Jeff Grogger
Harris School

Winter 2022
University of Chicago

PP 420: Problem Set 1

(due Tuesday, January 25, by 11:59pm via Canvas)

A. Analytical problems

1. Prove (P1)-(P5) from the lecture notes. Establish whether each property continues to hold if the model does not include a constant term.

2. One can establish that $0 \leq R^2 \leq 1$ when the model includes an intercept, where R^2 is defined in the class notes. Show why this result may fail if the model does not include a constant.

3. Consider the model:

$$y = X\beta + z\gamma + \varepsilon,$$

where z is an nx1 vector and γ is a scalar. Let $M=I-X(X'X)^{-1}X'$. Now consider the regressions:

(a) $My = z\gamma + \varepsilon_1$
(b) $y = z\gamma + \varepsilon_2$
(c) $y = Mz\gamma + \varepsilon_3$
(d) $My = Mz\gamma + \varepsilon_4$

Which of the regressions could be used to provide unbiased estimates of γ? Summarize in words the key lesson here regarding partial regression.

4. Consider the estimator

$\tilde{\beta} = [X'X + \lambda I]^{-1}X'y$, where $\lambda > 0$ is a known number and I is the identity matrix of order K.

(a) Is $\tilde{\beta}$ a linear estimator?
(b) Show that $\tilde{\beta}$ is biased. Provide an expression for the bias.
(c) Derive $V(\tilde{\beta}|X)$, the conditional covariance matrix of $\tilde{\beta}$.
(d) Show that $V(b|X) \geq V(\tilde{\beta}|X)$. *Hint*: Write $V(\tilde{\beta}|X)$ in terms of T, where $T = X'X[X'X + \lambda I]^{-1}$, and make use of the fact that $TT^{-1} = I$.
(e) Typically, the investigator chooses the value of λ. What happens to the bias of $\tilde{\beta}$ as λ grows? What happens to the difference between $V(b|X)$ and $V(\tilde{\beta}|X)$? Explain briefly the tradeoff that the choice of λ entails.

B. Computational Problems

For these problems (and many to come), you will want to download the Stata data set named pp420_data.dta from the course Canvas site.

1. Find the means of the following variables: earnings, welfare income, total income, education, age, and age squared. Comment on the sample sizes corresponding to the different variables.

2. Now separately regress earnings, welfare income, and total income on a constant. Compare the regression results to the means from above. Explain their relationship.

3. Now regress earnings on education, age, and age squared.

(a) Explain the sample size used to estimate the regression.
(b) Interpret the education coefficient.
(c) What are the TSS, RSS, and ESS?
(d) Use them to compute the R-square and compare it to the value reported in the regression output.

4. Retrieve the residuals and predicted values from the above regression, and use them to verify the following properties:

(a) the residuals sum to zero
(b) the mean of the predicted values equals the mean of the dependent variable.
(c) the residuals are orthogonal to the regressors and the predicted values.
(d) the square of the correlation coefficient between the dependent variable and the predicted values equals the R-square from the regression.

5. Re-estimate the regression from problem 3, this time omitting the constant term. Which of properties 4(a)-(d) still hold, if any? Explain.