Help the Stat Consulting Group by giving a gift

SAS Data Analysis Examples
Logit Regression

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 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:\data\binary";
  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:\data\binary";
  tables rank admit admit*rank;
run;
```

The FREQ Procedure

RANK	Frequency	Percent	Cumulative Frequency	Cumulative 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

ADMIT	Frequency	Percent	Cumulative Frequency	Cumulative 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.

- Logistic regression, the focus of this page.
- Probit regression. Probit analysis will produce results similar to logistic 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."

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:\data\binary" 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                        DATA.LOGIT           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.
```

- The first part of the above output tells us the file being analyzed (c:\data\binary) and the number of observations used. 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).
- We also see that SAS is modeling **admit** using a binary logit model and that the probability that of **admit** = 1 is being modeled. (If we omitted the **descending** option, SAS would model **admit** being 0 and our results would be completely reversed.)

```
                   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
```

- The portion of the output labeled Model Fit Statistics describes and tests the overall fit of the model. The -2 Log L (499.977) can be used in comparisons of nested models, but we won't show an example of that here.
- In the next section of output, the likelihood ratio chi-square of 41.4590 with a p-value of 0.0001 tells us that our model as a whole fits significantly better than an empty model. The Score and Wald tests are asymptotically equivalent tests of the same hypothesis tested by the likelihood ratio test, not surprisingly, these tests also indicate that the model is statistically significant.
- The section labeled Type 3 Analysis of Effects, shows the hypothesis tests for each of the variables in the model individually. The chi-square test statistics and associated p-values shown in the table indicate that each of the three variables in the model significantly improve the model fit. For **gre**, and **gpa**, this test duplicates the test of the coefficients shown below. However, for class variables (e.g., **rank**), this table gives the multiple degree of freedom test for the overall effect of the variable.

```
                         The LOGISTIC Procedure

                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
```

- The above table shows the coefficients (labeled Estimate), their standard errors (error), the Wald Chi-Square statistic, and associated p-values. The coefficients for **gre**, and **gpa** are statistically significant, as are the terms for **rank**=1 and **rank**=2 (versus the omitted category **rank**=4). 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 coefficients for the categories of rank have a slightly different interpretation. For example, having attended an undergraduate institution with a **rank** of 1, versus an institution with a **rank** of 4, increases the log odds of admission by 1.55.

Odds Ratio Estimates

Effect	Point Estimate	95% Wald 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	69.1	Somers' D	0.386
Percent Discordant	30.6	Gamma	0.387
Percent Tied	0.3	Tau-a	0.168
Pairs	34671	c	0.693

- The first table above gives the coefficients as odds ratios. An odds ratio is the exponentiated coefficient, and can be interpreted as the multiplicative change in the odds for a one unit change in the predictor variable. For example, for a one unit increase in **gpa**, the odds of being admitted to graduate school (versus not being admitted) increase by a factor of 2.24. For more information on interpreting odds ratios see our FAQ page: How do I interpret odds ratios in logistic regression?

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:\data\binary" 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

Contrast	DF	Wald Chi-Square	Pr > ChiSq
rank 2 vs. 3	1	5.5052	0.0190

Contrast Rows Estimation and Testing Results

Contrast	Type	Row	Estimate	Standard Error	Alpha	Confidence Limits		Cl
rank 2 vs. 3	PARM	1	0.6648	0.2833	0.05	0.1095	1.2200	

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:\data\binary" 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

Contrast	DF	Wald 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 Rows Estimation and Testing Results

Contrast	Type	Row	Estimate	Standard Error	Alpha	Confidence Limits		Chi-
gre=200	PROB	1	0.1844	0.0715	0.05	0.0817	0.3648	
gre=300	PROB	1	0.2209	0.0647	0.05	0.1195	0.3719	
gre=400	PROB	1	0.2623	0.0548	0.05	0.1695	0.3825	
gre=500	PROB	1	0.3084	0.0443	0.05	0.2288	0.4013	
gre=600	PROB	1	0.3587	0.0399	0.05	0.2847	0.4400	
gre=700	PROB	1	0.4122	0.0490	0.05	0.3206	0.5104	
gre=800	PROB	1	0.4680	0.0685	0.05	0.3391	0.6013	

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

- 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? for information on models with perfect prediction.
- Sample size: Both logit and probit 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 (available with the **exact** option in **proc logistic**). For more information see our data analysis example for 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 logit 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?
- Diagnostics: The diagnostics for logistic regression are different from those for OLS regression. For a discussion of model diagnostics for logistic regression, see Hosmer and Lemeshow (2000, Chapter 5). Note that diagnostics done for logistic regression are similar to those done for probit regression.
- By default, **proc logistic** models the probability of the lower valued category (0 if your variable is coded 0/1), rather than the higher valued category.

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

- How do I interpret odds ratios in logistic regression?
- Why are my logistic results reversed?
- SAS Annotated Output: proc logistic
- SAS Seminar: Logistic Regression in SAS

- SAS Links by Topic: Logistic Regression
- AS Textbook Examples: Applied Logistic Regression (Second Edition) by David Hosmer and Stanley Lemeshow
- A Tutorial on Logistic Regression (PDF) by Ying So, from SUGI Proceedings, 1995, courtesy of SAS).
- Some Issues in Using PROC LOGISTIC for Binary Logistic Regression (PDF) by David C. Schlotzhauer, courtesy of SAS).
- Logistic Regression Examples Using the SAS System by SAS Institute
- Logistic Regression Using the SAS System: Theory and Application by Paul D. Allison

How to cite this page Report an error on this page or leave a comment

The content of this web site should not be construed as an endorsement of any particular web site, book, or software product by the University of California.

© 2015 UC Regents Terms of Use & Privacy Policy