**Binary Dependent Variables**

In some cases the outcome of interest – rather than one of the right hand side variables is discrete rather than continuous.

The simplest example of this is when the Y variable is binary – so that it can take only 1 or 2 possible values (eg Pass/Fail, Profit/Loss, Win/Lose).

Representing this is easy – just use a dummy variable on the left hand side of your model, so that

\[
Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i
\]

Becomes

\[
D_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i
\]

Where \( D_i \) = 1 if the individual (or firm or country etc) has the characteristic (eg Win) = 0 if not (eg Lose).

Equation (2) can be estimated by OLS. When do this it is called a linear probability model and can interpret the coefficients in the same way as with other OLS models.

So \( \frac{dD}{dX_1} = \beta_1 \) now gives the impact of a unit change in the value of \( X_1 \) on the chances of belonging to the category coded \( D=1 \) (eg of winning) - hence the name linear probability model.

Example: Using the dataset marks.dta can work out the chances of getting a first depend on class attendance and gender using a linear probability model.

```plaintext
gen first=mark>=70
/* first set up a dummy variable that will become the dependent variable */
tab first

<table>
<thead>
<tr>
<th></th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>79</td>
<td>66.95</td>
<td>66.95</td>
</tr>
<tr>
<td>1</td>
<td>39</td>
<td>33.05</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>118</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>

So 39 students got a first out of 118. If we summarise this variable then the mean value of this (or any) binary variable is the proportion of the sample with that characteristic.

```plaintext
ci first

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>first</td>
<td>118</td>
<td>.3305085</td>
<td>.0434881</td>
<td>.2443825 .4166345</td>
</tr>
</tbody>
</table>
```

So in this case 33% of the course got a first class mark (\(.33 \equiv 33\%\)).

To see what determines this look at the OLS regression output.
This says that the chances of getting a first rise by 3.4 percentage points for every extra class attended
and that, on average, women are 21 percentage points more likely to get a first, even after allowing for the number of classes attended.

However, can show that OLS estimates when the dependent variable is binary
1. will suffer from heteroskedasticity, so that the t-statistics are biased
2. may not constrain the predicted values to lie between 0 and 1 (which need if going to predict behaviour accurately)

Using the example above
predict phat
(option xb assumed; fitted values)

. su phat

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>phat</td>
<td>118</td>
<td>.3305085</td>
<td>.1965232</td>
<td>-.1796094</td>
<td>.6510978</td>
</tr>
</tbody>
</table>

Can see that there are some negative predictions which is odd for a variable that is binary. Note that not many of the predicted values are zero (and none are 1). This is not unusual since the model is actually predicting the probability of belonging to one of the two categories.

Because of this however it is better to use a model that explicitly rather than implicitly models this probability and does not suffer from heteroskedasticity

So that model

Probability (D=1) as a function of the right hand side variables

There are 2 alternatives

1. Logit Model

Assumes that the Probability model is given by

\[
\text{Prob (D=1)} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}}
\]
2. The Probit Model

Assumes that the Probability model is given by

\[
\text{Prob (D_i=1)} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\beta_0 + \beta_1 x_1 + \beta_2 x_2)^2}
\]

The idea is to find the values of the coefficients that make this probability as close to the values in the dataset as possible. The technique is called **maximum likelihood estimation**

Example: Using the marks.dta data set above the logit and probit equivalents of the OLS linear probability estimates above are, respectively

```
logit first num_sems female
Iteration 0:   log likelihood = -74.875501
Iteration 1:   log likelihood = -63.834471
Iteration 2:   log likelihood = -62.907479
Iteration 3:   log likelihood = -62.868479
Iteration 4:   log likelihood = -62.868381

Logistic regression                               Number of obs   =        118
LR chi2(2)      =      24.01
Prob > chi2     =     0.0000
Log likelihood = -62.868381                       Pseudo R2       =     0.1604
------------------------------------------------------------------------------
  first |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
num_sems |   .2397919   .0747599     3.21   0.001     .0932652    .3863185
female |   1.054987   .4337783     2.43   0.015     .2047975    1.905177
    _cons |  -4.357462   1.084287    -4.02   0.000    -6.482626   -2.232298
------------------------------------------------------------------------------
```

```
probit first num_sems female
Iteration 0:   log likelihood = -74.875501
Iteration 1:   log likelihood = -63.637368
Iteration 2:   log likelihood = -63.127944
Iteration 3:   log likelihood = -63.122191
Iteration 4:   log likelihood =  -63.12219

Probit regression                                 Number of obs   =        118
LR chi2(2)      =      23.51
Prob > chi2     =     0.0000
Log likelihood =  -63.12219                       Pseudo R2       =     0.1570
------------------------------------------------------------------------------
  first |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
num_sems |   .1329781   .0397433     3.35   0.001     .0550827    .2108736
female |   .6198004   .2607587     2.38   0.017     .1087228    1.130878
    _cons |  -2.44597   .5545824    -4.41   0.000    -3.532932   -1.359009
------------------------------------------------------------------------------
```

The standard errors and t values on these variables should be free of the bias inherent in OLS – though they could still be subject to other types of heteroskedasticity so it is a good idea to use the ", robust" adjustment even with logit and probit estimators
logit first num_sems female, robust

Iteration 0:  log pseudolikelihood = -74.875501
Iteration 1:  log pseudolikelihood = -63.834471
Iteration 2:  log pseudolikelihood = -62.907479
Iteration 3:  log pseudolikelihood = -62.868479
Iteration 4:  log pseudolikelihood = -62.868381

Logistic regression                              Number of obs   =        118
Wald chi2(2)    =      14.65
Prob > chi2     =     0.0007
Log pseudolikelihood = -62.868381                 Pseudo R2       =     0.1604
------------------------------------------------------------------------------
|               Robust  |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>first</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
|num_sems   |   .2397919   .0770628     3.11   0.002     .0887516    .3908322
|female     |   1.054987   .4304162     2.45   0.014     .2113872    1.898588
|_cons      |  -4.357462   1.154993    -3.77   0.000    -6.621206   -2.093718
------------------------------------------------------------------------------

(in this example it make little difference)

However in both cases the estimated coefficients look very different from those of the OLS linear probability estimates. This is because they do not have the same interpretation as with OLS (they are simply values that maximise the likelihood function).

To obtain coefficients which can be interpreted in a similar way to OLS, need marginal effects

In the case of probit model this is given by

$$\frac{\delta \Pr ob(T_i = 1)}{\delta X_i} = \beta \phi(Z, \gamma)$$

and for the logit model this is given by

$$\frac{\delta \Pr ob(T_i = 1)}{\delta X_i} = \beta \frac{\exp(Z, \gamma)}{\left(1 + \exp(Z, \gamma)\right)^2} = \beta \Pr ob(T_i = 1) \cdot (1 - \Pr ob(T_i = 1))$$

In both cases the interpretation of these marginal effects is the impact that a unit change in the variable X_i has on the probability of belonging to the treatment group (just like OLS coefficients)
To obtain marginal effects in Stata run either the logit or probit command then simply type

```stata
logit first num_sems female
mfx
```

Marginal effects after logit

```
y  = Pr(first) (predict)
= .27708689
```

| variable | dy/dx    | Std. Err. | z    | P>|z|  | 95% C.I.   | X      |
|----------|----------|-----------|------|------|-----------|--------|
| num_sems | .0480326 | .01343    | 3.58 | 0.000| .021713   | 12.4576|
| female*  | .2192534 | .09097    | 2.41 | 0.016| .040952   | .397554|

(*) dy/dx is for discrete change of dummy variable from 0 to 1

```stata
probit first num_sems female
mfx
```

Marginal effects after probit

```
y  = Pr(first) (predict)
= .29192788
```

| variable | dy/dx    | Std. Err. | z    | P>|z|  | 95% C.I.   | X      |
|----------|----------|-----------|------|------|-----------|--------|
| num_sems | .0456601 | .01301    | 3.51 | 0.000| .020157   | 12.4576|
| female*  | .2177256 | .09231    | 2.36 | 0.018| .036807   | .398645|

(*) dy/dx is for discrete change of dummy variable from 0 to 1

(or with probit you can also type

```stata
dprobit first num_sems female
```

Iteration 0:  log likelihood = -74.875501
Iteration 1:  log likelihood = -63.637368
Iteration 2:  log likelihood = -63.127944
Iteration 3:  log likelihood = -63.122191
Iteration 4:  log likelihood = -63.12219

Probit regression, reporting marginal effects

```
Number of obs = 118
LR chi2(2)    = 23.51
Prob > chi2   = 0.0000
Log likelihood = -63.12219
Pseudo R2     = 0.1570
```

| first   | dF/dx    | Std. Err. | z    | P>|z|  | x-bar     | 95% C.I.   |
|---------|----------|-----------|------|------|-----------|-----------|
| num_sems| .0456601 | .0130119  | 3.35 | 0.001| 12.4576   | .020157   | .071163   |
| female* | .2177256 | .0923073  | 2.38 | 0.017| .389831   | .036807   | .398645   |

obs. P | .3305085
pred. P | .2919279 (at x-bar)

(*) dF/dx is for discrete change of dummy variable from 0 to 1

z and P>|z| correspond to the test of the underlying coefficient being 0

These estimates are similar to those of OLS (as they should be since OLS, logit and probit are unbiased)
Note also that the predicted values from the logit and probit regressions will lie between 0 and 1

predict phat_probit
(option p assumed; Pr(first))

su phat_probit phat_logit

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>phat_probit</td>
<td>118</td>
<td>.3330643</td>
<td>.1980187</td>
<td>.0072231</td>
<td>.7147911</td>
</tr>
<tr>
<td>phat_logit</td>
<td>118</td>
<td>.3305085</td>
<td>0</td>
<td>.3305085</td>
<td>.3305085</td>
</tr>
</tbody>
</table>

- while the means are the same the predictions are not identical for the two estimation techniques

**Goodness of Fit Tests**

The maximum likelihood equivalent to the F test of goodness of fit of the model as a whole is given by the Likelihood Ratio (LR) test which compares the value of the (log) likelihood when maximised with the likelihood function when all coefficients set to zero

Can show that

\[ 2 \left[ \log L_{\text{max}} - \log L_0 \right] \sim \chi^2_{(k-1)} \]

where \( k \) is the number of right hand side variables including the constant

If the estimate chi-squared value exceeds the critical value then reject the null that the model as no explanatory power

This value is given in the top right hand corner of the logit/probit output in Stata

Can also use the LR test to test restrictions on subsets of the coefficients in a similar way.

A maximum likelihood equivalent of the \( R^2 \) is the pseudo-\( R^2 \)

\[ = 1 - \frac{\log L_{\text{max}}}{\log L_0} \]

This value lies between 0 and 1 and the closer to one the better the fit of the model (again this value is given in the top right hand corner of the logit/probit output in Stata)