What is Logit Regression and how is it used in SAS data analysis?

Authored by

stats writer

June 29, 2024

Logit Regression is a statistical model used to analyze and predict binary or dichotomous outcomes, where the dependent variable can take only two possible values. It is a type of regression analysis that uses the logit function to transform the dependent variable into a continuous variable, which can then be modeled using standard regression techniques. In SAS data analysis, Logit Regression is commonly used to understand the relationship between a set of independent variables and a binary outcome, such as success or failure, presence or absence, etc. It is useful in various fields, including marketing, healthcare, and social sciences, for making predictions and identifying significant factors that influence the outcome of interest. SAS offers a range of powerful tools and procedures for conducting Logit Regression, making it a popular choice for data analysts and researchers.

Logit Regression | SAS Data Analysis Examples

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

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 outcome variable, admit/don't admit, is binary.

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 by clicking on

https://stats.idre.ucla.edu/wp-content/uploads/2016/02/binary.sas7bdat.

You can store this anywhere you like, but the syntax below assumes it has been stored in the directory c:data.

This data set has a binary response (outcome, dependent) variable called admit, which is equal to 1 if the individual was admitted to graduate school, and 0 otherwise. 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. We start out by looking at some descriptive statistics.

```
proc means data="c:databinary";
var gre gpa;
run;
```

The MEANS Procedure

```
Variable N Mean Std Dev Minimum Maximum
----------------------------------------------------------------------
GRE 400 587.7000000 115.5165364 220.0000000
800.0000000
GPA 400 3.3899000 0.3805668 2.2600000 4.0000000
----------------------------------------------------------------------

proc freq data="c:databinary";
tables rank admit admit*rank;
run;

The FREQ Procedure

Cumulative Cumulative
RANK Frequency Percent Frequency Percent
---------------------------------------------------
1 61 15.25 61 15.25
2 151 37.75 212 53.00
3 121 30.25 333 83.25
4 67 16.75 400 100.00

Cumulative Cumulative
ADMIT Frequency Percent Frequency Percent
---------------------------------------------------
0 273 68.25 273 68.25
```

1 127 31.75 400 100.00

Table of ADMIT by RANK

ADMIT RANK

Frequency|

Percent |

Row Pct |

Col Pct | 1| 2| 3| 4| Total

---------+--------+--------+--------+--------+-

0 | 28 | 97 | 93 | 55 | 273

| 7.00 | 24.25 | 23.25 | 13.75 | 68.25

| 10.26 | 35.53 | 34.07 | 20.15 |

| 45.90 | 64.24 | 76.86 | 82.09 |

---------+--------+--------+--------+--------+-

1 | 33 | 54 | 28 | 12 | 127

| 8.25 | 13.50 | 7.00 | 3.00 | 31.75

| 25.98 | 42.52 | 22.05 | 9.45 |

| 54.10 | 35.76 | 23.14 | 17.91 |

---------+--------+--------+--------+--------+-

Total 61 151 121 67 400

15.25 37.75 30.25 16.75 100.00

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.

Using the logit model

Below we run the logistic regression model. To model 1s rather than 0s, we
use the descending option. We do this because by default, proc logistic models
0s rather than 1s, in this case that would mean predicting the probability of
not getting into graduate school (admit=0) versus getting in (admit=1).
Mathematically, the models are equivalent, but conceptually, it probably makes
more sense to model the probability of getting into graduate school versus not getting in. The class statement tells SAS that rank is a
categorical variable. The param=ref option after the slash requests dummy

coding, rather than the default effects coding, for the levels of rank.
For more information on dummy versus effects coding in proc logistic, see
our FAQ page: In PROC LOGISTIC why aren't the coefficients consistent with the odds ratios?.

```
proc logistic data="c:databinary" descending;
class rank / param=ref ;
model admit = gre gpa rank;
run;
```

The output from proc logistic is broken into several sections each of which is discussed below.

The LOGISTIC Procedure

Model Information

Data Set c:databinary Written by SAS
Response Variable ADMIT
Number of Response Levels 2
Model binary logit
Optimization Technique Fisher's scoring

Number of Observations Read 400

Number of Observations Used 400

Response Profile

Ordered Total

Value ADMIT Frequency

1 1 127

2 0 273

Probability modeled is ADMIT=1.

Class Level Information

Class Value Design Variables

RANK 1 1 0 0

2 0 1 0

3 0 0 1

4 0 0 0

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Intercept
Intercept and
Criterion Only Covariates

AIC 501.977 470.517
SC 505.968 494.466
-2 Log L 499.977 458.517

Testing Global Null Hypothesis: BETA=0

Test Chi-Square DF Pr > ChiSq

Likelihood Ratio 41.4590 5 <.0001
Score 40.1603 5 <.0001
Wald 36.1390 5 <.0001

Type 3 Analysis of Effects

Wald
Effect DF Chi-Square Pr > ChiSq
GRE 1 4.2842 0.0385
GPA 1 5.8714 0.0154
RANK 3 20.8949 0.0001

Analysis of Maximum Likelihood Estimates

Standard Wald

Parameter DF Estimate Error Chi-Square Pr > ChiSq

Intercept 1 -5.5414 1.1381 23.7081 <.0001

GRE 1 0.00226 0.00109 4.2842 0.0385

GPA 1 0.8040 0.3318 5.8714 0.0154

RANK 1 1 1.5514 0.4178 13.7870 0.0002

RANK 2 1 0.8760 0.3667 5.7056 0.0169

RANK 3 1 0.2112 0.3929 0.2891 0.5908

Odds Ratio Estimates

Point 95% Wald

Effect Estimate Confidence Limits

GRE 1.002 1.000 1.004

GPA 2.235 1.166 4.282

RANK 1 vs 4 4.718 2.080 10.701

RANK 2 vs 4 2.401 1.170 4.927

RANK 3 vs 4 1.235 0.572 2.668

Association of Predicted Probabilities and Observed Responses

Percent Concordant 69.3 Somers' D 0.386
Percent Discordant 30.7 Gamma 0.386
Percent Tied 0.0 Tau-a 0.168
Pairs 34671 c 0.693

The output gives a test for the overall effect of rank, as well as coefficients
that describe the difference between the reference group (rank=4) and each of the other
three groups. We can also test for differences between the other levels of rank. For example,
we might want to test for a difference in coefficients for rank=2 and rank=3, that is, to
compare the odds of admission for students who attended a university with a rank of
2, to students who attended a
university with a rank of 3. We can test this type of hypothesis by adding a contrast
statement to the code for proc logistic. The syntax shown below is the
same as that shown above, except that it includes a contrast statement.
Following the word contrast, is the label that will appear in the output,

enclosed in single quotes (i.e., 'rank 2 vs. rank 3'). This is followed by the name of the variable we wish to test hypotheses about (i.e., rank), and a vector that describes the desired comparison (i.e., 0 1 -1). In this case the value computed is the difference between the coefficients for rank=2 and rank=3. After the slash (i.e., /) we use the estimate = parm option to request that the estimate be the difference in coefficients. For more information on use of the contrast statement, see our FAQ page: How can I create contrasts with proc logistic?.

```
proc logistic data="c:databinary" descending;
class rank / param=ref ;
model admit = gre gpa rank;
contrast 'rank 2 vs 3' rank 0 1 -1 / estimate=parm;
run;
```

Contrast Test Results

Wald

Contrast DF Chi-Square Pr > ChiSq

rank 2 vs 3 1 5.5052 0.0190

Contrast Estimation and Testing Results by Row

Standard Wald
Contrast Type Row Estimate Error Alpha Confidence Limits Chi-Square Pr > ChiSq

rank 2 vs 3 PARM 1 0.6648 0.2833 0.05 0.1095 1.2200 5.5052 0.0190

Because the models are the same, most of the output produced by the above proc logistic command is the same as before. The only difference is the additional output produced by the contrast statement. Under the heading Contrast Test Results we see the label for the contrast (rank 2 versus 3) along with its degrees of freedom, Wald chi-square statistic, and p-value. Based on the p-value in this table we know that the coefficient for rank=2 is significantly different from the coefficient for rank=3.

The second table, shows more detailed information, including the actual estimate of the difference (under Estimate), it's standard error, confidence limits, test statistic, and p-value. We can see that the estimated difference was 0.6648, indicating that having attended an undergraduate institution with a rank of 2, versus an institution with a rank of 3, increases the log odds of admission by 0.67.

You can also use predicted probabilities to help you understand the model. The contrast statement can be used to estimate predicted probabilities by specifying estimate=prob. In the syntax below we use multiple contrast statements to estimate the predicted probability of admission as gre changes from 200 to 800 (in increments of 100). When estimating the predicted probabilities we hold gpa constant at

3.39 (its mean), and rank at 2. The term intercept followed by a 1 indicates that the intercept for the model is to be included in estimate.

```
proc logistic data="c:databinary" descending;
class rank / param=ref ;
model admit = gre gpa rank;
contrast 'gre=200' intercept 1 gre 200 gpa 3.3899 rank 0 1 0 / estimate=prob;
contrast 'gre=300' intercept 1 gre 300 gpa 3.3899 rank 0 1 0 / estimate=prob;
contrast 'gre=400' intercept 1 gre 400 gpa 3.3899 rank 0 1 0 / estimate=prob;
contrast 'gre=500' intercept 1 gre 500 gpa 3.3899 rank 0 1 0 / estimate=prob;
contrast 'gre=600' intercept 1 gre 600 gpa 3.3899 rank 0 1 0 / estimate=prob;
contrast 'gre=700' intercept 1 gre 700 gpa 3.3899 rank 0 1 0 / estimate=prob;
contrast 'gre=800' intercept 1 gre 800 gpa 3.3899 rank 0 1 0 / estimate=prob;
run;
```

Contrast Test Results

Wald

Contrast DF Chi-Square Pr > ChiSq

gre=200 1 9.7752 0.0018
gre=300 1 11.2483 0.0008
gre=400 1 13.3231 0.0003
gre=500 1 15.0984 0.0001
gre=600 1 11.2291 0.0008
gre=700 1 3.0769 0.0794
gre=800 1 0.2175 0.6409

Contrast Estimation and Testing Results by Row

Standard Wald
Contrast Type Row Estimate Error Alpha Confidence Limits Chi-Square Pr > ChiSq

gre=200 PROB 1 0.1844 0.0715 0.05 0.0817 0.3648 9.7752 0.0018
gre=300 PROB 1 0.2209 0.0647 0.05 0.1195 0.3719 11.2483 0.0008
gre=400 PROB 1 0.2623 0.0548 0.05 0.1695 0.3825 13.3231 0.0003
gre=500 PROB 1 0.3084 0.0443 0.05 0.2288 0.4013 15.0984 0.0001

gre=600 PROB 1 0.3587 0.0399 0.05 0.2847 0.4400 11.2291 0.0008
gre=700 PROB 1 0.4122 0.0490 0.05 0.3206 0.5104 3.0769 0.0794
gre=800 PROB 1 0.4680 0.0685 0.05 0.3391 0.6013 0.2175 0.6409

As with the previous example, we have omitted most of the proc logistic output, because it is the same as before. The predicted probabilities are included in the column labeled Estimate in the second table shown above. Looking at the estimates, we can see that the predicted probability of being admitted is only 0.18 if one's gre score is 200, but increases to 0.47 if one's gre score is 800, holding gpa at its mean (3.39), and rank at 2.

Things to consider

References

Hosmer, D. and Lemeshow, S. (2000). Applied Logistic Regression (Second Edition).

New York: John Wiley and Sons, Inc.

Long, J. Scott (1997). Regression Models for Categorical and Limited Dependent Variables.

Thousand Oaks, CA: Sage Publications.

See also