POLS W4912 Spring 2008

Homework Assignment #3

**Show all work. For those questions that require you to write `R` code, hand in the programs that you wrote. This assignment is due on March 6.**

1. Show that the residuals are unbiased when you incorrectly include an irrelevant variable in a regression. Assume the population regression function is $\mathbf{y} = \mathbf{x}_1\beta_1 + \mathbf{x}_2\beta_2 + \boldsymbol{\varepsilon}$ and recognize that this is the true model so long as $\beta_2 = 0$ (i.e., $\mathbf{x}_2$ is irrelevant. Hint: while there are many ways to conduct the proof, you may just want to use the residual maker matrix $\mathbf{M}$ to derive the residuals using the correct and incorrect models and then take the expectation in each case. What does this imply for the biasedness or unbiasedness of the estimate of $\sigma^2$? (25 points)

2. Suppose a sample of adults is classified into groups 1, 2, or 3 depending on whether their education stopped during or at the end of elementary school, high school or college respectively. The relationship $y_i = \beta_0 + \beta_1 D_{i,2} + \beta_2 D_{i,3} + \varepsilon_i$ is specified where $y_i$ is income and $D_{i,k} = 1$ for all individuals in group $k$ and is zero otherwise (e.g., $D_{i,3} = 1$ if $i$ completed college and is zero otherwise).

   (a) In terms of the parameters of the model, what is the expected income of those whose education stopped in college?

   (b) In terms of the parameters of the model, what is the null hypothesis that going to college after high school makes no contribution to adult income?

   (c) Can the specified model be expressed in simpler, equivalent form $y_i = \alpha_0 + \alpha_1 x_i + \varepsilon_i$, where $x_i$ is years of education? (25 points)

3. You are reviewing a paper for a refereed journal in which the authors claim that high collinearity among explanatory variables leads to biased point estimates, since high collinearity increases the variance of the estimates. Do you agree or disagree with this argument? The file `multicollinearity.r` (available from the course web site) contains `R` code that performs a Monte Carlo analysis to check this argument. Run this program choosing three different experimental conditions to provide support for your position. Explain why your results make theoretical sense. (25 points)

4. From p. 74 of the notes, the equation

$$\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2$$

   can be used to indicate the bias in the OLS estimate of $\boldsymbol{\beta}_1$ when a relevant variable is omitted (think about $E[\hat{\boldsymbol{\beta}}_1]$). Perform a Monte Carlo analysis to verify that this equation is accurate. Set up the analysis so that $\mathbf{X}_1$ contains only one explanatory variable and $\mathbf{X}_2$ contains two explanatory variables, both of which are correlated with $\mathbf{X}_1$. I suggest modifying `multicollinearity.r` for this problem, rather than starting from scratch. (25 points)