Problem Set II

In the first two problems we use the same data as in problem set I. We will explore omitted variable bias and proxy variables as a way to mitigate that bias.

Problem 1 Omitted variable bias

1. Estimate a linear regression model that relates the log weekly wage to years of education. Report the OLS estimates and standard errors. What is the interpretation of the coefficient on education?

2. Ability is an important determinant of earnings and is likely to be correlated with years of education. What is the likely sign of the resulting omitted variable bias? The NLS data have a measure of ability, IQ. Add IQ to the regression and compare the coefficient of years of education before and after the addition. Regress IQ on years of education (remember that you need to have a good reason to omit the intercept) and show that the observed change in the education coefficient is as predicted by the omitted variable bias formula.

Problem 2 Proxy variables. Let y be the dependent variable, x and z be independent variables, of which only x is observed, e.g. x is years of education and z is ability. Although we do not observe z we have a proxy variable w for z.

1. We make the following assumptions on the proxy variable: (i) in the linear regression of y on x, z and w, the coefficient of w is 0, (ii) in a linear regression of z on x and w the coefficient of x is 0. Note that these regressions cannot be estimated because z is not observed. It is assumed that the coefficients in the linear regression of y on x, z and w (in which w can be omitted) are the true partial effects, and in particular the coefficient on years of education is the return to education that is free of ability bias. Now consider the linear regression of y on x and w in which z is omitted. Use the omitted variable bias formula of lecture 6 to show that OLS will consistently estimate the partial effect of x but that the coefficient of w is asymptotically biased. A variable w with properties (i) and (ii) is called a perfect proxy.

2. Now assume that w is an imperfect proxy. In particular we assume that although (i) is true, (ii) need not hold. In that case we have the choice to use the proxy w for the unobserved z or we can only use x as independent variable. Which is best depends on the size of the asymptotic bias in the two cases. The regression coefficient that we are interested in is β_x in the linear regression of y on x (that has coefficient β_x) and z (that has coefficient β_z). Express the bias due to omitting z in terms of β_z and the coefficient in the regression of z on x (call this κ_x).

3. Now consider the regression of y on x and w with coefficients γ_x and γ_w. In this regression z is omitted. Express the omitted variable bias in the OLS estimate of the

coefficient of x in terms of β_z and the coefficient of x in a regression of z on x and w (call this coefficient λ_x).

4. Compare the two bias formulas and find the condition under which the bias with a proxy is smaller than without a proxy in the regression.

5. In the NLS let IQ be a perfect measure (not proxy) for ability. If IQ were not observed we have a proxy variable KWW that is the score on a much simpler test. Should we use the proxy KWW for IQ? Show that the condition obtained in part 4 is satisfied.

Problem 3 Multicollinearity. A researcher estimates a regression with an intercept of (log) earnings on age, years of education and work-experience.

1. Since the data does not have work-experience the researcher proposes to use potential work-experience defined as age-years of schooling minus 6. STATA gives an error message when calculating the OLS estimates of the regression coefficients. Why? What went wrong?

2. A colleague suggests that most economists estimate a an (log) earnings regression with an intercept, age and years of education, also work-experience and its square. If only the squared work-experience is included will STATA still give the error message? Why (not)?

3. What hidden assumption makes this possible, i.e. can you change the regression model so that the STATA error message returns?