A Tutorial on PROC LOGISTIC

Arthur Li, City of Hope National Medical Center, Duarte, CA

ABSTRACT

In the pharmaceutical and health care industries, we often encounter data with dichotomous outcomes, such as having (or not having) a certain disease. This type of data can be analyzed by building a logistic regression model via the LOGISTIC procedure. In this paper, we will address some of the model-building issues that are related to logistic regression. In addition, some statements in PROC LOGISTIC that are new to SAS® 9.2 and ODS statistical graphics relating to logistic regression will also be introduced in this paper.

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

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

Two common ways to compare P_1 and P_0:

1. Relative Risk (RR) or Prevalence Ratio: P_1/P_0
2. Odds Ratio (OR): $[P_1/ 1 – P_1] / [P_0 /1 – P_0] = AD/BC$

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-up, 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 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 cannot be calculated for the case-control study because P_1 and P_0 cannot 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.

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:

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_{1i}$$

where *p* 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:

$$\text{logit}(p_i) = \log \frac{p_i}{1 - p_i} = \log(\text{odds}) \rightarrow p = \frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}}$$

Then the odds ratio can be simplified as the following:

$$OR(X = 1 \ vs \ X = 0) = \frac{\dfrac{e^{\beta_0 + \beta_1} / (1 + e^{\beta_0 + \beta_1})}{1/(1 + e^{\beta_0 + \beta_1})}}{\dfrac{e^{\beta_0} / (1 + e^{\beta_0})}{1/(1 + e^{\beta_0})}} = \frac{e^{\beta_0 + \beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

Based on the equation above, the β_1 from the logistic regression is the log odds comparing an individual with X = 1 to those with X = 0. When β_1 equals 0, the OR will become 1. Thus, testing for no association between the X and Y variables is the same as testing $\beta_1 = 0$. Similar to linear regression, the slope parameter β_1, 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.

SIMPLE LOGISTIC REGRESSION

THE PROSTATE CANCER STUDY

The Prostate Cancer Study (PCS) data is modified from an example in Hosmer and Lemeshow (2000). The goal of PCS is to investigate whether the variables measured at a baseline exam can be used to predict whether a tumor has penetrated the prostatic capsule. Among 380 patients in this data set, 153 had a cancer that penetrated the prostatic capsule. The description of the variables is listed in the following table:

Variable Name	Description	Codes/Value
ID	ID code	1 - 380
CAPSULE	Tumor Penetration of Prostatic Capsule (outcome)	0 - No Penetration 1 - Penetration
AGE	Age	years
ETHNIC	Ethnicity	"white", "black"
DIG_REC_EXAM	Results of the Digital Rectal Exam	"no nodule" "unilobar nodule" "bilobar nodule"
DCAPS	Detection of Capsular Involvement in Rectal Exam	1 = No 2 = Yes
PSA	Prostatic Specific Antigen Value	mg/ml
VOL	Tumor Volume Obtained from Ultrasound	cm^3
GLEASON	Total Gleason Score	0 - 10

LOGISTIC REGRESSION WITH A CONTINUOUS PREDICTOR

Prostatic Specific Antigen Value (PSA) is a known factor to the severity of prostate cancer. Suppose that you are using PSA as a predicting variable:

$$\text{logit}(p) = \beta_0 + \beta_{PSA} X_{PSA}$$
$$H_0 : \beta_{PSA} = 0$$

Program 1 uses the LOGISTIC PROCEDURE to model the probability of a tumor penetrating the prostatic capsule (CAPSULE) by using PSA as the predicting variable.

Program 1:

```
ods graphics on;
proc logistic data=prostate
         plots(only)=(effect oddsratio (type=horizontalstat));
    model capsule (event="1") = psa /clodds=both;
    unit psa = 10;
run;
ods graphics off;
```

Starting from SAS 9.2, PROC LOGISTIC can create statistical graphs automatically via ODS Statistical Graphics. The Output Delivery System (ODS) combines the data (for graphing) that is generated from PROC LOGISTIC with graphical templates and generates statistical graphics to the user-specified destination. To invoke ODS Statistical Graphics, you need to specify the following statement:

```
ODS GRAPHICS ON;
```

To turn off the ODS Statistical Graphics, you will use the following statement:

```
ODS GRAPHICS OFF;
```

The PLOTS= option in the PROC LOGISTIC statement is used to request specific plots. In order to use this option, you must invoke ODS Graphics first. The keyword ONLY is used for displaying only the requested plots. The EFFECT option is used for displaying the effect plots and the ODDSRATIO option is used for displaying odds ratio plots. The TYPE=HORIZONALSTAT option displays the odds ratio figure along the X-axis along with the odds ratio with the confidence limits on the right side of the graphics.

The EVENT= option in the MODEL statement is used to specify the category for which PROC LOGISTIC models the probability. This option is only applied for the binary response model. By default, PROC LOGISTIC uses the first ordered category as the event. In this example, the outcome variable CAPSULE is coded as 1 (event) or 0 (non-event). Thus, **event="1"** is used to model the probability for CAPSULE =1. The CLODDS=BOTH in the MODEL statement is used to specify confidence limits for both WALD and profile likelihood tests.

The UNIT statement enables you to acquire the odds ratio that is based on the specified units in a predictor variable.

Output from Program 1 (part 1):

```
               Model Information

Data Set                      WORK.PROSTATE
Response Variable             CAPSULE
Number of Response Levels     2
Model                         binary logit
Optimization Technique        Fisher's scoring

Number of Observations Read         380
Number of Observations Used         380
```

The “Model Information” table displays the information about the data set, the name of the response variable, the number of response levels, model types, the algorithm for obtaining parameter estimates, and the number of observations being used in the analysis.

Output from Program 1 (part 2):

```
           Response Profile

 Ordered                      Total
   Value      CAPSULE     Frequency

       1            0           227
       2            1           153

Probability modeled is CAPSULE=1.
```

The “Response Profile" table shows the levels and frequency of the response variable. Notice that Probability modeled is **CAPSULE=1** is printed in this table because **event="1"** is used in Program 2.1

Output from Program 1 (part 3):

```
          Model Fit Statistics

                             Intercept
               Intercept           and
Criterion           Only     Covariates

AIC              514.289        467.161
SC               518.229        475.041
-2 Log L         512.289        463.161
```

The statistics that are listed in the “Model Fit Statistics” table is more useful for comparing different models. AIC (Akaike Information Criterion) is a measure of the relative “goodness of fit” of a statistical model. SC (Schwarz criterion) is a criterion for model selection among a finite set of models. Given a set of candidate models for the data, the preferred model is the one with a minimum AIC or SC value. Note that these measures alone do not provide a test of a model in the sense of testing a null hypothesis.

Output from Program 1 (part 4):

```
         Testing Global Null Hypothesis: BETA=0

Test                 Chi-Square       DF     Pr > ChiSq

Likelihood Ratio        49.1277        1         <.0001
Score                   41.7430        1         <.0001
Wald                    29.4230        1         <.0001
```

The three tests in the table above test the null hypothesis that all regression coefficients are 0. A significant p-value indicates that there is at least one explanatory variable in the model that is statistically significant. These three tests are asymptotically similar. For small samples, the likelihood ratio test is the most reliable test.

Output from Program 1 (part 5):

```
              Analysis of Maximum Likelihood Estimates

                                  Standard          Wald
Parameter    DF    Estimate        Error    Chi-Square    Pr > ChiSq

Intercept     1     -1.1137       0.1616       47.5168        <.0001
PSA           1      0.0502      0.00925       29.4230        <.0001
```

The “Analysis of Maximum Likelihood Estimates” table contains the parameter estimates for the logistic regression model. The “Estimate” column contains the rate of change in the logit scale corresponding to a one unit change in the explanatory variable, controlling (adjusting) for the effects of other predicting variables. The Wald Chi-Square statistics and the corresponding p-values are results from the Wald Chi-Square tests, which are used to test whether the parameters are significantly different from 0. Based on the parameter estimates below, you will have the following model:

$$\text{logit}(p) = -1.11 + 0.05X_{PSA}$$

Based on this model, a one-unit increase in PSA corresponds to a 0.0502 increase in the log odds of capsular penetration.

Output from Program 1 (part 6):

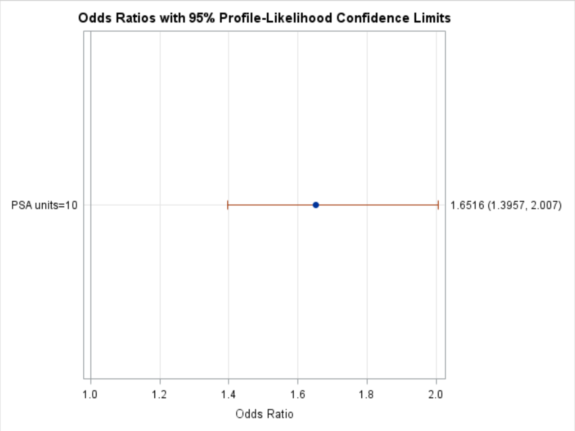
```
Association of Predicted Probabilities and Observed Responses

Percent Concordant     69.9    Somers' D    0.404
Percent Discordant     29.5    Gamma        0.406
Percent Tied            0.7    Tau-a        0.195
Pairs                 34731    c            0.702
```

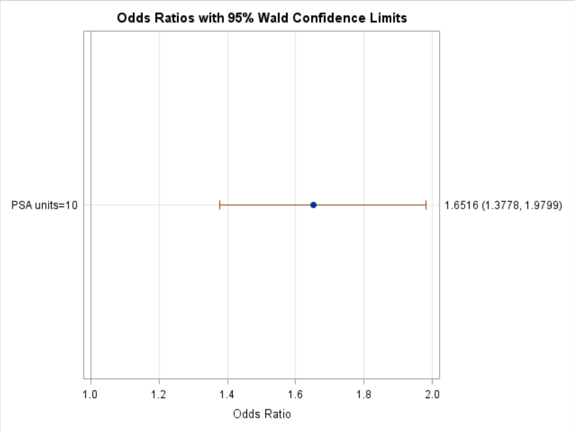
The “Association of Predicted Probabilities and Observed Responses" table lists several measures for predictive accuracy of the model. Pairs = 34731 refers to all possible pairs of observations with different outcomes. In this data set, 227 patients do not have capsular penetration, while 153 do. Thus, 227 X 153 = 34731. A pair of observations with different observed responses is said to be “concordant” if the observation without outcome (CAPSULE =0) has a lower predicted probability than the observation with outcome (CAPSULE =1). On the other hand, a pair of observations with different observed responses is said to be “discordant” if the observation without outcome (CAPSULE =0) has a higher predicted probability than the observation with outcome (CAPSULE =1). A pair with the same predicted probability is said to be “tied”. A preferred model is the one with a higher percentage of concordant pairs and a lower percentage of discordant pairs. This table also contains four rank correlation indexes. These indexes are also used to compare models for prediction purposes. A model with a higher index has better prediction. Among these four statistics, the c statistic is most commonly used. It estimates the probability of an individual with the outcome having a higher predicted probability than an individual without the outcome.

Output from Program 1 (part 7):

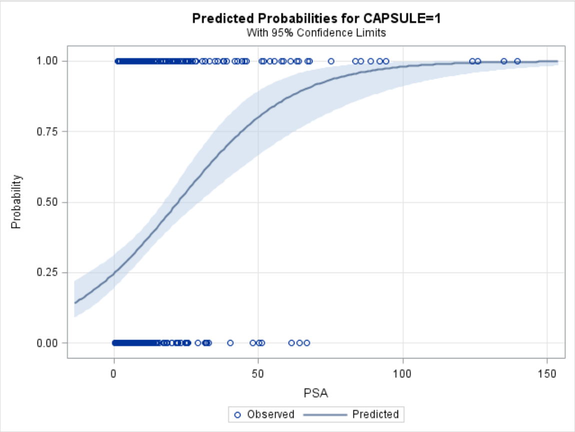
```
Odds Ratio Estimates and Profile-Likelihood Confidence Intervals

Effect         Unit     Estimate     95% Confidence Limits
PSA         10.0000        1.652        1.396        2.007


    Odds Ratio Estimates and Wald Confidence Intervals

Effect         Unit     Estimate     95% Confidence Limits
PSA         10.0000        1.652        1.378        1.980
```

The output above is generated from the CLODDS=BOTH option in the MODEL statement and the UNIT statement. The estimate column contains the odds ratio for every 10-unit increase in PSA ($\exp(10\beta_{\text{PSA}}) = \exp(10 \cdot 0.0502) = 1.652$). The odds of capsular penetration is 1.65 times greater for a man whose PSA measures 10 units larger than a man whose PSA measures 10 units smaller.





The two odds ratio plots are generated: one for the WALD confidence limit and one for the profile likelihood confidence limit. The odds ratio plots reflect the use of the UNIT statement. Without using the UNIT statement, the unit in the plot will be 1.



The effect plot above illustrates the predicted probability of the event via the values of PSA. As you can see, the predicted probability of capsular penetration increases as PSA increases.

LOGISTIC REGRESSION WITH A CATEGORICAL PREDICTOR

Suppose that you are interested in using the results of the digital rectal exam (DIG_REC_EXAM) as the predicting variable. There are three levels in this variable: "no nodule," "unilobar nodule," and "bilobar nodule." In this situation, you need to place the categorical variable in the CLASS statement. PROC LOGISTIC creates design variables for the categorical variable based on the coding scheme that you provided in the CLASS statement. The number of design variables being created is the number of levels in the categorical variable minus 1. Since there are three levels for the DIG_REC_EXAM variable, there will be 2 design variables being created.

The default coding scheme for the CLASS statement is the effect coding. For the effect coding scheme, the last level of all the design variables have a value of -1. The parameter estimates of the design variables estimate the difference between the effect of each level and the average effect over all levels.

Alternatively, you can also use the reference cell coding, which is preferable because it is easier to interpret. When you use reference cell coding, you can specify a baseline level and compare other levels in the categorical variable with the baseline. For example, supposing that you are using "no nodule" as the baseline level, you will have the following model:

$$\mathrm{logit}(p) = \beta_0 + \beta_{uni} X_{uni} + \beta_{bi} X_{bi}$$

$$X_{uni} = \begin{cases} 1 & X_{DIG\_REG\_EXAM} = \text{"}unilobar\text{"} \\ 0 & Otherwise \end{cases}$$

$$X_{bi} = \begin{cases} 1 & X_{DIG\_REG\_EXAM} = \text{"}bilobar\text{"} \\ 0 & Otherwise \end{cases}$$

$$H_0 : \beta_{uni} = \beta_{bi} = 0$$

Program 2 uses the LOGISTIC PROCEDURE to model the probability of a tumor penetrating the prostatic capsule (CAPSULE) by using DIG_REC_EXAM as the predicting variable.

Program 2:

```
ods graphics on;
proc logistic data=prostate plots(only)=(effect (clband)
                                         oddsratio (type=horizontalstat));
    class dig_rec_exam(param=ref ref="no nodule");
    model capsule (event="1") = dig_rec_exam/clodds=pl;
run;
ods graphics off;
```

In Program 2, the CLBAND option is used for adding the confidence limits to the effect plot. Without specifying the CLBAND option, only the predicted value for the categorical variable is plotted. The categorical variable needs to be listed in the CLASS statement. The PARAM= option is used to specify the parameterization method. In this example, the reference cell coding is used (REF). The REF= option is used to specify the baseline level. In the MODEL statement, CLODDS=PL is used to request the profile likelihood confidence interval. Partial outputs generated from Program 2 is listed below.

Output from Program 2 (part 1):

```
            Class Level Information

                                        Design
Class              Value               Variables

dig_rec_exam       bilobar nodule        1      0
                   no nodule             0      0
                   unilobar nodule       0      1
```

The “Class Level Information” table above shows that reference cell coding was used in the model and that the reference level is “no nodule.”

Output from Program 2 (part 2):

```
          Model Fit Statistics

                                 Intercept
                  Intercept            and
Criterion              Only     Covariates

AIC                 514.289        482.860
SC                  518.229        494.681
-2 Log L            512.289        476.860
```

Output from Program 2 (part 3):

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

Likelihood Ratio        35.4286        2         <.0001
Score                   33.8596        2         <.0001
Wald                    30.8133        2         <.0001
```

Output from Program 2 (part 4):

```
           Type 3 Analysis of Effects

                              Wald
Effect            DF    Chi-Square    Pr > ChiSq

dig_rec_exam       2       30.8133        <.0001
```

The “Type 3 Analysis of Effects” table is generated from the CLASS statement. The significant p-value indicates that at least one of the predictors (X_{uni} and X_{bi}) is associated with the outcome.

Output from Program 2 (part 5):

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

Intercept                     1  -1.4376   0.2552    31.7300     <.0001
dig_rec_exam bilobar nodule   1   2.1025   0.3863    29.6179     <.0001
dig_rec_exam unilobar nodule  1   1.1729   0.2881    16.5772     <.0001
```

Based on the parameter estimates, we have the following model:

$$\text{logit}(p) = -1.44 + 1.17X_{uni} + 2.10X_{bi}$$

The odds ratio to compare “bilobar nodule” and “unilobar nodule” with “no nodule” can also be calculated based on the model above:

$$\text{OR}_{\text{uni vs no}} = \exp(1.1729) = 3.23$$

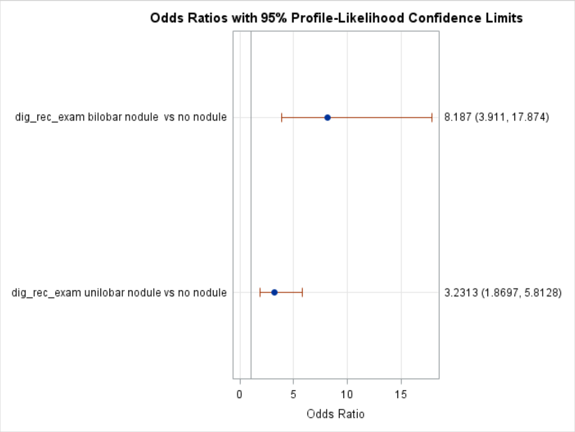
$$\text{OR}_{\text{bi vs no}} = \exp(2.1025) = 8.19$$

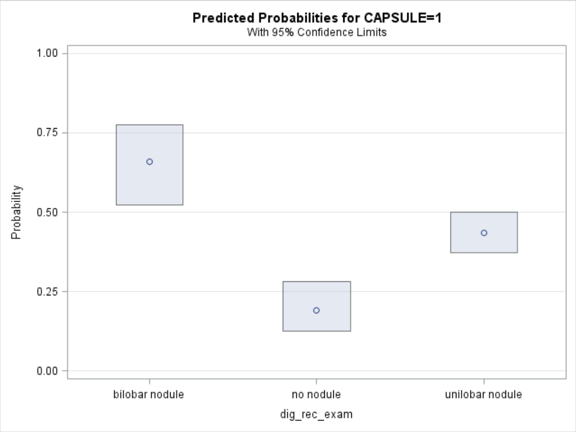
Output from Program 2 (part 6):

```
Odds Ratio Estimates and Profile-Likelihood Confidence Intervals
Effect                                          Unit     Estimate

dig_rec_exam bilobar nodule  vs no nodule       1.0000       8.187
dig_rec_exam unilobar nodule vs no nodule       1.0000       3.231

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals
95% Confidence Limits

   3.911       17.874
   1.870        5.813
```

The “Odds Ratio Estimates and Profile-Likelihood Confidence Intervals” table is generated from the CLODDS=PL option in the model statement. The odds of capsular penetration is 3.2 times higher in patients with unilobar nodule compared to patients with no nodule (95% CI: 1.87 - 5.81, p< 0.0001). The odds of capsular penetration is 8.2 times higher in patients with bilobar nodule compared to patients with no nodule (95% CI: 3.91 - 17.87, p< 0.0001).

Here are the odds ratio plot and the effect plot:





LOGISTIC REGRESSION WITH AN ORDERED CATEGORICAL PREDICTOR

When DIG_REC_EXAM is used as a categorical predictor, the effect plot shows that patients with bilobar nodule have the highest probability of having capsular penetration while patients with no nodule have the lowest probability. These suggest that DIG_REC_EXAM can also be treated as a linear variable in the model. This type of model is also called a grouped linear model:

$$\operatorname{logit}(p) = \beta_0 + \beta_g X_g$$

$$X_g = \begin{cases} 0 & X_{DIG\_REG\_EXAM} = \text{"}no\ nodule\text{"} \\ 1 & X_{DIG\_REG\_EXAM} = \text{"}unilobar\ nodule\text{"} \\ 2 & X_{DIG\_REG\_EXAM} = \text{"}bilobar\ nodule\text{"} \end{cases}$$

$$H_0 : \beta_g = 0$$

The DATA step in Program 3 creates a new variable (DIG_REC_EXAM_G) with linear coding. The variable DIG_REC_EXAM_G is then used as the predictor in PROC LOGISTIC.

Program 3:

```
data prostate1;
    set prostate;
    dig_rec_exam_g = (dig_rec_exam = "unilobar nodule") +
                   2*(dig_rec_exam = "bilobar nodule");
run;

proc logistic data=prostate1;
    model capsule (event="1") = dig_rec_exam_g/clodds=pl;
run;
```

Output from Program 3:

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

AIC              514.289        481.128
SC               518.229        489.009
-2 Log L         512.289        477.128

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

Likelihood Ratio        35.1606        1         <.0001
Score                   33.8363        1         <.0001
Wald                    30.8464        1         <.0001


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

Intercept          1     -1.3674    0.2117       41.7347        <.0001
dig_rec_exam_g     1      1.0632    0.1914       30.8464        <.0001


 Odds Ratio Estimates and Profile-Likelihood Confidence Intervals

Effect                 Unit     Estimate     95% Confidence Limits

dig_rec_exam_g       1.0000        2.896        2.010        4.267
```

Based on the output above, the result from the digital rectal exam by using linear coding is also associated with capsular penetration. The odds of having capsular penetration increases with an increasing number of nodules [OR = 2.9, 95%CI = (2.01 - 4.27), p <0.0001)]

USING THE LIKELIHOOD RATIO TEST TO COMPARE MODELS

In the previous section, variable DIG_REC_EXAM was coded differently in two different models and both showed that DIG_REC_EXAM significantly predicted outcome. To choose a model that is sufficient to describe the association between the predicting variable and the outcome, you can perform the likelihood ratio test (LRT). One of the conditions in performing LRT is that models need to be nested. It can be shown that the model with a reference cell coding scheme is a reparameterization of a model that adds terms to the grouped linear model:

Note that $X_g = X_{uni} + 2X_{bi} \quad \rightarrow \quad \frac{X_g}{2} = \frac{X_{uni}}{2} + X_{bi}$

Prove:

$$\text{logit}(\text{p}) = \beta_0 + \beta_{uni}X_{uni} + \beta_{bi}X_{bi} = \beta_0 + \left(\beta_{uni} - \frac{\beta_{bi}}{2}\right)X_{uni} + \beta_{bi}\left(\frac{X_{uni}}{2} + X_{bi}\right) = \beta_0 + \left(\beta_{uni} - \frac{\beta_{bi}}{2}\right)X_{uni} + \beta_{bi}\left(\frac{X_g}{2}\right)$$

$$= \beta_0 + \beta_1^{'} X_{uni} + \beta_2^{'} X_g = \text{grouped linear model} + \text{extra term}$$

The hypothesis for comparing these two models follows:

$$H_0 : \text{logit}(p) = \beta_0 + \beta_g X_g \,(\text{reduced model})$$

$$H_1 : \text{logit}(p) = \beta_0 + \beta_{uni} X_{uni} + \beta_{bi} X_{bi} \,(\text{full model})$$

The LR statistic can be calculated from the following formula:

LR = 2[logL(Full model) - logL(Reduced model)]
LR~X^2; with df = (df for the full model) - (df for the reduced model)

The LR statistic above can be obtained by taking the Likelihood Ratio chi-square statistic from the full model minus the Likelihood Ratio chi-square statistic from the reduced model. This test is demonstrated in Program 3.4.

Program 4:

```
proc logistic data=prostate1; * run full model;
    class dig_rec_exam(param=ref ref="no nodule");
    model capsule (event="1") = dig_rec_exam;
    ods output GlobalTests = GlobalTests_full;
run;

data _null_;
    set GlobalTests_full;
    if  Test = 'Likelihood Ratio' then do;
        call symput('ChiSq_full', ChiSq);
        call symput('df_full', DF);
    end;
run;

proc logistic data=prostate1;  * run reduced model;
    model capsule (event="1") = dig_rec_exam_g;
    ods output GlobalTests = GlobalTests_reduce;
run;

data _null_;
    set GlobalTests_reduce;
    if  Test = 'Likelihood Ratio' then do;
        call symput('ChiSq_reduce', ChiSq);
        call symput('df_reduce', DF);
    end;
run;

data result;    *LRT test;
    LR = &ChiSq_full - &ChiSq_reduce;
    df = &df_full - &df_reduce;
    p = 1-probchi(LR,df);
    label LR = 'Likelihood Ratio';
run;

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

Output from Program 4:

```
Likelihood ratio test


Likelihood
   Ratio      df       p


  0.26800      1    0.60468
```

Based on the result from LRT, we failed to reject the null hypothesis (LR = 0.27, df = 1, p = 0.6). A grouped linear variable is sufficient to describe the association between the digital rectal exam and capsular penetration.

MULTIPLE LOGISTIC REGRESSION

Suppose that you would like to study weather ethnicity modifies the association between PSA and capsular penetration. You would also like to adjust the result from the digital rectal exam (using linear coding) and gleason score in the model. This model is illustrated in Program 5.

$$\text{logit}(p) = \beta_0 + \beta_{PSA} X_{PSA} + \beta_{white} X_{white} + \beta_{\text{int}} X_{PSA} X_{white} + \beta_{DIG\_REC\_EXAM\_G} X_{DIG\_REC\_EXAM\_G} + \beta_{GLEASON} X_{GLEASON}$$

$$X_{white} = \begin{cases} 1 & X_{ETHNIC} = "white" \\ 0 & X_{ETHNIC} = "black" \end{cases}$$

Program 5:

```
ods graphics on;
proc logistic data=prostate1
    plots(only)=(effect(x=(psa) sliceby=ethnic) oddsratio (type=horizontalstat));
    class ethnic(param=ref ref="black");
    model capsule (event="1") = psa ethnic psa*ethnic
        dig_rec_exam_g gleason/clodds=pl;
    unit psa = 10/default=1;
    oddsratio 'psa 50 vs 40 for black' psa/at(ethnic="black" psa=40) cl=pl;
    oddsratio 'psa 50 vs 40 for white' psa/at(ethnic="white" psa=40) cl=pl;
run;
ods graphics off;
```

In program 5, the interaction effect between PSA and ethnicity is requested in the effect plot. The X= option is used to specify effects to be used on the X axis. When creating an effect plot for interaction, the continuous variable must be specified as main effects. The SLICEBY= option is used to illustrate predicted probabilities at each unique level of the variable that is listed in the SLICEBY= option. Other continuous variables in the MODEL statement that are not specified in the X= option will be fixed at their means.

When you include an interaction term in the model, the CLODDS= option in the MODEL statement does not compute an odds ratio for the variables that are involved with the interaction. In this situation, you must use the ODDSRATIO statement. The ODDSRATIO statement (new to 9.2) is used to produce the odds ratio for the listed variable. The AT (covariate=value-list) is used to specify fixed levels of the interacting covariates. Since the UNIT statement specifies the unit of PSA is 10, the OR is then computed for PSA = 50 vs PSA = 40 for "black" and "white" separately. The DEFAULT option in the UNIT statement provides the unit changes for all the predicting variables that are not listed in the UNIT statement. Without specifying this option, PROC LOGISTIC only generates the odds ratio for the predictors that are listed in the UNIT statement.

Partial Output from Program 5 (part 1):

```
             Type 3 Analysis of Effects
                                 Wald
Effect                 DF     Chi-Square     Pr > ChiSq

PSA                     1         0.0450         0.8319
ethnic                  1         0.0001         0.9943
PSA*ethnic              1         2.3100         0.1285
dig_rec_exam_g          1        13.1187         0.0003
GLEASON                 1        37.3187         <.0001
```

Partial Output from Program 5 (part 2):

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

Intercept               1     -7.9819      1.1856      45.3214       <.0001
PSA                     1     0.00358      0.0168       0.0450       0.8319
ethnic         white    1    -0.00438      0.6098       0.0001       0.9943
PSA*ethnic     white    1      0.0304      0.0200       2.3100       0.1285
dig_rec_exam_g          1      0.7990      0.2206      13.1187       0.0003
GLEASON                 1      0.9974      0.1633      37.3187       <.0001
```

Based on the output above, there is an interaction between PSA and ethnicity at a 15% level.

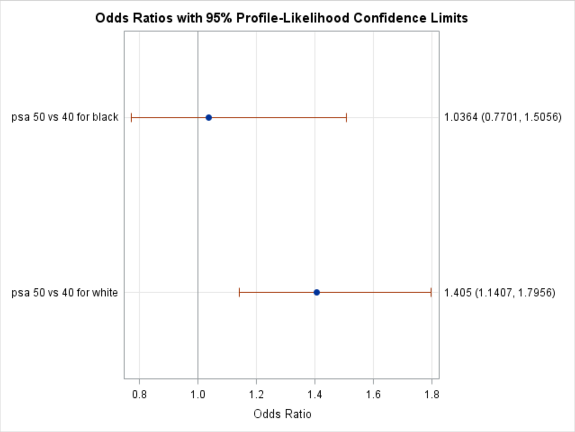
Partial Output from Program 5 (part 3):

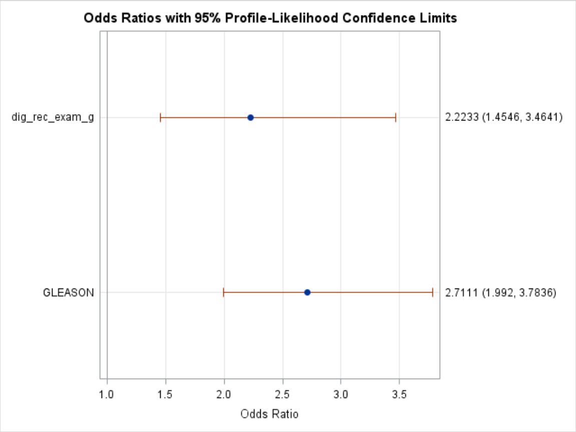
```
     Odds Ratio Estimates and Profile-Likelihood Confidence Intervals
                               Label                               Estimate

psa 50 vs 40 for black        PSA units=10 at ethnic=black           1.036
psa 50 vs 40 for white        PSA units=10 at ethnic=white           1.405

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals
95% Confidence Limits

   0.770       1.506
   1.141       1.796

 Odds Ratio Estimates and Profile-Likelihood Confidence Intervals
Effect                 Unit     Estimate     95% Confidence Limits

dig_rec_exam_g       1.0000        2.223        1.455        3.464
GLEASON              1.0000        2.711        1.992        3.784
```

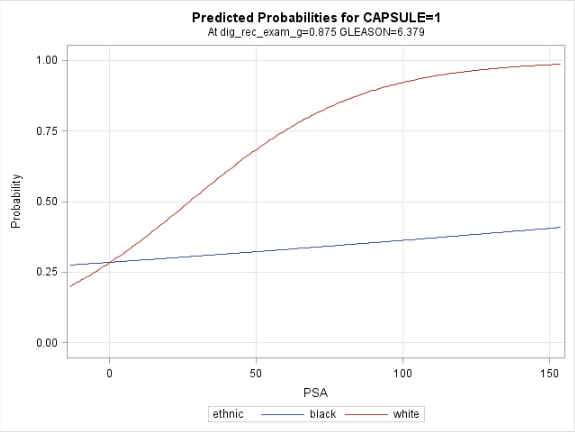
Based on the output above, the OR is calculated for PSA = 50 vs PSA = 40 for “black” and “white” separately. Because we’re using the DEFAULT option in the UNIT statement, the OR for both the DIG_REC_EXAM_G and GLEASON variables are calculated for a 1-unit increase.

The association between PSA and capsular penetration is not significant among blacks, but it is significant among white patients. Black patients who are 10 units higher in PSA are about 1.04 times as likely to have capsular penetration compared to black patients who are 10 units lower in PSA (95%CI: 0.77 - 1.51). On the other hand, white patients who are 10 units higher in PSA are about 1.41 times as likely to have capsular penetration compared to white patients who are 10 units lower in PSA (95%CI: 1.14 – 1.80). This magnitude of the OR is also demonstrated in the odds ratio plots below.

The effect plot below shows that the probability of capsular penetration increases with PSA; however, the increase is more pronounced in the white group.







GOODNESS-OF-FIT TEST

To examine how well your developed model fits the data, you can perform the goodness-of-fit test. The goodness-of-fit tests are used to examine how closely a model's fitted responses approximate observed responses. The hypothesis of the goodness-of-fit tests follows:

H_0 : The model fits the data
H_1 : The model does not fit the data

The way to access the performance of the model is to examine the fit of the model under different covariate patterns, such as using Pearson and Deviance statistics, which are summary statistics that are based on the differences in observed and fitted values. A covariate pattern is a set of values for the covariates in the model. For example, for two categorical variables with each having two levels, you will have four covariate patterns. However, when you include a continuous variable in the model, such as AGE, there could be as many covariate patterns as the number of observations in the data set.

The requirement for utilizing Pearson and Deviance statistics is that there must be at least 10 observations within each covariate pattern (on average) and at least 80% of the expected count within each covariate pattern is ≥ 5 and the remaining expected counts are > 2. These requirements make it impossible to achieve for a continuous variable.

When the sample size requirements for Deviance and Pearson chi-square statistics cannot be met for continuous variables, you can use the Hosmer-Lemeshow goodness-of-fit test. To perform Hosmer-Lemeshow goodness-of-fit test, all the subjects are divided into approximately 10 groups of roughly the same size based on the percentiles of the estimated probabilities. Then A Pearson chi-square statistic is then computed based on the observed and expected counts in each group.

$$\chi^2_{HL} = \sum_{i=1}^{10} \frac{\left(O_i - N_i \overline{p}_i\right)^2}{N_i \overline{p}_i \left(1 - \overline{p}_i\right)}, \quad df = 8$$

N_i = the total number of observatoins in the i^{th} group

O_i = the number of positive responses in observatoins in the i^{th} group

$\overline{p}_i$ = the average estimated probablity for subject in the i^{th} group based on the model

Program 6 generates goodness-of-fit statistics. The AGGREGATE option is used to treat each unique combination of the predictor values as a distinct group for calculating the Pearson chi-square test statistic and the Deviance. The Deviance and Pearson goodness-of-fit statistics are calculated only when SCALE=NONE is specified. The LACKFIT option is used to perform the Hosmer and Lemeshow goodness-of-fit test.

Program 6:

```
proc logistic data=prostate1;
    class ethnic(param=ref ref="black");
    model capsule (event="1") = psa ethnic psa*ethnic
          dig_rec_exam_g gleason/aggregate scale=none lackfit;
run;
```

Partial Output from Program 6 (part 1):

```
       Deviance and Pearson Goodness-of-Fit Statistics

Criterion          Value       DF     Value/DF     Pr > ChiSq

Deviance        329.7792      327       1.0085         0.4465
Pearson         307.7699      327       0.9412         0.7707

Number of unique profiles: 333
```

The p-values for both the Deviance and Pearson statistics are not significant, which indicates the model ft is adequate. However, the number of unique profiles is 333, which is too large for a sample size of 377. Thus, the results for these tests are not valid.

Partial Output from Program 6 (part 2):

```
              Partition for the Hosmer and Lemeshow Test

                              CAPSULE = 1             CAPSULE = 0
    Group        Total    Observed    Expected    Observed    Expected

        1           38           1        2.66          37       35.34
        2           38           3        4.73          35       33.27
        3           38           5        6.83          33       31.17
        4           38           9        9.59          29       28.41
        5           38          17       10.98          21       27.02
        6           38          15       14.52          23       23.48
        7           38          20       18.75          18       19.25
        8           38          24       22.28          14       15.72
        9           38          25       28.41          13        9.59
       10           35          32       32.24           3        2.76


 Hosmer and Lemeshow Goodness-of-Fit Test

 Chi-Square       DF     Pr > ChiSq

     9.2760        8         0.3196
```

The partition above does not show any violation for the requirements of the Hosmer and Lemeshow test (at least 80% expected count within each group > 5 and the remainder expected counts > 2). The insignificant p-value indicates the model fit is adequate.

ANALYSIS OF RESIDUALS AND INFLUENTIAL STATISTICS

PROC LOGISTIC generates different types of residuals and influential statistics that allow you to detect outliers and/or influential points of your model.

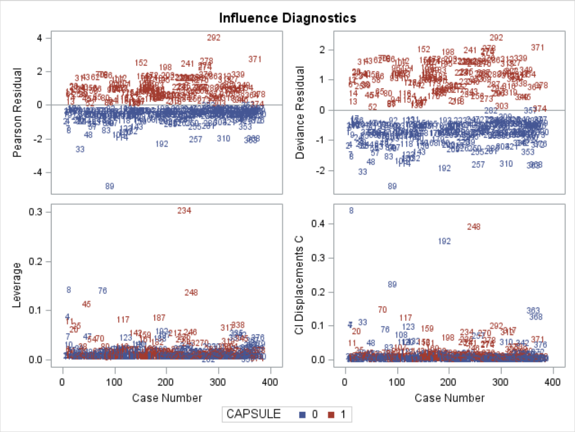
- **RESCHI**: Pearson (chi-square) residual, which is used for identifying observations that are poorly accounted for by the model
- **RESDEV**: deviance residual, which is used for identifying poorly-fitted observations
- **DIFCHISQ**: change in Pearson chi-square with the deletion of each observation
- **DIFDEV**: change in deviance with the deletion of each observation
- **DFBETAS**: these statistics provide influence on parameter estimates, more specifically, how much changes in each parameter with the deletion of each observation
- **C** and **CBAR**: confidence interval displacement diagnostic that measures how much the regression estimates (intercept, slopes) change with the deletion of each observation
- **H**: measures extremity of the observation in the design space of the explanatory variables

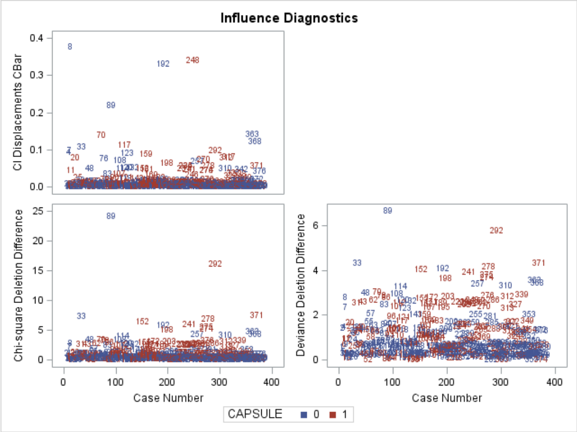
The residuals and influential statistics can be visualized from the diagnostic plots. Program 7 generates different types of diagnostic plots via ODS Statistical Graphics. The LABEL option is used to display the observation number on the diagnostic plots. DFBETAS generates plots with different DFBETAS in the y-axis and the observation number in the x-axis. INFLUENCE generates plots with RESCHI, RESDEV, H, confidence intervals for C and CBAR, DIFCHISQ, and DIFDEV in the y-axis, as well as the observation numbers in the x-axis. LEVERAGE generates plots with DIFCHISQ, DIFDEV, confidence interval for C, and the predictive probability in the y-axis and H in the x-axis. PHAT generates plots with DIFCHISQ, DIFDEV, confidence interval for C, and H (leverage) in the y-axis and the predictive probability in the x-axis. DPC generates plots with DIFCHISQ and DIFDEV in the y-axis and the predictive probability in the x-axis with colored markers according to the value of the confidence interval of displacement C.

Program 7:

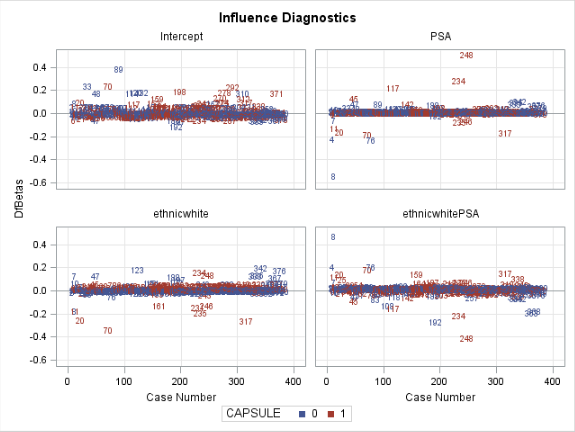
```
ods graphics on;
proc logistic data=prostate1 plots(only label)=
                              (dfbetas influence leverage phat dpc);
    class ethnic(param=ref ref="black");
    model capsule (event="1") = psa ethnic psa*ethnic dig_rec_exam_g gleason;
run;
ods graphics off;
```

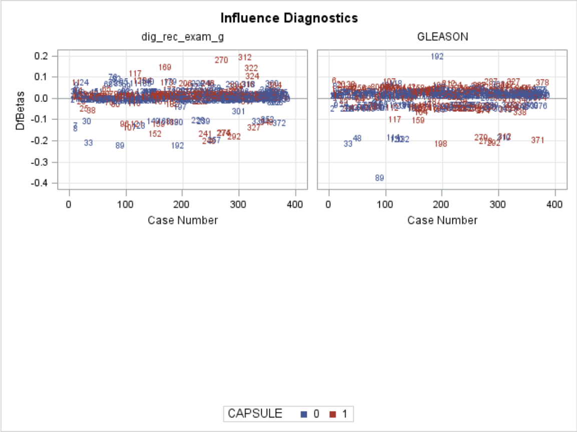
The two figures above are generated from the INFLUENCE option. The observations that are furthest away from 0 are the influential observations, such as observations 89, 292, 8, 234, etc.



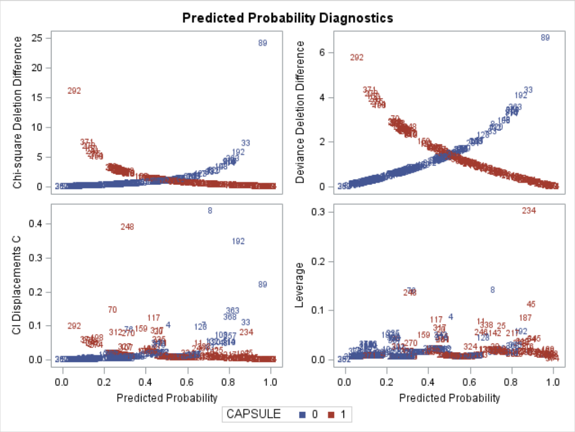


The two figures above are generated from the DFBETAS option.

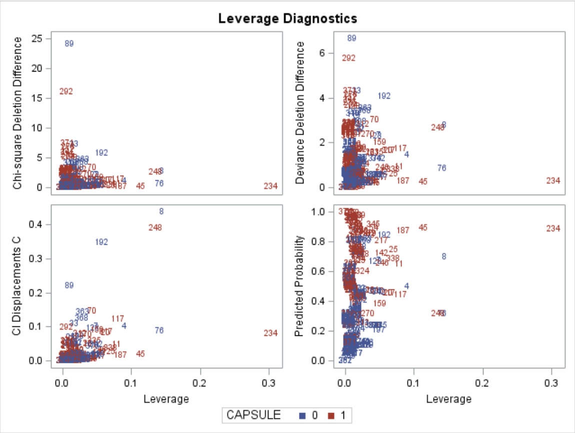




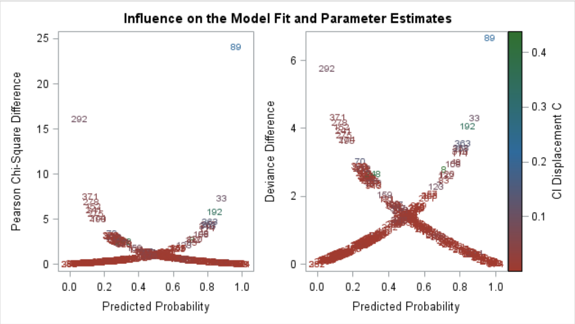
The next figure is generated from the PHAT option. The observation in the upper left corner (observation 292) had capsular penetration but with low predicted probability. The observation in the upper right corner (observation 89) did not have capsular penetration but has high predicted probability.



The next figure is generated from the LEVERAGE option.



The last figure is generated from the DPC option. The observations in the bottom “cup" in red need to be scrutinized more closely. These observations influence the parameter estimates to a relatively large extent but are not poorly fitted.



CONCLUSION

Analyzing variables with dichotomized outcomes is a common task for statisticians in the health care industry. With newly-added features in SAS/STAT 9.2, such as ODS GRAPHICS, PLOT= option, ODDSRATIO statement, and newly-implemented tests, PROC LOGISTIC becomes an even more powerful procedure which not only can be used to build a logistic model, but can also generate high quality figures.

REFERENCES

Agresti, A (1996), An Introduction to Categorical Data Analysis, New York: John Wiley & Sons.

Allison, P. (1999), Logistic Regression Using the SAS System: Theory and Application, Cary, N.C.: SAS Institute Inc.

Cook, R. D. and Weisberg, S. (1982), Residuals and Influence in Regression, New York: Chapman & Hall.

Fleiss, J. L. (1981), Statistical Methods for Rates and Proportions, Second Edition, New York: John Wiley & Sons.

Hosmer, D.W. and Lemeshow, S. (2000), Applied Logistic Regression, New York: John Wiley & Sons.

Kleinbaum, D.G., Kupper, L.L. and Muller, K.E. (1998), Applied Regression Analysis and Other Multivariable Methods, Boston : PWS-KENT Publishing Company

McCullagh, P. and Nelder, J. A. (1989), Generalized Linear Models, Second Edition, London: Chapman & Hall.

Rothman, K.J. (1986) Modern Epidemiology, Boston: Little, Brown and Company

SAS Institute Inc. (1995), Logistic Regression Examples Using the SAS System, Cary, NC: SAS Institute Inc.

Stokes, M. E., Davis, C. S., and Koch, G. G. (2000), Categorical Data Analysis Using the SAS System, Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

Arthur Li
City of Hope National Medical 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: arthurli@coh.org

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.