# What is dummy coding?

Authored by

**stats writer**

June 30, 2024

Dummy coding is a statistical technique used to represent categorical variables in a numerical format. It involves creating dummy variables, which are binary variables that indicate the presence or absence of a particular category within a variable. This allows for the inclusion of categorical variables in regression models and other statistical analyses, as they can be easily interpreted and compared with other variables. Dummy coding is commonly used in various fields such as psychology, sociology, and marketing research.

## FAQ: What is dummy coding?

**Dummy coding provides one way of using categorical predictor variables in various kinds of estimation models (see also effect coding), such as, linear regression. Dummy coding uses only ones and zeros to convey all of the necessary information on group membership. Consider the following example in which there are four observations within each of four groups:**

| group | g1 | g2 | g3 | g4 |
|---|---|---|---|---|
| | 1 | 2 | 5 | 10 |
| | 3 | 3 | 6 | 10 |
| | 2 | 4 | 4 | 9 |
| | 2 | 3 | 5 | 11 |
| mean | 2 | 3 | 5 | 10 |

+----------+

For this example we will need to create three dummy coded variables.
In general, with k groups there will be k-1 coded variables. Each of the
dummy coded variables uses one degree of freedom, so k groups has k-1 degrees of
freedom, just like in analysis of variance.

Here is how we will create the dummy variables which we will call d1, d2 and d3.
For d1, every observation in group 1 will be coded as 1 and 0 for all other groups
it will be coded as zero. We then code d2 with 1 if the observation is in group 2 and zero otherwise.
For d3, observations in group 3 will be coded 1 and zero for the other groups. For d4,
there is no d4. d4 is not needed because d1-d3 has all of the information needed
to determine which observation is in which group.

Here is how the data look when arranged for use with a regression procedure.

**y grp d1 d2 d3**

**1 1 1 0 0**

**3 1 1 0 0**

**2 1 1 0 0**

**2 1 1 0 0**

**2 2 0 1 0**

**3 2 0 1 0**

**4 2 0 1 0**

**3 2 0 1 0**

**5 3 0 0 1**

**6 3 0 0 1**

**4 3 0 0 1**

**5 3 0 0 1**

**10 4 0 0 0**

**10 4 0 0 0**

**9 4 0 0 0**

**11 4 0 0 0**



**Note that every observation in group 1 has the dummy code value of 1 for d1 and zero for the others. Those in group 2 have 1 for d2 and 0 otherwise, and for group 3 d3 equals 1 with zero for the others. Observations in group 4 have all**

zeros on d1, d2 and d3. These three dummy variables contain all of the information needed to
determine which observations are included in which group. If you are in group 1 then
d1 is equal to 1 while d2 and d3 are zero. Thus, each of the groups is defined by
having a one of the dummy variables equal to one except of one group which is all zero's.
The group with all zeros is known as the reference group, which
in our example is group 4. We will see exactly what this means after we look at
the regression analysis results.

F(3, 12) = 76.00 P = 0.0000 R-squared = 0.95

---

y | Coef. Std. Err. t P>|t|

---+---

d1 | -8 .5773503 -13.86 0.000

d2 | -7 .5773503 -12.12 0.000

d3 | -5 .5773503 -8.66 0.000

constant | 10 .4082483 24.49 0.000

---

With dummy coding the constant is equal to the mean of the reference group, i.e., the group with all dummy variables equal to zero. In this case, the value is equal to 10 which is the mean of group 4. The coefficients of each of the dummy variables is equal to the difference between the mean of the group coded 1 and the mean of the reference group. In our example the mean of group 1 is 2 and the difference of 2-10 is -8, which is the value of the regression coefficient for d1. The t-test associated with that coefficient is the test of group 1 versus group 4.

What if you used group 1 as the reference group? That is, what if group 1 was the group coded with all zeros? In that case, the value of the constant would be the mean of group 1 (which is 2) and the regression coefficients would be equal to the differences between the group mean and the mean of group 1. In all other respects the models are identical with the same F-ratio and R-squared

**regardless of which group is selected as the reference group.**

**What if you try to include dummy variables for all four groups? Some programs will refuse to run the analysis and some will run it but drop one of the dummy variables?**

**The fact is, you only need three dummy variables to determine membership in four groups.**

## See also