ECON 6511: Advanced Applied Econometrics
Homework 6

1. (Wooldridge, Chapter 15, Problem 1) Use the data in WAGE2.dta for this exercise.

   (a) In class we considered using *sibs* as a instrument for *educ* (which would yield an IV estimate of the return to education of 0.122). To convince yourself that using *sibs* as an IV for *educ* is <u>not</u> the same as just plugging *sibs* in for *educ* and running an OLS regression, run the regression of $\log(wage)$ on *sibs* and explain your findings.
   Answer: The regression of $\log(wage)$ on *sibs* gives:

   $$\log(wage) = 6.861 - 0.0279 sibs$$

   This is a reduced form simple regression equation. It shows that, controlling for no other factors, one more sibling in the family is associated with monthly salary that is about 2.8% lower. The t statistic on sibs is about -4.73. Of course *sibs* can be correlated with many things that should have a bearing on wage including, as we already saw, years of education.

   (b) The variable *brthord* is birth order (*brthord* is one for a first-born child, two for a second-born child, and so on.) Explain why *educ* and *brthord* might be negatively correlated. Regress *educ* on *brthord* to determine whether there is a statistically significant negative correlation.
   Answer: It could be that older children are given priority for higher education, and families may hit budget constraints and may not be able to afford as much education for children born later. The simple regression of *educ* on *brthord* gives

   $$educ = 14.15 - 0.283 brthord$$

   The equation predicts that every one-unit increase in *brthord* reduces predicted education by about .28 years. In particular, the difference in predicted education for a first-born and fourth-born child is about .85 years.

   (c) Use *brthord* as an IV for *educ* in the equation:

   $$\log(wage) = \beta_0 + \beta_1 educ + u.$$

   Report and interpret the results.
   Answer: When brthord is used as an IV for educ in the simple wage equation we get

   $$\log(wage) = 5.03 + 0.131 educ + u.$$

(The R-squared is negative.) This is much higher than the OLS estimate (.060) and even above the estimate when sibs is used as an IV for *educ* (.122). Because of missing data on *brthord*, we are using fewer observations than in the previous analyses.

(d) Now, suppose that we include number of siblings as an explanatory variable in the wage equation; this controls for family background, to some extent:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 sibs + u.$$

Suppose that we want to use *brthord* as an IV for *educ*, assuming that *sibs* is exogenous. The reduced form for *educ* is:

$$educ = \pi_0 + \pi_1 sibs + \pi_2 brthord + v$$

State and test the identification assumption.

Answer: In the reduced form equation

$$educ = \pi_0 + \pi_1 sibs + \pi_2 brthord + v$$

we need $\pi_2 \neq 0$ in order for the β_j to be identified. We take the null to be $H_0 : \pi_2 = 0$, and look to reject H_0 at a small significance level. The regression of *educ* on *sibs* and *brthord* (using 852 observations) yields $\hat{\pi}_2 = -0.153$ and $\text{se}(\hat{\pi}_2) = .057$. The t statistic is about -2.68, which rejects H_0 fairly strongly. Therefore, the identification assumptions appears to hold.

(e) Estimate the equation from part (d) using *brthord* an an IV for *educ* (and *sibs* as its own IV). Comment on the standard errors for $\hat{\beta}_{educ}$ and $\hat{\beta}_{sibs}$.

Answer: The equation estimated by IV is

$$\log(wage) = 4.94 + 0.137 educ + 0.0021 sibs + u.$$

The standard error on $\hat{\beta}_{educ}$ is much larger than we obtained in part (c). The 95% CI for β_{educ} is roughly $-.010$ to $.284$, which is very wide and includes the value zero. The standard error of $\hat{\beta}_{sibs}$ is very large relative to the coefficient estimate, rendering *sibs* very insignificant.

(f) Using the fitted values from part (d), $\widehat{educ}$, compute the correlation between $\widehat{educ}$ and *sibs*. Use this result to explain your findings from part (e).

Answer: Letting $\widehat{educ}$ be the first-stage fitted values, the correlation between $\widehat{educ}$ and $sibs_i$ is about $-.930$, which is a very strong negative correlation. This means that, for the purposes of using IV, multicollinearity is a serious problem here, and is not allowing us to estimate β_{educ} with much precision.

2. (Wooldridge, Chapter 15, Problem 2) The data in FERTIL2.dta include, for women in Botswana during 1988, information on number of children, years of education, age, and religious and economic status variables.

   (a) Estimate the model

   $$children = \beta_0 + \beta_1 educ + \beta_2 age + \beta_3 age^2 + u$$

   by OLS, and interpre the estimates. In particular, holding *age* fixed, what is the estimated effect of another year of education on fertility? If 100 women receive another year of education, how many fewer children are they expected to have?
   Answer: The equation estimated by OLS is:

   $$children = -4.138 - 0.0906 educ + 0.332 age - 0.00263 age^2 + u$$

   Another year of education, holding *age* fixed, results in about .091 fewer children. In other words, for a group of 100 women, if each gets another of education, they collectively are predicted to have about nine fewer children.

   (b) The variable $frsthalf$ is a dummy variable equal to one if the woman was born during the first six months of the year. Assuming that $frsthalf$ is uncorrelated with the error term from part (a), show that $frsthalf$ is a reasonable IV candidate for *educ*. (*Hint*: You need to do a regression.)
   Answer: The reduced form for *educ* is:

   $$educ = \pi_0 + \pi_1 age + \pi_2 age^2 + \pi_3 frsthalf + v$$

   and we need $\pi_3 \neq 0$. When we run the regression we obtain $\hat{\pi}_3 = -.852$ and se$(\hat{\pi}_3) =$ .113. Therefore, women born in the first half of the year are predicted to have almost one year less education, holding *age* fixed. The t statistic on $frsthalf$ is over 7.5 in absolute value, and so the identification condition holds.

   (c) Estimate the model from part (a) using $frsthalf$ as an IV for *educ*. Compare the estimated effect of education with the OLS estimate from part (a).
   Answer: The structural equation estimated by IV is:

   $$children = -3.388 - 0.1715 educ + 0.324 age - 0.00267 age^2$$

   The estimated effect of education on fertility is now much larger. Naturally, the standard error for the IV estimate is also bigger, about nine times bigger. This produces a fairly wide 95% CI for β_1.

3. (Wooldridge, Chapter 15, Problem 6) Use the data in MURDER.dta for this exercise. The variable *mrdrte* is the muder rate, that is, the number of murders per 100,000 people. The variable *exec* is the total number of prisoners executed for the current and prior two years; *unem* is the state unemployment rate.

   (a) How many states executed at least one prisoner in 1991, 1992, or 1993? Which state had the most executions?

   Answer: Sixteen states executed at least one prisoner in 1991, 1992, or 1993. (That is, for 1993, exec is greater than zero for 16 observations.) Texas had by far the most executions with 34.

   (b) Using the two years 1990 and 1993, do a pooled regression of *mrdrte* on *d*93, *exec*, and *unem*. What do you make of the coefficient on *exec*?

   Answer: The results of the pooled OLS regression are

$$mrdrte = -5.28 - 2.07d93 + 0.128exec + 2.53unem$$

   The positive coefficient on *exec* is no evidence of a deterrent effect. Statistically, the coefficient is not different from zero. The coefficient on *unem* implies that higher unemployment rates are associated with higher murder rates.

   (c) Using the changes from 1990 to 1993 only (for a total of 51 observations), estimate the equation

$$\Delta mrdrte = \beta_0 + \beta_1 \Delta exec + \beta_2 \Delta unem + \Delta u$$

   by OLS and report the results in the usual form. Now, does capital punishment appear to have a deterrent effect?

   Answer: When we difference (and use only the changes from 1990 to 1993), we obtain

$$\Delta mrdrte = 0.413 - 0.104\Delta exec - 0.067\Delta unem$$

   The coefficient on $\Delta exec$ is negative and statistically significant (p-value $\approx .02$ against a two-sided alternative), suggesting a deterrent effect. One more execution reduces the murder rate by about .1, so 10 more executions reduce the murder rate by one (which means one murder per 100,000 people). The unemployment rate variable is no longer significant.

   (d) The change in executions may be at least partly related to changes in the expected murder rate, so that $\Delta exec$ is correlated with Δu in part (c). It might be reasonable to assume that $\Delta exec_{-1}$ is uncorrelated with Δu. (After all, $\Delta exec_{-1}$ depends on executions that occurred three ore more years ago!) Regress $\Delta exec$ on $\Delta exec_{-1}$ to see

if they are sufficiently correlated; interpret the coefficient on $\Delta exec_{-1}$.
Answer: The regression $\Delta exec$ on $\Delta exec_{-1}$ yeilds:

$$\Delta exec = 0.350 - 1.08\Delta exec_{-1}$$

which shows a strong negative correlation in the change in executions. This means that, apparently, states follow policies whereby if executions were high in the preceding three-year period, they are lower, one-for-one, in the next three-year period. Technically, to test the identification condition, we should add $\Delta unem$ to the regression. But its coefficient is small and statistically very insignificant, and adding it does not change the outcome at all.

(e) Reestimate the equation in part (c), using $\Delta exec_{-1}$ as an IV for $\Delta exec$. Assume that $\Delta unem$ is exogenous. How do your conclusions change from part (c)?
Answer: When the differenced equation is estimated using $\Delta exec_{-1}$ as an IV for $\Delta exec$, we obtain

$$\Delta mrdrte = 0.411 - 0.100\Delta exec - 0.067\Delta unem$$

This is very similar to when we estimate the differenced equation by OLS. Not surprisingly, the most important change is that the standard error on $\hat{\beta}_1$ is now larger and reduces the statistical significance of $\hat{\beta}_1$.

4. The purpose of this exercise is to compare the estimates and standard errors obtained by correctly using 2SLS with those obtained using inappropriate procedures. Use the data file WAGE2.dta.

(a) Use a 2SLS routine to estimate the equation:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \beta_4 black + u,$$

where *sibs* is the IV for *educ*. Report the results in the usual form.
Answer: The IV (2SLS) estimates are

$$\log(wage) = 5.22 + 0.0936 educ + 0.0209 exper + 0.0115 tenure - 0.183 black$$

(b) Now, manually carry out 2SLS. That is, first regress $educ_i$ on $sibs_i$, $exper_i$, $tenure_i$, and $black_i$ and obtain the fitted values, $\widehat{educ}_i$, $i = 1, \ldots, n$. Then, run the second stage regression $\log(wage_i)$ on $\widehat{educ}_i$, $exper_i$, $tenure_i$, and $black_i$, $i = 1, \ldots, n$. Verify that the $\hat{\beta}_j$ are identical to those obtained from part (a), but that the standard errors are somewhat different. The standard errors obtained from the second stage regression when manually carrying out 2SLS are generally inappropriate.

Answer: The coefficient on $\widehat{educ_i}$ in the second stage regression is, naturally, .0936. But the reported standard error is .0353, which is slightly too large.