Mind Your Quants and Quals in Multiple Regression

T. Chris Oshima

John H. Neel

Georgia State University

Georgia Educational Research Association
Savannah, Georgia
November 14, 2003

oshima@gsu.edu
jneel@gsu.edu

Running head: MIND YOUR QUANTS AND QUALS

Abstract

It is basic knowledge in regression analysis that categorical or qualitative variables can be used provided that they are coded appropriately first. However, we have seen dreadful coding errors for categorical variables that have resulted in totally incorrect regression analysis. The paper will describe the process of dummy coding in simple language and demonstrate the misuse of coding schemes using a large real data set. It is our hope that this paper will raise awareness of qualitative vs. quantitative variables in regression analysis for novice researchers.

Mind Your Quants and Quals in Multiple Regression

At a dissertation defense, a student nervously presents her regression analysis used in her study. Then, a professor asks "Could you tell us how you coded the categorical variables in your study?" After a slight pause, she says "What do you mean by coding? The data were provided to me and they were already coded." Then, the professor's jaw drops in exasperation and she thinks, "Oh my goodness. She did not dummy code her categorical variables. She is not going to graduate..."

Although this is an extreme case, we have seen similar cases where categorical (or qualitative) variables are treated as if they were quantitative variables in regression analysis thus resulting in entirely incorrect interpretations (e.g., the race variable was coded 1 for White, 2 for African-American, and 3 for Asian-American instead of creating two dummy variables for race).

It is basic knowledge in regression analysis that categorical variables can be used provided that they are coded first. However, the coding schemes are generally introduced in a regression class which is often beyond the basic required research courses where multiple regression is briefly introduced. After a brief introduction of multiple regression, convinced with its usefulness, students may venture out their own analysis powered by easy-to-use SPSS programs. In SPSS, variables are simply listed,

regardless of their nature of variables (quantitative or qualitative), and SPSS will carry out the multiple regression analysis even when qualitative variables are not appropriately coded as long as they are numerically coded. Furthermore, there is no obvious "button" to push in SPSS to transform the qualitative variables into dummy variables. All these conditions contribute to the misuse of regression by novice researchers when the independent variables include some categorical variables.

The purpose of this paper is to describe the process of dummy coding in simple language and show what could happen if coding schemes are not properly applied to categorical variables. Then, it will be shown how to do the dummy coding on SPSS. In addition to a small data set for demonstrating the use of dummy coding, a large real data set will be used to demonstrate the misuse of coding schemes. It is our hope that this paper will be useful for novice researchers/students who have not had a formal training of multiple regression or for those students who are about to learn the coding schemes in their regression classes.

Demonstration of Dummy Coding Using Small Data Sets

According to Pedhazur (1997), the simplest method of coding a categorical variable is dummy coding. For those who are interested in other coding methods, see Pedhazur (1997). In dummy coding, g - 1 dummy variables are generated, where g is

the number of groups or levels within the categorical variable. For example, gender has two levels (males or females), thus g = 2. Therefore, one dummy variable containing 1 and 0 is created. Similarly, if a categorical variable has 5 levels (say, a race variable), four dummy variables are generated. In the regression analysis, all the categorical variables need to be transformed into dummy variables. We now turn to examples of dummy coding.

First, let us consider a case where a categorical variable has only two levels or groups (g = 2). Table 1 shows a very simple example. An independent-t test would show that there is a significant gender difference ($t_8 = 2.674$, p = .028). To approach the same problem in the regression framework, the gender variable is transformed into a new variable using dummy coding (z). A regression analysis would yield a regression equation y' = 4.6+ 3.8Z. The relationship between Z and score is significant ($F_{1,8} = 7.149$, p = .028) and R-square is .472. (Recall t^2 with m degrees of freedom equals F with 1 and m degrees of freedom.) The two approaches, of course, come to the same conclusion. Now, by using the dummy coding, the slope and intercept have some meanings. Here, the slope (3.8) is the difference of two means (males mean of 8.4 - females mean of 4.6) and the intercept (4.6) is the mean for the group who was assigned 0 (females). Therefore, testing the slope is equivalent to testing the mean difference of

the two groups.

Table 1
An Example When g = 2

obs	Y	Gender	Z
1	5	M	1
2	9	M	1
3	10	M	1
4	7	M	1
5	11	M	1
6	5	F	0
7	4	F	0
8	3	F	0
9	3	F	0
10	8	F	0

Now, what happens if 1 and 2 (or any other two numbers) are assigned for males and females respectively? In other words, one may have a data set where some kind of numerical coding is done and one may proceed to do the regression analysis without creating dummy coding. In the case of g = 2, the consequence is not too serious. If one's interest is obtaining and testing R-square, he or she still gets the same R-square as he or she did with dummy coding. The nice properties with dummy coding in terms of meanings of slopes and intercept would be lost. We would like to emphasize here that one would typically perform a t-test, not a regression analysis, for the data shown in Table 1. The regression analysis was applied here to illustrate dummy coding.

In a real life, it is likely that a mix of categorical and continuous variables would be included in the regression analysis. If a categorical variable has two levels (g = 2), then failing to transform it into dummy coding would not affect regression analysis that involves R-square. The story is very different if there are more than two levels (g > 2) within the categorical variable.

Table 2 shows the dummy coding scheme when g = 3. Two dummy variables Z1 and Z2 are created. As one can imagine from the previous example when g = 2, an ANOVA analysis and the regression analysis will result in the same conclusions. In the ANOVA approach, one would conclude that there is a significant mean difference for at least one pair of means ($F_{2, 12}$ = 10.670, p = .002). In the regression analysis, one would conclude that Group is a significant predictor for Y ($F_{2, 12}$ = 10.670, p = .002). The regression equation is Y' = 12.2 - 3.8Z1 - 7.6Z2 and R-square is .640. The intercept (12.2) is the mean of Group 3, the first slope (-3.8) is the mean difference between Groups 1 and 3, and the second slope (-7.6) is the mean difference between Groups 2 and 3.

Table 2

An Example When g = 3

obs	Y	Group	Z1	Z2
1	5	1	1	0
2	9	1	1	0
3	10	1	1	0
4	7	1	1	0
5	11	1	1	0
6	5	2	0	1
7	4	2	0	1
8	3	2	0	1
9	3	2	0	1
10	8	2	0	1
11	8	3	0	0
12	14	3	0	0
13	13	3	0	0
14	16	3	0	0
15	10	3	0	0

What happens then if one fails to transform the group variable into g - 1 dummy variables? The consequence is serious. On our data set, if the group variable (1,2, or 3) is used inadvertently in the regression analysis, the group variable would show no significance ($F_{1, 13} = 2.477$, $p = .140$). One would come to a totally wrong conclusion that Group is not a significant predictor of Y. One can imagine at this point if one has a mix of categorical and continuous variables and, furthermore, if one of the

categorical variables has more than two levels; if one fails to create a dummy variable for it, then the entire regression analysis would be affected seriously.

Even after the proper dummy coding is applied, one is not free from making serious errors. If the purpose of the regression analysis is to construct the best model with the fewest number of independent variables, one needs to be careful to move the dummy variables as a group. For example, if g = 3, one would have two dummy variables (Z1 and Z2). Z1 and Z2 always have to be removed together or entered together into the model. If one uses a computer-aided variable selection procedure such as stepwise regression, the resulting model may not make sense if dummy variables grouped together to represent a single categorical variable are separated in the process.

Practical Tips Using SPSS

Once a researcher understands the needs for dummy coding, the next step is how to implement it in SPSS. Although one could manually create g - 1 variables using "Transform-Compute-If" statement using the pull-down menu in the data view, we recommend using a syntax window and performing the transformation at once. The syntax is available at the following URLs: http://support.spss.com/answernet/details.asp?ID=17482 or

http://education.gsu.edu/coshima.

The syntax contains the following codes:

```
VECTOR nom(4).
LOOP #i = 1 to 4.
COMPUTE nom(#i) = (cat = #i).
END LOOP.
EXECUTE.
```

This is a template and changes need to be made to suit one's data. This template will create 4 dummy variables (nom1, nom2, nom3, and nom4) for a categorical variable called "cat" for which g = 5.

Suppose one has a categorical variable called "group" with three levels (g = 3) and wishes to create two dummy variables z1 and z2. Then, the syntax should be changed to:

```
VECTOR z(2).
LOOP #i = 1 to 2.
COMPUTE z(#i) = (group = #i).
END LOOP.
EXECUTE.
```

How to run the syntax window:

1. Download the syntax file "dummycoding.spo" from http://education.gsu.edu/coshima

on your hard drive.

2. Open the data view window in SPSS and open the data you are working on.

3. Go to File -Open-Syntax. Open the file "dummycoding.spo".

4. Make necessary changes in the syntax file and click Run-All. The new dummy variables will be calculated and inserted on the data view window with the given names.

In the next section, we will demonstrate the danger of coding errors of qualitative variables using a real data set.

Example Using a Real Data Set

This example uses data from the child file of the Early Childhood Longitudinal Study-Kindergarten Class of 1998-99 (ECLS-K) Public-Use Base Year Data Files available from the National Center for Educational Statistics. The child file contains one record for each of the 21,260 responding students. For these analyses we excluded the group whose race was not determined. In the data file each of these composite race groups is coded by a number in top to bottom order as is indicated in Table 3. We used a T-score in mathematics as the dependent variable in these analyses and we used Race or a derivative(s) of race as the independent variable as described below.

Table 3
The Original Coding Scheme

Code	Race
1	White, non Hispanic
2	Black, African American
3	Hispanic, Race Specified
4	Hispanic, Race Not Specified
5	Asian
6	Native Hawaiian, Other Pacific Islander
7	American Indian or Alaska Native
8	Multiracial

If one was to compare the means of these groups through regression, a naïve analysis would simply use the one to eight coding that is in the file. However, the proper analysis for this situation would be to create 7 dummy variables and to enter them into a regression analysis (see Kirk, 1994). Yet another naïve analysis would be to recode the data such that the racial group with the highest mean is coded 8, the next highest is coded 7, etc. We generated the recoded value to use in the analysis and termed the resulting coding scheme "reordered coding". We analyzed these scores using all three coding schemes: Naïve with original coding, naïve with ordered coding methods, and proper with dummy variables. The resulting analyses are listed in Table 4.

Table 4

Regression Results for Mathematics Under 3 Coding Schemes

Analysis	R	R^2	F for ANOVA
Naïve with given coding	.259	.067	$F_{1,19068} = 1366.4$
Naïve with ordered coding	.333	.111	$F_{1,19068} = 2370.3$
Proper with Dummy Variables	.340	.116	$F_{7,19062} = 356.598$

We can see several points illustrated in Table 4. First, note that the multiple correlations for each of the three analyses are different. The Naïve analyses depend upon the coding that is selected for the data. Who is to say that Whites should be coded as one or as eight? This determination of coding scheme is completely arbitrary with whoever creates the data set. Thus, either of the naïve analyses or one of 40,318 (8! = 40320) other schemes could result when the arbitrary coding is selected. Each of these schemes should have a different multiple correlation. Second, the naïve ordered coding resulted in a multiple correlation that was very close to that obtained from the proper dummy coding scheme. This will occur when the plot of means versus group number is close to a straight line. This is a result of our decision to use a coding scheme that numbers the groups from lowest to highest on the dependent variable. In most other naïve coding schemes, this will not occur and there will be a difference between the naïve coding scheme results and the proper dummy coding scheme results. The third observation is to note the large difference in F values. In our particular case,

all F values are significant due to the large sample size. The fact that there is a difference at all tells us that the naïve analysis will give incorrect results.

Summary

We have shown that improper or naïve coding of categorical variables can lead to incorrect results in multiple regression and in performing analysis of variance through regression. We believe that naïve coding is usually done when a researcher doesn't recognize its arbitrariness. The result of using such a coding scheme is to give incorrect results. These results are meaningless due to the arbitrary nature of the coding. The results do not correspond to any normal analysis and thus the results are themselves arbitrary; which is a poor state and one in which we hope researchers do not find themselves.

References

Early Childhood Longitudinal Study of the Kindergarten Class of 1998-99 [Data Base on CD-Rom. (2001). Washington, D.C.: National Center for Educational Statistics [Producer and Distributor].

Kirk, R. (1994). Experimental Design: Procedures for Behavioral Sciences. (3rd ed.). Belmont, CA: Wadsworth Publishing.

Pedhazur, E.J. (1997). Multiple regression in behavioral research: Explanation and prediction (3rd ed.). Fort Worth, TX: Harcourt Brace College Publishers.