Problem Set 2

Chris Conlon

Fall 2025

Econometrics I
NYU Stern

Professor Chris Conlon
Email: cconlon@stern.nyu.edu

1 Prediction errors

The Cornwell and Rupert data for this problem can be downloaded from `http://www.github.com/chrisconlon/applied_metrics`.

Source: Cornwell and Rupert, (1988) ”Efficient estimation with panel data: an empirical comparison of instrumental variables estimators”

(a) For this part of the assignment, you are to replicate the regression ”Mincerian Regression, Cornwell and Rupert Data” from the Linear Regression slides by obtaining the same coefficients and standard errors. If that is possible (which is often not the case), other differences in reported results can usually be explained. As part of your submission for this assignment, include the specific estimation results that you obtained for this regression.

Now that you have replicated the regression, we’ll consider a couple of minor extensions. (b) Functional Form. The example thus far computes a single, generic effect of education on LWAGE. We’re interested in determining if there is a different effect for men (FEM=0) and women (FEM=1). One compact way to do this is to add an interaction term, FEM*ED to the model. The different effects are the coefficient on ED which is for men and the sum of the two effects, ED and FEM*Ed, for women. Reestimate your model with this additional effect, and report your result. (How does this look for a year of schooling, vs the levels of schooling complted?)

(c) Standard Errors. Can you compute the bias-corrected bootstrap confidenfce interval for the difference in log wages between men and women college graduates (controlling for all other variables). Thinking about the right regression to run and the correct $g(\cdot)$ function should be what is tricky here.

3 Data warm-up

The data file is on the Course Page/Github Page and is a CSV file that contains data on fuel bills and number of rooms for 144 homes.

a) Produce a simple scatter (X-Y) plot with ROOMS on the horizontal axis and FUELBILL on the vertical axis. What conclusion do you draw about the relationship between number of rooms and fuelbill?

b) Note that ROOMS only takes a few values, 3,4,5,. . . ,11. Compute the mean value of FUELBILL for the different values of ROOMS. What do you conclude about the conditional mean? Plot the means against the number of rooms. What do you find?

4 Partitioned regression

Suppose a data set consists of $\boldsymbol{y}$ $(n \times 1)$, $\boldsymbol{X}_1$ $(n \times K_1)$ and $\boldsymbol{X}_2$ $(n \times K_2)$. Do the following four procedures produce the same value for the least squares coefficients on $\boldsymbol{X}_2$?

a) Regress $\boldsymbol{y}$ on both $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$.

b) Regress the residuals from a regression of $\boldsymbol{y}$ on $\boldsymbol{X}_1$ on the residuals (column by column) of regressions of $\boldsymbol{X}_2$ on $\boldsymbol{X}_1$.

c) Same as (b), but do not transform $\boldsymbol{y}$.

d) Same as (b), but do not transform $\boldsymbol{X}_2$.

Hint: $\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X'X}\right)^{-1}\boldsymbol{X}'$ is known as the *projection matrix* because

$$\boldsymbol{Py} = \boldsymbol{X}\left(\boldsymbol{X'X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b}_{OLS} = \hat{\boldsymbol{y}}.$$

Define $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{P} = \boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X'X}\right)^{-1}\boldsymbol{X}'$. This matrix is known as the *residual maker* because

$$\boldsymbol{My} = \left(\boldsymbol{I} - \boldsymbol{P}\right)\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{Py} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{e}.$$

You can make progress on this problem by using the residual maker. For example, the matrix of residuals from regressing $\boldsymbol{X}_2$ on $\boldsymbol{X}_1$ is given by

$$\left(\boldsymbol{I} - \boldsymbol{X}_1\left(\boldsymbol{X}_1{}'\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'\right)\boldsymbol{X}_2.$$

5 Change in the sum of squares

Suppose that $\boldsymbol{b}$ is the least squares coefficient vector in the regression of $\boldsymbol{y}$ on $\boldsymbol{X}$ and that $\boldsymbol{c}$ is any other $K \times 1$ vector. Prove that the difference in the two sums of squared residuals is

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{c})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{c}) - (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) = (\boldsymbol{c} - \boldsymbol{b})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{c} - \boldsymbol{b}).$$

A property of the matrix $\boldsymbol{X}'\boldsymbol{X}$ is that is *positive definite*. This means that for any vector $\boldsymbol{u} \neq 0$, $\boldsymbol{u}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{u} > 0$. How does this property and your result above connect to the definition of the least squares estimator?

5 OLS Residuals

(a) Consider the following table of data nad potential residuals:

y	x_1	x_2	e_1	e_2	e_3
?	1	0	1	2	3
?	1	-1	-3	-1	-2
?	1	1	2	-1	1

Which of the potential vectors of residuals $\boldsymbol{e}_1, \boldsymbol{e}_2$, and $\boldsymbol{e}_3$ (if any) could be from a regression of y on $\left[\begin{array}{cc} \boldsymbol{x}_1 & \boldsymbol{x}_2 \end{array}\right]$? Explain. (b) Show that the estimated OLS parameters are unchanged if the dependent variable and all dependent variables are transformed by subtracting their means.

That is, let

$$\left[\begin{array}{c} \beta_0 \\ \boldsymbol{\beta} \end{array}\right] = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{y},$$

where β_0 is the intercept, $\boldsymbol{\beta}$ is a column of the other parameter estimates, and the first column of $\boldsymbol{X}$ is a vector of 1 's as usual. Show that $\boldsymbol{\beta} = \left(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\right)^{-1} \widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{y}}$, where $\widetilde{\boldsymbol{X}}$ drops the first column of $\boldsymbol{X}$ and for the other columns, $\widetilde{\boldsymbol{x}}_k = \boldsymbol{x}_k - n^{-1}\sum_{i=1}^n x_{ki}$. Similarly, $\widetilde{\boldsymbol{y}} = \boldsymbol{y} - n^{-1}\sum_{i=1}^n y_i$.