Economics 675: Applied Microeconometrics – Fall 2018

Assignment 5 – Due date: Mon 19-Nov

Last updated: June 4, 2018

Contents

Guidelines:

- You may work in (small) groups while solving this assignment.
- Submit individual solutions via `http://canvas.umich.edu` in one PDF file collecting everything (e.g., derivations, figures, tables, computer code).
- Start each question on a separate page. Always add a reference section if you cite other sources.
- Clearly label all tables and figures, and always include a brief footnote with useful information.
- Always attach your computer code as an appendix, with annotations/comments as appropriate.
- Please provide as much detail as possible in your answers, both analytical and empirical.

1 Question 1: Many Instruments Asymptotics

This question considers some of the asymptotic properties of IV estimators with "many" instruments. Before working on this question, it may be useful to read Newey (2002) and Hansen, Hausman, and Newey (2008). Consider a textbook example of IV model with non-random instruments and the usual regularity conditions. Let $\{(u_i, v_i) : 1 \leq i \leq n\}$ be random sample of mean-zero finite fourth moments random variables, and consider the model:

$$\begin{aligned} &\mathbf{y} = \mathbf{x}\beta + \mathbf{u}, \qquad \mathbb{E}[u_i] = 0, \qquad \mathbb{V}[u_i] = \sigma_u^2 \\ &\mathbf{x} = \mathbf{Z}\boldsymbol{\pi} + \mathbf{v}, \qquad \mathbb{E}[v_i] = 0, \qquad \mathbb{V}[v_i] = \sigma_v^2, \qquad \mathbb{E}[u_i v_i] = \sigma_{uv}^2, \end{aligned}$$

where $\mathbf{y} = (y_1, \ldots, y_n)' \in \mathbb{R}^n$, $\mathbf{x} = (x_1, \ldots, x_n)' \in \mathbb{R}^n$, $\mathbf{u} = (u_1, \ldots, u_n)' \in \mathbb{R}^n$, $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n]' \in \mathbb{R}^{n \times K}$ with $\mathbf{z} = (z_1, \ldots, z_n)' \in \mathbb{R}^n$, $\mathbf{v} = (v_1, \ldots, v_n)' \in \mathbb{R}^n$, and $\{\mathbf{z}_n : n \geq 1\}$ a non-random sequence. Here, $\beta \in \mathbb{R}$ and $\boldsymbol{\pi} \in \mathbb{R}^K$. The only twist relative to the classical IV model is that we assume $K = K_n \to \infty$: we consider many (strong) instruments asymptotics. In particular, we assume:

$$\frac{K}{n} \to \rho \in [0, 1) \qquad \text{and} \qquad \frac{\boldsymbol{\pi}'\mathbf{Z}'\mathbf{Z}\boldsymbol{\pi}}{n} \to \mu > 0.$$

We consider two estimators, the classical 2SLS estimator and a bias-corrected version of it:

$$\widehat{\beta}_{\texttt{2SLS},n} = (\mathbf{x}'\mathbf{P}\mathbf{x})^{-1}\mathbf{x}'\mathbf{P}\mathbf{y} \qquad \text{and} \qquad \widehat{\beta}_{\texttt{2SLS}-\texttt{BC},n} = (\mathbf{x}'\check{\mathbf{P}}\mathbf{x})^{-1}\mathbf{x}'\check{\mathbf{P}}\mathbf{y},$$

where $\mathbf{P} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ and $\check{\mathbf{P}} = \mathbf{P} - \frac{K}{n}\mathbf{I}_n$, with $\mathbf{I}_n$ the $n \times n$ identity matrix.

We impose the following conditions throughout:

$$\mathbb{E}[\mathbf{u}|\check{\mathbf{v}}] = 0, \qquad \mathbb{V}[\mathbf{u}|\check{\mathbf{v}}] = \sigma_u^2\mathbf{I}_n, \qquad \check{\mathbf{v}} = \mathbf{v} - \frac{\sigma_{uv}^2}{\sigma_u^2}\mathbf{u}.$$

Please answer the following questions, providing sufficient conditions as needed and explaining your reasoning in detail.

1. Show that $\mathbb{E}[\mathbf{u}'\mathbf{u}/n] = \sigma_u^2$, $\mathbb{E}[\mathbf{v}'\mathbf{v}/n] = \sigma_v^2$,$\mathbb{E}[\mathbf{x}'\mathbf{u}/n] = \sigma_{uv}^2$, $\mathbb{E}[\mathbf{x}'\mathbf{P}\mathbf{u}/n] = K\sigma_{uv}^2/n$, and $\mathbb{E}[\mathbf{u}'\mathbf{P}\mathbf{u}/n] = K\sigma_u^2/n$.

2. Show that

$$\frac{\mathbf{x}'\mathbf{x}}{n} \to_p \mu + \sigma_v^2, \qquad \frac{\mathbf{x}'\mathbf{P}\mathbf{x}}{n} \to_p \mu + \rho\sigma_v^2, \qquad \frac{\mathbf{x}'\mathbf{P}\mathbf{u}}{n} \to_p \rho\sigma_{uv}^2.$$

3. Show that

$$\widehat{\beta}_{\texttt{2SLS},n} = \beta + \frac{\rho\sigma_{uv}^2}{\mu + \rho\sigma_v^2} + o_p(1).$$

   Under which conditions is $\widehat{\beta}_{\texttt{2SLS},n}$ consistent?

4. Show that $\widehat{\beta}_{\texttt{2SLS}-\texttt{BC},n} \to_p \beta$.

5. Employing the following steps, show that

$$\sqrt{n}(\widehat{\beta}_{\texttt{2SLS}-\texttt{BC},n} - \beta) \to_d \mathcal{N}(0, V(\rho)),$$

and characterize precisely the limiting variance $V(\rho) \in \mathbb{R}$.

(a) Verify that

$$\mathbf{x}'\check{\mathbf{P}}\mathbf{u} = \boldsymbol{\pi}'\mathbf{Z}'\left(\mathbf{P} - \frac{K}{n}\mathbf{I}_n\right)\mathbf{u} + \check{\mathbf{v}}'\left(\mathbf{P} - \frac{K}{n}\mathbf{I}_n\right)\mathbf{u} + \frac{\sigma_{uv}^2}{\sigma_u^2}\mathbf{u}'\left(\mathbf{P} - \frac{K}{n}\mathbf{I}_n\right)\mathbf{u}.$$

(b) Show that $\mathbb{E}[\boldsymbol{\pi}'\mathbf{Z}'(\mathbf{P} - (K/n)\mathbf{I}_n)\mathbf{u}] = 0$ and

$$\frac{1}{\sqrt{n}}\boldsymbol{\pi}'\mathbf{Z}'\left(\mathbf{P} - \frac{K}{n}\mathbf{I}_n\right)\mathbf{u} \to_d \mathcal{N}(0, V_1(\rho)), \qquad V_1(\rho) = \lim_{n\to\infty} \mathbb{V}\left[\frac{1}{\sqrt{n}}\boldsymbol{\pi}'\mathbf{Z}'\left(\mathbf{P} - \frac{K}{n}\mathbf{I}_n\right)\mathbf{u}\right],$$

and give the exact formula of $V_1(\rho)$.

(c) Show that $\mathbb{E}[\check{\mathbf{v}}'(\mathbf{P} - (K/n)\mathbf{I}_n)\mathbf{u}] = 0$ and $\check{\mathbf{v}}'(\mathbf{P} - (K/n)\mathbf{I}_n)\mathbf{u} = O_p(\sqrt{K})$.

(d) Show that $\mathbb{E}[\mathbf{u}'(\mathbf{P} - (K/n)\mathbf{I}_n)\mathbf{u}] = 0$ and $\mathbf{u}'(\mathbf{P} - (K/n)\mathbf{I}_n)\mathbf{u} = O_p(\sqrt{K})$.

(e) Show that $\mathbb{E}[\mathbf{x}'\check{\mathbf{P}}\mathbf{u}] = 0$, and compute $\vartheta_n^2 := \mathbb{V}[\mathbf{x}'\check{\mathbf{P}}\mathbf{u}/\sqrt{n}]$. Also, assume that

$$\frac{\mathbf{x}'\check{\mathbf{P}}\mathbf{u}/\sqrt{n}}{\vartheta_n} \to_d \mathcal{N}(0,1)$$

- **Extra credit**: Show this last result.

(f) Using the above, show that $\sqrt{n}(\widehat{\beta}_{\texttt{2SLS-BC},n} - \beta) \to_d \mathcal{N}(0, V(\rho))$.
Characterize the form of $V(\rho)$. What happens when $\frac{K}{n} \to \rho = 0$?

2 Question 2: Weak Instruments Simulations

In this problem we consider the finite sample properties of the two-stage least squares (2SLS) estimator for different identification strengths, that is, cases where the instrument is "strong" and cases where the instrument is "weak". Stock, Wright, and Yogo (2002) gives a comprehensive yet accessible discussion of this problem.

Consider the following data generating process:

$$y_i = \beta \cdot x_i + u_i,$$
$$x_i = \gamma \cdot z_i + v_i,$$

where we set, without loss of generality, $\beta = 0$, and consider a random sample $i = 1, 2, \cdots, n$ satisfying

$$\begin{bmatrix} z_i \\ u_i \\ v_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0.99 \\ 0 & 0.99 & 1 \end{bmatrix} \right).$$

The parameter γ gives a measure of instrument relevance, taking values

$$n \cdot \gamma^2 \in \{0,\ 0.25,\ 9,\ 99\} \qquad \Rightarrow \qquad \gamma \in \sqrt{\frac{1}{n} \cdot \{0,\ 0.25,\ 9,\ 99\}}.$$

Here we specify values of $n \cdot \gamma^2$ rather than γ because the former could be interpreted as the F-statistic of the first-stage regression.

Let the sample size n be 200. For each simulation replication, estimate the regression coefficient of y_i on x_i with either OLS or 2SLS, and report the following quantities:

- $\widehat{\beta}_{\texttt{OLS}}$: the OLS estimate
- $\mathsf{se}(\widehat{\beta}_{\texttt{OLS}})$: the standard error of the OLS estimate
- $\mathbf{1}_{\texttt{Rej,OLS}} \overset{\text{def}}{=} |\widehat{\beta}_{\texttt{OLS}}/\mathsf{se}(\widehat{\beta}_{\texttt{OLS}})| > 1.96$, i.e. whether the null hypothesis $\mathsf{H}_0 : \beta = 0$ is rejected at the 95% level
- $\widehat{\beta}_{\texttt{2SLS}}$: the 2SLS estimate
- $\mathsf{se}(\widehat{\beta}_{\texttt{2SLS}})$: the standard error of the 2SLS estimate
- $\mathbf{1}_{\texttt{Rej,2SLS}} \overset{\text{def}}{=} |\widehat{\beta}_{\texttt{2SLS}}/\mathsf{se}(\widehat{\beta}_{\texttt{2SLS}})| > 1.96$, i.e. whether the null hypothesis $\mathsf{H}_0 : \beta = 0$ is rejected at the 95% level
- $\widehat{F}_{\texttt{2SLS}}$: the first-stage F-statistic of the 2SLS.

Conduct 5,000 Monte Carlo simulations for each γ value, and complete Table 1 of summary statistics. Also give a discussion on how weak instrument affects estimation and inference, providing intuition of why the usual asymptotic approximations may not deliver a good approximation in some cases.

Table 1: Weak Instrument Summary Statistics

(a) $\gamma^2 = 0/n$ ($F \approx 1$)

	mean	**st.dev.**	**quantiles**		
			0.1	0.5	0.9
OLS					
$\widehat{\beta}$					
$\mathsf{se}(\widehat{\beta})$					
$\mathbf{1}_{\mathtt{rej}}$					
2SLS					
$\widehat{\beta}$					
$\mathsf{se}(\widehat{\beta})$					
$\mathbf{1}_{\mathtt{rej}}$					
$\widehat{F}$					

(b) $\gamma^2 = 0.25/n$ ($F \approx 1.25$)

	mean	**st.dev.**	**quantiles**		
			0.1	0.5	0.9
OLS					
$\widehat{\beta}$					
$\mathsf{se}(\widehat{\beta})$					
$\mathbf{1}_{\mathtt{rej}}$					
2SLS					
$\widehat{\beta}$					
$\mathsf{se}(\widehat{\beta})$					
$\mathbf{1}_{\mathtt{rej}}$					
$\widehat{F}$					

(c) $\gamma^2 = 9/n$ ($F \approx 10$)

	mean	**st.dev.**	**quantiles**		
			0.1	0.5	0.9
OLS					
$\widehat{\beta}$					
$\mathsf{se}(\widehat{\beta})$					
$\mathbf{1}_{\mathtt{rej}}$					
2SLS					
$\widehat{\beta}$					
$\mathsf{se}(\widehat{\beta})$					
$\mathbf{1}_{\mathtt{rej}}$					
$\widehat{F}$					

(d) $\gamma^2 = 99/n$ ($F \approx 100$)

	mean	**st.dev.**	**quantiles**		
			0.1	0.5	0.9
OLS					
$\widehat{\beta}$					
$\mathsf{se}(\widehat{\beta})$					
$\mathbf{1}_{\mathtt{rej}}$					
2SLS					
$\widehat{\beta}$					
$\mathsf{se}(\widehat{\beta})$					
$\mathbf{1}_{\mathtt{rej}}$					
$\widehat{F}$					

3 Question 3: Weak Instrument - Empirical Study

This question revisits two influential papers in the literature on weak/many IV models. See, e.g., Stock, Wright, and Yogo (2002), Hansen, Hausman, and Newey (2008), and references therein, for more details.

3.1 Angrist and Krueger (1991)

The Angrist and Krueger (1991) study uses quarter of birth (together with other variables) as instruments for educational attainment in the wage equation. See Appendix 4 for detailed data description. In all the following regressions, you should control for (i) `race`, (ii) `marrital status`, (iii) `SMSA`, (iv) dummies for `region`, and (iv) dummies for `YoB_ld`.

1. **OLS 1**: regress `log weekly wage` on `educational attainment`, with controls mentioned above.

   (This is Table V Column (5) in Angrist and Krueger (1991).)

2. **OLS 2**: regress `log weekly wage` on `educational attainment`, with additional control variables `age_q` and `age_sq`.

   (This is Table V Column (7) in Angrist and Krueger (1991).)

3. **2SLS 1**: regress `log weekly wage` on `educational attainment` with the default controls, and use cross dummies of `QoB` and `YoB_ld` as instruments. (Hint: to create cross dummies of two categorical variables, use `i.QoB##i.YoB_ld`.)

   (This is Table V Column (6) in Angrist and Krueger (1991).)

4. **2SLS 2**: regress `log weekly wage` on `educational attainment` with the default controls as well as `age_q` and `age_sq`, and use cross dummies of `QoB` and `YoB_ld` as instruments.

   (This is Table V Column (8) in Angrist and Krueger (1991).)

Report point estimates and corresponding standard errors. Provide a detailed discussion of the empirical findings, including the direction and magnitude of the OLS bias (if any). What is a reasonable estimate of the return to education? What are possible problems with the data, specification and distribution theory that could cast doubt on the validity of the empirical results?

3.2 Bound, Jaeger, and Baker (1995)

Bound, Jaeger, and Baker (1995) uses *permuted* quarter of birth (`QoB`) as instruments to address the weak instrument problem in Angrist and Krueger (1991) and its implications. More specifically, the quarter of birth variable is randomly permuted hence it is no longer correlated with educational attainment (i.e. no longer a relevant instrument). Then by using the permuted instrument in the 2SLS one learns the distributional property of the estimator under weak instrument (actually, "irrelevant" instruments).

5. **2SLS 1 (permute)**: for each simulation permute the quarter of birth variable `QoB` and run the model specified in **2SLS 1**. Repeat 500 times and report the average and standard deviation of the point estimates of the return to education.

   (This is Table 3 Column (1) in Bound, Jaeger, and Baker (1995).)

6. **2SLS 2 (permute)**: for each simulation permute the quarter of birth variable `QoB` and run the model specified in **2SLS 2**. Repeat 500 times and report the average and standard deviation of the point estimates of the return to education.

   (This is Table 3 Column (2) in Bound, Jaeger, and Baker (1995).)

What can you learn from this exercise? Give precise intuition on the empirical findings, and how these compare to the first part of this question. In particular, compare the standard errors in part 3.1 with the standard deviations of the permutation estimates herein. Does the standard error estimate give a good approximation to the finite sample variability of the 2SLS estimator?

NOTE: If you employ Stata to answer this question, you may notice that the command `ivregress` could be very slow for this permutation exercise. See Appendix 5 for a faster alternative.

4 Appendix: `Angrist_Krueger.dta` Data Description

Due to school start age policy and compulsory school attendance laws, the season of birth is related to the educational attainment. The typical compulsory school attendance law requires a student to start first grade in the fall of the calendar year in which he or she turns age 6 and to continue attending school until he or she turns 16. Thus an individual born in the early months of the year will usually enter first grade when he or she is close to age 7 and will reach age 16 in the middle of tenth grade. An individual born in the third or fourth quarter will typically start school either just before or just after turning age 6 and will finish tenth grade before reaching age 16 (i.e. more likely to finish tenth grade).

It is argued in Angrist and Krueger (1991) that the season of birth should not affect other outcome variables such as earnings, and quarter of birth (together with other variables) is used as instruments for educational attainment in the wage equation.

`Angrist_Krueger.dta` is a subsample of the original Angrist and Krueger (1991) dataset, which contains all individuals who were born between 1930 and 1939, and took the U.S. Census survey in 1980. The sample size is 329,509. The variables are summarized below:

Variable	**Mean**	**St.Dev.**	**Description**
l_w_wage	5.8999	.6788	log weekly wage
QoB	2.5064	1.1119	quarter of birth
age	44.645	2.9397	age measured in years
age_q	45.021	2.9207	age measured in quarters
age_sq			squared age (in quarters)
YoB	34.603	2.9050	year of birth
YoB_ld	4.6028	2.9050	last digit of YoB, x=193x
educ	12.770	3.2812	education attainment measured in years
married	.8626	.3443	dummy variable for marital status, 1 =married
non_white	.0817	.2739	dummy variable for race, 1 =black
SMSA	.1863	.3893	residence in SMSA, 1 =center city
region			2 =ESOCENT, 3 =MIDATL, 4 =MT, 5 =NEWENG, 6 =SOATL, 7 =WNOCENT, 8 =WSOCENT, 9 =ENOCE
ENOCENT	.2015	.4011	East North Central
ESOCENT	.0655	.2473	East South Central
MIDATL	.1617	.3682	Mid Atlantic
MT	.0494	.2167	Mountain
NEWENG	.0562	.2302	New England
SOATL	.1681	.3739	South Atlantic
WNOCENT	.0780	.2682	West North Central
WSOCENT	.0969	.2959	West South Central

5 Appendix: A (Slightly) Faster 2SLS Regression

The Stata command `ivregress` could be very slow when the sample size is large and there are many covariates/instruments. For the permutation exercise in Bound, Jaeger, and Baker (1995) we only need the point estimate. The following program is much faster for this purpose[1].

```
*******************************************************************************************
* A faster IV regression program
* Xinwei Ma
* Oct 28, 2015
*******************************************************************************************
capture program drop IV_quick
program define IV_quick, rclass
    syntax varlist(max=1) [, model(integer 1) ]
    local x "`varlist'"

    if (`model' == 1) {
        capture drop educ_hat
        qui reg educ non_white married SMSA i.region i.YoB_ld i.YoB_ld##i.`x'
        predict educ_hat
        qui reg l_w_wage educ_hat non_white married SMSA i.region i.YoB_ld
        return scalar beta = _b[educ_hat]
    }

    if (`model' == 2) {
        capture drop educ_hat
        qui reg educ non_white married SMSA age_q age_sq i.region i.YoB_ld i.YoB_ld##i.`x'
        predict educ_hat
        qui reg l_w_wage educ_hat non_white married SMSA age_q age_sq i.region i.YoB_ld
        return scalar beta = _b[educ_hat]
    }
end
*******************************************************************************************
```

Here `model(1)` and `model(2)` correspond to **2SLS 1** and **2SLS 2** in Question 3, respectively. The program takes only one variable, `QoB`, since it is what we would like to permute. Try the following,

```
* 2SLS 1, point estimate
IV_quick QoB, model(1)
display r(beta)

* 2SLS 2, point estimate
IV_quick QoB, model(2)
display r(beta)
```

[1] Of course the most time efficient way to do this is to code the estimator from scratch, by using matrix algebra.

References

Angrist, J. D., and A. B. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?," *The Quarterly Journal of Economics*, 106(4), 979–1014.

Bound, J., D. A. Jaeger, and R. M. Baker (1995): "Problems With instrumental Variables Estimation When the Correlation Between the instruments and the Endogenous Explanatory Variable is Weak," *Journal of the American Statistical Association*, 90(430), 443–450.

Hansen, C., J. Hausman, and W. K. Newey (2008): "Estimation With Many Instrumental Variables," *Journal of Business and Economic Statistics*, 26(4), 398–422.

Newey, W. K. (2002): "Many Instrument Asymptotics," Working paper, MIT.

Stock, J. H., J. H. Wright, and M. Yogo (2002): "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business & Economic Statistics*, 20(4).