

## Creating Dummy Variables

A dummy variable is a variable that takes on the values 1 and 0; 1 means something is true (such as age < 25, sex is male, or in the category "very much").

Some people also call dummy variables indicator variables.

---

Converting a categorical variable to dummy variables can be a tedious process when done using a series of series of if then statements.

---

Recoding a categorical SPSS variable into indicator (dummy) variables  
Q.

What is the SPSS command to transform a nominal variable of n classification groups into a series of n-1 indicator (or "dummy") variables?

A.

Unfortunately, there is no single command to do this. There are several short command sequences that can do it and examples are provided below. Of these, the DO REPEAT approach is somewhat more general, or at least easier if the reference category is not the lowest value.

```
* creating indicator variables.  
* all examples below generate indicators from a  
nominal variable, called cat, that is present in  
the active file.  
  
* create 4 indicator variables for categories 1 to 4  
of a 5-category variable called cat.
```

```
VECTOR nom(4).  
LOOP #i = 1 to 4.  
COMPUTE nom(#i) = (cat = #i).  
END LOOP.  
EXECUTE.
```

```
* alternatively .  
* create 4 indicator variables for categories 2 to 5  
of a 5-category variable called cat.
```

```
VECTOR ind(4).  
LOOP #i = 1 to 4.  
COMPUTE ind(#i) = (cat = #i + 1).  
END LOOP.  
EXECUTE.
```

```
* if you wanted to make the first category the reference  
category (0 on all indicator vars) with var names reflecting  
the original category : .
```

```
NUMERIC dum2 to dum5.  
VECTOR dumv = dum2 to dum5.  
LOOP #i = 1 to 4.  
COMPUTE dumv(#i) = (cat = #i + 1).
```

```

END LOOP.
EXECUTE.

* creating similar vars as above but using do repeat command.
DO REPEAT iv = indv2 to indv5
/ c = 2 to 5 .
COMPUTE iv = (cat = c).
END REPEAT.
EXECUTE.

* if reference category were neither first nor last, but 3rd,
DO REPEAT seems handier than VECTOR and LOOP.

DO REPEAT iv = c3i1 c3i2 c3i4 c3i5 / g = 1 2 4 5 .
COMPUTE iv = (cat = g).
END REPEAT.
EXECUTE.

```

---

## Working With Dummy Variables

- [Why use dummies?](#)
- [Nominal variables with multiple levels](#)
- [Interpreting results](#)

### Why use dummies?

Regression analysis is used with numerical variables. Results only have a valid interpretation if it makes sense to assume that having a value of 2 on some variable does indeed mean having twice as much of something as a 1, and having a 50 means 50 times as much as 1.

However, social scientists often need to work with categorical variables in which the different values have no real numerical relationship with each other. Examples include variables for race, political affiliation, or marital status. If you have a variable for political affiliation with possible responses including Democrat, Independent, and Republican, it obviously doesn't make sense to assign values of 1 - 3 and interpret that as meaning that a Republican is somehow three times as politically affiliated as a Democrat.

The solution is to use dummy variables - variables with only two values, zero and one. It does make sense to create a variable called "Republican" and interpret it as meaning that someone assigned a 1 on this variable is Republican and someone with an 0 is not.

### Nominal variables with multiple levels

If you have a nominal variable that has more than two levels, you need to create multiple dummy variables to "take the place of" the original nominal variable. For example, imagine that you wanted to predict depression from year in school:

freshman, sophomore, junior, or senior. Obviously, "year in school" has more than two levels.

What you need to do is to recode "year in school" into a set of dummy variables, each of which has two levels. The first step in this process is to decide the number of dummy variables. This is easy; it's simply  $k-1$ , where  $k$  is the number of levels of the original variable.

You could also create dummy variables for all levels in the original variable, and simply drop one from each analysis.

In this instance, we would need to create  $4-1=3$  dummy variables. In order to create these variables, we are going to take 3 of the levels of "year of school", and create a variable corresponding to each level, which will have the value of yes or no (i.e., 1 or 0). In this instance, we can create a variable called "sophomore," "junior," and "senior." Each instance of "year of school" would then be recoded into a value for "sophomore," "junior," and "senior." If a person were a junior, then "sophomore" would be equal to 0, "junior" would be equal to 1, and "senior" would be equal to 0.

### **Interpreting results**

The decision as to which level is not coded is often arbitrary. The level that is not coded is the category to which all other categories will be compared. As such, often the biggest group will be the not-coded category. For example, often "Caucasian" will be the not-coded group if that is the race of the majority of participants in the sample. In that case, if you have a variable called "Asian", the coefficient on the "Asian" variable in your regression will show the effect being Asian rather than Caucasian has on your dependant variable.

In our example, "freshman" was not coded so that we could determine if being a sophomore, junior, or senior predicts a different depressive level than being a freshman. Consequently, if the variable, "junior" was significant in our regression, with a positive beta coefficient, this would mean that juniors are significantly more depressed than freshman. Alternatively, we could have decided to not code "senior," if we thought that being a senior is qualitatively different from being of another year.