

Handout for PS104A

Dummy Variables

Dummy variables are simply variables that have been coded either a 0 or a 1 to indicate an observation falls into a certain category. They are also sometimes called indicator variables. We use dummy variables when we are interested in using a nominal level variable in our analysis.

Let's say in some dataset we're using, we have men coded as 1, and women coded as 2. We can't just put this into a regression --- the "1" and the "2" don't mean anything, except to mark which gender each survey respondent is. What we need to do is recode this variable into a dummy variable, with one category (say men) equal to zero, and the other category (women) equal to one. This is a variable we could put into a regression, and the coefficient we estimate on this variable will tell us how much we need to shift the intercept term for women relative to men.

For example, say I'm estimating a regression of the effect of education and gender on how conservative someone is. Education is coded as an interval variable (years of education), so I don't need to worry about that. But I need to recode gender into a dummy variable if I want to have meaningful regression results. My regression equation would be:

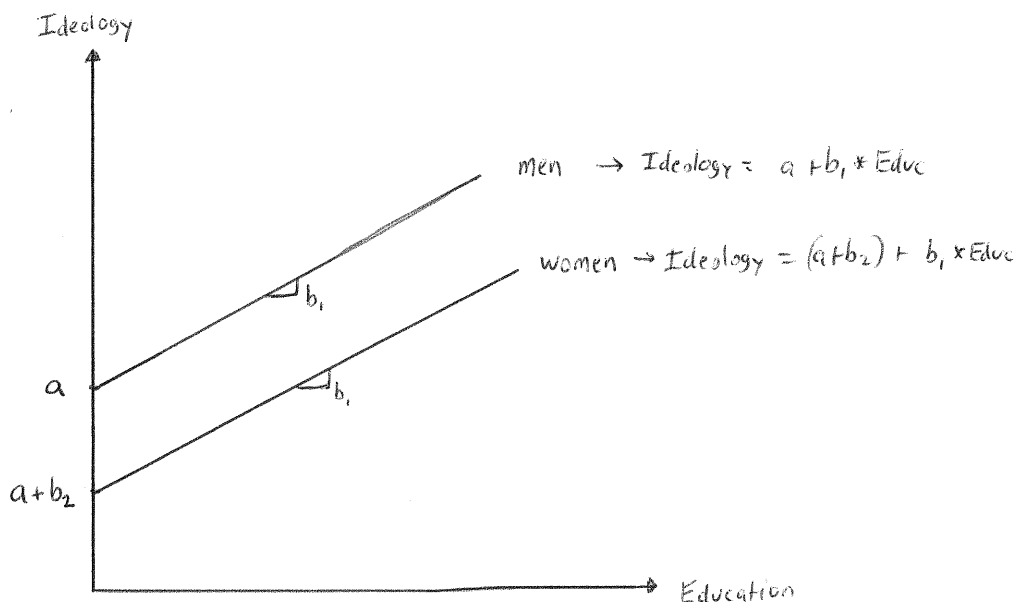
$$\text{Ideology} = a + b_1 * \text{education} + b_2 * \text{gender}$$

The b_2 term drops out for men (because they are coded zero), and adds b_2 to the intercept for women (because they are coded one).

The results of my regression of ideology on education and gender are:

<u>Variable</u>	<u>Coefficient</u>	<u>Std. Error</u>	<u>t</u>	<u>Sig</u>
Education	3	1	3	.000
Gender	-0.5	0.1	5	.000
Constant	1	1	1	.459

This tells me that increasing education increases how conservative someone is, and women are more liberal than men. The regression line would look like this:



Dummy variables can be used for nominal categories with more than 2 categories --- you'll just need to make additional dummy variables. For instance, if you had a 3 category nominal variable, you could see how the groups differ from each other with 2 dummy variables. One group must be left as a baseline group (equal to zero in all dummy variables).

You can also *interact* variables, meaning you can multiply 2 variables together. For instance, I might think that not only do education and gender affect ideology, but education affects the ideology of women differently than it does men. In that case I'd create an interaction term, multiplying education and gender, and include that in my regression model. My regression equation would be:

$$\text{Ideology} = a + b_1 \cdot \text{education} + b_2 \cdot \text{gender} + b_3 \cdot (\text{educ} \cdot \text{gender})$$

The b_2 term drops out for men (because they are coded zero), and adds b_2 to the intercept for women (because they are coded one). The b_3 term drops out for men, and adjusts the slope on education for women.

Say my results are:

Variable	Coefficient	Std. Error	t	Sig
Education	1	0.5	2	.030
Gender	-0.5	0.2	2.5	.000
Educ*Gender	0.5	0.2	2.5	.004
Constant	1	2	0.5	.894

This tells me the same thing as my previous results, plus education has more of an effect on being conservative for women than for men. The regression line would look like this:

