Problem Set 2

February 16, 2020

General Instructions

Complete the following assignments in a sweave file that shows your code, output, and discussion. Hand in the compiled (pdf) version. You can use this file to get you started. You can work with others, but please hand in your own version. Please report any glitches as soon as you discover them - thanks!

For the theory questions, you may want to work out your answer on paper first, then type up the key steps in LaTeX. You can insert your LaTeXcode directly in in your sweave file, or - if you prefer, in a separate file or files in TeXnicCenter. For full credit, typed answers are required.

Question 1: Orthogonality

Consider a CLRM with sample size n and k regressors. Show that if all columns in the data matrix $\mathbf{X}$ are orthogonal to one another, the resulting solution to the Least Square problem is equivalent to running k separate regressions, one per explanatory variable.

(Hint: Partition $\mathbf{X}$ *into columns and express the solution formula for* $\mathbf{b}$ *in terms of column interactions. Then examine what this solution looks like under the stated orthogonality assumption. Note: This is NOT asking you to do a partitioned regression. Just a basic regression, with* $\mathbf{X}$ *expressed in terms of its columns.)*

To get you started:

$$\begin{aligned}
&\mathbf{X} = \begin{bmatrix} \mathbf{c}_1 & \mathbf{c}_2 & \dots & \mathbf{c}_k \end{bmatrix} \\
&\mathbf{X}' = \begin{bmatrix} \mathbf{c}_1' \\ \mathbf{c}_2' \\ \vdots \\ \mathbf{c}_k' \end{bmatrix} \\
&\mathbf{X}'\mathbf{X} = \text{your turn...} \\
&\left(\mathbf{X}'\mathbf{X}\right)^{-1} = \text{your turn...} \\
&\text{etc.}
\end{aligned} \tag{1}$$

Question 2: Regression on a constant

Consider a classical linear regression model that regresses the dependent variable, $\mathbf{y}$, against a column of ones (call it $\mathbf{i}$), and no other explanatory variables. Assume the underlying theoretical population model has the usual OLS properties.

(a) Write down the underlying theoretical model for a single observation, and for the full sample of n observations. Clearly specify the vector dimensions of each element. Call the unknown population parameter μ).

(b) Derive the OLS estimator for this model (Call it b). (Hint: Start with the known formula of the OLS estimator and replace the $\mathbf{X}$ with the regressor for the current model. Then simplify until you're left with a scalar that has a very well known form).

(c) Derive the variance of this estimator. Assume that the error variance σ is known. (Hint: Proceed as above - start with the known form of V(b), then replace the $\mathbf{X}$ with the regressor for the current model).

(d) Show that the OLS estimator for your model is unbiased for the underlying population parameter.

(e) In light of your results, can you make a general statement regarding the best linear unbiased estimator for the population mean when no explanatory variables are available?

Question 3: Small Sample Properties of a Linear regression Model with non-zero-mean-error

Consider the TRUE linear regression model $\mathbf{y} = \mathbf{i}\beta_1 + \mathbf{X}\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$ where $\mathbf{i}$ is an n by 1 column of 1's, and $\mathbf{X}$ is an n by (k-1) matrix of regressors. Assume that the n elements of the random error vector $\boldsymbol{\epsilon}$ are distributed i.i.d normal with variance σ^2, but with mean $\mu \neq 0$. As usual, assume that the elements in $\boldsymbol{\epsilon}$ are uncorrelated with the elements in $\mathbf{X}$.

1. Which standard assumption underlying the classical linear regression model (CLRM) is violated by this specification?

2. Using Partitioned Regression results, show that the estimation of this model through simple OLS will produce unbiased slope coefficients, but a biased estimate for the regression intercept. (Denote the estimated intercept as b_1, and the vector of estimated slope coefficients as $\mathbf{b}_2$).

3. Now assume the true model is as before, but the analyst chooses to estimate it without an intercept. Continuing to assume the above given properties of the error term, show that in this case the OLS estimator $\mathbf{b}_2$ will be biased.

4. Would the bias vanish if the error term was well-behaved (i.e. had a $\mu = 0$)? Why or why not?

Question 4: Variance of OLS estimator

Consider script `mod1s3`, which deals with the small sample properties of the OLS estimator.
In this exercise you will show that the empirical standard deviations ("standard errors") of the OLS estimator depend on the variability in $\mathbf{X}$. Specifically, perform the following simulation tasks (make sure all your tables and figures are labeled clearly and correctly):

1. For all of the following steps, set $n = 1000, r1 = 500, \beta = \begin{bmatrix} 1 & 0.5 & 1.2 \end{bmatrix}$, and $\sigma^2 = 1$. Make sure to set your random number seed to (37) so your results match mine exactly. Define $x1$ as in the original script.

2. Draw $x2$ from a normal density with mean=3 and std=1, and $x3$ from a normal with mean=-2 and std=1. Simulate the sampling distribution of the OLS estimator, analogous to script `mod1s3` (but with only a single setting for σ^2). Summarize your results in a table (as we did in the original script). Call it *Table 1*. Collect your draws of the estimator in matrix *bmat1*.

3. Now boost both standard deviations of $x2$ and $x3$ to 3, and re-draw both variables for the same sample size as above. Simulate the conditional sampling distribution of the OLS estimator and capture your results in a new output table (call it *Table 2*). Collect your draws of the estimator in matrix *bmat2*.

4. Repeat, with settings for both standard deviations of 4. Call the output table *Table 3*. Collect your draws of the estimator in matrix *bmat3*.

5. Repeat, with settings for both standard deviations of 6. Call the output table *Table 4*. Collect your draws of the estimator in matrix *bmat4*.

6. Capture all 4 sampling distributions for the second element of the OLS solution (the coefficient on x2) in a single figure. Set the properties of the x and y axes such that all four plots fit comfortably into the figure window. (Keep in mind that now your first plot will have the widest spread!)

7. Repeat this figure for the third element of the OLS solution (the coefficient on x3).

8. Comment on your results.