

What is the Dummy Variable Trap?

Authored by
stats writer

December 10, 2025

RECOMMENDED CITATION

stats writer (2025). *What is the Dummy Variable Trap?*. PSYCHOLOGICAL SCALES.
Retrieved from <https://scales.arabpsychology.com/?p=107061>

The Dummy Variable Trap is a problem that can occur when creating a multiple linear regression model. It occurs when one or more of the independent variables are categorical variables, and one or more of the categories are redundant. This can lead to multicollinearity in the model, which can lead to incorrect results and inaccurate predictions.

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 house

Population size of a city

Age 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 . These are variables that we create specifically for regression analysis that 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.

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

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

When creating dummy variables, a problem that can arise is known as the **dummy variable trap**. This occurs when we create k dummy variables instead of $k-1$ dummy variables.

When this happens, at least two of the dummy variables will suffer from perfect . That is, they'll be perfectly correlated. This causes incorrect calculations of regression coefficients and their corresponding p-values.

Dummy Variable Trap: When the number of dummy variables created is equal to the number of values the categorical value can take on. This leads to multicollinearity, which causes incorrect calculations of regression coefficients and p-values.

For example, suppose we converted *marital status* into the following dummy variables:

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

In this case, *Single* and *Married* are perfectly correlated and have a correlation coefficient of -1.

Thus, when we go to perform multiple linear regression the calculations for the regression coefficients will be incorrect.

How to Avoid the Dummy Variable Trap

You only need to remember one rule to avoid the dummy variable trap:

If a categorical variable can take on k different values, then you should only create $k-1$ dummy variables to use in the regression model.

For example, suppose you'd like to convert a categorical variable "school year" into dummy variables. Suppose this variable takes on the following values:

Freshman
Sophomore
Junior
Senior

Since this variable can take on 4 different values, we will only create 3 dummy variables. For example, our dummy variables might be:

X1 = 1 if Sophomore; 0 otherwise

X2 = 1 if Junior; 0 otherwise

X3 = 1 if Senior; 0 otherwise

Since the number of dummy variables is one less than the number of values that "school year" can take on, we can avoid the dummy variable trap and the problem of multicollinearity.