Chapter 5. Categorical Data Analysis

5.1 Describing Categorical Data

Overview

	Type of Predictors		
Type of Response	**Categorical**	**Continuous**	**Categorical & Continuous**
Continuous	Analysis of Variance	Linear Regression	Analysis of Covariance (Regression with dummy variables)
Categorical	Logistic Regression or Contingency Tables	Logistic Regression	Logistic Regression

Scale of Measurement

Before analyzing, identify the measurement scale for each variable (continuous, nominal, or ordinal).

- *Nominal variables* have values with no logical ordering. (ex. Gender)
- *Ordinal variables* have values with a logical order. However, the relative distances between the values are not clear. (ex. Income)

Association

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

Frequency Tables (빈도표)

A frequency table shows the number of observations that fall in certain categories or intervals. A one-way frequency table examines one variable.

Income	Frequency	Percent	Cumulative Frequency	Cumulative Percent
High	155	36	155	36
Low	132	31	287	67
Medium	144	33	431	100

Crosstabulation Tables (분할표, 교차표; contingency table)

A crosstabulation table shows the number of observations for each combination of the row and column variables.

	Income	High	Low	Medium
Purchase	No	55	100	72
	Yes	100	32	72

The FREQ Procedure

```
PROC FREQ DATA=SAS-data-set;
    TABLES table-requests </options>;
RUN;
```

TABLES requests tables and specifies options for producing tests. The general form of a table request is *variable***variable2**···, where any number of these requests can be made in a single TABLES statement. For two-way crosstabulation tables, the first variable represents the rows and the second variable represents the columns.

5.2 Tests of Association

Null Hypothesis

- There is no association between **Gender** and **Purchase**.
- The probability of purchasing items of 100 dollars or more is the same whether you are male or female.

Alternative Hypothesis

- There is an association between **Gender** and **Purchase**.
- The probability of purchasing items over 100 dollars is different between males and females.

Chi-Square Test (카이제곱검정)

The chi-square test measures the difference between the observed cell frequencies and the cell frequencies that are expected if there is no association between the variables. If you have a significant chi-square statistic, there is strong evidence that an association exists between your variables.

NO ASSOCIATION
observed frequencies = expected frequencies
ASSOCIATION
observed frequencies ≠ expected frequencies

Chi-square tests and the corresponding *p*-values

- determine whether an association exists
- do not measure the strength of an association
  - the *p*-value for the chi-square test only indicates how confident you can be that the null hypothesis of no association is false. It does not tell you the magnitude of an association.
- depend on and reflect the sample size

Measure of Association

One measure of the strength of the association between two nominal variables is Cramer's V statistic. It is in the range of -1 to 1 for 2-by-2 tables and 0 to 1 for larger tables. Values further away from 0 indicate the presence of a relatively strong association. Cramer's V statistic is derived from Pearson chi-square statistic.

Odds Ratios

An *odds ratio* indicates how much more likely, with respect to odds, a certain event occurs in one group relative to its occurrence in another group.

Odds

The *odds* of an outcome are the ratio of the expected probability that the outcome will occur to the expected probability that the outcome will not occur.

Properties of the Odds Ratio, B to A

The odds ratio shows the strength of the association between the predictor variable and the outcome variable. If the odds ratio is 1, then there is no association between the predictor variable and the outcome. If the odds ratio is greater than 1, then group B is more likely to have the outcome. If the odds ratio is less than 1, then group A is more likely to have the outcome.

When Not to Use the Chi-Square Test

When more than 20% of the cells have expected cell frequencies of less than 5, the chi-square test might not be valid. This is because the *p*-values are based on the assumption that the test statistic follows a particular distribution when the sample size is sufficiently large.

Small Samples-Exact *p*-Values

Exact *p*-values are useful when the sample size is small, in which case the asymptotic *p*-values might not be useful. However, large data sets can require a prohibitive amount of time and memory for computing exact *p*-values.

Exact *p*-Values for Pearson Chi-Square

Observed Table		
0	3	3
2	2	4
2	5	7

χ^2=2.100
prob=.286

Possible Table 2		
1	2	3
1	3	4
2	5	7

χ^2=0.058
prob=.571

Possible Table 3		
2	1	3
0	4	4
2	5	7

χ^2=3.733
prob=.143

$$\chi^2 = \sum \frac{(\text{관측값} - \text{기댓값})^2}{\text{기댓값}}$$

Exact *p*-value is the sum of probabilities of all tables with χ^2 values as great or greater than that of the Observed Table:

p-value = .286 + .143 = .429

The exact p-value would be 0.429, which means you have a 42.9% chance of obtaining a table with at least as much of an association as the observed table simply by random chance.

Mantel-Haenszel Chi-Square Test

The Mantel-Haenszel chi-square test is particularly sensitive to ordinal associations. An ordinal association implies that as one variable increases, the other variable tends to increase or decrease.

The Mantel-Haenszel chi-square test

- determines whether an ordinal association exists
- does not measure the strength of the ordinal association
- depends upon and reflects the sample size

Spearman Correlation Statistic

To measure the strength of the ordinal association, you can use the Spearman correlation statistic. This statistic

- has a range between -1 and 1
- has values close to 1 if there is a relatively high degree of positive correlation
- has values close to -1 if there is a relatively high degree of negative correlation
- is appropriate only if both variables are ordinal scaled and the values are in logical order

Spearman versus Pearson

- The Spearman correlation uses ranks of the data
- The Pearson correlation uses the observed values when the variable is numeric

5.3 Introduction to Logistic Regression

Overview

Regression analysis enables you to characterize the relationship between a response variable and one or more predictor variables. In *linear regression*, the response variable is continuous. In *logistic regression*, the response variable is categorical.

Logit Transformation

Logistic regression models transform probabilities called logits*.

$$\text{logit}(\text{p}_i) = \ln\left(\frac{p_i}{(1-p_i)}\right)$$

where

i indexes all cases (observations)

p_i is the probability the event (a sale, for example) occurs in the i^{th} case

*the *logit* is the natural log of the odds.

Logistic Regression Model

$$\text{logit}(\text{p}_i) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

where

$\text{logit}(\text{p}_i)$ logit of the probability of the event

β_0 intercept of the regression equation

β_k parameter estimate of the k^{th} predictor variable

LOGISTIC Procedure

```
PROC LOGISTIC DATA=SAS-data-set <options>;
    CLASS variables </options>;
    MODEL response=predictors </options>;
    UNITS independent1=list... </options>;
    ODDSRATIO <'label'> variable </options>;
    OUTPUT OUT=SAS-data-set keyword=name </ options>;
RUN;
```

CLASS names the classification variables to be used in the analysis. The **CLASS** statement must precede the **MODEL** statement.

MODEL specifies the response variable and the predictor variables.

OUTPUT creates an output data set containing all the variables from the input data set and any requested statistics.

UNITS enables you to obtain an odds ratio estimate for a specified change in a predictor variable. The unit of change can be a number, standard deviation (SD) or a number times the standard deviation.

ODDSRATIO produces odds ratios for variables even when the variables are involved in interactions with other covariates, and for classification variables that use any parameterization. You can specify several **ODDSRATIO** statements.

Effect (Default) Coding: Three Levels

			Design Variables	
CLASS	Value	Label	1	2
IncLevel	1	Low Income	1	0
	2	Medium Income	0	1
	3	High Income	-1	-1

For *effect coding* (also called *deviation from the mean coding*), the number of design variables created is the number of the **CLASS** variable minus 1. For the last level of the **CLASS** variable (High), all the design variables have a value of −1. Parameter estimates of the **CLASS** main effects using this coding scheme estimate the difference between the effect of each level and the average effect over all levels.

Effect Coding: An Example

$$\text{logit}(p) = \beta_0 + \beta_1 * D_{\text{Low income}} + \beta_2 * D_{\text{Medium income}}$$

β_0 = the average value of the logit across all categories

β_1 = the difference between the logit for Low income and the average logit

β_2 = the difference between the logit for Medium income and the average logit

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.5363	0.1015	27.9143	<.0001
IncLevel	1	1	-0.2259	0.1481	2.3247	0.1273
IncLevel	2	1	-0.2200	0.1447	2.3111	0.1285

If you use Effect Coding for a **CLASS** variable, then the parameter estimates and *p*-values (as well as odds ratios) reflect differences from the overall mean value over all levels. So, for **IncLevel1**, the Estimate shows the estimated difference in logit values between **IncLevel**=1 (Low Income) and the average logit across all income levels.

Reference Cell Coding: Three Levels

			Design Variables	
CLASS	Value	Label	1	2
IncLevel	1	Low Income	1	0
	2	Medium Income	0	1
	3	High Income	0	0

For *reference cell coding*, parameter estimates of the **CLASS** main effects estimate the difference between the effect of each level and the last level, called the reference level. For example, the effect for the level **Low** estimates the difference between **Low** and **High**.

Reference Cell Coding: An Example

$$\text{logit}(p) = \beta_0 + \beta_1 * D_{\text{Low income}} + \beta_2 * D_{\text{Medium income}}$$

β_0 = the value of the logit when income is High

β_1 = the difference between the logits for Low and High income

β_2 = the difference between the logits for Medium and High income

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.0904	0.1608	0.3159	0.5741
IncLevel	1	1	-0.6717	0.2465	7.4242	0.0064
IncLevel	2	1	-0.6659	0.2404	7.6722	0.0056

Now, the parameter Estimate and *p*-value for **IncLevel**=1 reflect the difference between **IncLevel**=1 and **IncLevel**=3 (the reference level). It is important to know what type of parameterization you are using in order to interpret and report the results of this table.

Odds Ratio Calculation from the Current Logistic Regression Model

Logistic regression model:

$$\text{logit}(\hat{p}) = \log(odds) = \beta_0 + \beta_1 * (gender)$$

Odds ratio (females to males)"

$$\text{odds}_{females} = e^{\beta_0+\beta_1}$$

$$\text{odds}_{males} = e^{\beta_0}$$

$$\text{odds ratio} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

Model Assessment: Comparing Pairs

- Counting concordant, discordant, and tied pairs is a way to assess how well the model predicts its own data and therefore how well the model fits.
- In general, you want a high percentage of concordant pairs and a low percentage of discordant and tied pairs.

Comparing Pairs

To find concordant, discordant, and tied pairs, compare everyone who had the outcome of interest against everyone who did not.

- Concordant Pair
  - For all pairs of observations with different values of the response variable, a pair is *concordant* if the observation with the outcome has a **higher** predicted outcome probability (based on the model) than the observation without the outcome.
- Discordant Pair
  - A pair is *discordant* if the observation with the outcome has a **lower** predicted probability than the observation without the outcome.
- Tied Pair
  - A pair is *tied* if it is neither concordant nor discordant (the probabilities are the same).

Model: Concordant, Discordant, and Tied Pairs

In general, higher percentages of concordant pairs and lower percentages of discordant pairs indicate a more desirable model.

The four rank correlation indices (Somer's D, Gamma, Tau-a, and c) are computed from the numbers of concordant, discordant, and tied pairs of observations. In general, a model with higher values for these indices has better predictive ability than a model with lower values for these indices.

5.4 Multiple Logistic Regression

In multiple logistic regression models, several continuous or categorical predictor variables are trying to explain the variability of the response variable. The goal in multiple logistic regression is similar to that in linear multiple regression. Find the best subset of variables by eliminating unnecessary ones.

Adjusted Odds Ratio

Adjusted odds ratios measure the effect between a predictor variable and a response variable while holding all the other predictor variables constant.

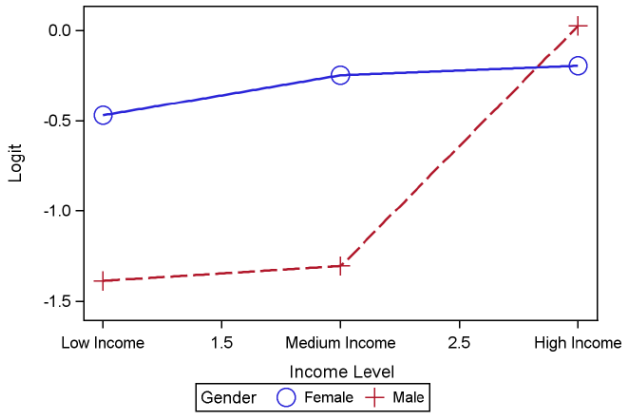
For example, the odds ratio for the variable **Gender** would measure the effect of **Gender** on **Purchase** while holding **Income** and **Age** constant.

Multiple Logistic Regression

A multiple logistic regression model with only the main effects might not be enough in explaining the effect of a variable on the outcome. In this case, you might want to fit a more complex model that has interactions. An *interaction* between two variables A and B is said to occur when the effect of A on the outcome depends on the observed level of B, or when the effect of B on the outcome depends on the observed level of A.

When you use the backward elimination method with interactions in the model, note that all effects contained by that effect must also be in the model. This requirement is called *model hierarchy*.

Interaction Plot



To visualize the interaction between Gender and Income, you could produce an interaction plot. If there is no interaction between Gender and Income, then the slopes should be relatively parallel. However, the graph above shows that the slopes are not parallel. The reason for the interaction is that the probability of making purchases of 250 dollars or more is highly related to income for men but is weakly related to income for women.