AMERICAN JOURNAL OF EPIDEMIOLOGY
Vol. 125, No. 2
Printed in U.S.A.

CODING ORDINAL INDEPENDENT VARIABLES IN MULTIPLE REGRESSION ANALYSES

S. D. WALTER,[1] A. R. FEINSTEIN,[2] AND C. K. WELLS[2]

Walter, S. D. (Dept. of Clinical Epidemiology and Biostatistics, McMaster U., Hamilton, Ontario, Canada L8N 3Z5), A. R. Feinstein, and C. K. Wells. Coding ordinal independent variables in multiple regression analyses. *Am J Epidemiol* 1987;125:319-23.

The authors present a coding scheme for ordinal independent variables which may be used in various forms of regression analysis. The scheme is useful in dose-response analyses, when the objective is to identify contrasts in the dependent (or response) variable between successive levels of the independent variable, or to identify critical threshold values of the independent variables at which significant changes occur in the response. An example is given of evaluating the survival of lung cancer patients according to their stage of symptomatology. The authors discuss the interpretation of the regression coefficients when this coding scheme is used with linear regression, logistic regression, or in the proportional hazards regression model.

biostatistics; regression analysis

Several investigators (1–3) have discussed the question of coding schemes for discrete independent variables in multiple logistic regression and multiple linear regression analyses. They have described various ways to characterize the contrasts between the responses at the different levels (or strata) of the independent variables. Two common options are 1) to compare the responses in each stratum with the responses in a single referent or control stratum or 2) to compare the responses in each stratum with the average response for the whole sample.

We would like to draw attention to an additional coding scheme that we have found useful in defining contrasts for regression analyses involving categorical independent variables that have discrete *ordered* categories. If the categories are coded by using a single variable which takes values in a direct numerical sequence, such as 1,2,3,4, . . . , the regression procedure will analyze these numbers as though they represented a true dimension, with equal intervals measured between adjacent categories. If the ordinal values are arbitrarily assigned rather than actually measured, the regression coefficient is difficult or impossible to interpret. Also, depending on how closely the assumed scaling (1,2,3, . . .) agrees with the actual dose-response relationship, the test of significance of the coefficient may or may not be powerful against alternative hypotheses of interest. If the true dose-response relationship is linear, with equal increments between successive categories, the one degree of freedom test

Received for publication March 31, 1986, and in final form July 2, 1986.

[1] Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada.

[2] Clinical Epidemiology Unit and Robert Wood Johnson Clinical Scholars Program, Yale University, New Haven, CT.

Reprint requests to Dr. S. D. Walter, Department of Clinical Epidemiology and Biostatistics, McMaster University, 1200 Main Street West, Hamilton, Ontario, Canada L8N 3Z5.

This research was supported in part by the National Health Research and Development Program through a National Health Scientist award (to S. D. W.) and in part by Grant 6309 from the Robert Wood Johnson Foundation.

of the regression coefficient will be relatively powerful. On the other hand, if a threshold exists, the overall linear test of trend may have low power, as opposed to a coding scheme that can completely reproduce the threshold relationship.

Objectives of proposed coding scheme

The coding we have developed is designed to preserve the ordinal ranks of the independent variable, while identifying the ranked categories in a manner suitable for subsequent analysis. With this type of data, a frequent goal is to evaluate the contrasts between the responses in the successive ordered categories. This goal involves an examination of the progressive changes in the dependent variable as one moves through the strata of the independent variable in their defined order. In a related but different type of evaluation, the goal is to find a *particular* point in the sequence of strata at which a substantial change occurs in the dependent variable, for example, at a critical "threshold" value of the independent variable.

As an example, we have been considering the relationship of symptom stage to survival in a series of lung cancer patients (4). The four levels of symptoms are asymptomatic, primary symptoms only, systemic symptoms (with or without primary symptoms), and metastatic symptoms (with or without primary and/or systemic symptoms). Although these categories are not expressed quantitatively, their order corresponds to the usual clinical concepts of severity. Although this order is also likely to correlate with prognosis for survival, the correlation is of course not a requirement of our proposed coding method.

We required a coding scheme that would fulfill at least two objectives. First, we wanted to estimate the prognostic differences in survival between patients in successive symptom categories, for example, between the systemic and metastatic groups. Second, we wished to identify any important thresholds in symptoms that would correspond to the most significant differences in survival. This approach might, for instance, conclude that the largest differences in the outcome were between patients with any symptoms (primary, systemic, or metastatic) as opposed to those with no symptoms.

Definition of coding scheme

Our general coding scheme is as follows. Suppose there are $(k + 1)$ ordered levels of the independent variable; this implies that we may define (up to) k dummy independent variables $X_1, X_2, \ldots X_k$ to describe between-strata differences. The strata will be labeled in sequence as $0,1,2, \ldots k$. Then, the values of the independent variables in each stratum are

Stratum no.	Dummy variable					
	X_1	X_2	X_3		X_{k-1}	X_k
0	0	0	0	...	0	0
1	1	0	0	...	0	0
2	1	1	0	...	0	0
.	.	.	.		.	.
.	.	.	.		.	.
.	.	.	.		.	.
$k-1$	1	1	1	...	1	0
k	1	1	1	...	1	1

In our lung cancer example, therefore, we defined three variables as follows:

Symptom stage	X_1	X_2	X_3
Asymptomatic	0	0	0
Primary	1	0	0
Systemic	1	1	0
Metastatic	1	1	1

Analysis of survival data requires the use of special regression techniques, such as the Cox proportional hazards model (5). It is, however, somewhat simpler if we introduce the general principles of our coding scheme by using ordinary linear regression. In a later section, we will discuss the interpretation of the regression coefficients when using proportional hazards or logistic regression models.

INTERPRETATION OF GENERAL REGRESSION MODELS

We now consider, in general, regression models that may include $X_1, X_2, \ldots X_k$ in various combinations. Two specific models are of particular interest. First, we may fit the model with all k independent variables included; in this model, the coefficient of X_i corresponds to the difference in responses between levels i and $(i - 1)$, and all such differences are estimated. Consider, for example, the hypothetical situation shown as pattern 1 in table 1. Here, there is a progressive increase in the dependent variable going through categories A, B, C, and D of the independent variable. The largest difference appears to be between individuals in categories C and D. Fitting the model with all three independent variables included yields the regression equation

$$Y = 5 + X_1 + 2X_2 + 7X_3. \qquad (1)$$

By substituting appropriate values of X_1, X_2, and X_3 into equation 1, one may calculate the average response for individuals in each of the categories A to D. For example, in category C, we have $X_1 = 1$, $X_2 = 1$, and $X_3 = 0$, thus, the predicted response for level C in equation 1 is $Y = 5 + 1 + 2 + 0 = 8$. (The reader may verify that the average values may be computed similarly for the other groups.) The large difference between categories C and D is reflected in the large coefficient for X_3.

Second, the model may include only a single independent variable, X_i say; here, the coefficient of X_i defines the contrast between the aggregated responses in level i and above, and the aggregated responses in level $(i - 1)$ and below. For instance, in the previous example, the model with X_2 alone would provide a comparison between individuals in C and D (combined) and individuals in A and B (combined). This model would be appropriate for pattern 2 of table 1, where a jump in the level of response occurs between the two highest and two lowest of the ordered strata. The coefficient of X_2 (equal to 4) represents the difference in response between these two sets. Individuals in categories A or B have $X_2 = 0$ and average response values of 5; in contrast, the other two categories have $X_2 = 1$, yielding a prediction of $Y = 5 + 4 = 9$.

This second model is attractive when "stepwise" methods are used. By adopting a forward inclusion rule for model selection, one will identify at the first step the cut point defining the most significant dichotomy on the scale of the independent variable X. Subsequent steps in the process will identify any other demarcations in the scale where meaningful changes occur in the response. Suppose pattern 3 of table 1 applied. With sufficient data, a stepwise algorithm will select X_1 and X_3 into the model. We would conclude that there is no prognostic difference between categories B and C, but that individuals in categories A and D differ from this combined "middle" stratum (B and C) by amounts that are represented by the coefficients of X_1 and X_3, equal to 3 and 7, respectively.

We will now discuss the interpretation of the coefficients b_i in three particular types of regression model, the linear, the proportional hazards, and the logistic.

TABLE 1

Three hypothetical response patterns and associated regression models

Pattern	Average value of dependent variable in various categories (A–D) of an independent variable				Model
	A	B	C	D	
1	5	6	8	15	$Y = 5 + X_1 + 2X_2 + 7X_3$
2	5	5	9	9	$Y = 5 + 4X_2$
3	5	8	8	15	$Y = 5 + 3X_1 + 7X_3$

INTERPRETATION OF COEFFICIENTS IN SPECIFIC REGRESSION MODELS

Linear regression

The examples discussed above and shown in table 1 have a simple interpretation in the context of ordinary linear regression. The dependent variable Y is

modeled directly, and the regression equation gives the expected value of Y itself. Thus, the expected value of Y is given by the regression equation $Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_kX_k$ in the full model, for instance. This approach is appropriate if the dependent variable is continuous, for instance, systolic blood pressure or serum cholesterol.

Proportional hazards regression

In the analysis of survival data, the proportional hazards regression model uses as its dependent variable the hazard function $h(t, \mathbf{X})$, where $\mathbf{X}$ represents the set of exposure variables for an individual. Under the proportional hazards assumption, this function may be represented as

$$h(t, \mathbf{X}) = h_0(t) \exp(b_0 + b_1X_1 + b_2X_2 + \ldots + b_kX_k), \quad (2)$$

where $h_0(t)$ is an underlying hazard, a function of time t only. The second component of $h(t, \mathbf{X})$ represents the way in which the hazard varies according to an individual's values of $X_1, X_2, \ldots X_k$. Overall, $h(t, \mathbf{X})$ gives the instantaneous probability of death at time t, for an individual with exposure values $\mathbf{X}$.

If there is a single regression variable X_1, which takes values 0 and 1, the hazard function for individuals with $X_1 = 1$ (the "exposed," say) is from equation 2:

$$h_E = h_0(t) \exp(b_0 + b_1),$$

while the hazard function for those with $X_1 = 0$ (the "unexposed") is

$$h_{\bar{E}} = h_0(t) \exp(b_0).$$

Thus, $h_E/h_{\bar{E}} = \exp(b_1)$, or $\log(h_E/h_{\bar{E}}) = b_1$. Hence, we may think of b_1 as a "log relative risk" of the "exposed" to the "unexposed" which, strictly speaking, applies to an infinitesimally short interval of time. Equivalently, $\exp(b_1)$ is the relative risk. Note that this relative risk is constant over the time t of follow-up; this is in fact the essence of the proportional hazards assumption, that the relative risk of X_1 is constant, regardless of the absolute risk (proportional to $h_0(t)$) or the time instant in question.

Now with a general number of ordered strata as discussed earlier, the exponentials of the coefficients $b_1, b_2, \ldots b_k$ represent relative risk values, comparing various of the strata in the same way as for linear regression. For instance, if all the variables $X_1, X_2, \ldots X_k$ are included in the model (analogous to pattern 1 of table 1), $\exp(b_i)$ gives the relative risk between stratum i and stratum $i - 1$.

Logistic regression

In logistic regression, the model is specified by

$$\log(p/1 - p) = b_0 + b_1X_1 + \ldots + b_kX_k,$$

where p is the probability of an event under study (e.g., the probability that a certain birth will have a congenital anomaly). This method is widely used to analyze case-control data (6).

As before, if there is a single predictor X_1 with values 0 and 1, the probability p_E for the exposed ($X_1 = 1$) is defined by

$$\log(p_E/1 - p_E) = b_0 + b_1,$$

and the probability $p_{\bar{E}}$ for the unexposed ($X_1 = 0$) is given by

$$\log(p_{\bar{E}}/1 - p_{\bar{E}}) = b_0.$$

Hence,

$$\log(p_E(1 - p_{\bar{E}})/p_{\bar{E}}(1 - p_E) = b_1,$$

showing that b_1 may be thought of here as a log odds ratio; similarly, $\exp(b_1)$ is the odds ratio itself.

Finally, if we return again to the situation with k ordered strata, the coefficients b_i now represent log odds ratios between appropriate strata. For instance, if only X_i is fitted in the model (analogous to pattern 2 of table 1), $\exp(b_i)$ is the odds ratio between the combined strata i and above, and the combined strata $(i - 1)$ and below.

Discussion

We have presented a coding scheme for regression modeling, which is particularly

useful in examining dose-response relationships for data with ordered categorical exposure strata.

It should be pointed out that if one has $(k + 1)$ strata and one parameterizes a full set of k independent variables, it is always possible to "convert" from any one coding scheme to any other, obtaining point estimates and variances; this will typically involve some algebraic manipulation of the codings, and of the corresponding variance-covariance matrices for the parameters. However, the overall significance test of association on k degrees of freedom, comparing differences between all the strata simultaneously, is the same for any valid coding scheme.

The choice between alternative codings should be based on which particular contrasts between the various strata are required. The correct choice will lead directly to relevant parameters, such as mean differences or odds ratios, even if only a partial model is fitted. An incorrect coding choice in a partial model will yield results that are generally not "convertible" to another coding scheme, so reanalysis may be needed.

As usual, one must be conscious of the danger of overanalysis of data, in this context by considering a variety of coding schemes to find "the most significant effect." With $(k + 1)$ strata, there are only k degrees of freedom available to examine between-strata differences. This means that a maximum of k statistically independent statements about the data can be made. Ideally, the (up to) k contrasts to be tested should be specified a priori, and the coding scheme should be chosen accordingly.

References

1. Lemeshow S, Hosmer DW Jr. Estimating odds ratios with categorically scaled covariates in multiple logistic regression analysis. Am J Epidemiol 1984:119:147–51.
2. Fleiss JL. Re: "Estimating odds ratios with categorically scaled covariates in multiple logistic regression analysis." (Letter). Am J Epidemiol 1985:121:476–7.
3. Dulberg C. Another view. (Letter). Am J Epidemiol 1985:121:477–8.
4. Feinstein AR, Wells CK. Lung cancer staging: a critical evaluation. Clinics Chest Med 1982:3:291–305.
5. Lee ET. Statistical methods for survival data analysis. Belmont, CA: Lifetime Learning Publications, 1980.
6. Breslow NE, Day NE. Statistical methods in cancer research. Vol 1. The analysis of case-control studies. Lyon: International Agency for Research on Cancer, 1980.