Stata Data Analysis Examples: Logistic Regression (from Idre, UCLA website)

Version info: Code for this page was tested in Stata 12.

Logistic regression, also called a logit model, is used to model dichotomous outcome variables. In the logit model the log odds of the outcome is modeled as a linear combination of the predictor variables.

Examples of logistic regression

Example 1: Suppose that we are interested in the factors that influence whether a political candidate wins an election. The outcome (response) variable is binary (0/1); win or lose. The predictor variables of interest are the amount of money spent on the campaign, the amount of time spent campaigning negatively and whether or not the candidate is an incumbent.

Example 2: A researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution, effect admission into graduate school. The response variable, admit/don't admit, is a binary variable.

Description of the data

For our data analysis below, we are going to expand on Example 2 about getting into graduate school. We have generated hypothetical data, which can be obtained from our website.

```
use http://www.ats.ucla.edu/stat/stata/dae/binary.dta, clear
```

This data set has a binary response (outcome, dependent) variable called **admit**. There are three predictor variables: **gre**, **gpa** and **rank**. We will treat the variables **gre** and **gpa** as continuous. The variable **rank** takes on the values 1 through 4. Institutions with a rank of 1 have the highest prestige, while those with a rank of 4 have the lowest.

```
summarize gre gpa
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         gre |       400       587.7    115.5165       220        800
         gpa |       400      3.3899    .3805668      2.26          4


tab rank
       rank |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         61       15.25       15.25
          2 |        151       37.75       53.00
          3 |        121       30.25       83.25
          4 |         67       16.75      100.00
------------+-----------------------------------
      Total |        400      100.00


tab admit
```

```
      admit |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         273       68.25       68.25
          1 |         127       31.75      100.00
------------+-----------------------------------
      Total |         400      100.00
```

Logistic regression

Below we use the **logit** command to estimate a logistic regression model. The **i.** before **rank** indicates that **rank** is a factor variable (i.e., categorical variable), and that it should be included in the model as a series of indicator variables. Note that this syntax was introduced in Stata 11.

```
logit admit gre gpa i.rank
Iteration 0:   log likelihood = -249.98826
Iteration 1:   log likelihood = -229.66446
Iteration 2:   log likelihood = -229.25955
Iteration 3:   log likelihood = -229.25875
Iteration 4:   log likelihood = -229.25875
 Logistic regression                                     Number of obs   =        400
                                                         LR chi2(5)      =      41.46
                                                         Prob > chi2     =     0.0000
                                                         Log likelihood  = -229.25875
                                                         Pseudo R2       =     0.0829
 ------------------------------------------------------------------------------
       admit |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
 ------------+----------------------------------------------------------------
         gre |   .0022644    .001094     2.07   0.038     .0001202    .0044086
         gpa |   .8040377   .3318193     2.42   0.015     .1536838    1.454392
             |
        rank |
          2  |  -.6754429   .3164897    -2.13   0.033    -1.295751   -.0551346
          3  |  -1.340204   .3453064    -3.88   0.000    -2.016992   -.6634158
          4  |  -1.551464   .4178316    -3.71   0.000    -2.370399   -.7325287
             |
       _cons |  -3.989979   1.139951    -3.50   0.000    -6.224242   -1.755717
 ------------------------------------------------------------------------------
```

- In the output above, we first see the iteration log, indicating how quickly the model converged. The log likelihood (-229.25875) can be used in comparisons of nested models, but we won't show an example of that here.
- Also at the top of the output we see that all 400 observations in our data set were used in the analysis (fewer observations would have been used if any of our variables had missing values).
- The likelihood ratio chi-square of 41.46 with a p-value of 0.0001 tells us that our model as a whole fits significantly better than an empty model (i.e., a model with no predictors).
- In the table we see the coefficients, their standard errors, the z-statistic, associated p-values, and the 95% confidence interval of the coefficients. Both **gre** and **gpa** are statistically significant, as are the three indicator variables for **rank**. The logistic regression coefficients give the

change in the log odds of the outcome for a one unit increase in the predictor variable.

- For every one unit change in **gre**, the log odds of admission (versus non-admission) increases by 0.002.
- For a one unit increase in **gpa**, the log odds of being admitted to graduate school increases by 0.804.
- The indicator variables for **rank** have a slightly different interpretation. For example, having attended an undergraduate institution with **rank** of 2, versus an institution with a **rank** of 1, decreases the log odds of admission by 0.675.

We can test for an overall effect of **rank** using the **test** command. Below we see that the overall effect of **rank** is statistically significant.

```
test 2.rank 3.rank 4.rank
( 1)  [admit]2.rank = 0
( 2)  [admit]3.rank = 0
( 3)  [admit]4.rank = 0
          chi2(  3) =   20.90
        Prob > chi2 =    0.0001
```

We can also test additional hypotheses about the differences in the coefficients for different levels of rank. Below we test that the coefficient for **rank**=2 is equal to the coefficient for **rank**=3. (Note that if we wanted to estimate this difference, we could do so using the **lincom** command.)

```
test 2.rank = 3.rank
( 1)  [admit]2.rank - [admit]3.rank = 0
                chi2(  1) =    5.51
          Prob > chi2 =    0.0190
```

You can also exponentiate the coefficients and interpret them as odds-ratios. Stata will do this computation for you if you use the **or** option, illustrated below. You could also use the **logistic** command.

```
logit , or
Logistic regression                               Number of obs   =        400
                                                  LR chi2(5)      =      41.46
                                                  Prob > chi2     =     0.0000
                                                  Log likelihood  = -229.25875
                                                  Pseudo R2       =     0.0829
------------------------------------------------------------------------------
       admit | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gre |   1.002267   .0010965     2.07   0.038      1.00012    1.004418
         gpa |   2.234545   .7414652     2.42   0.015     1.166122    4.281877
             |
```

```
      rank |
        2  |   .5089309   .1610714    -2.13   0.033     .2736922    .9463578
        3  |   .2617923   .0903986    -3.88   0.000     .1330551    .5150889
        4  |   .2119375   .0885542    -3.71   0.000     .0934435    .4806919
------------------------------------------------------------------------------
```

Now we can say that for a one unit increase in **gpa**, the odds of being admitted to graduate school (versus not being admitted) increase by a factor of 2.23. For more information on interpreting odds ratios see our FAQ page How do I interpret odds ratios in logistic regression? .

You can also use predicted probabilities to help you understand the model. You can calculate predicted probabilities using the **margins** command. Below we use the **margins** command to calculate the predicted probability of admission at each level of **rank**, holding all other variables in the model at their means. For more information on using the **margins** command to calculate predicted probabilities, see our page Using margins for predicted probabilities.

```
margins rank, atmeans
```

```
Adjusted predictions                              Number of obs   =        400
Model VCE    : OIM
Expression   : Pr(admit), predict()
at           : gre             =       587.7 (mean)
               gpa             =      3.3899 (mean)
               1.rank          =       .1525 (mean)
               2.rank          =       .3775 (mean)
               3.rank          =       .3025 (mean)
               4.rank          =       .1675 (mean)
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       rank  |
         1   |   .5166016   .0663153     7.79   0.000     .3866261    .6465771
         2   |   .3522846   .0397848     8.85   0.000     .2743078    .4302614
         3   |    .218612   .0382506     5.72   0.000     .1436422    .2935819
         4   |   .1846684   .0486362     3.80   0.000     .0893432    .2799937
------------------------------------------------------------------------------
```

In the above output we see that the predicted probability of being accepted into a graduate program is 0.51 for the highest prestige undergraduate institutions (rank=1), and 0.18 for the lowest ranked institutions (rank=4), holding **gre** and **gpa** at their means.

Below we generate the predicted probabilities for values of **gre** from 200 to 800 in increments of 100. Because we have not specified either **atmeans** or used **at(**...**)** to specify values at with the other predictor variables are held, the values in the table are average predicted probabilities calculated using the sample values of the other predictor variables. For example, to calculate the average predicted probability when **gre** = 200, the predicted probability was calculated for each case, using that case's values of **rank** and **gpa**, with **gre** set to 200.

```
margins , at(gre=(200(100)800))  vsquish
```

```
Predictive margins                                Number of obs   =        400
Model VCE    : OIM
```

```
Expression   : Pr(admit), predict()
1._at        : gre             =         200
2._at        : gre             =         300
3._at        : gre             =         400
4._at        : gre             =         500
5._at        : gre             =         600
6._at        : gre             =         700
7._at        : gre             =         800
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   .1667471   .0604432     2.76   0.006     .0482807    .2852135
          2  |    .198515   .0528947     3.75   0.000     .0948434    .3021867
          3  |   .2343805   .0421354     5.56   0.000     .1517966    .3169643
          4  |   .2742515   .0296657     9.24   0.000     .2161078    .3323951
          5  |   .3178483    .022704    14.00   0.000     .2733493    .3623473
          6  |   .3646908   .0334029    10.92   0.000     .2992224    .4301592
          7  |   .4141038   .0549909     7.53   0.000     .3063237    .5218839
------------------------------------------------------------------------------
```

In the table above we can see that the mean predicted probability of being accepted is only 0.167 if one's GRE score is 200 and increases to 0.414 if one's GRE score is 800 (averaging across the sample values of **gpa** and **rank**).

It can also be helpful to use graphs of predicted probabilities to understand and/or present the model.

Another example for logit interpretation
http://www.ats.ucla.edu/stat/stata/output/stata_logistic.htm

Probit regression

Below we use the **probit** command to estimate a probit regression model. The **i.** before **rank** indicates that **rank** is a factor variable (i.e., categorical variable), and that it should be included in the model as a series of indicator variables. Note that this syntax was introduced in Stata 11.

```
probit admit gre gpa i.rank
Iteration 0:   log likelihood = -249.98826
Iteration 1:   log likelihood = -229.29667
Iteration 2:   log likelihood = -229.20659
Iteration 3:   log likelihood = -229.20658

Probit regression                                 Number of obs   =        400
                                                  LR chi2(5)      =      41.56
                                                  Prob > chi2     =     0.0000
                                                  Log likelihood  = -229.20658
                                                  Pseudo R2       =     0.0831
------------------------------------------------------------------------------
       admit |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gre |   .0013756   .0006489     2.12   0.034     .0001038    .0026473
         gpa |   .4777302   .1954625     2.44   0.015     .0946308    .8608297
```

```
        rank |
           2 |  -.4153992   .1953769    -2.13   0.033    -.7983308   -.0324675
           3 |   -.812138   .2085956    -3.89   0.000    -1.220978   -.4032981
           4 |   -.935899   .2456339    -3.81   0.000    -1.417333   -.4544654
             |
       _cons |  -2.386838   .6740879    -3.54   0.000    -3.708026   -1.065649
------------------------------------------------------------------------------
```

- In the output above, we first see the iteration log, indicating how quickly the model converged. The log likelihood (-229.20658) can be used in comparisons of nested models, but we won't show an example of that here.
- Also at the top of the output we see that all 400 observations in our data set were used in the analysis (fewer observations would have been used if any of our variables had missing values).
- The likelihood ratio chi-square of 41.56 with a p-value of 0.0001 tells us that our model as a whole is statistically significant, that is, it fits significantly better than a model with no predictors.
- In the table we see the coefficients, their standard errors, the z-statistic, associated p-values, and the 95% confidence interval of the coefficients. Both **gre**, **gpa**, and the three indicator variables for **rank** are statistically significant. The probit regression coefficients give the change in the z-score or probit index for a one unit change in the predictor.
    - For a one unit increase in **gre**, the z-score increases by 0.001.
    - For each one unit increase in **gpa**, the z-score increases by 0.478.
    - The indicator variables for **rank** have a slightly different interpretation. For example, having attended an undergraduate institution of **rank** of 2, versus an institution with a **rank** of 1 (the reference group), decreases the z-score by 0.415.

Another example for probit interpretation
http://www.ats.ucla.edu/stat/stata/output/Stata_Probit.htm