Paying Attention to the Default Reference Category in Several SPSS Statistics Procedures: An Example of Coding Reversal

Hongwei Yang
University of Kentucky

This paper reviews two common approaches to handling categorical predictors in regression analysis and how they are implemented in several Statistical Package for the Social Sciences (SPSS) procedures. Through this review, the aim is to revisit a warning from the SPSS literature regarding dummy coding for categorical predictors with just two classes. In this problem, the original coding of a two-category dummy variable is reversed by the software program without any alert and such a change is likely to result in an incorrect interpretation of the regression coefficient estimate of the corresponding categorical predictor. Further, the paper discusses the generalizability of this conclusion to several other statistics programs: Statistical Analysis System (SAS), JMP, and STATA.

In regression analysis, categorical predictors are very commonly seen. A categorical predictor is usually presented as a single variable in one of two formats: 1) A string variable, and 2) a numeric variable. Here, the most commonly used two-class categorical predictor gender is used as an example. When it is presented as a string variable, its values are in the form of letters: male and female, for example. On the other hand, when it is presented as a numeric variable, its values are usually in the form of allocated (numeric) codes (Kutner, Nachtsheim, & Neter, 2004). The allocated codes are arbitrarily selected, and could be any sets of numbers. For the two values of the gender variable, possible sets of codes are (-1) and (-2), 0 and 1, 2 and 3, 99 and 100, 999 and 1000, etc. In fact, as long as two numbers are distinct from each other, they could be allocated to represent the two different values of the gender variable.

Although a categorical predictor could be presented as either a single string variable or a single numeric variable, in many cases neither format could be directly used as an input in a regression model without properly performing further coding of those (string/numeric) categories. For a string variable, the underlying reason is obvious because its values are non-numeric. For the other approach where a categorical predictor is allocated arbitrarily selected numeric codes, the problem is that those codes define a metric for the categories of the predictor that may not be reasonable. This is because the spacing of categories indicated by those allocated codes may not be in accord with the reality (Kutner et al., 2004). With that described, in order for a categorical predictor to be used as a direct input in a regression model, it should be properly recoded. And this process is known as dummy coding.

When dummy-coding a categorical predictor, the most common scheme is the 0-1 coding. Within the 0-1 scheme, for a categorical variable with c categories, a total of (c - 1) dummy/indicator variables are needed with each one representing one category of the predictor. The only ignored category for which there is no dummy variable created is the reference level or base level. All other coded categories are compared with the reference category in terms of the average change in the outcome when it moves from the reference category to that particular non-reference category. Unlike a single categorical predictor in either string or numeric format, its dummy-coded indicator variables can be used as direct inputs for a regression model.

The way the 0-1 coding scheme is implemented in the software Statistical Package for the Social Sciences (SPSS) varies from one regression procedure to another. Related procedures can generally be classified into two categories: 1) Those that are not capable of automatically creating dummy variables, and 2) those that have this particular capability. The REGRESSION procedure is an example from the first category, whereas the GENLIN, LOGISTIC REGRESSION, NOMREG procedures, etc. belong to the second category.

The paper focuses on the use of a categorical predictor in those SPSS procedures that can automatically perform dummy coding. A procedure in this category has to be informed of the categorical nature of a predictor before it automatically recodes the variable into one or more dummy variables in the background. For a categorical predictor presented in a string format, this is not an issue at all because its values are non-numeric. SPSS identifies all such non-numeric variables as categorical without having to

be told. However, when a categorical predictor is presented in numeric format using allocated codes, things could become more complex.

A particular confusing case is when the allocated codes for a two-category binary predictor are selected to be identical to its 0-1 dummy codes and, at the same time, the predictor is still specified as categorical in an SPSS procedure with an automatic dummy coding capability. Norusis (2003) provides a warning on this issue, saying that there is nothing to be gained by declaring such a predictor as categorical and since such a specification prevents the original coding (already in 0-1 format) from being preserved, it is not a recommended practice. However, some, like Field (2009), think that declaring a two-class binary predictor as categorical or non-categorical in a SPSS procedure should not make any difference when the variable is already coded as 0 and 1, which clearly violates this warning. Considering such an unfortunate fact, this paper elaborates on this warning and uses an example to demonstrate it with the hope of helping practitioner comprehension related to the issue.

As is known, numeric values associated with a categorical variable are often coded by the researcher in a manner that does not reflect substantive meaning. They are different from numeric values of a non-categorical predictor that are actual measures of an attribute. Unfortunately, SPSS is not capable of distinguishing the former from the latter without additional information from the outside. So, a SPSS procedure with an automatic dummy-coding capability needs to be informed of the categorical nature of a predictor before the procedure generates dummy indicators for it. When a single numeric categorical variable is entered into a SPSS procedure, it should usually be specified either as a factor (e.g., in the GENLIN procedure) or as a categorical covariate (e.g., in the binary logistic procedure). According to the default settings (although these default settings could be overridden or controlled in many cases), SPSS will then identify the "last" level of this predictor as the reference level and create dummy variables for all other categories of this predictor. By default, the "last" level is defined in ascending (from lowest to highest) order of the alpha-numeric coding. So, the highest numeric coding is the default "last" level, and it corresponds to the default reference category selected by a SPSS procedure with an automatic dummy coding capability (SPSS Inc., 2010).

Suppose the two-class categorical predictor gender is presented using two allocated codes: 99 (male) and 100 (female). After specifying this variable either as a factor or as a categorical covariate, SPSS will create (2-1) = 1 dummy variable for this categorical predictor. By default, it first identifies 100 as the last level of this predictor because 100 (for female) > 99 (for male), then it specifies this level (female) as the reference level by assigning values of 0 to the dummy variable, while the other level (male) is coded as 1. It is this newly created dummy variable for gender that is used in the regression model. Furthermore, it is this dummy variable (not the original gender variable) that has a regression coefficient to estimate. The interpretation of the regression coefficient estimate for this dummy-coded gender predictor variable should be made for the male category as is compared with the female category (i.e., reference level).

With the allocated codes of 99 and 100 described for the gender variable, this coding process and the final dummy coding result should remain the same regardless of its allocated codes: As the default option, the last level or, in ascending alpha-numeric order, the highest coding is selected as the reference level before the dummy variable is created for the other level of the two-class categorical predictor. This is even true when the allocated codes are 0 and 1; the same two values that are used in the 0-1 dummy coding scheme.

Suppose that the gender variable is originally allocated two numeric codes: 0 for male and 1 for female. This is similar to the previous example where 99 was allocated to male and 100 to female in the sense that the male category is always allocated the lower coding (i.e., 99 and 0, respectively), whereas the female category is allocated the higher coding (i.e., 100 and 1, respectively). For the second case, after this gender variable is designated as categorical in a SPSS regression procedure with an automatic dummy-coding capability, the procedure identifies the last level or by default, in ascending alpha-numeric order, the higher code of 1 (i.e., the female category) as the reference level and assigns values of 0 to the new dummy variable to be internally computer-generated. The other (non-reference) level of the dummy variable will be created for the other category (i.e., the male category) that is originally allocated the code of 0, so that the male category is now coded as 1 in the newly-generated dummy variable. So, in this new dummy variable, the coding of the gender variable is reversed from the original coding scheme. The dummy variable can then be used as a direct input for a regression model. When it comes to parameter

interpretation of the dummy variable, it should be made for the male category (i.e., coded as 1 in the new variable) as is compared with the female category (i.e., coded as 0 in the new variable).

Such a change in coding may be difficult to spot for practitioners of regression analysis, particularly those who are not familiar with SPSS procedures that can perform dummy-coding automatically. When a practitioner who knows about the 0-1 coding intentionally allocates 0 to male and 1 to female because the interest is in the female group rather than the male group, the above-described coding reversal in some SPSS procedures with an automatic dummy-coding capability causes the final parameter estimate to focus on the male group, instead. When this coding change goes unnoticed, the interpretation of the parameter estimate is very likely to be made still regarding the female group to stay consistent with the original coding, which of course is incorrect. An example follows that demonstrates this point.

Heuristic Example

The data analyzed here as an example comes from Kutner et al. (2004). The data are results from an economist who studied 10 mutual firms and 10 stock firms. The economist was most interested in the relationship between the elapsed time for the innovation to be adopted (Y), size of firm (X_1), and type of firm (X_{Type}). Type of firm (X_{Type}) is a two-class categorical predictor, so it has to be recoded into a numeric variable X_2. Y is expressed in number of months and X_1 in millions of dollars.

Figure 1 presents a screenshot of the data set that has a total of five columns. In the data set, the first two columns are the dependent variable (*ElapsedTime_y*) and one of the two predictors (*X1_Size*). They are not categorical, so no special recoding is needed for any of them. The next 3 columns are all about the other predictor variable; type of firm. The column called *X2_Type* presents the predictor variable in string format. And the other two columns present this predictor in numeric format with the column called *X2_Type99100* using the allocated codes of 99 and 100 and the column called *X2_Type01* using the allocated codes of 0 and 1. In both numeric columns, the lower code (99 and 0) is assigned to the mutual category and the higher code (100 and 1) is assigned to the stock category. Two distinct sets of allocated codes are used here for the purpose of comparing the final modeling results. It is anticipated that, if done properly, the results should be the same because the values allocated to represent company type are just numeric codes without any substantive meaning.

StockMutual.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

23 :

	ElapsedTime_y	X1_Size	X2_Type	X2_Type99100	X2_Type01	var
1	17	151	Mutual	99	0	
2	26	92	Mutual	99	0	
3	21	175	Mutual	99	0	
4	30	31	Mutual	99	0	
5	22	104	Mutual	99	0	
6	0	277	Mutual	99	0	
7	12	210	Mutual	99	0	
8	19	120	Mutual	99	0	
9	4	290	Mutual	99	0	
10	16	238	Mutual	99	0	
11	28	164	Stock	100	1	
12	15	272	Stock	100	1	
13	11	295	Stock	100	1	
14	38	68	Stock	100	1	
15	31	85	Stock	100	1	
16	21	224	Stock	100	1	
17	20	166	Stock	100	1	
18	13	305	Stock	100	1	
19	30	124	Stock	100	1	
20	14	246	Stock	100	1	
21						

Figure 1. A Screen Shot of Data.

To concisely present the information without having to burden the readers with unnecessary details, a simple, first-order regression model is fitted here that does not contain any complex terms like two-way interactions:

$$Mean\ Y = \alpha + \beta_1 * X_1 + \beta_2 * X_2 \quad (1)$$

where Y is the number of months for the innovation to be adopted, X_1 is the (non-categorical) company size variable measured in millions of dollars, and X_2 is the categorical company type variable. This company type variable could take one of the two forms: 1) A string variable in the form of letters (e.g., *X2_Type* in Figure 1), and 2) a numeric variable with allocated codes (e.g., *X2_Type99100*, and *X2_Type01* in Figure 1).

This model is estimated in SPSS in four different ways using two different procedures. Each way of fitting the model features a different approach to the categorical predictor representing type of company. The two procedures used here are the REGRESSION procedure and the GENLIN procedure. They are briefly compared below:

- The former procedure is designed for multiple linear regression based on general linear models. The procedure requires all inputs should be numeric, because it is not capable of automatically handling a categorical predictor in string format. Additionally, this procedure can analyze dummy indicators and/or numeric values of a non-categorical predictor that are actual measures of an attribute, but it does not allow the use of allocated codes representing classes of a categorical predictor. The only exception is when the allocated codes for the classes of a binary categorical predictor are selected to be the same as the dummy codes for that predictor.
- The latter procedure is designed for multiple regression based on generalized linear models that incorporate the type of models for the previous procedure as a special case. Not only is this procedure capable of handling a broader family of regression models, but it is also able to do the two things that the previous procedure cannot do: 1) Handling categorical predictors in non-numeric format, and 2) Handling allocated codes by automatically converting them to dummy codes in the form of 0 and 1.

With that described about the two procedures, they are used to estimate Equation 1 in four different ways. During the four analyses, the REGRESSION procedure uses the *X2_Type01* variable whereas the GENLIN procedure uses each of the three variables (*X2_Type*, *X2_Type99100*, and *X2_Type01*) as predictors in each of three regressions.

Note that when using the GENLIN procedure, both *X2_Type99100*, and *X2_Type01* are entered as Factors under the Predictors tab. This is so because this paper assumes the following: It is ***intuitive*** for a practitioner to think that either one of the two (*X2_Type99100*, and *X2_Type01*) has a categorical nature because it represents a categorical predictor: Type of company and, with that thinking in mind, he or she is likely to tell SPSS about the belief by entering either of the two into the program as a factor. Figures 2 and 3 are screenshots of the Predictors tab when entering *X2_Type99100*, and *X2_Type01*, respectively. In the interest of space, the screenshots of the other two (more straightforward) analyses are omitted from here.

The modeling results from all four analyses are presented in Table 1. The focus is on the parameter estimate for the predictor representing company type. As is indicated, all four analyses have produced similar results. The four estimates for the parameter of the corresponding predictor indicating company type are identical in absolute value. It is just that the parameter estimate from the REGRESSION procedure is (+8.055) whereas all others are (-8.055). Such a difference is due to the fact that the two procedures by default use different reference levels. The

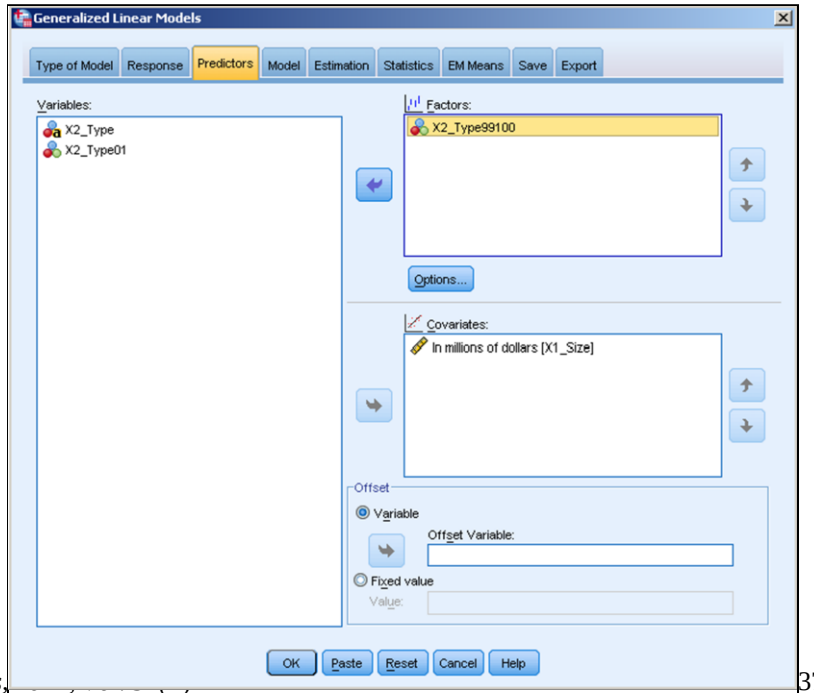

Figure 2. A Screenshot of Entering X2_Type99100.

REGRESSION procedure uses *X2_Type01*=0 (mutual companies) as the reference level and the parameter estimate for the company type variable measures the change in the outcome for stock companies as is compared to mutual companies. The GENLIN procedure does exactly the opposite by default because it uses the *X2_Type01*=1 (or *X2_Type99100*=100, *X2_Type* = Stock) as the reference level, or the last level given an ascending alpha-numeric order of categorical values. So, the parameter estimate for the company type variable of each analysis under the GENLIN procedure measures the change in the outcome for mutual companies as is compared to stock companies. Because the direction of change for the company type predictor is opposite to each other under the two procedures, the estimates for the corresponding parameter have opposite signs, but are identical in absolute value.

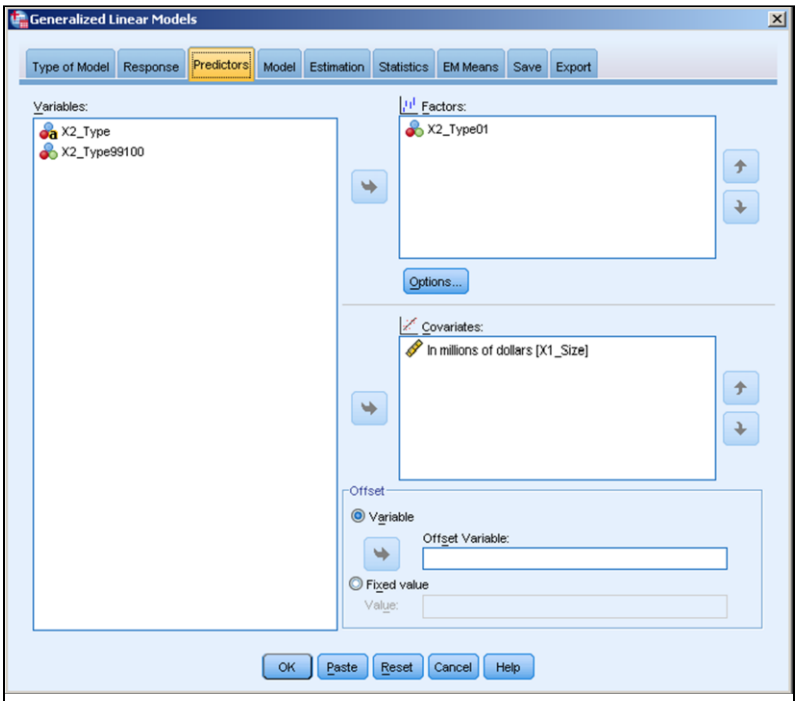

Figure 3. A Screenshot of Entering X2_Type01.

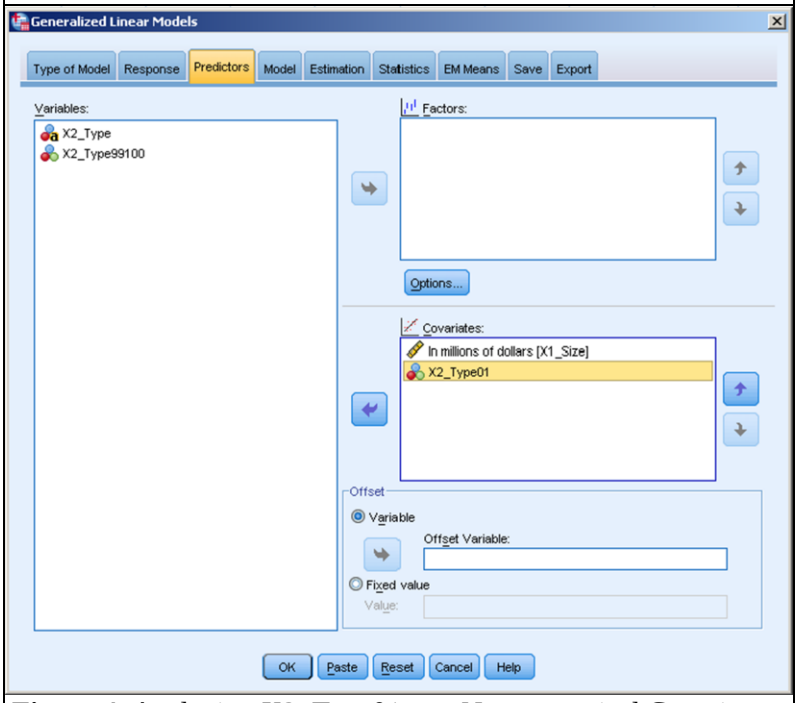

Figure 4. Analyzing X2_Type01 as a Noncategorical Covariate.

Special attention should be paid to the first and fourth analysis. In the first analysis, *X2_Type01* is analyzed in the REGRESSION procedure as an input for the Independent(s) box. In the fourth analysis, the *same* variable is analyzed in the GENLIN procedure (using its default settings) as an input in the Factors box under the Predictors tab. Although it is the same predictor (*X2_Type01*) that is used in both analyses that aim to fit the same regression model as is described by Equation 1, the parameter estimates are exactly opposite to each other due to the reasons outlined above: ((+8.055) in the first analysis versus (-8.055) in the fourth analysis). Both estimates are correct and are equivalent of each other, but they should be interpreted from different perspectives. That is, (+8.055) indicates that, on average, stock companies need 8.055 more months than mutual companies in adopting an innovation whereas (-8.055) suggests the average amount of time mutual companies take to adopt an innovation is 8.055 months less than stock companies.

In fact, it would have been unnecessary to declare *X2_Type01* as categorical because its allocated codes have already been selected to be identical to its dummy codes. In such a case, this variable could be just used as a (non-categorical) covariate in the GENLIN procedure. Figure 4 provides another analysis of the model in Equation 1 using the GENLIN procedure. In this analysis, the *X2_Type01* variable is entered

into the same box as the *X1_Size* variable, which is different from analyses 2, 3, and 4. This time, the parameter estimate for the company type variable becomes (+8.055), the same result as analysis 1. In this case, the interpretation should be made consistently, as noted with the original coding of the *X2_Type01* variable, because no coding reversal has been done.

Another point that is worth noting is, as long as the order of the coding remains the same, the choice of allocated codes for classes of a categorical predictor does not affect the parameter estimates. The third and the fourth analyses use different allocated codes for the two categories of the respective company type variable, but their parameter estimates are the same: (-8.055).

Table 1. Analysis Results from Two SPSS Procedures

	SPSS Statistics Procedure Used			
Items	**REGRESSION**	**GENLIN**		
Analysis	Analysis 1	Analysis 2	Analysis 3	Analysis 4
Variable	*X2_Type01*	*X2_Type*	*X2_Type99100*	*X2_Type01*
Data type	Numeric	String	Numeric	Numeric
Estimate	**(+8.055)**	**(-8.055)**	**(-8.055)**	**(-8.055)**
Base level	*X2_Type01*=0	*X2_Type*=Stock	*X2_Type99100*=100	*X2_Type01*=1

Discussion

The paper focuses on the coding reversal issue that happens to binary categorical predictors in some SPSS procedures that have an automatic dummy coding capability. The paper alerts practitioners of regression analysis to this issue, particularly when the allocated codes for the binary predictor are selected to be the same as the codes for the 0-1 dummy coding scheme. Although there is nothing wrong to think of this binary predictor as categorical in this situation, it requires special attention to find out what category it is that is being used as the reference level by SPSS if the predictor is indeed declared as categorical in the computer program.

When analyzing a two-class categorical predictor that has already been dummy-coded in a SPSS procedure with an automatic dummy-coding capability, the best strategy is not to let the program recode it again. To prevent the program from performing the recoding automatically, the categorical predictor under discussion should be used as a (non-categorical) covariate but not as a factor or a categorical covariate.

With the conclusion drawn based on SPSS, it may also be generalized to other statistics programs. Like SPSS, a few Statistical Analysis System (SAS) procedures by default take a similar approach to a declared categorical predictor; selecting its last category in ascending alpha-numeric order as the reference level, coding it into 0, and using 1 to represent all other non-reference categories. Therefore, the aforementioned analyses 2 to 4 where a two-level predictor is declared as categorical can be duplicated using such SAS procedures that include PROC GLM and PROC GENMOD, which allow the specification of a predictor as categorical using the CLASS statement (SAS Documentation, 2010). In these three analyses using either SAS procedure, the issue of coding reversal as described in this paper also exists. Analysis 1 where a two-level categorical predictor presented in 0 and 1 is analyzed as a non-categorical covariate can be duplicated using either of the above SAS procedures (without declaring the predictor as categorical) or using another one called PROC REG; the counterpart of the REGRESSION procedure in SPSS.

However, there are also procedures in SAS that work differently than PROC GLM or PROC GENMOD, and among them is PROC LOGISTIC for logistic regression analysis. With PROC LOGISTIC, after declaring a two-level predictor already in the form of 0 and 1 as categorical, the interpretation of its parameter estimate should be made relative to the average effect across both levels rather than relative to an internal, computer-generated 0 category that in fact does not exist with PROC LOGISTIC. This is so because PROC LOGISTIC uses a different dummy coding scheme than the other two procedures. Whereas PROC GLM and PROC GENMOD use the 0-1 coding, PROC LOGISTIC performs the (-1)-1 coding where the last category in ascending alpha-numeric order is coded as (-1), instead of 0 (SAS Documentation, 2010). Further, the last category is no longer the reference level with which the other category is compared. It is the average effect across both levels of the categorical

predictor that serves as the baseline of comparison for its two categories. Another statistics program that handles a declared categorical predictor in the same manner as PROC LOGISTIC in SAS is JMP (SAS Institute Inc., 2008). Several of JMP's regression modules (like the Fit Model module) by default also perform the (-1)-1 coding and code into (-1) the last category in ascending alpha-numeric order. Although the issue around the (-1)-1 coding for a declared categorical predictor in a statistics program does not quite fall into the coding reversal issue as discussed in this paper, both cases are likely to cause the original coding of the categorical predictor to change without any alert. Therefore, cautions should be taken when interpreting such parameter estimates.

Finally, the last statistics program that should be discussed here is STATA because it is almost as comprehensive and popular as SPSS and SAS. STATA provides a xi command that is capable of using the 0-1 coding scheme to automatically create dummy variables that can next be analyzed by the regress command for regression modeling. Unlike SPSS, whose default setting in many of its procedures is to pick up the last category in ascending alpha-numeric order as the reference, the xi command in STATA by default does exactly the opposite by selecting the first category as the reference (Hamilton, 2004). Therefore, the coding reversal issue for a declared two-level categorical predictor already in the form of 0 and 1 as described in the paper does not exist with the xi command in STATA.

References

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). Thousand Oaks, CA: Sage Publications Inc.

Hamilton, L. C. (2004). *Statistics with STATA: Updated for version 8*. Belmont, CA: Thomson Learning.

Kutner, M., Nachtsheim, C., & Neter, J. (2004). *Applied linear regression models* (4th ed.). New York: McGraw-Hill.

Norusis, M. J. (2003). *SPSS 12.0 statistical procedures companion*. Chicago: SPSS, Inc.

SAS Documentation. (2010). *SAS/STAT 9.22 user's guide*. Cary, NC: SAS Institute, Inc.

SAS Institute Inc. (2008). JMP (Version 8) [Compute software]. Cary, NC: SAS Institute Inc.

SPSS Inc. (2010). *IBM SPSS Advanced Statistics 19*. Chicago, IL: SPSS, Inc.

Send correspondence to:
Hongwei Yang
University of Kentucky
Email: hya222@uky.edu