Ch 7 in Wooldridge"Multiple Regression Analysis with Qualitative Information:

Binary (or Dummy) Variables"

Read Section 7.1

Use data in "W_wage1.dta" to estimate the following equation by OLS:

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u. \quad (7.1)$$

Read Section 7.2: What is the following equal to? Why?

$$\mathrm{E}(wage|female = 1, educ) - \mathrm{E}(wage|female = 0, educ).$$

$$\widehat{wage} = \underset{(0.72)}{-1.57} - \underset{(0.26)}{1.81}\, female + \underset{(.049)}{.572}\, educ + \underset{(.012)}{.025}\, exper + \underset{(.021)}{.141}\, tenure \quad (7.4)$$

$$n = 526,\ R^2 = .364.$$

Read the explanations to Example 7.1

Read "Interpreting Coefficients on Dummy Explanatory Variables When the Dependent Variable Is log(y)"

EXAMPLE 7.5
(Log Hourly Wage Equation)

Let us reestimate the wage equation from Example 7.1, using log(*wage*) as the dependent variable and adding quadratics in *exper* and *tenure*:

$$\begin{aligned}\widehat{\log(wage)} = {} & .417 - .297\,female + .080\,educ + .029\,exper \\ & \quad (.099) \quad (.036) \quad\quad (.007) \quad\quad (.005) \\ & - .00058\,exper^2 + .032\,tenure - .00059\,tenure^2 \\ & \quad (.00010) \quad\quad (.007) \quad\quad (.00023) \\ & n = 526,\ R^2 = .441.\end{aligned} \tag{7.9}$$

Using the same approximation as in Example 7.4, the coefficient on *female* implies that, for the same levels of *educ*, *exper*, and *tenure*, women earn about 100(.297) = 29.7% less than men. We can do better than this by computing the exact percentage difference in predicted wages. What we want is the proportionate difference in wages between females and males, holding other factors fixed: $(\hat{wage}_F - \hat{wage}_M)/\hat{wage}_M$. What we have from (7.9) is

$$\widehat{\log(wage_F)} - \widehat{\log(wage_M)} = -.297.$$

Exponentiating and subtracting one gives

$$(\hat{wage}_F - \hat{wage}_M)/\hat{wage}_M = \exp(-.297) - 1 \approx -.257.$$

This more accurate estimate implies that a woman's wage is, on average, 25.7% below a comparable man's wage.

Read Section 7.3:

What is a "base group" or a "base category"?

What is an "ordinal" variable?

** Use "W_lawsch85" data to estimate the equation in the Example 7.8:
Two alternative specifications:
1) Use "rank" as a single variable
2) Create "rank dummies"

EXAMPLE 7.8
(Effects of Law School Rankings on Starting Salaries)

Define the dummy variables *top10*, *r11_25*, *r26_40*, *r41_60*, *r61_100* to take on the value unity when the variable *rank* falls into the appropriate range. We let schools ranked below 100 be the base group. The estimated equation is

$$\begin{aligned}\widehat{\log(salary)} = {} & 9.17 + .700\ top10 + .594\ r11\_25 + .375\ r26\_40 \\ & \quad (0.41) \quad\ (.053) \quad\ (.039) \quad\ (.034) \\ & + .263\ r41\_60 + .132\ r61\_100 + .0057\ LSAT \\ & \quad (.028) \quad\ (.021) \quad\ (.0031) \\ & + .014\ GPA + .036 \log(libvol) + .0008 \log(cost) \\ & \quad (.074) \quad\ (.026) \quad\ (.0251) \\ & n = 136,\ R^2 = .911,\ \bar{R}^2 = .905.\end{aligned} \tag{7.13}$$

We see immediately that all of the dummy variables defining the different ranks are very statistically significant. The estimate on *r61_100* means that, holding *LSAT*, *GPA*, *libvol*, and *cost* fixed, the median salary at a law school ranked between 61 and 100 is about 13.2% higher than that at a law school ranked below 100. The difference between a top 10 school and a below 100 school is quite large. Using the exact calculation given in equation (7.10) gives exp(.700) − 1 ≈ 1.014, and so the predicted median salary is more than 100% higher at a top 10 school than it is at a below 100 school.

As an indication of whether breaking the rank into different groups is an improvement, we can compare the adjusted *R*-squared in (7.13) with the adjusted *R*-squared from including *rank* as a single variable: the former is .905 and the latter is .836, so the additional flexibility of (7.13) is warranted.

Interestingly, once the rank is put into the (admittedly somewhat arbitrary) given categories, all of the other variables become insignificant. In fact, a test for joint significance of *LSAT*, *GPA*, log(*libvol*), and log(*cost*) gives a *p*-value of .055, which is borderline significant. When *rank* is included in its original form, the *p*-value for joint significance is zero to four decimal places.

Section 7.4 Interactions Involving Dummy Variables

Interactions Among Dummy Variables

Allowing for Different Slopes

EXAMPLE 7.10
(Log Hourly Wage Equation)

$$\widehat{\log(wage)} = \underset{(.119)}{.389} - \underset{(.168)}{.227}\,female + \underset{(.008)}{.082}\,educ - \underset{(.0131)}{.0056}\,female\text{-}educ + \underset{(.005)}{.029}\,exper - \underset{(.00011)}{.00058}\,exper^2 + \underset{(.007)}{.032}\,tenure - \underset{(.00024)}{.00059}\,tenure^2 \qquad (7.18)$$

$$n = 526,\ R^2 = .441.$$

The estimated return to education for men in this equation is .082, or 8.2%. For women, it is .082 − .0056 = .0764, or about 7.6%. The difference, −.56%, or just over one-half a percentage point less for women, is not economically large nor statistically significant: the t statistic is −.0056/.0131 ≈ −.43. Thus, we conclude that there is no evidence against the hypothesis that the return to education is the same for men and women.

The coefficient on *female*, while remaining economically large, is no longer significant at conventional levels ($t = -1.35$). Its coefficient and t statistic in the equation without the interaction were −.297 and −8.25, respectively [see equation (7.9)]. Should we now conclude that there is no statistically significant evidence of lower pay for women at the same levels of *educ*, *exper*, and *tenure*? This would be a serious error. Since we have added the interaction *female-educ* to the equation, the coefficient on *female* is now estimated much less precisely than it was in equation (7.9): the standard error has increased by almost five-fold (.168/.036 ≈ 4.67). The reason for this is that *female* and *female-educ* are highly correlated in the sample. In this example, there is a useful way to think about the multicollinearity: in equation (7.17) and the more general equation estimated in (7.18), δ_0 measures the wage differential between women and men when $educ = 0$. As there is no one in the sample with even close to zero years of education, it is not surprising that we have a difficult time estimating the differential at $educ = 0$ (nor is the differential at zero years of education very informative). More interesting would be to estimate the gender differential at, say, the average education level in the sample (about 12.5). To do this, we would replace *female-educ* with *female*-(*educ* − 12.5) and rerun the regression; this only changes the coefficient on *female* and its standard error. (See Exercise 7.15.)

If we compute the F statistic for H_0: $\delta_0 = 0$, $\delta_1 = 0$, we obtain $F = 34.33$, which is a huge value for an F random variable with numerator $df = 2$ and denominator $df = 518$: the p-value is zero to four decimal places. In the end, we prefer model (7.9), which allows for a constant wage differential between women and men.

7.12 Use the data in GPA2.RAW for this exercise.

(i) Consider the equation

$$colgpa = \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + \beta_3 hsperc + \beta_4 sat + \beta_5 female + \beta_6 athlete + u,$$

where *colgpa* is cumulative college grade point average, *hsize* is size of high school graduating class, in hundreds, *hsperc* is academic percentile in graduating class, *sat* is combined SAT score, *female* is a binary gender variable, and *athlete* is a binary variable, which is one for student-athletes. What are your expectations for the coefficients in this equation? Which ones are you unsure about?

(ii) Estimate the equation in part (i) and report the results in the usual form. What is the estimated GPA differential between athletes and nonathletes? Is it statistically significant?

(iii) Drop *sat* from the model and reestimate the equation. Now what is the estimated effect of being an athlete? Discuss why the estimate is different than that obtained in part (ii).

(iv) In the model from part (i), allow the effect of being an athlete to differ by gender and test the null hypothesis that there is no ceteris paribus difference between women athletes and women nonathletes.

(v) Does the effect of *sat* on *colgpa* differ by gender? Justify your answer.

7.15 Use the data in WAGE1.RAW for this exercise.

(i) Use equation (7.18) to estimate the gender differential when $educ = 12.5$. Compare this with the estimated differential when $educ = 0$.

(ii) Run the regression used to obtain (7.18), but with $female \cdot (educ - 12.5)$ replacing $female \cdot educ$. How do you intepret the coefficient on *female* now?

(iii) Is the coefficient on *female* in part (ii) statistically significant? Compare this with (7.18) and comment.

Testing for Differences in Regression Functions Across Groups

Read the text.

Use the data in "W_gpa3.dta" to estimate equation (7.20).

Test the hypothesis in equation (7.21).

Examine the F-test in equation (7.24)

What is a Chow test? How is it applied? How is is different from the F-test that we used before?

Section 7.5 A Binary Dependent Variable: The Linear Probability Model

Now suppose that the dependent variable has a qualitative meaning (rather than a quantitative meaning).

Consider the case of the binary outcome.

Why is P(y=1|x) = E(y|x) in equation (7.27) ?

$$P(y = 1|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k, \qquad (7.27)$$

$$\Delta P(y = 1|\boldsymbol{x}) = \beta_j \Delta x_j. \qquad (7.28)$$

What is the main advantage of the linear probability model?

What is the main disadvantage of the linear probability model?

Use "W_mroz.dta" to estimate equation (7.29).

Use "W_crime1.dta" to estimate equation (7.31).