

# PLSC 503: Problem Set 2

Thad Dunning

Department of Political Science

Yale University

Spring 2010

Due January 26, 2010

# 1 Theoretical/Conceptual Problems

Before completing this problem set, you should work all of Exercise Set A as well as problems 1-11 in Exercise Set B in Freedman (2009, Chapter 3). For questions 1-7 below, let  $X$  be an  $n \times 2$  matrix, where the first column is all 1's and the second column is  $(x_1, x_2, \dots, x_{n-1}, x_n)'$ . Let  $Y$  be an  $n \times 1$  column vector consisting of  $(y_1, y_2, \dots, y_{n-1}, y_n)'$ . Remember to show your work in problems 2-9!

1. What is the size of  $X'X$ ? Of  $X'Y$ ? What about  $(X'X)^{-1}$  and  $(X'X)^{-1}(X'Y)$ ? Can you multiply  $X$  and  $Y$ ? Why or why not?
2. Find  $X'X$ . (That is, write out  $X'X$ , with typical elements given by  $n$ ,  $\sum_{i=1}^n (x_i)^2$ , and so on).
3. Find  $(X'X)^{-1}$ .
4. Find  $(X'Y)$ .
5. Find  $(X'X)^{-1}X'Y$ .
6. Show that the (2, 1) element of  $(X'X)^{-1}X'Y = \frac{\text{Cov}(x,y)}{\text{Var}(x)}$ . That is, when there is a constant and one variable in an  $n \times 2$  design matrix, the matrix representation reduces to the usual formula for the slope coefficient of the bivariate regression line.

In working this problem, it will be helpful to remember the alternate definitions of covariance and variance:

$$\begin{aligned}\text{Cov}(x, y) &= \frac{1}{n} \sum_{i=1}^n x_i y_i - \left( \frac{1}{n} \sum_{i=1}^n x_i \right) \left( \frac{1}{n} \sum_{i=1}^n y_i \right) \\ &= \overline{xy} - (\bar{x})(\bar{y})\end{aligned}\tag{1}$$

and

$$\begin{aligned}\text{Var}(x) &= \frac{1}{n} \sum_{i=1}^n (x_i)^2 - \left( \frac{1}{n} \sum_{i=1}^n x_i \right) \left( \frac{1}{n} \sum_{i=1}^n x_i \right) \\ &= \overline{x^2} - (\bar{x})^2.\end{aligned}\tag{2}$$

7. Show that the (1, 1) element of  $(X'X)^{-1}X'Y = \bar{y} - b\bar{x}$ , where  $b = \frac{\text{Cov}(x,y)}{\text{Var}(x)}$ .
8. Work exercise 3.B.12 in Freedman (2009), that is, exercise B.12 in Chapter 3.
9. Work exercise 3.B.13 in Freedman (2009).
10. Work exercise 3.B.14 in Freedman (2009).
11. Work exercise 3.B.17 in Freedman (2009).

## 2 Computer lab: Yule's regression

### 2.1 Bivariate regression

Read Section 1.4 in Freedman (2009). Then open the dataset Yule.dta or Yule.csv, which are available under the "Resources" tab on the classes v2 server (<https://classesv2.yale.edu/portal/>). (These file types seem excluded by my website backend and thus are not posted to the usual course website). Yule.dta is a Stata file, while Yule.csv is a comma-separated values file. These are the data for 32 metropolitan unions from 1871-81, given in Table 1.3 in Freedman (2009). You will need to subtract 100 from each entry to get the percentage change. Here,  $\Delta\text{Paup}$  is the percentage change in the number of paupers;  $\Delta\text{Out}$  is the percentage change in the ratio of paupers outside the poor house to those inside,  $\Delta\text{Old}$  is the percentage change in the population over 65, and  $\Delta\text{Pop}$  is the percentage change in the population. Refer to Freedman (2009, Chapter 1) for more information.

1. Compute the means and SDs of  $\Delta\text{Paup}$ ,  $\Delta\text{Out}$ ,  $\Delta\text{Pop}$ , and  $\Delta\text{Old}$ .
2. Compute all 6 correlations between  $\Delta\text{Paup}$ ,  $\Delta\text{Out}$ ,  $\Delta\text{Pop}$ , and  $\Delta\text{Old}$ .
3. Make a scatter plot of  $\Delta\text{Paup}$  against  $\Delta\text{Out}$ .
4. Use the means, SDs, and the correlation between  $\Delta\text{Paup}$  and  $\Delta\text{Out}$  to run a regression of  $\Delta\text{Paup}$  on  $\Delta\text{Out}$ , i.e, find the slope and intercept of the regression line. (Do not use Stata's "regress" command, though you may use other commands to manipulate means, SDs, and the like; show your work). Also compute the SD of the residuals.
5. Now use Stata's "regress" command (or an equivalent command in another statistical package) to run a regression of  $\Delta\text{Paup}$  on  $\Delta\text{Out}$ . (Make sure your statistical software includes an intercept, which is the default option in Stata).

Turn in your output (including standard errors and/or t-statistics if you wish, though these can be ignored for now.) How does the slope coefficient in your regression output compare to the slope of the regression line you found in 4? Comment.

## 2.2 Multiple regression

6. Use Stata's matrix commands to generate a  $32 \times 1$  vector  $Y$  whose typical element is  $\Delta\text{Paup}_i$  and a  $32 \times 4$  design matrix  $X$  whose typical row is given by  $[1 \ \Delta\text{Out}_i \ \Delta\text{Old}_i \ \Delta\text{Pop}_i]$ .

Calculate  $(X'X)^{-1}X'Y$ . Compare the elements of this vector to the fitted coefficients of Yule's regression (see Section 1.4 of Freedman 2009). Your answers may differ from Yule's, as reported in Freedman (2009, Section 1.4). Why might this be?

7. Use the computer to form a new matrix, deleting the second column of  $X$ , that is, the column  $\Delta\text{Out}$ . Call this  $32 \times 3$  new matrix  $\tilde{X}$ ; its typical row is given by  $[1 \ \Delta\text{Old}_i \ \Delta\text{Pop}_i]$ . Now,

regress  $Y$  on  $\tilde{X}$ , using the standard regression commands. Store the residuals from this regression in a vector called  $\tilde{f}$ .

Then, regress  $\Delta\text{Out}$  (that is, the second column of  $X$ ) on  $\tilde{X}$ . Store the residuals from this regression in another vector called  $\tilde{g}$ .

Finally, regress  $\tilde{f}$  on  $\tilde{g}$ . How does your answer compare to the second element of  $(X'X)^{-1}X'Y$ ? Using the computer, check also that the residuals from this final regression are orthogonal to  $X$  (that is, the original design matrix.) Comment on your findings.

8. How do your answers in 6 and 7 compare to the regression slope coefficients you found in 4 and 5 above? If these answers are all the same, say why they are the same; if they are different, say why they are different.
9. Calculate the Residual Sum of Squares and the Total Sum of Squares for Yule's multiple regression (that is, the regression in 6). Use this to calculate the  $R^2$  of the regression.