Lecture 19

chapter 7
outline

we are going to cover Sections 7.1 - 7.4 only (pp. 218-240) in Chapter 7. We will leave section 7.5 and the Linear Probability Model until the sequel to this course.

I. Dummy Variables
   A. Additive Dummy Variable
   B. Interpreting coefficients on Dummy variables when the Dependent variable is log(y)
   C. Dummy Variables for Multiple categories
      i. The Reference (Base) Group
      ii. Avoiding the Dummy Variable Trap
   D. Incorporating Ordinal Information using Dummy variables - using a Rank Variable is equivalent to putting restrictions on Group Dummy Variables, see p. 229.

E. Multiplicative Dummies

II. Chow Test: Testing for Differences in Regression Functions Across Groups

A. Separate Regressions Approach

B. Additive / Multiplicative Dummy Variable Approach

C. Both Approaches provide same calculated F-statistic but Additive / Multiplicative Dummy Variable Approach offers more direct integretation of nature of structural change, if there is any.

Two types of qualitative variables - qualitative dependent variable, qualitative explanatory variable. A *quanitative* variable is a variable like wage or income that is measured on a continuous scale. A *qualitative variable* is one that *describes a condition* like gender, y = 1 if male, 0 if female, or y = 1 if a person defaults on a loan, 0 otherwise. A qualitative variable can also consist of multiple categories like y = -1 if a person drives to work, 0 if person walks to work, and 1 if persons takes public transportation to work. An *ordinal variable* is a qualitative variable with categories that are *ranked*. For example, the Moody's Financial Rating service gives ordinal rankings of municipalities ranging from say 0 to 4 with a zero rating indicating municipalities with a poor credit rating and a high probability of default on their municipal bonds while a rating (ranking) of 4 indicates a municipality with an excellent credit rating and a very low probability of default on their municipal bonds.

In this lecture we will be more interested in understanding the nature of qualitative and ordinal explanatory variables and leave the study of qualitative dependent variables to the next course.

Binary qualitative variables (male, female for example) are often called dummy variables. When defining dummy variables it must be decided which category or event will be assigned the value one and which category or event will be assigned the value of zero. One possible definition of

The gender dummy variable could be $X = \begin{cases} 1 \text{ if male} \\ 0 \text{ if female} \end{cases}$

or, just as well, $X^* = \begin{cases} 0 \text{ if male} \\ 1 \text{ if female} \end{cases}$. Im statistical analysis terms it makes no ~~definition~~ difference which definition you use, The definition will only affect the sign of a coefficient associated with a dummy variable not the statistical significance of the dummy variable. The category associated with the 0 value of the dummy variable is called the base group (or reference group).

Example of use of Additive Dummy

$$wage = \beta_0 + \delta_0\, female + \beta_1\, educ + u$$

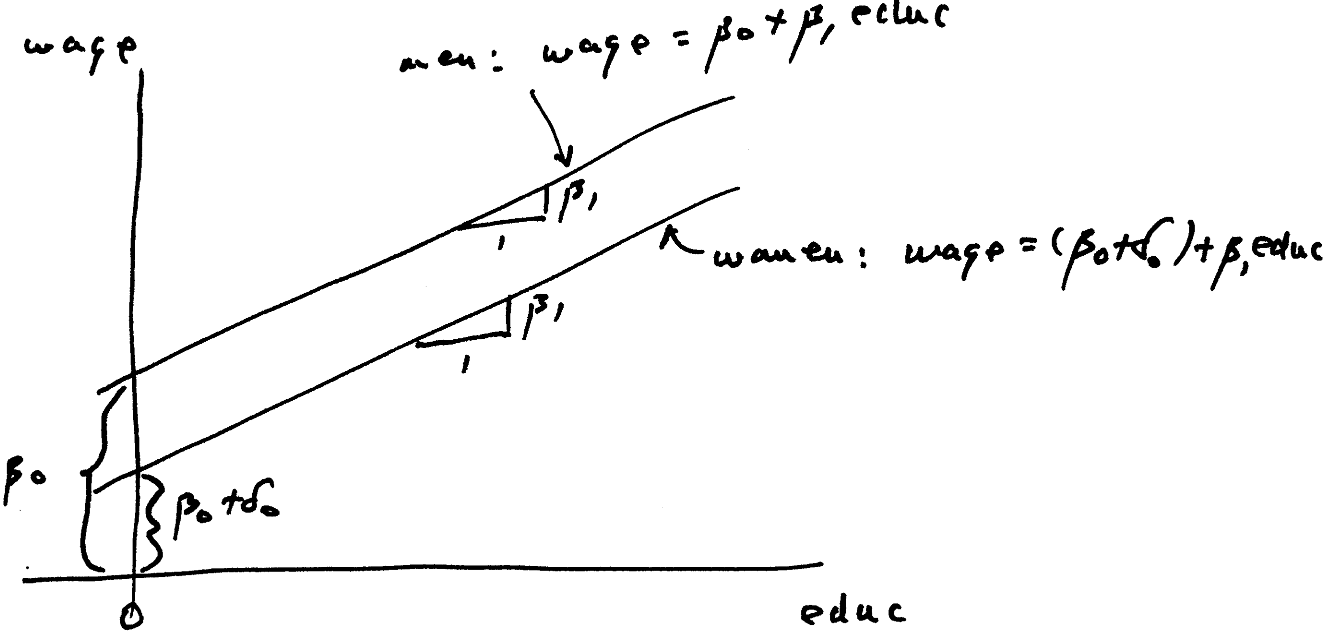
female is a dummy variable where female = {1 if female, 0 if male}. Here males is the reference group. we could have just as easily modeled the male vs. female difference in wage by

$$wage = \beta_0 + \gamma_0 male + \beta_1 educ + u$$

where male is a dummy variable defined as

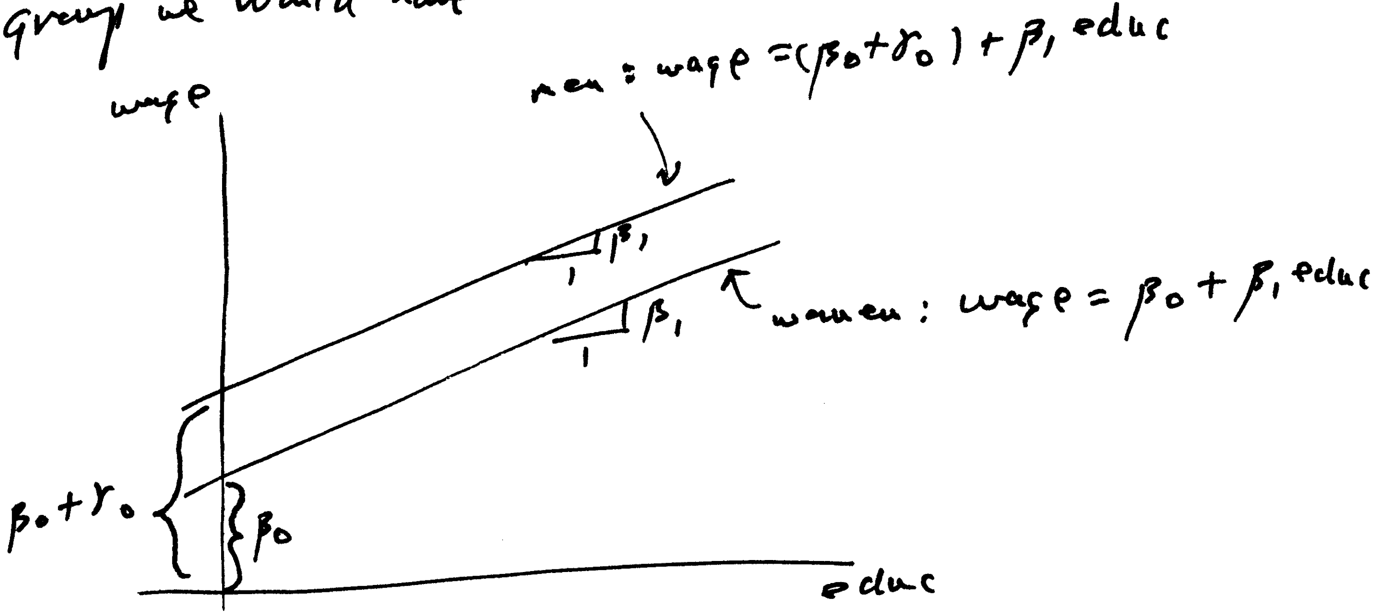
male = {1 if male, 0 if female}.

The first dummy variable representation is graphed as follows:



β_1 is the slope of both PRFs ~~for~~ men and women. ~~while~~ The intercept of the mens' PRF is β_0 while the intercept of the women's PRF is $(\beta_0 + \delta_0)$. In this representation, $\delta_0 < 0$. δ_0 represents the difference of the women's intercept from the men's intercept.

If instead we had used the women as the reference group we would have


men: wage $=(\beta_0+\gamma_0)+\beta_1$ educ
wage
β_1
1
β_1
1
women: wage $=\beta_0+\beta_1$ educ
$\beta_0+\gamma_0$
β_0
educ

Here $\gamma_0 > 0$ $(= -\delta_0)$. For statistical inference purposes it doesn't make any difference which group you choose as the reference group, any statistical significance between the groups will be discerned regardless.

In policy analysis dummy variables are often used to indicate the individuals who received a treatment (like an education training course). For example, treat could be a dummy variable like treat = {1 if individual participated in the training program or 0 if individual did not participate in the training program.}

For some good examples of how dummy explanatory variables can be used in regression equations see examples 7.1 - 7.5 in the wooldridge textbook.

Now the use of several dummy variables at once can be used to categorize multiple groups, for example, married men, married women, single men, and single women. Before defining appropriate dummy variables for these four categories we first have to pick a reference group (it doesn't make any difference which one we pick, we just have to remember which group we decide to choose as the reference group). Suppose we choose single males as the base group. Then we must define 3 remaining dummy variables marrmale = {1 if married male, 0 otherwise}, marrfem = {1 if married female, 0 otherwise} and singfem = {1 if single female, 0 otherwise}. See equal (7.11) in Wooldrige for the analysis of a log(wage) equation vis-a-vis these four groups. We cannot include a fourth dummy variable, say singmale, because if we did so, the four dummy variables would be perfectly collinear with the intercept term in the equation and ordinary least squares coefficient estimates would not be computable. (Recall Assumption MLR.4, p.86, in Wooldridge.) Inappropriately including dummy variables for all categories of an explanatory variable is called the dummy variable trap and is to be avoided at all costs. (When the computer doesn't return coefficient estimates, standard errors, and t-statistics for some of the variables in your regression and if you are using dummy variables to represent multiple categories you might suspect that you have fallen into the dummy variable trap!)

Incorporating Ordinal Information by using Dummy Variables

Wooldridge gives an excellent example on p. 229. Suppose that local govenments are rated on a scale from 0 to 4 concerning the quality of their debt, 0 representing the lowest quality debt versus 4 representing the highest quality debt with a very low probability of default. Call this variable CR.

Consider MBR = municipal bond interest rate and the regression function

$$MBR = \beta_0 + \beta_1 CR + \text{other factors}. \qquad (1)$$

β_1 is the percentage point change that we would expect in MBR if CR was increased by one unit (rank). We would expect $\beta_1 < 0$. But using a ranked variable like CR really represents a restriction on the behavior of MBR as you from one rank to the next. Each increase in rank gives you the same decrease in MBR whether going from $CR = 0$ to $CR = 1$ versus, say $CR = 3$ versus $CR = 4$. ~~What~~ It is interesting that we can statistically test this "equal incremental effect" assumption using the CR ordinal rank variable by considering the unrestricted equation

$$MBR = \beta_0 + \delta_1 CR_1 + \delta_2 CR_2 + \delta_3 CR_3 + \delta_4 CR_4 + \text{other factors} \qquad (2)$$

where the multiple category dummy variables CR_1, CR_2, CR_3, and CR_4 are defined as $CR_i = 1$ if $CR = i$ (rank $= i$) and zero otherwise, $i = 1, 2, 3, 4$. Notice in the above regression function $CR = 0$ local governments ~~are~~ is the reference group and that we avoided the dummy variable trap by leaving out the category $CR = 0$ in the regression equation.

To test the appropriateness of equation (1) we simply need to test the null hypothesis $H_0: \delta_2 = 2\delta_1$, $\delta_3 = 3\delta_1$, and $\delta_4 = 4\delta_1$ in equation (2). The resulting F-test would therefore have 4 numerator degrees of freedom and $N-k-1$ denominator degrees of freedom where $k = 4$ + no. of other factors. The sum of squared residuals from the first equation (1) would represent the restricted residual sum of squares, SSR_r, whereas the sum of squared residuals from the second equation (2) would represent the unrestricted sum of squared residuals, SSR_{ur}. If the p-value of the F-statistic is greater than α (say .05) then we accept the null hypothesis and therefore using the rank variable CR is OK. Otherwise, we should model the incremental effects of the various credit ratings to be unequal and we would need to use the unrestricted regression function (2).

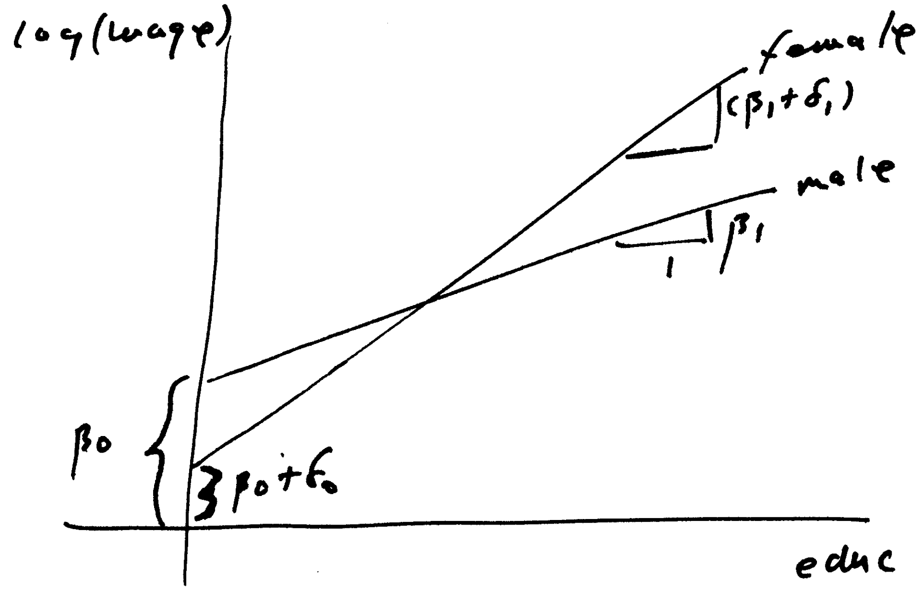
(Multiplicative)

Interactive Dummy Variables

Consider the following additive / multiplicative dummy variable specification of the log(wage) equation:

$$\log(wage) = \beta_0 + \delta_0 \, female + \beta_1 \, educ + \delta_1 \, female \cdot educ + u \quad (3)$$

The PDFs of the male and female groups are graphically represented by (if $\delta_2 < 0$, $\delta_1 > 0$)



Using both multiplicative and additive dummies we allow the groups to have different intercepts and different slopes in their PRFs.

For exaples of the use of interactive (multiplicative) dummies see exaples 7.10 and 7.11 in Wooldridge.

Testing for Differences in Regression Functions Across Groups: The Chow Test

In equation (3) above, the log (wage) equations of the male and female groups are the same if $H_0: \delta_0 = \delta_1 = 0$. This is the called the Chow Test and represents testing for Differences in Regression equations across groups. Such testing can be carried out in either of two equivalent ways: By the (1) "Separate Regressions" Approach or by the (2) "Additive / Multiplicative Dummy Variable" Approach.

Separate Regressions Approach

Consider the log(wage) equation

$$\log(wage)_i = \beta_0 + \beta_1 educ_i + u_i$$

where we have arranged the data so that the female observations, say n_f, are first in order and the male observation, say n_m, follow. Then $N = n_f + n_m$. First we apply OLS to the female regression

$$\log(wage)_i = \beta_0 + \beta_1 educ_i + u_i \quad , \quad i = 1, 2, \cdots, n_f,$$

and obtain the sum of squared residuals, say SSR_1. Then we apply OLS to the male regression

$$\log(wage_i) = \gamma_0 + \gamma_1 educ_i + u_i \, , \quad i = n_f + 1, \cdots, n_f + n_m = N,$$

and obtain the sum of squared residuals, say SSR_2.

Finally, we apply OLS to the combined (pooled) regression

$$\log(wage)_i = \beta_0 + \beta_1 educ_i + u_i \ , \ i = 1, 2, \cdots, n_f, n_f+1, \cdots, n_f+n_m = N,$$

where we have assumed $\beta_0 = \gamma_0$ and $\beta_1 = \gamma_1$ in pooling the regression. Let the sum of squared residuals of the pooled regression be denoted by SSR_r (the restricted residual sum of squares). Then the Chow F-test of the regression functions being the same is computed as

$$F_{2,N-4} = \frac{[SSR_r - (SSR_1 + SSR_2)]/2}{(SSR_1 + SSR_2)/(N-4)} .$$

If the p-value of this F-statistic is greater than α (say $\alpha = .05$) then we accept the null hypothesis that $H_0 : \beta_0 = \gamma_0$ and $\beta_1 = \gamma_1$ and the two regression functions are the same for the two groups. Otherwise we reject H_0 and accept the alternative hypothesis that the regression functions of the two groups are different in some way, either different intercepts ($\beta_0 \neq \gamma_0$) or different slopes ($\beta_1 \neq \gamma_1$) or both.

Multiplicative/Additive Dummy Variable Approach

Consider the Additive/multiplicative Dummy Variable specification

$$\log(wage) = \beta_0 + \delta_0 \, female + \beta_1 \, educ + \delta_1 \, female \cdot educ + u. \quad (3)$$

Here we have the "additive" dummy, female, and the "multiplicative" (interactive) dummy, female·educ. We pool all of the data together, regardless of the order of the male and female observations, and apply OLS to equation (3). Label the sum of squared residuals, SSR_{ur}, the unrestricted SSR. (up to minute rounding error, we should find that $SSR_1 + SSR_2$ of the separate regressions approach should equal SSR_{ur} calculated in the OLS estimation of (3).) To conduct the Chow Test we compute

$$F_{2,N-4} = \frac{(SSR_r - SSR_{ur})/2}{SSR_{ur}/(N-4)} .$$

This test statistic is, up to minute rounding error, the same as the F-statistic we get using the "separate regressions" approach.

For an example of the Chow Test see eqs. (7.20) - (7.22) in Wooldridge.

To generalize the Chow test to $k > 1$ explanatory variables, let there be two groups $g = 1$ and $g = 2$ and let the two separate regression functions be denoted by

$$y = \beta_{g,0} + \beta_{g,1} x_1 + \beta_{g,2} x_2 + \cdots + \beta_{g,k} x_k + u,$$

for $g = 1$ and $g = 2$. The Chow Test then consists of the null hypothesis (of no difference)

$$H_0 : \beta_{0,1} = \beta_{0,2}, \beta_{1,1} = \beta_{1,2}, \cdots, \beta_{1,k} = \beta_{2k}.$$

The Chow F-statistic is then

$$F = \frac{[SSR_R - (SSR_1 + SSR_2)]/(k+1)}{(SSR_1 + SSR_2)/(N-2(k+1))}$$

if we use the "separate regressions" approach. (Wooldridge uses SSR_p for the above restricted sum of squared residuals. SSR_p stands for the "pooled" sum of squared residuals.)

When conducting a Chow Test, my preference is to work with the Additive/multiplicative Dummy variable specification (see eq. (3) for example) because of two primary reasons:

(i) Although it takes a little effort to create the requisite additive and multiplicative dummy variables, once done, it is easy to use a test statement in SAS PROC REG to compute the Chow F-statistic.

(ii) If the Chow F-statistic is statistically significant (i.e. p-value $< \alpha$), then individual t-statistics on the additive and multiplicative dummies can be used to pinpoint the specific nature of the structural change, i.e. the structural ~~change~~ difference may only occur in the intercept or in particular differential slopes when ~~talking about~~ considering k explanatory variables.