# How can I perform logistic regression using Stata for data analysis?

Authored by

**stats writer**

June 29, 2024

Logistic regression is a statistical analysis technique used to model the relationship between a categorical dependent variable and one or more independent variables. It is commonly used in data analysis to predict the likelihood of a certain outcome based on a set of explanatory variables. Stata is a popular statistical software that offers a user-friendly interface for performing logistic regression. To perform logistic regression using Stata, the user must first import their data into the software and then select the appropriate logistic regression command. Stata provides options for both binary and multinomial logistic regression, as well as various model diagnostic tools to ensure the accuracy of the results. The output of the analysis includes coefficients, odds ratios, and statistical significance levels, which can be interpreted to understand the relationship between the variables and the predicted outcome. Overall, Stata's efficient and comprehensive features make it a reliable tool for conducting logistic regression and gaining insights into the data.

## Logistic Regression | Stata Data Analysis Examples

## Logistic Regression

## 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.

## 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 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 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 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

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.

### 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|
-------------+----------------------------------------------------------------
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
------------------------------------------------------------------------------

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) 2.rank = 0
( 2) 3.rank = 0
( 3) 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) 2.rank - 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|

---+---

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,**
**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.**

**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|

---+---

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|

---+---

_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.**

**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? for more information about using search).**

**fitstat**

**Measures of Fit for logit of admit**

**Log-Lik Intercept Only: -249.988 Log-Lik Full Model: -229.259**
**D(393): 458.517 LR(5): 41.459**

**Prob > LR: 0.000**

**McFadden's R2: 0.083 McFadden's Adj R2: 0.055**

**ML (Cox-Snell) R2: 0.098 Cragg-Uhler(Nagelkerke) R2: 0.138**

**McKelvey & Zavoina's R2: 0.142 Efron's R2: 0.101**

**Variance of y*: 3.834 Variance of error: 3.290**

**Count R2: 0.710 Adj Count R2: 0.087**

**AIC: 1.181 AIC*n: 472.517**

**BIC: -1896.128 BIC': -11.502**

**BIC used by Stata: 494.466 AIC used by Stata: 470.517**

## Things to consider

## See also

## References