J Chron Dis Vol. 40, No. 7, pp. 697–704, 1987
Printed in Great Britain

0021-9681/87 $3.00 + 0.00
Pergamon Journals Ltd

LARGE SAMPLE CONFIDENCE INTERVALS FOR REGRESSION STANDARDIZED RISKS, RISK RATIOS, AND RISK DIFFERENCES

W. DANA FLANDERS and PHILIP H. RHODES

U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control, Center for Environmental Health, Division of Chronic Disease Control, Agent Orange Projects, Atlanta, GA 30333, U.S.A.

(*Received in revised form* 11 *August* 1986)

Abstract—Several methods have been proposed for standardizing risks, risk ratios, and risk differences based on the results of logistic regression. These methods provide an alternative to direct standardization, a particularly useful approach when there are many covariates. In this paper, methods for calculating approximate confidence limits for these standardized measures are presented. A simple example, in which published data are used, illustrates the techniques and allows comparison with confidence limits calculated from the directly standardized risk ratio.

Standardization Logistic analyses Adjusted disease risks Confidence intervals Epidemiologic methods

INTRODUCTION

Standardization is a frequently used epidemiologic technique that permits comparison of disease risks in an index group with those in a comparison group after adjustment for confounding [1–3]. A standardized risk can be viewed as a weighted average of category-specific risks, a standardized risk ratio, as the ratio of two such risks, and a standardized risk difference as the difference between two such risks. When direct standardization is used, the standardized risk (R_i) for the index group ($i = 1$) or the comparison group ($i = 0$) can be written as follows:

$$R_i = \sum_j w_j r_{i,j} \quad (1)$$

where the subscript j indicates stratum of the confounders, w_j is the weight for the jth stratum, and $r_{1,j}$ and $r_{0,j}$ are the risks in the index and comparison groups for the jth stratum. These techniques have been extended so that they may be applied when logistic or other multivariate methods have been used to analyze the data, an application particularly useful when there are many strata or potential confounders [2, 4–6]. Wilcosky and Chambless recently considered three such techniques and illustrated their use by comparing directly standardized risks with those standardized by using results of logistic regression [2]. These and related methods have also been discussed by Greenland [7, 8] and others [6, 9]. The purpose of this paper is threefold. First, equations for approximate confidence interval estimation, applicable to each of the three techniques discussed by Wilcosky and Chambless, are presented. Second, a simple program that can be used to calculate the regression standardized risks, risk ratios, and risk differences with associated confidence limits is presented. Third, published data are analyzed to illustrate the methods of interval estimation and to provide a comparison of confidence intervals calculated for directly standardized risks with those calculated for risks standardized using results of logistic regression.

NOTATION, ASSUMPTIONS, AND OVERVIEW

We assume that interest is in comparing the disease risk in an index with that in a comparison population, after adjustment for con-

founding. The logistic model for the simple case, assuming M confounders or covariates $(E_2, \ldots, E_{M+1})$, is given by:

$$\log(p/1-p) = \beta_0 + \beta_1 E_1 + \beta_2 E_2 + \cdots + \beta_{M+1} E_{M+1} \quad (2)$$

where p is the probability of disease and E_1 is an indicator variable that is 1 for subjects in the index group and 0 otherwise. Depending on study design, p may be interpretable as a prevalence, a risk, or some other measure of disease frequency, but it is referred to here as a disease "risk". An estimate of risk for specified values of $E_1, E_2, \ldots, E_{M+1}$ is given by:

$$\hat{R} = 1/(1 + \exp(-(b_0 + b_1 E_1 + b_2 E_2 + \cdots + b_{M+1} E_{M+1}))), \quad (3)$$

where b_m $(m = 0, \ldots, M+1)$ are parameter estimates. This formulation implies, of course, that risks, the risk difference, and the risk ratio comparing the index with the comparison group will differ for different values of the covariates. To present summary estimates based on logistic regression results, three methods have recently been proposed [2]. The methods are similar to direct standardization in that each approach yields a summary measure that has been averaged over different combinations of the covariates with each combination weighted according to a specified standard or set of weights. These standardization techniques for logistic regression have been termed "conditional" prediction, "marginal" prediction, and "stratified" prediction [2]. These methods, their interrelationships, applications, and limitations have recently been discussed [2].

To apply the conditional approach, the investigator chooses a "standard" reference value for each of the covariates, say $E_2^*, \ldots, E_{M+1}^*$. Summary measures are then given by:

$$\hat{R}_{i,c} = 1/(1 + \exp(-(b_0 + b_1 E_1 + b_2 E_2^* + \cdots + b_{M+1} E_{M+1}^*))), \quad i = 0, 1 \quad (4)$$

where E_1 is 1 for the index group and 0 otherwise.

To apply the "stratified" method, the investigator chooses standard weights for each stratum or combination of covariates. Summary estimates are then given by:

$$\hat{R}_{i,s} = \sum_j w_j/(1 + \exp(-(b_0 + b_1 E_{1,j} + b_2 E_{2,j} + \cdots + b_{M+1} E_{M+1,j}))) \quad (i = 0, 1) \quad (5)$$

where the summation is over the J strata, $j = 1, \ldots, J$; $w_j, E_{2,j}, \ldots, E_{M+1,j}$ are the weights and covariates, respectively, for stratum j. Possible choices of weights for the stratified method are the relative size of strata in either the comparison group or the entire study population.

To apply the marginal method, the investigator selects a standard population for which the distribution of the covariates is known or specified. The study population itself is the usual choice. Summary measures are then given by equation (6), which is similar to equation (5), except that the subscript k refers to individuals in the standard population, and $w_k = 1/N$, where N is the number of individuals.

$$\hat{R}_{i,m} = \sum_k w_k/(1 + \exp(-(b_0 + b_1 i + b_2 E_{2,k} + \cdots + b_{M+1} E_{M+1,k}))) \quad (6)$$

Interaction or product terms may be included, as appropriate, in the model. To evaluate the standardized risks using equations (4–6), values for such interaction terms must, of course, be consistent with the modeled interactions. In particular, if $E_{l,j} = E_l^* E_{m,j}$ for some l and m in equation (5) or (6), $E_{l,j}$ must equal $E_{m,j}$ when evaluating the risk in the index group and must equal 0 when evaluating the risk in the comparison group.

The standardized risk ratio, $\hat{R}_{2,h}$, for standardization method h ($h = c$, m or s) is given by $\hat{R}_{1,h}/\hat{R}_{0,h}$ and the standardized risk difference, $\hat{R}_{3,h}$, by $\hat{R}_{1,h} - \hat{R}_{0,h}$. This formulation highlights the similarity with direct standardization, since equations (5) and (6) are analogous to equation (1) with observed risks corresponding to model-predicted risks.

Consistent with common practice, we assume that parameters of the logistic model are estimated by the method of maximum likelihood. The asymptotic distribution of the parameter estimates is then approximately normal under reasonable assumptions, with the variances and covariances given, approximately, by the inverse of the observed information matrix. Commercially available logistic regression programs (such as "proc logist" in SAS [10, 11]) print this estimated covariance matrix as an option. The weights used in standardization are frequently treated as constants [2, 6, 12], a convention that we adopt. Alternatively, results may be viewed as conditional upon the chosen standard.

APPROXIMATE VARIANCE AND CONFIDENCE LIMITS

Estimates of the standardized risks, $R_{1,h}$, $R_{0,h}$, $R_{2,h}$ and $R_{3,h}$, are thus given by simple

functions of the parameter estimates and the set of (constant) weights. The asymptotic variance ($V_{i,h}$, $i = 0, 1, 2, 3$) of the risks ($R_{i,h}$, $i = 0, 1, 2$; $h = c, m, s$) is estimated by:

$$\hat{V}_{i,h} = \sum_{kl}\sum (\widehat{dR_{i,h}/d\beta_k})(\widehat{dR_{i,h}/d\beta_1})(v_{k,1}) \quad (7)$$

where $(\widehat{dR_{i,h}/d\beta_k})$ is the partial derivative of $R_{i,h}$ with respect to β_k, $k = 0, 1, \ldots, M+1$ evaluated at the estimated value for each parameter and $(v_{k,1})$ is the $(k, 1)$th term of the estimated covariance matrix (9). Formulae for these partial derivatives, obtained by differentiation of equations (4), (5) and (6), are given in Appendix 1. (In matrix notation, $\hat{V}_i = (\widehat{dR_i/db})(C)(\widehat{dR_i/db})'$, where C is the estimated covariance matrix and $(\widehat{dR_{i,h}/d\beta})$ is the $1 \times M+2$ matrix of estimated partial derivatives of $R_{i,h}$. This expression is simply a first-order Taylor series approximation, generalized to several variables (9).)

Under the assumptions stated previously, $\hat{R}_{0,h}$, $\hat{R}_{1,h}$, $\hat{R}_{2,h}$ and $\hat{R}_{3,h}$ are asymptotically normally distributed. If the disease risk is small, however, confidence limits for $\hat{R}_{0,h}$ and $\hat{R}_{1,h}$ may be based on the logarithm of the risk [13]. Because of the restricted range and skewed distribution of the risk ratio, confidence limits for $\hat{R}_{2,h}$ may also be based on a logarithmic transformation [12, 13], so that approximate $(1-\alpha)$ 100% confidence limits are given by;

$$R_{i,h}, \bar{R}_{i,h} = \hat{R}_{i,h} \exp(\pm z_{\alpha/2} \hat{V}_{i,h}^{1/2}/\hat{R}_{i,h}), \quad (i = 0, 1, 2;\ h = c, m, s), \quad (8)$$

where $z_{\alpha/2}$ is the value that cuts off area $\alpha/2$ in the upper tail of the standard normal distribution, and $\hat{V}_{i,h}$ is given by equation (7) [$\hat{V}_{i,h}^{1/2}/\hat{R}_{i,h}$ is the estimated standard error of $\log(\hat{R}_{i,h})$].

Approximate $(1-\alpha)$ 100% confidence limits for the standardized risk difference are given by:

$$R_{3,h}, \bar{R}_{3,h} = \hat{R}_{3,h} \pm (z_{\alpha/2} * \hat{V}_{3,h}^{1/2}) \quad (9)$$

A simple computer program can be run with "proc logist" in SAS [10, 11] to calculate regression standardized risks, risk ratios, risk differences, and associated confidence limits. The program listing and a sample of the resulting printout is given in Appendix 2.

Example

To illustrate calculation of confidence intervals, we analyzed data published by Wilcosky and Chambless [2], which relate prevalence of electrocardiographic abnormalities to replacement estrogen usage, obesity, and age. (Since the published data that we used in our analyses [2, Table 1] did not include the age of individuals, we assigned age 55, 65 and 80 to subjects in the 50–59, 60–69 and 70+ categories, respectively. This procedure presumably explains minor differences between our results and those published by Wilcosky and Chambless [2].) The logistic model used for these data was:

$$\log(p/1-p) = \beta_0 + \beta_1 \text{HORM} + \beta_2 \text{OBESITY} + \beta_3 \text{AGE} \quad (10)$$

where p is the prevalence of R-wave abnormalities, "HORM" indicates usage of hormones (1 = no, 0 = yes), "OBESITY" indicates obesity (1 = present, 0 = absent), and "AGE" is a variable for age (treated as continuous). Our estimates for the parameters β_0 to β_3 were -1.231, 0.554, -1.525 and -0.027, respectively, similar (except for sign) to results obtained by Wilcosky and Chambless, who used the exact age of each subject [2].

Table 1. Weights used to analyze data of Wilcosky and Chambless [2]

Stratum No.	Body mass	Age	Weight
1	obese	55	0.23
2	obese	65	0.13
3	obese	80	0.07
4	nonobese	55	0.33
5	nonobese	65	0.18
6	nonobese	80	0.07

Standardized estimates of prevalence according to hormone usage were calculated with the program in Appendix 2, treating obesity and age as confounders. For direct and for stratified standardization, the number of subjects (hormone users plus nonusers) in each stratum was used as the weight (Table 1). For marginal standardization, the study population itself was used as the standard. Estimates of the prevalence among users and among nonusers and the prevalence ratio, obtained with different methods of standardization, are similar to the corresponding estimates of Wilcosky and Chambless (Table 2). (Stratified standardization was equivalent to marginal standardization in this example, since "AGE" assumes only three values.)

The variance–covariance matrix, printed by SAS, is reproduced in Table 3. Only the calculation of $(dR_{1,m}/d\beta_2)$ is illustrated, because other calculations are similar.

$$dR_{1,m}/db_2 = \sum_j \frac{w_j E_{2,k} \exp(-(b_0 + b_1 + b_2 E_{2,k} + b_3 E_{3,k}))}{(1 + \exp(-(b_0 + b_1 + b_2 E_{2,k} + b_3 E_{3,k})))^2}$$

Table 2. Estimated standardized risks, ratios and confidence limits for data of Wilcosky and Chambless [2]

Method of adjustment	Prevalence		Ratio	Difference
	User	Nonusers		
Direct adjustment	0.036(0.024, 0.055)	0.020(0.010, 0.042)	1.8(0.7, 4.1)	
Regression adjustment				
Stratified	0.037(0.024, 0.056)	0.022(0.010, 0.048)	1.7(0.7, 4.2)	0.015(−0.008, 0.038)
Marginal	0.037(0.024, 0.056)	0.022(0.010, 0.048)	1.7(0.7, 4.2)	0.015(−0.008, 0.038)
Conditional	0.030(0.018, 0.049)	0.017(0.007, 0.041)	1.7(0.7, 4.3)	0.012(−0.007, 0.031)

Table 3. Estimated covariance matrix, output of SAS program

	β_0	β_1	β_2	β_3
β_0:	2551	−0.117	−0.032	−0.042
β_1:		0.225	0.020	0.001
β_2:			0.301	−0.0003
β_3:				0.0007

Table 4. Partial derivatives of $R^*_{i,m}$ data of Wilcosky and Chambless [2]

	$k=0$	$k=1$	$k=2$	$k=3$
$dR_{1,m}/d\beta_k$	0.021	0.021	0.003	1.257
$dR_{0,m}/d\beta_k$	0.035	0	0.005	2.096

*Evaluated at parameter estimates.

$$= \frac{0.23*118}{(1+118)^2} + \frac{0.13*155}{(1+155)^2} + \frac{0.07*230}{(1+230)^2} + \frac{0.33*26}{(1+33)^2} + \frac{0.18*34}{(1+34)^2} + \frac{0.07*50}{(1+50)^2}$$

$$= 0.021. \quad (11)$$

Results for other combinations of i and j are given in Table 4.

Substitution into equation (7) gives; $\widehat{\text{Var}}(\hat{R}_{1,m}) = 0.000072$, $\widehat{\text{Var}}(\hat{R}_{0,m}) = 0.000063$, and $\widehat{\text{Var}}(\hat{R}_{2,m} = 0.072$. Approximate confidence intervals calculated from equation (8) are summarized in Table 2. The confidence limits for the directly standardized risk, calculated as suggested by Flanders [12], are close to those obtained for regression-standardized risks. Because of the relatively low prevalence of electrocardiographic abnormalities, these confidence limits are only approximate.

DISCUSSION

These formulas, or alternatively the program in Appendix 2, provide a straightforward method for calculating large sample confidence intervals for risks, risk ratios, and risk differences that have been standardized by using results of logistic regression. Such an approach should be especially useful when adjustment must be made for many confounders simultaneously. Even though hypothesis testing is probably better done with test statistics more directly related to the odds ratios, presentation of confidence intervals is important because it allows for random variation in an easily interpretable way [14].

An alternative to the standardization procedure described in this paper and in that of Wilcosky and Chambless [2], would be to assume uniformity of the risk ratio or the risk difference, permitting direct estimation of the measure using maximum likelihood techniques [7, 8, 15]. This approach would be particularly useful if the measure were uniform across strata. If the measure of interest is heterogeneous, however, Greenland has argued that "the only meaningful summary estimates will be standardized estimates' [7, 8]. In that case, a standardized procedure based on "smoothed" estimates from a logistic or other appropriate model would permit appropriate point and confidence interval estimation. This paper provides the equations and a computer program to implement that approach when a logistic model is used.

In the example presented here, disease is rare (prevalence < 0.07), so that uniformity of the odds ratio, implied by the logistic model that we used, is reasonably consistent with uniformity of the risk ratio. In this situation, the directly estimated risk ratio [15, 16] should be approximately equal to standardized estimate that is based on results of logistic regression. In fact, with this alternative approach, the maximum likelihood estimate of the (uniform) risk ratio was 1.7, the same as the standardized estimate (95% confidence limits from 0.7 to 4.1). In general, though, estimates that are based on an assumption of uniformity will differ from standardized estimates, in part because the latter may depend on the choice of standard.

Summary measures, such as standardized risks or risks, are probably best interpreted as a weighted average [8]. As such, a standardized

risk characterizes the average risk or risk for a population with the same distribution of covariates as the standard but may characterize individual risk poorly. For example, if substantial variability of stratum-specific risks exists or is predicted by a statistical model, such as the logistic model, a standardized risk may be highly dependent on the choice of weights [2]. If risks vary substantially, presentation of stratum-specific estimates in addition to, or instead of, summary measures may be more appropriate.

REFERENCES

1. Miettinen OS: Standardization of risk ratios. **Am J Epidemiol** 96: 383–388, 1972
2. Wilcosky TC, Chambless LE: A comparison of direct adjustment and regression adjustment of epidemiologic measures. **J Chron Dis** 38: 849–856, 1985
3. Kleinbaum DG, Kupper LL, Morgenstern J: **Epidemiologic Research: Principles and Quantitative Methods.** Belmont, Calif. Lifetime Learning Publications, 1982
4. Lee J: Covariance adjustment of rates based on the multiple logistic regression model. **J Chron Dis** 34: 415–426, 1981
5. Lane PW, Nelder JA: Analysis of covariance and standardization as instances of prediction. **Biometrics** 38: 613–621, 1982
6. Freeman DH, Holford TR: Summary rates. **Biometrics** 36: 195–205, 1980
7. Greenland S: Interpretation and estimation of summary ratios under heterogeneity. **Stat Med** 1: 217–227, 1982
8. Greenland S: Estimating variances of standardized estimators in case-control studies and sparse data. **J Chron Dis** 39: 473–477, 1986
9. Bishop YMM, Fienberg SE, Holland PW: **Discrete Multivariate Analysis: Theory and Practice.** Cambridge, Mass. MIT Press, 1975. pp. 487–494
10. Harrell FE: The PHLGM Procedure. In **SUGI Supplemental Library Users Guide.** Cary, N.C.: SAS Institute, Inc., 1983. pp. 267–294
11. **SAS User's Guide: Basics,** 1982 edn. Cary, N.C.: SAS Institute, Inc., 1982
12. Flanders WD: Approximate variance formulas for standardized rate ratios. **J Chron Dis** 37: 449–453, 1984
13. Miettinen OS: Estimability and estimation in case-referent studies. **Am J Epidemiol** 103: 226–235, 1976
14. Rothman KJ: A show of confidence. **N Engl J Med** 299: 1362–1363, 1978
15. Wacholder S: Binomial regression in GLIM: estimating risk ratios and risk differences. **Am J Epidemiol** 123: 174–184, 1986
16. Baker RJ, Nelder JA: **The GLIM System Release 3 Manual.** Oxford: Numerical Algorithms Group, 1978

APPENDIX 1

Evaluation of the Derivatives

For stratified sampling, $dR_{0,s}/d\beta_k$ and $dR_{1,s}/d\beta_k$ evaluated at the parameter estimates are given by:

$$\widehat{dR_{i,s}/d\beta_k} = \sum_j \frac{w_j E_{k,j} \exp(-(b_0 + b_1 E_{1,j} + b_2 E_{2,j} + \cdots + b_{M+1} E_{M+1,j}))}{(1 + \exp(-b_0 + \beta_1 E_{1,j} + b_2 E_{2,j} + \cdots + b_{M+1} E_{M+1,j})))^2} \quad \text{(A1)}$$

for $i = 0, 1$ and $k = 0, 1, \ldots, M+1$. In this expression, $E_{1,j} = 1$ for the index group and 0 otherwise, $E_{2,j}, \ldots, E_{M+1,j}$ are the values of the covariates for the jth stratum, and summation is over all strata ($j = 1, 2, \ldots, J$). An alternative expression for equation (A1) is given by:

$$\widehat{dR_{i,s}/d\beta_k} = \sum_j w_j E_{k,j} \hat{p}_j (1 - \hat{p}_j) \quad \text{(A1')}$$

where $\hat{p}_j$ is the model-predicted risk for stratum j.

For the marginal method, $\widehat{dR_{i,m}/d\beta_k}$ ($i = 0, 1$; $k = 0, 1, \ldots, M+1$) is given by equation (A2), where $E_{2,j}, \ldots, E_{M+1,j}$ now refer to the value of the covariates for the jth individual, and summation is over all individuals in the standard population.

$$\widehat{dR_{i,m}/d\beta_k} = \sum_j \frac{w_j E_{k,j} \exp(-(b_0 + b_1 E_{1,j} + b_2 E_{2,j} + \cdots + b_{M+1} E_{M+1,j}))}{(1 + \exp(-(b_0 + b_1 E_{1,j} + b_2 E_{2,j} + \cdots + b_{M+1} E_{M+1,j})))^2} \quad \text{(A2)}$$

For conditional standardization, $\widehat{dR_{i,c}/db_k}$ ($i = 0, 1$; $k = 0, 1, 2, 3$) is given by equation (A3), where $E^*_{2,1}, \ldots, E^*_{M+1,1}$ are the selected reference values.

$$\widehat{dR_{i,c}/d\beta_k} = \frac{E^*_k \exp(-(b_0 + b_1 E^*_1 + b_2 E^*_2 + \cdots + b_{M+1} E^*_{M+1}))}{(1 + \exp(-(b_0 + b_1 E^*_1 + b_2 E^*_2 + \cdots + b_{M+1} E^*_{M+1})))^2} \quad \text{(A3)}$$

The estimated derivative $\widehat{dR_{2,h}/d\beta_k}$ is given by:

$$\widehat{dR_{2,h}/d\beta_k} = (\widehat{dR_{1,h}/d\beta_k})/\hat{R}_{0,h} - (\widehat{dR_{0,h}/d\beta_k})(\hat{R}_{1,h})/\hat{R}^2_{0,h} \quad \text{(A4)}$$

where $\widehat{dR_{1,h}/d\beta_{k,h}}$ and $\widehat{dR_{0,h}/d\beta_{k,h}}$ are given by equations (A1–A3). The estimated derivative $dR_{3,h}/d\beta_k$ is given by:

$$\widehat{dR_3/d\beta_k} = (\widehat{dR_{1,h}/d\beta_k}) - (\widehat{dR_{0,h}/d\beta_k}). \quad \text{(A5)}$$

APPENDIX 2

SAS Program for Standardization

The following program can be used with "proc logist" (SAS) to compute standardized risks, risk differences, and risk ratios using results of logistic regression. To use the program, the (independent) variables in the "model" statement of "proc logist" must be listed in a specific order: the exposure of interest, E_1, must be listed first, variables for which no interaction terms with E_1 are used must be listed next, and variables for which interaction terms with E_1 are used must be last, each followed immediately by the interaction term with E_1. To illustrate using the example presented in the text, suppose that three new variables, race, horm*age, and horm*obese were to be included in the model, where horm*age and horm*obese represent interaction terms between hormone usage and age or obesity, respectively. An appropriate model statement would then be: "model disease = horm race age horm*age obese horm*obese." The confidence level is selected by setting "ZA" equal to the appropriate z score in line 16 of the program and the number of interaction terms that involve E_1 is indicated in line 14. The following SAS data sets are required:

Stratified standardization

"input": a data set with one "observation" for each stratum, including a variable encoding the level of each covariate and the weight for each stratum. If the number of covariates, not including

any interaction terms with E_1, is M, then the covariates must be the second through the $M+1$-th variables in the data set and must be in the same order as they are listed in the "model' statement. The weight is the $M+2$-th variable in the data set. If interaction between E_1 any of the covariates has been modeled with product terms, the corresponding variables must not be one of the first $M+2$-th variables in the data set. (The required values for these interaction terms are generated by the program.)

"covar": a data set with the parameter estimates as the first observation and the variance–covariance matrix as the remaining $M+2$ observations, created by proc logist if output = covar is included as an option.

Marginal standardization

"input": a data set with one observation for each subject in the standard population. The variables encode the level of the covariates and the weight for each subject, with each weight typically equal to 1 divided by the number of subjects. The order of the variables must be as described previously for stratified standardization.

"covar": as above

Conditional standardization

"input": a data set with one observation, giving the selected level for each of the covariates and a "dummy" weight equal to 1, with the same variable order as used previously.

"covar": as above

```
1          SAS          LOG          OS ASA 82.4                MVS/XA

NOTE: THE JOB WDF1FFFF HAS BEEN RUN UNDER RELEASE 82.4 OF SAS AT CENTER FOR DISEASES

1
2          DATA INPUT; SET IN1.WILC;
3

NOTE: DATA SET WORK.INPUT HAS 844 OBSERVATIONS AND 5 VARIABLES. 1066 OBS/TRK.
NOTE: THE DATA STATEMENT USED 0.10 SECONDS AND 320 K.

4          PROC LOGIST DATA=INPUT PCOV OUT=COVAR;
5            MODEL DISEASE=HORM OBESE AGE;
6
7          * CALCULATE STANDARDIZED RATES, RATE RATIOS, RATE DIFFERENCES AND;
8          * ASSOCIATED CONFIDENCE LIMITS;

NOTE: LOGIST IS SUPPORTED BY THE AUTHOR, NOT BY SAS.
NOTE: DATA SET WORK.COVAR HAS 5 OBSERVATIONS AND 4 VARIABLES. 1304 OBS/TRK.
NOTE: FRANK E. HARRELL, JR.
NOTE: CLINICAL BIOSTATISTICS
NOTE: BOX 3337, DUKE UNIVERSITY MEDICAL CENTER
NOTE: DURHAM, NC 27710
NOTE: THE PROCEDURE LOGIST USED 0.93 SECONDS AND 484 K AND PRINTED PAGES 1 TO 2.

9          PROC MATRIX;
10         * DATA SET INPUT IS THE STANDARD;
11         * DATA SET COVAR IS THE VARIANCE-COVARIANCE MATRIX;
12           FETCH X DATA=INPUT; FETCH C DATA=COVAR;
13         * INDICATE THE NUMBER OF INTERACTION TERMS;
14           M3=0;
15         * INPUT NORMAL DEVIATE FOR DESIRED ALPHA LEVEL;
16           ZA=1.96;
17         * GET MATRIX DIMENSIONS;
18           N=NROW(X); M=NCOL(C); MO=M+1;
19         * GET BETA'S, WEIGHTS, COVARIANCE MATRIX, COVARIATE VALUES;
20           BETA=(C(1,)');                    C=C(2:MO, 1:M);
21           M4=M-M3;                          W=X(, M4);
22           M4=M4-1;                          X=X(, 2:M4);
23         * CREATE MATRICES WITH COLUMNS OF 1'S OR 0'S, NEEDED FOR CALCULATIONS;
24           B1=J(N,1,1);                      BO=J(N,1,0);
25           M4=M4-M3-1;
26           X1=B1||B1;                        XO=B1||BO;
27           IF M4 GE 1 THEN DO;
28             X1=X1||X(,1:M4);                X0=X0||X(,1:M4);
29           END;
30         * APPROPRIATE EVALUATION OF INTERACTION TERMS WITH E1;
31           DO J=1 TO M3;
32             M4=M4+1;
33             X1=X1||(X(,M4)#X(,M4));
34             X0=X0||(X(,M4)#B0);
35           END;
36         * CALCULATE STANDARDIZED RATES;
37             XX1=1+EXP(-X1*BETA);            XX0=1+EXP(-XC*BETA);
38             R1=1#/XX1;                      R0=1#/XX0;
```

```
2          SAS          LOG          OS SAS 82.4          MVS/XA JOB WDF1FFFF STEP SAS          PR

39            SR1 = DET((W')*R1);                SR0 = DET((W')*R0);
40          * CALCULATE STANDARDIZED RATE RATIO AND RATE DIFFERENCE;
41            SR2 = SR1 # /SR0;                  SR3 = SR1 - SR0;
42          * CALCULATE PARTIAL DERIVATIVES;
43            R1 = (R1 # (1 - XX1) # /XX1) # W;   R0 = (R0 # (1 - XX0) # /XX0) # W;
44            PD1 = (R1')*X1;                    PD0 = (R0')*X0;
45            PD2 = PD1 # /SR0 - PD0 # SR1 # /SR0 # # 2;
46            PD3 = PD1 - PD0;
47          * CALCULATE VARIANCES;
48            VAR1 = PD1*C*(PD1');               VAR0 = PD0*C*(PD0');
49            VAR2 = PD2*C*(PD2');               VAR3 = PD3*C*(PD3');
50          * CALCULATE CONFIDENCE INTERVALS AND OUTPUT RESULTS;
51            NOTE STANDARDIZED RATE, REFERENCE GROUP:; R = 'LIMIT'; RR = ' '; S = 'ESTIMATE'
52              R = 'LIMIT'; RR = ' '; S = 'ESTIMATE';
53              MM = EXP(ZA # SQRT(VAR0) # /SR0);
54              LOWER = SRC # /MM; UPPER = SR0 # MM;
55              PRINT SR0 ROWNAME = RR COLNAME = S;
56              PRINT LOWER UPPER ROWNAME = RR COLNAME = R;
57            NOTE STANDARDIZED RATE, INDEX GROUP:;
58              MM = EXP(ZA # SQRT(VAR1) # /SR1);
59              LOWER = SR1 # /MM; UPPER = SR1 # MM;
60              PRINT SR1 ROWNAME = RR COLNAME = S;
61              PRINT LOWER UPPER ROWNAME = RR COLNAME = R;
62            NOTE STANDARDIZED RATE RATIO:;
63              MM = EXP(ZA # SQRT(VAR2) # /SR2);
64              LOWER = SR2 # /MM; UPPER = SR2 # MM;
65              PRINT SR2 ROWNAME = RR COLNAME = S;
66              PRINT LOWER UPPER ROWNAME = RR COLNAME = R;
67            NOTE STANDARDIZED RATE DIFFERENCE:;
68              MM = ZA*SQRT(VAR3);
69              LOWER = SR3 - MM; UPPER = SR3 + MM;
70              PRINT SR3 ROWNAME = RR COLNAME = S;
71              PRINT LOWER UPPER ROWNAME = RR COLNAME = R;

NOTE: THE PROCEDURE MATRIX USED 0.69 SECONDS AND 490 K AND PRINTED PAGES 3 TO 4.
      NOTE: SAS USED 490 K MEMORY.

NOTE: SAS INSTITUTE INC.
      SAS CIRCLE
      PO BOX 8000
      CARY, N.C. 27511-8000
```

LOGISTIC REGRESSION PROCEDURE

DEPENDENT VARIABLE: DISEASE

844 OBSERVATIONS
817 DISEASE = 0
27 DISEASE = 1
0 OBSERVATIONS DELETED DUE TO MISSING VALUES

VARIABLE	MEAN	MINIMUM	MAXIMUM	RANGE
HORM	0.71327	0	1	1
OBESE	0.42654	0	1	1
AGE	61.5581	55	80	25

−2 LOG LIKELIHOOD FOR MODEL CONTAINING INTERCEPT ONLY = 239.01

MODEL CHI-SQUARE = 11.05 WITH 3 D.F. (SCORE STAT.) P = 0.0115.
CONVERGENCE OBTAINED IN 6 ITERATIONS. R = 0.164.
MAX ABSOLUTE DERIVATIVE 0.3332D-08. −2 LOG L = 226.61.
MODEL CHI-SQUARE = 12.40 WITH 3 D.F. (−2 LOG L.R.) P = 0.0061.

VARIABLE	BETA	STD. ERROR	CHI-SQUARE	P	R
INTERCEPT	−1.78461299	1.59712597	1.25	0.2638	
HORM	0.55380834	0.47459977	1.36	0.2433	0.000
OBESE	−1.52488251	0.54850896	7.73	0.0054	−0.155
AGE	−1.02663475	0.02629423	1.03	0.3111	0.000

FRACTION OF CONCORDANT PAIRS OF PREDICTED PROBABILITIES AND RESPONSES :0.588
RANK CORRELATION BETWEEN PREDICTED PROBABILITY AND RESPONSE :0.361

LOGISTIC REGRESSION PROCEDURE

DEPENDENT VARIABLE: DISEASE

COVARIANCE MATRIX OF ESTIMATES

	INTERCEPT	HORM	OBESE	AGE
INTERCEPT	2.550811	−0.108003	−0.0125265	−0.0405273
HORM	−0.108003	0.2252449	−0.0199102	−0.00105917
OBESE	−0.0125265	−0.0199102	0.3008621	−0.000303813
AGE	−0.0405273	−0.00105917	−0.000303813	0.0006913866

STANDARDIZED RATE, REFERENCE GROUP:

SR0	ESTIMATE
	0.021718
LOWER	LIMIT
	0.00984148
UPPER	LIMIT
	0.0479271

STANDARDIZED RATE, INDEX GROUP:

SR1	ESTIMATE
	0.0369663
LOWER	LIMIT
	0.0243473
UPPER	LIMIT
	0.0561254

STANDARDIZED RATE RATIO:

SR2	ESTIMATE
	1.7021
LOWER	LIMIT
	0.693896
UPPER	LIMIT
	4.17518
	SAS

STANDARDIZED RATE DIFFERENCE:

SR3	ESTIMATE
	0.0152482
LOWER	LIMIT
	−0.00792938
UPPER	LIMIT
	0.0384258