Problem Set 5
Due Monday, May 11 at the beginning of lecture
ECON 381-2, Northwestern University, Spring 2015

You may work in groups of 2 or 3 as long as each group member turns in their own handwritten copy and all group members are clearly indicated at the top of every copy. No credit is given if no work is shown. For Stata/R problems include both your code and the output.

1. Consider the instrumental variables model

$$Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + U_i,$$

where we observe an i.i.d. sample $(Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i), i = 1, \ldots, N$ and we assume that these variables have finite fourth moments. Denote the dimension of $\boldsymbol{X}_i$ by K and the dimension of $\boldsymbol{Z}_i$ by M.

   (a) Define $\boldsymbol{C}_{tsls}$ as in the lecture notes, and let $\widetilde{\boldsymbol{X}}_i \equiv \boldsymbol{C}_{tsls}\boldsymbol{Z}_i$. Assume that $\mathbb{E}[\boldsymbol{Z}_i\boldsymbol{Z}_i']$ is invertible, and that $M \geq K$. Show that $\mathbb{E}[\widetilde{\boldsymbol{X}}_i\widetilde{\boldsymbol{X}}_i']$ is invertible if and only if $\mathbb{E}[\boldsymbol{X}_i\boldsymbol{Z}_i']$ has rank K. Explain the significance of this result.

   (b) Let $\boldsymbol{C}$ be any known or estimated $K \times M$–dimensional matrix with full rank and define $\widehat{\boldsymbol{\beta}}_{iv,\boldsymbol{C}}$ as the instrumental variables estimator that uses $\boldsymbol{C}\boldsymbol{Z}_i$ as an instrument for $\boldsymbol{X}_i$. Show that if $M = K$ then $\widehat{\boldsymbol{\beta}}_{iv,\boldsymbol{C}} = \widehat{\boldsymbol{\beta}}_{tsls} = \widehat{\boldsymbol{\beta}}_{ogmm}$ for any choice of $\boldsymbol{C}$, as long as all of these estimators exist. Explain the significance of this result.

   (c) Let $\widehat{U}_i \equiv Y_i - \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}_{iv,\boldsymbol{C}}$ for any choice of $\boldsymbol{C}$ for which $\widehat{\boldsymbol{\beta}}_{iv,\boldsymbol{C}}$ exists. Show that if $M = K$ then $\sum_{i=1}^N \boldsymbol{Z}_i\widehat{U}_i = \boldsymbol{0}$. Explain the significance of this result.

2. The following questions are about "The Effect of Prison Population Size on Crime Rates: Evidence From Prison Overcrowding Litigation" by Steven D. Levitt, which was published in *The Quarterly Journal of Economics* in May 1996. The article is available on Canvas. *Hint: An academic paper like this is very dense. I suggest you read the questions below, then quickly read/skim the paper, then go back and carefully read the parts that are pertinent to the questions.*

   (a) From Figure 1 we can see that between 1970 and 1992 the size of the prison population increased greatly while the number of violent and property crimes remained roughly flat. Can we conclude from this that imprisonment has a small effect on the crime rate? Explain why or why not.

   (b) At the bottom of pg. 322 Levitt says:

> *Consequently, OLS estimates of the effect of prisons on crime are likely to understate the true magnitude of the effect, perhaps dramatically.*

   Explain his reasoning. *Hint: Writing down some equations may help.*

   (c) Explain the last paragraph on pg. 335. What is Levitt doing here and why?

(d) Given Levitt's analysis, how would he argue that we should view the aggregate time series in Figure 1? Do you find Levitt's analysis convincing? Why or why not? *Note: There is no "right" answer on whether you find his analysis convincing or not, but there are good (well-reasoned) and bad (poorly-reasoned) arguments.*

3. Suppose that $\pi > 0$ is a constant. Show that for any random variable V_i,

$$\mathbb{1}[V_i \leq \pi] - \mathbb{1}[V_i \leq 0] = \mathbb{1}[V_i \in (0, \pi]].$$

Stata/R

4. This problem uses the data set `card.dta` (after economist David Card), available on Canvas.

   (a) Familiarize yourself with the data.

   (b) Regress log wage (`lwage`) on years of schooling (`educ`), while controlling for experience (`exper`) and experience squared (`expersq`). Do you think the coefficient on schooling is a good estimate of the causal effect of education on wages? If not, what is the likely direction of the asymptotic bias? In either case, how would you interpret its magnitude?

   (c) The data set contains a dummy named `nearc4` for whether the individual lived near a four-year college when they were younger (in 1966). Do you think that this is an exogenous instrument when controlling for `exper` and `expersq`? Why or why not? Regardless of your answer, compute the TSLS estimator with `lwage` as the dependent variable, using `nearc4` to instrument for `educ` and controlling for both `exper` and `expersq`. Discuss your results.

   (d) The data set contains the following additional variables: `black`, `smsa`, `south`, `smsa66`. What do these variables stand for? *Hint: Wikipedia may be helpful.*

   (e) Is it important to control for `black`, `smsa`, `south` and `smsa66` in an instrumental variables regression using `nearc4` to instrument for `educ`? Why or why not?

   (f) Compute the TSLS estimator with `lwage` as the dependent variable, using `nearc4` to instrument for `educ` and controlling for `exper`, `expersq`, `black`, `smsa`, `south` and `smsa66`. Discuss your results and compare them to the linear regression in part (b) and the instrumental variables regression(s) in part (**??**).

   (g) Add `nearc2` (dummy for near a 2-year college) as a second instrument in the previous part and compute the TSLS estimator. Discuss your results relative to the previous estimates.

   (h) Re-run all of the instrumental variables regressions (parts c), f) and g)) using OGMM instead. When are your results numerically identical and when are they different? Explain.

   (i) Determine whether there is a weak instruments problem (according to the rule of thumb) for all of the instrumental variables regressions in parts c), f) and g).

(j) (Not graded: *Conduct an overidentification test for specification in part g). Do you reject the null hypothesis in an overidentification test at the 5% level? Do you reject it at the 10% level?*)

5. This question is about the dataset `fish.dta`, which is available on Canvas. It contains 97 daily fish price and quantity observations at the Fulton Fish Market in New York City.

   (a) Familiarize yourself with the data set.

   (b) Run an OLS regression of log total quantity on log average price.

      i. If this were a supply curve, what would be the estimated price elasticity of supply? Does this make sense?

      ii. If this were a demand curve, what would be the estimated price elasticity of demand? Does this make sense? Is this a reliable estimate of the price elasticity of demand?

      iii. Explain how simultaneity bias can explain these findings.

   (c) Regress log total quantity on `speed3`. What is the coefficient on `speed3`? Is it significant? Do you think that `speed3` could affect fish supply? Could it also affect fish demand?

   (d) Run an instrumental variables regression of log total quantity on log average price, using `wave3` as an instrument for log average price. How should the coefficient on log average price be interpreted? Is there a weak instruments problem?

   (e) Run the same TSLS regression but now use both `wave3` and `speed3` as instruments. Compare your findings to part (d). Do you have a weak instrument problem? (Not graded: *Can you reject the null hypothesis (at 5% significance) that both instruments are exogenous?*) Do your results change if you use OGMM instead of TSLS? Provide a plausible explanation for your findings. Was adding the second instrument a good or bad idea?

   (f) Run the same TSLS regression using both `wave3` and `wave2` as instruments. (Not graded: *Can you reject the hypothesis that these instruments are exogenous?*) Do you have a weak instruments problem? Do you prefer this specification or the specification in (d)? Do your results change if you use OGMM instead of TSLS?