AMERICAN
Journal of Epidemiology
Formerly AMERICAN JOURNAL OF HYGIENE

VOL. 119 FEBRUARY 1984 NO. 2

Reviews and Commentary

ESTIMATING ODDS RATIOS WITH CATEGORICALLY SCALED COVARIATES IN MULTIPLE LOGISTIC REGRESSION ANALYSIS[1]

STANLEY LEMESHOW AND DAVID W. HOSMER, JR.

An important use of multiple logistic regression analysis by epidemiologists is to obtain estimates of odds ratios controlling for other variables. This paper explains how to use the output from logistic regression computer programs to obtain odds ratio estimates and associated confidence intervals.

Problems in computation of the odds ratio estimate can arise for categorically scaled covariates measured at only two levels; as the number of levels increases, these computational problems may become extensive. The appropriate way to include a categorically scaled variable with K distinct categories in a statistical model is to construct $K - 1$ design (dummy or indicator) variables. There are several methods for creating these variables. The choice of method depends on a number of considerations including the goals of the analysis and ease of incorporation into the statistical software package being used. There is no one standard method for creating these design variables. In fact, some programs provide a choice of methods, each requiring a different technique for obtaining estimated odds ratios. The inconvenience of some of these techniques should persuade users to prefer one method over another.

The methodology in this paper is illustrated using data from a study considered by Lemeshow and Hosmer (1). That study involved the prediction of hospital mortality in 558 patients admitted to the general medical/surgical intensive care unit at Baystate Medical Center in Springfield, Massachusetts. Only patients with complete data on all condition and treatment variables were included in the analysis. This resulted in a reduction in the sample size from 558 to 540. For the purposes of this paper, it suffices to consider only the three independent variables given in table 1.

THREE COMMON METHODS FOR CODING DESIGN VARIABLES

Through user-specified transformations, each of the methods described in this section may be used with any of the more popular statistical packages having logistic regression capabilities (e.g., BMDP (2), GLIM (3), SAS (4)).

Since the 1983 release of the program BMDPLR (BMDP (2)) provides the option

[1] From the Biostatistics/Epidemiology Program, Division of Public Health, University of Massachusetts, Amherst, MA 01003. (Send reprint requests to Dr. Lemeshow at this address.)

TABLE 1

Summary of variables used in multiple logistic regression models for data collected at Baystate Medical Center, Springfield, Massachusetts

Variable	Code		No. (%)	Mean	Standard deviation
Age (years)				54.9	18.5
Coma	1	None or coma <48 hours	489 (90.6)		
	2	Coma ≥48 hours	51 (9.4)		
Mechanical ventilation	1	None or <24 hours	386 (71.5)		
	2	MVENT 1–4 days, no PEEP*	61 (11.3)		
	3	MVENT ≥5 days or PEEP	93 (17.2)		

* PEEP, positive end-expiratory pressure.

of using any one of the three most common methods for creating design variables without resorting to user-specified transformations, discussion of the methodology will center on this program. It should be stressed that the considerations addressed in this paper will apply to the analysis of design variables with any software package—not just BMDP.

Consider the trichotomous variable mechanical ventilation (MVENT). The three available options for coding this variable in BMDPLR (BMDP (2, p. 339)) are denoted "marginal" (MARG), "partial" (PART), and "orthogonal" (ORTH). Application of each of these methods to MVENT would yield the coding for two design variables w_1 and w_2 (table 2).

The method denoted MARG in BMDPLR creates two contrasts, w_1 and w_2, which compare each of the higher order levels of the categorical response variable to the baseline (or lowest) level. The method denoted PART creates two statistically independent variables which are particularly convenient when computing the log odds between any two levels of the categorically scaled covariate. Finally, the method denoted ORTH creates two design variables whose estimated logistic regression coefficients may be used to test for linear and quadratic trends, respectively.

TABLE 2

Design variables created for the trichotomous variable MVENT

Level of MVENT	Method					
	MARG		PART		ORTH	
	w_1	w_2	w_1	w_2	w_1	w_2
1	−1	−1	0	0	−0.7071	0.4082
2	1	0	1	0	0.0	−0.8165
3	0	1	0	1	0.7071	0.4082

Before proceeding, it is of interest to note what the values for a design variable, v, would be for a dichotomous variable such as COMA (table 3). In this case, the MARG and ORTH methods are equivalent to one another since they differ only by a multiplicative constant proportional to $\sqrt{2}$. Hence, the estimated logistic regression coefficients will differ by this same constant. The ORTH method of coding design variables should be used only when the original variable is at least ordinal scaled, with categories in some sense equally spaced. Because of these restrictions, we will not consider the ORTH method further in this paper. We note that users of the 1980 and 1981 releases of BMDP have only the method MARG available to them. To use PART or ORTH requires user-specified transformations.

TABLE 3

Design variables created for the dichotomous variable COMA

Level of COMA	Method		
	MARG	PART	ORTH
1	−1	0	−0.7071
2	1	1	0.7071

ESTIMATING ODDS RATIOS AND CONFIDENCE INTERVALS

Suppose the logistic regression model is used to obtain estimates of the probability that a patient will die ($Y = 1$) in the hospital given the values of AGE = a, MVENT, and COMA. This model states that

$$\Pr(Y = 1|a,w_1,w_2,v) = \frac{e^{l(a,w_1,w_2,v)}}{1 + e^{l(a,w_1,w_2,v)}},$$

where $l(a,w_1,w_2,v) = \beta_0 + \beta_1 a + \gamma_1 w_1 + \gamma_2 w_2 + \delta v$. The values of the estimated coefficients and their estimated standard errors for the two methods MARG and PART are given in table 4.

The estimated logit, or log odds, of dying versus living is

$$\hat{l}(a,w_1,w_2,v) = \hat{\beta}_0 + \hat{\beta}_1 a + \hat{\gamma}_1 w_1 + \hat{\gamma}_2 w_2 + \hat{\delta} v.$$

USE OF METHOD MARG

The log of the estimated odds ratio for level 2 of MVENT versus level 1 of MVENT, $\hat{\Psi}(2,1)$, holding AGE and COMA constant, is the difference between the respective logits. That is

$$\begin{aligned}\ln(\hat{\Psi}(2,1)) = {} & \{\hat{\beta}_0 + \hat{\beta}_1 a + \hat{\gamma}_1(1) + \hat{\gamma}_2(0) \\ & + \hat{\delta} v\} - \{\hat{\beta}_0 + \hat{\beta}_1 a \\ & + \hat{\gamma}_1(-1) + \hat{\gamma}_2(-1) + \hat{\delta} v\} \\ = {} & 2\hat{\gamma}_1 + \hat{\gamma}_2.\end{aligned}$$

Similarly, it can be shown that $\ln(\hat{\Psi}(3,1)) = 2\hat{\gamma}_2 + \hat{\gamma}_1$, $\ln(\hat{\Psi}(2,3)) = \hat{\gamma}_1 - \hat{\gamma}_2$.

The estimate of the odds ratio $\Psi(2,1)$ is $\exp(\ln(\hat{\Psi}(2,1)))$. Estimates of the other odds ratios are obtained in a similar manner. In the current example, these computations yield the following estimates:

$$\begin{aligned}\ln(\hat{\Psi}(2,1)) &= 2(0.124) + (0.714) \\ &= 0.962; \\ \hat{\Psi}(2,1) &= \exp(0.962) = 2.62, \\ \ln(\hat{\Psi}(3,1)) &= 2(0.714) + (0.124) \\ &= 1.552; \\ \hat{\Psi}(3,1) &= \exp(1.552) = 4.72, \\ \ln(\hat{\Psi}(2,3)) &= 0.124 - 0.714 = \\ &\quad -0.59; \\ \hat{\Psi}(2,3) &= \exp(-0.59) = 0.554.\end{aligned}$$

To assess the significance of, or to compute, confidence intervals for these estimated odds ratios, we need estimates of their standard errors. For the estimated log odds $\ln(\hat{\Psi}(2,1))$, it can be shown that

TABLE 4

Estimated logistic regression coefficients and standard errors (SE) for two methods, MARG and PART, of creating design variables

Variable and design variable	Method			
	MARG		PART	
	Coefficient	$\widehat{SE}$(coefficient)	Coefficient	$\widehat{SE}$(coefficient)
AGE	0.034	0.009	0.034	0.009
MVENT				
w_1	0.124	0.216	0.962	0.332
w_2	0.714	0.244	1.552	0.385
COMA				
v	1.429	0.209	2.858	0.418
Constant	−2.216	0.523	−4.483	0.579

$$\hat{SE}\{\ln(\hat{\Psi}(2,1))\} = \{4\,\hat{SE}(\hat{\gamma}_1)^2 + \hat{SE}(\hat{\gamma}_2)^2 + 4\,\hat{Cov}(\hat{\gamma}_1,\hat{\gamma}_2)\}^{1/2},$$

$$\hat{SE}\{\ln(\hat{\Psi}(3,1))\} = \{4\,\hat{SE}(\hat{\gamma}_2)^2 + \hat{SE}(\hat{\gamma}_1)^2 + 4\,\hat{Cov}(\hat{\gamma}_1,\hat{\gamma}_2)\}^{1/2},$$

and

$$\hat{SE}\{\ln(\hat{\Psi}(2,3))\} = \{\hat{SE}(\hat{\gamma}_1)^2 + \hat{SE}(\hat{\gamma}_2)^2 - 2\,\hat{Cov}(\hat{\gamma}_1,\hat{\gamma}_2)\}^{1/2},$$

where

$$\hat{Cov}(\hat{\gamma}_1,\hat{\gamma}_2) = \hat{Corr}(\hat{\gamma}_1,\hat{\gamma}_2) \cdot \hat{SE}(\hat{\gamma}_1) \cdot \hat{SE}(\hat{\gamma}_2).$$

Using the estimated correlation matrix of the estimated coefficients and the estimated standard errors which are part of the BMDPLR output, we find that $\hat{Corr}(\hat{\gamma}_1,\hat{\gamma}_2) = -0.644$ and, hence, $\hat{Cov}(\hat{\gamma}_1,\hat{\gamma}_2) = -0.644(0.216)(0.244) = -0.34$. The resulting estimated standard error of $\ln(\hat{\Psi}(2,1))$ is

$$\hat{SE}\{\ln(\hat{\Psi}(2,1))\} = \{4(0.216)^2 + (0.244)^2 + 4(-0.034)\}^{1/2} = 0.332.$$

Similarly, it can be shown that $\hat{SE}\{\ln(\hat{\Psi}(3,1))\} = 0.386$ and $\hat{SE}\{\ln(\hat{\Psi}(2,3))\} = 0.417$. The confidence limits for an approximate 95 per cent confidence interval for $\ln(\Psi(2,1))$ are

$$\ln(\hat{\Psi}(2,1)) \pm 1.96\,\hat{SE}\{\ln(\hat{\Psi}(2,1))\}.$$

The resulting 95 per cent confidence interval is $0.311 \leq \ln(\hat{\Psi}(2,1)) \leq 1.613$, and by exponentiating these limits, we obtain $1.37 \leq \Psi(2,1) \leq 5.02$. Similar calculations can be performed for the other odds ratios $\Psi(3,1)$ and $\Psi(2,3)$.

For the dichotomous variable COMA, the estimated log odds of COMA $\geq$48 hours ($v = +1$) versus COMA <48 hours or no COMA ($v = -1$), holding AGE and MVENT fixed, is

$$\ln(\hat{\Psi}(2,1)) = (\hat{\beta}_0 + \hat{\beta}_1 a + \hat{\gamma}_1 w_1 + \hat{\gamma}_2 w_2 + \hat{\delta}(1)) - (\hat{\beta}_0 + \hat{\beta}_1 a + \hat{\gamma}_1 w_1 + \hat{\gamma}_2 w_2 + \hat{\delta}(-1)) = 2\hat{\delta}.$$

Thus, the estimated odds ratio is

$$\exp(2\hat{\delta}) = \exp(2(1.429)) = 17.43.$$

To obtain the corresponding approximate 95 per cent confidence estimate of $\Psi(2,1)$, an estimate of the standard error of $\ln(\hat{\Psi}(2,1)) = 2\hat{\delta}$ is needed. This is $2\{\hat{SE}(\hat{\delta})\}$, where $\hat{SE}(\hat{\delta})$ is obtained from the BMDPLR output. Hence, the approximate 95 per cent confidence interval is

$$\exp\{2(1.429) - 1.96(2)(0.209)\} \leq \Psi(2,1) \leq \exp\{2(1.429) + 1.96(2)(0.209)\}$$

or

$$7.681 \leq \Psi(2,1) \leq 39.54.$$

As can be seen, a significant amount of computational effort is necessary if the BMDP method of MARG is used for creating design variables. Use of the PART method will simplify the necessary computations.

Use of method PART

When the same development used for the presentation of the results for the MARG method is followed, it can be shown that the estimated log odds and odds ratios for MVENT are $\ln(\hat{\Psi}(2,1)) = \hat{\gamma}_1$, $\ln(\hat{\Psi}(3,1)) = \hat{\gamma}_2$, and $\ln(\hat{\Psi}(2,3)) = \hat{\gamma}_1 - \hat{\gamma}_2$. The estimated standard errors of the log odds are: $\hat{SE}\{\ln(\hat{\Psi}(2,1))\} = \hat{SE}(\hat{\gamma}_1)$, $\hat{SE}\{\ln(\hat{\Psi}(3,1))\} = \hat{SE}(\hat{\gamma}_2)$, $\hat{SE}\{\ln(\hat{\Psi}(2,3))\} = \{\hat{SE}(\hat{\gamma}_1)^2 + \hat{SE}(\hat{\gamma}_2)^2 - 2\hat{Cov}(\hat{\gamma}_1,\hat{\gamma}_2)\}^{1/2}$, where the standard errors are obtained from the computer output and the covariance is obtained as the product of the correlation and the standard errors. For the dichotomous variable COMA, $\ln(\hat{\Psi}(2,1)) = \hat{\delta}$ and $\hat{SE}\{\ln(\hat{\Psi}(2,1))\} = \hat{SE}(\hat{\delta})$. In this example, these quantities are given in table 4.

The reader may verify that computations with these expressions yield precisely the same estimates and confidence intervals for the odds ratios as given in the discussion of the MARG method.

Discussion

When a goal of the analysis is to compare all levels of a categorically scaled

variable with a referent level, much less computational effort is required to obtain odds ratio estimates and confidence intervals with the PART method than with the MARG method. This is because the MARG method yields more complex expressions for the standard errors. These computations become particularly tedious when the categorical variable is measured at four or more levels. For comparisons of any other two levels, the computations required are equivalent for the two methods.

It has been our experience that many users of the 1980 and 1981 versions of BMDPLR, when declaring variables to be categorical, assume that the method being used to create design variables is PART when in fact it is MARG. They then proceed to calculate odds ratios using the computations for PART which are incorrect and lead to erroneous odds ratio estimates.

In summary, the PART method of specifying design variables will save a significant amount of computational effort when there is interest in comparing the levels of the categorical variable with a single referent level. The PART method leads to direct estimates of the log odds ratios and the estimated standard errors may be obtained directly from the output. The PART method of specifying design variables may be used in the 1980 and 1981 versions of BMDPLR by having users create their own design variables coded (0,1) and declare them to be "interval" scaled. This strategy becomes cumbersome when considering numerous models and/or many variables.

Our recommendation is that at the model building stage any of the methods for creation of design variables may be used. Once a final model has been selected, the user should form the design variables using the PART method to simplify the computations necessary to obtain estimated odds ratios and their associated confidence intervals.

References

1. Lemeshow S, Hosmer DW Jr. A review of goodness of fit statistics for use in the development of logistic regression models. Am J Epidemiol 1982:115:92-106.
2. Dixon WJ. BMDP statistical software. Berkeley, CA: University of California Press, 1983.
3. Baker RJ, Nelder JA. The GLIM system release 3 generalized linear interactive modeling. Oxford, England: Numerical Algorithms Group, 1978.
4. Helwig JT, Council KA. SAS users guide. Raleigh, NC: SAS Institute, Inc, 1979.