Chapter 9

Binary Choice Models

Some time we are interested in analyzing binary response or qualitative response variables that have outcomes $Y$ equal to 1 when the event occurs and equal to 0 when the event does not occur. Some example include going to college, getting married, buying a house, or getting a job. All these cases involve a yes/no answer. How is this yes/no answer affected by other variables? That is the subject matter of this chapter.

9.1 The linear probability model

9.1.1 The model

The simplest binary choice model is the linear probability model, where as its name suggests, the probability of the event occurring, $p$, is assumed to be a linear function of a set of explanatory variable. If we only have one variable the model is

$$p_i = p(Y_i = 1) = \beta_1 + \beta_2 X_i.$$  \hfill (9.1)

The response variable $Y_i$ can be written as the summation of its deterministic and its random component,

$$Y_i = E(Y_i|X_i) + u_i.$$  \hfill (9.2)

It is simple to compute $E(Y_i|X_i)$, the expected value of $Y_i$ given $X_i$, because $Y$ takes only two values. It is 1 with probability $p_i$ and 0 with probability $1 - p_i$,

$$E(Y_i|X_i) = 1 \times p_i + 0 \times (1 - p_i) = p_i = \beta_1 + \beta_2 X_i.$$  \hfill (9.3)

This means that we can write the model as

$$Y_i = \beta_1 + \beta_2 X_i + u_i.$$  \hfill (9.4)


that is just the same model we have been estimating in previous chapters. The big
difference is that $Y_i$ takes only the values 0 and 1.

9.1.2 The linear probability model in Gretl

Let’s estimate the following model

$$\text{inlf}_i = \beta_1 + \beta_2 \text{educ}_i + \beta_3 \text{faminc}_i + u_i,$$

where $\text{inlf}$ is equal to one if individual $i$ is in the labor force, zero otherwise,
educ is the number of years of education and faminc is the family income. The
regression command in Gretl is the same as before.

Model 1: OLS, using observations 1-753
Dependent variable: inlf

<table>
<thead>
<tr>
<th>coefficient</th>
<th>std. error</th>
<th>t-ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>const 0.0689887</td>
<td>0.0973736</td>
<td>0.7085</td>
<td>0.4789</td>
</tr>
<tr>
<td>educ 0.0379040</td>
<td>0.00835610</td>
<td>4.536</td>
<td>6.67e-06 ***</td>
</tr>
<tr>
<td>faminc 1.45940e-06</td>
<td>1.56306e-06</td>
<td>0.9337</td>
<td>0.3508</td>
</tr>
</tbody>
</table>

Mean dependent var 0.568393 S.D. dependent var 0.495630
Sum squared resid 178.0367 S.E. of regression 0.487219
R-squared 0.036221 Adjusted R-squared 0.033651
F(2, 750) 14.09349 P-value(F) 9.81e-07
Log-likelihood -525.5192 Akaike criterion 1057.038
Schwarz criterion 1070.911 Hannan-Quinn 1062.383

$$\hat{\text{inlf}} = 0.0689887 + 0.0379040 \text{educ} + 1.45940e-06 \text{faminc}$$

$$N = 753 \quad R^2 = 0.0337 \quad F(2, 750) = 14.093 \quad \hat{\sigma} = 0.48722$$

(standard errors in parentheses)

For example, the coefficient on educ indicates that every additional years of edu-
cation increases the probability of being in the labor force by about 4%. This informa-
tion is graphed in Figure 9.1.

There are two main problems with a linear probability model such as the one
presented in Equation 9.4.

1. The model will predict unrealistic probabilities, beyond 1 and below 0 (see Fig-
ure 9.1).
2. Because $Y_i$ only takes the values of 0 and 1, the error term $u$ will be far from
following a normal distribution.

The solution is to transform the linear probability model. Two common transforma-
tions are the logit and the probit.
9.2 Logit analysis

9.2.1 The logit transformation

Let $Z_i$ be,

$$Z_i = \beta_1 + \beta_2 X_i$$

(9.6)

The logit model hypothesizes that the probability of occurrence of the event $Y = 1$ is determined by the function

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

(9.7)

where

$$\frac{\partial p}{\partial Z} = f(Z) = \frac{e^{-Z}}{(1 + e^{-Z})^2}$$

(9.8)

and

$$\frac{\partial p}{\partial X} = \frac{\partial p}{\partial Z} \frac{\partial Z}{\partial X} = f(Z)\beta_2$$

(9.9)

This means that the marginal effect of variable $X$ on the probability of $Y = 1$ is $f(Z)\beta_2$, where $f(Z)$ needs to be evaluated on some specific value of $X$, let’s say, the mean of $X$. 

---

**Fig. 9.1** inlf$_i = \beta_1 + \beta_2 \text{educ}_i + \beta_3 \text{faminc}_i + u_i$
9.2.2 Logit regression in Gretl

Fortunately, all these calculations are done automatically by Gretl. If we want to obtain the logit estimates of Equation 9.5 in the main Gretl window we have to go to Model \(\rightarrow\) Nonlinear models \(\rightarrow\) Logit \(\rightarrow\) Binary... and select the option “Show p-values” to obtain

Convergence achieved after 4 iterations

Model 2: Logit, using observations 1-753
Dependent variable: inlf

<table>
<thead>
<tr>
<th>coefficient</th>
<th>std. error</th>
<th>z</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>const</td>
<td>-1.85287</td>
<td>0.428444</td>
<td>-4.325</td>
</tr>
<tr>
<td>educ</td>
<td>0.161773</td>
<td>0.0367856</td>
<td>4.398</td>
</tr>
<tr>
<td>faminc</td>
<td>6.58050e-06</td>
<td>6.83134e-06</td>
<td>0.9633</td>
</tr>
</tbody>
</table>

Mean dependent var 0.568393  S.D. dependent var 0.244933
McFadden R-squared 0.027185  Adjusted R-squared 0.021359
Log-likelihood -500.8762  Akaike criterion 1007.752
Schwarz criterion 1021.625  Hannan-Quinn 1013.097

Number of cases ‘correctly predicted’ = 449 (59.6%)
f(beta’x) at mean of independent vars = 0.245
Likelihood ratio test: Chi-square(2) = 27.9939 [0.0000]

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 1</td>
<td>0 69 256</td>
</tr>
</tbody>
</table>

Gretl actually estimates this model using an estimation technique called Maximum Likelihood Estimation, that is why the computer iterates before giving the estimates. The output is very similar as the one obtained in previous chapters. The effect of educ on inlf is statistically significant. However, the key difference in this output is that the coefficients are not interpreted as the marginal effects. Recall that the marginal effects are calculated using Equation 9.9. To make Gretl obtain this marginal effects you need to reestimate the model and select the option “Show slopes at mean” to obtain

<table>
<thead>
<tr>
<th>coefficient</th>
<th>std. error</th>
<th>z</th>
<th>slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>const</td>
<td>-1.85287</td>
<td>0.428444</td>
<td>-4.325</td>
</tr>
<tr>
<td>educ</td>
<td>0.161773</td>
<td>0.0367856</td>
<td>4.398</td>
</tr>
<tr>
<td>faminc</td>
<td>6.58050e-06</td>
<td>6.83134e-06</td>
<td>0.9633</td>
</tr>
</tbody>
</table>

The marginal effect of educ on inlf is actually 0.0396. An additional year of education will increase the probability that you are in the labor force by about 4%.
9.3 Probit analysis

9.3.1 The probit transformation

The probit model is similar in spirit as the logit model. Let \( Z_i \) be,

\[
Z_i = \beta_1 + \beta_2 X_i
\]  

(9.10)

The probit model hypothesizes that the probability of occurrence of the event \( Y = 1 \) is determined by the function

\[
p_i = F(Z_i)
\]  

(9.11)

where \( F(\cdot) \) is actually the cumulative standardized normal distribution. Then,

\[
\frac{\partial p}{\partial Z} = f(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z^2}
\]  

(9.12)

is just the derivative of \( F(\cdot) \). As in the logit case,

\[
\frac{\partial p}{\partial X} = \frac{\partial p}{\partial Z} \frac{\partial Z}{\partial X} = f(Z)\beta_2
\]  

(9.13)

Again, this means that the marginal effect of variable \( X \) on the probability of \( Y = 1 \) is \( f(Z)\beta_2 \), where \( f(Z) \) needs to be evaluated on some specific value of \( X \), let’s say, the mean of \( X \).

9.3.2 Probit regression in Gretl

If we want to obtain the probit estimates of Equation 9.5 in the main Gretl window we have to go to Model \( \rightarrow \) Nonlinear models \( \rightarrow \) Probit \( \rightarrow \) Binary... and select the option “Show p-values” to obtain

Convergence achieved after 5 iterations

Model 3: Probit, using observations 1-753
Dependent variable: inlf

<table>
<thead>
<tr>
<th>coefficient</th>
<th>std. error</th>
<th>z</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>const</td>
<td>-1.14768</td>
<td>0.261470</td>
<td>-4.389</td>
</tr>
<tr>
<td>educ</td>
<td>0.100666</td>
<td>0.0224296</td>
<td>4.488</td>
</tr>
<tr>
<td>faminc</td>
<td>3.84752e-06</td>
<td>4.09647e-06</td>
<td>0.9392</td>
</tr>
</tbody>
</table>

Mean dependent var 0.568393 S.D. dependent var 0.392673
McFadden R-squared 0.027216 Adjusted R-squared 0.021389
Log-likelihood -500.8606 Akaike criterion 1007.721
Schwarz criterion 1021.593 Hannan-Quinn 1013.065
Number of cases ‘correctly predicted’ = 449 (59.6%)
f(β’x) at mean of independent vars = 0.393
Likelihood ratio test: Chi-square(2) = 28.0252 [0.0000]

<table>
<thead>
<tr>
<th>Predicted</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td>0</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>48</td>
</tr>
</tbody>
</table>

Once again, the effect of educ on inlf is statistically significant. To make Gretl obtain this marginal effects using Equation 9.13 you need to reestimate the model and select the option “Show slopes at mean” to obtain

<table>
<thead>
<tr>
<th>coefficient</th>
<th>std. error</th>
<th>z</th>
<th>slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>const</td>
<td>-1.14768</td>
<td>0.261470</td>
<td>-4.389</td>
</tr>
<tr>
<td>educ</td>
<td>0.100666</td>
<td>0.0224296</td>
<td>4.488</td>
</tr>
<tr>
<td>faminc</td>
<td>3.84752e-06</td>
<td>4.09647e-06</td>
<td>0.9392</td>
</tr>
</tbody>
</table>

The marginal effect of educ on inlf is 0.0395. We obtain almost the same results as before. The probit model predicts that an additional year of education will increase the probability that you are in the labor force by about 4%.