Introduction to Econometrics
Problem Set 5
Due date: April 4th at the beginning of class

You may collaborate with other students. Please hand in your own work. At the top of your answer sheet, please identify the students with whom you have consulted to complete this problem set.

The data for this problem set are available on the class webpage. Click on the link to the data file and save the data on your local disk.

---

1. In each case state which, if any, of the "Big 4" assumptions you think are violated.

   (a) $y_i = \beta_0 + \beta_1 x_{i1} + u_i$ where $u_i \sim N(0, 5\sigma^2)$.

   (b) $y_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 z_{i3} + u_i$ where $E(u_i) = 0$ and $Var(u_i) = \sigma_i^2$.

   (c) $income_i = \beta_0 + \beta_1 education_i + u_i$ where $u_i \sim N(0, \sigma^2)$ and u_i includes unobserved information like mother's education and father's education.

2. Multiple choice - *briefly* explain your answer:

   (a) In the multivariate regression model $\log(wages_i) = \beta_0 + \beta_1 age_i + u_i$, if $u_i \sim N(0, 0.2\ \sigma_i^2)$ and A4 is satisfied:

      i. The regression is not possible with OLS
      ii. The estimated slope parameter will be biased
      iii. The estimated slope parameter will be unbiased
      iv. All of the above
      v. None of the above

   (b) For a two sided t test of $H_0 : \beta_1 = 0$ at the 5% level, if n is "large enough" and the t-statistic = -2, you can:

i. Reject test of $H_0 : \beta_1 = 0$ and Reject test of $H_0 : \beta_1 \leq 0$

ii. Reject test of $H_0 : \beta_1 = 0$ and Fail to reject test of $H_0 : \beta_1 \leq 0$

iii. Fail to reject test of $H_0 : \beta_1 = 0$

iv. Say nothing - you would need to know more information

(c) In the regression model $y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (D_i \cdot X_i) + u_i$, where X is a continuous variable and D is a binary (1 or 0) variable, β_3:

i. Indicates the slope of the regression when $D = 1$.

ii. Has a standard error that is not normally distributed as D is not normally distributed.

iii. Indicates the difference in the slopes for those with $D_i = 1$ and those with $D_i = 0$.

iv. Has no meaning since $D_i \cdot X_i = 0$ when $D_i = 0$.

3. For this question use the data in **ps5q3.dta** [1].

(a) Begin by estimating and reporting the results from the following equation:

$$log(wage)_i = \beta_0 + \beta_1 educ_i + u_i \tag{1}$$

use the command "describe" in Stata to see the description of each variable.

(b) Interpret $\hat{\beta}_1$. Is education a significant predictor of log(wages) at a 5% significance level? How do you know?

(c) Add the variable $educ^2$. Is education *educ* a significant predictor of log(wages) at a 5% significance level? Is your result consistent with your answer in (b)? Why or why not?

(d) How would you change your test in (c) to know whether education significantly affects log(wages)? Why is this change necessary?

(e) Extend the original model to allow the initial wages to depend on race and the return to education to depend on race and test whether the return to education does depend on race (use a 5% size of the test). Report the value of your test-statistic and it's p-value.

[1] These data was used in M. Blackburn and D. Neumark (1992), "Unobserved Ability, Efficiency Wages, and Interindustry Wage Differentials," Quarterly Journal of Economics 107, 1421-1436. This question is based on question C7.2 from Wooldridge.

(f) What is the approximate difference in monthly salary between blacks and nonblacks with the same level of education? Is this difference statistically significant (use a 5% size of the test). Report the value of your test-statistic and it's p-value.

(g) Suppose we wanted to look at regression results by race. Estimate the following model separately for blacks and nonblacks (you will run 2 different regressions where you use the **reg** command with an "if" statement).

$$log(wage)_i = \beta_0 + \beta_1 educ_i + \beta_2 exper + v_i \tag{2}$$

Now we'll use interaction terms in (equation (2)) to look at differences across race (you will need to generate new variables that reflect the proper interactions).

(h) Estimate and report the following equation:

$$log(wage)_i = \alpha_0 + \alpha_2 black_i + \alpha_3 educ_i + \alpha_4 educ_i * black_i + \alpha_5 exper_i + \alpha_6 exper_i * black_i + e_i \tag{3}$$

(i) Is the impact of education statistically different for blacks and nonblacks at a 5% significance level? Is the impact of experience statistically different for blacks and non blacks at a 5% significance level? How do you know?

(j) Test the null hypothesis that there are no overall differences by race (use the appropriate STATA command). What is the F statistic? Based on the p-value, can you reject your null? How do you interpret these results?

(k) Estimate the Restricted and Unrestricted model and report its SSRs.

(l) Calculate the F statistic according to the SSR formula. Your results should match what Stata calculated for the F stat in question (j).

(m) How are the estimates from equation (3) related to those from part g)?

4. Economists are interested in examining the connection between local economic conditions and obesity. There exists data with state-level obesity rates and economic conditions across a series of years. Data from 2011 is included in **ps5q4.dta**. Indicator variables for each of the 4 regions of the United States are included that equal 1 if the observation is from a state in a particular region and 0 otherwise. For example, "northeast" equals 1 if the observation is from Connecticut, Delaware, Maine, Maryland, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, or Vermont and 0 otherwise.

We'll start with describing the data to get a sense of the prevalence of obesity in America.

(a) Estimate and report the following regression for obesity rate (0 to 100%) in state s and report your results (each of the indicator variables already exists in the data):

$$obesrt_s = \alpha_0 + \alpha_1 southeast_s + \alpha_2 midwest_s + \alpha_3 west_s + u_s$$

(b) How do your coefficient estimates compare to the mean obesity rate in each region? Briefly explain why you see the pattern that you see.

(c) To get a sense of how economic conditions (in this case the unemployment rate (0 to 100%), log of median income, and high school graduation rate) connect to obesity rates, estimate and report the following regression (the geographic abbreviations represent the same indicators as above):

$$obesrt_s = \beta_0 + \beta_1 urate_s + \beta_2 \log(medinc_s) + \beta_3 hsgrad_s + \beta_4 southeast_s + \beta_5 midwest_s + \beta_6 west_s + u_s$$

(d) How do you interpret the estimates of β_1 and β_2 (careful of the units)? Are they each statistically significant at the 5% level?

(e) There's a debate in economics about whether recessions are good or bad for your health. Provide you own intuition for why you might think the sign of $\hat{\beta}_1$ is correct.

(f) Write down the joint null hypothesis that corresponds to the following statement: 'The obesity rate has no important regional differences"

(g) Write down and estimate the restricted model to match your hypothesis in (f).

(h) Calculate the F statistic for the test using the SSR's from your unrestricted model in (c) and the restricted model in (g). (You can check this is correct by using the `test` command with your null hypothesis after running the unrestricted model.)

5. The data in **ps5q5.dta** contains school level data on test performance and other school characteristics from the 1994 4th grade Michigan Education Assessment Program (MEAP) standardized test for a random sample of Michigan elementary schools. This is a common type of data that education policy-makers would use to analyze features such as student to teacher ratios, the impact of subsidized lunch programs, etc.

(a) Plot and display a scatter plot of the percent of students who passed the math exam (`math4`) against the percent of students in a school who are eligible for free or reduced lunch (`lunch`).

(b) Based on your graph, do you think that you should be worried about heteroskedasticity in a regression of test scores on lunch eligibility? Why or why not?

(c) Run the following regression assuming homoskedasticity:

$$math4_s = \beta_0 + \beta_1 lunch_s + \beta_2 exppp_s + u_s \tag{4}$$

where $lunch_s$ is the percent of students eligible for reduced lunches and $exppp_s$ is expenditure per pupil (in \$100s). Report your results.

(d) Conduct the Breusch-Pagan and White Tests on your model. What would you conclude about Assumption 3?

(e) Reestimate equation (4) using robust standard errors (`reg y x1 x2, robust`). Report your results.

(f) How do the coefficient estimates compare?

(g) How do the p-values compare? Does allowing for heteroskedasticity change your interpretation of the regression? If so, how?