What is effect coding?

Authored by

stats writer

June 30, 2024

Effect coding is a statistical technique used to represent categorical variables in regression models. It involves assigning numerical codes to different categories of a variable, with one category designated as the baseline. The effect of each category is then measured relative to the baseline, allowing for the comparison of different categories in terms of their impact on the outcome variable. This method is particularly useful for analyzing data with multiple categorical variables and can help identify significant effects and interactions between variables.

FAQ: What is effect coding?

Effect coding provides one way of using categorical predictor variables in various kinds of estimation models (see also dummy coding), such as, linear regression.

Effect coding uses only ones, zeros and minus ones 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 |
+----------+

Grand mean = 5

For this example we will need to create three effect coded variables.
In general, with k groups there will be k-1 coded variables. Each of the
effect 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 effect variables which we will call e1, e2 and e3.
For e1, every observation in group 1 will be coded as 1, 0 for groups 2 and 3, and -1
for group 4.
We then code e2 with 1 if the observation is in group 2 and 0 for groups 1 and 3, and
-1 for group 4.
For e3, observations in group 3 will be coded 1 and 0 for groups 1 and 2, and

-1 for group 4. For e4,
there is no e4. e4 is not needed because e1-e3 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 e1 e2 e3
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 -1 -1 -1
10 4 -1 -1 -1
9 4 -1 -1 -1

11 4 -1 -1 -1

Note that every observation in group 1 has the effect code value of 1 for e1 and 0 for the others. Those in group 2 have 1 for e2 and 0 otherwise, and for group 3 e3 equals 1 with 0 for the others. Observations in group 4 have all -1s on e1, e2 and e3. These three effect variables contain all of the information needed to determine which observations are included in which group. If you are in group 1 then e1 is equal to 1 while e2 and e3 are 0. Thus, each of the groups is defined by having a one of the effect variables equal to one except of one group which is all -1s. The group with all -1s 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|
---------+--------------------------------------
e1 | -3 .3535534 -8.49 0.000
e2 | -2 .3535534 -5.66 0.000
e3 | 2.36e-16 .3535534 0.00 1.000
constant | 5 .2041241 24.49 0.000
------------------------------------------------
```

With effect coding the constant is equal to the grand mean of all of the observations. In this case, the value is equal to 5 which is the grand mean. The coefficients of each of the effect variables is equal to the difference between the mean of the group coded 1 and the grand mean. In our example the mean of group 1 is 2 and the difference of 2-5 is -3, which is the value of the regression coefficient for e1. The t-test associated with that coefficient is the test of group 1 versus the grand mean. The coefficient for e3 is 2.36e-16 which is just a computer rounding way of saying

zero, i.e., 5-5 equals 0 give or take some rounding error.

What if you used group 1 as the reference group? That is, what if group 1 was the
group coded with all -1s? In that case, the value of the constant would still be the
grand mean. The coefficients for e2 and e3 will have the same values, but the
coefficients for e1 will be replaced by the coefficient for e4, the difference
between the grand mean and group 4. 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.

Unbalanced data

By unbalanced data we mean unequal group sizes. There are a couple of differences when
using effect coding with unbalanced designs. Consider the following four group design:

+---------------------------------+
| group | g1 | g2 | g3 | g4 |

-------	------+------+------+------
	1			10
	3	3	6	10
	2	4		9
	2		5	11
+---------------------------------+
| mean | 2 | 3.5 | 5.5 | 10 |
+---------------------------------+

Grand mean = 5.5 -- Unweighted grand mean = 5.25

In the table above, the grand mean (5.5) is the overall mean of the 12 observations while
the unweighted grand mean (5.25) is just the simple average of the four group means.

Now if we do the standard effect coding and run the regression we get the
following summary table.

F(3, 12) = 73.60 P = 0.0000 R-squared = 0.965

---

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

---+---

e1 | -3.25 .369755 -8.79 0.000
e2 | -1.75 .4635124 -3.78 0.005
e3 | .25 .4635124 0.54 0.604
constant | 5.25 .2420615 21.69 0.000

Now the constant for the model is the unweighted grand mean, i.e., the mean of means.
In all other respects the coefficients are interpreted in the way except that you
replace grand mean with the term unweighted grand mean. Of course, what is really going on
is that for balanced groups, the weighted and unweighted means are the same.

Why use effect coding?

Here's a good question, why use effect coding instead of dummy coding? If you have several
categorical variables in a model it often doesn't make much difference whether you use
effect coding or dummy coding. However, if you have an interaction of two categorical
variables then effect coding may provide some benefits. The primary benefit is that you

get reasonable estimates of both the main effects and interaction using effect coding.
With dummy coding the estimate of the interaction is fine but main effects are not "true"
main effects but rather what are called simple effects, i.e., the effect of one variable
at one level of the other variable. This is why most analysis of variance
programs use some type of effect
coding when estimating the various effects in an ANOVA model.

See also