PROBIT REGRESSION | STATA DATA ANALYSIS EXAMPLES

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

Probit regression, also called a probit model, is used to model dichotomous or binary outcome variables. In the probit model, the inverse standard normal distribution of the probability is modeled as a linear combination of the predictors.

Please Note: The purpose of this page is to show how to use various data analysis commands. It does not cover all aspects of the research process which researchers are expected to do. In particular, it does not cover data cleaning and checking, verification of assumptions, model diagnostics and potential follow-up analyses.

Examples of probit 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 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 https://stats.idre.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** is ordinal, it 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. We will treat **rank** as categorical.

```
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

tab admit rank

           |                    rank
     admit |         1          2          3          4 |     Total
-----------+--------------------------------------------+----------
         0 |        28         97         93         55 |       273
         1 |        33         54         28         12 |       127
-----------+--------------------------------------------+----------
     Total |        61        151        121         67 |       400
```

Analysis methods you might consider

Below is a list of some analysis methods you may have encountered. Some of the methods listed are quite reasonable while others have either fallen out of favor or have limitations.

- Probit regression, the focus of this page.
- Logistic regression. A logit model will produce results similar

probit regression. The choice of probit versus logit depends largely on individual preferences.

- OLS regression. When used with a binary response variable, this model is known as a linear probability model and can be used as a way to describe conditional probabilities. However, the errors (i.e., residuals) from the linear probability model violate the homoskedasticity and normality of errors assumptions of OLS regression, resulting in invalid standard errors and hypothesis tests. For a more thorough discussion of these and other problems with the linear probability model, see Long (1997, p. 38-40).
- Two-group discriminant function analysis. A multivariate method for dichotomous outcome variables.
- Hotelling's T^2. The 0/1 outcome is turned into the grouping variable, and the former predictors are turned into outcome variables. This will produce an overall test of significance but will not give individual coefficients for each variable, and it is unclear the extent to which each "predictor" is adjusted for the impact of the other "predictors."

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.

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) =   21.32
         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.

```
test 2.rank = 3.rank

 ( 1)  [admit]2.rank - [admit]3.rank = 0

           chi2(  1) =    5.60
         Prob > chi2 =    0.0179
```

You can also use predicted probabilities to help you understand the model. You can calculate predicted probabilities using the **margins** command, which was introduced in Stata 11. 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 (/stata/dae/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  |   .5163741   .0656201     7.87   0.000     .3877611    .6449871
          2  |   .3540742   .0394725     8.97   0.000     .2767096    .4314388
          3  |   .2203289   .0383674     5.74   0.000     .1451302    .2955277
          4  |   .1854353   .0487112     3.81   0.000     .0899631    .2809075
------------------------------------------------------------------------------
```

In the above output we see that the predicted probability of being accepted into a graduate program is 0.52 for the highest prestige undergraduate institutions (rank=1), and 0.19 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 which 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 value of **rank** and **gpa**, and setting **gre** 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  |   .1621325   .0621895     2.61   0.009     .0402434    .2840216
          2  |   .1956415    .053758     3.64   0.000     .0902777    .3010054
          3  |   .2330607   .0422138     5.52   0.000     .1503231    .3157983
          4  |   .2741667   .0293439     9.34   0.000     .2166537    .3316797
          5  |   .3185876   .0226349    14.08   0.000     .2742239    .3629512
          6  |    .365808   .0333436    10.97   0.000     .3004557    .4311603
          7  |   .4151847   .0541532     7.67   0.000     .3090463    .5213231
------------------------------------------------------------------------------
```

In the table above we can see that the mean predicted probability of being accepted is only 0.16 if one's **GRE** score is 200 and increases to 0.42 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.

We may also wish to see measures of how well our model fits. This can be particularly useful when comparing competing models. The user-written command **fitstat** produces a variety of fit statistics. You can find more information on **fitstat** by typing **search fitstat** (see How can I use the search command to search for programs and get additional help? (https://stats.idre.ucla.edu/stata/faq/search-faq/) for more information about using **search**).

```
fitstat

Measures of Fit for probit of admit

Log-Lik Intercept Only:       -249.988   Log-Lik Full Model:           -229.207
D(393):                        458.413   LR(5):                          41.563
                                         Prob > LR:                       0.000
McFadden's R2:                   0.083   McFadden's Adj R2:               0.055
ML (Cox-Snell) R2:               0.099   Cragg-Uhler(Nagelkerke) R2:      0.138
McKelvey & Zavoina's R2:         0.165   Efron's R2:                      0.101
Variance of y*:                  1.197   Variance of error:               1.000
Count R2:                        0.710   Adj Count R2:                    0.087
AIC:                             1.181   AIC*n:                         472.413
BIC:                         -1896.232   BIC':                          -11.606
BIC used by Stata:             494.362   AIC used by Stata:             470.413
```

Things to consider

- Empty cells or small cells: You should check for empty or small cells by doing a crosstab between categorical predictors and the outcome variable. If a cell has very few cases (a small cell), the model may become unstable or it might not run at all.
- Separation or quasi-separation (also called perfect prediction), a condition in which the outcome does not vary at some levels of the independent variables. See our page FAQ: What is complete or quasi-complete separation in logistic/probit regression and how do we deal with them? (https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faqwhat-is-complete-or-quasi-complete-separation-in-logisticprobit-regression-and-how-do-we-deal-with-them/) for information on models with perfect prediction.
- Sample size: Both probit and logit models require more cases than OLS regression because they use maximum likelihood estimation techniques. It is sometimes possible to estimate models for binary outcomes in datasets with only a small number of cases using exact logistic regression (using the **exlogistic** command). For more information see our data analysis example for exact logistic regression (/stata/dae/exact-logistic-regression/). It is also important to keep in mind that when the outcome is rare, even if the overall dataset is large, it can be difficult to estimate a probit model.
- Pseudo-R-squared: Many different measures of psuedo-R-squared exist. They all attempt to provide information similar to that provided by R-squared in OLS regression; however, none of them can be interpreted exactly as R-squared in OLS regression is interpreted. For a discussion of various pseudo-R-squareds see Long and Freese (2006) or our FAQ page What are pseudo R-squareds? (https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/)

- In Stata, values of 0 are treated as one level of the outcome variable, and all other non-missing values are treated as the second level of the outcome.
- Diagnostics: The diagnostics for probit regression are different from those for OLS regression. The diagnostics for probit models are similar to those for logit models. For a discussion of model diagnostics for logistic regression, see Hosmer and Lemeshow (2000, Chapter 5).

See also

- Stata help for probit (http://www.stata.com/help.cgi?logit)
- Annotated output for the probit command (/stata/output/probit-regression/)
- Stat Books for Loan, Logistic Regression and Limited Dependent Variables (/books/#Logistic Regression and Related Methods)

References

- Hosmer, D. & Lemeshow, S. (2000). Applied Logistic Regression (Second Edition). New York: John Wiley & Sons, Inc.
- Long, J. Scott (1997). Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks, CA: Sage Publications.

Click here to report an error on this page or leave a comment

How to cite this page (https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-cite-web-pages-and-programs-from-the-ucla-statistical-consulting-group/)