How can dummy variables be used in regression analysis?

Authored by

stats writer

April 25, 2024

Dummy variables are a commonly used tool in regression analysis to represent categorical variables in a quantitative model. They are used to assign numerical values to different categories or groups, allowing for the inclusion of qualitative data in the regression model. This enables the model to account for the effects of categorical variables on the dependent variable, providing a more accurate and comprehensive analysis. Dummy variables are particularly useful in situations where the categories cannot be ranked or are not inherently numerical, such as gender, race, or geographical region. By using dummy variables, regression analysis can incorporate both quantitative and qualitative data, leading to more robust and reliable results.

Use Dummy Variables in Regression Analysis

is a method we can use to quantify the relationship between one or more predictor variables and a .

Typically we use linear regression with . Sometimes referred to as "numeric" variables, these are variables that represent a measurable quantity. Examples include:

Number of square feet in a housePopulation size of a cityAge of an individual

However, sometimes we wish to use categorical variables as predictor variables. These are variables that take on names or labels and can fit into categories. Examples include:

Eye color (e.g. "blue", "green", "brown")Gender (e.g.

"male", "female")Marital status (e.g. "married", "single", "divorced")

When using categorical variables, it doesn't make sense to just assign values like 1, 2, 3, to values like "blue", "green", and "brown" because it doesn't make sense to say that green is twice as colorful as blue or that brown is three times as colorful as blue.

Instead, the solution is to use dummy variables. These are variables that we create specifically for regression analysis that take on one of two values: zero or one.

Dummy Variables: Numeric variables used in regression analysis to represent categorical data that can only take on one of two values: zero or one.

The number of dummy variables we must create is equal to *k*-1 where *k* is the number of different values that the categorical variable can take on.

The following examples illustrate how to create dummy variables for different datasets.

Example 1: Create a Dummy Variable with Only Two Values

Suppose we have the following dataset and we would like to use *gender* and *age* to predict *income*:

Income	Age	Gender
$45,000	23	Male
$48,000	25	Female
$54,000	24	Male
$57,000	29	Female
$65,000	38	Female
$69,000	36	Female
$78,000	40	Male
$83,000	59	Female
$98,000	56	Male
$104,000	64	Male
$107,000	53	Male

To use *gender* as a predictor variable in a regression model, we must convert it into a dummy variable.

Since it is currently a categorical variable that can take on two different values ("Male" or "Female"), we only need to create *k*-1 = 2-1 = 1 dummy variable.

To create this dummy variable, we can choose one of the values ("Male" or "Female") to represent 0 and the other to represent 1.

In general, we usually represent the most frequently

occurring value with a 0, which would be "Male" in this dataset.

Income	Age	Gender
$45,000	23	Male
$48,000	25	Female
$54,000	24	Male
$57,000	29	Female
$65,000	38	Female
$69,000	36	Female
$78,000	40	Male
$83,000	59	Female
$98,000	56	Male
$104,000	64	Male
$107,000	53	Male

Income	Age	Gender_Dummy
$45,000	23	0
$48,000	25	1
$54,000	24	0
$57,000	29	1
$65,000	38	1
$69,000	36	1
$78,000	40	0
$83,000	59	1
$98,000	56	0
$104,000	64	0
$107,000	53	0

We could then use *Age* and *Gender_Dummy* as predictor variables in a regression model.

Example 2: Create a Dummy Variable with Multiple Values

Suppose we have the following dataset and we would like to use *marital status* and *age* to predict *income*:

Income	Age	Marital Status
$45,000	23	Single
$48,000	25	Single
$54,000	24	Single
$57,000	29	Single
$65,000	38	Married
$69,000	36	Single
$78,000	40	Married
$83,000	59	Divorced
$98,000	56	Divorced
$104,000	64	Married
$107,000	53	Married

To use marital status as a predictor variable in a regression model, we must convert it into a dummy variable.

Since it is currently a categorical variable that can take on three different values ("Single", "Married", or "Divorced"), we need to create *k*-1 = 3-1 = 2 dummy variables.

To create this dummy variable, we can let "Single" be our baseline value since it occurs most often. Thus, here's how we would convert *marital status* into dummy variables:

Income	Age	Marital Status
$45,000	23	Single
$48,000	25	Single
$54,000	24	Single
$57,000	29	Single
$65,000	38	Married
$69,000	36	Single
$78,000	40	Married
$83,000	59	Divorced
$98,000	56	Divorced
$104,000	64	Married
$107,000	53	Married

Income	Age	Married	Divorced
$45,000	23	0	0
$48,000	25	0	0
$54,000	24	0	0
$57,000	29	0	0
$65,000	38	1	0
$69,000	36	0	0
$78,000	40	1	0
$83,000	59	0	1
$98,000	56	0	1
$104,000	64	1	0
$107,000	53	1	0

We could then use *Age*, *Married*, and *Divorced* as predictor variables in a regression model.

How to Interpret Regression Output with Dummy Variables

Suppose we fit a model using the dataset in the previous example with *Age*, *Married*, and *Divorced* as the predictor variables and *Income* as the response variable.

Here's the regression output: