Problem Set 1 - Week 38, 2013

Problem 1

Consider the linear regression model

$$y_i = \beta_1 + f_i\beta_2 + \epsilon_i,$$

where y_i is the height of individual i, and f_i is a gender dummy, which takes values $f_i = 1$ for females and $f_i = 0$ for males. We observe n_f females and n_m males. The total sample size is $n = n_f + n_m$. Let $\bar{y}_f$ be the average of y_i in the female subsample, and $\bar{y}_m$ be average of y_i in the male subsample. Let $x_i = (1, f_i)$, and let y be the $n \times 1$ vector with entries y_i, and X be the $n \times 2$ matrix with rows x_i.

(a) Somebody proposes to also include a male dummy $m_i = 1 - f_i$ into the model and to estimate the regression $y_i = \beta_1 + f_i\beta_2 + m_i\beta_3 + \epsilon_i$ by OLS. Explain why this is problematic.

(b) Somebody proposes to also include the square f_i^2 of the female dummy into the model and the estimate the regression$y_i = \beta_1 + f_i\beta_2 + f_i^2\beta_3 + \epsilon_i$ by OLS. Explain why this is problematic.

(c) Give expressions for the 2×2 matrix $X'X$ and the 2×1 vector $X'y$ in terms of n_f , n_m, y_f and y_m.

(d) Now assume that $n_f = n_m = 100$, that $\bar{y}_f = 165$, and that $\bar{y}_m = 175$. Calculate the OLS estimator for β_1 and β_2.

(e) In addition to the assumptions in (d) now also assume that $Var(\epsilon_i|f_i) = 50$. Calculate the estimated standard error for $\hat{\beta}_2$.

(f) Consider the null hypothesis $H_0 : \beta_2 = 0$. Using your results in (d) and (e) calculate the t-test statistics for testing H_0. Would you reject H_0 at 5% significance level?

Problem 2

(a) Show that, under random sampling and the zero conditional mean assumption $E(\epsilon|X) = 0$, $E(\hat{\beta}|X) = \beta$ if $X'X$ is nonsingular.

(b) In addition to the assumptions from part (1), assume that $Var(\epsilon|X) = \sigma^2$. Show that $Var(\hat{\beta}|X) = \sigma^2(X'X)^{-1}$

Problem 3

(a) Load the MROZ.csv data into R (can be found on blackboard)

(b) Run some summary statistics. Are there any variables with missing values? Why might there be missing values in this (these) variables?

(c) Estimate the following model[1]

$$\log(wage) = \beta_0 + \beta_1 exper + \beta_2 exper^2 + \beta_3 educ + \beta_4 age + \beta_5 kidslt6 + \beta_6 kidsge6 + u$$

with the normal standard errors, for the 428 employed women in the sample. Compare your results with Example 4.1 in Graduate Wooldridge (2002).

(d) Estimate the model in problem 3 with heteroscedasticity robust standard errors. Compare your results with Example 4.1.

(e) Write down the hypothesis that education does not affect wages in both cases, and use the t-values from the above regressions to reach a conclusion about the significance of education.

(f) Use the F-statistics to test the hypothesis of $\beta_4 = \beta_5 = \beta_6 = 0$ as in the example.

(f.1) Run and save the restricted regression

(f.2) Use the 'anova' command to perform the F-test: "anova(*restricted_model, unrestricted_model*)

(g) Test the same hypothesis using the LM statistic

(g.1) Extract the residuals from the restricted regression

(g.2) Regress the restricted residuals on the full set of explanatory variables (including the variables you are testing)

(g.3) The test statistic is then $N * R^2$, where N is the number of observations and R^2 is the R squared from the regression from step b. Under the null, this is chi-squared distributed with k degrees of freedom, where k is the number of restrictions (in this case k=3).[2]

(h) Plot wage-experience profiles for different education levels. Interpret them.

(i) Test the hypothesis of no effect of experience on wages (note that there is both an '*exper*' and '*expersq*' term in the regression).

[1] Equation (4.16) in Wooldrige (2010,2002)

[2] Use the *pchisq(X, df=k)* command to display the cumulative chi-squared distribution function where X is the test statistic and k is the degree of freedom. The probability of the null hypothesis is then *1-pchisq(X, df=k)*.

(j) Include an interaction term between education and experience. How would you interpret that? Is it significant?

(k) Can you make a model that fits better than the one in problem 3?

(l) Discuss whether assumptions OLS.1 - OLS.3 are likely to be fulfilled.