MCSTRAT: A SAS Macro to Analyze Data From a Matched or Finely Stratified Case-Control Design

Robert A. Vierkant Terry M. Therneau
Jon L. Kosanke James M. Naessens

Contents

1 Introduction

A case-control design is a common approach used to assess disease-exposure relationships, and the logistic regression model is the most common framework for the analysis of such data. This model expresses the logit transform of the disease probability as a linear combination of independent, or exposure, variables. Let y be a disease outcome of interest, taking on a value of 1 if the disease is present and 0 is the disease is absent. Let $\mathbf{x}$ be a vector of independent variables, and let β be a vector of (unknown) coefficients corresponding to $\mathbf{x}$. A logistic regression analysis models the probability of an outcome as

$$Pr(y) = \frac{\exp(\beta'\mathbf{x})}{1 + \exp(\beta'\mathbf{x})}.$$

When designing case-control studies, it is often useful to match controls to cases based on certain factors in order to minimize inherent variation within these factors. However, for the valid analysis of such an approach, a modeling technique that correctly incorporates the matched nature of the data is needed. This prohibits the use of a standard unconditional logistic regression analysis generally available with the SAS procedure LOGISTIC [1]. A stratified conditional logistic regression analysis has the same modeling flexibility as an unconditional analysis, but can also take into account the correlation structure attributable to matching without violation of model assumptions. This report presents a SAS macro that fits a conditional logistic regression model to matched or finely stratified case-control data using SAS procedure PHREG, taking advantage of an identity between the conditional logistic likelihood and the Cox proportional hazards likelihood. The macro enhances standard PHREG output by producing summary tables and statistics used to describe the matched sets. It also calculates several regression diagnostics, some not available in PHREG, that can be used to assess model fit. The macro serves as an update of the supplemental SAS procedure MCSTRAT [2], which cannot be run with present versions of SAS. Many of the tables and summary statistics produced are exact replications of the supplemental procedure.

2 Conditional Logistic Regression Analysis

2.1 Rationale

An unconditional logistic regression analysis is not a valid method of analyzing matched or finely stratified data because the optimality properties of its maximum likelihood method do not hold when the number of subjects in a stratum or matched set becomes small. A conditional analysis circumvents this problem by forming an "exact" likelihood function. Within each stratum, a likelihood function is formed based on an exhaustive enumeration of all possible combinations of cases and controls, conditional on the total number of cases and controls in the stratum. Assume that the case-control data set is composed of K matched sets, each with n_{1k} cases and n_{0k} controls, $k = 1, 2, \ldots, K$. Let $n_k = n_{1k} + n_{0k}$. Then the likelihood function for the kth stratum is

$$l_k(\beta) = \frac{\prod_{i=1}^{n_{1k}} P(\mathbf{x}_i|y=1) \prod_{i=n_{1k}+1}^{n_k} P(\mathbf{x}_i|y=0)}{\sum_j \{\prod_{i_j=1}^{n_{1k}} P(\mathbf{x}_{ji_j}|y=1) \prod_{i_j=n_{1k}+1}^{n_k} P(\mathbf{x}_{ji_j}|y=0)\}}$$

where the summation over j in the denominator is over all $\binom{n_k}{n_{1k}}$ combinations. Application of Bayes theorem to each term above gives the following equation:

$$l_k(\beta) = \frac{\prod_{i=1}^{n_{1k}} \exp(\beta' \mathbf{x}_i)}{\sum_j \prod_{i_j=1}^{n_{1k}} \exp(\beta' \mathbf{x}_{ji_j})}.$$

The full conditional likelihood is the product of the individual likelihoods over the K matched sets,

$$l(\beta) = \prod_{k=1}^{K} l_k(\beta).$$

More detail can be found elsewhere [3, 4, 5].

2.2 Similarities with the Cox Model

There are striking similarities between the conditional logistic likelihood function and the partial likelihood function used to fit a Cox proportional hazards model. Each is fit using a conditional likelihood technique. Let y and $\mathbf{x}$ be defined as earlier, but assume that individuals are now followed prospectively until either they develop the disease of interest or are censored. Denote the survival time of a patient as t. The Cox model is estimating a hazard function of developing the disease of interest using a log-linear combination of independent variables,

$$\lambda(t) = \lambda_0(t) \exp(\beta' \mathbf{x}),$$

where λ_0 is an arbitrary baseline hazard function.

The Cox model likelihood is a product of conditional probabilities. Assume first that survival times are continuously distributed and the possibility of tied event times can be ignored. At each event time t_l, a risk set is formed that consists of all subjects still under study just prior to t_l. The conditional probability that individual i has an event at time t_l given one individual in the risk set has an event at that time is then calculated. The likelihood is formed by multiplying these conditional probabilities over all events,

$$l(\beta) = \prod_{l=1}^{L} \frac{\exp(\beta' \mathbf{x}_l)}{\sum\limits_{R} \exp(\beta' \mathbf{x}_r)}$$

where the summation over R in the denominator is over all individuals in the risk set.

Several methods of forming a likelihood for the Cox model in the presence of tied event times have been developed. One such method uses an exact partial likelihood and involves an exhaustive enumeration of the possible events at each tied event time given the total number of events and the number of individuals in the risk set. Assume that event time l corresponds to n_{1l} events. Then the exact partial likelihood is formed as

$$l(\beta) = \prod_{l=1}^{L} \frac{\prod_{i=1}^{n_{1l}} \exp(\beta' \mathbf{x}_i)}{\sum_j \prod_{i_j=1}^{n_{1l}} \exp(\beta' \mathbf{x}_{ji_j})}$$

where the summation over j in the denominator is over all possible combinations of n_{1l} events among all individuals in the risk set. If there is only a single event time, this equation reduces to

$$l(\beta) = \frac{\prod_{i=1}^{n_1} \exp(\beta' \mathbf{x}_i)}{\sum_j \prod_{i_j=1}^{n_1} \exp(\beta' \mathbf{x}_{ji_j})}. \tag{1}$$

In a stratified Cox analysis, the likelihood function is the product of the partial likelihoods for the individual strata. The likelihood for such an analysis with K strata then becomes

$$l(\beta) = \prod_{k=1}^{K} l_k(\beta)$$

where

$$l_k(\beta) = \frac{\prod_{i=1}^{n_{1k}} \exp(\beta' \mathbf{x}_i)}{\sum_j \prod_{i_j=1}^{n_{1k}} \exp(\beta' \mathbf{x}_{ji_j})}$$

which is exactly the same as the likelihood function for a conditional logistic regression analysis. Thus, a matched case-control data set can be fit using a Cox proportional hazards analysis if (1) each matched set is treated as a stratum, (2) all cases within a given matched set are assumed to have the same event time, and (3) the exact partial likelihood method is used to model the data. The following SAS code fits a conditional logistic regression model to matched case-control data using PROC PHREG.

```
proc phreg;
   model TIME*CASE(0)=X1 X2 / ties=discrete;
   strata SET;
```

Here case refers to case-control status, with zero indicating the variable level for controls. TIME is a dummy variable in this application and should be coded so that all cases and controls have the same non-zero value. X1 and X2 are the independent variables of interest. The variable SET is used in the strata statement to uniquely define each matched set. Finally, the ties=discrete option specifies the use of the exact partial likelihood to fit the data in the presence of tied event times. This option is necessary when there exist strata that contain more than one case. For 1:k matching, however (i.e., one case per stratum), equation (1) holds and all of the available methods for handling ties give the same result.

3 Diagnostics

Diagnostics can be used in regression analyses to assess influence of an observation or a matched set on model fit. An introduction to linear regression diagnostics can be found in Neter, Wasserman, and Kutner [6]. Many of these techniques were first applied to logistic regression analyses by Pregibon [7], and extended to conditional logistic regression analyses by Pregibon [8] and Moolgavkar et al [9]. The SAS macro MCSTRAT produces several measures similar to these in addition to others. Some of the diagnostics produced are available through SAS PROC PHREG. These PHREG diagnostics were created specifically for Cox proportional hazards analyses and have been proven to be effective in such analyses. They have not been readily used in conditional logistic analyses, although one would expect they would be valid tools for such an approach, based on the similarities in the likelihood functions between the two models. The macro also produces several regression diagnostics which were created specifically for conditional logistic regression analyses. These "established" diagnostics have been shown to be effective tools for examining model fit in a logistic regression framework and can serve as a benchmark for the PHREG diagnostics.

3.1 The Cox Model in a Cross-Sectional Context

Before discussion of the diagnostics available using PROC PHREG, it is necessary to re-express some of the Cox proportional hazards equations in a cross-sectional context. Consider a set of n_k subjects in stratum k such that the counting process $N_i \equiv \{N_i(t), t \geq 0\}$ for the ith subject represents the number of observed events experienced over time t. The sample paths of the process N_i are step functions with jumps of size +1, with $N_i(0) = 0$. Let $Y_i(t)$ indicate whether the ith subject is at risk at time t,

$$Y_i(t) = \begin{cases} 1 & \text{subject is at risk} \\ 0 & \text{otherwise} \end{cases} .$$

Then the cumulative baseline hazard function for stratum k is estimated by

$$\hat{\Lambda}_{0k}(t) = \sum_{i=1}^{n_k} \int_0^t \frac{dN_i(s)}{\sum_{j=1}^{n_k} Y_j(s) \exp(\hat{\beta}' \mathbf{x}_j(s))}.$$

However, for a matched case-control data set, $Y_i \equiv 1$ for every individual i. Also, $N_i \equiv 1$ for all cases and 0 for all controls. Finally, since a case-control study is cross-sectional, there is no integration over time. Thus, the cumulative baseline hazard function reduces to

$$\hat{\Lambda}_{0k}(t) = \hat{\Lambda}_{0k} = \sum_{i=1}^{n_k} \frac{N_i}{\sum_{j=1}^{n_k} \exp(\hat{\beta}' \mathbf{x_j})}$$

where the sum is over all observations in a stratum or matched set.

For a Cox model with no time-dependent covariates, the martingale residuals can be expressed as

$$\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(s) \exp(\hat{\beta}' \mathbf{x}_i(s)) d\hat{\Lambda}_{0k}(s),$$

which for a matched case-control study reduces to

$$\hat{M}_i = N_i - \frac{c \ \exp(\hat{\beta}' \mathbf{x}_i)}{\sum_{j=1}^{n_k} \exp(\hat{\beta}' \mathbf{x}_j)}$$

where c is the number of cases in the matched set. This can be re-expressed as

$$\hat{M}_i = N_i - \hat{\xi}_i,$$

where the $\hat{\xi}_i$ are the model's "fitted values" and can be interpreted as the estimated probability that a given individual is a case.

The vector of score residuals is

$$\mathbf{L}_i(t) = \int_0^t [\mathbf{x}_i(s) - \overline{\mathbf{x}}(s)] d\hat{M}_i(s)$$

where $\overline{\mathbf{x}}(t)$ is a weighted average of covariates in risk set,

$$\overline{\mathbf{x}}(t) = \frac{\sum_{i=1}^{n_k} \mathbf{x}_i(t) \exp(\hat{\beta}' \mathbf{x}_i(t))}{\sum_{i=1}^{n_k} \exp(\hat{\beta}' \mathbf{x}_i(t))}.$$

For a matched case-control study, this reduces to

$$\mathbf{L}_i = [\mathbf{x}_i - \overline{\mathbf{x}}] \hat{M}_i.$$

3.2 PHREG Diagnostics

The following diagnostics are produced using PROC PHREG.

- DFBETA Statistics ($\hat{\mathbf{\Delta}}_i$).
  This vector of diagnostics approximates the changes in individual parameter estimates due to the deletion of a subject [10]. The approximation is derived as a function of weighted score residuals, and is calculated as

  $$\hat{\mathbf{\Delta}}_i = \hat{\mathbf{V}} \mathbf{L}_i$$

  where $\hat{\mathbf{V}}$ is the estimated covariance matrix of $\hat{\beta}$ from the Cox model and $\mathbf{L}_i$ is the vector of score residuals for the *ith* subject. A diagnostic

is calculated for each independent variable in the model, and values can be negative or positive depending on how the individual affects model fit. A positive value reflects a positive effect on $\hat{\beta}$; in other words, the subject makes $\hat{\beta}$ larger, and removal of the subject will decrease $\hat{\beta}$.

Scaled DFBETA statistics are also useful tools for assessing model fit, and are calculated simply by dividing the original DFBETA statistic by the standard error of the regression coefficient obtained from the model's covariance matrix.

- Likelihood Displacement Statistic (LD).
  The LD statistic approximates the likelihood displacement, which is the change in twice the log-likelihood due to the deletion of a subject. It is calculated as

$$\mathbf{LD}_i = \mathbf{L}_i' \hat{\mathbf{V}} \mathbf{L}_i.$$

  It is a global diagnostic in that it assesses influence of an individual on the overall fit of the model rather than on individual regression coefficients. Its values are always positive.

- LMAX Statistic.
  The LMAX statistic was derived as the Cox proportional hazards equivalent of Cook's distance [11]. It is based on the matrix

$$\mathbf{B} = \mathbf{L} \hat{\mathbf{V}} \mathbf{L}'$$

  where $\mathbf{L}$ is a matrix with rows that correspond to the score residual vectors $\mathbf{L}_i$. $LMAX$ is the unit length eigenvector of $\mathbf{B}$ which has the largest eigenvalue λ_{max}. $LMAX_i$ is then the absolute value of the ith element of $LMAX$, corresponding to individual i. It is a global diagnostic that only takes on positive values.

3.3 Established Conditional Logistic Diagnostics

These diagnostics are reviewed in detail by Hosmer and Lemeshow [5]. Each can be considered a global diagnostic and each takes on only positive values.

- Leverage Values (h).
  These values are the diagonal elements of the hat matrix as derived by Pregibon [7], and are calculated as

$$h_i = \hat{\xi}_i[\mathbf{x}_i - \overline{\mathbf{x}}]\mathbf{V}^{-1}[\mathbf{x}_i - \overline{\mathbf{x}}]'.$$

- Delta Chi-Square (ΔX_i^2).
  This diagnostic evaluates the decrease in the Pearson chi-square statistic due to the deletion of a subject.

$$\Delta X_i^2 = \frac{\hat{M}_i^2}{\hat{\xi}_i(1-h_i)} = \frac{(N_i - \hat{\xi}_i)^2}{\hat{\xi}_i(1-h_i)}$$

- Influence Statistic (INFL).
  This diagnostic is similar to the PHREG likelihood displacement statistic and assesses the composite change in covariate estimates due to the deletion of a subject. It is calculated as

$$INFL_i = \Delta X_i^2 \frac{h_i}{1-h_i}.$$

3.4 Miscellaneous Diagnostic Information

The diagnostics listed above are often plotted against the model's fitted values ($\hat{\xi}_i$), against independent variables already in the model, and against other variables that may affect model fit. It is also helpful in a matched

case-control study to evaluate diagnostics on a matched set level. This is accomplished by summing up the values of the diagnostics over all individuals in a matched set.

4 SAS Macro MCSTRAT

The macro MCSTRAT (available in the SAS autocall library /usr/local/sasmac) contains code to fit a conditional logistic regression model and generate the regression diagnostics $\hat{\boldsymbol{\Delta}}_i$, scaled $\hat{\boldsymbol{\Delta}}_i$, LD, LMAX, h, ΔX_i^2, and INFL as well as the fitted values $\hat{\xi}$ for matched or finely stratified case-control data. The macro first generates tables that describe the matched sets and the independent variables included in the logistic model. It then uses PROC PHREG along with the OUTPUT statement to fit the model and generate the PHREG diagnostics. Next, IML code is used to generate the remaining diagnostics and the model's fitted values. The macro call statement and a description of the keyword macro parameters follows. Required parameters must by specified by the user when calling the macro in order for the program to execute correctly. Optional parameters customize and enhance the regression output. Many of the parameters are given default values that can be overridden by the user.

```
%mcstrat(data=, setid=, case=, indvar=, uni=, cov=,
         mincntl=, mincase=, outdata=, id=, maxiter=,
         epsilon=, tables=, diag=)
```

- DATA (required).
  Specifies the name of the input data set to be used.

- SETID (required).
  Specifies the name of the variable which identifies the matched sets in the input data set. This variable simply indicates to which matched set each observation belongs. All observations in the same matched set should have the same value for this variable. Default name is SETID.

- CASE (required).
  Specifies the name of the case-control indicator variable. Values for this variable must be 1 for cases and 0 for controls.

- INDVAR (required).
  Specifies the names of the independent variables to be included in the logistic model. When specifying more than one independent variable, variable names should be separated by blanks. All variable names should be 7 characters or less, if possible. This allows the macro to reserve a character when naming the DFBETA and scaled DFBETA statistics. If limiting variable names to 7 characters or less is not possible, just make sure that the first 7 characters uniquely distinguish the names of each independent variable from each other.

- UNI (optional).
  Requests that univariate descriptive statistics be printed for each independent variable in the model. If this option is chosen, the macro simply runs PROC MEANS on all independent variables, broken down by case-control status. Care should be taken when interpreting these values, since the output is presented without accounting for the matched nature of the data. Valid parameter values are NO and YES (default).

- COV (optional).
  Requests that the model's estimated covariance matrix be printed. Valid parameter values are NO (default) and YES.

- MINCNTL (optional).
  Specifies the minimum number of controls required in each matched set. Any matched sets not meeting this criterion are excluded from the analysis. Default is 1.

- MINCASE (optional).
  Specifies the minimum number of cases required in each matched set.

Any matched sets not meeting this criterion are excluded from the analysis. Default is 1.

- OUTDATA (optional).
  Names a SAS data set to be created which contains all observations used in the model. This is useful if the number of observations used in the model differs from the number of observations in the original data set (based on restrictions imposed by the MINCNTL and MINCASE parameter values).

- ID (optional).
  Requests that a list of matched sets not included in the model be printed. Valid parameter values are NO (default) and YES.

- MAXITER (optional).
  Specifies the maximum number of iterations to be performed when fitting the logistic model in PHREG. Default is 10.

- EPSILON (optional).
  Specifies the difference in the log-likelihood used to determine model convergence. Default is .000001.

- TABLES (optional).
  Specifies a list of independent variables for which frequency tables should be created. The list should be a subset of the independent variables in the model and should contain only 0/1 or 1/2 indicators. When specifying more than one variable, variable names should be separated by blanks. These tables report how many cases and controls had a value of "1" in each matched set.

- DIAG (optional).
  Requests that output data sets containing regression diagnostics be created. Valid parameter values are NO and YES (default). If this option is chosen, two data sets are created. The first one, called SUBDIAG, contains information on an individual, or subject, level. All individuals used to fit the model and the following variables are included in the data set.

  1. All independent variables in the logistic model (specified with the INDVAR macro parameter).
  2. The case-control variable (specified with the CASE macro parameter).
  3. The set id variable (specified with the SETID macro parameter).
  4. XI $\rightarrow$ The model's fitted values, $\hat{\xi}_i$.
  5. DELTAX2 $\rightarrow$ The ΔX_i^2 statistic assessing the effect of the observation on overall model fit.
  6. INFL $\rightarrow$ The influence statistic assessing the effect of the observation on overall fit of the model.
  7. HAT $\rightarrow$ The leverage values.
  8. LD $\rightarrow$ The likelihood displacement statistic.
  9. LMAX $\rightarrow$ The LMAX statistic assessing the effect of the observation on overall model fit.
  10. D(var) $\rightarrow$ The individual DFBETA statistics assessing the effect of the observation on a particular parameter estimate in the model. One variable is created for each independent variable in the model. (var) corresponds to the first seven characters of the independent variable in the logistic model.
  11. S(var) $\rightarrow$ The scaled DFBETA statistics, created by dividing the original DFBETA by the corresponding independent variable's standard error from the estimated covariance matrix. One variable is created for each independent variable. (var) corresponds to the first seven characters of the independent variable in the logistic model.

The second data set, called SETDIAG, contains diagnostic information on a stratum or matched set level. It contains the sums of the diagnostics from the SUBDIAG data set (summed over all observations in the matched set). All matched sets used to fit the model and the following variables are included in the data set.

1. The set id variable.
2. DELTAX2 $\rightarrow$ The sum of the DELTAX2 diagnostics from SUBDIAG.
3. INFL $\rightarrow$ The sum of the INFL diagnostics from SUBDIAG.
4. HAT $\rightarrow$ The sum of the leverage values from SUBDIAG.
5. LD $\rightarrow$ The sum of the LD diagnostics from SUBDIAG.
6. LMAX $\rightarrow$ The sum of the LMAX diagnostics from SUBDIAG.
7. D(var) $\rightarrow$ The sum of the DFBETA statistics for a particular parameter estimate in the model. One variable is created for each independent variable.

5 Example

The example presented here uses the low birth weight data found in Appendix 4 of Hosmer and Lemeshow [5]. In this example, mothers of low birth weight babies (cases) were matched to three mothers of normal birth weight babies of the same age (controls). Twenty-nine matched sets, each containing one case and three controls, were created. Variables in the final model include smoking status (SMOKE), uterine irritability (UI), presence of a previous pre-term delivery (PTD), and low maternal weight at the last menstrual period (LWD), dichotomized as the lower 25^{th} percentile vs. the upper 75^{th} percentile. All of the variables are dichotomous, taking on values of 1 when the condition is present and 0 when the condition is absent. Independent variable names are exactly the same as the example shown in Hosmer and Lemeshow in order to allow direct comparisons of the results. Variables are

stored in the SAS data set LOWWGT. This data set also contains a variable distinguishing the matched sets (SET) as well as a variable that indicates case-control status (LBW, coded as 1 for cases and 0 for controls). The first part of the code that reads in the data set is shown below.

```
data lowwgt;
   input set lbw lwd smoke ui ptd;

   cards;
1   1    0    0    0     0
1   0    0    0    0     0
1   0    0    1    0     0
1   0    1    0    0     0
2   1    0    1    1     1
2   0    1    0    0     0
2   0    0    1    0     0
2   0    0    0    0     0
3   1    0    0    0     0
3   0    0    0    0     0
3   0    0    0    0     0
3   0    0    0    0     0
.   .    .    .    .     .
.   .    .    .    .     .
.   .    .    .    .     .
```

The macro call to generate tables and univariate descriptive statistics, fit the regression model, and generate regression diagnostics for these data is as follows.

```
%mcstrat(data=lowwgt,setid=set,case=lbw,indvar=smoke ui ptd lwd,
       uni=yes,diag=yes,tables=smoke ui ptd lwd)
```

Table 1 contains matched set summary information automatically produced by the macro, and Table 2 contains univariate statistics for the independent variables of interest, produced by the macro if the parameter UNI is set to YES. Notice that since each of these independent variables is a 0/1 variable, the mean values presented can be interpreted as proportions.

Table 3 contains the frequency table that summarizes the numbers of cases and controls in each matched set for which the variable UI was equal to 1. This table was created by including the variable UI in the TABLES parameter. Notice that data represented here are on a matched set level. The blank value in the lower right hand cell indicates there were no matched sets for which one of the cases and two of the controls had UI=1.

Table 4 contains the SAS output created in the macro using the PHREG procedure, including regression coefficients, standard errors, and corresponding p-values. Also printed are the odds ratios (labeled risk ratios by SAS) and 95% confidence intervals (defined by the values labeled Lower and Upper by SAS). Notice that each variable is potentially a risk factor for having a low birth weight baby, as all parameter estimates are positive. However, the only variable significant at the $\alpha = .05$ level is the presence of a previous pre-term delivery.

Tables 5 and 6 display the contents of the diagnostic data sets SUBDIAG and SETDIAG, respectively. The labels attached to each of the diagnostics are created automatically by the macro.

Table 1

MCSTRAT: LINEAR LOGISTIC REGRESSION ANALYSIS FOR MATCHED SETS

SETID = SET

CASE/CONTROL INDICATOR = LBW

OF OBSERVATIONS READ = 116

OF OBSERVATIONS USED = 116

OF MATCHED SETS READ = 29

OF MATCHED SETS USED = 29

SUMMARY OF MATCHED SETS ANALYZED

# CASES	# CONTROLS	# MATCHED SETS
1	3	29
29	87	29

Table 2

Univariate Statistics for Matched Sets Used

LBW	N Obs	Variable	N	Mean	Std Dev
Control	87	SMOKE	87	0.3448276	0.4780675
		UI	87	0.1494253	0.3585739
		PTD	87	0.0804598	0.2735805
		LWD	87	0.2183908	0.4155492
Case	29	SMOKE	29	0.5862069	0.5012300
		UI	29	0.3448276	0.4837253
		PTD	29	0.3793103	0.4938040
		LWD	29	0.4137931	0.5012300

LBW	N Obs	Variable	Minimum	Maximum
Control	87	SMOKE	0	1.0000000
		UI	0	1.0000000
		PTD	0	1.0000000
		LWD	0	1.0000000
Case	29	SMOKE	0	1.0000000
		UI	0	1.0000000
		PTD	0	1.0000000
		LWD	0	1.0000000

Table 3

Cases vs. # Controls Per Matched Set Where UI=1

		# Controls		
		0	1	2
		Frequency Count	Frequency Count	Frequency Count
Case-Control Ratio	# Cases			
1 : 3	0	11	6	2
	1	7	3	

Table 4

Results of Modeling
The PHREG Procedure

Data Set: WORK.__DAT
Dependent Variable: __TIME
Censoring Variable: LBW
Censoring Value(s): 0
Ties Handling: DISCRETE

Iteration History

Iter	Step	Log Likelihood	SMOKE	UI
0	INITIAL	-40.202536472	0	0
1	NEWTON	-32.102060263	0.491228	0.488163
2	NEWTON	-32.051407729	0.554653	0.498640
3	NEWTON	-32.051370228	0.554248	0.500048
4	NEWTON	-32.051370228	0.554248	0.500050

Iteration History

Iter	PTD	LWD
0	0	0
1	1.668123	0.414989
2	1.521927	0.524032
3	1.525913	0.521318
4	1.525917	0.521315

Last Evaluation of the Gradient

SMOKE	UI	PTD	LWD
3.629097E-12	1.064905E-11	1.196765E-11	-8.4182E-12

Table 4 (cont.)

Results of Modeling
The PHREG Procedure

Testing Global Null Hypothesis: BETA=0

Criterion	Without Covariates	With Covariates	Model Chi-Square
-2 LOG L	80.405	64.103	16.302 with 4 DF (p=.00264)
Score			17.260 with 4 DF (p=.00172)
Wald			12.639 with 4 DF (p=.01318)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
SMOKE	1	0.554248	0.48123	1.32650	.24943
UI	1	0.500050	0.54080	0.85499	.35514
PTD	1	1.525917	0.63518	5.77130	.01629
LWD	1	0.521315	0.51520	1.02389	.31160

Variable	Risk Ratio	Lower	Upper
SMOKE	1.741	0.678	4.470
UI	1.649	0.571	4.759
PTD	4.599	1.324	15.972
LWD	1.684	0.614	4.623

Table 5: Contents of Data Set SUBDIAG

CONTENTS PROCEDURE

Data Set Name:	WORK.SUBDIAG	Observations:	116
Member Type:	DATA	Variables:	21
Engine:	V612	Indexes:	0
Created:	10:16 Mon, Dec 6, 99	Observation Length:	168
Last Modified:	10:16 Mon, Dec 6, 99	Deleted Observations:	0
Protection:		Compressed:	NO
Data Set Type:		Sorted:	YES
Label:			

-----Alphabetic List of Variables and Attributes-----

#	Variable	Type	Len	Pos	Label
12	DELTAX2	Num	8	88	delta chi-square
20	DLWD	Num	8	152	delta beta for variable LWD
18	DPTD	Num	8	136	delta beta for variable PTD
14	DSMOKE	Num	8	104	delta beta for variable SMOKE
16	DUI	Num	8	120	delta beta for variable UI
11	HAT	Num	8	80	leverage value from hat matrix
13	INFL	Num	8	96	overall influence statistic
2	LBW	Num	8	8	
8	LD	Num	8	56	likelihood displacement
9	LMAX	Num	8	64	LMAX global influence statistic
6	LWD	Num	8	40	
5	PTD	Num	8	32	
1	SET	Num	8	0	
21	SLWD	Num	8	160	scaled delta beta for variable LWD
3	SMOKE	Num	8	16	
19	SPTD	Num	8	144	scaled delta beta for variable PTD
15	SSMOKE	Num	8	112	scaled delta beta for variable SMOKE
17	SUI	Num	8	128	scaled delta beta for variable UI
4	UI	Num	8	24	
7	XBETA	Num	8	48	Linear Predictor
10	XI	Num	8	72	fitted values

Table 6: Contents of Data Set SETDIAG

CONTENTS PROCEDURE

Data Set Name:	WORK.SETDIAG	Observations:	29
Member Type:	DATA	Variables:	10
Engine:	V612	Indexes:	0
Created:	10:16 Mon, Dec 6, 99	Observation Length:	80
Last Modified:	10:16 Mon, Dec 6, 99	Deleted Observations:	0
Protection:		Compressed:	NO
Data Set Type:		Sorted:	NO
Label:			

-----Alphabetic List of Variables and Attributes-----

#	Variable	Type	Len	Pos	Label
5	DELTAX2	Num	8	32	sum of delta chi-square
10	DLWD	Num	8	96	sum of delta beta for LWD
9	DPTD	Num	8	80	sum of delta beta for PTD
7	DSMOKE	Num	8	48	sum of delta beta for SMOKE
8	DUI	Num	8	64	sum of delta beta for UI
4	HAT	Num	8	24	sum of leverage values
6	INFL	Num	8	40	sum of overall influence statistic
2	LD	Num	8	8	sum of likelihood displacement
3	LMAX	Num	8	16	sum of LMAX values
1	SET	Num	8	0	

Examination of diagnostics is relatively simple when using the diagnostic data sets created by the macro. All figures presented here were produced with PROC GPLOT programming statements. Figure 1 presents a plot of the DELTAX2 statistic with the model's fitted values XI using the data set SUBDIAG. Notice the plot seems to be composed of two separate lines: one seemingly exponential and extending from the upper left to the lower right of the graph and the other less noticable and extending from the extreme lower left to the right and slightly upward. The first such line contains the cases and the second, less extreme line, contains the controls. Large values of the DELTAX2 diagnostic represent individuals who disproportionately affect model fit. Since the fitted values can be interpreted as the estimated probability an individual is a case, it makes sense that the cases with small fitted values (those observations in the upper left part of the plot) and the controls with large fitted values (those in the upper right part of the plot) are the most influential. The individual with the largest values of ΔX_i^2 in this plot is a case who has none of the risk factors of interest, and is matched to a control with three of the four risk factors of interest (smoking, presence of a previous pre-term delivery, and low maternal weight). The fact that the line containing the cases is much more extreme than the one containing the controls is explained by the derivation of the ΔX_i^2 statistic. Upon examination this diagnostic (section 3.3), it can be seen that the value for a poorly fit case is approximately $1/\hat{\xi}$ (when N=1 and $\hat{\xi}$ approaches 0). Similarly, the value of ΔX_i^2 for a poorly fit control approaches $\hat{\xi}$ (when N=0 and $\hat{\xi}$ approaches 1). Thus, cases can potentially have values of ΔX_i^2 much greater than 1, as opposed to controls. Because of this apparent difference in scale, it may be helpful to plot the controls separately, as in Figure 2.

Figure 3 presents a plot of the global influence statistic (INFL) vs. the PHREG likelihood displacement statistic (LD), again using the SUBDIAG data set. Notice the very high correlation between these two diagnostics (in fact, $r > .99$). Figure 4 plots the INFL statistic with the PHREG LMAX statistic. Again, there is a positive correlation between the two diagnostics, although not nearly as striking as seen in Figure 3. Further examination revealed relatively strong positive associations between all global diagnostics produced by the macro. The similarities between the PHREG diagnostics and the more established influence statistics seem to indicate that the

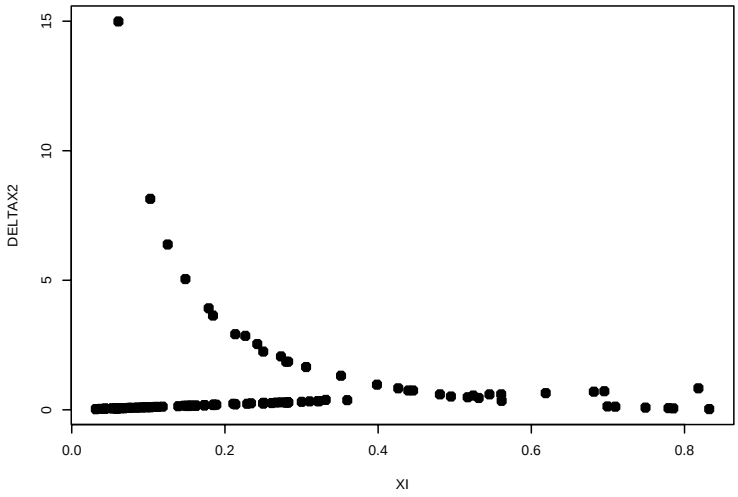

Figure 1: plot of DELTAX2 by fitted values, cases and controls

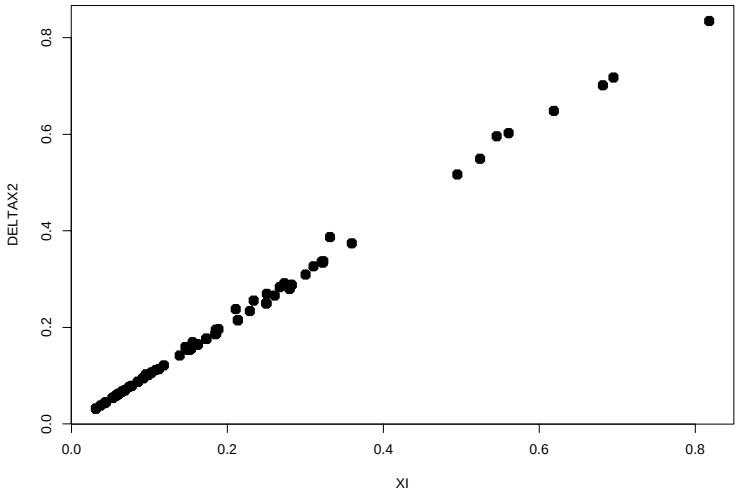

Figure 2: plot of DELTAX2 by fitted values, controls only

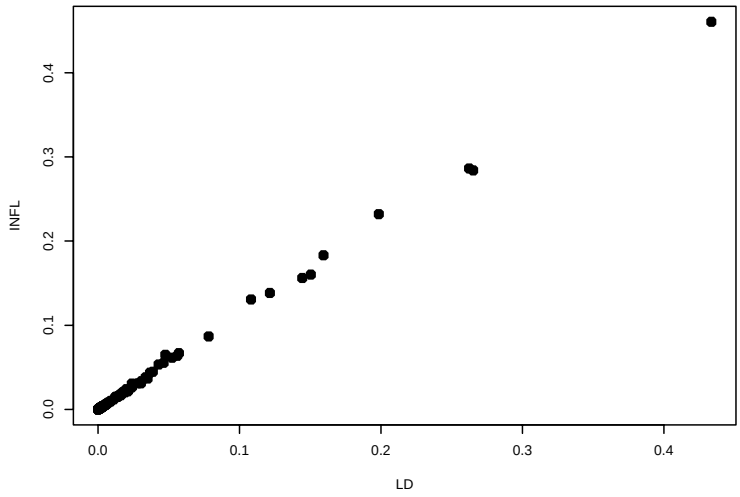

Figure 3: plot of INFL statistic with likelihood displacement

PHREG diagnostics are valid tools for examining model fit in a conditional logistic model, at least for this particular example.

Figure 5 plots the individual DFBETA statistic for the variable PTD vs. the set ID variable using the SUBDIAG data set. The dark circles represent cases and the light circles represent controls. Four individuals seem to have relatively large negative values for this diagnostic, and all are cases who did not previously have a pre-term delivery.

Finally, Figure 6 presents a plot of the likelihood displacement statistic vs. the set ID variable using the data set SETDIAG. Three of the matched sets have relatively large values for this diagnostic. The most extreme value (LD=0.45) is the matched set that contains the same woman found earlier to have an extremely large ΔX_i^2 value. Deletion of this matched set from the data increases the odds ratios for smoking status, presence of a previous pre-term delivery, and low maternal weight by 17%, 42%, and 27%, respectively.

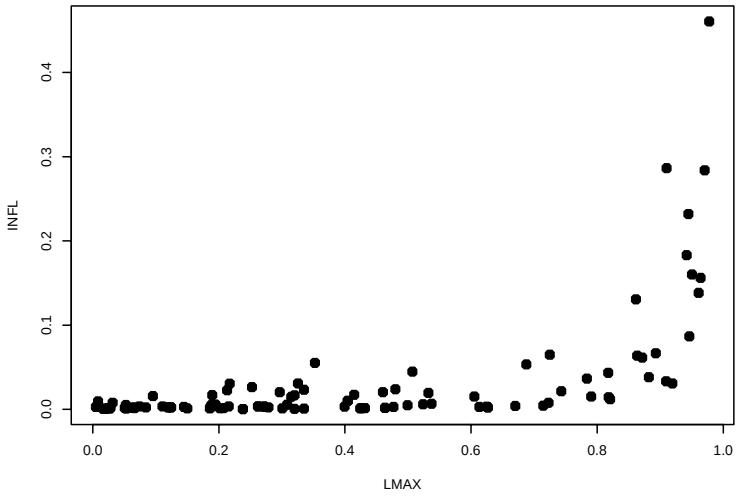

Figure 4: plot of INFL statistic with LMAX statistic

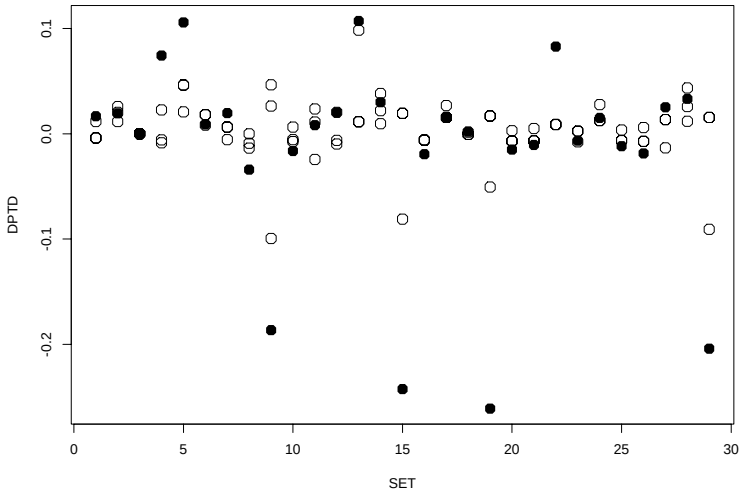

Figure 5: plot of DFBETA statistic for PTD with SET ID variable

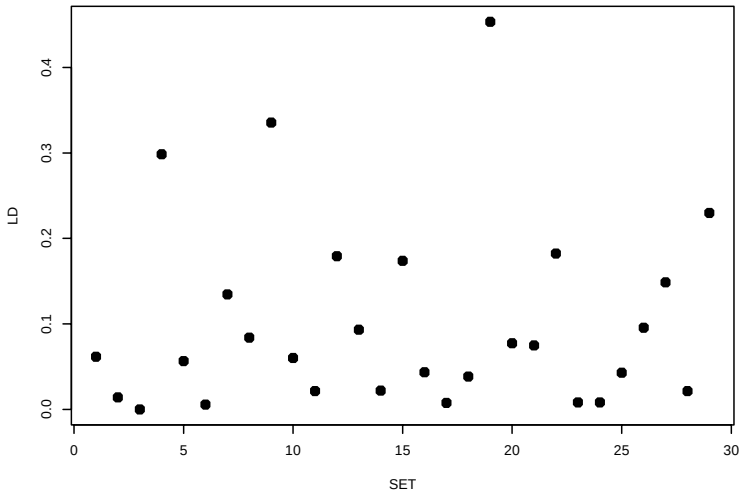

Figure 6: plot of likelihood displacement with SET ID, SETDIAG data set

6 Conclusion

Analyzing data from a matched case-control study requires specialized approaches not readily accessible using PROC LOGISTIC. The SAS macro MCSTRAT provides an easy and effective way to describe the data, fit a model, and calculate regression diagnostics for matched or finely stratified data.

References

[1] SAS Institute, Inc. *SAS/STAT Software: Changes and Enhancements through Release 6.12*. Cary, NC: SAS Institute, Inc., 1997.

[2] Naessens J.M., Offord K.P., Scott W.F., Daood S.L. *PROC MCSTRAT*. Technical Report Series Number 28, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, 1984.

[3] Cox D.R. and Hinkley D.V. *Theoretical Statistics*. New York: John Wiley & Sons, Inc., 1974.

[4] Breslow N.E. and Day N.E. *Statistical Methods in Cancer Research: Volume 1–The Analysis of Case-Control Studies*. Lyon: International Agency for Research on Cancer, 1980.

[5] Hosmer D.W. and Lemeshow S. *Applied Logistic Regression*. New York: John Wiley & Sons, Inc., 1989.

[6] Neter J., Wasserman W., and Kutner M. *Applied Linear Regression Models, Second Edition*. Boston: Irwin, Inc., 1989.

[7] Pregibon D. Logistic Regression Diagnostics. *Annals of Statistics* 1981; 9: 705-24.

[8] Pregibon D. Data Analytic Methods for Matched Case-Control Studies. *Biometrics* 1984; 40: 639-51.

[9] Moolgavkar S.H., Lustbader E.D., and Venzon D.J. Assessing the Adequacy of the Logistic Model for Matched Case-Control Studies. *Statistics in Medicine* 1985; 4: 425-35.

[10] Cain K.C. and Lange N.T. Approximate Case Influence for the Proportional Hazards Regression Model with Censored Data. *Biometrics* 1984; 40: 493-99.

[11] Pettitt A.N. and bin Daud I. Case-Weighted Measures of Influence for Proportional Hazards Regression. *Applied Statistics* 1989; 38: 313-29.