Multiple Regression Analysis

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + u$
- 5. Dummy Variables

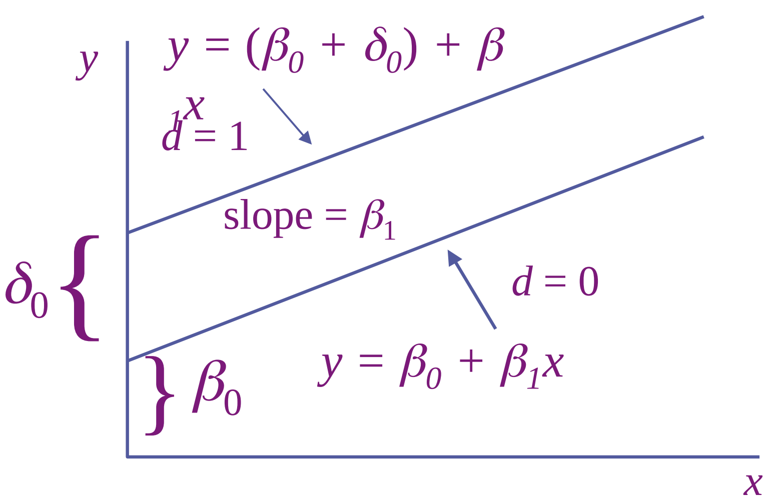
Dummy Variables

- A dummy variable is a variable that takes on the value 1 or 0
- Examples: male (= 1 if are male, 0 otherwise), south (= 1 if in the south, 0 otherwise), etc.
- Dummy variables are also called binary variables, for obvious reasons

A Dummy Independent Variable

- Consider a simple model with one continuous variable (*x*) and one dummy (*d*)
- $y = \beta_0 + \delta_0 d + \beta_1 x + u$
- This can be interpreted as an intercept shift
- If d = 0, then $y = \beta_0 + \beta_1 x + u$
- If d = 1, then $y = (\beta_0 + \delta_0) + \beta_1 x + u$
- The case of d = 0 is the base group

Example of $\delta_0 > 0$



Dummies for Multiple Categories

- We can use dummy variables to control for something with multiple categories
- Suppose everyone in your data is either a HS dropout, HS grad only, or college grad
- To compare HS and college grads to HS dropouts, include 2 dummy variables
- hsgrad = 1 if HS grad only, 0 otherwise; and colgrad = 1 if college grad, 0 otherwise

Multiple Categories (cont)

- Any categorical variable can be turned into a set of dummy variables
- Because the base group is represented by the intercept, if there are n categories there should be n – 1 dummy variables
- If there are a lot of categories, it may make sense to group some together
- Example: top 10 ranking, 11 – 25, etc.

Interactions Among Dummies

- Interacting dummy variables is like subdividing the group
- Example: have dummies for male, as well as hsgrad and colgrad
- Add male*hsgrad and male*colgrad, for a total of 5 dummy variables –> 6 categories
- Base group is female HS dropouts
- hsgrad is for female HS grads, colgrad is for female college grads
- The interactions reflect male HS grads and male college grads

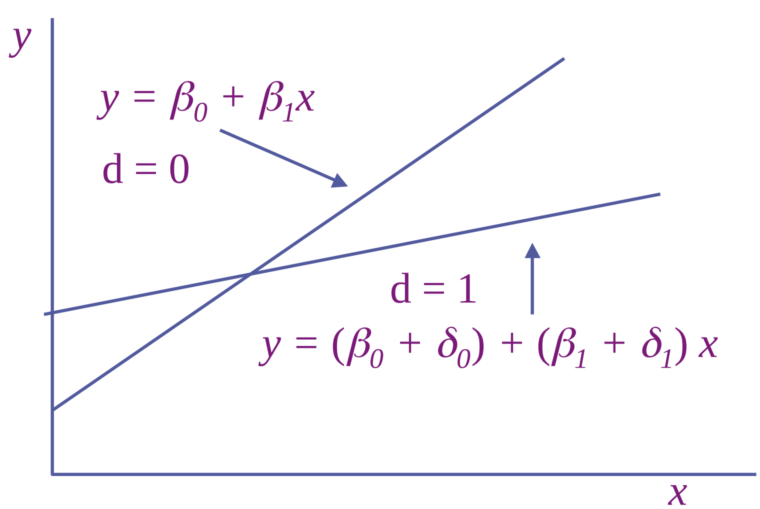
More on Dummy Interactions

- Formally, the model is $y = \beta_0 + \delta_1 male + \delta_2 hsgrad + \delta_3 colgrad + \delta_4 male*hsgrad + \delta_5 male*colgrad + \beta_1 x + u$, then, for example:
- If male = 0 and hsgrad = 0 and colgrad = 0
- $y = \beta_0 + \beta_1 x + u$
- If male = 0 and hsgrad = 1 and colgrad = 0
- $y = \beta_0 + \delta_2 hsgrad + \beta_1 x + u$
- If male = 1 and hsgrad = 0 and colgrad = 1
- $y = \beta_0 + \delta_1 male + \delta_3 colgrad + \delta_5 male*colgrad + \beta_1 x + u$

Other Interactions with Dummies

- Can also consider interacting a dummy variable, *d*, with a continuous variable, *x*
- $y = \beta_0 + \delta_1 d + \beta_1 x + \delta_2 d^*x + u$
- If $d = 0$, then $y = \beta_0 + \beta_1 x + u$
- If $d = 1$, then $y = (\beta_0 + \delta_1) + (\beta_1 + \delta_2)\, x + u$
- This is interpreted as a change in the slope

Example of $\delta_0 > 0$ and $\delta_1 < 0$



Testing for Differences Across Groups

- Testing whether a regression function is different for one group versus another can be thought of as simply testing for the joint significance of the dummy and its interactions with all other *x* variables
- So, you can estimate the model with all the interactions and without and form an *F* statistic, but this could be unwieldy

The Chow Test

- Turns out you can compute the proper F statistic without running the unrestricted model with interactions with all *k* continuous variables
- If run the restricted model for group one and get SSR_1, then for group two and get SSR_2
- Run the restricted model for all to get SSR, then

$$F = \frac{\left[SSR - \left(SSR_1 + SSR_2\right)\right]}{SSR_1 + SSR_2} \bullet \frac{\left[n - 2(k+1)\right]}{k+1}$$

The Chow Test (continued)

- The Chow test is really just a simple *F* test for exclusion restrictions, but we've realized that $SSR_{ur} = SSR_1 + SSR_2$
- Note, we have $k + 1$ restrictions (each of the slope coefficients and the intercept)
- Note the unrestricted model would estimate 2 different intercepts and 2 different slope coefficients, so the df is $n - 2k - 2$

Linear Probability Model

- $P(y = 1|x) = E(y|x)$, when y is a binary variable, so we can write our model as
- $P(y = 1|x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$
- So, the interpretation of β_j is the change in the probability of success when x_j changes
- The predicted *y* is the predicted probability of success
- Potential problem that can be outside [0,1]

Linear Probability Model (cont)

- Even without predictions outside of [0,1], we may estimate effects that imply a change in *x* changes the probability by more than +1 or –1, so best to use changes near mean
- This model will violate assumption of homoskedasticity, so will affect inference
- Despite drawbacks, it's usually a good place to start when *y* is binary

Caveats on Program Evaluation

- A typical use of a dummy variable is when we are looking for a program effect
- For example, we may have individuals that received job training, or welfare, etc
- We need to remember that usually individuals choose whether to participate in a program, which may lead to a self-selection problem

Self-selection Problems

- If we can control for everything that is correlated with both participation and the outcome of interest then it's not a problem
- Often, though, there are unobservables that are correlated with participation
- In this case, the estimate of the program effect is biased, and we don't want to set policy based on it!