Problem Set 1

Econometrics I
NYU Stern

Professor Paul T. Scott
Email: ptscott@stern.nyu.edu

* Problems marked with an asterisk are optional and will count only for imaginary bonus points.

Problem 1 Distribution Theory

Hint: this problem can be solved using conditional distributions, but it is much easier if you use shortcuts. Consider equation (4-34) from the textbook.

The random variables y, x, z have a multivariate normal distribution with mean vector $\mu' = [1, 2, 4]$ and covariance matrix

$$\Sigma = \begin{bmatrix} 2 & 3 & 1 \\ 3 & 5 & 2 \\ 1 & 2 & 6 \end{bmatrix}$$

1. Compute the slope and intercept in the conditional mean function $E[y|x]$. Compute the two slopes and the intercept in the conditional mean function $E[y|x, z]$. *Hint: you may assume both functions are linear.* Is the slope on x the same in the two functions? Explain.

2. Compute the conditional variance $Var[y|x]$. Compute the conditional variance $Var[y|x, z]$.

3. Compute the squared correlation between y and x. Compute the squared correlation between y and $E[y|x]$. (Hint: You found in part 1 that $E[y|x] = \alpha + \beta x$.)

4. Compute the squared correlation between y and $E[y|x, z]$. (Hint: You found in part 1, $E[y|x, z] = \alpha + \beta x + \gamma z$.)

Problem 2 Regression

(You will need software to do this exercise. You may use any computer program that you wish. The computations are straightforward.)

The data file (which you should download)

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

is an Excel (portable) CSV file that contains data on fuel bills and number of rooms for 144 homes.

1. 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?

2. 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?

Problem 3* (optional) More Distribution Theory

Consider the joint distribution of two random variables, y, which is the number of failures of some component (disk drive) in a brand of computer per unit of time and x, the average lifetime of some different but related component (a chip). Note that y is a discrete random variable and x is a continuous random variable.

Suppose that the conditional density of y given x is

$$f(y|x) = \frac{e^{-\beta x}(\beta x)^y}{y!}, \quad y = 0, 1, ..., \ x \geq 0, \ \beta > 0,$$

while the marginal density of x is

$$f(x) = \theta e^{-\theta x}, \quad x \geq 0, \ \theta > 0.$$

Thus, conditioned on x, y has a Poisson distribution with parameter βx, while x, unconditionally, has an exponential distribution.

1. What is the joint density of these two random variables, $f(y, x)$?

2. Show that the unconditional density of y is $f(y) = \delta(1 - \delta)^y$ where $\delta = \frac{\theta}{(\beta+\theta)}$. (Hint: Integrate x out of $f(y, x)$)

3. Show that $E[x] = \frac{1}{\theta}$ and $Var[x] = \frac{1}{\theta^2}$.

   For a discrete random variable z that has a Poisson distribution with parameter α

$$f(z) = \frac{e^{-\alpha}\alpha^z}{z!}, \quad E[z] = Var[z] = \alpha.$$

   It follows then, that in our conditional distribution,

$$E[y|x] = Var[y|x] = \beta x.$$

   Note that this "regression" model has a linear conditional mean function. You could obtain $E[y]$, $Var[y]$, and $Cov[y,x]$ from the marginal distribution $f(y)$ and the joint distribution $f(x,y)$ by summing and integrating using the definitions. But, there is a much easier way.

4. Using the fundamental results:

$$\begin{aligned} E[y] &= E_x[E[y|x]] \\ Var[y] &= E_x[Var[y|x]] + Var_x[E[y|x]] \\ Cov[x,y] &= Cov[x, E[y|x]], \end{aligned}$$

   show that

$$\begin{aligned} E[y] &= \frac{\beta}{\theta} = \gamma \\ Var[y] &= \frac{\beta}{\theta} + \left(\frac{\beta}{\theta}\right)^2 = \gamma(1+\gamma) \\ Cov\,[x,y] &= \frac{\beta}{\theta^2} = \frac{\gamma}{\theta}. \end{aligned}$$

Problem 4 Least Squares Algebra

1. In the December, 1969, American Economic Review (pp. 886-896), Nathaniel Leff reports the following least squares regression results for a cross section study of the effect of age composition on savings in 74 countries in 1964:

$$\log S/Y = 7.3439 + 0.1596 \log Y/N + 0.0254 \log G - 1.3520 \log D_1 - 0.3990 \log D_2 \quad (R^2 = 0.57)$$
$$\log S/N = 8.7851 + 1.1486 \log Y/N + 0.0265 \log G - 1.3438 \log D_1 - 0.3966 \log D_2 \quad (R^2 = 0.96)$$

   where S/Y = domestic savings ratio, S/N = per capita savings, Y/N = per capita income, D_1 = percentage of the population under 15, D_2 = percentage of the population over 64, and G = growth rate of per capita income.

Are these results correct? Explain.

Arthur Goldberger raised this question in a comment on Leff's paper in a comment in the 1973 American Economic Review. The (2 page) paper can be downloaded from https://www.jstor.org/stable/1803140. Leff's (1 page) reply is at https://www.jstor.org/stable/1803141. Read these two papers. What's your opinion? Specifically, what about Leff's reaction to the comment?

2. *** (Optional) Regression without a constant term.** What is the effect on R^2 of computing a linear regression without a constant term? (Note, it depends on how R^2 is computed.)

3. *** (Optional) Partitioned regression.** Suppose a data set consists of n observations on y, K_1 variables in X_1 and K_2 variables in X_2. Do the following four procedures produce the same value for the least squares coefficients on X_2?

   (a) Regress y on both X_1 and X_2.

   (b) Regress the residuals from a regression of y on X_1 on the residuals (column by column) of regressions of X_2 on X_1.

   (c) Same as (b), but do not transform y.

   (d) Same as (b), but do not transform X_2.

4. **Residual makers.** Define

$$M \equiv I - X\left(X'X\right)X',$$

and

$$M_1 \equiv \left(I - X_1\left(X_1'X_1\right)X_1'\right),$$

where the columns of X_1 are a subset of the columns X.
This follows the textbook's equation (3-14) for M and (3-19) for M_1. What is the result of the matrix product M_1M?

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

$$(y–Xc)'(y–Xc)–(y–Xb)'(y–Xb) = (c–b)'X'X(c–b).$$

A property of the matrix $X'X$ is that is *positive definite.* This means that for any vector $u \neq 0$, $uX'Xu > 0$. 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 X matrix which has 5 columns (variables) $X = [\text{a constant term, income}, P_D, P_N, P_S]$ where P_m is a price index for the m^{th} expenditure group. Denote the the m^{th} least squares coefficient vector by $b_m m = D, N, S$.

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

   $$b_D + b_N + b_S = [1, 0, 0, 0, 0]'.$$

   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.

7. *** (Optional) Multicollinearity**. The regression model of interest is

   $$y = X_1\beta_1 + X_2\beta_2 + \epsilon,$$

   where X_1 is K_1 variables, including a constant and X_2 is K_2 variables not including a constant. It is believed that multicollinearity between the columns of X_1 and X_2 is adversely affecting the regression.

   Consider the following 'cure.' We will first regress each variable in X_2 on all of the variables in X_1. By construction, the residuals in these regressions, call them $Z_2 = (z_1, ..., z_{K_2})$, are orthogonal to every variable in X_1. So, instead of regressing y on X_1 and X_2, we linearly regress y on X_1 and Z_2.

   Denote by $b = (b_1, b_2)$ the least squares coefficients in the original regression, and by $c = (c_1, c_2)$ the least squares coefficients in the regression of y on X_1 and Z_2.

   Show the algebraic relation between b and c. Is c unbiased?

   Use the gasoline data

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

   Let y be the variable $GASEXP$ in the data set, let X_2 denote the three macroeconomic price indices, PD, PN, and PS, and let X_1 denote the other independent variables, constant, $GASP$, and $PCINCOME$. Carry out the computations listed above and verify that the algebraic results you obtained do appear in the empirical results.

Problem 5* (optional) Textbook Questions

Please do problems 1, 6, and 10 from Greene, Chapter 3.