# Models for a binary dependent variable

A binary dependent variable is one that can only take on values 0 or 1 at each observation; typically it's a coding of something qualitative (e.g. married versus not married, approved for a loan versus not approved).

## 1 The Linear Probability Model (LPM)

In the LPM we estimate the standard linear model

$$y = X\beta + u \tag{1}$$

using OLS.

Under the unbiasedness assumption $E(u|X) = 0$ we see that $E(y) = X\beta$. But, since $y$ is discrete, we can also compute $E(y)$ from first principles as the probability-weighted sum of its possible values:

$$E(y) = 1 \times P(y = 1) + 0 \times P(y = 0) = P(y = 1) \tag{2}$$

It follows that $X\beta = P(y = 1)$. For convenience, we'll refer to this common value as $p$ (and the value at each observation as $p_i$). Given an estimate of the parameter vector, $\hat{\beta}$, the fitted value, $X_i\hat{\beta}$, represents the estimated probability that $y$ takes on the value 1 at observation $i$.

This raises a first objection to the LPM: while the fitted values should be interpretable as probabilities, there's nothing in the OLS mechanism to ensure that $0 \leq X_i\hat{\beta} \leq 1$ at all observations, i.e. nothing to ensure that the fitted values are "legitimate" probabilities.

Now consider the error term in (1), namely,

$$u = y - X\beta = y - p$$

If $y_i = 1$ then $u_i = 1 - p_i$, and if $y_i = 0$ then $u_i = -p_i$. Still assuming $E(u|X) = 0$, the variance of $u_i$ can be calculated as the probability-weighted sum of the squares of the possible values of $u$:

$$\sigma_i^2 = E(u_i^2) = (1 - p_i)^2 \times p_i + (-p_i)^2 \times (1 - p_i)$$

which simplifies to $p_i\,(1 - p_i)$. This means that if $p_i$ differs across observations depending on the values of $X_i$ (as it must if the model is to be any use), the variance of the error term is non-constant. That is, the LPM is inherently heteroskedastic.

## 2 Maximum Likelihood estimation

The standard ML approach to the model with a binary dependent variable is to postulate a continuous "latent variable", $y^*$, such that

$$y^* = X\beta + u \tag{3}$$

where $u$ follows some well-defined probability distribution. The observed dependent variable, $y$, is then linked to the unobserved $y^*$ via

$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases} \tag{4}$$

If the $y_i$ values represent the outcomes of decisions made by individuals (buy a car, emigrate, vote in the Presidential election or whatever) then the latent variable can be interpreted as the perceived net benefit or utility of the action—if the net benefit is positive, perform the action, otherwise don't.

The distribution of the error term is generally assumed to be either standard normal (giving the *probit* model) or the logistic distribution (*logit* model).

How can we express the probabilities of the two outcomes in this sort of model? Well, we have

$$\begin{aligned} P(y_i = 1) &= P(y_i^* > 0) \\ &= P(X_i\beta + u_i > 0) \\ &= P(u_i > -X_i\beta) \\ &= P(u_i \leq X_i\beta) \end{aligned}$$

The last line works on condition that the distribution of $u$ is symmetrical about zero (which is the case for both probit and logit). For example, let $X_i\beta = 2.15$; then we'll get $y_i^* > 0$, and hence $y_i = 1$, so long as $u_i > -2.15$. But by symmetry, that has the same probability as $u_i < 2.15$.

The expression $P(u_i \leq X_i\beta)$ denotes the CDF of $u$ evaluated at $X_i\beta$. In the probit model this is the normal CDF, $\Phi(X_i\beta)$, and in the logit model it is the logistic CDF, $\Lambda(X_i\beta)$ where $\Lambda(x) = 1/(1 + e^{-x})$. We'll use the symbol $F(X_i\beta)$ to indicate either of these functions, depending on the context.

By the complementarity of probabilities, $P(y_i = 0)$ must equal $1 - P(y_i = 1) = 1 - F(X_i\beta)$.

To obtain estimates for the model composed of equations (3) and (4), we employ an algorithm that selects $\hat{\beta}$ to as to maximize the joint likelihood of the sample data (as a practical matter we in fact work with the log-likelihood). Each data point makes the following contribution to the log-likelihood:

$$\begin{aligned} \ell_i &= y_i \times \log P(y_i = 1) + (1 - y_i) \times \log P(y_i = 0) \\ &= y_i \log F(X_i\beta) + (1 - y_i) \log(1 - F(X_i\beta)) \end{aligned}$$

At each observation, either $y_i = 1$, in which case the first term is non-zero and the second zero, or $y_i = 0$, in which case the reverse is true. The joint log-likelihood is:

$$\ell = \sum_{i=1}^{n} \left[ y_i \log F(X_i\beta) + (1 - y_i) \log(1 - F(X_i\beta)) \right]$$

This is what we're asking the computer to maximize for us (with respect to $\hat{\beta}$) when we call for probit or logit estimates.

## 3 Reading probit and logit estimates

Some care is needed in interpreting the estimated coefficients ($\hat{\beta}$) from these models. Note that these do *not* represent the marginal effects of the independent variables on the probability that $y_i = 1$. Rather they are just the effects on the sum $X_i\hat{\beta}$. To get the marginal effects we're interested in we need to find

$$\frac{\partial F(X_i\hat{\beta})}{\partial X_i}$$

for the nonlinear function $F$. These effects will not be constant. For convenience, gretl prints each effect evaluated at the means of all the independent variables.

It often makes sense to compute a series of effects to get a fuller sense of what the model implies. For example, consider a probit model using using a dataset from T. A. Mroz (gretl's `mroz87.gdt`) containing information on 753 women. The binary dependent variable, `LFP`, takes a value of 1 if the woman participated in the labor force in 1975, otherwise 0. The dataset contains several covariates

which might plausibly influence a woman's decision to seek paid employment: we'll use KL6 (number of children under the age of 6); WA, the woman's age; WE, her education level; and MTR, the marginal tax rate she faces.

Let's see how we can find the probability that a woman is in the labor force for each case in KL6 = (0, 1, 2, 3). We need to choose values for the other elements of the $X$ matrix, and the simplest thing to do is set them at their sample means. The following gretl script will do the job.

```
open mroz87.gdt

# list of independent variables
list Xlist = KL6 WA WE MTR

# set variables other than KL6 to their sample means
matrix Xrow = { 1, 0, mean(WA), mean(WE), mean(MTR) }
matrix X = Xrow | Xrow | Xrow | Xrow
# set KL6 values from 0 to 3
matrix Kcol = { 0, 1, 2, 3 }'
X[,2] = Kcol
# check X
print X

# run probit
probit LFP 0 Xlist
# compute Phi(X\beta)
matrix P = Kcol ~ cnorm(X*$coeff)
# give names to the columns of P (optional)
colnames(P, "KL6 P(LFP=1)")
print P
g1 <- gnuplot 2 1 --matrix=P --suppress --with-lines
g1.show
```

The output reads, in part:

```
X (4 x 5)

      1.0000       0.0000       42.538       12.287      0.67886
      1.0000       1.0000       42.538       12.287      0.67886
      1.0000       2.0000       42.538       12.287      0.67886
      1.0000       3.0000       42.538       12.287      0.67886

Model 1: Probit, using observations 1-753
Dependent variable: LFP

             coefficient   std. error   t-ratio     slope
  ------------------------------------------------------------
  const       1.32275      0.705847     1.874
  KL6        -0.852144     0.111001    -7.677    -0.334180
  WA         -0.0347571    0.00673706  -5.159    -0.0136305
  WE          0.105956     0.0240874    4.399     0.0415522
  MTR        -1.11720      0.639522    -1.747    -0.438127

P (4 x 2)

         KL6     P(LFP=1)
      0.0000      0.65088
      1.0000      0.32116
      2.0000     0.093988
      3.0000     0.015052
```

Thus we see that for the first child under 6, the probability of participating in the labor force drops off from 0.65 to 0.32; for the second it falls to 0.094; and for the third, to 0.015. Not surprisingly, we get a diminishing marginal impact, as shown in Figure 1.

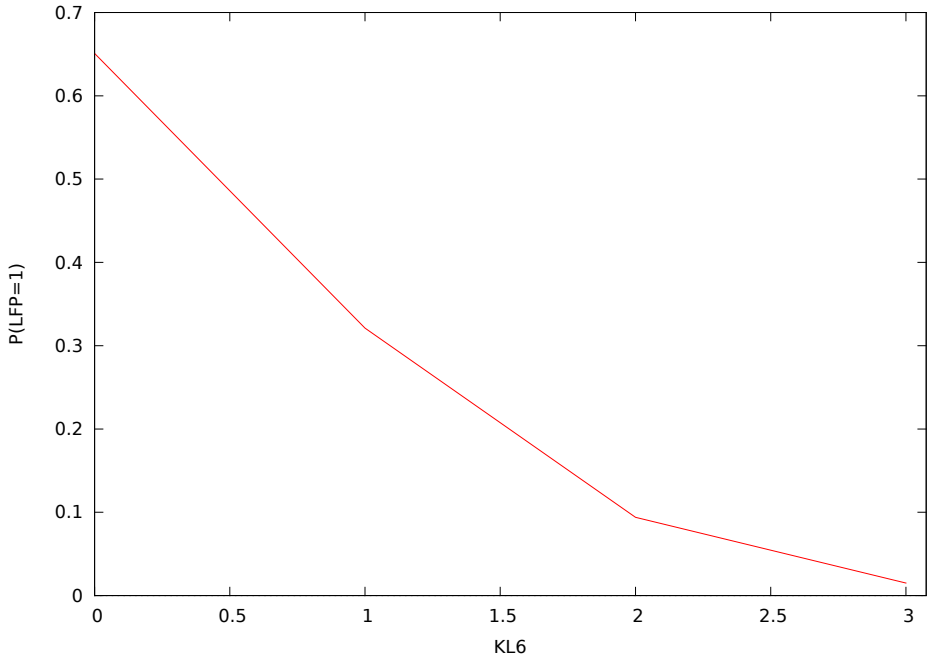


Figure 1: Probability of LFP as a function of KL6

Note that the Linear Probability Model, by contrast, imposes a constant marginal impact, which may be unrealistic. If we apply OLS to the model specified above and repeat the calculation of the effect of varying KL6 from 0 to 3 while holding the other regressors at their sample means, we get the following rather unsatisfactory result:

```
     KL6      P(LFP=1)
  0.0000       0.63928
  1.0000       0.34107
  2.0000      0.042847
  3.0000      -0.25537
```