Problem Set 5

This problem set is due at 8:30 am on 10/17

Please upload both Rmd and PDF files on Sakai

Do not show the code in the pdf, show outputs and write-up only

Total points: 20

Matrix Basics (5 points, 1 for each sub-question)

$$S = \begin{bmatrix} 1 & 800 & 3.8 & 1 \\ 1 & 450 & 2.7 & 0 \\ 1 & 730 & 4.0 & 1 \\ 1 & 650 & 2.5 & 0 \end{bmatrix}; g = \begin{bmatrix} 3.75 \\ 2.5 \\ 3.9 \\ 3.5 \end{bmatrix}; F = \begin{bmatrix} 1 & 2 & 19 \\ 0 & 3 & 21 \\ 0 & 3 & 21 \end{bmatrix}; W = \begin{bmatrix} 1 & 80 & 1 & 20 \\ 1 & 600 & 3 & 150 \\ 1 & 20 & 2 & 5 \\ 1 & 128 & 2 & 32 \end{bmatrix}$$

Input these four matrices into R, complete the following calculations:

1. Add the two matrices that are technically possible to add.
2. Why can we not add other matrices?
3. Multiply matrices S and g.
4. Why can we multiply these two matrices and not S and F?
5. Take the transpose of S

Matrix Algebra Regression 1 (1 point)

Suppose S is a matrix of explanatory variables, where column 1 is a constant, column, 2 is SAT, column 3 is H.S. GPA, and column 4 is whether the student is male or female. Now, suppose that g is a vector of college GPA scores for four different people. Using matrix commands, calculate (1) the intercept, (2) the slope, and (3) the SSR of the predicted line resulting from a regression of g on SAT (g is the dependent variable, column 3 is the independent variable).

1. Now, convert the matrices in question 3 to variables and confirm your matrix calculations using the regress command.

Matrix Algebra Regression Part 2 (1 point)

Repeat the matrix and regression operations from part 1, except this time use the entire matrix S and not just the first two columns.

1. What are the intercept, slope, and SSR of this new model?

Matrix Algebra Regression Part 3 (2 points)

Suppose W is a matrix of demographic explanatory variables, where column 1 is a constant, column 2 is the annual income of the parents in units of $10,000, column 3 is variable coding the ethnicity of the student, and column 4 is the number of non-fiction books owned by the student's parents (units are 100s of books).

1. Try to calculate the impact of the parents' income and number of books on the student's college GPA using matrix commands. (Hints: The calculation may not work. Simply tell us what you tried and why it didn't work in the write-up).

2.Why can the intercept and slope estimates not be calculated?

Matrix Calculations on an actual data set. (3 points)

Use `VOTE1.dta` for the following analysis:

1. Create a y vector that includes the variable VoteA. Then create a constant term and store it, `expendA`, and `expendB` in an `X` matrix.
2. Now, use R matrix commands to calculate the slope estimates, intercept, and SSR of regression of `VoteA` on `expendA` and `expendB`.
3. Use R's regression commands to verify your result.

Omitted Variable Bias (5 points)

Use `BWGHT.DTA` for the following analysis:

A problem of interest to health officials (and others) is to determine the effects of smoking during pregnancy on infant health. One measure of infant health is birth weight; a birth weight that is too low can put an infant at risk for contracting various illnesses.

Use `data('bwght')` in the wooldridge package. Variable defined as follows:

- faminc: 1988 family income, $1000s
- cigtax: cig. tax in home state, 1988
- cigprice: cig. price in home state, 1988
- bwght: birth weight, ounces
- fatheduc: father's yrs of educ
- motheduc: mother's yrs of educ
- parity: birth order of child
- male: =1 if male child
- white: =1 if white
- cigs: cigs smked per day while preg
- lbwght: log of bwght
- bwghtlbs: birth weight, pounds
- packs: packs smked per day while preg
- lfaminc: log(faminc)

$$bwght = \beta_0 + \beta_1 cigs + u.$$

1. Estimate this model by OLS and report the results. Do you think the model has omitted variable bias? If so, can you list three potential omitted variables?

Then, we add $faminc$ into the model.

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 faminc + u.$$

2. Estimate this model by OLS and report the results.
3. Compare the estimate with previous bivariate regression. Is the effect of smoking larger or smaller when you control for income?
4. Estimate the following model:

$$log(bwght) = \beta_0 + \beta_1 cigs + \beta_2 log(faminc) + u.$$

If $faminc$ increases by 0.10 (10 percentage points), what is the estimated percentage change in `bwght`?

5. Find the correlation between `log(faminc)` and `cigs`. Is it roughly what you expected? Can you sign the bias of the bivariate regression $log(bwght) = \beta_{cigs}$? Discuss how the bias affects our interpretation.

Multicollinearity (3 points)

Again, we focus on this model:

$$log(bwght) = \beta_0 + \beta_1 cigs + \beta_2 log(faminc) + u.$$

1. Manually compute the Variance Inflation Factors.
2. Using the built-in function to verify.
3. Do you think this model has a multicollinearity problem? If so, would it affect our interpretation?