The Use of Dummy Variables in Regression Analysis

By Smita Skrivanek, Principal Statistician, MoreSteam.com LLC

What is a Dummy variable?

A **Dummy variable** or **Indicator Variable** is an artificial variable created to represent an attribute with two or more distinct categories/levels.

Why is it used?

Regression analysis treats all independent (X) variables in the analysis as numerical. Numerical variables are interval or ratio scale variables whose values are directly comparable, e.g. '10 is twice as much as 5', or '3 minus 1 equals 2'. Often, however, you might want to include an attribute or nominal scale variable such as 'Product Brand' or 'Type of Defect' in your study. Say you have three types of defects, numbered '1', '2' and '3'. In this case, '3 minus 1' doesn't mean anything… you can't subtracting defect 1 from defect 3. The numbers here are used to *indicate* or identify the levels of 'Defect Type' and do not have intrinsic meaning of their own. Dummy variables are created in this situation to 'trick' the regression algorithm into correctly analyzing attribute variables.

How is a dummy variable created?

We will illustrate this with an example: Let's say you want to find out whether the location of a house in the East, Southeast or Northwest side of a development and whether the house was built before or after 1990 affects its sale price. The image below shows a portion of the Sale Price dataset:

Sale Price in $ thousands

SalePrice	Y1990	E	SE
370	1	1	0
315	0	0	1
310	1	0	0
305	0	0	0
305	1	0	1
300	1	0	0
300	0	0	1
295	0	0	0
295	1	0	0
293	0	1	0
290	0	1	0
290	0	0	1
290	0	1	0
290	0	0	1
288	0	0	0
	1		0

SalePrice is the numerical response variable. The dummy variable **Y1990** represents the binary independent variable 'Before/After 1990'. Thus, it takes two values: '1' if a house was built after 1990 and '0' if it was built before 1990. Thus, a single dummy variable is needed to represent a variable with two levels.

Notice, there are only two dummy variables left, East (E) and SouthEast (SE). Together, they represent the Location variable with three levels (E, SE, NW). They're constructed so that

E = '1' if the house falls on the East side and '0' otherwise, and

SE = '1' if the house falls on the Southeast side and '0' otherwise

What happened to the third location, NW? Well, it turns out we don't need a third dummy variable to represent it. Setting both E and SE to '0' indicates a house on the NW side. Notice that this coding only works if the three levels are mutually exclusive (so not overlap) and exhaustive (no other levels exist for this variable).

The regression of SalePrice on these dummy variables yields the following model:

SalePrice = 258 + 33.9*Y1990 - 10.7*E + 21*SE

The constant intercept value 258 indicates that houses in this neighborhood start at $258 K irrespective of location and year built. The coefficient of Y1990 indicates that other things being equal, houses in this neighborhood built after 1990 command a $33.9 K premium over those built before 1990.

Similarly, houses on the East side cost $10.7 K lower (it has a negative sign) than houses on the NW side and houses on the SE side cost $21 K higher than houses on the NW side. Thus, NW serves as the baseline or reference level for E and SE.

We can estimate the sale price for a house built before 1990 and located on the East side from this equation by substituting Y1990 = 0, E = 1 and SE = 0, giving SalePrice = $247.3 K.

Things to keep in mind about dummy variables

Dummy variables assign the numbers '0' and '1' to indicate membership in any mutually exclusive and exhaustive category**.**

1. The number of dummy variables necessary to represent a single attribute variable is equal to the number of levels (categories) in that variable minus one.
2. For a given attribute variable, none of the dummy variables constructed can be redundant. That is, one dummy variable can not be a constant multiple or a simple linear relation of another.
3. The interaction of two attribute variables (e.g. Gender and Marital Status) is represented by a third dummy variable which is simply the product of the two individual dummy variables.