Notes 9: Qualitative variables

ECO 231W - Undergraduate Econometrics

Prof. Carolina Caetano

1 Qualitative variables

At this point, you must have realized that there exist some variables which are not exactly numbers. When I say that we should control for the person's income or education, it's easy to imagine a number. Income can be measured in dollars per year, education can be measured in years, and both would be numbers that make sense. However, I also mentioned the mother's race and marital status. Race is a category. Is she white, black, latino, asian, etc.? How can we give these variables a meaningful number? We could say: white=1, black=2, latino=3, etc. but that doesn't really make any sense. The same goes for marital status. Legally, a person is either single, married, divorced, or widowed. No number can really describe those states. Yet, these are important variables, and they often show up on data sets. How do we deal with them? Without realizing, you have already been dealing with qualitative data. The way we deal with them is by transforming them in $0-1$ variables, the so called dummies.

Dummy variables are _____

_____

The way to understand dummies is to think that they answer a yes or no question, where 1 is yes, and 0 is no. For example, we could have a dummy variable called *single*. The way to understand it is to think of the following question:

- _____

  _____

When we are using dummy variables (qualitative variables), the interpretation of the coefficients of the linear model is slightly different. In truth it is mathematically the same, but one must be careful with the language and the meaning, which change. Take the model

$$birthweight = \beta_0 + \beta_1 cigarettes + \beta_2 income + \beta_3 single + u.$$

with $\mathbb{E}[u|cigarettes, income, single] = 0$. Sure, this is probably not a good model of reality, but bear with me.

- What is the interpretation of β_2?_____

  _____

- What is the interpretation of β_3? _____

  _____

  Let's check the equations. What is the expected *birthweight* if the woman is single?

  and if the woman is not single, leaving everything else constant:

  subtracting the first from the second:

- What is the interpretation of β_0? _____

  _____

  _____

  Remember that

In the general case, suppose that x_2 is a dummy variable. Consider he model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u,$$

with $\mathbb{E}[u|x_1, x_2] = 0$. Then

- What is β_2? _____

_____

- What is β_0? _____

_____

There is nothing special about x_2. I could be saying the same for x_1.

How do we control for marital status? Marital status is not just whether the person is single or not. It is whether the person is specifically single, married, divorced or widowed. There are 4 possibilities. How do we incorporate them into the model? The answer seems complex at first, but later you will see that it makes a lot of sense. What you do is create dummies for all categories but one. For example, you create 3 dummies:

$$single,\ married,\ divorced$$

Now we have 3 new variables in the model.

- A single woman is characterized by $single = 1$. Observe that the categories are excluding, so that a woman that has $single = 1$ automatically has $married = 0$, and $divorced = 0$.
- Analogously, a married woman is characterized by $married = 1$, and automatically, $single = 0$, and $divorced = 0$.
- A divorced woman is characterized by _____

_____

- A widowed woman is characterized by _____

_____

So, you can see that there is something pretty special happening here. We have 3 dummy variables that completely describe 4 categories. In general, if we have M excluding categories, use $M - 1$ dummy variables, for all categories but one. Observe that

1. _____
2. _____

_____

The interpretation of the coefficients of the linear model is still mathematically the same, but the language and meaning changes yet again. Take the model

$$\begin{aligned} birthweight = \beta_0 + \beta_1 cigarettes + \beta_2 income \\ + \beta_3 single + \beta_4 married + \beta_5 divorced + u. \end{aligned}$$

with $\mathbb{E}[u|cigarettes, income, single, married, divorced] = 0$.

- What is the interpretation of β_3?_____

  _____

  Let's check the equations. What is the expected *birthweight* if the woman is single?

  if the woman is widowed, leaving everything else constant:

  subtracting the first from the second:

- What is the interpretation of β_0?_____

  _____

  Remember that

Observe that the interpretation is always with respect to the excluded category. This means that although it doesn't really matter which category you exclude, the interpretation of the coefficients does change when you change the excluded category. For example, suppose that I excluded the divorced category, and instead created 3 dummies:

$$single,\ married,\ widowed$$

Now, in the model

$$\begin{aligned} birthweight = \beta_0 + \beta_1 cigarettes + \beta_2 income \\ + \beta_3 single + \beta_4 married + \beta_5 widowed + u. \end{aligned}$$

with $\mathbb{E}[u|cigarettes, income, single, married, widowed] = 0$.

- What is β_3?_____

- What is β_0? _____

- What is $\beta_4 - \beta_3$? _____

  Check mathematically that this is true.

Why do we create $M-1$ dummies, why not M? _____