Problem Set 1

Chris Conlon

Fall 2025

Econometrics I
NYU Stern

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

1 Prediction errors

Define $u \equiv y - E[y|x]$. Show that for any function $g(x)$, $E[g(x)u] = 0$.

2 Multivariate normal distributions

The random variables y, x have a multivariate normal distribution with mean vector $\mu' = [1, 2]$ and covariance matrix

$$\Sigma = \begin{bmatrix} 2 & 3 \\ 3 & 6 \end{bmatrix}$$

a) Define

$$\varepsilon = y - E[y|x]$$

Show that $E[\varepsilon|x] = 0$ and $E[\varepsilon] = 0$

b) It's true that $E[y|x]$ is linear in x, i.e., $E[y|x] = \alpha + \beta x$ (Optional: show this). Given the mean vector and covariance matrix above, what are the values for α and β?

c) Compute the conditional variance $Var[y|x]$. (Hint: you may appeal to the Law of Total Variance.)

d) Compute the squared correlation R^2 between y and x.

e) Compute the squared correlation R^2 between y and $E[y|x]$. Comment, comparing your answer here to part (d).

3 Data warm-up

The data file (which you should download)

http://ptscott.com/teaching/data/fuelbills.csv

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

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

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

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

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?

6 The budget model.

Consider a plan to fit least squares regressions using three dependent variables $y_1,\ y_2,\ y_3$, where y_j is the share of total expenditure on durables, nondurables, and services, respectively. Note that the three budget shares sum to 1.

All three regressions will use the same $\boldsymbol{X}$ matrix which has 5 columns (variables) $\boldsymbol{X} =$ [a constant term, income, P_D, P_N, P_S] where P_m is a price index for the m^{th} expenditure group. Denote the the m^{th} least squares coefficient vector by $\boldsymbol{b}_m$, where $m = D, N, S$.

Prove that the sum of the three least squares coefficient vectors is

$$\boldsymbol{b}_D + \boldsymbol{b}_N + \boldsymbol{b}_S = [1, 0, 0, 0, 0]'.$$

That is, the constant terms sum to 1 and the other coefficients sum to zero.

Now, suppose instead of budget shares, we have expenditure data. Moreover, though we would like to use income as the second independent variable, we have only total expenditure, the sum of the three expenditures. Now, what do you get when you add the three least squares coefficient vectors? Prove your answer.