ECON 535 Spring 2026: Problem Set 1
Econometrics II
Dr. Qureshi Due: 2026-02-17

Analytic Problems

1. Consider the model

$$\underbrace{y}_{n\times 1} = \underbrace{\mathbf{X}}_{n\times k_1}\underbrace{\beta}_{k_1\times n} + \underbrace{\mathbf{Z}}_{n\times k_2}\underbrace{\delta}_{n\times k_2} + \underbrace{\varepsilon}_{n\times 1}$$

   Let $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Now consider the regressions:

   (a) $\mathbf{M}y = \mathbf{Z}\gamma + \varepsilon_1$
   (b) $y = \mathbf{Z}\gamma + \varepsilon_2$
   (c) $y = \mathbf{MZ}\gamma + \varepsilon_3$
   (d) $\mathbf{M}y = \mathbf{MZ}\gamma + \varepsilon_4$

   For each of the regressions, show whether γ yields unbiased estimates of δ. If a regression yields biased estimates, use your derivation to discuss the conditions under which it would yield unbiased estimates.

2. Show that the regression of y on a constant, x_2, and x_3 while imposing the restriction $\beta_2 - \beta_3 = 1$ leads to the regression of $y - x_2$ on $x_2 + x_3$.

3. Consider the model:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i$$

   For the hypothesis $H_0 : \beta_2 = 1,\ \beta_3 = \beta_4$:

   (a) propose a test statistic (with specific, rather than general, notation)
   (b) state what is its distribution under the null hypothesis and why
   (c) explain the conditions under which you would reject H_0 and explain the intuition behind using these conditions
   (d) explain and show your test statistic can be computed from two regressions

4. Suppose you are interested in the causal effect of belonging to a labor union on an individual's earnings and have a random sample of workers with which to study this problem. Let Y_i be the earnings of person i; U_i is an indicator for union status. Suppose $Y_{1,i}$ is the potential earnings of person i if her union status is equal to one and $Y_{0,i}$ is her potential earnings if union status is equal to zero.

   (a) Give an intuitive explanation (i.e. words, not an equation) for the difference between $\mathbb{E}[Y_{1,i} - Y_{0,i}]$ and $\mathbb{E}[Y_{1,i} - Y_{0,i} | U_i = 1]$.
   (b) Give an intuitive explanation for the difference between $(Y_{1,i} - Y_{0,i})$ and $\mathbb{E}[Y_{1,i} - Y_{0,i}]$.
   (c) Describe what the quantity $\mathbb{E}[Y_i | U_i = 1]$ is.
   (d) Provide an intuitive explanation for the quantity $\mathbb{E}[Y_i | U_i = 1] - \mathbb{E}[Y_i | U_i = 0]$.

(e) What is the relationship between $(\mathbb{E}[Y_i|U_i = 1] - \mathbb{E}[Y_i|U_i = 0])$ and $\mathbb{E}[Y_{1,i} - Y_{0,i}|U_i = 1]$?

(f) Suppose you run a regression $Y_i = c + \beta U_i + \varepsilon_i$. Explain precisely what the estimate of the constant term and the estimate of β will be.

(g) Explain the following two statements. Then discuss whether condition ii. is satisfied if condition i. is satisfied:

i. "The OLS estimator of β will be unbiased if $\mathbb{E}[\varepsilon_i|U_i] = 0$."

ii. "The difference in average earnings between union members and non-members will be an unbiased estimate of the causal effect of unions on union members if $\mathbb{E}[Y_{0,i}|U_i = 1] = \mathbb{E}[Y_{0,i}|U_i = 0]$."

Computational Problems

1. For these problems, download the Stata dataset `ec535_hw1.dta` from Blackboard.

(a) Use the summary command to find the means of the following variables: earnings, welfare income, total income, education, age, and age squared. Comment on the sample size corresponding to the different variables.

(b) Now use the regress command to separately regress earnings, welfare income, and total income on a constant. Compare the regression results to the means from above. Explain their relationship.

(c) Now regress earnings (`earn`) on education (`higrade`), age, and age squared (`agesq`). Variable names given in parentheses.

i. Consider the role of a constant term:

A. Verify that you get the same coefficient of the effect of education if you regress earnings on education, age, and age squared (that you have already done) and when you run the following regression (this requires running multiple regressions before you can run the following one):

$$earn^* = \gamma_0 + \gamma_1 higrade^* + \eta,$$

where $earn^*$ are the residuals from regressing earnings on age and age squared, and $higrade^*$ are the residuals from regressing education on age and age squared.

B. Verify that the regression in part 1(c)iA gives you the same answers as running the same regression without a constant, and explain why this is the case.

ii. Verify that when you regress earnings on education, age, and age squared with a constant, and when you run the following demeaned regression (without a constant), you get the same slope coefficients:

$$\widetilde{earn} = \alpha_1 \widetilde{higrade} + \alpha_2 \widetilde{age} + \alpha_3 \widetilde{agesq} + \tilde{v},$$

where $\tilde{w}$ is defined as the demeaned variable w, i.e., $\tilde{w} = w - \bar{w}$.

(d) Retrieve the residuals and predicted values from the above regression, and use them to verify the following properties:

i. the residuals sum to zero

ii. the mean of the predicted values equals the mean of the dependent variables

iii. the residuals are orthogonal to the regressors and the predicted values

iv. the square of the correlation coefficient between the dependent variable and the predicted values equals the R^2 from the regression.

(e) Re-estimate the regression from part 1(c)iA, this time omitting the constant term. Which of properties from part 1d still hold, if any? Explain.

2. For these problems, reuse the same dataset from the previous problem (`ec535_hw1.dta`).

(a) Regress earnings on education, age, and age squared.

i. What is the standard error of the regression? What is the standard error of the education coefficient? The age coefficient?

ii. Interpret the education coefficient.

(b) Based on the same regression, assume that the GM assumptions hold and carry out tests of the following hypothesis:

i. Carry out the following test using a command to test whether the coefficient is statistically different from zero (in Stata, this would be the `test` command).

$$H_0 : \beta_{educ} = 0, H_A : \beta_{educ} \neq 0$$

ii. Now, test the above hypothesis by running two separate regressions and calculating the F statistic using information from the regression output as shown in class.