Print

Summary: Lesson 5: Categorical Data Analysis

This summary contains topic summaries, syntax, and sample programs.

Topic Summaries

To go to the movie where you learned a task or concept, select a link.

Describing Categorical Data

A one-way frequency table displays frequency statistics for a categorical variable.

An association exists between two variables if the distribution of one variable changes when the value of the other variable changes. If there's no association, the distribution of the first variable is the same regardless of the level of the other variable.

To look for a possible association between two or more categorical variables, you can create a crosstabulation table. A crosstabulation table shows frequency statistics for each combination of values (or levels) of two or more variables.

To create frequency and crosstabulation tables in SAS, and request associated statistics and plots, you use the TABLES statement in the FREQUENCY procedure. You can use the PLOTS= option in the TABLES statement to request specific plots for frequency and crosstabulation tables.

When ordinal values are ordered logically, you can use more powerful statistical tests that can detect linear (ordinal) associations instead of only general associations. To logically order the values of a variable for calculations and output, you can create a new variable or you can apply a temporary format to an existing variable. The ORDER=FORMATTED option in the PROC FREQ statement tells PROC FREQ to perform calculations and display output by using the formatted values instead of the stored values.

Tests of Association

To perform a formal test of association between two categorical variables, you use the chi-square test. The Pearson chi-square test is the most commonly used of several chi-square tests. The chi-square statistic indicates the difference between observed frequencies and expected frequencies. Neither the chi-square statistic nor its *p*-value indicates the magnitude of an association.

Cramer's V statistic is one measure of the strength of an association between two categorical variables. Cramer's V statistic is derived from the Pearson chi-square statistic.

To measure the strength of the association between a binary predictor variable and a binary outcome variable, you can use an odds ratio. An odds ratio indicates how much more likely it is, with respect to odds, that a certain event, or outcome, occurs in one group relative to its occurrence in another group.

To perform a Pearson chi-square test of association and generate related measures of association, you specify the CHISQ option and other options in the TABLES statement in PROC FREQ.

For ordinal associations, the Mantel-Haenszel chi-square test is a more powerful test than the Pearson chi-square test. The Mantel-Haenszel chi-square statistic and its *p*-value indicate whether an association exists but not the magnitude of the association.

To measure the strength of the linear association between two ordinal variables, you can use the Spearman correlation statistic. The Spearman correlation is considered to be a rank correlation because it provides a degree of linearity between the ordinal variables.

To perform a Mantel-Haenszel chi-square test of association and generate related measures of association, you specify the CHISQ option and other options in the TABLES statement in PROC FREQ.

Introduction to Logistic Regression

Logistic regression is a type of statistical model that you can use to predict a categorical response, or outcome, on the basis of one or more continuous or categorical predictor variables. You select one of three types of logistic regression — binary, nominal, or ordinal — based on your response variable.

Although linear and logistic regression models have the same structure, you can't use linear regression with a binary response variable. Binary logistic regression uses a predictor variable to estimate the probability of a specific outcome. To directly model the relationship between a continuous predictor and the probability of an event or outcome, you must use a nonlinear function: the inverse logit function.

To model categorical data, you use the LOGISTIC procedure. The two required statements are the PROC LOGISTIC statement and the MODEL statement. Depending on the complexity of your analysis, you can use additional statements in PROC LOGISTIC. If your model has one or more categorical predictor variables, you must specify them in the CLASS statement. The MODEL statement specifies the response variable and can specify other information as well, such as the response variable. In the MODEL statement, the EVENT= option specifies the event category for a binary response model. To specify the type of confidence intervals you want to use, you add the CLODDS= option to the MODEL statement. PROC LOGISTIC computes Wald confidence intervals by default. You can use the PLOTS= option in the PROC LOGISTIC statement to request specific plots.

Instead of working directly with the categorical predictor variables in the CLASS statement, PROC LOGISTIC first parameterizes each predictor variable. The CLASS statement creates a set of one or more design variables that represent the information in each specified classification variable. PROC LOGISTIC uses the design variables, and not the original variables, in model calculations. Two common parameterization methods are effect coding (the method that PROC LOGISTIC uses by default) and reference cell coding. To specify a parameterization method other than the default, you use the PARAM= option in the CLASS statement. If you want to specify a reference level other than the default for a classification variable, you use the REF= variable option in the CLASS statement.

Akaike's information criterion (AIC) and the Schwarz criterion (SC) are goodness-of-fit measures that you can use to compare models. -2Log L is a goodness-of-fit measure that is not commonly used to compare models. Comparing pairs is another goodness-of-fit measure that you can use to compare models.

PROC LOGISTIC uses a 0.05 significance level and a 95% confidence interval by default. If you want to specify a different significance level for the confidence interval, you can use the ALPHA= option in the MODEL statement.

For a continuous predictor variable, the odds ratio measures the increase or decrease in odds associated with a one-unit difference of the predictor variable by default.

Multiple Logistic Regression

A multiple logistic regression model characterizes the relationship between a categorical response variable and multiple predictor variables.

One method of selecting a subset of predictor variables for a multiple logistic regression model is the backward elimination method. To specify the variable selection method in PROC LOGISTIC, you add the SELECTION= option to the MODEL statement. By default, for the backward elimination method, PROC LOGISTIC uses a 0.05 significance level to determine which variables remain in the model. If you want to change the significance level, you can use the SLSTAY= (or SLS=) option in the MODEL statement.

Multiple logistic regression uses adjusted odds ratios, which measure the effect of a single predictor variable on a response variable while holding all the other predictor variables constant.

In PROC LOGISTIC, the UNITS statement enables you to obtain customized odds ratio estimates for a specified unit of change in one or more continuous predictor variables.

In the CLASS statement, when you use the REF= option with a variable that has either a temporary or a permanent format assigned to it, you must specify the formatted value of the level instead of the stored value.

When you fit a multiple logistic regression model, the simplest approach is to consider only the main effects—the effect of each predictor individually—on the response. If you suspect that there are interactions between predictor variables, you can fit a more complex logistic regression model that includes interactions. When you use the backward elimination method with interactions in the model, PROC LOGISTIC must preserve the model hierarchy when eliminating main effects. You specify interactions in the MODEL statement.

By default, PROC LOGISTIC produces the odds ratio only for variables that are not involved in an interaction.To tell PROC LOGISTIC to produce the odds ratios for each value of a variable that is involved in an interaction, you can use the ODDSRATIO statement. To specify whether PROC LOGISTIC computes the odds ratios for a categorical variable against the reference level or against all of its levels, you can use the DIFF= option. The AT option specifies fixed levels of one or more interacting variables (also called covariates). PROC LOGISTIC computes odds ratios at each of the specified levels.

To visualize the interaction between two categorical variables, you can produce an interaction plot.

Syntax

To go to the movie where you learned a statement or option, select a link.

```
PROC FREQ DATA=SAS-data-set 'SAS-library' <option(s)>;
          TABLES=table-request(s) </ option(s)>;
          additional statements;
RUN;
```

Selected Options in PROC FREQ

Statement	Option
PROC FREQ	ORDER=
TABLES	CELLCHI2 CHISQ (Pearson and Mantel-Haenszel) CL EXPECTED MEASURES NOCOL NOPERCENT PLOTS= RELRISK

```
PROC LOGISTIC DATA=SAS-data-set<options>;
          CLASS variable <(variable_option(s)> ... </ options>;
          MODEL response<(variable_options)>=predictors </ options>;
          UNITS independent1=list ... </ options>;
          STORE <OUT=>item-store-name </ LABEL='label'>;
          ODDSRATIO <'label'> variable </ options>;
RUN;
```

Selected Options in PROC LOGISTIC

Statement	Option
PROC LOGISTIC	PLOTS=
CLASS	PARAM= REF= (general usage and usage with a formatted variable)

MODEL	ALPHA= CLODDS= EVENT= SELECTION= SLSTAY= | SLS=
ODDSRATIO	AT CL= DIFF=

Sample Programs

Examining the Distribution of Categorical Variables

```
proc freq data=statdata.sales;
   tables Purchase Gender Income
          Gender*Purchase
          Income*Purchase /
          plots=(freqplot);
   format Purchase purfmt.;
   title1 'Frequency Tables for Sales Data';
run;

ods select histogram probplot;

proc univariate data=statdata.sales;
   var Age;
   histogram Age / normal (mu=est
                    sigma=est);
   probplot Age / normal (mu=est
                   sigma=est);
   title1 'Distribution of Age';
run;

title;
```

Ordering the Values in a Frequency or Crosstabulation Table

```
data statdata.sales_inc;
   set statdata.sales;
   if Income='Low' then IncLevel=1;
   else If Income='Medium' then IncLevel=2;
   else If Income='High' then IncLevel=3;
run;

proc freq data=statdata.sales_inc;
   tables IncLevel*Purchase / plots=freq;
   format IncLevel incfmt. Purchase purfmt.;
   title1 'Create variable IncLevel to correct Income';
run;

title;
```

Performing a Pearson Chi-Square Test of Association

```
proc freq data=statdata.sales_inc;
   tables Gender*Purchase /
          chisq expected cellchi2 nocol nopercent
          relrisk;
   format Purchase purfmt.;
   title1 'Association between Gender and Purchase';
run;
```

```
title;
```

Performing a Mantel-Haenszel Chi-Square Test of Ordinal Association

```
proc freq data=statdata.sales_inc;
   tables IncLevel*Purchase / chisq measures cl;
   format IncLevel incfmt. Purchase purfmt.;
   title1 'Ordinal Association between IncLevel and Purchase?';
run;

title;
```

Fitting a Binary Logistic Regression Model

```
proc logistic data=statdata.sales_inc
              plots(only)=(effect);
   class Gender (param=ref ref='Male');
   model Purchase(event='1')=Gender;
   title1 'LOGISTIC MODEL (1):Purchase=Gender';
run;

title;
```

Fitting a Multiple Logistic Regression Model

```
proc logistic data=statdata.sales_inc
              plots(only)=(effect oddsratio);
   class Gender (param=ref ref='Male')
         IncLevel (param=ref ref='1');
   units Age=10;
   model Purchase(event='1')=Gender Age IncLevel /
         selection=backward clodds=pl;
   title1 'LOGISTIC MODEL (2):Purchase=Gender Age IncLevel';
run;

title;
```

Fitting a Multiple Logistic Regression Model with Interactions

```
proc logistic data=statdata.sales_inc
              plots(only)=(effect oddsratio);
   class Gender (param=ref ref='Male')
         IncLevel (param=ref ref='1');
   units Age=10;
   model Purchase(event='1')=Gender | Age | IncLevel @2 /
         selection=backward clodds=pl;
   title1 'LOGISTIC MODEL (3): Main Effects and 2-Way Interactions';
   title2 '/ sel=backward';
run;

title;
```

Fitting a Multiple Logistic Regression Model with All Odds Ratios

```
ods select OddsRatiosPL ORPlot;

proc logistic data=statdata.sales_inc
              plots(only)=(oddsratio);
   class Gender (param=ref ref='Male')
         IncLevel (param=ref ref='1');
   units Age=10;
   model Purchase(event='1')=Gender | IncLevel Age;
   oddsratio Age / cl=pl;
   oddsratio Gender / diff=ref at (IncLevel=all) cl=pl;
   oddsratio IncLevel / diff=ref at (Gender=all) cl=pl;
   title1 'LOGISTIC MODEL (3a): Significant Terms and All Odds Ratios';
   title2 '/ sel=backward';
run;

title;
```

Generating Predictions Using PROC PLM

```
ods select none;
proc logistic data=statdata.ameshousing3;
   class Fireplaces (ref='0') Lot_Shape_2 (ref='Regular') / param=ref;
   model Bonus(event='1')=Basement_Area|Lot_Shape_2 Fireplaces;
   units Basement_Area=100;
   store out=isbonus;
run;
ods select all;

data newhouses;
   length Lot_Shape_2 $9;
   input Fireplaces Lot_Shape_2 $ Basement_Area;
   datalines;
   0  Regular   1060
   2  Regular    775
   2  Irregular 1100
   1  Irregular  975
   1  Regular    800
   ;
run;

proc plm restore=isbonus;
   score data=newhouses out=scored_houses / ILINK;
   title 'Predictions using PROC PLM';
run;

proc print data=scored_houses;
run;
```

Statistics I: Introduction to ANOVA, Regression, and Logistic Regression

Close