THE UNIVERSITY *of* EDINBURGH | usher institute

Categorical variables

Steven Kerr

Supported by

THE UNIVERSITY *of* EDINBURGH | DDI Data-Driven Innovation

Categorical variables

- Categorical variables are variables which indicate membership in some categories.
- For example, the variable *fruit* might take one of the values {apple, banana, orange}.
- Believe it or not, categorical variables often generate confusion.

Trap 1: meaningless operations

- One source of confusion is that categories are often represented numerically. So, for example, instead of using the labels {apple, banana, orange}, we might use the labels {1, 2, 3}.

- The problem is that numbers have more *mathematical structure* than category labels. For example, it is meaningful to add together 1 and 2; it is not meaningful to add together apple and banana.

- Unfortunately, this often leads to operations being done on numerical labels that are not meaningful.

Trap 1: meaningless operations

- For example, we might wish to calculate something like a covariance between a categorical variable and a numerical variable.
- Let's say we have {apple, banana, orange} labelled as {1, 2, 3}, and we naively calculate the covariance with a numerical tastiness rating between 0 and 10.
- The tastiness ratings are {9, 8, 7} (these are my personal ratings, I mean no ill will towards bananas or oranges).

Trap 1: meaningless operations

- We get $$\frac{(1-2)(9-8)+(2-2)(8-9)+(3-2)(7-8)}{3}=-\frac{2}{3}$$

- But remember the numerical labels for {apple, banana, orange} are arbitrary; we could equally well choose {1,000, 2,000, 3,000}.

- Repeating the above calculation with these labels gives an answer of $-\frac{2{,}000}{3}$.

- You can get any ‘answer’ you want by re-labelling.

Trap 1: meaningless operations

- If arbitrarily re-assigning numerical labels to categories affects the answer, then you are doing something that is not meaningful. And you should stop it. Right now.

- It is surprisingly easy to fall into this trap. Watch out for it!

Trap 2: The dummy variable trap

- Let's say I have a categorical variable for sex, and I do a linear regression with height as the dependent variable,

$$height_i = \beta_0 + \beta_1 male_i + \beta_2 female_i$$

Looks innocent enough, right?

- Attempting to train this model will cause your computer to crap its proverbial pants.

Trap 2: The dummy variable trap

- The problem is that the variables included on the right hand side are not a linearly independent set.
- This is sometimes referred to as *multicollinearity*.
- In essence, there are multiple values for the parameters $\beta_0, \beta_1, \beta_2$ that will result in identical models.

Trap 2: The dummy variable trap

- For example, $\beta_0 = 0, \beta_1 = 1, \beta_2 = 2$ will result in the following predictions:
  - $height = 1$ for males, $height = 2$ for females

- However, $\beta_0 = 1, \beta_1 = 0, \beta_2 = 1$ will result in the exact same predictions! (check for yourself)

- Indeed there are infinitely many values that result in identical models.

- How is your poor computer meant to choose between these equally fine, upstanding models?

- The model is *underdetermined*.

Trap 2: The dummy variable trap

- The solution is to remove a redundant variable. The following models are all perfectly ok:

$$height_i = \beta_1 male_i + \beta_2 female_i$$
$$height_i = \beta_0 + \beta_2 female_i$$
$$height_i = \beta_0 + \beta_1 male_i$$

- Fortunately, R is nice and takes care of this for you. It won't let you do something stupid. For example, if you try:
  - lm(height ~ 1 + sex, data = data)

  It will automatically correct your mistake.

- You have to try harder to break it. Which brings me to my next topic.

One-hot encoding

- Let's go back to our fruit variable. Say we have a column in our data like this:

fruit
apple
apple
banana

- One-hot encoding switches this to:

apple	banana	orange
1	0	0
1	0	0
0	1	0

One-hot encoding

- If I was really intent on breaking R, I could one-hot encode the fruit variable and force R to fall into the dummy variable trap,
  - lm(tastiness ~ 1 + apple + banana + orange, data = data)
- That wouldn't be a good use of time. However, one-hot encoding is often useful/essential.
- One-hot encoding categorical variables helps us avoid trap 1: meaningless operations.
- However, amusingly it can make us more likely to fall into trap 2: the dummy variable trap.
- Conclusion: life is hard.