

# **Calculating the Odds Ratio and Goodness of Fit Statistics for Logistic Regression Models using PROC LOGISTIC**

**Chuan-Chuan C. Wun, Ph.D.  
Houston Center for Quality of Care and Utilization Studies,  
VA HSR&D Field Program, VA Medical Center, Houston, TX**

## **Abstract**

The logistic regression model is used to assess the relationship between a categorical outcome variable and a set of either categorical or continuous predictors. The degree of association between the outcome variable and the predictors is commonly measured by the odds ratio. Several goodness-of-fit statistics have been proposed to test how well the model predicts the observed data.

This paper presents SAS programs that calculate the odds ratio, the confidence interval for the odds ratio, and some commonly used goodness-of-fit statistics: Pearson chi-square, likelihood ratio chi-square, Hosmer-Lemeshow test statistic and  $R^2$ . These programs may be used to supplement the output generated by PROC LOGISTIC in SAS 6.06.

## **1. Introduction**

Regression methods have become an integral component of any data analysis concerned with describing the relationship between a response variable and one or more explanatory variables. It is often the case that the outcome variable is discrete, taking on two or more possible values. In the early 1960s, the logistic model was proposed<sup>1,2</sup> and has become the standard method of analysis in this situation.

What distinguishes a logistic regression model from the linear regression model is that the outcome variable in logistic regression is binary or dichotomous. In any regression problem the key quantity is the mean value of the outcome variable, given the value of independent variable. This quantity is called the conditional mean and can be expressed as " $E(Y|x)$ " where  $Y$  denotes the outcome variable and  $x$  denotes a value of the independent variable. In linear regression, this conditional mean may be expressed as an equation linear in  $x$ , such as:

$$E(Y|x) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k \quad \dots\dots\dots (1)$$

This expression implies that it is possible for  $E(Y|x)$  to take on any value as  $x$  ranges from  $-\infty$  to  $+\infty$ . With dichotomous outcome data the conditional mean of the regression equation must be formulated to be bound between zero and one.

To simplify the notation, the quantity  $\pi(x) = E(Y|x)$  is used to represent the conditional mean of  $Y$  given  $x$  in logistic distribution. The specific form of logistic regression model is:

$$\pi(x) = \frac{e^{\beta_0 + \sum_{i=1}^k \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^k \beta_i x_i}} \dots\dots\dots (2)$$

The logit transformation of  $\pi(x)$  is central to the study of logistic regression. This transformation is defined as follows:

$$\begin{aligned} g(x) &= \ln \left[ \frac{\pi(x)}{1 - \pi(x)} \right] \\ &= \beta_0 + \sum_{i=1}^k \beta_i x_i \end{aligned} \dots\dots\dots (3)$$

The importance of this transformation is that  $g(x)$  has many of the desirable properties of a linear regression model. The logit,  $g(x)$  is linear in its parameters, may be continuous, and may range from  $-\infty$  to  $+\infty$ .

## 2. Testing for the Significance of the Coefficients

The maximum likelihood approach is used to estimate the parameters in the logistic regression model. After estimating the coefficients, the first look at the fitted model commonly concerns whether the independent variables are "significantly" related to the outcome variable.

Several testing methods were proposed. In SAS 6.06, PROC LOGISTIC uses the Wald chi-squared statistic to test the significance of the  $\beta$  coefficients in the logistic model. The Wald chi-squared statistic is computed as the square of the parameter estimate divided by its standard error estimate, and will assume a chi-square distribution with one degree of freedom.

### 3. Odds Ratio: A Measure of Association

After assessing the significance of the coefficients, usually we also want to quantify the degree of association between the outcome and the independent variables. The odds ratio is a measure of association which has found wide use, especially in epidemiology.

The odds ratio can be easily interpreted by a logistic regression model with a dichotomous independent variable  $x$ . Assuming that  $x$  is coded as either zero or one. The odds of the outcome being present among individuals with  $x=1$  is defined as  $\pi(1)/[1-\pi(1)]$ . Similarly the odds of the outcome being present among individuals with  $x=0$  is defined as  $\pi(0)/[1-\pi(0)]$ . Using the expression for the logistic regression model shown in equation (2), the odds ratio is:

$$\begin{aligned}\psi &= \frac{\left( \frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}} \right) / \left( \frac{1}{1 + e^{\beta_0 + \beta_1}} \right)}{\left( \frac{e^{\beta_0}}{1 + e^{\beta_0}} \right) / \left( \frac{1}{1 + e^{\beta_0}} \right)} \dots\dots\dots (4) \\ &= \frac{e^{\beta_0 + \beta_1}}{e^{\beta_0}} = e^{\beta_1}\end{aligned}$$

And the log odds is:

$$\ln(\psi) = \ln(e^{\beta_1}) = \beta_1 \dots\dots\dots (5)$$

Thus the odds ratio approximates how much more likely (or unlikely) it is for the outcome to be present among those with  $x=1$  than among those with  $x=0$ . This fact concerning the interpretability of the coefficients is the fundamental reason why logistic regression has proven such a powerful analytic tool for epidemiologic research.

Along with the point estimation of odds ratio, a confidence interval estimate may also provide additional information. In general, the 95% confidence interval of the odds ratio is given by the following expression:

$$\exp [\beta_1 \pm 1.96 * SE(\beta_1)] \dots \dots (6)$$

#### 4. Example

To illustrate how to compute the odds ratio, 95% confidence interval of the odds ratio and later on goodness of fit statistics in order to supplement the output generated by PROC LOGISTIC in SAS 6.06, sample data collected at Baystate medical center in Springfield, Massachusetts, during 1986<sup>3</sup> will be used in this paper.

This data was designed to study the association between low infant birth weight and a list of risk factors. The dependent variable: LOW is dichotomous (0= birth weight  $\geq 2500$  grams, 1= birth weight  $< 2500$  grams). The risk factors of interest include: AGE (age of mother in years), LWD (Weight at the last menstrual period: 1=  $< 110$  pounds, 0=  $\geq 110$  pounds), RACE (1=white, 2=black, 3=other), SMOKE (smoking status during pregnancy: 1=yes, 0=no), PTD (history of premature labor: 1=yes, 0=no), HT (history of hypertension: 1=yes, 0=no), UI (presence of uterine irritability: 1=yes, 0=no), FTV (number of physician visits during the first trimester).

Based on the estimated covariance matrix produced by PROC LOGISTIC, the following program computes the odds ratio and the corresponding 95% confidence interval for the sample data. The results are presented in Table 1.

---

## PROGRAM: COMPUTE ODDS RATIO AND 95% CI OF ODDS RATIO

---

```
LIBNAME LL 'C:\GF';
PROC SORT DATA=LL.LOWBW;
  BY DESCENDING LOW;
PROC LOGISTIC DATA=LL.LOWBW ORDER=DATA OUTEST=BETAS COVOUT;
  MODEL LOW=LWT RACE1 RACE2 SMOKE PTL HT UI;
PROC PRINT DATA=BETAS;
TITLE 'THE COVARIANCE MATRIX OF THE BETA COEFFICIENTS';
```

```
DATA ODDS ;
  ARRAY B{8} INTERCEP LWT RACE1 RACE2 SMOKE PTL HT UI;
  ARRAY NEWB{8} B1-B8;
  ARRAY VAR{8} V1-V8;
  DO I=1 TO 9;
    SET BETAS (KEEP=INTERCEP-UI) END=EOF;
    IF I=1 THEN DO;
      DO J=1 TO 8;
        NEWB{J}=B{J};
      END;
    END;
    IF I>1 THEN DO;
      DO J=1 TO 8 ;
        IF J=I-1 THEN DO ;
          VAR{J}=B{J};
        END;
      END;
    END;
  END;
  IF EOF THEN return;
DROP INTERCEP LWT RACE1 RACE2 SMOKE PTL HT UI I J ;
```

```
DATA ODDS2;
  SET ODDS;
  ARRAY BETA{8} B1-B8;
  ARRAY V{8} V1-V8;
DO I=1 TO 8;
  COEFF=BETA{I};
  VAR=V{I};
  SE=V{I}**0.5;
  OUTPUT;
END;
```

```
DATA ODDS3;
  SET ODDS2;
  OR=EXP(COEFF);
  LCIOR=EXP(COEFF-(1.96*SE));
  UCIOR=EXP(COEFF+(1.96*SE));
  IF _N_=1 THEN VARIABLE='INTERCEP';
  IF _N_=2 THEN VARIABLE='LWT';
  IF _N_=3 THEN VARIABLE='RACE1';
  IF _N_=4 THEN VARIABLE='RACE2';
  IF _N_=5 THEN VARIABLE='SMOKE';
  IF _N_=6 THEN VARIABLE='PTL';
  IF _N_=7 THEN VARIABLE='HT';
  IF _N_=8 THEN VARIABLE='UI';
  LABEL COEFF='Beta Coefficient'
        SE='Standard Error: Betas'
        OR='Estimated Odds Ratio'
        LCIOR='Lower 95% CI: Odds Ratio'
        UCIOR='Upper 95% CI: Odds Ratio';
  KEEP VARIABLE COEFF SE OR LCIOR UCIOR;
PROC PRINT Label;
  VAR VARIABLE COEFF SE OR LCIOR UCIOR;
TITLE 'ODDS RATIO AND 95% CI FOR ODDS RATIO';
```

**Table 1**

**Output: Odds Ratio and 95% CI for the Odds Ratio**

**The LOGISTIC Procedure**

**Response Profile**

| Ordered<br>Value | LOW | Count |
|------------------|-----|-------|
| 1                | 1   | 59    |
| 2                | 0   | 130   |

**Simple Statistics for Explanatory Variables**

| Variable | Mean       | Standard<br>Deviation | Minimum | Maximum |
|----------|------------|-----------------------|---------|---------|
| LWT      | 129.814815 | 30.579380             | 80.0000 | 250.000 |
| RACE1    | 0.137566   | 0.345359              | 0.0000  | 1.000   |
| RACE2    | 0.354497   | 0.479631              | 0.0000  | 1.000   |
| SMOKE    | 0.391534   | 0.489390              | 0.0000  | 1.000   |
| PTL      | 0.195767   | 0.493342              | 0.0000  | 3.000   |
| HT       | 0.063492   | 0.244494              | 0.0000  | 1.000   |
| UI       | 0.148148   | 0.356190              | 0.0000  | 1.000   |

**Criteria for Assessing Model Fit**

| Criterion | Intercept<br>Only | Intercept<br>and<br>Covariates | Chi-Square for Covariates   |
|-----------|-------------------|--------------------------------|-----------------------------|
| AIC       | 236.672           | 217.986                        | .                           |
| SC        | 239.914           | 243.920                        | .                           |
| -2 LOG L  | 234.672           | 201.986                        | 32.686 with 7 DF (p=0.0001) |
| Score     | .                 | .                              | 30.658 with 7 DF (p=0.0001) |

**Analysis of Maximum Likelihood Estimates**

| Variable | Parameter<br>Estimate | Standard<br>Error | Wald<br>Chi-Square | Pr ><br>Chi-Square | Standardized<br>Estimate |
|----------|-----------------------|-------------------|--------------------|--------------------|--------------------------|
| INTERCPT | -0.0865               | 0.9518            | 0.0083             | 0.9275             | .                        |
| LWT      | -0.0159               | 0.00686           | 5.3831             | 0.0203             | -0.268152                |
| RACE1    | 1.3257                | 0.5222            | 6.4440             | 0.0111             | 0.252425                 |
| RACE2    | 0.8971                | 0.4339            | 4.2748             | 0.0387             | 0.237218                 |
| SMOKE    | 0.9387                | 0.3987            | 5.5430             | 0.0186             | 0.253282                 |
| PTL      | 0.5032                | 0.3412            | 2.1747             | 0.1403             | 0.136871                 |
| HT       | 1.8550                | 0.6951            | 7.1217             | 0.0076             | 0.250053                 |
| UI       | 0.7857                | 0.4564            | 2.9630             | 0.0852             | 0.154294                 |

**Table 1 (continued)**

**Association of Predicted Probabilities and Observed Responses**

|                    |                   |
|--------------------|-------------------|
| Concordant = 74.3% | Somers' D = 0.491 |
| Discordant = 25.2% | Gamma = 0.493     |
| Tied = 0.5%        | Tau-a = 0.212     |
| (7670 pairs)       | c = 0.746         |

**THE COVARIANCE MATRIX OF THE BETA COEFFICIENTS**

| OBS | _LINK_ | _TYPE_ | _NAME_   | INTERCEP | LWT       | RACE1    |
|-----|--------|--------|----------|----------|-----------|----------|
| 1   | LOGIT  | PARMS  | ESTIMATE | -0.08655 | -0.015905 | 1.32572  |
| 2   | LOGIT  | COV    | INTERCPT | 0.90586  | -0.005928 | -0.01419 |
| 3   | LOGIT  | COV    | LWT      | -0.00593 | 0.000047  | -0.00071 |
| 4   | LOGIT  | COV    | RACE1    | -0.01419 | -0.000707 | 0.27274  |
| 5   | LOGIT  | COV    | RACE2    | -0.17726 | 0.000418  | 0.08526  |
| 6   | LOGIT  | COV    | SMOKE    | -0.12587 | 0.000145  | 0.04383  |
| 7   | LOGIT  | COV    | PTL      | -0.03011 | 0.000144  | 0.00198  |
| 8   | LOGIT  | COV    | HT       | 0.14996  | -0.001562 | 0.01516  |
| 9   | LOGIT  | COV    | UI       | -0.05529 | 0.000132  | 0.01141  |

| OBS | RACE2    | SMOKE    | PTL      | HT       | UI       |
|-----|----------|----------|----------|----------|----------|
| 1   | 0.89708  | 0.93873  | 0.50321  | 1.85504  | 0.78570  |
| 2   | -0.17726 | -0.12587 | -0.03011 | 0.14996  | -0.05529 |
| 3   | 0.00042  | 0.00015  | 0.00014  | -0.00156 | 0.00013  |
| 4   | 0.08526  | 0.04383  | 0.00198  | 0.01516  | 0.01141  |
| 5   | 0.18826  | 0.08278  | -0.00505 | -0.00190 | 0.00083  |
| 6   | 0.08278  | 0.15898  | -0.01976 | 0.00133  | 0.00356  |
| 7   | -0.00505 | -0.01976 | 0.11644  | 0.00016  | -0.02444 |
| 8   | -0.00190 | 0.00133  | 0.00016  | 0.48319  | 0.03341  |
| 9   | 0.00083  | 0.00356  | -0.02444 | 0.03341  | 0.20834  |

**ODDS RATIO AND 95% CI FOR ODDS RATIO**

| OBS | VARIABLE | Beta Coefficient | Standard Error: Betas | Estimated Odds Ratio | Lower 95% CI: Odds Ratio | Upper 95% CI: Odds Ratio |
|-----|----------|------------------|-----------------------|----------------------|--------------------------|--------------------------|
| 1   | INTERCEP | -0.08655         | 0.95177               | 0.91709              | 0.14199                  | 5.9234                   |
| 2   | LWT      | -0.01591         | 0.00686               | 0.98422              | 0.97108                  | 0.9975                   |
| 3   | RACE1    | 1.32572          | 0.52225               | 3.76489              | 1.35272                  | 10.4785                  |
| 4   | RACE2    | 0.89708          | 0.43388               | 2.45243              | 1.04777                  | 5.7402                   |
| 5   | SMOKE    | 0.93873          | 0.39872               | 2.55672              | 1.17027                  | 5.5857                   |
| 6   | PTL      | 0.50321          | 0.34123               | 1.65403              | 0.84738                  | 3.2285                   |
| 7   | HT       | 1.85504          | 0.69512               | 6.39196              | 1.63657                  | 24.9651                  |
| 8   | UI       | 0.78570          | 0.45644               | 2.19394              | 0.89679                  | 5.3673                   |

## 5. Goodness of Fit

In addition to testing hypothesis concerning the significance of individual independent variables, we would like to determine how adequate the model we have is in describing the outcome variable. This is referred to as its goodness of fit. This corresponds to the global F test in linear regression.

The two major components for assessing the fit of the model are summary measures of goodness-of-fit and logistic regression diagnostics for individual observations. The procedure: PROC LOGISTIC in SAS 6.06 produces quite complete results in the logistic regression diagnostic part. Therefore, this paper will only concentrate on how to compute various summary goodness-of-fit statistics by using PROC LOGISTIC.

In general, we will conclude that the model fits if the summary measures of the distance between observed and predicted values are small. In logistic regression, there are several possible measures of difference between the observed and predicted values.

### 5.1 Pearson and Log-likelihood Chi-Square Statistics

Before discussing specific goodness-of-fit statistics, the term covariate pattern is commonly used to describe each unique combination of predictors (or covariates) in the model.

The number of covariate patterns, denoted by J, equals the product of the number of levels of each predictor. And the number of subjects within covariate pattern j is denoted by  $m_j$ . For each covariate pattern, we can predict the probability of being a success or failure from the model. The predicted probability for covariate pattern can be denoted by  $\hat{\pi}_j$ . Let  $y_j$  denotes the number of positive response,  $y=1$ , among the  $m_j$  subjects in the jth covariate pattern, the Pearson chi-square statistic is:

$$X^2 = \sum_{j=1}^J \left[ \frac{(y_j - m_j \hat{\pi}_j)^2}{m_j \hat{\pi}_j} + \frac{((m_j - y_j) - m_j (1 - \hat{\pi}_j))^2}{m_j (1 - \hat{\pi}_j)} \right] \dots \dots (7)$$

The log-likelihood chi-square statistic can be defined as follows:

$$D = \sum_{j=1}^J 2 \left[ y_j \ln \left( \frac{y_j}{m_j \hat{\pi}_j} \right) + (m_j - y_j) \ln \left( \frac{(m_j - y_j)}{m_j (1 - \hat{\pi}_j)} \right) \right] \dots \dots (8)$$

The distribution of the statistics  $X^2$  and  $D$  is supposed to be chi-square with degree of freedom equal to  $J-(p+1)$ , where  $p$  is the number of predictors fitted in the model. The null hypothesis is that the model fits well. Since we want our model to fit well, we will not conclude that it fits well unless the  $p$  value is greater than 0.1 or even 0.2. Setting  $\alpha$  at these level allows us to avoid type two error.

The following program shows the computation of Pearson and log-likelihood chi-square statistics for the logistic regression model fitted in the low birth weight data. The results are presented in Table 2.

---

**PROGRAM: Assessing Fit-The Pearson and Likelihood Ratio Chi-Square**

---

```

PROC SORT DATA=LL.LOWBW;
  BY DESCENDING LOW;
PROC LOGISTIC DATA=LL.LOWBW ORDER=DATA;
  MODEL LOW=LWD SMOKE PTD;
  OUTPUT OUT=PHATOUT4 PRED=PHAT;
TITLE 'ASSESSING FIT: PEARSON CHI-SQUARE AND LOG-LIKELIHOOD RATIO STATISTICS';
PROC SUMMARY SUM;
  CLASS LWD SMOKE PTD ;
  VAR LOW PHAT;
  OUTPUT OUT=PHATOUT6 SUM=OSUCCESS ESUCCESS;
DATA PHATOUT7;
  SET PHATOUT6 END=EOF;
  IF _TYPE_=7;
  OFAILURE=_FREQ_-OSUCCESS;
  EFAILURE=_FREQ_-ESUCCESS;
  CHISQi=(((OSUCCESS-ESUCCESS)**2)/ESUCCESS) + (((OFAILURE-EFAILURE)**2)/EFAILURE);
  LLRi=2*((OSUCCESS*(LOG(OSUCCESS/ESUCCESS)))+(OFAILURE*(LOG(OFAILURE/EFAILURE))));
  CHISQ+CHISQi;
  LLR+LLRi;
  IF EOF THEN OUTPUT;
DATA L;
  SET PHATOUT7;
  CHIPVAL=1-PROBCHI(CHISQ,4);
  LLPVAL=1-PROBCHI(LLR,4);
  LABEL CHISQ='Pearson Chi-Square'
        LLR='Log-Likelihood Ratio'
        CHIPVAL='P-value: Pearson Chi-Square'
        LLPVAL='P-value: Log-Likelihood Ratio';
PROC PRINT LABEL;
  Var CHISQ CHIPVAL LLR LLPVAL;

```

## Table 2

### Output: Pearson and Log-likelihood Chi-Square Statistics

#### The LOGISTIC Procedure

#### Response Profile

| Ordered Value | LOW | Count |
|---------------|-----|-------|
| 1             | 1   | 59    |
| 2             | 0   | 130   |

#### Simple Statistics for Explanatory Variables

| Variable | Mean     | Standard Deviation | Minimum | Maximum |
|----------|----------|--------------------|---------|---------|
| LWD      | 0.222222 | 0.416844           | 0       | 1.00000 |
| SMOKE    | 0.391534 | 0.489390           | 0       | 1.00000 |
| PTD      | 0.158730 | 0.366395           | 0       | 1.00000 |

#### Criteria for Assessing Model Fit

| Criterion | Intercept Only | Intercept and Covariates | Chi-Square for Covariates   |
|-----------|----------------|--------------------------|-----------------------------|
| AIC       | 236.672        | 221.318                  | .                           |
| SC        | 239.914        | 234.285                  | .                           |
| -2 LOG L  | 234.672        | 213.318                  | 21.354 with 3 DF (p=0.0001) |
| Score     | .              | .                        | 22.053 with 3 DF (p=0.0001) |

#### Analysis of Maximum Likelihood Estimates

| Variable | Parameter Estimate | Standard Error | Wald Chi-Square | Pr > Chi-Square | Standardized Estimate |
|----------|--------------------|----------------|-----------------|-----------------|-----------------------|
| INTERCPT | -1.4574            | 0.2495         | 34.1078         | 0.0001          | .                     |
| LWD      | 0.9299             | 0.3771         | 6.0813          | 0.0137          | 0.213715              |
| SMOKE    | 0.4703             | 0.3403         | 1.9101          | 0.1669          | 0.126907              |
| PTD      | 1.2950             | 0.4287         | 9.1231          | 0.0025          | 0.261588              |

#### Association of Predicted Probabilities and Observed Responses

|                    |                   |
|--------------------|-------------------|
| Concordant = 59.4% | Somers' D = 0.396 |
| Discordant = 19.8% | Gamma = 0.500     |
| Tied = 20.8%       | Tau-a = 0.171     |
| (7670 pairs)       | c = 0.698         |

| OBS | Pearson Chi-Square | P-value: Pearson Chi-Square | Log-Likelihood Ratio | P-value: Log-Likelihood Ratio |
|-----|--------------------|-----------------------------|----------------------|-------------------------------|
| 1   | 5.00564            | 0.28672                     | 4.95335              | 0.29212                       |

The following program shows how to compute Hosmer-Lemeshow statistic to assess the fit of a logistic model containing continuous covariate (AGE) for the low birth weight data. The results are present in Table 3.

---

**PROGRAM: Assessing Fit- The Hosmer-Lemeshow Test**

---

```

PROC LOGISTIC DATA=LLLOWBW ORDER=DATA;
  MODEL LOW=AGE RACE1 RACE2 SMOKE HT UI LWD PTD AGELWD SMOKELWD;
  OUTPUT OUT=LLLOWBWOUT PRED=PHAT;
  TITLE 'ASSESSING FIT: HOSMER-LEMESHOW TEST STATISTICS';

PROC RANK DATA=LLLOWBWOUT OUT=PHATOUT1 GROUPS=10;
  VAR PHAT;
  RANKS GP;

PROC SUMMARY DATA=PHATOUT1 SUM;
  CLASS GP;
  VAR LOW PHAT;
  OUTPUT OUT=PHATOUT2 SUM=OSUCCESS ESUCCESS ;

DATA PHATOUT3;
  SET PHATOUT2 END=EOF;
  IF _TYPE_=0 THEN DELETE;
  OFAILEURE=_FREQ_-OSUCCESS;
  EFAILEURE=_FREQ_-ESUCCESS;
  CHISQ1=(((OSUCCESS-ESUCCESS)**2)/ESUCCESS) + (((OFAILEURE-EFAILEURE)**2)/EFAILEURE);
  CHISQ+CHISQ1;
  IF EOF THEN OUTPUT;
DATA P;
  SET PHATOUT3;
  PVALUE=1-PROBCHI(CHISQ,8);
  LABEL CHISQ='HOSMER-LEMESHOW STATISTICS'
  PVALUE='P-VALUE:HOSMER-LEMESHOW STATISTICS';
PROC PRINT LABEL;
  VAR CHISQ PVALUE;
RUN;

```

## 5.2 The Hosmer-Lemeshow Test

If a continuous predictor is included in the model, there will be a large number of covariate patterns. Under this circumstance, the number of covariate patterns:  $J$  approximately equals the number of subjects:  $n$  ( $J \approx n$ ) and the expected frequency in each cell will be small. Since chi-square tests require that 75-80% of the cells have expected values greater than 5, the p-values calculated for Pearson ( $X^2$ ) and log-likelihood chi-square (D) statistics, using  $X^2$  ( $J-P-1$ ) distribution, are incorrect.

One way to avoid the above noted difficulties with the distribution of  $X^2$  and D when  $J \approx n$  is to group the data. Hosmer and Lemeshow (1980)<sup>4</sup> and Lemeshow and Hosmer (1982)<sup>5</sup> proposed grouping based on the values of the predicted probabilities. The strategy they suggested was to rank the predicted probabilities from lowest to highest, and then collapse the data based on percentiles of the predicted probabilities into 10 groups. This method results in the first group containing  $n/10$  subjects having the smallest estimated probabilities, and the last group also containing  $n/10$  subjects having the largest estimated probabilities. Also with this method, we can create a table with the 10 groups displayed in the column and 2 rows- one for success ( $y=1$ ) and the other for failure ( $y=0$ ). For the  $y=1$  row, estimates of the expected frequencies for each of the 10 groups are obtained by summing the estimated probabilities over all subjects in the group. For  $y=0$  row, the estimated expected frequencies is obtained by summing, over all subjects in the group, one minus the estimated probability. The Hosmer-Lemeshow goodness-of-fit statistic,  $\hat{C}$ , is obtained by calculating the Pearson chi-square statistic for the  $2 \times 10$  table of observed and estimated expected frequencies.

Using an extensive set of simulation, Hosmer and Lemeshow (1980)<sup>4</sup> demonstrated that the distribution of  $\hat{C}$  is well approximated by the chi-square distribution with  $g-2$  degrees of freedom. The notation,  $g$ , represents the number of groups and is 10 usually.

# Table 3

## Output: The Hosmer-Lemeshow Test

### The LOGISTIC Procedure

Data Set: LL.LOWBW  
 Response Variable: LOW  
 Response Levels: 2  
 Number of Observations: 189  
 Link Function: Logit

### Response Profile

| Ordered<br>Value | LOW | Count |
|------------------|-----|-------|
| 1                | 1   | 59    |
| 2                | 0   | 130   |

### Simple Statistics for Explanatory Variables

| Variable | Mean      | Standard<br>Deviation | Minimum | Maximum |
|----------|-----------|-----------------------|---------|---------|
| AGE      | 23.238095 | 5.298678              | 14.0000 | 45.0000 |
| RACE1    | 0.137566  | 0.345359              | 0.0000  | 1.0000  |
| RACE2    | 0.354497  | 0.479631              | 0.0000  | 1.0000  |
| SMOKE    | 0.391534  | 0.489390              | 0.0000  | 1.0000  |
| HT       | 0.063492  | 0.244494              | 0.0000  | 1.0000  |
| UI       | 0.148148  | 0.356190              | 0.0000  | 1.0000  |
| LWD      | 0.222222  | 0.416844              | 0.0000  | 1.0000  |
| PTD      | 0.158730  | 0.366395              | 0.0000  | 1.0000  |
| AGELWD   | 4.941799  | 9.552865              | 0.0000  | 33.0000 |
| SMOKELWD | 0.111111  | 0.315104              | 0.0000  | 1.0000  |

### Criteria for Assessing Model Fit

| Criterion | Intercept<br>Only | Intercept<br>and<br>Covariates | Chi-Square for Covariates    |
|-----------|-------------------|--------------------------------|------------------------------|
| AIC       | 236.672           | 214.012                        | .                            |
| SC        | 239.914           | 249.672                        | .                            |
| -2 LOG L  | 234.672           | 192.012                        | 42.660 with 10 DF (p=0.0001) |
| Score     | .                 | .                              | 39.204 with 10 DF (p=0.0001) |

**Table 3 (continued)**

**The LOGISTIC Procedure**

**Analysis of Maximum Likelihood Estimates**

| Variable | Parameter Estimate | Standard Error | Wald Chi-Square | Pr > Chi-Square | Standardized Estimate |
|----------|--------------------|----------------|-----------------|-----------------|-----------------------|
| INTERCPT | -0.5118            | 1.0875         | 0.2214          | 0.6380          | .                     |
| AGE      | -0.0840            | 0.0456         | 3.3966          | 0.0653          | -0.245327             |
| RACE1    | 1.0831             | 0.5189         | 4.3566          | 0.0369          | 0.206230              |
| RACE2    | 0.7597             | 0.4640         | 2.6802          | 0.1016          | 0.200885              |
| SMOKE    | 1.1531             | 0.4584         | 6.3270          | 0.0119          | 0.311132              |
| HT       | 1.3592             | 0.6615         | 4.2224          | 0.0399          | 0.183217              |
| UI       | 0.7282             | 0.4795         | 2.3063          | 0.1288          | 0.142996              |
| LWD      | -1.7299            | 1.8683         | 0.8574          | 0.3545          | -0.397574             |
| PTD      | 1.2316             | 0.4714         | 6.8259          | 0.0090          | 0.248784              |
| AGELWD   | 0.1474             | 0.0829         | 3.1650          | 0.0752          | 0.776381              |
| SMOKELWD | -1.4074            | 0.8187         | 2.9553          | 0.0856          | -0.244498             |

**Association of Predicted Probabilities and Observed Responses**

|                    |                   |
|--------------------|-------------------|
| Concordant = 78.2% | Somers' D = 0.569 |
| Discordant = 21.3% | Gamma = 0.571     |
| Tied = 0.5%        | Tau-a = 0.245     |
| (7670 pairs)       | c = 0.784         |

**ASSESSING FIT: HOSMER-LEMESHOW TEST STATISTICS**

| OBS | HOSMER-LEMESHOW STATISTICS | P-VALUE:HOSMER-LEMESHOW STATISTICS |
|-----|----------------------------|------------------------------------|
| 1   | 5.23078                    | 0.73265                            |

### 5.3 Other Summary Statistic: $R^2$

For sake of completeness, Hosmer (1989)<sup>6</sup> also suggested  $R^2$ -type measure for use with logistic regression. According to them, the  $R^2$  for the logistic regression model is:

$$R_L^2 = 100 (L_0 - L_p) / (L_0 - L_s) \dots\dots (9)$$

Where  $L_0$  and  $L_p$  denote the log-likelihoods for models containing only the intercept, and the model containing the intercept plus the  $p$  covariates respectively.  $L_s$  is the log-likelihood for the saturated model, which can be obtained from the following formula:

$$L_s = L_p + 0.5D \dots\dots\dots (10)$$

$D$  stands for the log-likelihood chi-square statistic and can be calculated based on equation (8) as shown in the above.

Thus, Hosmer commented that the quantity  $R^2$  is nothing more than an expression of the likelihood ratio test. For this reason,  $R^2$  is not a measure of goodness-of-fit itself but can only be used to supplement the results with other goodness-of-fit statistics.

Based on equation (9), the following program calculates the  $R^2$  for the logistic regression model in low birth weight data. The results are presented in Table 4.

---

## PROGRAM: To Compute the R-Square

---

```
PROC LOGISTIC DATA=LL.LOWBW ORDER=DATA NOPRINT;
  MODEL LOW=;
  OUTPUT OUT=LL.MODEL0 PRED=PHAT0;

DATA COMBINE ;
  MERGE LL.LOWBWOUT
        LL.MODEL0 END=EOF;
  LL0i=LOG((PHAT0**LOW)*((1-PHAT0)**(1-LOW)));
  LLPi=LOG((PHAT**LOW)*((1-PHAT)**(1-LOW)));
  LL0+LL0i;
  LLP+LLPi;
  IF EOF THEN OUTPUT;
  KEEP LOW PHAT0 PHAT LL0 LLP;

PROC SUMMARY DATA=LL.LOWBWOUT SUM;
  CLASS AGE RACE1 RACE2 SMOKE HT UI LWD PTD AGELWD SMOKELWD;
  VAR LOW PHAT;
  OUTPUT OUT=PHATOUT4 SUM=OSUCCESS ESUCCESS ;

DATA PHATOUT5;
  SET PHATOUT4 END=EOF;
  IF _TYPE_=1023 ;
  OFAILURE=_FREQ_-OSUCCESS;
  EFAILURE=_FREQ_-ESUCCESS;
  DSQi=2*(OSUCCESS*(LOG(OSUCCESS/ESUCCESS)))+2*(OFAILURE*(LOG(OFAILURE/EFAILURE)));
  CHISQi=(((OSUCCESS-ESUCCESS)**2)/ESUCCESS) + (((OFAILURE-EFAILURE)**2)/EFAILURE);
  IF OSUCCESS=0 THEN DO;
    DSQi=2*(OFAILURE*(LOG(OFAILURE/EFAILURE)));
  END;
  IF OFAILURE=0 THEN DO;
    DSQi=2*(OSUCCESS*(LOG(OSUCCESS/ESUCCESS)));
  END;
  DSQ+DSQi;
  CHISQ+CHISQi;
  IF EOF THEN OUTPUT;

DATA FINAL;
  MERGE COMBINE PHATOUT5;
  LLS=LLP+(0.5*DSQ);
  RSQUARE=100*(LL0-LLP)/(LL0-LLS);
  LABEL LL0='LOG-LIKELIHOOD:INTERCEPT ONLY'
        LLP='LOG-LIKELIHOOD:INTERCEPT+COVARIATES'
        LLS='LOG-LIKELIHOOD:SATURATED MODEL'
        RSQUARE='R-SQUARE';
PROC PRINT LABEL;
  VAR LL0 LLP LLS RSQUARE;
  TITLE 'ASSESSING FIT: R-SQUARE STATISTICS';
RUN;
```

**Table 4**  
**Output: R<sup>2</sup>**

| OBS | LOG-LIKELIHOOD: INTERCEPT<br>ONLY  | LOG-LIKELIHOOD: INTERCEPT+COVARIATES |
|-----|------------------------------------|--------------------------------------|
| 1   | -117.336                           | -96.0062                             |
| OBS | LOG-LIKELIHOOD: SATURATED<br>MODEL | R-SQUARE                             |
| 1   | -20.7073                           | 22.0740                              |

## 6. Conclusion

With minor modifications, the programs presented in this paper can be used to compute the odds ratio and goodness-of-fit statistics for logistic regression model in various data sets other than the low birth weight data. In the future, I plan to convert these programs into SAS MACROS so that they can be applied universally to various logistic models in various data sets.

## References

1. Cornfield, J., Gordon, T., Smith, W.W. (1961). Quantal response curves for experimentally uncontrolled variables. *Bulletin of the International Statistical Institute* 38:97-115.
2. Cornfield, J. (1962). Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis. *Federation Proceedings* 21:58-61.
3. Hosmer, D.W. and Lemeshow, S. (1989). *Applied logistic regression*. John Wiley & Sons, Inc. New York. Appendix 1, 247-252.
4. Hosmer, D.W. and Lemeshow, S. (1980). A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics*, A10, 1043-1069.
5. Lemeshow, S. and Hosmer, D.W. (1982). The use of goodness-of-fit statistics in the development of logistic regression models. *American Journal of Epidemiology*, 115, 92-106.
6. Hosmer, D.W. and Lemeshow, S. (1989). *Applied logistic regression*. John Wiley & Sons, Inc. New York. 148-149.