Paper GH-09

Analysis of a Binary Outcome Variable Using the FREQ and the LOGISTIC Procedures

Arthur X. Li, City of Hope Comprehensive Cancer Center, Duarte, CA

ABSTRACT

A common application in the health care industry is to investigate the association between a response variable interest that has dichotomous outcome such as having or not having a certain disease with one or more variables. This type of study can be analyzed by using the FREQ procedure if there are only one or two explanatory categorical variables. A more general approach to study a binary outcome variable would be building a logistic regression model by using the LOGISTIC procedure which can handle one or more categorical or continuous independent variables. In this talk, in addition to reviewing both PROC FREQ and PROC LOGISTIC, other model building issues including detecting confounding variables and identifying effect modifiers will also be addressed.

BACKGROUND

RELATIVE RISK AND ODDS RATIOS

One of the starting points in analyzing the association between two categorical variables is to construct a contingency table, which is a format for displaying data that is classified by two different variables. For purposes of simplicity, assume that the outcome variable (Y) takes on only two possible values and the explanatory variable (X) has only two levels.

		Variable Y	
		Y = 1	Y = 0
Variable X	X = 1	*A*	*B*
	X = 0	*C*	*D*

When analyzing data in a contingency table, you will often want to compare the proportions of having a certain outcome (Y = 1) across different levels of explanatory variables (Variable X). For example, in the above contingency table, you are interested in comparing P_1 and P_0, where

$P_1 = A/(A+B)$
$P_0 = C/(C+D)$

There are three ways to compare P_1 and P_0:

1. Difference in Proportions: $P_1 - P_0$
2. Relative Risk (RR) or Prevalence Ratio: P_1/P_0
3. Odds Ratio (OR): $[P_1/ 1 - P_1] / [P_0 /1 - P_0] = AD/BC$

The difference in proportions is more useful when both P_1 and P_0 are close to either 0 or 1 than in the middle of the range. For example, if you compare the proportion of people who had adverse effects in the drug1 group is 0.01 and 0.001 in the drug2 group, which provides a difference of 0.009. On the other hand, if the proportion is changed to 0.5 for the drug1 group and 0.491 for the drug2 group, the difference is still 0.009. However, the difference in the first comparison is more remarkable since about 10 times as many people had adverse effects with the drug1 as the drug2 group did. In this situation, comparing the proportion by using the ratio of the proportion is a much better choice.

By looking at the equation, relative risk is a ratio of the probability of the event occurring in the exposed group versus a non-exposed group. The odds ratio is the ratio of the odds of an event occurring in the exposed group compared to the odds of the event occurring in the non-exposed group. Both measurements are commonly used in clinical trials and epidemiological studies. When RR or OR equals 1, it means that there is no association between the X and the Y variables.

Compared to OR, RR is easier to interpret; it is close to what most people think when they compare relative probably of an event. Furthermore, OR tends to generate more pronounced numbers compared to RR. However, calculating OR is more common in an observational study because RR can not be calculated in all study designs.

STUDY DESIGN

An observational study can be categorized into three main study designs: cross-sectional, cohort (prospective), and case-control (retrospective) study. For example, to study the association between oral contraceptive (OC) use and having breast cancer, you can use either one of these three study designs.

For a cross-sectional study, women are recruited at a given time point and asked whether they are using OC and whether they have breast cancer. You are not taking into account whether the OC use preceded the breast cancer or having breast cancer preceded the OC use. For the cohort study, you will start with a group of women without breast cancer and assign a subgroup of them into a trial that does not use OC and assign the rest of the women to a different trial that uses OC. After a certain number of years of follow-ups, you compare the proportion of cancer cases between the OC users and the non-OC users. For the case-control study, you start with a group of women with breast cancer and a group of women without breast cancer, then look back to determine whether or not they took OC in previous years.

Regardless of the study design, you can compute the chi-square statistics (formula can be found in SAS® documentation) to test the association between the X and Y variables. You can also calculate the odds ratio for all three of these study designs, but you can only calculate relative risks for the cohort. In the cross-sectional study, P_1/P_0 is called the prevalence ratio, which is not a risk because the disease and the risk factor are collected at the same time. The relative risk can not be calculated for the case-control study because P_1 and P_0 can not be estimated. However, it can be shown that OR = RR $[(1 – P_0)/ (1 – P_1)]$, which suggests that OR can approximate RR when P_1 and P_0 are close to 0.

The focus of this paper is on computing the odds ratio, which can be done by using either the FREQ procedure or the LOGISTIC procedure.

CALCULATING THE ODDS RATIOS FROM A LOGISTIC REGRESSION MODEL

A logistic regression is used for predicting the probability occurrence of an event by fitting data to a logit function. It describes the relationship between a categorical outcome variable with one or more explanatory variables. The following equation illustrates the relationship between an outcome variable and one explanatory variable X:

$$logit(\pi) = \alpha + \beta X$$

where π is the probability of occurrence of an event (Y = 1) in the population. The logit function on the left side of the equation is defined as the following:

$$\text{logit}(\pi) = \log\frac{\pi}{1-\pi} = \log(\text{odds})$$

Solving π will yield the following equation:

$$\pi = \frac{e^{\alpha+\beta}}{1+e^{\alpha+\beta}}$$

Then the odds ratio can be simplified as the following:

$$OR(X = 1 \; vs \; X = 0) = \frac{\frac{e^{\alpha+\beta}/(1+e^{\alpha+\beta})}{1/(1+e^{\alpha+\beta})}}{\frac{e^{\alpha}/(1+e^{\alpha})}{1/(1+e^{\alpha})}} = \frac{e^{\alpha+\beta}}{e^{\alpha}} = e^{\beta}$$

Based on the equation above, the β from the logistic regression is the log odds comparing an individual with X = 1 to those with X = 0. When β equals 0, the OR will become 1. Thus, testing for no association between the X and Y variables is the same as testing $\beta = 0$. Similar to linear regression, the slope parameter β, that provides the measure of the relationship between X and Y, is used for testing the association hypothesis. For logistic regression, the maximum likelihood procedure is used to estimate the parameters.

CONFOUNDING AND INTERACTION

When studying the association between an explanatory variable (X) and the outcome of the interest (Y), we often encounter extraneous variables (Z) that may affect association of our main interest. The variable Z can be either a confounder or an effect modifier.

If variable Z is associated with both X and Y variables, then not including variable Z in the analysis can cause either over- or under-estimates of the relationship between the variable X and the variable Y, which will lead to incorrect conclusions from the hypothesis test. In this situation, variable Z is called a confounder. Consequently, adjustments for variable Z in a statistical model can remove its confounding effect.

If the relationship between the X and Y variables differs depending upon whether the Z variable is absent or not, you should report the test results separately for each level of variable Z. In this situation, variable Z is referred to as an effect modifier. Furthermore, effect modification is equivalent to "interaction between the X and the Z variable," which can be identified using statistical testing.

You can use PROC FREQ to analyze the association of your interest when there is only one confounder or one effect modifier. If you want to control multiple confounder variables or include multiple effect modifiers in your model, you need to use the PROC LOGISTIC.

THE PURPOSES AND STRATEGIES FOR MODEL BUILDING

The methods of fitting a regression model differ depending upon your research purpose. The goal of modeling building can be categorized into two reasons: one is to study the essential association between an outcome variable with a set of explanatory variables, which is commonly used in the epidemiologic field; the other is to predict the outcome variable by using a set of explanatory variables. For purposes of estimating association, interaction and confounding issues must be considered in the model building process; however, building a prediction model only needs to consider the interaction effect.

Building a prediction model is often used in situations where the research interest is more focused on statistical decision making or generating (not testing) hypotheses for a future study in a different sample. Furthermore, a prediction model needs to be validated in an independent sample to evaluate its usefulness. There are many model selection techniques for building a prediction model, such as forward, backward, and stepwise regression. The focus of this paper is not on building a prediction model but rather estimating the relationship between a main explanatory variable and an outcome variable.

For estimating relationship (the focus of this paper), all variables that are sensibly confounding the main effect should be investigated for their effects as confounders by including them in the models. For example, if variable Z alters the estimate of the odds ratio of variable Y and variable X by a certain degree or its standard error, variable Z should be controlled as a confounder in the model even though its coefficient is not statistically significant. If variable Z does not change the odds ratio of main interest or its standard error greatly, variable Z can or cannot be controlled depending on modeling tradition and the believability of an adjusted estimate comparing an unadjusted estimate of odds ratio, statistical significance of the main or adjusting variables, and the number of other confounders must be included in the model. On the other hand, if variable Z should not be a confounder, it should not be controlled in the model, regardless of whether or not variable Z changes the estimate of the odds ratio of interest and/or its standard error or is being statistically significant itself.

ANALYZING A CONTINGENCY TABLE BY USING THE FREQ PROCEDURE

ANALYZING A SIMPLE CONTINGENCY TABLE

Table 1 is taken from a study reported by Forthofer & Lehnen (1981) (Agresti, 1990). The table records the measures of Caucasians who work in certain industrial plants in Houston. The response variable of the study is breathing test results, which are either normal or abnormal. The explanatory variable is smoking status (never smoked or currently is a smoker).

Table 1. Contingency table of breathing test by smoking status

		Breathing Test	
		Abnormal	Normal
Smoking Status	Current	131	927
	Never	38	741

There are several methods of testing the association between the smoking status and the breathing test. These tests are based on the chi-square statistics, which can be calculated from the FREQ procedure. In

addition, PROC FREQ can also produce contingency tables for you. The assumption of utilizing the chi-square statistic is that the expected cell counts for each cell should exceed 5.

Program 1 starts with creating a SAS dataset by entering cell counts from the contingency table, followed by the FREQ procedure. The WEIGHT statement in PROC FREQ is used because the data is entered directly from the cell count of the table. The TABLE statement creates a 2-way contingency table with each variable separated by an asterisk (*). In the TABLE statement, the CHISQ option is used to calculate the chi-square statistics, which includes Pearson chi-square, likelihood-ratio chi-square, and Mantel-Haenszel chi-squares. These statistics are asymptotically equivalent. The formula for these statistics can be found from SAS documents. The RELRISK option is used to compute relative risks for column1 and column2, along with the odds ratio and their confidence intervals for the contingency tables.

Program 1:

```
data breathTest;
    input test $ 1-8 neversmk $ 10-16  count;
datalines;
abnormal current 131
normal   current 927
abnormal never   38
normal   never   741
;

proc freq data=breathTest;
    weight count;
    tables neversmk*test/chisq relrisk;
run;
```

Output 1.1

```
The FREQ Procedure

Table of neversmk by test

neversmk     test

Frequency|
Percent  |
Row Pct  |
Col Pct  |abnormal|normal  |  Total
---------+--------+--------+-------
current  |    131 |    927 |   1058
         |   7.13 |  50.46 |  57.59
         |  12.38 |  87.62 |
         |  77.51 |  55.58 |
---------+--------+--------+-------
never    |     38 |    741 |    779
         |   2.07 |  40.34 |  42.41
         |   4.88 |  95.12 |
         |  22.49 |  44.42 |
---------+--------+--------+-------
Total    |    169 |   1668 |   1837
         |   9.20 |  90.80 | 100.00
```

When creating a two-way contingency table, by default, PROC FREQ generates cell frequencies, cell percentages of the total frequency, and cell percentages of row and column frequencies (Output 1.1).

Output 1.2 is generated from the CHISQ option. Pearson chi-square statistics is labeled “Chi-square.” All these chi-square statistics indicate that there is a strong association between smoking status and the breathing test. Output 1.3 is generated from the RELRISK option. The calculated odds ratio is 2.8, which means that the odds of having an abnormal test result are about 2.8 times higher for current smokers compared to those who have never smoked (95% Confidence interval: 1.9 – 4.0).

Since this is a case-control study, the relative risk estimates, labeled “Cohort (Col1 Risk)” and Cohort (Col2 Risk)” should not be considered.

Output 1.2:

```
Statistics for Table of neversmk by test

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     1     30.2421    <.0001
Likelihood Ratio Chi-Square    1     32.3820    <.0001
Continuity Adj. Chi-Square     1     29.3505    <.0001
Mantel-Haenszel Chi-Square     1     30.2257    <.0001
Phi Coefficient                      0.1283
Contingency Coefficient              0.1273
Cramer's V                           0.1283
```

Output 1.3:

```
           Estimates of the Relative Risk (Row1/Row2)

Type of Study                  Value       95% Confidence Limits
-----------------------------------------------------------------
Case-Control (Odds Ratio)     2.7557        1.8962        4.0047
Cohort (Col1 Risk)            2.5383        1.7904        3.5987
Cohort (Col2 Risk)            0.9211        0.8960        0.9470

Sample Size = 1837
```

CONFOUNDING AND INTERACTION IN THE CONTINGENCY TABLE – BREATHING TEST STUDY

In addition to the breathing test and the smoking status in the breathing test study, the age variable is also collected. Age variable is categorized into two groups: "less than 40" and "40 or above." The data is displayed in Table 2. Suppose that the researcher wants to consider the confounding or the interaction effect of the age group.

Table 2. Contingency table of breathing test by smoking status stratified by age.

		Breathing Test			
		Age < 40		Age ≥ 40	
		Abnormal	Abnormal	Normal	Normal
Smoking Status	Current	57	682	74	245
	Never	34	577	4	164

In Program 2, the variable for the age group OVER40 is entered in the DATA step. In addition, the CMH option is added in the TABLES statement, which computes Cochran-Mantel-Haenszel statistics (test for association between the row and column variables after adjusting for the remaining variables). The CMH option also provides the adjusted Mantel-Haenszel and logit estimates of the odds ratio and relative risks and the Breslow-Day test for homogeneity of odds ratios.

Program 2:

```
data breathTestAge;
    input test $ 1-8 neversmk $ 10-16 over40 $ 18-20 count;
datalines;
normal   never   no  577
abnormal never   no  34
normal   current no  682
abnormal current no  57
normal   never   yes 164
abnormal never   yes 4
normal   current yes 245
abnormal current yes 74
;

proc freq data=breathTestAge;
    weight count;
    tables over40*neversmk*test/chisq relrisk cmh;
run;
```

Output 2.1:

The FREQ Procedure

Table 1 of neversmk by test
Controlling for over40=no

neversmk test

Frequency / Percent / Row Pct / Col Pct	abnormal	normal	Total
current	57	682	739
	4.22	50.52	54.74
	7.71	92.29	
	62.64	54.17	
never	34	577	611
	2.52	42.74	45.26
	5.56	94.44	
	37.36	45.83	
Total	91	1259	1350
	6.74	93.26	100.00

Statistics for Table 1 of neversmk by test
Controlling for over40=no

Statistic	DF	Value	Prob
Chi-Square	**1**	**2.4559**	**0.1171**
Likelihood Ratio Chi-Square	1	2.4893	0.1146
Continuity Adj. Chi-Square	1	2.1260	0.1448
Mantel-Haenszel Chi-Square	1	2.4541	0.1172
Phi Coefficient		0.0427	
Contingency Coefficient		0.0426	
Cramer's V		0.0427	

Statistics for Table 1 of neversmk by test
Controlling for over40=no

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	**1.4184**	**0.9144**	**2.2000**
Cohort (Col1 Risk)	1.3861	0.9190	2.0906
Cohort (Col2 Risk)	0.9772	0.9499	1.0054

Sample Size = 1350

Output 2.1 and Output 2.2 report stratum-specific chi-square statistics and odds ratio for each age group. Output 2.3 contains Mantel-Haenszel statistics and the adjusted odds ratio, along with the Breslow-Day Test for Homogeneity of the odds ratios, which resulted from the CMH option. The Mantel-Haenszel statistic is used to test the association between two categorical variables (smoking status and breathing test), adjusting for a third categorical variable (age group). The statistics and its adjusted odds ratio are only useful if there is a homogeneity in the odds ratios across each category of the adjusting variable. In other words, if stratum-specific odds ratios differ from one another, the Mantel-Haenszel adjusted odds ratio should not be reported. Instead, you need to report the stratum-specific odds ratios along with their respective p-values for association.

In this example, the Breslow-Day test of homogeneity is clearly significant ($p < 0.0001$), which means that the association between smoking status and the breathing test are not the same across different age groups. The magnitude changes in the odds ratio in different age groups further demonstrates that there is an interaction between age groups and smoking status. For people who were younger than 40, there is not a significant association between smoking status and the breathing test ($p = 0.11$), with the odds ratio equaling 1.4 (95% CI = 0.9 – 2.2); on the other hand, in the age group with people that are 40 or older, there is a strong association between smoking status and breathing test with an odds ratio equaling 12.4 (95% CI = 4.4 – 34.5).

Output 2.2:

Table 2 of neversmk by test
Controlling for over40=yes

neversmk test

Frequency Percent Row Pct Col Pct	abnormal	normal	Total
current	74	245	319
	15.20	50.31	65.50
	23.20	76.80	
	94.87	59.90	
never	4	164	168
	0.82	33.68	34.50
	2.38	97.62	
	5.13	40.10	
Total	78	409	487
	16.02	83.98	100.00

Statistics for Table 2 of neversmk by test
Controlling for over40=yes

Statistic	DF	Value	Prob
Chi-Square	**1**	**35.4510**	**<.0001**
Likelihood Ratio Chi-Square	1	45.1246	<.0001
Continuity Adj. Chi-Square	1	33.9203	<.0001
Mantel-Haenszel Chi-Square	1	35.3782	<.0001
Phi Coefficient		0.2698	
Contingency Coefficient		0.2605	
Cramer's V		0.2698	

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	**12.3837**	**4.4416**	**34.5272**
Cohort (Col1 Risk)	9.7429	3.6253	26.1844
Cohort (Col2 Risk)	0.7868	0.7374	0.8394

Sample Size = 487

Output 2.3:

```
Summary Statistics for neversmk by test
Controlling for over40

  Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic    Alternative Hypothesis    DF       Value      Prob
──────────────────────────────────────────────────────────────
    1        Nonzero Correlation        1     25.2444    <.0001
    2        Row Mean Scores Differ     1     25.2444    <.0001
    3        General Association        1     25.2444    <.0001

           Estimates of the Common Relative Risk (Row1/Row2)

Type of Study     Method                  Value     95% Confidence Limits
──────────────────────────────────────────────────────────────────────
Case-Control      Mantel-Haenszel        2.5683       1.7618       3.7441
  (Odds Ratio)    Logit                  1.9840       1.3252       2.9702

Cohort            Mantel-Haenszel        2.4174       1.6754       3.4879
  (Col1 Risk)     Logit                  1.8475       1.2641       2.7001

Cohort            Mantel-Haenszel        0.9289       0.9046       0.9538
  (Col2 Risk)     Logit                  0.9437       0.9195       0.9686

     Breslow-Day Test for
Homogeneity of the Odds Ratios
──────────────────────────────
Chi-Square              18.0829
DF                            1
Pr > ChiSq                <.0001

Total Sample Size = 1837
```

CONFOUNDING AND INTERACTION IN THE CONTINGENCY TABLE – NURSE HEALTH STUDY

The data from the following table is from the Nurses Health Study, which includes nurses aged 30 to 55 who were enrolled in 1976 and were followed-up at 2-year intervals. Part of the study investigated the association between oral contraceptive use and the effect of a woman's age.

Table 3. Contingency table of breast cancer by OC use stratified by age group

		Breast Cancer			
		Age 30 – 39		Age 40 – 55	
		Case	Control	Case	Control
OC Use	Yes	71	28418	143	20651
	No	35	12267	321	44424

In Program 3, the first TABLE statement calculates the association between OC use and having breast cancer without considering the age effect. Based on the result from Output 3.1, there is a strong association between OC use and having breast cancer (p <0.0001, OR =0.69, 95%CI = 0.59 – 0.82).

The second TABLE statement in PROC FREQ calculates the association between OC use and breast cancer considering the age effect. Output 3.2 and Output 3.3 report stratum-specific chi-square statistics and odds ratios for each age group. In output 3.4, the Breslow-Day test for homogeneity did not indicate significant departure from homogeneity (p=0.70), which means that there is no interaction effect between age group and OC use.

To exam whether or not age is a confounder, we checked the adjusted and unadjusted odds ratio of our main interest. There is no gold standard of how the degree in changes in the odds ratio from unadjusted measurements to adjusted measurements determines a confounder. For instance, a 10% or more change from an unadjusted odds ratio to an adjusted odds ratio can be considered a confounder. In this example, the age-adjusted odds ratio is 0.94, which increases more than 35% from the unadjusted odds ratio (0.69), which suggests that age is a confounder. In this situation, the age-adjusted statistics and its odds ratio should be reported. In conclusion, after adjusting for age, the odds ratio of breast cancer for oral

contraceptive users compared to nonusers is not significant (p = 0.51; age adjusted OR = 0.94, 95% CI = 0.79 – 1.13).

Program 3:

```
data nurse_study;
    input bc age oc count @@;
datalines;
1 0 1 71  0 0 1 28418
1 0 0 35  0 0 0 12267
1 1 1 143 0 1 1 20661
1 1 0 321 0 1 0 44424
;
proc freq data=nurse_study order=data;
    weight count;
    tables oc*bc/chisq relrisk;
    tables age*oc*bc/chisq relrisk cmh;
run;
```

Output 3.1:

```
The FREQ Procedure
Table of oc by bc

oc        bc

Frequency|
Percent  |
Row Pct  |
Col Pct  |       1|       0|  Total
---------+--------+--------+
       1 |     214|   49079|   49293
         |    0.20|   46.15|   46.35
         |    0.43|   99.57|
         |   37.54|   46.40|
---------+--------+--------+
       0 |     356|   56691|   57047
         |    0.33|   53.31|   53.65
         |    0.62|   99.38|
         |   62.46|   53.60|
---------+--------+--------+
Total         570   105770   106340
              0.54    99.46   100.00

Statistics for Table of oc by bc

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     1     17.8881    <.0001
Likelihood Ratio Chi-Square    1     18.1401    <.0001
Continuity Adj. Chi-Square     1     17.5337    <.0001
Mantel-Haenszel Chi-Square     1     17.8879    <.0001
Phi Coefficient                     -0.0130
Contingency Coefficient              0.0130
Cramer's V                          -0.0130

Statistics for Table of oc by bc

           Estimates of the Relative Risk (Row1/Row2)

Type of Study                  Value       95% Confidence Limits
-----------------------------------------------------------------
Case-Control (Odds Ratio)     0.6944        0.5858        0.8230
Cohort (Col1 Risk)            0.6957        0.5874        0.8239
Cohort (Col2 Risk)            1.0019        1.0010        1.0028

Sample Size = 106340
```

Output 3.2:

Table 1 of oc by bc
Controlling for age=0

oc bc

Frequency Percent Row Pct Col Pct	1	0	Total
1	71 0.17 0.25 66.98	28418 69.67 99.75 69.85	28489 69.84
0	35 0.09 0.28 33.02	12267 30.07 99.72 30.15	12302 30.16
Total	106 0.26	40685 99.74	40791 100.00

Statistics for Table 1 of oc by bc
Controlling for age=0

Statistic	DF	Value	Prob
Chi-Square	1	0.4128	0.5206
Likelihood Ratio Chi-Square	1	0.4058	0.5241
Continuity Adj. Chi-Square	1	0.2879	0.5916
Mantel-Haenszel Chi-Square	1	0.4128	0.5206
Phi Coefficient		-0.0032	
Contingency Coefficient		0.0032	
Cramer's V		-0.0032	

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	0.8757	0.5838	1.3133
Cohort (Col1 Risk)	0.8760	0.5847	1.3124
Cohort (Col2 Risk)	1.0004	0.9992	1.0015

Sample Size = 40791

Output 3.3:

Table 2 of oc by bc
Controlling for age=1

oc bc

Frequency Percent Row Pct Col Pct	1	0	Total
1	143 0.22 0.69 30.82	20661 31.52 99.31 31.74	20804 31.74
0	321 0.49 0.72 69.18	44424 67.77 99.28 68.26	44745 68.26
Total	464 0.71	65085 99.29	65549 100.00

Statistics for Table 2 of oc by bc
Controlling for age=1

Statistic	DF	Value	Prob
Chi-Square	1	0.1822	0.6695
Likelihood Ratio Chi-Square	1	0.1832	0.6687
Continuity Adj. Chi-Square	1	0.1420	0.7063
Mantel-Haenszel Chi-Square	1	0.1822	0.6695
Phi Coefficient		-0.0017	
Contingency Coefficient		0.0017	
Cramer's V		-0.0017	

Statistics for Table 2 of oc by bc
Controlling for age=1

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	0.9578	0.7860	1.1673
Cohort (Col1 Risk)	0.9581	0.7873	1.1660
Cohort (Col2 Risk)	1.0003	0.9989	1.0017

Sample Size = 65549

Output 3.4:

```
Summary Statistics for oc by bc
Controlling for age

  Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic    Alternative Hypothesis    DF       Value      Prob
------------------------------------------------------------------
    1        Nonzero Correlation        1      0.4361    0.5090
    2        Row Mean Scores Differ     1      0.4361    0.5090
    3        General Association        1      0.4361    0.5090


           Estimates of the Common Relative Risk (Row1/Row2)

Type of Study     Method                  Value     95% Confidence Limits
------------------------------------------------------------------------
Case-Control      Mantel-Haenszel        0.9419       0.7882       1.1256
  (Odds Ratio)    Logit                  0.9415       0.7882       1.1246

Cohort            Mantel-Haenszel        0.9422       0.7897       1.1243
  (Col1 Risk)     Logit                  0.9419       0.7894       1.1238

Cohort            Mantel-Haenszel        1.0003       0.9994       1.0013
  (Col2 Risk)     Logit                  1.0003       0.9995       1.0012


    Breslow-Day Test for
Homogeneity of the Odds Ratios
--------------------------------
Chi-Square              0.1521
DF                           1
Pr > ChiSq              0.6966

Total Sample Size = 106340
```

ANALYZING THE BINARY OUTCOME USING THE LOGISTIC PROCEDURE

A SIMPLE LOGISITIC REGRESSION

The following equation can be used to model the relationship between having breast cancer and OC use in the Nurse Health Study. The X variable from the dataset is entered as 1 for OC users and 0 for non-OC users.

$logit\ (p) = \beta_0 + \beta_1 X$

The DESCENDING option in Program 4 is used to model probability of Y equaling 1 since the outcome variable BC (breast cancer) is coded as 1 for breast cancer cases. Without specifying the DESCENDING option, SAS will model the probability of the lowest value of Y.

Similar to PROC FREQ, the WEIGHT statement is used since the data is entered from the cell count data. The MODEL statement is used to analyze the relationship between an outcome variable and one or more independent variables.

Program 4:

```
proc logistic data=nurse_study descending;
    weight count;
    model bc = oc;
run;
```

Output 4.1

```
The LOGISTIC Procedure

             Model Information

Data Set                      WORK.NURSE_STUDY
Response Variable             bc
Number of Response Levels     2
Weight Variable               count
Model                         binary logit
Optimization Technique        Fisher's scoring

Number of Observations Read           8
Number of Observations Used           8
Sum of Weights Read              106340
Sum of Weights Used              106340
```

Output 4.2

```
                Response Profile
 Ordered                       Total            Total
   Value           bc     Frequency           Weight

       1            1             4           570.00
       2            0             4        105770.00

Probability modeled is bc=1.
```

Output 4.3

```
         Model Fit Statistics
                            Intercept
               Intercept            and
Criterion          Only     Covariates

AIC            7099.726       7083.586
SC             7099.806       7083.745
-2 Log L       7097.726       7079.586
```

Output 4.4

```
        Testing Global Null Hypothesis: BETA=0
Test                 Chi-Square       DF     Pr > ChiSq

Likelihood Ratio        18.1401        1         <.0001
Score                   17.8881        1         <.0001
Wald                    17.6834        1         <.0001
```

Output 4.5

```
             Analysis of Maximum Likelihood Estimates
                                Standard          Wald
Parameter    DF    Estimate       Error    Chi-Square    Pr > ChiSq

Intercept     1     -5.0704      0.0532     9095.8096        <.0001
oc            1     -0.3646      0.0867       17.6834        <.0001
```

Output 4.6

```
           Odds Ratio Estimates
             Point          95% Wald
Effect    Estimate      Confidence Limits

oc           0.694       0.586       0.823
```

Output 4.1 displays the dataset information about the dataset that you used for the logistic regression. Output 4.2 shows how the outcome variable BC is coded and the probability modeled is for BC = 1 because the DESCENDING option is used. The three statistics in output 4.3 displays criteria for assessing the model fit; -2 Log L can be used to compare the model fit between a full model to a reduced model.

Output 4.4 displays three types of statistics. The likelihood ratio test is equivalent to the likelihood ratio chi-square reported in output 3.1 from PROC FREQ. The score is identical to Pearson's chi-square test in output 3.1. The Wald test is equivalent to the Wald test in output 4.5, which gives the parameter estimate for β. You can write out the logistic model based on output 4.5, which is

logit (p) = -5.07 – 0.365X

Based on this equation, the odds ratio can be calculated as exp(-0.365) = 0.69, which is the same in output 4.6. Also notice that the odds ratio that calculated from PROC LOGISTIC is identical to the one that calculated from PROC FREQ.

For the Nurse Health Study, data was entered numerically (1 or 0) for both the outcome and explanatory variables. However, in the breathing test study, the outcome variable (TEST) and the explanatory variables (NEVERSMK and OVER40) were entered as characters. When the outcome variable is a character variable, by default, PROC LOGISTIC created ordered values based on the alphabetical order of the outcome variable. For character explanatory variables, a CLASS statement can be used to convert (recode) the characters to ordered numerical values. A common way to recode the character explanatory variables with two levels is to use 1 for the exposed group and 0 for the non-exposed group. For example, for the breathing test study, the model can be written as the following:

$logit\ (p) = \beta_0 + \beta_1 X$

$$X = \begin{cases} 1 & current \\ 0 & never \end{cases}$$

which is equivalent to the following (if you prefer the matrix form):

$$\begin{bmatrix} Logit(p_1) \\ Logit(p_0) \end{bmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_0 \end{pmatrix}$$

This type of parameterization is called *incremental effects* parameterization, which consists of a design matrix with 0s and 1s. β_1 is the increment in log odds for current smokers compared to those that never smoked.

Program 5 illustrates how to analyze the association between two character variables. The DESCENDING option is not used this time because we are modeling probability of being "abnormal" versus "normal" (the default alphabetic order). The probability that is modeled, test='abnormal', is shown in output 5.1.

The CLASS statement is used to list the classification variable. The REF option within the parenthesis is used to indicate which value is the reference. Since we are comparing "current" to "never," "never" is listed. By default, the last ordered value of the classification variable is considered the reference level. So not specifying the REF option will yield the same result. In order to use the incremental effects parameterization, the PARAM = REF option needs to be specified after the "/" in the CLASS statement. Output 5.2 displays how the NEVERSMK variable is coded.

Since the CLASS statement is used in the PROC LOGISTIC, the TYPE 3 Analysis of Effects table (Output 5.5) is also created, which is the Wald test for the effect. Since the test for the NEVERSMK variable has only 1 degree of freedom, this test result will be identical to the one from output 5.6. Again, the odds ratio and its confidence interval calculated in output 5.7 is identical to the one calculated from PROC FREQ in output 1.3.

Program 5:

```
proc logistic data=breathTestAge;
    class neversmk (ref="never")/param=ref;
    weight count;
    model test = neversmk;
run;
```

Output 5.1

The LOGISTIC Procedure

Response Profile

Ordered Value	test	Total Frequency	Total Weight
1	abnormal	4	169.0000
2	normal	4	1668.0000

Probability modeled is test='abnormal'.

Output 5.2

Class Level Information

Class	Value	Design Variables
neversmk	current	1
	never	0

Output 5.3

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	1130.417	1100.035
SC	1130.497	1100.194
-2 Log L	**1128.417**	**1096.035**

Output 5.4

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	32.3820	1	<.0001
Score	30.2421	1	<.0001
Wald	28.2434	1	<.0001

Output 5.5

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
neversmk	1	28.2434	<.0001

Output 5.6

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.9704	0.1663	318.9365	<.0001
neversmk	current	1	1.0136	0.1907	28.2434	<.0001

Output 5.7

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
neversmk current vs never	2.756	1.896	4.004

INVESTIGATING THE INTERACTION EFFECT BY USING PROC LOGISTIC

To test for effect modification of the variable X by variable Z, you need to include variable Z and the interaction term of X and Z (a product of X and Z). For example,

$logit\ (p) = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 X*Z$

One way to test for the interaction is to use the Wald test, which tests the hypothesis of β_3. The result will be similar to the Breslow-Day test of homogeneity of the odds ratio. Program 6 tests the interaction effect between the NEVERSMK and OVER40 in the breathing test study. The Wald test from output 6.2 shows that there is a significant interaction between smoking status and age (p = 0.0001), which is consistent with the Breslow-Day test from PROC FREQ.

Program 6:

```
proc logistic data=breathTestAge;
    class neversmk (ref="never") over40 (ref="no")/param=ref;
    weight count;
    model test = neversmk over40 neversmk*over40;
run;
```

Output 6.1

```
The LOGISTIC Procedure
         Model Fit Statistics
                            Intercept
              Intercept           and
Criterion          Only     Covariates

AIC            1130.417      1055.467
SC             1130.497      1055.785
-2 Log L       1128.417      1047.467
```

Output 6.2

```
                     Analysis of Maximum Likelihood Estimates
                                                Standard          Wald
Parameter                        DF    Estimate       Error    Chi-Square    Pr > ChiSq

Intercept                         1     -2.8315      0.1765      257.4193        <.0001
neversmk         current          1      0.3495      0.2240        2.4355        0.1186
over40           yes              1     -0.8820      0.5359        2.7086        0.0998
neversmk*over40  current yes       1      2.1668      0.5691       14.4985        0.0001
```

Instead of using the Wald test, you can also use the likelihood ratio test (LRT). To use the LRT, you need to compare a full model to a reduced model. The full model contains the interaction term, while the reduced model does not. The null hypothesis is the reduced model fitting the data better compared to the full model.

H_0: $logit\ (p) = \beta_0 + \beta_1 X + \beta_2 Z$ (Reduced model)
H_1: $logit\ (p) = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 X*Z$ (Full model)

Based on the result from the two regressions, you can calculate the likelihood ratio (LR) by using the following equation:

LR = 2 [logL(Full model) – log L(Reduced model)]
= [– 2log L(Reduced model)] – [– 2 logL(Full model)]

LR follows χ^2 distribution with degrees of freedom equaling the number of terms from the full model minus the number of terms from the reduced model. To calculate the likelihood ratio test, the reduced model must be a subset of the full model. If a variable that is in the reduced model cannot be found in the full model, then you cannot use the LRT.

PROC LOGISTIC calculates – 2 times log likelihood statistics under the model fit statistics table. For example, – 2 logL(Full model) can be obtained from output 6,1, which is 1047.467. A complete procedure for calculating the likelihood ratio test is illustrated in Program 7. The result is shown in output 7.1 (LR = 20.66, with 1 df, and p < 0.0001). The result is consistent to the one from the Wald test.

Program 7:

```
*run full model with the interaction term;
proc logistic data=breathTestAge;
    class neversmk (ref="never") over40 (ref="no")/param=ref;
    weight count;
    model test = neversmk over40 neversmk*over40;
    ods output FitStatistics = log2Ratio_full GlobalTests = df_full;

*create a macro variable 'neg2L_full' containing -2logL from the full model;
data _null_;
    set log2Ratio_full;
    if  Criterion = '-2 Log L';
    call symput('neg2L_full', InterceptAndCovariates);

*create a macro variable 'df_full' containing degrees of freedom from the full model;
data _null_;
    set df_full;
    if  Test = 'Likelihood Ratio';
    call symput('df_full', DF);

*run reduced model without the interaction term;
proc logistic data=breathTestAge;
    class neversmk (ref="never") over40 (ref="no")/param=ref;
    weight count;
    model test = neversmk over40;
    ods output FitStatistics = log2Ratio_reduce GlobalTests = df_reduce;

*create a macro variable 'neg2L_reduce' containing -2logL from the reduced model;
data _null_;
    set log2Ratio_reduce;
    if  Criterion = '-2 Log L';
    call symput('neg2L_reduce', InterceptAndCovariates);

*create a macro variable 'df_reduce' containing degrees of freedom from the reduced
model;
data _null_;
    set df_reduce;
    if  Test = 'Likelihood Ratio';
    call symput('df_reduce', DF);
run;

*calculate the LRT;
data result;
    LR = &neg2L_reduce - &neg2L_full;
    df = &df_full - &df_reduce;
    p = 1-probchi(LR,df);
    label LR = 'Likelihood Ratio';

proc print data=result label noobs;
    title "Likelihood ratio test";
run;
```

Output 6.1:

```
Likelihood ratio test

Likelihood
   Ratio      df            p

  20.6558      1    .000005497
```

Since there is an interaction between smoking status and age, you need to report age-specific odds ratio, which can be obtained by running logistic regression within each level of age group (see Program 8). The results are not shown for this program.

Program 8:

```
proc sort data=breathTestAge;
    by over40;
run;

proc logistic data=breathTestAge;
    by over40;
    class neversmk (ref="never")/param=ref;
    weight count;
    model test = neversmk;
run;
```

INVESTIGATING THE CONFOUNDING EFFECT BY USING PROC LOGISTIC

To determine if variable Z is a confounder, you can compare the odds ratio for the main explanatory variable X from the model that includes the variable Z and one from the model that does not include variable Z. For example, to test whether age confounds the association between OC use and having breast cancer, you need to include age within the model (see program 9). The age-adjusted odds ratio is 0.94 (from output 9.1), and the unadjusted odds ratio is 0.69 (output 4.6), which suggests that age is a confounder. This result is consistent with the one that was generated from the FREQ procedure.

Program 9:

```
proc logistic data=nurse_study descending;
    weight count;
    model bc = oc age;
run;
```

Output 9.1

```
           Odds Ratio Estimates
             Point          95% Wald
Effect    Estimate      Confidence Limits

oc           0.942       0.788       1.126
age          2.674       2.141       3.338
```

CONCLUSION

Analyzing variables with dichotomized outcomes by using the FREQ and LOGISTIC procedures is a common task for statisticians in the health care industry. Simply knowing how to use the procedures and how to interpret the results from the procedures is not sufficient. Understanding the goal of model building and following correct model building steps are extremely important in order to obtain accurate and unbiased results.

REFERENCES

Agresti A. (1990), Categorical Data Analysis. Wiley, New York.
Rothman K. and Greenland S (1998), Modern Epidemiology. Lippincott-Raven, Philadelphia, PA
Stokes M, Davis C and Koch G, Categorical Data Analysis Using the SAS System, SAS Institute Inc. Cary, NC: SAS Institute Inc. 2006. SAS OnlineDoc® 9.1.3. Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

Arthur X. Li
City of Hope Comprehensive Cancer Center
Department of Information Science
1500 East Duarte Road
Duarte, CA 91010 - 3000
Work Phone: (626) 256-4673 ext. 65121
Fax: (626) 471-7106
E-mail: xueli@coh.org