

How can we use Probit Regression in Stata to analyze our data?

Authored by
stats writer

June 29, 2024

RECOMMENDED CITATION

stats writer (2024). *How can we use Probit Regression in Stata to analyze our data?*.
PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=160213>

Probit Regression is a statistical method used to analyze binary or categorical data in Stata. It allows us to understand the relationship between a binary response variable and a set of explanatory variables. The Probit model estimates the probability of the binary response variable taking a specific value, given the values of the explanatory variables. This method is particularly useful in situations where the dependent variable is a binary outcome, such as yes or no, success or failure, etc. In Stata, we can use the "probit" command to run the Probit Regression and obtain results such as coefficient estimates, standard errors, and significance levels. This allows us to identify the significant factors that influence the probability of the binary outcome, making it a valuable tool for data analysis.

Probit Regression | Stata Annotated Output

This page shows an example of probit regression analysis with footnotes explaining the output in Stata. The data in this example were gathered on undergraduates applying to graduate school and includes undergraduate GPAs, the reputation of the school of the undergraduate (a topnotch indicator), the students' GRE score, and whether or not the student was admitted to graduate school. Using this dataset, we can predict admission to graduate school using undergraduate GPA, GRE scores, and the reputation of the school of the undergraduate. Our outcome variable is binary, and we will use a probit

model. Thus, our model will calculate a predicted probability of admission based on our predictors. The probit model does so using the cumulative distribution function of the standard normal.

First, let us examine the dataset and our response variable. Our binary outcome variable must be coded with zeros and ones, so we will check this before proceeding.

```
use https://stats.idre.ucla.edu/stat/stata/dae/logit.dta,
clear
```

```
summarize
```

```
Variable | Obs Mean Std. Dev. Min Max
```

```
-----+-----
```

```
admit | 400 .3175 .4660867 0 1
```

```
gre | 400 587.7 115.5165 220 800
```

```
topnotch | 400 .1625 .3693709 0 1
```

```
gpa | 400 3.3899 .3805668 2.26 4
```

tabulate admit

admit | Freq. Percent Cum.

-----+-----			
0		273 68.25	68.25
1		127 31.75	100.00
-----+-----			
Total		400	100.00

To run the model in Stata, we first give the response variable (admit), followed by our predictors (gre, topnotch and gpa).

probit admit gre topnotch gpa

Iteration 0: log likelihood = -249.98826

Iteration 1: log likelihood = -238.97735

Iteration 2: log likelihood = -238.94339

Iteration 3: log likelihood = -238.94339

Probit regression Number of obs = 400

LR chi2(3) = 22.09

Prob > chi2 = 0.0001

Log likelihood = -238.94339 Pseudo R2 = 0.0442

admit | Coef. Std. Err. z P>|z|

-----+-----

gre | .0015244 .0006382 2.39 0.017 .0002736 .0027752

topnotch | .2730334 .1795984 1.52 0.128 -.078973

.6250398

gpa | .4009853 .1931077 2.08 0.038 .0225012 .7794694

_cons | -2.797884 .6475363 -4.32 0.000 -4.067032

-1.528736

Iteration Historya

Iteration 0: log likelihood = -249.98826

Iteration 1: log likelihood = -238.97735

Iteration 2: log likelihood = -238.94339

Iteration 3: log likelihood = -238.94339

a. Iteration History - This is a listing of the log likelihoods at each iteration for the probit model. Remember that probit regression uses maximum likelihood estimation, which is an iterative procedure. The first iteration (called

Iteration 0) is the log likelihood of the "null" or "empty" model; that is, a model with no predictors. At the next iteration (called Iteration 1), the specified predictors are included in the model. In this example, the predictors are gre, topnotch and gre. At each iteration, the log likelihood increases because the goal is to maximize the log likelihood. When the difference between successive iterations is very small, the model is said to have "converged" and the iterating stops. For more information on this process for binary outcomes, see

Regression Models for Categorical and Limited Dependent Variables by J. Scott Long (page 52-61).

Model Summary

Probit regression Number of obs = 400

LR chi2(3) = 22.09

Prob > chi2 = 0.0001

Log likelihood = -238.94339 Pseudo R2 = 0.0442

b. Log likelihood - This is the log likelihood of the fitted model. It is used in the Likelihood Ratio Chi-Square test of whether all predictors' regression coefficients in the model are simultaneously zero.

c. Number of obs - This is the number of observations in the dataset for which all of the response and predictor variables are non-missing.

d. LR chi2(3) - This is the Likelihood Ratio (LR) Chi-Square test that at least one of the predictors' regression coefficient is not equal to zero. The number in the parentheses indicates the degrees of freedom of the Chi-Square distribution used to test the LR Chi-Square statistic and is defined by the number of predictors in the model (3).

e. Prob > chi2 - This is the probability of getting a LR test statistic as extreme as, or more so, than the observed

statistic under the null hypothesis; the null hypothesis is that all of the regression coefficients are simultaneously equal to zero. In other words, this is the probability of obtaining this chi-square statistic (22.09) or one more extreme if there is in fact no effect of the predictor variables. This p-value is compared to a specified alpha level, our willingness to accept a type I error, which is typically set at 0.05 or 0.01. The small p-value from the LR test, 0.0001, would lead us to conclude that at least one of the regression coefficients in the model is not equal to zero. The parameter of the chi-square distribution used to test the null hypothesis is defined by the degrees of freedom in the prior line, $\chi^2(3)$.

f. Pseudo R² - This is McFadden's pseudo R-squared. Probit regression does not have an equivalent to the R-

squared that is found in OLS regression; however, many people have tried to come up with one. There are a wide variety of pseudo-R-square statistics. Because this statistic does not mean what R-square means in OLS regression (the proportion of variance of the response variable explained by the predictors), we suggest interpreting this statistic with great caution. For more information on pseudo R-squareds, see [What are Pseudo R-Squareds?](#).

Parameter Estimates

```
-----+-----
admitg| Coef. Std. Err. z P>|z|
-----+-----
gre | .0015244 .0006382 2.39 0.017 .0002736 .0027752
topnotch | .2730334 .1795984 1.52 0.128 -.078973
        .6250398
gpa | .4009853 .1931077 2.08 0.038 .0225012 .7794694
_cons | -2.797884 .6475363 -4.32 0.000 -4.067032
        -1.528736
```

g. admit - This is the binary response variable predicted by the model.

h. Coef. - These are the regression coefficients. The predicted probability of admission can be calculated using these coefficients. For a given record, the predicted probability of admission is

$$F(-2.797884 + gre \times 0.0015244 + topnotch \times 0.2730334 + gpa \times 0.4009853)$$

where F is the cumulative distribution function of the standard normal. However, interpretation of the coefficients in probit regression is not as straightforward as the interpretations of coefficients in linear regression or logit regression. The increase in probability attributed to a one-unit increase in a given predictor is dependent both on the values of the other predictors and the starting value of

the given predictors.

For example, if we hold gre and topnotch constant at zero, the one

unit increase in gpa from 2 to 3 has a different effect than the one unit

increase from 3 to 4 (note that the probabilities do not change by a common difference or common factor):

$$F(-2.797884 + 2 \times 0.4009853) = 0.02297167$$

$$F(-2.797884 + 3 \times 0.4009853) = 0.05536416$$

$$F(-2.797884 + 4 \times 0.4009853) = 0.11625017$$

and the effects of these one unit increases are different if we

hold gre and topnotch constant at their respective means instead of zero:

$$F(-2.797884 + 587.7 \times 0.0015244 + 0.1625 \times 0.2730334 + 2 \times 0.4009853) = 0.1455628$$

$$F(-2.797884 + 587.7 \times 0.0015244 + 0.1625 \times 0.2730334 + 3 \times 0.4009853) = 0.25634003$$

$$F(-2.797884 + 587.7 \times 0.0015244 + 0.1625 \times 0.2730334 + 4 \times 0.4009853) = 0.39986946.$$

However, there are limited ways in which we can interpret the

individual regression coefficients. A positive coefficient means that an increase in the predictor leads to an increase in the predicted probability. A negative coefficient means that an increase in the predictor leads to a decrease in the predicted probability.

gre - The coefficient of gre is 0.0015244. This means that an increase in GRE score increases the predicted probability of admission.

topnotch - The coefficient of topnotch is 0.2730334. This means attending a top notch institution as an undergraduate increases the predicted probability of admission.

gpa - The coefficient of gpa is 0.4009853. This means that an increase in GPA increases the predicted probability of admission.

_cons - The constant term is -2.797884. This means that if all of the predictors (gre, topnotch and

gpa) are evaluated at zero, the predicted probability of admission is $F(-2.797884) = 0.002571929$. So, as expected, the predicted probability of a student with a GRE score of zero and a GPA of zero from a non-topnotch school has an extremely low predicted probability of admission.

To generate values from F in Stata, use the normal function. For example,

`display normal(0)`

will display .5, indicating that $F(0) = .5$ (i.e., half of the area under the standard normal distribution curve falls to the left of zero). The first student in our dataset has a GRE score of 380, a GPA of 3.61, and a topnotch indicator value of 0. We could multiply these values by their corresponding coefficients,

`display -2.797884 + (.0015244*380) + (.2730334*0) +`

(.4009853*3.61)

to determine that the predicted probability of admittance is $F(-0.77105507)$. To find this value, we type

display normal(-0.77105507)

and arrive at a predicted probability of 0.22033715.

i. Std. Err. - These are the standard errors of the individual regression coefficients. They are used in both the calculation of the z test statistic, superscript j, and the confidence interval of the regression coefficient, superscript i.

j. z - The test statistic z is the ratio of the Coef. to the Std. Err. of the respective predictor. The z value follows a standard normal distribution which is used to test against a two-sided alternative hypothesis that the Coef. is not equal to zero.

k. $P > |z|$ - This is the probability the z test statistic (or a more extreme test statistic) would be observed under the null hypothesis that a particular predictor's regression coefficient is zero, given that the rest of the predictors are in the model. For a given alpha level, $P > |z|$ determines whether or not the null hypothesis can be rejected. If $P > |z|$ is less than alpha, then the null hypothesis can be rejected and the parameter estimate is considered statistically significant at that alpha level.

gre - The z test statistic for the predictor gre is $(0.0015244/0.0006382) = 2.39$ with an associated p-value of 0.017. If we set our alpha level to 0.05, we would reject the null hypothesis and conclude that the regression coefficient for gre has been found to be statistically different from zero given topnotch and gpa

are in the
model.

topnotch -

The z test

statistic for the predictor topnotch is $(0.2730334/0.1795984) = 1.52$ with an associated p-value of 0.128. If we set our alpha level to 0.05, we would

fail to reject the null hypothesis and conclude that the regression coefficient for

topnotch has not been

found to be statistically different from zero given gre and gpa

are in the

model.

gpa -

The z test

statistic for the predictor gpa is $(0.4009853/0.1931077) = 2.08$ with an

associated p-value of 0.038. If we set our alpha level to 0.05, we would

reject the null hypothesis and conclude that the

regression coefficient for gpa has been found to be statistically different from zero given gre and topnotch are in the model.

_cons -The z test statistic for the intercept, **_cons**, is $(-2.797884/0.6475363) = -4.32$ with an associated p-value of < 0.001 . With an alpha level of 0.05, we would reject the null hypothesis and conclude that **_cons** has been found to be statistically different from zero given gre, topnotch and gpa are in the model and evaluated at zero.

I. - This is the Confidence Interval (CI) for an individual coefficient given that the other predictors are in the model. For a given predictor with a level of 95% confidence, we'd say that we are 95% confident that the "true" coefficient lies between the

lower and upper limit of the interval. It is calculated as the Coef. $(z_{\alpha/2}) \cdot (\text{Std.Err.})$, where $z_{\alpha/2}$ is a critical value on the standard normal distribution.

The CI is equivalent to the z test statistic: if the CI includes zero, we'd fail to reject the null hypothesis that a particular regression coefficient is zero given the other predictors are in the model. An advantage of a CI is that it is illustrative; it provides a range where the "true" parameter may lie.