Problem Set 6

1. Suppose that you wish to estimate the effect of class attendance on student performance. A basic model is

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

where we have included previous GPA and ACT (a test score from school).

(i) Let *dist* be the distance from the students' living quarters to the lecture hall. Do you think *dist* is uncorrelated with u?

(ii) Assuming that *dist* and u are uncorrelated, what other assumption must *dist* satisfy to be a valid IV for *atndrte*?

(iii) Suppose we add the interaction term $priGPA \cdot atndrte$ as a control. If *atndrte* is correlated with u, then, in general, so is $priGPA \cdot atndrte$. What might be a good IV for $priGPA \cdot atndrte$? [Hint: If $E(u|priGPA, ACT, dist) = 0$, as happens when $priGPA$, ACT, and *dist* are all exogenous, then any function of $priGPA$ and *dist* is uncorrelated with u.]

2. Consider a simple model where the explanatory variable has classical measurement error:

$$\begin{aligned} y &= \beta_0 + \beta_1 x + u \\ w &= x + e, \end{aligned}$$

where u has zero mean and is uncorrelated with x and e. We observe y and w only. Assume that e has zero mean and is uncorrelated with x, and that x also has zero mean (this last assumption is only to simplify the algebra)

(i) Write $x = w - e$ and plug this into the true model. Show that the error term in the new equation, say, v, is negatively correlated with w if $\beta_1 > 0$. What does this imply about the OLS estimator of β_1 from the regression of y on w?

(ii) What happens to the bias if $\beta_1 < 0$? What do you conclude about the effect of measurement error on the estimate?

(iii) Suppose we have a second measurement of w, call this $\tilde{w}$, where $\tilde{w} = x + \tilde{e}$ and $\tilde{e}$ is independent of x, u, and e. Show that $\tilde{w}$ is a valid and relevant instrument for w.

(iv) Confirm your theoretical findings by using the data in 'wage2'. Choose a suitable measure of wages and compare an OLS regression using IQ, to an IV regression where you use KWW (knowledge of the world of work) as a second measure of intelligence as an instrument.

3. Use the data in 'card' for this exercise.

TABLE 15.1 Dependent Variable: log(*wage*)

Explanatory Variables	OLS	IV
educ	.075 (.003)	.132 (.055)
exper	.085 (.007)	.108 (.024)
$exper^2$	−.0023 (.0003)	−.0023 (.0003)
black	−.199 (.018)	−.147 (.054)
smsa	.136 (.020)	.112 (.032)
south	−.148 (.026)	−.145 (.027)
Observations *R*-squared	3,010 .300	3,010 .238
Other controls: *smsa66, reg662, …, reg669*		

(From Wooldridge 2016, Introductory Econometrics 6th Edition)

(i) In the table above, the difference between the IV and OLS estimates of the return to education is economically important. The IV results are obtained by using $nearc4$ as an instrument (an dummy variable for if the person lives near a 4-year college) Obtain the reduced form residuals, $\hat{v}$, from the reduced form regression: $educ$ on $nearc4$, $exper$, $expersq$, $black$, $smsa$, $south$, $smsa66$, $reg662$, …, $reg669$—see the table above ($smsa$ is an indicator for whether they lived in an urban area, and $reg\ldots$ are regional dummies). Use this to test whether $educ$ is exogenous; that is, determine if the difference between OLS and IV is statistically significant.

(ii) Estimate the equation by 2SLS, adding $nearc2$ as an extra instrument. Does the coefficient on *educ* change much?

(iii) Test the single over-identifying restriction from part (ii), i.e. test the validity of .

4. The data in 'fertil2' include, for women in Botswana during 1988, information on number of children, years of education, age, and religious and economic status variables.

   (i) Estimate the model

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

   by OLS and interpret 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?

   (ii) 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 (i), show that $frsthalf$ is a reasonable IV candidate for *educ*.

   (iii) Estimate the model from part (i) by using $frsthalf$ as an IV for *educ*. Compare the estimated effect of education with the OLS estimate from part (i).

   (iv) Add the binary variables *electric*, *tv*, and *bicycle* to the model and assume these are exogenous. Estimate the equation by OLS and 2SLS and compare the estimated coefficients on *educ*. Interpret the coefficient on *tv* and explain why television ownership has a negative effect on fertility.

5. A simple model to determine the effectiveness of condom usage on reducing sexually transmitted diseases (STDs) among sexually active high school students is

   $$infrate = \beta_0 + \beta_1 conuse + \beta_2 avginc + \beta_3 city + u_1,$$

   where $infrate$ = the percentage of sexually active students who have contracted and STD. *conuse* = the percentage of boys who claim to use condoms regularly. *avginc* = average family income. *city* = a dummy variable indicating whether a school is in a city. The model is at the school level.

   (i) Interpreting the preceding equation in a causal fashion, what should be the sign of β_1?

(ii) Why might $infrate$ and $conuse$ be jointly determined?

(iii) If condom usage increases with the rate of STDs, so that $\gamma_1 > 0$ in the equation

$$conuse = \gamma_0 + \gamma_1 infrate + u_2,$$

what is the likely bias in estimating β_1 by OLS?

(iv) Let $condis$ be a binary variable equal to 1 if a school has a program to distribute condoms. Explain how this can be used to estimate β_1 (and the other betas) by IV. What do we have to assume about $condis$ in each equation?

6. Use 'smoke' for this exercise.

(i) A model to estimate the effects of smoking on annual income (perhaps through lost work days due to illness or productivity effects) is

$$log(incom) = \beta_0 + \beta_1 cigs + \beta_2 educ + \beta_3 age + \beta_4 age^2 + u_1,$$

where $cigs$ is number of cigarettes smoked per day, on average. How do you interpret β_1?

(ii) To reflect the fact that cigarette consumption might be jointly determined with income, a demand for cigarettes equation is

$$\begin{aligned} cigs &= \beta_0 + \beta_1 log(income) + \beta_2 educ + \beta_3 age \\ &\quad + \beta_4 age^2 + \beta_5 log(cigpric) + \beta_6 restaurn + u_2, \end{aligned}$$

where $cigpric$ is the price of a pack of cigarettes (in cents) and $restaurn$ is a binary variable equal to 1 if the person lives in a state with restaurant smoking restrictions. Assuming these are exogenous to the individual, what signs would you expect for β_5 and β_6?

(iii) Thinking of the equations in part (i) and part (ii) as a system of two simultaneous equations, under what assumption is the income equation from part (i) identified?

(iv) Estimate the income equation by OLS and discuss the estimate of β_1.

(v) Estimate the reduced form for $cigs$. (Recall that this entails regressing $cigs$ on all exogenous variables.) Are $log(cigpric)$ and $restaurn$ significant in the reduced form?

(vi) Now, estimate the income equation by 2SLS. Discuss how the estimate of β_1 compares with the OLS estimate.

(vii) Do you think that cigarette prices and restaurant smoking restrictions are exogenous in the income equation?