Problem Set 7
17.874
Spring, 2004

1. A colleague has written a paper using polling data that were using representative sampling rather than random sampling. In representative sampling, a set of characteristics are chosen, such as gender, income, and education, and then interviewers select people matching those characteristics. In the study in question, income was not measured, but the sampling characteristic selected on education, which is positively correlated with income. Concerned about bias, your colleague decides to include the matching variable (call it Q) as a control variable in the regression to capture the effects of income. Your colleague wishes the measure the effect of X on Y, controlling for other factors (including income). He states in the paper that using Q as a control variable in lieu of income yields unbiased estimates of the effect of X on Y. Q is called a *proxy* variable.
   a. Is this argument right? Do proxy variables eliminate bias? Show your answer by deriving the formula for the unbiasedness of the regression of Y on X and Q (i.e., E[b]), where the true regression is of Y on X and W (income).
   b. Compare the formula for unbiasedness of the regression above for the formula when only X is included in the regression.

2. Greene, Chapter 21, problem 1.