ECON 695 Problem Set 2

September 10, 2025

Due: 11:59PM Central Time on Friday, 09/26/2025. Upload a photocopy of your handwritten solutions, or a PDF generated by LaTex/lyx to Canvas.

Note 1: I highly recommend using lyx to answer non-coding questions. Lyx is a free document processor that interfaces with LaTeX. You will be able to type and see your equations directly, and export the .lyx file to a PDF.

Note 2: I provide a Jupyter notebook (**PS2_oaxaca.ipynb**) to guide the coding exercise (question 4). Under "lecture2" on Canvas, I also uploaded a Jupyter notebook for running regressions in Python. You can use it as reference.

1 Population Regression vs CEF

In lecture 2, we considered three things:y_i, $E[y_i|x_i]$ and $x_i'\beta^*$ the "infeasible" or population OLS estimator. We showed that we can write:

$$y_i = x_i'\beta^* + u_i$$
$$\text{where } \beta^* = argmin_b \, E[(y_i - x_i'b)^2]$$

1. Show that $E[x_i u_i] = 0$.
2. Show $\beta^* = E[x_i x_i']^{-1} E[x_i y_i]$.
3. Show $E[u_i|x_i] = 0 \Rightarrow E[x_i u_i] = 0$. Hint: Apply the Law of Iterated Expectations.
4. List two cases where $E[u_i|x_i] = 0$. Clearly state your assumption and explain why $E[u_i|x_i] = 0$ under such assumptions.

2 Regressions with Dummies

Suppose we have a sample of size N made up of two groups, denoted 0 and 1. Let D_i be a dummy variable with $D_i = 1$ indicating membership in group 1. Finally, let N_0 and N_1 represent the size of the two subgroups (so $N = N_0 + N_1$).

a) Consider an OLS regression model:

$$y_i = \alpha + \beta D_i + u_i$$

Using the first order conditions for the OLS estimates ($\widehat{\alpha}$, $\widehat{\beta}$), show that

$$\begin{aligned} \widehat{\alpha} &= \frac{1}{N_0} \sum_{i \in 0} y_i, \\ \widehat{\alpha} + \widehat{\beta} &= \frac{1}{N_1} \sum_{i \in 1} y_i \end{aligned}$$

b) Now consider a case where there are 3 groups: $0, 1, 2$ and there are 2 dummies: $D_{1i} = 1$ if i is in group 1, and 0 otherwise, and $D_{2i} = 1$ if i is in group 2, and 0 otherwise, with the regression model:

$$y_i = \alpha + \beta_1 D_{1i} + \beta_2 D_{2i} + u$$

i) find the first order conditions for minimizing the sum of squared residuals, which define the OLS estimator.

ii) using the FOC show that the OLS estimators for the 3-group model will have

$$\begin{aligned} \widehat{\alpha} &= \frac{1}{N_0} \sum_{i \in 0} y_i \\ \widehat{\alpha} + \widehat{\beta}_1 &= \frac{1}{N_1} \sum_{i \in 1} y_i \\ \widehat{\alpha} + \widehat{\beta}_2 &= \frac{1}{N_2} \sum_{i \in 2} y_i \end{aligned}$$

3 Residualized Regression (Frisch-Waugh)

In lecture4, we showed that the j^{th} row of the population regression coefficient β^* from the model

$$y_i = x_i \beta^* + u_i$$

can be obtained by first getting the residual from an auxiliary regression of x_{ji} on all the other $x's$:

$$x_{ji} = x'_{(\sim j)i} \pi + \xi_i$$

then forming:

$$\beta_j^* = E[\xi_i^2]^{-1} E[\xi_i y_i]$$

Suppose that we also "residualized" the dependent variable using an auxiliary regression of y_i on all the other $x's$:

$$y_i = x'_{(\sim j)i} \lambda + \phi_i$$

Show that its also true that:

$$\beta_j^* = E[\xi_i^2]^{-1} E[\xi_i \phi_i]$$

(In other words, we could residualize BOTH x_{ji} AND y_i and still get the same right answer).

4 Oaxaca Decomposition

The goal of this problem set is to extend the analysis of public-private pay gaps in Lecture 3 to *male workers*.

The file **pay.csv** contains 45,404 observations on male workers, age 30-50, with education of 12 or more years, in the March 2012 and 2013 CPS files. The variables are:

govt=1 if government worker
logwage=log hourly wage based on earnings, weeks worked, and average hours per week last year
educ = years of education
hs = 1 if high school grad (12 years of schooling)
somecoll=1 if education is 13-14 years.
college = 1 if BA (16 years of education)
ma = 1 if masters degree (18 years of schooling)
phd = 1 if phd or LLD or medical degree (20 years of schooling)
age= age in years (range 30-50)

NOTE: note that "somecoll" combines 2 levels of education, 13 and 14 years (13 years are people with some college but no degree, 14 years are people with an AA degree from a community college).

Please create 2 new variables: educ13=1[educ=13] and educ14=1[educ=14] so you will have a total of 6 possible categories for education: **phd, ma, college, educ14, educ13, hs.**

1. Construct a table like the 1st table (page 7) in Lecture 3. Show the means of all the variables for workers in the private (govt=0) and public (govt=1) sectors, and the differences in these means.
2. Construct a table like the 2nd table (page 8) in Lecture 3. Show the fractions of private and government workers at each education level, and the mean wages of the two groups in each education level, and the public-private wage gap at each level.
3. Using the notation of Lecture 3, let private sector workers be called "group a" (who are the reference group), let government workers be called "group b", and define the 6 category variables D_{gi} for each level of education $g = 1...6$. Let $x_i' = (D_{1i}, D_{2i}, ...D_{6i})$. Find $\overline{x}^a$ and $\overline{x}^b$.
   (a) Fit an OLS regression model for wages on x_i for each group, and estimate the coefficients $\hat{\beta}^a$ and $\hat{\beta}^b$. Verify that these coefficients are the mean log wages of groups a and b, respectively, for each level of education.
   (b) Using your estimates of $\overline{x}^a$, $\overline{x}^b$, $\hat{\beta}^a$ and $\hat{\beta}^b$, construct $\overline{y}^a = \sum_g \overline{p}_g^a \overline{y}_g^a = (\overline{x}^a)'\hat{\beta}^a$ and $\overline{y}^b = \sum_g \overline{p}_g^b \overline{y}_g^b = (\overline{x}^b)'\hat{\beta}^b$. Verify that these match the means of wages for groups a and b that you construct directly.
   (c) Using your estimates of $\overline{x}^a$, and $\hat{\beta}^b$ construct

$$\overline{y}^b_{counterf} = \sum_g \overline{p}_g^a \overline{y}_g^b = (\overline{x}^a)'\hat{\beta}^b$$

   Find the adjusted wage gap: $\overline{y}^b_{counterf} - \overline{y}^a$. How does this compare to the actual gap $\overline{y}^b - \overline{y}^a$?
   (d) Construct the weight $w_g = N_g^a / N_g^b$ for education categories $g = 1...6$. Use this to construct $\overline{y}^b_{counterf2} = \frac{\sum_{i \in b} w_g y_i}{\sum_{i \in b} w_g}$ and verify that:

$$\overline{y}^b_{counterf} = \overline{y}^b_{counterf2}.$$

$$\text{that is, } \sum_g \overline{p}_g^a \overline{y}_g^b = \frac{\sum_{i \in b} w_g y_i}{\sum_{i \in b} w_g}$$