Due: Mar. 26

Econ 4223 - Fall 2018 **Problem Set 4 - Solutions** **beginning of class**

Warm Up

1. Wooldridge (2015), exercise 9.2

Solution: Part (i): The coefficient on $voteA88$ implies that if candidate A had one more percentage point of the vote in 1988, she/he is predicted to have only .067 more percentage points in 1990. Or, 10 more percentage points in 1988 implies .67 points, or less than one point, in 1990. The t statistic is only about 1.26, and so the variable is insignificant at the 10% level against the positive one-sided alternative. (The critical value is 1.282.) While this small effect initially seems surprising, it is much less so when we remember that candidate A in 1990 is always the incumbent. Therefore, what we are finding is that, conditional on being the incumbent, the percent of the vote received in 1988 does not have a strong effect on the percent of the vote in 1990.

Part (ii): Naturally, the coefficients change, but not in important ways, especially once statistical significance is taken into account. For example, while the coefficient on log(expendA) goes from -.929 to -.839, the coefficient is not statistically or practically significant anyway (and its sign is not what we expect). The magnitudes of the coefficients in both equations are quite similar, and there are certainly no sign changes. This is not surprising given the insignificance of voteA88.

Exercises

2. *Wooldridge*, 9.4

Solution: Part (i): For the CEV assumptions to hold, we must be able to write $tvhours = tvhours^* + e_0$, where the measurement error e_0 has zero mean and is uncorrelated with $tvhours^*$ and each explanatory variable in the equation. (Note that for OLS to consistently estimate the parameters, we do not need e_0 to be uncorrelated with $tvhours^*$.)

Part (ii): The CEV assumptions are unlikely to hold in this example. For children who do not watch TV at all, $tvhours^* = 0$, and it is very likely that reported TV hours is zero. So if $tvhours^* = 0$, then $e_0 = 0$ with high probability. If $tvhours^* > 0$, the measurement error can be positive or negative, but since $tvhours \geq 0$, e_0 must satisfy $e_0 \geq -tvhours^*$. So e_0 and $tvhours^*$ are likely to be correlated. As mentioned in part (i), because it is the dependent variable that is measured with error, what is important is that e_0 is uncorrelated with the explanatory variables. But this is unlikely to be the case, because *tvhours** depends directly on the explanatory variables. Or, we might argue directly that more highly educated parents tend to underreport how much television their children watch, which means e_0 and the education variables are negatively correlated.

3. *Wooldridge*, 15.2

Solution: Part (i): It seems reasonable to assume that *dist* and u are uncorrelated because classrooms are not usually assigned with convenience for particular students in mind.

Part (ii): The variable *dist* must be partially correlated with *atndrte*. More precisely, in the reduced form

$$atndrte = \pi_0 + \pi_1 priGPA + \pi_2 ACT + \pi_3 dist + v,$$

we must have $\pi_3 \neq 0$. Given a sample of data, we can test $H_0 : \pi_3 = 0$ against $H_1 : \pi_3 \neq 0$ using a t test.

Part (iii): We now need instrumental variables for *atndrte* and the interaction term $priGPA \cdot atndrte$. (Even though *priGPA* is exogenous, *atndrte* is not, and so $priGPA \cdot atndrte$ is generally correlated with u.) Under the exogeneity assumption that $E(u|priGPA, ACT, dist) = 0$, any function of *priGPA*, *ACT*, and *dist* is uncorrelated with u. In particular, the interaction $priGPA \cdot dist$ is uncorrelated with u. If *dist* is partially correlated with *atndrte*, then $priGPA \cdot dist$ is partially correlated with $priGPA \cdot atndrte$. So, we can estimate the equation

$$stndfnl = \beta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 ACT + \beta_4 priGPA \cdot atndrte + u,$$

by 2SLS using IVs *dist*, *priGPA*, *ACT*, and $priGPA \cdot dist$. It turns out this is not generally optimal. It may be better to add $priGPA^2$ and $priGPA \cdot ACT$ to the instrument list. This would give us overidentifying restrictions to test. See Wooldridge (2002, Chapters 5 and 9) for further discussion.

Computer Exercises You should use `R` to complete these exercises. Any data set referred to in the question should be available in the `wooldridge` package in `R`. You do not need to turn in an R-script for these questions, but you are welcome to do so if you would like to.

4. *Wooldridge*, 9.C2 (Chapter 9, Computer Exercise 2)

Solution: Part (i): We estimate the model from column (2) but with KWW in place of IQ. The coefficient on *educ* becomes about .058 (se $\approx$ .006), so this is similar to the estimate obtained with IQ, although slightly larger and more precisely estimated.

Part (ii): When KWW and IQ are both used as proxies, the coefficient on *educ* becomes about .049 (se $\approx$ .007). Compared with the estimate when only KWW is used as a proxy, the return to education has fallen by almost a full percentage point.

Part (iii): The t statistic on IQ is about 3.08 while that on KWW is about 2.07, so each is significant at the 5% level against a two-sided alternative. They are jointly very significant, with $F(2, 925) \approx 8.59$ and p-value $\approx$ .0002.

5. *Wooldridge*, 15.C3

Solution: Part (i): IQ scores are known to vary by geographic region and so does the availability of four year colleges. It could be that, for a variety of reasons, people with higher abilities grow up in areas with four-year colleges nearby.

Part (ii): The simple regression of IQ on nearc4 gives

$$\widehat{IQ} = \underset{(0.63)}{100.61} + \underset{(0.74)}{2.60}\, nearc4$$
$$n = 2,061, R^2 = .0059,$$

which shows that predicted IQ score is about 2.6 points higher for a man who grew up near a four-year college. The difference is statistically significant (t statistic $\approx$ 3.51).

Part (iii): When we add $smsa66, reg662, \ldots, reg669$ to the regression from part (ii), we get

$$\widehat{IQ} = \underset{(1.62)}{104.77} + \underset{(.814)}{.348}\, nearc4 + \underset{(0.81)}{1.09}\, smsa66 + \ldots$$
$$n = 2,061, R^2 = .0626,$$

where, for brevity, the coefficients on the regional dummies are not reported. Now, the relationship between IQ and $nearc4$ is much weaker and statistically insignificant. In other words, once we control for region and environment while growing up, there is no apparent link between IQ score and living near a four-year college.

Part (iv): The findings from parts (ii) and (iii) show that it is important to include $smsa66, reg662, \ldots, reg669$ in the wage equation to control for differences in access to colleges that might also be correlated with ability

Cool Down

6. *Wooldridge*, 15.1.

Solution: Part (i): It has been fairly well established that socioeconomic status affects student performance. The error term u contains, among other things, family income, which has a positive effect on GPA and is also very likely to be correlated with PC ownership.

Part (ii): Families with higher incomes can afford to buy computers for their children. Therefore, family income certainly satisfies the second requirement for an instrumental variable: it is correlated with the endogenous explanatory variable [see (15.5) with $x = PC$ and $z = faminc$]. But as we suggested in part (i), $faminc$ has a positive affect on GPA, so the first requirement for a good IV, (15.4), fails for $faminc$. If we had $faminc$ we would include it as an explanatory variable in the equation; if it is the only important omitted variable correlated with PC, we could then estimate the expanded equation by OLS.

Part (iii): This is a natural experiment that affects whether or not some students own computers. Some students who buy computers when given the grant would not have without the grant. (Students who did not receive the grants might still own computers.) Define a dummy variable, $grant$, equal to one if the student received a grant, and zero otherwise. Then, if $grant$ was randomly assigned, it is uncorrelated with u. In particular, it is uncorrelated with family income and other socioeconomic factors in u. Further, $grant$ should be correlated with PC: the probability of owning a PC should be significantly higher for student receiving grants. Incidentally, if the university gave grant priority to low-income students, $grant$ would be negatively correlated with u, and IV would be inconsistent.

References

Wooldridge, Jeffrey M. 2015. *Introductory Econometrics: A Modern Approach*. Cengage Learning, 6 ed.