

# A SAS<sup>®</sup> Macro to Analyze Data From a Matched or Finely Stratified Case-Control Design

Robert A. Vierkant, Terry M. Therneau, Jon L. Kossanke, James M. Naessens  
Mayo Clinic, Rochester, MN

## ABSTRACT

A matched case-control design is a common approach used to assess disease-exposure relationships, and is often a more efficient method than an unmatched design. However, for the valid analysis of such an approach, a modeling technique that incorporates the matched nature of the data is needed. This prohibits the use of a standard unconditional logistic regression analysis generally available in PROC LOGISTIC. A stratified conditional logistic model has the same flexibility as an unconditional model, yet can still take into account the correlation structure attributable to matching. This paper presents a SAS macro that fits a conditional logistic regression model to matched or finely stratified data using the PHREG procedure. 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. This paper is intended for an audience with a working knowledge of statistical modeling.

## INTRODUCTION

The logistic regression model  $\Pr(\text{outcome}) = \exp(\beta'X) / (1 + \exp(\beta'X))$  is a common framework for the analysis of data with a binary outcome. For a discrete covariate,  $\exp(\beta)$  represents the odds ratio for presence vs. absence of a certain characteristic. In 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. This type of design is known as a matched case-control study and is most often analyzed using stratified conditional logistic regression models that take into account the matched nature of the data. The SAS procedure LOGISTIC is often used to analyze data arising from a case-control study, but cannot take into account the correlation structure of a matched or finely stratified design. This paper presents a SAS macro that uses the PHREG procedure to fit a conditional logistic regression model to matched data. The macro produces tables and statistics summarizing each matched set. It also calculates regression diagnostics using SAS/IML software and output data sets created by the PHREG procedure. It expands upon work previously reported by Naessens et al (1983).

## CONDITIONAL LOGISTIC REGRESSION ANALYSIS

Special conditional likelihood techniques have been developed to estimate parameters in a matched study when the more general techniques are inadequate. Briefly, in a matched case-control study with  $K$  strata or matched sets, the conditional likelihood for the  $k$ th set is the probability of the observed data conditional on the subjects in the set. The full conditional likelihood is then the product of the individual likelihoods across all  $K$  strata. More detail can be found in Cox and Hinkley (1974).

Although PROC LOGISTIC cannot fit a conditional logistic regression model, there are several other methods of performing such an analysis in SAS. One method makes use of an identity between the form of the matched case-control log-likelihood function and the partial likelihood for a Cox model when using the 'discrete' method that corrects for ties. The following SAS code fits a conditional logistic regression model to matched case-control data.

```
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.

## MODEL FIT AND REGRESSION DIAGNOSTICS

After the model building stage, it is generally a good idea to determine how effective the final logistic model is in describing the dependent variable. This is referred to as goodness-of-fit or model fit. Model fit can be assessed both on a summary level, and on either an individual subject or matched set level. Summary measures of goodness-of-fit for logistic models include the Pearson chi-square and deviance statistics, and the Hosmer-Lemeshow goodness of fit test (Hosmer and Lemeshow, 1989). These measures are valuable tools in giving an overall indication of fit, but may not be specific about individual model components. Thus, it is a good idea to couple the summary statistics with an evaluation of model fit over each set of observed covariate combinations. Measures used to investigate these individual components

are called regression diagnostics. An introduction to linear regression diagnostics can be found in Neter, Wasserman, and Kutner (1989). Many of these techniques were first applied to logistic regression analyses by Pregibon (1981), and extended to conditional logistic regression analyses by Pregibon (1984) and Moolgavkar, Lustbader, and Venzon (1985). Some of these diagnostics are available in PROC PHREG by issuing an output statement within the procedure. They include the influence statistics derived by Cain and Lange (1984), which approximate changes in the individual parameter estimates due to deletion of a subject ( $\Delta\beta_i$ , or alternatively DFBETA). Here the subscript  $i$  refers to a specific independent variable in the regression model. Scaled influence statistics can also be calculated by dividing the individual  $\Delta\beta_i$  values by the parameter estimate's standard error. Diagnostics available in PHREG that assess global measures of influence include the likelihood displacement statistic (LD), which approximates the amount by which minus twice the log likelihood changes due to deletion of a subject, and the LMAX statistic, derived as the Cox proportional hazards equivalent to Cook's distance (Pettitt and bin Daud, 1989). The diagnostics  $\Delta\beta_i$ , LD, and LMAX are all related in that each is a function of the model's weighted score residuals and estimated covariance matrix. In fact, the value of LD is by definition virtually identical to the sum of the squared  $\Delta\beta_i$ . Other useful diagnostics are presented in Hosmer and Lemeshow (1989) and include the "leverage" values ( $h$ ) obtained from the hat matrix as derived by Pregibon (1981), the decrease in the value of the Pearson chi-square statistic due to deletion of a subject ( $\Delta X^2$ ), and the influence statistic assessing the overall change in covariate estimates due to deletion of a subject (INFL). This overall influence statistic approximates the global effect of deleting a subject by incorporating information from each of the

individual  $\Delta\beta_i$ . Not surprisingly, its values correlate very highly with the global diagnostic LD mentioned earlier. Many of the diagnostics mentioned above are often plotted against the model's fitted values ( $\xi$ ), which are estimates of conditional probability assuming the logistic regression model is correct, and against other external influences that may affect model fit, such as time.

### SAS MACRO STRAT

We have developed a SAS macro called STRAT (program available from contact author) that contains code to fit a conditional logistic regression model and generate the regression diagnostics  $\Delta\beta_i$ , scaled  $\Delta\beta_i$ , LD, LMAX,  $\Delta X^2$ , h, and INFL as well as the fitted values  $\xi$  for matched or finely stratified case-control data. It expands on work previously reported by Naessens et al (1983). 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 some of the regression diagnostics. Next, IML code is used to generate the remaining diagnostics and model's fitted values as presented in Hosmer and Lemeshow (1989). Keyword parameters are required to specify the data set (DATA), the stratification or set ID variable (SETID), the variable that distinguishes cases from controls (CASE), and the independent variables of interest (INDVAR). The values for the CASE variable must be 1 for cases and 0 for controls. All independent variable names should be included in the INDVAR parameter, separated by blank spaces. Optional keyword parameters specify whether the covariance matrix be printed (COV—default is no), the maximum number of iterations to be performed when fitting the model (MAXITER—default is 10), the difference in log likelihood used to determine model convergence

(EPSILON—default is .000001), whether a list of matched sets not included in the model be printed (ID—default is no), whether univariate statistics be printed for independent variables (UNI—default is yes), and whether output data sets containing the regression diagnostics mentioned earlier be created (DIAG—default is yes). Another optional keyword parameter specifies a list of independent variables for which frequency tables will be created (TABLES). The tables created show how many cases and controls had a value of 1 in each set, so they are best used with 0/1 indicator variables. If the user requests diagnostics output by specifying DIAG=yes, the program creates two data sets of regression diagnostics. The first data set, SUBDIAG, contains the original independent variables and set ID variable as well as diagnostic information on a subject level, including the values of  $\Delta X^2$ , INFL, h, LD, LMAX, and the fitted values  $\xi$ . Also included are each of the individual influence statistics  $\Delta\beta_i$  and the scaled influence statistics. The second data set, SETDIAG, contains information on a matched set level, including the set ID variable, the sums of the values  $\Delta X^2$ , INFL, LD, LMAX, and h over all observations in the set, and the sums of the squared values of the individual influence statistics and the scaled influence statistics.

### EXAMPLE

The example presented here uses the low birth weight data found in Appendix 4 of Hosmer and Lemeshow (1989). In this example, mothers who gave birth to a low birth weight baby (cases) were matched to three mothers of normal birth weight babies of the same age (controls). Twenty-nine strata, each containing 1 case and 3 controls, were created. Variables included in the final model were smoking status (SMOKE), uterine irritability (UI), presence or absence of a previous pre-term delivery (PTD), and maternal weight at the last

menstrual period, dichotomized as the lower 25<sup>th</sup> percent vs. the upper 75<sup>th</sup> percent (LWD). All of the independent variables are dichotomous, taking on values of 1 when the condition is present and 0 when the condition is absent. Note that the STRAT macro can fit a model using continuous variables, but dichotomous variables are used here to allow direct comparison with the analysis presented in Hosmer and Lemeshow (1989). 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 (CASE, coded as 1 for cases and 0 for controls). The macro call to generate tables and univariate statistics describing the data, fit the regression model, and generate regression diagnostics for these data is as follows:

```
%strat(data=lowwgt, setid=set, case=case,
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 here can also 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. Notice that data represented here are on a matched set level. This table is created by including the variable UI in the TABLES parameter.

TABLE 1  
 STRAT: LINEAR LOGISTIC REGRESSION  
 ANALYSIS FOR MATCHED SETS  
 =====  
 SETID = SET  
 CASE/CONTROL INDICATOR = CASE  
  
 # 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

OBS	Var	N	Mean	Std Dev	Min
Control	SMOKE	87	0.345	0.478	0
	UI	87	0.149	0.359	0
	PTD	87	0.080	0.274	0
	LWD	87	0.218	0.416	0
Case	SMOKE	29	0.586	0.501	0
	UI	29	0.345	0.484	0
	PTD	29	0.379	0.494	0
	LWD	29	0.414	0.501	0

OBS	Variable	Max
Control	SMOKE	1.0
	UI	1.0
	PTD	1.0
	LWD	1.0
Case	SMOKE	1.0
	UI	1.0
	PTD	1.0
	LWD	1.0

Table 4 contains the SAS output created in the macro using the PHREG procedure, including regression coefficients, standard errors, and corresponding P-values.

TABLE 3  
# Cases vs. # Controls Where UI=1

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

Corresponding graphs can be found in Figures 1 and 2. In each graph, one observation (a case) is found to be both influential and an outlier (the observation positioned in the upper left-hand part of each graph). This case has none of the risk factors included in the final model, and she belongs to a stratum that has a control with three of the risk factors (smoking, uterine irritability, and low maternal weight).

### CONCLUSION

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

TABLE 4: Results of Modeling  
Maximum Likelihood Estimates

Var	DF	Parameter	Std	Wald $\chi^2$	P-value
		Estimate	Error		
SMOKE	1	0.554	0.481	1.326	.2494
UI	1	0.500	0.540	0.854	.3551
PTD	1	1.525	0.635	5.771	.0162
LWD	1	0.521	0.515	1.023	.3116

Conditional Risk Ratio and  
95% Confidence Limits

Var	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

Notice that each variable is potentially a risk factor for low birth weight, as all parameter estimates are positive. However, the only variable significant at the  $\alpha=.05$  level was occurrence of a previous pre-term delivery.

The following SAS code produces scatterplots of  $\Delta X^2$  and INFL by the fitted values  $\hat{\xi}$  using the subject-specific diagnostic data set SUBDIAG:

```
proc gplot data=subdiag;
  plot deltax2*xi;
  title 'Figure 1';
proc gplot data=subdiag;
  plot infl*xi;
  title 'Figure 2';
```

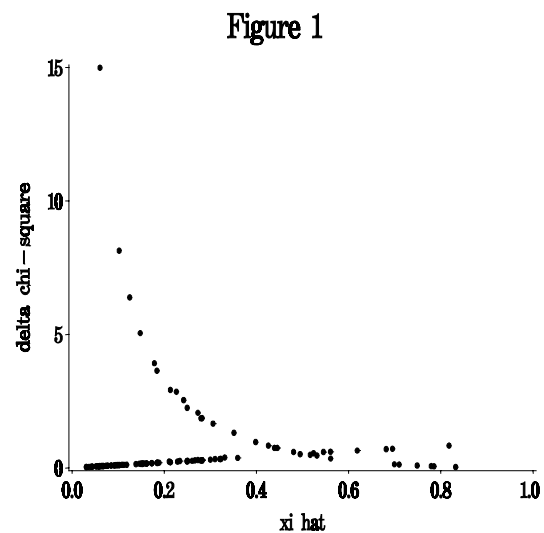
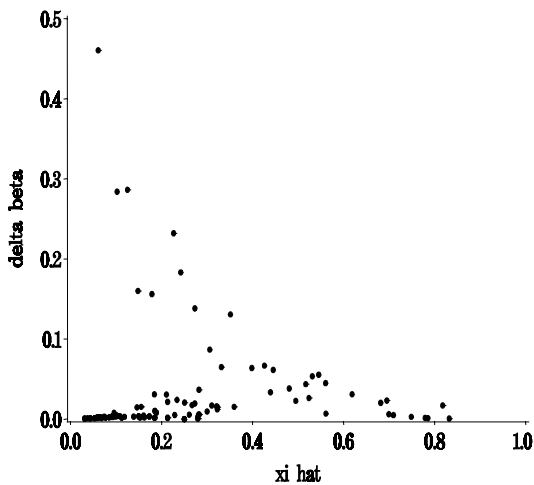


Figure 2



**REFERENCES**

Cain K.C. and Lange N.T. (1984). *Approximate Case Influence for the Proportional Hazards Regression Model with Censored Data*. Biometrics 40, 493-499.

Cox D.R. and Hinkley D.V. (1974). *Theoretical Statistics*. Chapman Hall, London.

Hosmer D.W. and Lemeshow S. (1989). *Applied Logistic Regression*. John Wiley and Sons, Inc., New York, NY.

Moolgavkar SH, Lustbader ED, and Venzon (1985). *Assessing the Adequacy of the Logistic Model for Matched Case-Control Studies*. Statistics in Medicine 4, 425-435.

Naessens JM, Offord KP, Scott WF, and Daoud SL (1983). *A Computer Program for the Analysis of Case-Control Studies*. In: SAS Institute, Inc., Proceedings of the 8<sup>th</sup> Annual SAS User's Group International Conference, 611-618.

Neter J., Wasserman W., and Kutner M. (1989). *Applied Linear Regression Models, Second Edition*. Irwin, Inc., Boston, MA.

Pregibon D (1981). *Logistic Regression Diagnostics*. Annals of Statistics 9, 705-724.

Pettitt A.N. and bin Daud I. (1989). *Case-Weighted Measures of Influence for Proportional Hazards Regression*. Applied Statistics 38, 313-329.

Pregibon D (1984). *Data Analytic Methods for Matched Case-Control Studies*. Biometrics 40, 639-651.

SAS Institute, Inc. (1992). SAS Technical Report P-229, *SAS/STAT Software: Changes and Enhancements, Release 6.07*. Cary, NC: SAS Institute, Inc.

SAS and SAS/IML are registered trademarks of SAS Institute Inc., in the USA and other countries. ® indicates USA registration.

**CONTACT INFORMATION**

Robert A. Vierkant, MAS  
 Harwick 7, Biostatistics  
 Mayo Clinic  
 200 First Street SW  
 Rochester, MN 55905  
 (507) 284-8993  
[vierkant.robert@mayo.edu](mailto:vierkant.robert@mayo.edu)