**Dummy Coding of Categorical Variables**

Dummy coding is a common method for including categorical variables in regression models. In dummy coding, for a categorical variable with 'n' categories, 'n-1' binary (0/1) variables, called dummy variables, are created to represent the categories. One category serves as the reference or baseline category, and the others are compared to it.

Here's how dummy coding works with an example:

Suppose you have a categorical variable "Color" with three categories: Red, Green, and Blue. To include this variable in a regression model using dummy coding, you would create two dummy variables: "Is\_Green" and "Is\_Blue." The reference category (baseline) can be "Red."

* "Is\_Green": It takes the value 1 if the color is Green and 0 otherwise.
* "Is\_Blue": It takes the value 1 if the color is Blue and 0 otherwise.

Now, let's say you want to predict the price of a product based on its color. Your regression model might look like this:

Price = *β*0​ + *β*1​X is green + *β*2​ X is blue + *ϵ*

* *β*0​ is the intercept of the model, representing the expected price when the color is Red (the reference category).
* *β*1​ represents the change in the expected price when the color is Green compared to Red.
* *β*2​ represents the change in the expected price when the color is Blue compared to Red.
* *ϵ* is the error term.

Interpreting the coefficients:

* *β*0​​: This is the expected price when the color is Red (the reference category). It's the baseline price.
* *β*1​​: This tells you how much the price is expected to change when the color is Green compared to Red.
* *β*2​​: This tells you how much the price is expected to change when the color is Blue compared to Red.

To interpret the effect of Green or Blue compared to the reference category (Red):

* If *β*1​​ is positive and statistically significant, it means that Green is associated with a higher price than Red.
* If *β*2​​​ is positive and statistically significant, it means that Blue is associated with a higher price than Red.

In this way, dummy coding allows you to compare the effect of each category relative to the reference category while accounting for the categorical nature of the variable. The reference category is the baseline against which the other categories are compared in terms of their impact on the dependent variable.

Top of Form