases the probability by 0.3 percent, every additional year of schooling of the father increases the probability by a negligible amount, and being male reduces the probability by 1.4 percent. Generally logit and probit analysis yield similar marginal effects. However, the tails of the

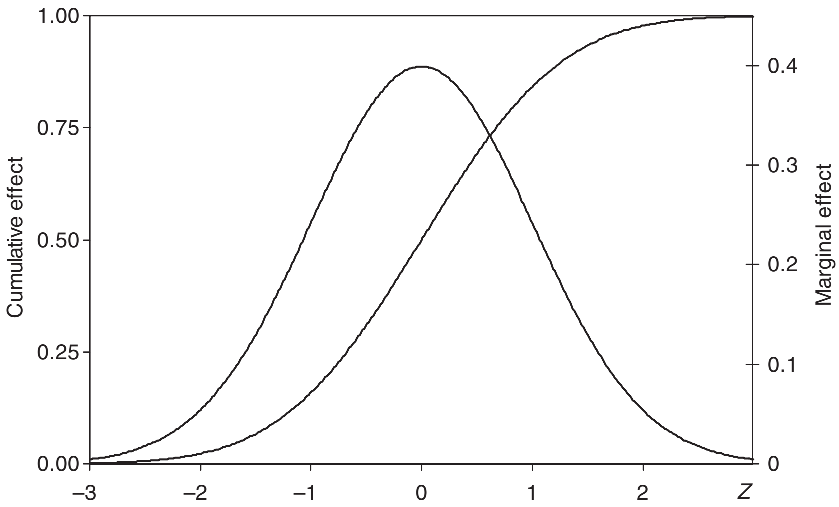


Figure 11.6 Cumulative and marginal normal effects of Z

Table 11.5

```
. probit GRAD ASVABC SM SF MALE

Probit Estimates                                  Number of obs =    570
                                                  chi2(4)       =  94.12
                                                  Prob > chi2   = 0.0000
Log Likelihood = -115.23672                       Pseudo R2     = 0.2900

--------------------------------------------------------------------------
    GRAD |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
---------+----------------------------------------------------------------
  ASVABC |   .0831963   .0117006     7.110   0.000     .0602635     .106129
      SM |   .0353463   .0425199     0.831   0.406    -.0479913    .1186838
      SF |   .0057229    .032375     0.177   0.860    -.0577309    .0691766
    MALE |  -.1883038   .1873426    -1.005   0.315    -.5554885     .178881
   _cons |  -2.702067   .5335551    -5.064   0.000    -3.747816   -1.656318
```

Table 11.6 Probit estimation. Dependent variable: *GRAD*

Variable	Mean	b	Mean × b	$f(Z)$	$bf(Z)$
ASVABC	50.151	0.0832	4.1726	0.0732	0.0061
SM	11.653	0.0353	0.4114	0.0732	0.0026
SF	11.818	0.0057	0.0674	0.0732	0.0004
MALE	0.570	−0.1883	−0.1073	0.0732	−0.0138
Constant	1.000	−2.7021	−2.7021		
Total			1.8418		

logit and probit distributions are different and they can give different results if the sample is unbalanced, with most of the outcomes similar and only a small minority different. This is the case in the present example because only 8 percent of the respondents failed to graduate, and in this case the estimates of the marginal effects are somewhat larger for the probit regression.

Exercises

11.3 Regress the variable *COLLEGE* defined in Exercise 11.1 on *ASVABC*, *MALE*, *SM*, and *SF* using probit analysis. Calculate the marginal effects and compare them with those obtained using OLS and logit analysis.

11.4* The following probit regression, with iteration messages deleted, was fitted using 2726 observations on females in the NLSY in 1994.

WORKING is a binary variable equal to 1 if the respondent was working in 1994, 0 otherwise. *CHILDL06* is a dummy variable equal to 1 if there was a child aged less than 6 in the household, 0 otherwise.

```
.probit WORKING S AGE CHILDL06 CHILDL16 MARRIED ETHBLACK ETHHISP if MALE==0

Probit estimates                                  Number of obs   =       2726
                                                  LR chi2(7)      =     165.08
                                                  Prob > chi2     =     0.0000
Log likelihood = -1403.0835                       Pseudo R2       =     0.0556

------------------------------------------------------------------------------
 WORKING |      Coef.   Std. Err.      z     P>|z|      [95% Conf. Interval]
---------+--------------------------------------------------------------------
       S |   .0892571   .0120629    7.399   0.000      .0656143       .1129
     AGE |  -.0438511    .012478   -3.514   0.000     -.0683076   -.0193946
CHILDL06 |  -.5841503   .0744923   -7.842   0.000     -.7301525   -.4381482
CHILDL16 |  -.1359097   .0792359   -1.715   0.086     -.2912092    .0193897
 MARRIED |  -.0076543   .0631618   -0.121   0.904     -.1314492    .1161407
ETHBLACK |  -.2780887    .081101   -3.429   0.001     -.4370436   -.1191337
 ETHHISP |  -.0191608   .1055466   -0.182   0.856     -.2260284    .1877068
   _cons |    .673472   .2712267    2.483   0.013      .1418775    1.205066
------------------------------------------------------------------------------
```

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 *EAEF Regression Exercises*. The mean values of the variables are given in the output below:

```
.sum WORKING S AGE CHILDL06 CHILDL16 MARRIED ETHBLACK ETHHISP if MALE==0

Variable |     Obs        Mean   Std. Dev.       Min        Max
---------+-----------------------------------------------------
 WORKING |    2726    .7652238   .4239366          0          1
       S |    2726    13.30998   2.444771          0         20
     AGE |    2726    17.64637    2.24083         14         22
CHILDL06 |    2726    .3991196   .4898073          0          1
CHILDL16 |    2726    .3180484   .4658038          0          1
 MARRIED |    2726    .6228907   .4847516          0          1
ETHBLACK |    2726    .1305943   .3370179          0          1
 ETHHISP |    2726    .0722671   .2589771          0          1
```

Calculate the marginal effects and discuss whether they are plausible. [The data set and a description are posted on the website.]

11.4 Censored regressions: Tobit analysis

Suppose that one hypothesizes the relationship

$$Y^* = \beta_1 + \beta_2 X + u, \qquad (11.17)$$

with the dependent variable subject to either a lower bound Y_L or an upper bound Y_U. In the case of a lower bound, the model can be characterized as

$$\begin{aligned} Y^* &= \beta_1 + \beta_2 X + u \\ Y &= Y^* \qquad \text{for } Y^* > Y_L \\ Y &= Y_L \qquad \text{for } Y^* \leq Y_L \end{aligned} \tag{11.18}$$

and similarly for a model with an upper bound. Such a model is known as a censored regression model because Y^* is unobserved for $Y^* < Y_L$ or $Y^* > Y_U$. It is effectively a hybrid between a standard regression model and a binary choice model, and OLS would yield inconsistent estimates if used to fit it. To see this, consider the relationship illustrated in Figure 11.7, a one-shot Monte Carlo experiment where the true relationship is

$$Y = -40 + 1.2X + u, \tag{11.19}$$

the data for X are the integers from 11 to 60, and u is a normally distributed random variable with mean 0 and standard deviation 10. If Y were unconstrained, the observations would be as shown in Figure 11.7. However we will suppose that Y is constrained to be non-negative, in which case the observations will be as shown in Figure 11.8. For such a sample, it is obvious that an OLS regression that included those observations with Y constrained to be 0 would yield inconsistent estimates, with the estimator of the slope downwards biased and that of the intercept upwards biased.

The remedy, you might think, would be to use only the subsample of unconstrained observations, but even then the OLS estimators would be biased.

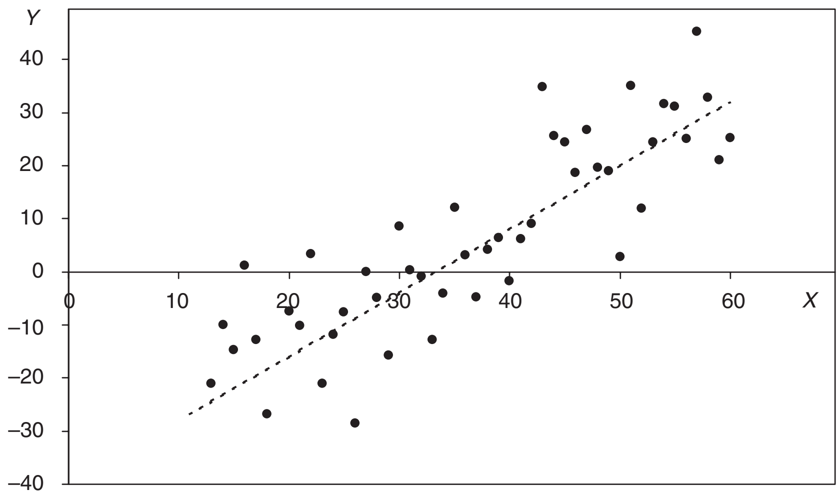


Figure 11.7

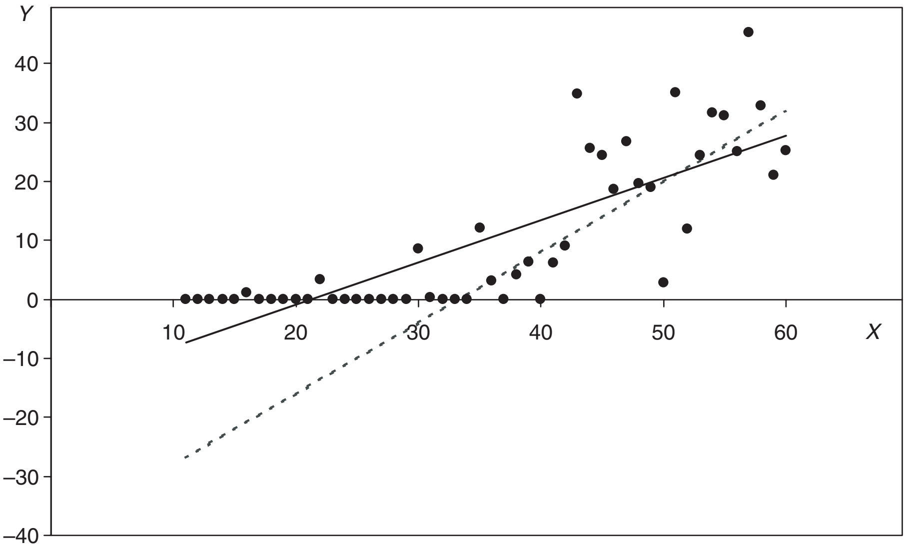


Figure 11.8

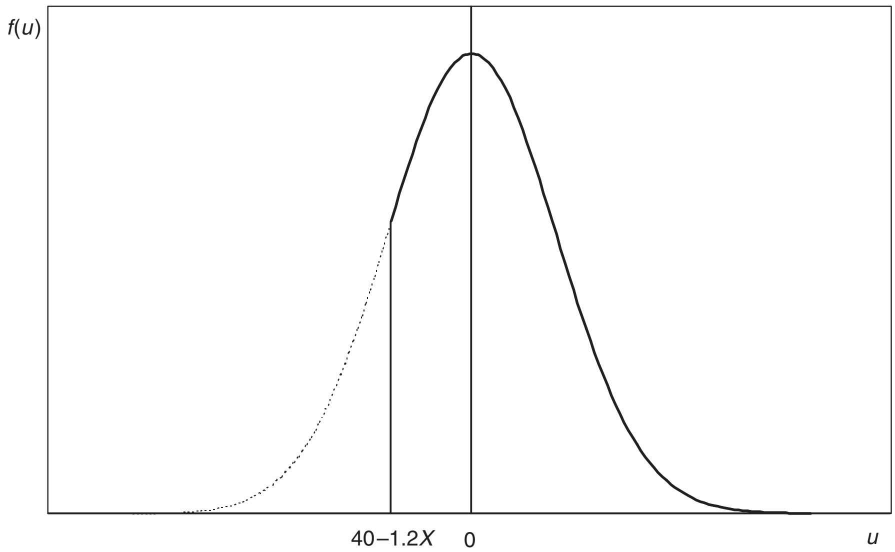


Figure 11.9

An observation i will appear in the subsample only if $Y_i > 0$, that is, if

$$-40 + 1.2X_i + u_i > 0. \tag{11.20}$$

This requires

$$u_i > 40 - 1.2X_i \tag{11.21}$$

and so u_i must have the truncated distribution shown in Figure 11.9. In this example, the expected value of u_i must be positive and a negative function of X_i.

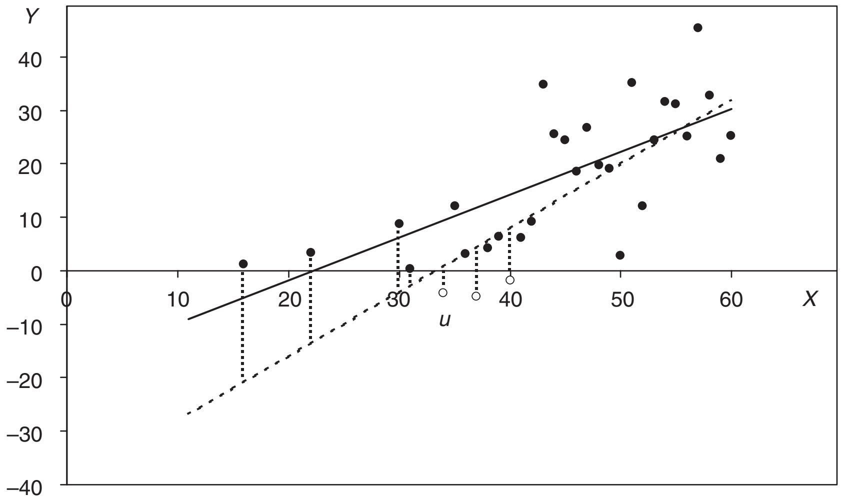


Figure 11.10

Since u_i is negatively correlated with X_i, the fourth Gauss–Markov condition is violated and OLS will yield inconsistent estimates.

Figure 11.10 displays the impact of this correlation graphically. The observations with the four lowest values of *X* appear in the sample only because their disturbance terms (marked) are positive and large enough to make *Y* positive. In addition, in the range where *X* is large enough to make the nonstochastic component of *Y* positive, observations with large negative values of the disturbance term are dropped. Three such observations, marked as circles, are shown in the figure. Both of these effects cause the intercept to tend to be overestimated, and the slope to be underestimated, in an OLS regression.

If it can be assumed that the disturbance term has a normal distribution, one solution to the problem is to use tobit analysis, a maximum likelihood estimation technique that combines probit analysis with regression analysis. A mathematical treatment will not be attempted here. Instead it will be illustrated using data on expenditure on household equipment from the Consumer Expenditure Survey data set. Figure 11.11 plots this category of expenditure, *HEQ*, and total household expenditure, *EXP*. For 86 of the 869 observations, expenditure on household equipment is 0. The output from a tobit regression is shown (Table 11.7). In Stata the command is tobit and the point of left-censoring is indicated by the number in parentheses after '`ll`'. If the data were right-censored, '`ll`' would be replaced by '`ul`'. Both may be included.

OLS regressions including and excluding the observations with 0 expenditure on household equipment yield slope coefficients of 0.0472 and 0.0468 respectively, both of them below the tobit estimate, as expected. The size of the bias tends to increase with the proportion of constrained observations. In this case only 10 percent are constrained, and hence the difference between the tobit and OLS estimates is small.

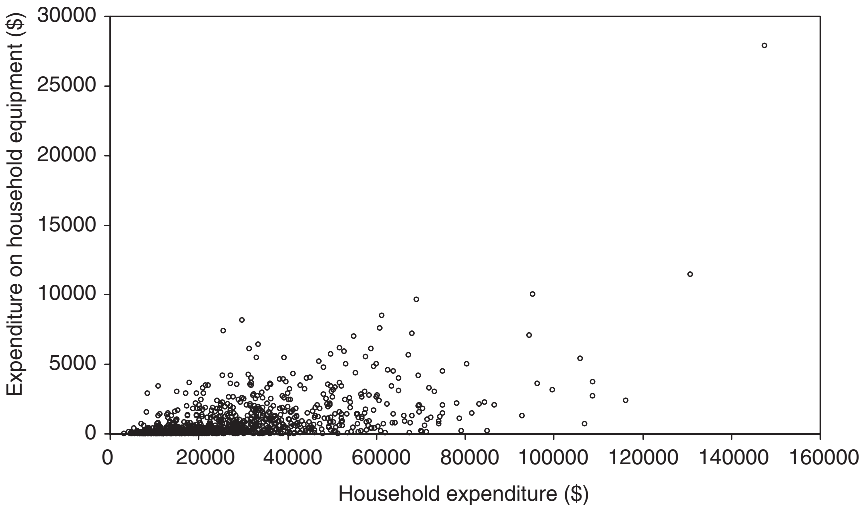


Figure 11.11 Expenditure on household equipment and total household expenditure

Table 11.7

```
. tobit HEQ EXP, ll(0)

Tobit Estimates                                   Number of obs =    869
                                                  chi2(1)       = 315.41
                                                  Prob > chi2   = 0.0000
Log Likelihood = -6911.0175                       Pseudo R2     = 0.0223

------------------------------------------------------------------------
     HEQ |      Coef.  Std. Err.      t     P>|t|    [95% Conf. Interval]
---------+--------------------------------------------------------------
     EXP |   .0520828   .0027023   19.273   0.000    .0467789    .0573866
   _cons |  -661.8156   97.95977   -6.756   0.000   -854.0813   -469.5499
---------+--------------------------------------------------------------
     _se |   1521.896    38.6333            (Ancillary parameter)
------------------------------------------------------------------------

Obs. summary:        86 left-censored observations at HEQ<=0
                    783 uncensored observations
```

Tobit regression yields inconsistent estimates if the disturbance term does not have a normal distribution or if it is subject to heteroscedasticity (Amemiya, 1984). Judging by the plot in Figure 11.11, the observations in the example are subject to heteroscedasticity and it may be preferable to use expenditure on household equipment as a proportion of total expenditure as the dependent variable, in the same way that in his seminal study, which investigated expenditure on consumer durables, Tobin (1958) used expenditure on durables as a proportion of disposable personal income.

Exercise

11.5 Using the CES data set, perform a tobit regression of expenditure on your commodity on total household expenditure, and compare the slope coefficient with those obtained in OLS regressions including and excluding observations with 0 expenditure on your commodity.

11.5 Sample selection bias

In the tobit model, whether or not an observation falls into the regression category ($Y > Y_L$ or $Y < Y_U$) or the constrained category ($Y = Y_L$ or $Y = Y_U$) depends entirely on the values of the regressors and the disturbance term. However, it may well be that participation in the regression category may depend on factors other than those in the regression model, in which case a more general model specification with an explicit two-stage process may be required. The first stage, participation in the regression category, or being constrained, depends on the net benefit of participating, B^*, a latent (unobservable) variable that depends on a set of $m - 1$ variables Q_j and a random term ε:

$$B_i^* = \delta_1 + \sum_{j=2}^{m} \delta_j Q_{ji} + \varepsilon_i. \tag{11.22}$$

The second stage, the regression model, is parallel to that for the tobit model:

$$\begin{aligned} &Y_i^* = \beta_1 \sum_{j=2}^{k} \beta_j X_{ji} + u_1 \\ &Y_i = Y_i^* && \text{for } B_i^* > 0, \\ &Y_i \text{is not observed} && \text{for } B_i^* \leq 0. \end{aligned} \tag{11.23}$$

For an observation in the sample,

$$E(u_i | B_i^* > 0) = E\left(u_i | \varepsilon_{\underline{i}} > -\delta_1 - \sum_{j=2}^{m} \delta_j Q_{ji}\right). \tag{11.24}$$

If $\varepsilon_{\underline{i}}$ and $u_{\underline{i}}$ are distributed independently, $E(u_i | \varepsilon_{\underline{i}} > -\delta_1 - \sum_{j=2}^{m} \delta_j Q_{ji})$ reduces to the unconditional $E(u_i)$ and the selection process does not interfere with the regression model. However if $\varepsilon_{\underline{i}}$ and $u_{\underline{i}}$ are correlated, $E(u_i)$ will be nonzero and problems parallel to those in the tobit model arise, with the consequence that OLS estimates are inconsistent (see Box 11.1 on the Heckman two-step procedure). If it can be assumed that $\varepsilon_{\underline{i}}$ and $u_{\underline{i}}$ are jointly normally distributed

BOX 11.1 The Heckman two-step procedure

The problem of selection bias arises because the expected value of u is nonzero for observations in the selected category if u and ε are correlated. It can be shown that, for these observations,

$$E\left(u_i \mid \varepsilon_i > -\delta_1 - \sum_{j=2}^{m} \delta_j Q_{ji}\right) = \frac{\sigma_{u\varepsilon}}{\sigma_\varepsilon}\lambda_i,$$

where $\sigma_{u\varepsilon}$ is the population covariance of u and $\varepsilon, \sigma_\varepsilon$ is the standard deviation of ε, and λ_i, described by Heckman (1976) as the inverse of Mill's ratio, is given by

$$\lambda_i = \frac{f(v_i)}{F(v_i)},$$

where

$$v_i = \frac{\varepsilon_i}{\sigma_\varepsilon} = \frac{-\delta_1 - \sum_{j=2}^{m} \delta_j Q_{ji}}{\sigma_\varepsilon}$$

and the functions f and F are as defined in the section on probit analysis: $f(v_i)$ is the density function for ε normalized by its standard deviation and $F(v_i)$ is the probability of B_i^* being positive. It follows that

$$E\left(Y_i \mid \varepsilon_i > -\delta_1 - \sum_{j=2}^{m} \delta_j Q_{ji}\right) = E\left(\beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + u_i \mid \varepsilon_i > -\delta_1 - \sum_{j=2}^{m} \delta_j Q_{ji}\right)$$

$$= \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + \frac{\sigma_{u\varepsilon}}{\sigma_\varepsilon}\lambda_i.$$

The sample selection bias arising in a regression of Y on the X variables using only the selected observations can therefore be regarded as a form of omitted variable bias, with λ the omitted variable. However, since its components depend only on the selection process, it can be estimated from the results of probit analysis of selection (the first step). If it is included as an explanatory variable in the regression of Y on the X variables, least squares will then yield consistent estimates.

As Heckman acknowledges, the procedure was first employed by Gronau (1974), but it is known as the Heckman two-step procedure in recognition of its development by Heckman into an everyday working tool, its attraction being that it is computationally far simpler than maximum likelihood estimation of the joint model. However, with the improvement in computing speeds and the development of appropriate procedures in regression applications, maximum likelihood estimation of the joint model is no more burdensome than the two-step procedure and it has the advantage of being more efficient.

with correlation ρ, the model may be fitted by maximum likelihood estimation, with null hypothesis of no selection bias H_0: $\rho = 0$. The Q and X variables may overlap, identification requiring in practice that at least one Q variable is not also an X variable.

The procedure will be illustrated by fitting an earnings function for females on the lines of Gronau (1974), the earliest study of this type, using the LFP94 subsample from the NLSY data set described in Exercise 11.4 (Table 11.8). *CHILDL06* is a dummy variable equal to 1 if there was a child aged less than

Table 11.8

```
. heckman LGEARN S ASVABC ETHBLACK ETHHISP if MALE==0, select(S AGE CHILDL06
> CHILDL16 MARRIED ETHBLACK ETHHISP)

Iteration 0:   log likelihood = -2683.5848  (not concave)
...
Iteration 8:   log likelihood = -2668.8105

Heckman selection model                         Number of obs   =    2661
(regression model with sample selection)        Censored obs    =     640
                                                Uncensored obs =    2021

                                                Wald chi2(4)    =  714.73
Log likelihood =  -2668.81                      Prob > chi2     =  0.0000

------------------------------------------------------------------------
         |      Coef.   Std. Err.     z     P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------
LGEARN   |
       S |    .095949   .0056438   17.001   0.000     .0848874    .1070106
  ASVABC |   .0110391   .0014658    7.531   0.000     .0081663    .0139119
ETHBLACK |   -.066425   .0381626   -1.741   0.082    -.1412223    .0083722
 ETHHISP |   .0744607   .0450095    1.654   0.098    -.0137563    .1626777
   _cons |   4.901626   .0768254   63.802   0.000     4.751051    5.052202
---------+--------------------------------------------------------------
select   |
       S |   .1041415   .0119836    8.690   0.000     .0806541    .1276288
     AGE |  -.0357225    .011105   -3.217   0.001    -.0574879   -.0139572
CHILDL06 |  -.3982738   .0703418   -5.662   0.000    -.5361412   -.2604064
CHILDL16 |   .0254818   .0709693    0.359   0.720    -.1136155     .164579
 MARRIED |   .0121171   .0546561    0.222   0.825    -.0950069    .1192412
ETHBLACK |  -.2941378   .0787339   -3.736   0.000    -.4484535   -.1398222
 ETHHISP |  -.0178776   .1034237   -0.173   0.863    -.2205843    .1848292
   _cons |   .1682515   .2606523    0.646   0.519    -.3426176    .6791206
---------+--------------------------------------------------------------
 /athrho |    1.01804   .0932533   10.917   0.000     .8352669    1.200813
/lnsigma |  -.6349788   .0247858  -25.619   0.000    -.6835582   -.5863994
---------+--------------------------------------------------------------
     rho |    .769067   .0380973                       .683294    .8339024
   sigma |   .5299467   .0131352                      .5048176    .5563268
  lambda |   .4075645     .02867                      .3513724    .4637567
------------------------------------------------------------------------
LR test of indep. eqns. (rho = 0):  chi2(1) = 32.90   Prob > chi2 = 0.0000
```

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 other variables have the same definitions as in the *EAEF* data sets. The Stata command for this type of regression is '`heckman`' and as usual it is followed by the dependent variable and the explanatory variables and qualifier, if any (here the sample is restricted to females). The variables in parentheses after select are those hypothesized to influence whether the dependent variable is observed. In this example it is observed for 2,021 females and is missing for the remaining 640 who were not working in 1994. Seven iteration reports have been deleted from the output.

First we will check whether there is evidence of selection bias, that is, that $\rho \neq 0$. For technical reasons, ρ is estimated indirectly through atanh $\rho = \frac{1}{2}\log((1+\rho)/(1-\rho))$, but the null hypothesis H_0: atanh $\rho = 0$ is equivalent to H_0: $\rho = 0$. atanh ρ is denoted '`athrho`' in the output and, with an asymptotic t statistic of 10.92, the null hypothesis is rejected. A second test of the same null hypothesis that can be effected by comparing likelihood ratios is described in Section 11.6.

The regression results indicate that schooling and the ASVABC score have highly significant effects on earnings, that schooling has a positive effect on the probability of working, and that age, having a child aged less than 6, and being black have negative effects. The probit coefficients are different from those reported in Exercise 11.4, the reason being that, in a model of this type, probit analysis in isolation yields inefficient estimates (Table 11.9).

Table 11.9

```
. reg LGEARN S ASVABC ETHBLACK ETHHISP if MALE==0

  Source |       SS       df       MS              Number of obs =    2021
---------+------------------------------           F(  4,  2016) =  168.55
   Model |  143.231149     4  35.8077873           Prob > F      =  0.0000
Residual |  428.301239  2016  .212451012           R-squared     =  0.2506
---------+------------------------------           Adj R-squared =  0.2491
   Total |  571.532389  2020  .282936826           Root MSE      =  .46092

------------------------------------------------------------------------
  lgearn |      Coef.  Std. Err.       t     P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------
       S |   .0807836    .005244    15.405   0.000     .0704994    .0910677
  ASVABC |   .0117377   .0014886     7.885   0.000     .0088184     .014657
ETHBLACK |  -.0148782   .0356868    -0.417   0.677    -.0848649    .0551086
 ETHHISP |   .0802266    .041333     1.941   0.052    -.0008333    .1612865
   _cons |   5.223712   .0703534    74.250   0.000     5.085739    5.361685
------------------------------------------------------------------------
```

It is instructive to compare the regression results with those from an OLS regression not correcting for selection bias. The results are in fact quite similar, despite the presence of selection bias. The main difference is in the coefficient of *ETHBLACK*. The probit regression indicates that black females are significantly less likely to work than whites, controlling for other characteristics. If this is the case, black females, controlling for other characteristics, may require higher wage offers to be willing to work. This would reduce the apparent earnings discrimination against them, accounting for the smaller negative coefficient in the OLS regression. The other difference in the results is that the schooling coefficient in the OLS regression is 0.081, a little lower than that in the selection bias model, indicating that selection bias leads to a modest underestimate of the effect of education on female earnings.

One of the problems with the selection bias model is that it is often difficult to find variables that belong to the selection process but not the main regression. Having a child aged less than 6 is an excellent variable because it clearly affects the willingness to work of a female but not her earning power while working, and for this reason the example discussed here is very popular in expositions of the model.

One final point, made by Heckman (1976): if a selection variable is illegitimately included in a least squares regression, it may appear to have a significant effect. In the present case, if *CHILDL06* is included in the earnings function, it has a *positive* coefficient significant at the 5 percent level. The explanation would appear to be that females with young children tend to require an especially attractive wage offer, given their education and other endowments, to be induced to work.

Exercise

11.6* Using your *EAEF* data set, investigate whether there is evidence that selection bias affects the least squares estimate of the returns to college education. Define $COLLYEAR = S - 12$ if $S > 12, 0$ otherwise, and $LGEARNCL = LGEARN$ if $COLLYEAR > 0$, missing otherwise. Use the Heckman procedure to regress *LGEARNCL* on *COLLYEAR*, *ASVABC MALE*, *ETHBLACK*, and *ETHHISP*, with *ASVABC SM*, *SF*, and *SIBLINGS* being used to determine whether the respondent attended college. Run the equivalent-regression using least squares. Comment on your findings.

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

11.8 Investigate whether having a child aged less than 6 is likely to be an especially powerful deterrent to working if the mother is unmarried by downloading the LFP94 data set from the website and repeating the regressions in this section adding an interactive dummy variable *MARL06* defined as the product of *MARRIED* and *CHILDL06* to the selection part of the model.

11.6 An introduction to maximum likelihood estimation

Suppose that a random variable X has a normal distribution with unknown mean μ and standard deviation σ. For the time being we will assume that we know that σ is equal to 1. We will relax this assumption later. You have a sample of two observations, values 4 and 6, and you wish to obtain an estimate of μ. The common-sense answer is 5, and we have seen that this is scientifically respectable as well since the sample mean is the least squares estimator and as such an unbiased and efficient estimator of the population mean, provided certain assumptions are valid.

However, we have seen that in practice in econometrics the necessary assumptions, in particular the Gauss–Markov conditions, are often not satisfied and as a consequence least squares estimators lose one or more of their desirable properties. We have seen that in some circumstances they may be inconsistent and we have been concerned to develop alternative estimators that are consistent. Typically we are not able to analyze the finite-sample properties of these estimators and we just hope that the estimators are well behaved.

Once we are dealing with consistent estimators, there is no guarantee that those based on the least squares criterion of goodness of fit are optimal. Indeed it can be shown that, under certain assumptions, a different approach, maximum likelihood estimation, will yield estimators that, besides being consistent, are asymptotically efficient (efficient in large samples).

To return to the numerical example, suppose for a moment that the true value of μ is 3.5. The probability density function of the normal distribution is given by

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2((X-\mu)/\sigma)^2}. \tag{11.25}$$

Figure 11.12 shows the distribution of X conditional on $\mu = 3.5$ and $\sigma = 1$. In particular, the probability density is 0.3521 when $X = 4$ and 0.0175 when $X = 6$. The joint probability density for the two observations is the product, 0.0062.

Now suppose that the true value of μ is 4. Figure 11.13 shows the distribution of X conditional on this value. The probability density is 0.3989 when $X = 4$ and 0.0540 when $X = 6$. The joint probability density for the two observations is now 0.0215. We conclude that the probability of getting values 4 and 6 for the two observations would be three times as great if μ were 4 than it would be if μ were 3.5. In that sense, $\mu = 4$ is more likely than $\mu = 3.5$. If we had to choose between these estimates, we should therefore choose 4. Of course we do not have to choose between them. According to the maximum likelihood principle, we should consider all possible values of μ and select the one that gives the observations the greatest joint probability density.

Table 11.10 computes the probabilities of $X = 4$ and $X = 6$ for values of μ from 3.5 to 6.5. The fourth column gives the joint probability density,

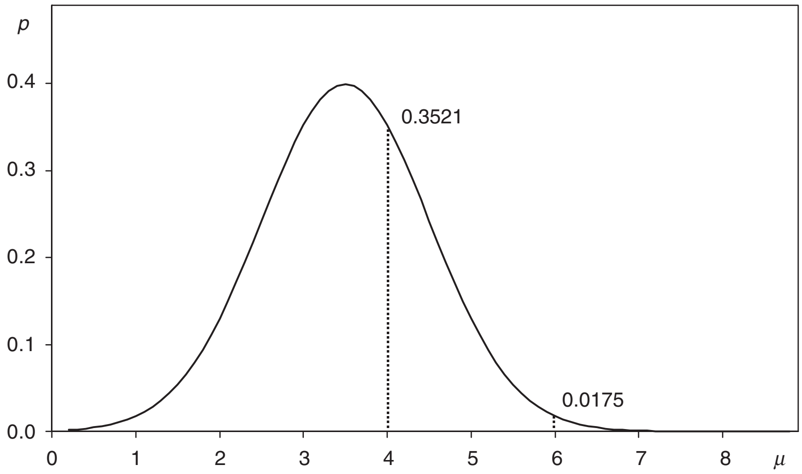


Figure 11.12 Probability densities at $X_1 = 4$ and $X_2 = 6$ conditional on $\mu = 3.5$

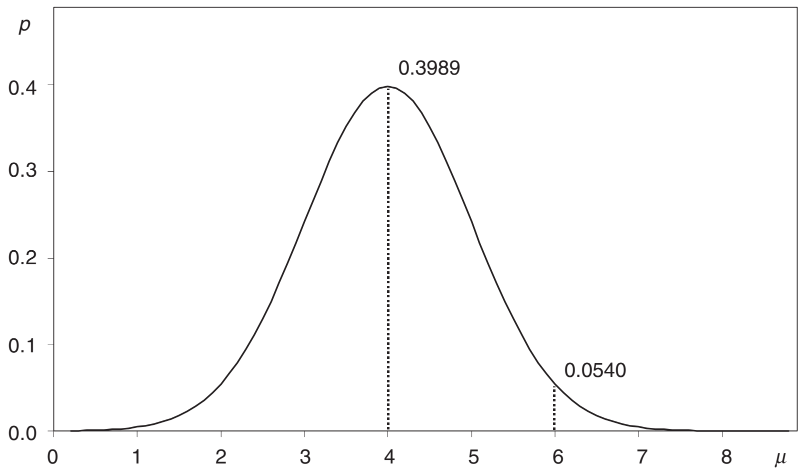


Figure 11.13 Probability densities at $X_1 = 4$ and $X_2 = 6$ conditional on $\mu = 4.0$

which is known as the likelihood function. The likelihood function is plotted in Figure 11.14. You can see that it reaches a maximum for $\mu = 5$, the average value of the two observations. We will now demonstrate mathematically that this must be the case.

First, a little terminology. The likelihood function, written $L(\mu \mid X_1 = 4, X_2 = 6)$ gives the joint probability density as a function of μ, given the sample observations. We will choose μ so as to maximize this function.

In this case, given the two observations and the assumption $\sigma = 1$, the likelihood function is given by

$$L(\mu) = \left(\frac{1}{\sqrt{2\pi}} e^{-1/2(4-\mu)^2}\right)\left(\frac{1}{\sqrt{2\pi}} e^{-1/2(6-\mu)^2}\right). \qquad (11.26)$$

Table 11.10

μ	$p(4\|\mu)$	$p(6\|\mu)$	L	$\log L$
3.5	0.3521	0.0175	0.0062	−5.0879
4.0	0.3989	0.0540	0.0215	−3.8379
4.5	0.3521	0.1295	0.0456	−3.0879
4.6	0.3332	0.1497	0.0499	−2.9979
4.7	0.3123	0.1714	0.0535	−2.9279
4.8	0.2897	0.1942	0.0563	−2.8779
4.9	0.2661	0.2179	0.0580	−2.8479
5.0	0.2420	0.2420	0.0585	−2.8379
5.1	0.2179	0.2661	0.0580	−2.8479
5.2	0.1942	0.2897	0.0563	−2.8779
5.3	0.1714	0.3123	0.0535	−2.9279
5.4	0.1497	0.3332	0.0499	−2.9979
5.5	0.1295	0.3521	0.0456	−3.0879
6.0	0.0540	0.3989	0.0215	−3.8379
6.5	0.0175	0.3521	0.0062	−5.0879

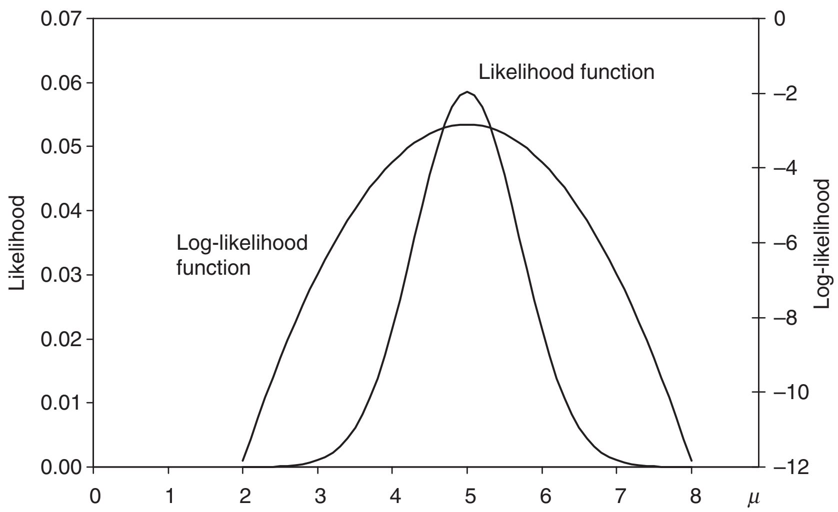


Figure 11.14 Likelihood and log-likelihood functions for μ

We will now differentiate this with respect to μ and set the result equal to 0 to obtain the first-order condition for a maximum. We will then differentiate a second time to check the second-order condition. Well, actually we won't. Even with only two observations in the sample, this would be laborious, and when we generalize to n observations it would be very messy. We will use a trick to simplify the proceedings. $\log L$ is a monotonically increasing function of L. So the value of μ that maximizes L also maximizes $\log L$, and vice versa. $\log L$ is much easier to work with, since

$$\begin{aligned}\log L &= \log\left[\left(\frac{1}{\sqrt{2\pi}}e^{-1/2(4-\mu)^2}\right)\left(\frac{1}{\sqrt{2\pi}}e^{-1/2(6-\mu)^2}\right)\right]\\ &= \log\left(\frac{1}{\sqrt{2\pi}}e^{-1/2(4-\mu)^2}\right)+\log\left(\frac{1}{\sqrt{2\pi}}e^{-1/2(6-\mu)^2}\right)\\ &= \log\left(\frac{1}{\sqrt{2\pi}}\right)-\frac{1}{2}(4-\mu)^2+\log\left(\frac{1}{\sqrt{2\pi}}\right)-\frac{1}{2}(6-\mu)^2. \qquad (11.27)\end{aligned}$$

The maximum likelihood estimator, which we will denote $\hat{\mu}$, is the value of μ that maximizes this function, given the data for X. It is given by the first-order condition

$$\frac{d\log L}{d\mu}=(4-\hat{\mu})+(6-\hat{\mu})=0. \qquad (11.28)$$

Thus $\hat{\mu}=5$. The second derivative is -2, so this gives a maximum value for $\log L$, and hence L. [Note that $-\frac{1}{2}(a-\mu)^2=-\frac{1}{2}a^2+a\mu-\frac{1}{2}\mu^2$. Hence the differential with respect to μ is $(a-\mu)$.]

Generalization to a sample of n observations

Consider a sample that consists of n observations $X_1,\ldots,X_n$. The likelihood function $L(\mu|X_1,\ldots,X_n)$ is now the product of n terms:

$$L(\mu)=\left(\frac{1}{\sqrt{2\pi}}e^{-1/2(X_1-\mu)^2}\right)\times\cdots\times\left(\frac{1}{\sqrt{2\pi}}e^{-1/2(X_n-\mu)^2}\right). \qquad (11.29)$$

The log-likelihood function is now the sum of n terms:

$$\begin{aligned}\log L &= \log\left(\frac{1}{\sqrt{2\pi}}e^{-1/2(X_1-\mu)^2}\right)+\cdots+\log\left(\frac{1}{\sqrt{2\pi}}e^{-1/2(X_n-\mu)^2}\right)\\ &= \log\left(\frac{1}{\sqrt{2\pi}}\right)-\frac{1}{2}(X_1-\mu)^2+\cdots+\log\left(\frac{1}{\sqrt{2\pi}}\right)-\frac{1}{2}(X_n-\mu)^2.\end{aligned} \qquad (11.30)$$

Hence the maximum likelihood estimator of μ is given by

$$\frac{d\log L}{d\mu}=(X_1-\hat{\mu})+\cdots+(X_n-\hat{\mu})=0. \qquad (11.31)$$

Thus

$$\sum_{i=1}^{n}X_i-n\mu=0 \qquad (11.32)$$

and the maximum likelihood estimator of μ is the sample mean. Note that the second derivative is $-n$, confirming that the log-likelihood has been maximized.

Generalization to the case where σ is unknown

We will now relax the assumption that σ is equal to 1 and accept that in practice it would be unknown, like μ. We will investigate the determination of its maximum likelihood graphically using the two-observation example and then generalize to a sample of n observations.

Figure 11.15 shows the probability distribution for X conditional on μ being equal to 5 and σ being equal to 2. The probability density at $X_1 = 4$ and $X_2 = 6$ is 0.1760 and the joint density 0.0310. Clearly we would obtain higher densities, and higher joint density, if the distribution had smaller variance. If we try σ equal to 0.5, we obtain the distribution shown in Figure 11.16. Here the individual densities are 0.1080 and the joint density 0.0117. Clearly we have made the distribution too narrow, for X_1 and X_2 are now in its tails with even lower density than before.

Figure 11.17 plots the joint density as a function of σ. We can see that it is maximized when σ is equal to 1, and this is therefore the maximum likelihood estimate, provided that we have been correct in assuming that the maximum likelihood estimate of μ is 5.

We will now derive the maximum likelihood estimators of both μ and σ simultaneously, for the general case of a sample of n observations. The likelihood function is

$$L(\mu, \sigma \mid X_1, \ldots, X_n) = \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-1/2((X_1-\mu)/\sigma)^2} \right) \times \cdots \times \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-1/2((X_n-\mu))/\sigma^2} \right) \tag{11.33}$$

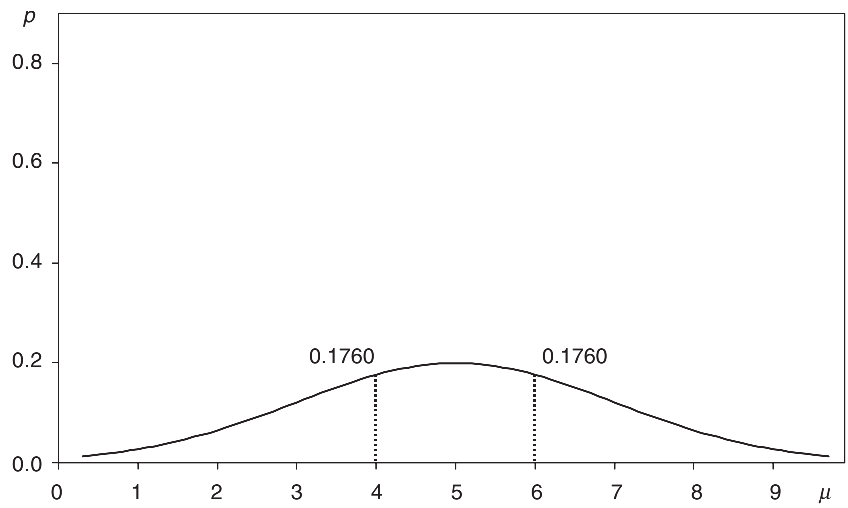


Figure 11.15 Probability densities at $X_1 = 4$ and $X_2 = 6$ conditional on $\sigma = 2$

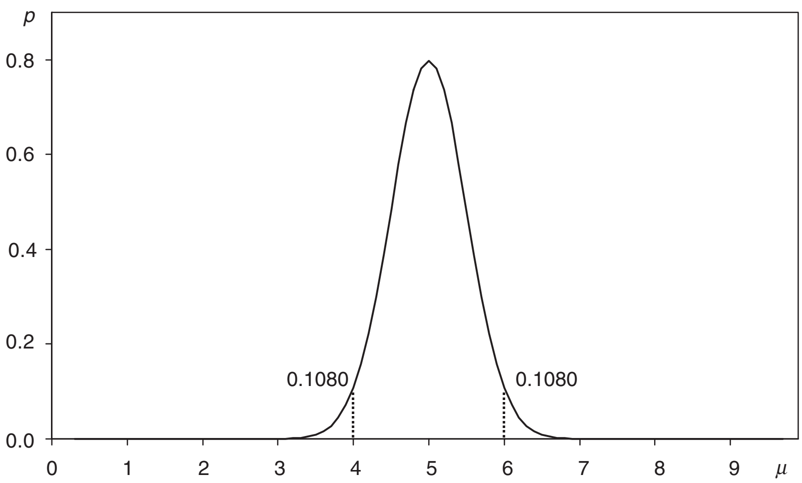


Figure 11.16 Probability densities at $X_1 = 4$ and $X_2 = 6$ conditional on $\sigma = 0.5$

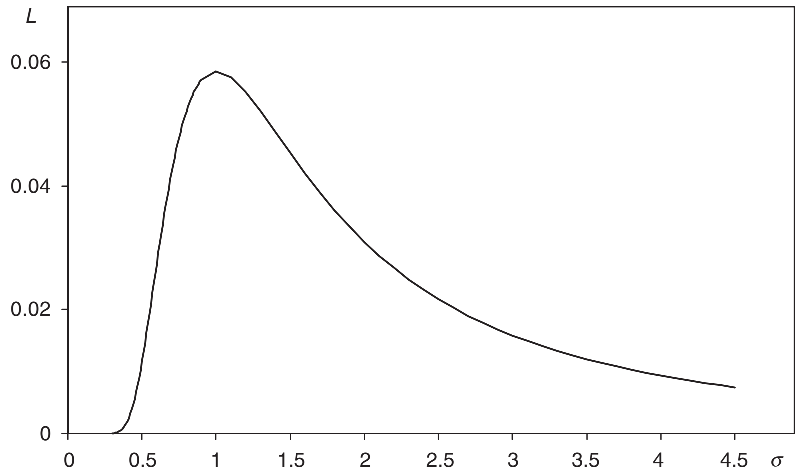


Figure 11.17 Likelihood function for σ

and so the log-likelihood function is

$$\begin{aligned}\log L &= \log\left[\left(\frac{1}{\sigma\sqrt{2\pi}}e^{-1/2((X_1-\mu)/\sigma)^2}\right)\times\cdots\times\left(\frac{1}{\sigma\sqrt{2\pi}}e^{-1/2((X_n-\mu)/\sigma)^2}\right)\right]\\ &= \log\left(\frac{1}{\sigma\sqrt{2\pi}}e^{-1/2(X_1-\mu)/\sigma^2}\right)+\cdots+\log\left(\frac{1}{\sigma\sqrt{2\pi}}e^{-1/2(X_n-\mu)/\sigma^2}\right)\\ &= n\log\left(\frac{1}{o\sqrt{2\pi}}\right)-\frac{1}{2}\left(\frac{X_1-\mu}{\sigma}\right)^2-\cdots-\frac{1}{2}\left(\frac{X_n-\mu}{\sigma}\right)^2\\ &= n\log\frac{1}{\sigma}+n\log\frac{1}{\sqrt{2\pi}}+\frac{1}{\sigma^2}\left(-\frac{1}{2}(X_1-\mu)^2-\cdots-\frac{1}{2}(X_n-\mu)^2\right).\end{aligned} \tag{11.34}$$

The partial derivative of this with respect to μ is

$$\frac{\partial \log L}{\partial \mu} = \frac{1}{\sigma^2}[(X_1 - \mu) + \cdots + (X_n - \mu)]. \quad (11.35)$$

Setting this equal to 0, one finds that the maximum likelihood estimator of μ is the sample mean, as before. The partial derivative with respect to σ is

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(X_i - \mu)^2. \quad (11.36)$$

Substituting its maximum likelihood estimator for μ and putting the expression equal to 0, we obtain

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2. \quad (11.37)$$

Note that this is actually biased downwards in finite samples, the unbiased estimator being given by the same expression with n replaced by $(n - 1)$. However it is asymptotically more efficient using the mean square error criterion, its smaller variance more than compensating for the bias. The bias in any case attenuates as the sample size becomes large.

Application to the simple regression model

Suppose that Y_i depends on X_i according to the simple relationship

$$Y_i = \beta_1 + \beta_2 X_i + u_i. \quad (11.38)$$

Potentially, before the observations are generated, Y_i has a distribution around $(\beta_1 + \beta_2 X_i)$, according to the value of the disturbance term. We will assume that the disturbance term is normally distributed with mean 0 and standard deviation σ, so

$$f(u) = \frac{1}{\sigma\sqrt{2\pi}}e^{-1/2(u/\sigma)^2}. \quad (11.39)$$

The probability that Y will take a specific value Y_i in observation i is determined by the probability that u_i is equal to $(Y_i - \beta_1 - \beta_2 Xi)$. Given the expression above, the corresponding probability density is

$$\frac{1}{\sigma\sqrt{2\pi}}e^{-1/2((Y_i - \beta_1 - \beta_2 X_i)/\sigma)^2}. \quad (11.40)$$

The joint probability density function for the observations in the sample is the product of the terms for each observation. Taking the observations as given, and treating the unknown parameters as variables, we say that the likelihood function for β_1, β_2 and σ is given by

$$L(\beta_1, \beta_2, \sigma \,|\, Y_1, \ldots, Y_n) = \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-1/2((\frac{Y_1-\beta_1-\beta_2X_1}{)}/\sigma)^2}\right) \times \cdots \times \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-1/2((Y_n-\beta_1-\beta_2X_n)/\sigma)^2}\right). \qquad (11.41)$$

The log-likelihood function is thus given by

$$\log L = n \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2\sigma}[(Y_1 - \beta_1 - \beta_2X_1)^2 + \cdots + (Y_n - \beta_1 - \beta_2X_n)^2]. \qquad (11.42)$$

The values of β_1 and β_2 that maximize this function are exactly the same as those obtained using the least squares principle. However, the estimate of σ is slightly different.

Goodness of fit and statistical tests

As noted in the discussion of logit analysis, there is no measure of goodness of fit equivalent to R^2 in maximum likelihood estimation. The pseudo-R^2 seen in some regression output, including that of Stata, compares its log-likelihood, $\log L$, with the log-likelihood that would have been obtained with only the intercept in the regression, $\log L_0$. A likelihood, being a joint probability, must lie between 0 and 1, and as a consequence a log-likelihood must be negative. The pseudo-R^2 is the proportion by which $\log L$ is smaller, in absolute size, than $\log L_0$:

$$\text{pseudo-}R^2 = 1 - \frac{\log L}{\log L_0}. \qquad (11.43)$$

While it has a minimum value of 0, its maximum value must be less than 1 and unlike R^2 it does not have a natural interpretation. However variations in the likelihood, like variations in the residual sum of squares in a standard regression, can be used as a basis for tests. In particular the explanatory power of the model can be tested via the likelihood ratio statistic.

$$2 \log \frac{L}{L_0} = 2(\log L - \log L_0). \qquad (11.44)$$

This distributed as a chi-squared statistic with $k - 1$ degrees of freedom, where $k - 1$ is the number of explanatory variables, under the null hypothesis that the coefficients of the variables are all jointly equal to 0. Further, the validity of a restriction can be tested by comparing the constrained and unconstrained likelihoods, in the same way that it can be tested by comparing the constrained and unconstrained residual sum of squares in a least squares regression model. For example, the null hypothesis H_0: $\rho = 0$ in the selection bias model can be tested by comparing the unconstrained likelihood L_U with the likelihood L_R when the model is fitted assuming that u and ε are distributed independently. Under the null hypothesis H_0: $\rho = 0$, the test statistic $2 \log L_U/L_R$ is distributed as a chi-squared statistic with one degree of freedom. In the example in Section 11.4 the test statistic, 32.90, appears in the last line of the output and the null hypothesis is rejected, the critical value of chi-squared with one degree of freedom being 10.83 at the 0.1 percent level.

As was noted in Section 11.2, the significance of an individual coefficient can be evaluated via its asymptotic t statistic, so-called because the standard error is valid only in large samples. Since the t distribution converges on the normal distribution in large samples, the critical values of the latter should be used.

Exercise

11.8* An event is hypothesized 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.

11.9* In Exercise 11.4, $\log L_0$ is the log-likelihood reported on iteration 0. Compute the pseudo-R^2 and confirm that it is equal to that reported in the output.

11.10* In Exercise 11.4, 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.