September, 1982

Regression analysis in health services research: the use of dummy variables

Paula Diehr, *University of Washington*
Lincoln Nayak Polissar

Reprinted from MEDICAL CARE, September 1982
Vol. 20, No. 9

Regression Analysis in Health Services Research: The Use of Dummy Variables

LINCOLN POLISSAR, PH.D.,* AND PAULA DIEHR, PH.D.†

Dummy variables frequently are used in regression analysis but often in an incorrect fashion. A brief review of examples in the medical care literature showed that the interpretation of dummy variable regression coefficients and their significance was often incorrect or unclear. This article shows how dummy variables can be used and assessed properly. The importance of testing for the joint effect of a group of dummy variables is stressed. It also gives a standard and useful extension of the dummy variable technique to testing for the effect of collections of variables.

IN SEVERAL RECENT issues of *Medical Care* and *Health Services Research*, nine articles were found that used dummy variables in regression analysis to represent categorical variables in a regression model.[1–9] The results often were interpreted incorrectly and the appropriate statistical tests were not always performed. Dummy variables provide a powerful and useful tool for analysis, but their misuse by researchers produces unnecessary misinformation. The correct use of dummy variables is based on analysis of variance and covariance, as well as tests of the general linear model and the construction of F-tests. Although all of these topics are accessible individually,[10,11] they rarely are presented with direct reference to dummy variable regression and some of the subtleties of the subject are lost. In this paper the proper use and interpretation of "dummies" is presented with illustrations from the nine articles reviewed, in order to help the readers of medical care literature assess the use of dummy variables in regression analysis more accurately, and as an aid to those whose research requires their use.

As an example, which shall be carried throughout, suppose that the number of patient visits to a clinic (VISITS) is the dependent variable and the type of insurance carried by the patient (INSURANCE) is the independent variable. INSURANCE can be coded in categories "health maintenance organization (HMO)," "Blue Cross" and "none"; but these categories have no numerical values associated with them and therefore cannot be put directly into a regression model.

The situation can be handled by introducing a group of dichotomous variables to represent the effect of INSURANCE, as follows:

HMO = 1 if INSURANCE is coded "health maintenance organization" for a subject
= 0 for all other subjects

BLUE = 1 if INSURANCE is coded "Blue Cross" for a subject
= 0 for all other subjects

NONE = 1 if INSURANCE is coded "none" for a subject
= 0 for all other subjects

Note that the type of insurance for a subject is uniquely specified by the values of any two of these three variables.

* Research Associate Professor, Department of Biostatistics, University of Washington and the Fred Hutchinson Cancer Research Center, Seattle, Washington.

† Associate Professor, Department of Biostatistics, University of Washington, Seattle, Washington.

Reprint requests: Dr. Lincoln Polissar, Fred Hutchinson Cancer Research Center, 1124 Columbia St., Seattle, WA 98104.

0025-7079/82/0900/0959/$01.20

No term in the literature has been found to distinguish between the categorical variable, such as INSURANCE, and the dummy variables (HMO, BLUE, NONE) derived from it. The term construct has been chosen to denote the categorical variable in this article as opposed to the individual dummies. Marital status is another example of a construct with commonly used dummy variables MARRIED, DIVORCED, WIDOWED and SEPARATED. Some constructs used in the papers reviewed were time,[1] treatment plan,[3] distance from a clinic, years employed, number of family members with insurance,[6] per capita income, perceived severity of illness,[5] and geographic regions of the country.[7] The distinction between constructs and dummies is important, because theories and hypotheses usually involve an entire construct, but researchers typically have reported results only for individual dummies, a pitfall which we shall discuss below.

Regression Models With Dummy Variables

In the example above, the construct INSURANCE with three categories can be represented in a regression model by using two dummy variables, for example, HMO and NONE. The category whose dummy is omitted from the regression model is referred to as the reference category. It is often good practice to choose the most frequent category as the reference category, so that the dummy regression coefficients represent deviations of smaller groups from the largest group. In one study,[6] the reference category was the small group of subjects with missing values. This was a poor choice since there was no particular interest in detecting differences between respondents and nonrespondents.

In our example we will presume that Blue Cross members are the largest group in our study. Suppose that in addition to the independent variable INSURANCE, there is a second independent variable of known importance: AGE. Besides the construct of interest, there are usually other variables, such as AGE, in a regression model. The simplest linear regression model involving all of these variables is:

$$\text{VISITS} = B_0 + B_1(\text{HMO}) + B_2(\text{NONE}) + B_3(\text{AGE}) + e \quad (1)$$

where e is a random error term, and the B's are the usual regression coefficients. Each dummy variable coefficient represents the mean difference in the dependent variable between a category and the reference category. For example, if the estimated values from the fitted regression equations are $\hat{B}_1 = -1.1$ and $\hat{B}_2 = -2.3$, then HMO participants and uninsured persons have, respectively, a mean of 1.1 and 2.3 fewer visits than Blue Cross participants during the period of the study, after controlling for the effect of age.

Notice that the effect of the dummy variables HMO and NONE is either to leave out or to include the coefficients B_1 and B_2 in the model. Thus, though there is only one regression model, there are three equations corresponding to different values of HMO and NONE. For example, HMO participants always will have a value of one for HMO and zero for NONE, so that the equation for HMO participants will be: $\text{VISITS} = B_0 + B_1(1) + B_2(0) + B_3(\text{AGE}) + e$.

The reduced regression equations for the three cases are as follows:

Blue Cross	$\text{Visits} = B_0 + B_3(\text{AGE}) + e$
Health Maintainence Organization	$\text{Visits} = B_0 + B_1 + B_3(\text{AGE}) + e$
No Insurance	$\text{Visits} = B_0 + B_2 + B_3(\text{AGE}) + e$

If VISITS are plotted against AGE, the fitted regression lines for the three insurance plans will be parallel with the same slope, $\hat{B}_3$. They will have different intercepts, $\hat{B}_0$, $(\hat{B}_0 + \hat{B}_1)$ and $(\hat{B}_0 + \hat{B}_2)$. respectively.

Testing the Construct

In the nine articles that used multiple dummy variables to represent a single con-

struct, only one tested for the effect of the entire construct.[1] Typically, the other papers only commented on the significance of the individual dummy variables, yet the hypotheses were implicitly formulated in terms of entire constructs and it is possible that some incorrect conclusion were drawn.

It may be argued that tests of constructs are rarely of interest, since the investigator usually would be interested in knowing which of the categories were significantly different from others, and thus would look only at the individual significance levels of dummies. Dummy variable methodology, however, is based on analysis of variance and covariance models. In the strict application of these models it generally is considered incorrect to test for differences between categories of a construct when the overall construct is not significant. Thus, what is done often in practice—looking only at individual dummies and their "p values"—may not be on firm statistical ground.

Moreover, whenever a regression equation is presented, with the p values for each of its coefficients, it is easy for a reader to incorrectly infer the significance of the construct. For example, if one sees a regression equation in which neither HMO or NONE has a significant regression coefficient, a natural, though possibly incorrect, conclusion would be that the construct INSURANCE was not important in predicting the dependent variable. In fact, the combined effect of HMO and NONE might be highly significant. Future research then could neglect to control for INSURANCE based on this misunderstanding.

As noted in the next section, interpretation of individual dummy variables is a fairly subtle process and signficance of dummies does not necessarily imply significance of constructs and vice versa.

Before looking at the significance of individual dummies then, it must be determined whether the construct INSURANCE has a nonrandom effect on VISITS. The correct way to examine this is to determine whether the variables HMO and NONE taken together (keeping BLUE as the reference category) significantly improve the prediction of VISITS. To test the construct we would compare two regression models: one with and one without the dummy variables HMO and NONE, as follows:

$$\text{VISITS} = B_0 + B_3(\text{AGE}) + e$$

compared to

$$\text{VISITS} = B_0 + B_1(\text{HMO}) + B_2(\text{NONE}) + B_3(\text{AGE}) + e.$$

We would like to know if the prediction of VISITS is better in the second model than in the first. If not, we accept the null hypothesis H_0, that $B_1 = B_2 = 0$; in this case, we conclude that the separation of the three estimated regression lines has arisen randomly and INSURANCE has no discernable effect on VISITS. Under the usual assumption that e (error in measuring visits) is independently and normally distributed with mean zero and a constant variance, the test statistic for the effect of INSURANCE is the familiar F statistic, which can be calculated as follows. Let R_1 be the value of the multiple correlation coefficient with only AGE included and R_3 be the multiple correlation coefficient with the construct INSURANCE also included via HMO and NONE. The F-statistic for testing H_0: $B_1 = B_2 = 0$ with n observations is:

$$F_{2,n-4} = \frac{(R_3^2 - R_1^2)/2}{(1 - R_3^2)/(n - 4)} \quad . \qquad (2)$$

$F_{2,n-4}$ can be used to enter a table of F values to get the significance level (the "p value") for H_0.

In the general case, suppose that a construct "CAT" has k + 1 categories so that k dummy variables have been created from it for inclusion in the regression model. Label these dummies as D1, D2, . . ., Dk with regression coefficients $B_1, \ldots, B_k$,

and suppose that m additional noncategorical variables such as AGE, INCOME, etc., are included in the regression equation. The main hypothesis of interest is $H_0: B_1 = \ldots = B_k = 0$. This hypothesis states that the construct CAT has no effect on the dependent variable after controlling for the other m variables. The F statistic for n observations is

$$F_{k,n-k-m-1} = \frac{(R^2_{k+m} - R^2_m)/k}{(1 - R^2_{k+m})/(n - k - m - 1)}, \quad (3)$$

where R^2_m and R^2_{k+m} are the familiar squared multiple correlation coefficients for equations with, respectively, m independent variables and with these m plus the k dummies. The significance level of the test can be found in an F table under (k) numerator and $(n - k - m - 1)$ denominator degrees of freedom. If the F statistic is not significant, then we fail to reject the null hypothesis that the true coefficients for the dummies are zero.

Most multiple regression computer packages provide tests of significance for the individual dummies, but do not, without further manipulation, provide tests of significance for the entire construct. The information needed can be obtained usually in a single computer run; SPSS and BMD are useful computer packages for doing this.

Interpretation of Individual Dummy Variables

Although examining the statistical significance of the entire construct is the first priority, researchers will naturally want to look at the parts of the construct. However, interpretations of the statistical significance of individual dummy variables must be done very carefully for the following reasons.

Effect of the Reference Category

The meaning, numerical value, and statistical significance of dummy variable coefficients depends on the reference category. Dummies that are statistically significant with one reference category can be nonsignificant with a different choice. The significance level of the dummy depends on the mean difference in dependent variable values between observations in the reference category and observations in the category represented by the dummy. However, the significance of the construct will not be affected by a change in reference category, nor will the mean of the dependent variable (adjusted or unadjusted for other regression variables) be affected in any category.

Obviously, changing the reference category means changing the groups of observations that are being compared and this will change the significance level of the dummy.

As an example, suppose that the mean number of visits to a clinic is highest for Blue Cross enrollees, intermediate for HMO enrollees and lowest for uninsured persons. If NONE is the reference category, then BLUE might be statistically significant because it is the highest category and is being compared to the lowest category. On the other hand, if HMO is chosen as the reference category, then BLUE might not be significant because an intermediate and more similar category has been chosen as the reference. The reference category always should be presented clearly along with the results. In one paper [5] the reference category had to be guessed at by eliminating other categories. Even then some knowledge of the subject matter was required.

Significance of Constructs Versus Significance of Dummies

Although the significance of the construct is independent of the reference category, the construct can be significant with or without significant dummies. In particular, it is possible to have some dummies significant with a nonsignificant construct, and it is also possible to have all dummies nonsignificant with a significant

construct. Significance of some dummies without significance of the construct can occur if the mean value of the dependent variable for some dummies is quite different from the mean for the reference category, but the R^2 explained by the construct is small compared to the number of dummies used. Note that in equation (3) the change in R^2 due to the construct, $R^2_{k+m} - R^2_m$, is divided by k, the number of dummies used. Thus, it is possible to have a striking increase in R^2 for one or more dummies, which is then greatly diluted by an unimpressive increase in R^2 from other dummies. In this case, the increase in R^2 per dummy is small. This suggests using as few dummies as appropriate to represent a construct when the effect of the construct is not expected to be large.

In the second case, all dummies can be nonsignificant with a significant construct when the number of dummies is large enough that a smaller set of dummies would provide almost as much information as the full set. In this case, the removal of any one dummy variable would make little change in the R^2; i.e. that dummy would be "not significant." However, the effect of removing all dummies (i.e., effect of the construct) would be large and significant. A general discussion of significance of dummies vs. significance of constructs in given by Cramer.[12]

Dependence Among Dummy Variables

Intercorrelation among the independent variables in a regression equation causes problems in the interpretation of the regression coefficients. This is true with any type of independent variable; however, intercorrelation among dummies takes a more obvious form than dependence among other types of variables. For example, consider the pair of variables HMO and NONE as defined earlier, along with the pair AGE and INCOME. Notice that any values of AGE and INCOME might occur together, although some combinations would be more likely than others. In the case of HMO and NONE, however, the combination HMO = 1 and NONE = 1 (an HMO member without health insurance) is impossible. Hence, HMO and NONE have a specific logical dependence that is not found among continuous variables, so that the discussion of HMO requires acknowledgment of NONE. The simplified idea of testing for the effect of HMO "while holding NONE constant" represents an impossible situation and a more careful interpretation of the significance tests is required. Analysis of the effect of AGE also may need to take account of INCOME, but the dependence is not as obvious.

Interpretation of the Coefficients

Another difference between dummies and other variables involves the meaning of the coefficients. For example, if INCOME is dropped from a regression model including AGE, then the coefficient for AGE (although numerically different) still retains its interpretation, namely, the way the number of years since birth affects the dependent variable. But if NONE is dropped from a model that includes HMO, then the coefficient for HMO changes its meaning substantially, as follows. (Again, suppose that the dependent variable is VISITS and that HMO, BLUE and NONE are as defined earlier).

Case	Meaning of Coefficient of HMO
1. Regression model includes both HMO and NONE.	The average difference in VISITS between HMO members and Blue Cross members.
2. Regression model includes HMO but omits NONE.	The average difference in VISITS between HMO members and the combination of uninsured persons and Blue Cross members.

Thus, the coefficient of HMO has a qualitatively different meaning in the two

cases. Also, the significance level of HMO in case 1 refers to a test of the model with both NONE and HMO versus a model with only NONE. Such a test, in effect, determines whether it is appropriate to combine the BLUE and the HMO categories as having a similar number of visits. This implication of the significance test is not at all intuitive, and the importance of thinking carefully about any statements about individual dummies must be stressed. Also note that in stepwise regression, the meaning of a dummy that already has entered will change at each subsequent step where a new dummy for the construct enters.

Once a construct is shown to be statistically significant, the investigator may wish to test whether the mean of the dependent variable is significantly different between categories of the construct. For instance, 1) are HMO and Blue Cross enrollees significantly different? 2) Are HMO members significantly different from the combined Blue Cross and uninsured subjects? 3) Are HMO members significantly different from the uninsured persons? As noted in the previous paragraph, the significance level of the HMO variable in the regression when NONE also is included (case 1) provides the answer to the first question. Similarly, the significance level of HMO when NONE is not included in the regression (case 2) answers the second question. Thus, a judicious choice of the reference category, combined with control over the order in which dummy variables are allowed to enter the regression equation, will allow the automatic testing of some category differences. To answer the third question, we would need to rerun the regression, using the uninsured persons as the reference category. A side note to this procedure is that when a large number of comparisons are made between categories, the p values may be incorrect if no adjustment is made for the number of comparisons. This "multiple comparisons" problem is beyond the scope of this article.

In a standard regression program we cannot test for significant differences among all possible categories in a single computer run since the necessary information is not produced. However, we can always estimate the magnitude of the difference in the mean of the dependent variable between any categories (adjusted for other regression variables) by subtracting the coefficients for the dummies of the two categories. For example, in the case 1 regression above, $\hat{B}_1 - \hat{B}_2$ would estimate the difference in VISITS between HMO enrollees and uninsured persons, adjusting for other variables in the equation.

If a large number of group means are to be compared, it may be more convenient to use an analysis of covariance program such as BMDP1V[13] which will test for differences between all the categories (or combinations of categories) of interest.

Testing Collections of Variables

The same procedure for testing a construct can be applied to any other collection of variables. Suppose, for example, that the construct "age" is represented in a regression model by the variables AGE and AGESQ = AGE.[2] The significance of the construct "age," represented here in a quadratic form, can be tested by using equation (3) with k = 2 for AGE and AGESQ.

Diverse variables can be tested together. Socioeconomic status, for example, might be represented by INCOME (current monthly income), OCCUPATION (a rank score) and EDUCATION (years completed). The combined effect of these three variables can be computed by using equation (3) with k = 3. Four of the articles reviewed[2,4,7,8] used quadratic or cubic terms of age or bed-size, yet none of these articles presented tests for the entire effect of the construct. Also, some articles referred to economic, demographic or other groups of variables, yet none of them presented tests for the effect of an entire group of variables to determine, for example, if

the economic variables significantly affected the dependent variable.

Increasing the Power of Tests

Using as few dummies as necessary to represent a particular construct generally will increase the chance of detecting a significant effect of the construct; that is, the test is more powerful. Note that in equation (3) the change in R^2 in the numerator is divided by the number of dummies used. A construct which is statistically significant when represented by one or two dummies may not achieve significance when represented by four or five dummies. Of course, fewer dummies may decrease significance if the wrong categories are combined.

If the dependent variable is related to the categories in a monotonic way and if it generally increases or decreases across the categories, then another way to gain power is to enter ordinal constructs into the regression model as a single variable rather than as a set of dummies. For example, the construct "perceived health status" with levels "excellent," "good," "fair" and "poor," can be coded as three dummy variables or as a single ordered variable with values scaled as 1, 2, 3, 4. A reduction from three dummy variables to one ordered variable would increase the power in detecting a true effect of the ordinal construct. However, different methods of scaling the ordinal construct might produce different results; a sensitivity analyis is recommended in this case with several different trial values assigned to the ordinal construct (e.g., try 1, 2, 4, 8). Such an analysis will show the effects of various types of scaling. Of course, if the relationship of the ordinal construct to the dependent variable is possible nonmonotonic or if the detailed changes of the dependent variables across construct categories are of interest, then several dummies should be used in place of a single variable.

Only one article that was reviewed entered an ordinal construct as an ordered variable and that was social class. Another paper, referred to earlier, had several ordinal constructs that could have been used in

TABLE 1. Summary of Use of Dummy Variables in Nine Papers

Type of Use	Number of Papers (out of nine)	References
Groups of nondummy variables used	1	6
Constructs not tested	8	5–12
Potentially significant constructs missed due to lack of test of construct	4	6 (Interaction of hospital and type of treatment)
		8 (Usual source of care for children, insurance coverage for children, prior treatment)
		9 (# members covered)
		12 (duration of disability)
Significance or nonsignificance of dummies treated as significance of construct	4	5, 8, 9, 12
Ordinal variables used as dummies (potential loss of power)	2	8, 12

this way[9] but were represented by dummies instead. The effect of duration of disability on self-assessed health status in this study probably would have been significant if used as a single, scaled variable. It also may have been significant as used—a construct with three dummies—but no test of the entire construct was performed.

Decisions on the number of dummies to use for a construct and whether or not to use scaled variables rather than dummies for ordinal constructs are matters of judgment. The researcher should be aware, however, that the choice will affect the power of the regression analysis to detect the true effect of the construct.

Some Examples From the Literature

Table 1 shows how the nine articles reviewed handled dummy variables. Notice that only one tested for the significance of a group of variables. Perhaps some of the other recommended procedures were performed by the authors, but were not reported in the papers. Nevertheless, it is possible that significant effects of constructs went undetected.

Conclusions

Dummy variables provide a powerful method of data analysis. The single greatest fault in their use is the lack of testing for an effect of the construct they represent. Perhaps this is a natural oversight stemming from the habit of thinking of each variable singly. What has been shown is that a construct is a tightly knit family and that the meaning of each family member must be understood in the context of the whole.

Acknowledgments

The authors would like to thank Elaine Nasco-Egnew for help in the preparation of many drafts and Elaine Eldridge, Ph.D., for editorial assistance.

References

1. Hornbrook MC. Market structure and advertising in the U.S. pharmaceutical industry. Med Care 1978;16:90.

2. Lairson DR, Swint JM. A multivariate analysis of the likelihood and volume of preventive visits demand in a group practice. Med Care 1978;16:730.

3. Mitchell JB. Patient outcomes in alternative long-term care settings. Med Care 1978;16:439.

4. Rutten FFH, van der Gaag J. Referrals and demand for specialist care in the Netherlands. Health Serv Res 1977;12:233.

5. Salkever DA, German PS, Shapiro S, Horky R, Skinner EA. Episodes of illness and access to care in the inner city: a comparison of HMO and non-HMO populations. Health Serv Res 1976;11:252.

6. Scitovsky AA, McCall N, Benham L. Factors affecting the choice between two prepaid plans. Med Care 1978;16:660.

7. Thomas DE, Stokes HH. How many beds should a hospital department serve? Health Serv Res 1976;11:241.

8. van der Gaag J, van de Ven W. The demand for primary health care. Med Care 1978;16:299.

9. Wan TTH. Predicting self-assessed health status: a multivariate approach. Health Serv Res 1976;11:464.

10. Draper N, Smith H. Applied regression analysis. New York: John Wiley and Sons, 1966.

11. Kleinbaum DG, Kupper LL. Applied regression analysis and other multivariate models. North Scituate, Mass.: Duxbury Press, 1978.

12. Cramer EM. Significance tests and tests of models in multiple regression. Am Statistician 1972;26(4):26.

13. Dixon WJ. BMDP-77. Biomedical Computer Programs. P-Series. Berkely: University of California press, 1977.