PRINCETON UNIVERSITY

Getting Started in Logit and Ordered Logit Regression

(ver. 3.1 *beta*)

Oscar Torres-Reyna
Data Consultant
otorres@princeton.edu

http://dss.princeton.edu/training/

Logit model

- Use logit models whenever your dependent variable is binary (also called dummy) which takes values 0 or 1.
- Logit regression is a nonlinear regression model that forces the output (predicted values) to be either 0 or 1.
- Logit models estimate the probability of your dependent variable to be 1 (Y=1). This is the probability that some event happens.

Logit model

From Stock & Watson, key concept 9.3. The logit model is:

$$\Pr(Y=1 \mid X1, X2, \ldots X_k) = F(\beta_0 + \beta_1 X1 + \beta_2 X2 + \ldots + \beta_K X_K)$$

$$\Pr(Y=1 \mid X1, X2, \ldots X_k) = \frac{1}{1+e^{-(\beta_0 + \beta_1 X1 + \beta_2 X2 + \ldots + \beta_K X_K)}}$$

$$\Pr(Y=1 \mid X1, X2, \ldots X_k) = \frac{1}{1+\left(\frac{1}{e^{(\beta_0 + \beta_1 X1 + \beta_2 X2 + \ldots + \beta_K X_K)}}\right)}$$

Logit and probit models are basically the same, the difference is in the distribution:

- Logit – Cumulative standard logistic distribution (*F*)
- Probit – Cumulative standard normal distribution (Φ)

Both models provide similar results.

Logit model

In Stata you run the model as follows:

Dependent variable

Independent variable(s)

```
. logit  y_bin x1 x2 x3 x4 x5 x6 x7

Iteration 0:   log likelihood =  -251.9712
Iteration 1:   log likelihood =  -192.3814
Iteration 2:   log likelihood = -165.56847
Iteration 3:   log likelihood = -160.76756
Iteration 4:   log likelihood = -160.44413
Iteration 5:   log likelihood =   -160.442

Logistic regression                               Number of obs   =        490
                                                   LR chi2(7)      =     183.06
                                                   Prob > chi2     =     0.0000
Log likelihood =   -160.442                        Pseudo R2       =     0.3633

------------------------------------------------------------------------------
       y_bin |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .2697623   .1759677     1.53   0.125    -.0751281    .6146527
          x2 |  -.2500592   .1459846    -1.71   0.087    -.5361837    .0360653
          x3 |   .1150445   .1486181     0.77   0.439    -.1762417    .4063306
          x4 |   .3649722    .153434     2.38   0.017     .0642472    .6656973
          x5 |  -.3131214   .1467796    -2.13   0.033    -.6008042   -.0254386
          x6 |  -.1361499   .1566993    -0.87   0.385    -.4432749    .1709752
          x7 |   3.206987   .3631481     8.83   0.000     2.495229    3.918744
       _cons |    1.58614     .39927     3.97   0.000     .803585    2.368695
------------------------------------------------------------------------------
Note: 1 failure and 1 success completely determined.
```

It tests whether the combined effect, of all the variables in the model, is different from zero. If, for example, < 0.05 then the model have some relevant explanatory power, which does not mean it is well specified or at all correct.

Logit coefficients are in log-odds units and cannot be read as regular OLS coefficients. To interpret you need to estimate the predicted probabilities of Y=1 (see next page)

Test the hypothesis that each coefficient is different from 0. To reject this, the t-value has to be higher than 1.96 (for a 95% confidence). If this is the case then you can say that the variable has a significant influence on your dependent variable (y). The higher the z the higher the relevance of the variable.

Two-tail p-values test the hypothesis that each coefficient is different from 0. To reject this, the p-value has to be lower than 0.05 (95%, you could choose also an alpha of 0.10), if this is the case then you can say that the variable has a significant influence on your dependent variable (y)

Logit: predicted probabilities

After running the model:

```
logit y_bin x1 x2 x3 x4 x5 x6 x7
```

Type

```
predict y_bin_hat   /*These are the predicted probabilities of Y=1 */
```

Here are the estimations for the first five cases, type:

```
browse y_bin x1 x2 x3 x4 x5 x6 x7 y_bin_hat
```

Predicted probabilities

y_bin	x1	x2	x3	x4	x5	x6	x7	y_bin_hat
1	3	.2779036	-1.107956	.2825536	-2.971267	.554832	-.5820704	.7841014
0	3	.3206847	-.94872	.4925385	-1.371243	-.0959275	-.6641465	.6678266
0	3	.3634657	-.789484	.7025234	.2287798	-.7466869	-.7462227	.5267279
1	3	.246144	-.885533	-.0943909	-.3198499	-.3573879	.0628607	.9274359
1	3	.424623	-.7297683	.9461306	.1230506	-.0358964	.095743	.9439594
1	3	.4772141	-.723246	1.02968	.1175985	-.0022627	.0965806	.9448991

To estimate the probability of Y=1 for the first row, replace the values of X into the logit regression equation. For the first case, given the values of X there is 79% probability that Y=1:

$$\Pr(Y=1 \mid X_1, X_2, \ldots X_7) = \frac{1}{1+\left(\dfrac{1}{e^{(1.58+0.26X_1-.25X_2+0.11X_3+0.36X_4-0.31X_5-0.13X_6+3.20X_7)}}\right)} = 0.7841014$$

Logit: Odds ratio

You can request odds ratio rather than logit coefficients by adding the option `or` (after comma)

Dependent variable

Independent variable(s)

Getting odds ratios

```
. logit  y_bin x1 x2 x3 x4 x5 x6 x7, or

Iteration 0:   log likelihood =  -251.9712
Iteration 1:   log likelihood =  -192.3814
Iteration 2:   log likelihood = -165.56847
Iteration 3:   log likelihood = -160.76756
Iteration 4:   log likelihood = -160.44413
Iteration 5:   log likelihood =   -160.442

Logistic regression                               Number of obs   =        490
                                                   LR chi2(7)      =     183.06
                                                   Prob > chi2     =     0.0000
Log likelihood =   -160.442                        Pseudo R2       =     0.3633
```

y_bin	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
x1	1.309653	.2304567	1.53	0.125	.9276246	1.849014
x2	.7787547	.1136862	-1.71	0.087	.5849765	1.036724
x3	1.121923	.1667381	0.77	0.439	.8384153	1.501299
x4	1.440474	.2210176	2.38	0.017	1.066356	1.945847
x5	.7311612	.1073196	-2.13	0.033	.5483705	.9748823
x6	.8727118	.1367534	-0.87	0.385	.6419307	1.186461
x7	24.70453	8.971405	8.83	0.000	12.12451	50.33718

```
Note: 1 failure and 1 success completely determined.
```

It tests whether the combined effect, of all the variables in the model, is different from zero. If, for example, < 0.05 then the model have some relevant explanatory power, which does not mean it is well specified or at all correct.

They represent the *odds of Y=1 when X increases by 1 unit*. These are the exp(logit coeff).

If the OR > 1 then the odds of Y=1 increases

If the OR < 1 then the odds of Y=1 decreases

Look at the sign of the logit coefficients

Test the hypothesis that each coefficient is different from 1. To reject this, the t-value has to be higher than 1.96 (for a 95% confidence). If this is the case then you can say that the variable has a significant influence on your dependent variable (y). The higher the z the higher the relevance of the variable.

Two-tail p-values test the hypothesis that each coefficient is different from 1. To reject this, the p-value has to be lower than 0.05 (95%, you could choose also an alpha of 0.10), if this is the case then you can say that the variable has a significant influence on your dependent variable (y)

Predicted probabilities and marginal effects

For the latest procedure see the following document:

http://dss.princeton.edu/training/Margins.pdf

The procedure using `prvalue` in the following pages does not work with Stata 13.

Ordinal logit

When a dependent variable has more than two categories and the values of each category have a meaningful sequential order where a value is indeed 'higher' than the previous one, then you can use ordinal logit.

Here is an example of the type of variable:

```
. tab y_ordinal
```

Agreement level	Freq.	Percent	Cum.
Disagree	190	38.78	38.78
Neutral	104	21.22	60.00
Agree	196	40.00	100.00
Total	490	100.00	

Ordinal logit: the setup

Dependent variable

Independent variable(s)

```
. ologit  y_ordinal x1 x2 x3 x4 x5 x6 x7

Iteration 0:   log likelihood = -520.79694
Iteration 1:   log likelihood = -475.83683
Iteration 2:   log likelihood = -458.82354
Iteration 3:   log likelihood = -458.38223
Iteration 4:   log likelihood = -458.38145

Ordered logistic regression                       Number of obs   =        490
                                                  LR chi2( 7)      =     124.83
                                                  Prob > chi2     =     0.0000
Log likelihood = -458.38145                        Pseudo R2       =     0.1198
```

y_ordinal	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
x1	.220828	.0958182	2.30	0.021	.0330279	.4086282
x2	-.0543527	.0899153	-0.60	0.546	-.2305834	.1218779
x3	.1066394	.0925103	1.15	0.249	-.0746775	.2879563
x4	.2247291	.0913585	2.46	0.014	.0456697	.4037885
x5	-.2920978	.0910174	-3.21	0.001	-.4704886	-.113707
x6	.0034756	.0860736	0.04	0.968	-.1652255	.1721767
x7	1.566211	.1782532	8.79	0.000	1.216841	1.915581
/cut1	-.5528058	.103594			-.7558463	-.3497654
/cut2	.5389237	.1027893			.3374604	.740387

Note: 1 observation completely determined. Standard errors questionable.

If this number is < 0.05 then your model is ok. This is a test to see whether all the coefficients in the model are different than zero.

Logit coefficients are in log-odds units and cannot be read as regular OLS coefficients. To interpret you need to estimate the predicted probabilities of Y=1 (see next page)

Ancillary parameters to define the changes among categories (see next page)

Test the hypothesis that each coefficient is different from 0. To reject this, the t-value has to be higher than 1.96 (for a 95% confidence). If this is the case then you can say that the variable has a significant influence on your dependent variable (y). The higher the z the higher the relevance of the variable.

Two-tail p-values test the hypothesis that each coefficient is different from 0. To reject this, the p-value has to be lower than 0.05 (95%, you could choose also an alpha of 0.10), if this is the case then you can say that the variable has a significant influence on your dependent variable (y)

Ordinal logit: predicted probabilities

Following Hamilton, 2006, p.279, `ologit` estimates a score, *S*, as a linear function of the X's:

$$S = 0.22X_1 - 0.05X_2 + 0.11X_3 + 0.22X_4 - 0.29X_5 + 0.003X_6 + 1.57X_7$$

Predicted probabilities are estimated as:

P(y_ordinal="disagree") = *P*(*S* + *u* ≤ _cut1) = *P*(*S* + *u* ≤ -0.5528058)

P(y_ordinal="neutral") = *P*(_cut1 < *S* + *u* ≤ _cut2) = *P*(-0.5528058 < *S* + *u* ≤ 0.5389237)

P(y_ordinal="agree") = *P*(_cut2 < *S* + *u*) = *P*(0.5389237 < *S* + *u*)

To estimate predicted probabilities type `predict` right after `ologit` model. Unlike `logit`, this time you need to specify the predictions for all categories in the ordinal variable (y_ordinal), type:

```
predict disagree neutral agree
```

Ordinal logit: predicted probabilities

To read these probabilities, as an example, type

```
browse country disagree neutral agree if year==1999
```

In 1999 there is a 62% probability of 'agreement' in Australia compared to 58% probability in 'disagreement' in Brazil while Denmark seems to be quite undecided.

country	disagree	neutral	agree
Australia	.1700809	.2090298	.6208892
Austria	.17576	.2127421	.6114978
Belgium	.3058564	.2617683	.4323753
Botswana	.1215602	.1703741	.7080657
Brazil	.5808533	.2241725	.1949743
Bulgaria	.3134856	.2628762	.4236383
Burundi	.5940011	.2193996	.1865993
Canada	.1627286	.2039865	.6332849
Chile	.1998139	.2267881	.5733979
Denmark	.3604209	.2663039	.3732751

Predicted probabilities and marginal effects

For the latest procedure see the following document:

http://dss.princeton.edu/training/Margins.pdf

The procedure using `prvalue` in the following pages does not work with Stata 13.

Predicted probabilities: using `prvalue`

After runing `ologit` (or `logit`) you can use the command prvalue to estimate the probabilities for each event.

`Prvalue` is a user-written command, if you do not have it type `findit spost` , select `spost9_ado from http://www.indiana.edu/~jslsoc/stata` and click on "`(click here to install)`"

If you type `prvalue` without any option you will get the probabilities for each category when all independent values are set to their mean values.

```
. prvalue

ologit: Predictions for y_ordinal

Confidence intervals by delta method

                                 95% Conf. Interval
     Pr(y=Disagree|x):  0.3627  [ 0.3159,    0.4094]
     Pr(y=Neutral|x):  0.2643   [ 0.2197,    0.3090]
     Pr(y=Agree|x):    0.3730   [ 0.3262,    0.4198]

             x1          x2          x3          x4          x5          x6          x7
x=    2.0020408  -8.914e-10  -1.620e-10  -1.212e-10   2.539e-09  -9.744e-10  -6.040e-10
```

You can also estimate probabilities for a particular profile (type `help prvalue` for more details).

```
. prvalue , x(x1=1 x2=3 x3=0 x4=-1 x5=2 x6=2 x6=9 x7=4)

ologit: Predictions for y_ordinal

Confidence intervals by delta method

                                 95% Conf. Interval
     Pr(y=Disagree|x):  0.0029  [-0.0033,    0.0090]
     Pr(y=Neutral|x):  0.0055   [-0.0061,    0.0172]
     Pr(y=Agree|x):    0.9916   [ 0.9738,    1.0094]

      x1   x2   x3   x4   x5   x6   x7
x=     1    3    0   -1    2    9    4
```

For more info go to: http://www.ats.ucla.edu/stat/stata/dae/probit.htm

Predicted probabilities: using `prvalue`

If you want to estimate the impact on the probability by changing values you can use the options `save` and `dif` (type `help prvalue` for more details)

```
. prvalue , x(x1=1) save

ologit: Predictions for y_ordinal

Confidence intervals by delta method

                                      95% Conf. Interval
     Pr(y=Disagree|x):  0.3837   [ 0.3098,    0.4576]
     Pr(y=Neutral|x):  0.2641    [ 0.2195,    0.3087]
     Pr(y=Agree|x):    0.3522    [ 0.2806,    0.4238]

            x1          x2          x3          x4          x5          x6          x7
x=           1  -8.914e-10  -1.620e-10  -1.212e-10   2.539e-09  -9.744e-10  -6.040e-10
```

Probabilities when x1=1 and all other independent variables are held at their mean values. Notice the `save` option.

```
. prvalue , x(x1=2) dif

ologit: Change in Predictions for  y_ordinal

Confidence intervals by delta method

                         Current     Saved     Change    95% CI for Change
     Pr(y=Disagree|x):  0.3627     0.3837    -0.0210   [-0.0737,    0.0317]
     Pr(y=Neutral|x):  0.2643      0.2641     0.0003   [-0.0026,    0.0031]
     Pr(y=Agree|x):    0.3730      0.3522     0.0208   [-0.0299,    0.0714]

                  x1          x2          x3          x4          x5          x6          x7
Current=           2  -8.914e-10  -1.620e-10  -1.212e-10   2.539e-09  -9.744e-10  -6.040e-10
  Saved=           1  -8.914e-10  -1.620e-10  -1.212e-10   2.539e-09  -9.744e-10  -6.040e-10
   Diff=           1           0           0           0           0           0           0
```

Probabilities when x1=2 and all other independent variables are held at their mean values. Notice the `dif` option.

Here you can see the impact of x1 when it changes from 1 to 2.

For example, the probability of y=Agree goes from 35% to 37% when x1 changes from 1 to 2 (and all other independent variables are held at their constant mean values.

NOTE: You can do the same with logit or probit models

Useful links / Recommended books

- DSS Online Training Section http://dss.princeton.edu/training/
- UCLA Resources to learn and use STATA http://www.ats.ucla.edu/stat/stata/
- DSS help-sheets for STATA http://dss/online_help/stats_packages/stata/stata.htm
- *Introduction to Stata* (PDF), Christopher F. Baum, Boston College, USA. "A 67-page description of Stata, its key features and benefits, and other useful information." http://fmwww.bc.edu/GStat/docs/StataIntro.pdf
- STATA FAQ website http://stata.com/support/faqs/
- Princeton DSS Libguides http://libguides.princeton.edu/dss

Books

- *Introduction to econometrics* / James H. Stock, Mark W. Watson. 2nd ed., Boston: Pearson Addison Wesley, 2007.
- *Data analysis using regression and multilevel/hierarchical models* / Andrew Gelman, Jennifer Hill. Cambridge ; New York : Cambridge University Press, 2007.
- *Econometric analysis* / William H. Greene. 6th ed., Upper Saddle River, N.J. : Prentice Hall, 2008.
- *Designing Social Inquiry: Scientific Inference in Qualitative Research* / *Gary King, Robert O. Keohane, Sidney Verba, Princeton University Press*, 1994.
- *Unifying Political Methodology: The Likelihood Theory of Statistical Inference* / Gary King, Cambridge University Press, 1989
- *Statistical Analysis: an interdisciplinary introduction to univariate & multivariate methods /* Sam Kachigan, New York : Radius Press, c1986
- *Statistics with Stata (updated for version 9) /* Lawrence Hamilton, Thomson Books/Cole, 2006