StatNews #72

Coding Categorical Variables in Regression Models: Dummy and Effect Coding May 2008

In this issue of *StatNews*, we explore methods for incorporating categorical variables into a linear regression model. We offer examples of the application of these methods and tips for using them in statistical software.

In linear regression, the relationship between a dependent variable, *Y*, and an independent variable *X*, is modeled by $Y = a+\beta X$. The slope of the line, β, reflects the change in the dependent variable *Y* as a result of a one-unit change in the independent variable *X*. This interpretation of the slope, or coefficient, in linear regression is appropriate only when the independent variable is continuous (quantitative). How can we incorporate categorical (qualitative) independent variables into the regression model and formulate the model so the variables have interpretable coefficients? There are two commonly used methods for coding categorical variables so they can be used in regression models, *dummy coding* and *effect coding*.

Suppose we are investigating the effects of a vitamin supplementation treatment for pregnant women on the birth weight of infants. Expectant mothers were randomly assigned to 3 levels of vitamin supplementation, A, B and C, and birth weights of infants were subsequently recorded[1]. Simply adding treatment as a predictor in linear regression will not yield a meaningful result. The coefficient will reflect the change in infant birth weight as a result of a "unit change" in treatment; however, the "unit change" is not well defined in this case, because each vitamin supplementation treatment level can have a different effect (in magnitude) on infant birth weights.

One way to code the treatment variable so the linear regression methodology can be used to predict the infant birth weight is by using **dummy coding**, which assigns values "1" and "0" to reflect the presence and absence, respectively, of a treatment level. Since in this example we have 3 treatment levels, we need two dummy variables so each level is uniquely defined by combining the two dummy variables; these will be the predictors of the regression model. Each dummy variable will be compared to the reference level, which will be coded as "0" for both dummy variables. In this case, for example, treatment C can be considered as the reference level by the coding:

Treatment	Dummy1	Dummy2
A	1	0
B	0	1
C	0	0

Fitting the linear regression yields: $Y = 3201.50 - 574.83\ (D1) - 375.16\ (D2)$

The intercept is equal to 3201.50. This is the mean birth weight of the reference group (infants whose mother had treatment C). Note that this is so because the reference group is defined when both *D1* and *D2* are equal to zero. The coefficient of *D1* (– 574.83) reflects that the mean birth weight of the infants for the treatment A group is 574.83 less than the mean for the reference group and so is equal to 2626.67 (= 3201.5 – 574.83). The coefficient of *D2* (–375.16) reflects that the mean birth weight of infants for the treatment B group is 2826.34 (= 3201.5 – 375.16).

[1] Data Source : Wayne W. Daniel : Biostatistics, A Foundation for Analysis in the Health Sciences

Another way to represent a categorical variable in a regression model is by using **effect coding**. With effect coding the variables take values "-1", "0" and "1". One level will be coded as "–1" for both variables.

Treatment	Effect1	Effect2
A	1	0
B	0	1
C	–1	–1

The intercept in this case represents the mean of the means of the 3 treatment levels. When the data are balanced, this is the same as the overall mean, but this is not the case when the data are unbalanced (that is, when there are unequal numbers of observations in each group). The coefficients of the effect variables in the regression model signify the deviations of the mean of each level from this mean.

Fitting linear regression with effect coding to the above example yields:

$$Y = 2884.83 – 258.16\ (E1) – 58.5\ (E2)$$

The intercept of 2884.83 is the overall mean of infant birth weight since the data are balanced in this example. The regression coefficient of *E1* is –258.16 and represents the deviation of the mean of treatment A group from the overall mean. The mean birth weight for the treatment A group is 2626.67 (= 2884.83 – 258.16). Similarly, the coefficient of *E2* (–58.5), gives the deviation of the mean of treatment B group from the overall mean. The mean birth weight for the treatment B group is thus 2826.33 (= 2884.83 – 58.5). The deviation from the overall mean birth weight of the treatment C group is 316.66, which results from the model when *E1* and *E2* are both equal to –1: (– 258.16(–1) – 58.5(–1) = 258.16 + 58.5 = 316.66). The mean birth weight for the treatment C group is 3201.5 (= 2884.83 + 316.66). Note that these results are the same as those obtained from the model using dummy coding.

When using statistical software, it is important to know: (1) what the default coding method is for categorical variable in regression models and (2) which level is being treated as the reference level. SAS, STATA, SPSS and R, for example, use dummy coding, whereas JMP uses effect coding by default. In SAS and SPSS by default, the last level (of the alphanumeric order) of the categorical variable is considered as the reference level; in STATA and R it is the first level. In JMP by default it is the last level which will be coded -1 for each effect variable.

If you need help with the coding, analysis and interpretation of categorical variables in regression, do not hesitate to contact a statistical consultant at the Cornell Statistical Consulting Unit.

Author: Resmi Gupta

(This newsletter was distributed by the Cornell Statistical Consulting Unit. Please forward it to any interested colleagues, students, and research staff. Anyone not receiving this newsletter who would like to be added to the mailing list for future newsletters should contact us at cscu@cornell.edu. Information about the Cornell Statistical Consulting Unit and copies of previous newsletters can be obtained at http://www.cscu.cornell.edu).