Categorical Variables in Regression: Dummy and Effect Coding

Statnews #72

Cornell Statistical Consulting Unit

Created May 2008. Last updated September 2020

Introduction

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 $Y = \alpha + \beta X$. The slope of the line, β, reflects the change in the dependent variable associated with a one-unit change in the independent variable. 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. These two coding schemes will be illustrated below.

Categorical independent variables

Suppose we are investigating the effects of a vitamin supplementation treatment for pregnant women on the birth weight of infants (this example is drawn from Daniel (1978)). Expectant mothers were randomly assigned to 3 levels of vitamin supplementation, A, B and C, and birth weights of infants were subsequently recorded. We cannot simply include the 3-level categorical treatment variable as the independent variable X in the equation $Y = \alpha + \beta X$. The coefficient β is supposed to reflect the change in infant birth weight as a result of a "one-unit change" in X; however, the "one-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.

Dummy coding

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 the values 1 and 0 to

indicate the presence and absence, respectively, of a specific treatment level. Since in this example we have 3 treatment levels, we need two dummy variables. Each treatment level is uniquely defined by combining the two dummy variables, and these dummy variables become the predictors in the regression model.

With 3 treatment levels, we can define the following 2 dummy variables: Dummy1 is equal to 1 if the treatment is A, and is equal to 0 otherwise; Dummy2 is equal to 1 if the treatment is B, and is equal to 0 otherwise. When both of these variables are equal to zero, then the treatment must be equal to C. Treatment C is the "reference level", the level indicated when all dummy variables are equal to zero:

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

Fitting the linear regression yields $Y = 3201.50 - 574.83 \times \text{Dummy1} - 375.16 \times \text{Dummy2}$.

The intercept is equal to 3201.50. This is the mean birth weight of the reference group (infants whose mother had treatment C). As described above, this is because the reference group is indicated when both dummy variables are equal to zero. The coefficient of Dummy1, -574.83, indicates that the mean birth weight of the treatment A group is 574.83 units less than the mean for the reference group. Thus, the mean of the treatment A group is $2626.67 = 3201.5 - 574.83$. The coefficient of Dummy2, -375.16, indicates that the mean birth weight of the treatment B group is $2826.34 = 3201.5 - 375.16$.

Effect coding

Another way to represent a categorical variable in a regression model is by using effect coding. With effect coding the indicator variables for the treatment in the regression model have the values -1, 0 or 1. One treatment level will correspond to all indicadtor variables having the value -1. In our example, effect coding looks like this:

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

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

$Y = 2884.83 - 258.16 \times \text{Effect1} - 58.5 \times \text{Effect2}$

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 represent the deviations of the mean of each level from this overall mean.

In this example, the data are balanced and the intercept of 2884.83 is the overall mean of infant birth weight. The regression coefficient of Effect1 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 Effect2 is -58.5 and 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 is calculated from the model when Effect1 and Effect2 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.

Dummy and effect coding in tatistical software

When using statistical software, it is important to know: (1) what the default coding method is for categorical variables 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, the last level, based on alphanumeric order. of the categorical variable is considered as the reference level by default. In STATA and R it is the first level. In JMP by default it is the last alphanumeric level which will be indicated by all effect variables equal to -1. 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

Reference: Daniel, W. W. (1978). Biostatistics: a foundation for analysis in the health sciences. Wiley.