Economics 675: Applied Microeconometrics
Fall 2018 - Assignment 5

Paul R. Organ*

November 19, 2018

Contents

1 Question 1: Many Instruments Asymptotics **2**

2 Question 2: Weak Instruments Simulations **5**

3 Question 3: Weak Instruments - Empirical Study **7**
3.1 Angrist and Krueger (1991) 7
3.2 Bound, Jaeger, and Baker (1995) 7

A R Code **9**

B Stata Code **14**

*prorgan@umich.edu

1 Question 1: Many Instruments Asymptotics

1. We have the following:

- $\mathbb{E}[\frac{\mathbf{u}'\mathbf{u}}{n}] = \mathbb{E}[\frac{1}{n}\sum_i u_i^2] = \mathbb{E}[u_i^2] = \mathbb{V}[u_i] = \sigma_u^2$
- $\mathbb{E}[\frac{\mathbf{v}'\mathbf{v}}{n}] = \mathbb{E}[\frac{1}{n}\sum_i v_i^2] = \mathbb{E}[v_i^2] = \mathbb{V}[v_i] = \sigma_v^2$
- $\mathbb{E}[\frac{\mathbf{x}'\mathbf{u}}{n}] = \mathbb{E}[\frac{\mathbf{v}'\mathbf{u}}{n}] = \sigma_{uv}^2$, since $\mathbf{Z}$ is non-random
- $\mathbb{E}[\frac{\mathbf{x}'\mathbf{P}\mathbf{x}}{n}] = \mathbb{E}[\frac{\mathbf{v}'\mathbf{P}\mathbf{u}}{n}] = \frac{K\sigma_{uv}^2}{n}$, since $\mathbb{E}[v_i u_j] = 0$ for all $i \neq j$, and $\sum_{i=1}^n p_{ii} = K$
- $\mathbb{E}[\frac{\mathbf{u}'\mathbf{P}\mathbf{u}}{n}] = \frac{K\sigma_u^2}{n}$, following the same argument just above this

2. First, by the LLN we have

$$\frac{\mathbf{x}'\mathbf{x}}{n} = \frac{\mathbf{v}'\mathbf{v}}{n} + 2\frac{\mathbf{v}'\mathbf{Z}\boldsymbol{\pi}}{n} + \frac{\boldsymbol{\pi}'\mathbf{Z}'\mathbf{Z}\boldsymbol{\pi}}{n} \to_p \sigma_v^2 + \mu$$

since the first term goes to σ_v^2 as shown above, the second term drops out since $e[\mathbf{v}'\mathbf{Z}\boldsymbol{\pi}] = 0$, and the third term goes to μ by assumption.

Next, again using the LLN we have

$$\frac{\mathbf{x}'\mathbf{P}\mathbf{x}}{n} = \frac{\mathbf{v}'\mathbf{P}\mathbf{v}}{n} + 2\frac{\mathbf{v}'\mathbf{Z}\boldsymbol{\pi}}{n} + \frac{\boldsymbol{\pi}'\mathbf{Z}'\mathbf{Z}\boldsymbol{\pi}}{n} \to_p \rho\sigma_v^2 + \mu$$

using that $\mathbf{P}\mathbf{Z} = \mathbf{Z}$, so the arguments for the second and third term are the same. For the first term, we have $\mathbb{E}[\mathbf{v}'\mathbf{P}\mathbf{v}] = \sum_{i=1}^n p_{ii}\sigma_v^2/n = K\sigma_v^2/n = \rho$.

Finally, by LLN and using the equality shown in part (1), we have

$$\frac{\mathbf{x}'\mathbf{P}\mathbf{u}}{n} \to_p \mathbb{E}\Big[\frac{\mathbf{x}'\mathbf{P}\mathbf{u}}{n}\Big] = \rho\sigma_{uv}^2$$

3. Using our results from above, and the Continuous Mapping Theorem, we have

$$\hat{\beta}_{2SLS,n} = \beta + (\mathbf{x}'\mathbf{P}\mathbf{x}/n)^{-1}(\mathbf{x}'\mathbf{P}\mathbf{u}/n) \to_p \beta + \frac{\rho\sigma_{uv}^2}{\mu + \rho\sigma_u^2} + o_p(1)$$

For $\hat{\beta}_{2SLS,n}$ to be consistent, we need the second term to drop out. This requires either (1) ρ is 0 (i.e., if n is large and K is small, so we have many fewer instruments than observations) or (2) σ_{uv}^2 is 0 (i.e., exogeneity).

4. Using the definition of $\check{\mathbf{P}}$ and the previous results, we have

$$\begin{aligned}
\hat{\beta}_{2SLS-BC,n} &= \beta + \Big(\mathbf{x}'\big(\mathbf{P} - \frac{K}{n}\mathbf{I}_n\big)\mathbf{x}/n\Big)^{-1}\Big(\mathbf{x}'\big(\mathbf{P} - \frac{K}{n}\mathbf{I}_n\big)\mathbf{u}/n\Big) \\
&= \beta + \Big(\mathbf{x}'\mathbf{P}\mathbf{x}/n - \frac{K}{n}\mathbf{x}'\mathbf{x}/n\Big)^{-1}\Big(\mathbf{x}'\mathbf{P}\mathbf{u}/n - \frac{K}{n}\mathbf{x}'\mathbf{u}/n\Big) \\
&\to_p \beta + \Big(\mu + \rho\sigma_v^2 - \rho(\mu + \sigma_v^2)\Big)^{-1}\Big(\rho\sigma_{uv}^2 - \rho\sigma_{uv}^2\Big) + o_p(1)
\end{aligned}$$

Note that the denominator of the second term is zero, and thus we have consistency.

5a. First, note that $\check{\mathbf{v}} = \mathbf{v} - \frac{\sigma_{uv}^2}{\sigma_u^2}\mathbf{u} \implies \mathbf{v} = \check{\mathbf{v}} + \frac{\sigma_{uv}^2}{\sigma_u^2}\mathbf{u}$. Then it is straightforward to verify:

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

5b. For the first result, we have that $\mathbf{u}$ itself is mean-zero, hence this expectation will be zero as well.

The second result (asympotic distribution) follows by the law of large numbers. To calculate the variance, we have:

$$\begin{aligned}
\mathbb{V}\Big[\frac{1}{\sqrt{n}}\boldsymbol{\pi}'\mathbf{Z}'\big(\mathbf{P} - \frac{K}{n}\mathbf{I}_n\big)\mathbf{u}\Big] &= \frac{1}{n}\boldsymbol{\pi}'\mathbf{Z}'\Big(\mathbf{P} - \frac{K}{n}\mathbf{I}_n\Big)(\sigma_u^2\mathbf{I}_n)\Big(\mathbf{P} - \frac{K}{n}\mathbf{I}_n\Big)\mathbf{Z}\boldsymbol{\pi} \\
&= \sigma_u^2\frac{1}{n}\boldsymbol{\pi}'\mathbf{Z}'\Big(\mathbf{P} - 2\frac{K}{n}\mathbf{P} + \frac{K^2}{n^2}\mathbf{I}_n\Big)\mathbf{Z}\boldsymbol{\pi} \\
&= \sigma_u^2\mu(1-\rho)^2 + o(1)
\end{aligned}$$

so that $V_1(\rho) = \sigma_u^2\mu(1-\rho)^2$.

5c. The first result follows from the assumption that $\mathbb{E}[\mathbf{u}|\hat{\mathbf{v}}] = 0$.

For the second result, we first estimate the corresponding variance:

$$\begin{aligned}
\mathbb{V}\big[\hat{\mathbf{v}}'(\mathbf{P} - \frac{K}{n}\mathbf{I}_n)\mathbf{u}\big] &= \mathbb{E}\big[\hat{\mathbf{v}}'(\mathbf{P} - \frac{K}{n}\mathbf{I}_n)\mathbf{u}\mathbf{u}'(\mathbf{P} - \frac{K}{n}\mathbf{I}_n)\hat{\mathbf{v}}\big] \\
&= \mathbb{E}\big[\hat{\mathbf{v}}'(\mathbf{P} - \frac{K}{n}\mathbf{I}_n)(\sigma_u^2\mathbf{I}_n)(\mathbf{P} - \frac{K}{n}\mathbf{I}_n)\hat{\mathbf{v}}\big] \\
&= \sigma_u^2\Big(\mathbb{E}[\hat{\mathbf{v}}'\mathbf{P}\hat{\mathbf{v}}] - 2\frac{K}{n}\mathbb{E}[\hat{\mathbf{v}}'\mathbf{P}\hat{\mathbf{v}}] + \frac{K^2}{n^2}\mathbb{E}[\hat{\mathbf{v}}'\hat{\mathbf{v}}]\Big) \\
&= \sigma_u^2\Big(\sigma_{\hat{\mathbf{v}}}^2\sum_{i=1}^n P_{ii} - 2\frac{K}{n}\sigma_{\hat{\mathbf{v}}}^2\sum_{i=1}^n P_{ii} + K\frac{K}{n}\sigma_{\hat{\mathbf{v}}}^2\Big) \\
&= K\sigma_u^2\sigma_{\hat{\mathbf{v}}}^2(1 - 2\rho + \rho) + o(1) \\
&= K\sigma_u^2\sigma_{\hat{\mathbf{v}}}^2(1-\rho) + o(1) \\
&= O(K)
\end{aligned}$$

We can then apply the Markov inequality to show the desired result.

5d. For the first result, we have

$$\mathbb{E}\Big[\mathbf{u}'(\mathbf{P} - \frac{K}{n}\mathbf{I}_n)\mathbf{u}\Big] = \sigma_u^2 \cdot \operatorname{tr}\Big(\mathbf{P} - \frac{K}{n}\mathbf{I}_n\Big) = \sum_{i=1}^n\Big(P_{ii} - \frac{K}{n}\Big) = 0$$

For the second result, we again estimate the corresponding variance:

$$
\begin{aligned}
\mathbb{V}\Big[\mathbf{u}'\big(\mathbf{P}-\frac{K}{n}\mathbf{I}_n\big)\mathbf{u}\Big] &= \mathbb{E}\big[(\mathbf{u}'\check{\mathbf{P}}\mathbf{u})^2\big] \\
&= \mathbb{E}\Big[\big(\sum_{1\le i,j\le n}\check{P}_{ij}u_iu_j\big)^2\Big] \\
&= \mathbb{E}\Big[\sum_{1\le i\le n}\check{P}_{ii}^2u_i^2\Big] + \mathbb{E}\Big[\sum_{1\le i\ne j\le n}(2\check{P}_{ij}+\check{P}_{ii}\check{P}_{jj})u_i^2u_j^2\Big] \\
&= \sigma_u^2\Big[\sum_{1\le i\le n}\check{P}_{ii}^2\Big] + \sigma_u^4\Big[\sum_{1\le i\ne j\le n}(2\check{P}_{ij}+\check{P}_{ii}\check{P}_{jj})\Big] \\
&= O(K)
\end{aligned}
$$

And again, we can apply the Markov inequality to show the desired result.

5e. This result follows by combining what we shouled in parts (a) through (d) above.

Looking at $\mathbb{E}[\mathbf{x}'\check{\mathbf{P}}\mathbf{u}]$, note that if we plug in each of the three components, they are additively separable, and the expecation of each component is zero, as we showed above. Hence the total expectation is zero. The variance result follows similarly.

2 Question 2: Weak Instruments Simulations

See the underlying code in the appendices (Appendix A for R and Appendix B for STATA).

My results are shown in Table 1 for R and in Table 2 for STATA. We see that when the instruments are weak (i.e., when $\gamma^2 \cdot n$ is small), the 2SLS estimate for β is closer to the OLS estimate.

Table 1: Weak Instrument Summary Statistics: R

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

	mean	sd	q1	q5	q9
OLS_Beta	0.990	0.010	0.977	0.990	1.002
OLS_seBeta	0.010	0.001	0.009	0.010	0.011
OLS_1Rej	1.000	0.000	1.000	1.000	1.000
2SLS_Beta	1.049	5.772	0.544	0.991	1.451
2SLS_seBeta	241.695	6828.546	0.093	0.298	6.902
2SLS_1Rej	0.631	0.483	0.000	1.000	1.000
2SLS_F	1.019	1.437	0.015	0.479	2.690

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

	mean	sd	q1	q5	q9
OLS_Beta	0.990	0.010	0.977	0.990	1.003
OLS_seBeta	0.010	0.001	0.009	0.010	0.011
OLS_1Rej	1.000	0.000	1.000	1.000	1.000
2SLS_Beta	0.709	20.896	0.016	0.880	1.916
2SLS_seBeta	1662.438	70047.601	0.111	0.479	13.509
2SLS_1Rej	0.493	0.500	0.000	0.000	1.000
2SLS_F	1.287	1.804	0.020	0.594	3.474

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

	mean	sd	q1	q5	q9
OLS_Beta	0.705	0.034	0.662	0.705	0.748
OLS_seBeta	0.033	0.003	0.029	0.033	0.037
OLS_1Rej	1.000	0.000	1.000	1.000	1.000
2SLS_Beta	-0.014	0.119	-0.175	-0.000	0.126
2SLS_seBeta	0.116	0.031	0.082	0.110	0.157
2SLS_1Rej	0.056	0.231	0.000	0.000	0.000
2SLS_F	10.080	6.560	2.631	8.921	19.002

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

	mean	sd	q1	q5	q9
OLS_Beta	0.020	0.010	0.007	0.020	0.033
OLS_seBeta	0.010	0.001	0.009	0.010	0.011
OLS_1Rej	0.514	0.500	0.000	1.000	1.000
2SLS_Beta	-0.000	0.010	-0.013	-0.000	0.013
2SLS_seBeta	0.010	0.001	0.009	0.010	0.011
2SLS_1Rej	0.029	0.167	0.000	0.000	0.000
2SLS_F	100.601	24.993	70.511	98.559	133.274

Table 2: Weak Instrument Summary Statistics: STATA

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

	mean	sd	q1	q5	q9
OLS_Beta	0.9901	0.0102	0.9771	0.9902	1.0034
OLS_seBeta	0.0100	0.0010	0.0088	0.0099	0.0112
OLS_1Rej	1.0000	0.0000	1.0000	1.0000	1.0000
2SLS_Beta	0.8311	5.6715	0.5172	0.9834	1.4323
2SLS_seBeta	295.8764	1490.0000	0.0939	0.3116	6.6295
2SLS_1Rej	0.6178	0.4860	0.0000	1.0000	1.0000
2SLS_F	1.0095	1.4276	0.0142	0.4499	2.7896

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

	mean	sd	q1	q5	q9
OLS_Beta	0.9889	0.0105	0.9756	0.9889	1.0024
OLS_seBeta	0.0103	0.0010	0.0090	0.0102	0.0116
OLS_1Rej	1.0000	0.0000	1.0000	1.0000	1.0000
2SLS_Beta	1.5299	50.0462	-0.8195	0.6554	2.7305
2SLS_seBeta	5106.0382	20300.0000	0.1552	0.8783	26.1890
2SLS_1Rej	0.3158	0.4649	0.0000	0.0000	1.0000
2SLS_F	1.2523	1.7685	0.0171	0.5515	3.4064

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

	mean	sd	q1	q5	q9
OLS_Beta	0.9476	0.0176	0.9246	0.9478	0.9702
OLS_seBeta	0.0172	0.0017	0.0151	0.0171	0.0194
OLS_1Rej	1.0000	0.0000	1.0000	1.0000	1.0000
2SLS_Beta	-0.1619	3.4814	-0.7381	-0.0040	0.3013
2SLS_seBeta	4.6675	160.3499	0.1645	0.3361	1.0125
2SLS_1Rej	0.0884	0.2839	0.0000	0.0000	0.0000
2SLS_F	10.0068	6.4326	2.8591	8.7944	18.4916

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

	mean	sd	q1	q5	q9
OLS_Beta	0.6624	0.0346	0.6177	0.6623	0.7065
OLS_seBeta	0.0338	0.0034	0.0296	0.0337	0.0381
OLS_1Rej	1.0000	0.0000	1.0000	1.0000	1.0000
2SLS_Beta	-0.0113	0.1050	-0.1476	-0.0014	0.1139
2SLS_seBeta	0.1039	0.0235	0.0780	0.1007	0.1334
2SLS_1Rej	0.0498	0.2176	0.0000	0.0000	0.0000
2SLS_F	100.3726	24.9273	70.4628	94.1650	133.3342

3 Question 3: Weak Instruments - Empirical Study

3.1 Angrist and Krueger (1991)

I report my results from R in Table 3 and from STATA in Table 4. See the Appendices for the underlying code: Appendix A (R) and Appendix B (STATA).

We would expect the OLS estimate of the return to education to be biased upwards, since it is likely most OLS estimates will omit important unobserved variables (such as talent, ability, etc.) and those are positively correlated with education. Interestingly, here we see that the 2SLS estimates are close to the OLS estimates (even slightly higher than the OLS estimates, in the 2SLS 1 model).

The F statistics from the IV regressions are relatively small (4.747 and 1.613 for IV models 1 and 2, respectively), suggesting we may have an issue with weak instruments.

Table 3: Replication of Angrist and Krueger (1991), Table V: R

	OLS 1	2SLS 1	OLS 2	2SLS 2
	(5)	(6)	(7)	(8)
educ	0.0632	0.0806	0.0632	0.0600
	(0.0003)	(0.0164)	(0.0003)	(0.0290)
non_white	−0.2575	−0.2302	−0.2575	−0.2626
	(0.0040)	(0.0261)	(0.0040)	(0.0458)
SMSA	−0.1763	−0.1581	−0.1763	−0.1797
	(0.0029)	(0.0174)	(0.0029)	(0.0305)
married	0.2479	0.2440	0.2479	0.2486
	(0.0032)	(0.0049)	(0.0032)	(0.0073)
age_q			−0.0760	−0.0741
			(0.0604)	(0.0626)
age_sq			0.0008	0.0007
			(0.0007)	(0.0007)
9 Year-of-birth dummies	Yes	Yes	Yes	Yes
8 Region-of-residence dummies	Yes	Yes	Yes	Yes

3.2 Bound, Jaeger, and Baker (1995)

I report the results of my permutation analysis using both R and STATA in Table 5. See the underlying code in the appendices for details of implementation: Appendix A (R) and Appendix B (STATA).

Table 4: Replication of Angrist and Krueger (1991), Table V: STATA

VARIABLES	(5) OLS 1	(6) 2SLS 1	(7) OLS 2	(8) 2SLS 2
educ	0.0632	0.0806	0.0632	0.0600
	(0.0003)	(0.0164)	(0.0003)	(0.0290)
non_white	-0.2575	-0.2302	-0.2575	-0.2626
	(0.0040)	(0.0261)	(0.0040)	(0.0458)
SMSA	-0.1763	-0.1581	-0.1763	-0.1797
	(0.0029)	(0.0174)	(0.0029)	(0.0305)
married	0.2479	0.2440	0.2479	0.2486
	(0.0032)	(0.0049)	(0.0032)	(0.0073)
age_q			-0.0760	-0.0741
			(0.0604)	(0.0626)
age_sq			0.0008	0.0007
			(0.0007)	(0.0007)
Constant	5.0164	4.7850	6.8896	6.9095
	(0.0069)	(0.2192)	(1.3570)	(1.3684)
9 Year-of-birth dummies	Yes	Yes	Yes	Yes
8 Region-of-residence dummies	Yes	Yes	Yes	Yes

Note: Standard errors in parentheses

Table 5: Permutation Results: Summary of Coefficient on Education

Software	Model	Average	Standard Deviation
R	2SLS 1	0.0636229	0.0378478
R	2SLS 2	0.0636069	0.0378167
STATA	2SLS 1	0.0625177	0.0392369
STATA	2SLS 2	0.0625487	0.0392468

A R Code

```
##############################################################################
# Author: Paul R. Organ
# Purpose: ECON 675, PS5
# Last Update: Nov 19, 2018
##############################################################################
# Preliminaries
options(stringsAsFactors = F)

# packages
require(tidyverse) # data cleaning and manipulation
require(magrittr)  # syntax
require(ggplot2)   # plots
require(sandwich)  # robust standard errors
require(ivpack)    # IV regression
require(xtable)    # tables for LaTeX
require(stargazer) # tables for LaTeX
require(boot)      # bootstrapping

options(scipen = 999)
setwd('C:/Users/prorgan/Box/Classes/Econ 675/Problem Sets/PS5')

##############################################################################
# Question 2) Weak Instrument Simulations
##############################################################################

# sample size
n <- 200

# set of gammas we will test
gammas <- sqrt(c(0/n,0.25/n,9/n,99/n))

# define data-generating process
dgp <- function(n, gamma){
  z <- rnorm(n, 0, 1)
  u <- rnorm(n, 0, 1)
  v <- .99*u + sqrt(1-.99^2)*rnorm(n, 0, 1)

  x <- gamma * z + v
  y <- u # beta is 0

  out = data.frame(y = y, x = x, z = z)
  return(out)
}

# define function to estimate everything we want
bigFunction <- function(rep, gamma){
  df <- dgp(n, gamma)

  # ols regression
  ols <- lm(y ~ x, data = df)

  # ols coef and se
```

```
  beta_ols <- ols$coefficients['x']
  se_ols   <- sqrt(vcovHC(ols,'HC1')['x','x'])

  # ols rejection indicator (null beta=0)
  rej_ols <- 1 * (beta_ols/se_ols > 1.96)

  # iv regression
  iv <- ivreg(y ~ x | z, data = df)

  # iv coef and se
  beta_iv <- iv$coefficients['x']
  se_iv   <- sqrt(vcovHC(iv,'HC1')['x','x'])

  # iv rejection indicator (null beta=0)
  rej_iv <- 1 * (beta_iv/se_iv > 1.96)

  # iv F stat
  F_iv <- summary(iv, diagnostics=T)$diagnostics[1,3]

  # single row dataframe to report, which we can combine over replications
  out <- data.frame(rep = rep, gamma = gamma,
                    beta_ols=beta_ols, se_ols=se_ols, rej_ols=rej_ols,
                    beta_iv=beta_iv, se_iv=se_iv, rej_iv=rej_iv, F_iv=F_iv)
  return(out)
}

# run function M times for each gamma, save results
set.seed(22)

M    <- 5000
reps <- 1:M

ptm <- proc.time()
results1 <- lapply(reps, FUN = bigFunction, gamma=gammas[1]) %>% bind_rows
results2 <- lapply(reps, FUN = bigFunction, gamma=gammas[2]) %>% bind_rows
results3 <- lapply(reps, FUN = bigFunction, gamma=gammas[3]) %>% bind_rows
results4 <- lapply(reps, FUN = bigFunction, gamma=gammas[4]) %>% bind_rows
proc.time() - ptm
# runtime is 4 minutes

# summarize results of a single column
summarizeResult <- function(col){
  m  <- mean(col)
  sd <- sd(col)
  q1 <- quantile(col, 0.1)
  q5 <- quantile(col, 0.5)
  q9 <- quantile(col, 0.9)
  out = c(m, sd, q1, q5, q9)
  return(out)
}

# summarize results for a gamma simulation
summarizeResults <- function(data, gamma){
  df <- data
```

```
  # empty matrix to fill
  tab <- matrix(NA,7,5)

  tab[1,] <- summarizeResult(df$beta_ols)
  tab[2,] <- summarizeResult(df$se_ols)
  tab[3,] <- summarizeResult(df$rej_ols)
  tab[4,] <- summarizeResult(df$beta_iv)
  tab[5,] <- summarizeResult(df$se_iv)
  tab[6,] <- summarizeResult(df$rej_iv)
  tab[7,] <- summarizeResult(df$F_iv)

  # append gamma as a column to keep track
  out <- cbind(gamma, tab) %>% as.data.frame()
  names(out) <- c('gamma', 'mean', 'sd', 'q1', 'q5', 'q9')
  return(out)
}

# apply functions for each gamma
summ1 <- summarizeResults(results1, gammas[1])
summ2 <- summarizeResults(results2, gammas[2])
summ3 <- summarizeResults(results3, gammas[3])
summ4 <- summarizeResults(results4, gammas[4])

# output for LaTeX
summ1 %<>% select(-gamma)
rownames(summ1) <- c('OLS_Beta','OLS_seBeta','OLS_1Rej',
                     '2SLS_Beta','2SLS_seBeta','2SLS_1Rej','2SLS_F')
xtable(summ1, digits=c(0,3,3,3,3,3))

summ2 %<>% select(-gamma)
rownames(summ2) <- c('OLS_Beta','OLS_seBeta','OLS_1Rej',
                     '2SLS_Beta','2SLS_seBeta','2SLS_1Rej','2SLS_F')
xtable(summ2, digits=c(0,3,3,3,3,3))

summ3 %<>% select(-gamma)
rownames(summ3) <- c('OLS_Beta','OLS_seBeta','OLS_1Rej',
                     '2SLS_Beta','2SLS_seBeta','2SLS_1Rej','2SLS_F')
xtable(summ3, digits=c(0,3,3,3,3,3))

summ4 %<>% select(-gamma)
rownames(summ4) <- c('OLS_Beta','OLS_seBeta','OLS_1Rej',
                     '2SLS_Beta','2SLS_seBeta','2SLS_1Rej','2SLS_F')
xtable(summ4, digits=c(0,3,3,3,3,3))

##############################################################################
# Question 3: Weak Instrument - Empirical Study
##############################################################################

# clean up
rm(list = ls())
gc()

df <- read_csv('Angrist_Krueger.csv')
```

```
##############################################################################################
# Question 3.1) Angrist and Krueger
# OLS1
ols1 <- lm(l_w_wage ~ educ + non_white + SMSA + married +
             factor(region) + factor(YoB_ld), data = df)

# 2SLS1
iv1 <- ivreg(l_w_wage ~ educ + non_white + SMSA + married +
               factor(region) + factor(YoB_ld) |
               factor(QoB)*factor(YoB_ld) + non_white + SMSA + married +
               factor(region) + factor(YoB_ld), data = df)

# look at F statistic for weak instrument check
summary(iv1, diagnostics = T)

# OLS2
ols2 <- lm(l_w_wage ~ educ + non_white + SMSA + married + age_q + age_sq +
             factor(region) + factor(YoB_ld), data = df)

# 2SLS1
iv2 <- ivreg(l_w_wage ~ educ + non_white + SMSA + married + age_q + age_sq +
               factor(region) + factor(YoB_ld) |
               factor(QoB)*factor(YoB_ld) + non_white + SMSA + married +
               age_q + age_sq + factor(region) + factor(YoB_ld), data = df)

# look at F statistic for weak instrument check
summary(iv2, diagnostics = T)

# output for LaTeX
vars <- c('educ', 'non_white', 'SMSA', 'married', 'age_q', 'age_sq')

stargazer(ols1, iv1, ols2, iv2, keep = vars, # show all four regs
          dep.var.labels.include = F, # drop the l_w_wage heading
          model.names = F, # we replace them below
          column.labels = c('OLS 1', '2SLS 1', 'OLS 2', '2SLS 2'),
          add.lines = list(c('9 Year-of-birth dummies', rep('Yes', 4)),
                           c('8 Region-of-residence dummies', rep('Yes', 4))),
          omit.table.layout = 'sn', # get rid of end table stuff (N, R2, etc.)
          star.char = c('', '', ''), # remove asterisks
          digits = 4) # decimals

##############################################################################
# Question 3.2) Bound, Jaeger, and Baker

# first for the smaller set of covariates
boot_iv1 <- function(df, i){
  # permute only QoB
  df$QoB <- df$QoB[i]

  # iv reg (2SLS 1)
  reg <- ivreg(l_w_wage ~ educ + non_white + SMSA + married +
                 factor(region) + factor(YoB_ld) |
                 factor(QoB)*factor(YoB_ld) + non_white + SMSA + married +
```

```
                factor(region) + factor(YoB_ld), data = df)

  return(reg$coefficients['educ'])
}

# run bootstrap, 500 replications
ptm <- proc.time()
set.seed(22)
iv1_results <- boot(data = df, R = 500, statistic = boot_iv1)
proc.time() - ptm
# runtime is 28 minutes

iv1_avg <- mean(iv1_results$t)
iv1_sd  <- sd(iv1_results$t)

# second for the larger set of covariates
boot_iv2 <- function(df, i){
  # permute only QoB
  df$QoB <- df$QoB[i]

  # iv reg (2SLS 2)
  reg <- ivreg(l_w_wage ~ educ + non_white + SMSA + married + age_q + age_sq +
                 factor(region) + factor(YoB_ld) |
                 factor(QoB)*factor(YoB_ld) + non_white + SMSA + married +
                 age_q + age_sq + factor(region) + factor(YoB_ld), data = df)

  return(reg$coefficients['educ'])
}

# run bootstrap, 500 replications
ptm <- proc.time()
set.seed(22)
iv2_results <- boot(data = df, R = 500, statistic = boot_iv2)
proc.time() - ptm
# runtime is 31 minutes

iv2_avg <- mean(iv2_results$t)
iv2_sd  <- sd(iv2_results$t)

################################################################################
```

B Stata Code

```
*******************************************************************************
* Author: Paul R. Organ
* Purpose: ECON 675, PS5
* Last Update: Nov 19, 2018
*******************************************************************************
clear all
set more off
capture log close

cd "C:\Users\prorgan\Box\Classes\Econ 675\Problem Sets\PS5"
log using ps5.log, replace

*******************************************************************************
*** Question 2: Weak Instrument Simulations
*******************************************************************************

* from Yingjie
program define weak_IV, rclass
    syntax [, obs(integer 200) f_stat(real 10) ]
        drop _all

        set obs `obs'

        * DGP
        gen u = rnormal()
        gen v = 0.99 * u + sqrt(1-0.99^2) * rnormal()
        gen z = rnormal()

        local gamma_0 = sqrt((`f_stat' - 1) / `obs')
        gen x = `gamma_0' * z + v
        gen y = u

        * OLS
        qui reg y x, robust
        return scalar OLS_b   = _b[x]
        return scalar OLS_se  = _se[x]
        return scalar OLS_rej = abs(_b[x]/_se[x]) > 1.96

        * 2SLS
        qui ivregress 2sls y (x = z)
        return scalar TSLS_b   = _b[x]
        return scalar TSLS_se  = _se[x]
        return scalar TSLS_rej = abs(_b[x]/_se[x]) > 1.96
        qui reg x z
        return scalar TSLS_F   = e(F)
end

* simulation 1: F = 1
simulate OLS_b=r(OLS_b) OLS_se=r(OLS_se) OLS_rej=r(OLS_rej) ///
    TSLS_b=r(TSLS_b) TSLS_se=r(TSLS_se) TSLS_rej=r(TSLS_rej) TSLS_F=r(TSLS_F), ///
    reps(5000) seed(22) nodots: ///
    weak_IV, f_stat(1)
```

```
local k = 1
matrix Results = J(7, 5, .)

qui sum OLS_b, detail
matrix Results[`k',1] = r(mean)
matrix Results[`k',2] = r(sd)
matrix Results[`k',3] = r(p10)
matrix Results[`k',4] = r(p50)
matrix Results[`k',5] = r(p90)
local k = `k' + 1

qui sum OLS_se, detail
matrix Results[`k',1] = r(mean)
matrix Results[`k',2] = r(sd)
matrix Results[`k',3] = r(p10)
matrix Results[`k',4] = r(p50)
matrix Results[`k',5] = r(p90)
local k = `k' + 1

qui sum OLS_rej, detail
matrix Results[`k',1] = r(mean)
matrix Results[`k',2] = r(sd)
matrix Results[`k',3] = r(p10)
matrix Results[`k',4] = r(p50)
matrix Results[`k',5] = r(p90)
local k = `k' + 1

qui sum TSLS_b, detail
matrix Results[`k',1] = r(mean)
matrix Results[`k',2] = r(sd)
matrix Results[`k',3] = r(p10)
matrix Results[`k',4] = r(p50)
matrix Results[`k',5] = r(p90)
local k = `k' + 1

qui sum TSLS_se, detail
matrix Results[`k',1] = r(mean)
matrix Results[`k',2] = r(sd)
matrix Results[`k',3] = r(p10)
matrix Results[`k',4] = r(p50)
matrix Results[`k',5] = r(p90)
local k = `k' + 1

qui sum TSLS_rej, detail
matrix Results[`k',1] = r(mean)
matrix Results[`k',2] = r(sd)
matrix Results[`k',3] = r(p10)
matrix Results[`k',4] = r(p50)
matrix Results[`k',5] = r(p90)
local k = `k' + 1

qui sum TSLS_F, detail
matrix Results[`k',1] = r(mean)
```

```
matrix Results[`k',2] = r(sd)
matrix Results[`k',3] = r(p10)
matrix Results[`k',4] = r(p50)
matrix Results[`k',5] = r(p90)
local k = `k' + 1

mat2txt, matrix(Results) saving(q2_1.txt) format(%9.4f) replace

* now run for F=1.25, 10, and 100

* simulation 2: F = 1.25
simulate OLS_b=r(OLS_b) OLS_se=r(OLS_se) OLS_rej=r(OLS_rej) ///
    TSLS_b=r(TSLS_b) TSLS_se=r(TSLS_se) TSLS_rej=r(TSLS_rej) TSLS_F=r(TSLS_F), ///
    reps(5000) seed(22) nodots: ///
    weak_IV, f_stat(1.25)

local k = 1
matrix Results = J(7, 5, .)

qui sum OLS_b, detail
matrix Results[`k',1] = r(mean)
matrix Results[`k',2] = r(sd)
matrix Results[`k',3] = r(p10)
matrix Results[`k',4] = r(p50)
matrix Results[`k',5] = r(p90)
local k = `k' + 1

qui sum OLS_se, detail
matrix Results[`k',1] = r(mean)
matrix Results[`k',2] = r(sd)
matrix Results[`k',3] = r(p10)
matrix Results[`k',4] = r(p50)
matrix Results[`k',5] = r(p90)
local k = `k' + 1

qui sum OLS_rej, detail
matrix Results[`k',1] = r(mean)
matrix Results[`k',2] = r(sd)
matrix Results[`k',3] = r(p10)
matrix Results[`k',4] = r(p50)
matrix Results[`k',5] = r(p90)
local k = `k' + 1

qui sum TSLS_b, detail
matrix Results[`k',1] = r(mean)
matrix Results[`k',2] = r(sd)
matrix Results[`k',3] = r(p10)
matrix Results[`k',4] = r(p50)
matrix Results[`k',5] = r(p90)
local k = `k' + 1

qui sum TSLS_se, detail
matrix Results[`k',1] = r(mean)
matrix Results[`k',2] = r(sd)
```

```
matrix Results[`k',3] = r(p10)
matrix Results[`k',4] = r(p50)
matrix Results[`k',5] = r(p90)
local k = `k' + 1

qui sum TSLS_rej, detail
matrix Results[`k',1] = r(mean)
matrix Results[`k',2] = r(sd)
matrix Results[`k',3] = r(p10)
matrix Results[`k',4] = r(p50)
matrix Results[`k',5] = r(p90)
local k = `k' + 1

qui sum TSLS_F, detail
matrix Results[`k',1] = r(mean)
matrix Results[`k',2] = r(sd)
matrix Results[`k',3] = r(p10)
matrix Results[`k',4] = r(p50)
matrix Results[`k',5] = r(p90)
local k = `k' + 1

mat2txt, matrix(Results) saving(q2_2.txt) format(%9.4f) replace

* simulation 3: F = 10
simulate OLS_b=r(OLS_b) OLS_se=r(OLS_se) OLS_rej=r(OLS_rej) ///
    TSLS_b=r(TSLS_b) TSLS_se=r(TSLS_se) TSLS_rej=r(TSLS_rej) TSLS_F=r(TSLS_F), ///
    reps(5000) seed(22) nodots: ///
    weak_IV, f_stat(10)

local k = 1
matrix Results = J(7, 5, .)

qui sum OLS_b, detail
matrix Results[`k',1] = r(mean)
matrix Results[`k',2] = r(sd)
matrix Results[`k',3] = r(p10)
matrix Results[`k',4] = r(p50)
matrix Results[`k',5] = r(p90)
local k = `k' + 1

qui sum OLS_se, detail
matrix Results[`k',1] = r(mean)
matrix Results[`k',2] = r(sd)
matrix Results[`k',3] = r(p10)
matrix Results[`k',4] = r(p50)
matrix Results[`k',5] = r(p90)
local k = `k' + 1

qui sum OLS_rej, detail
matrix Results[`k',1] = r(mean)
matrix Results[`k',2] = r(sd)
matrix Results[`k',3] = r(p10)
matrix Results[`k',4] = r(p50)
matrix Results[`k',5] = r(p90)
```

```
local k = `k' + 1

qui sum TSLS_b, detail
matrix Results[`k',1] = r(mean)
matrix Results[`k',2] = r(sd)
matrix Results[`k',3] = r(p10)
matrix Results[`k',4] = r(p50)
matrix Results[`k',5] = r(p90)
local k = `k' + 1

qui sum TSLS_se, detail
matrix Results[`k',1] = r(mean)
matrix Results[`k',2] = r(sd)
matrix Results[`k',3] = r(p10)
matrix Results[`k',4] = r(p50)
matrix Results[`k',5] = r(p90)
local k = `k' + 1

qui sum TSLS_rej, detail
matrix Results[`k',1] = r(mean)
matrix Results[`k',2] = r(sd)
matrix Results[`k',3] = r(p10)
matrix Results[`k',4] = r(p50)
matrix Results[`k',5] = r(p90)
local k = `k' + 1

qui sum TSLS_F, detail
matrix Results[`k',1] = r(mean)
matrix Results[`k',2] = r(sd)
matrix Results[`k',3] = r(p10)
matrix Results[`k',4] = r(p50)
matrix Results[`k',5] = r(p90)
local k = `k' + 1

mat2txt, matrix(Results) saving(q2_3.txt) format(%9.4f) replace

* simulation 4: F = 100
simulate OLS_b=r(OLS_b) OLS_se=r(OLS_se) OLS_rej=r(OLS_rej) ///
    TSLS_b=r(TSLS_b) TSLS_se=r(TSLS_se) TSLS_rej=r(TSLS_rej) TSLS_F=r(TSLS_F), ///
    reps(5000) seed(22) nodots: ///
    weak_IV, f_stat(100)

local k = 1
matrix Results = J(7, 5, .)

qui sum OLS_b, detail
matrix Results[`k',1] = r(mean)
matrix Results[`k',2] = r(sd)
matrix Results[`k',3] = r(p10)
matrix Results[`k',4] = r(p50)
matrix Results[`k',5] = r(p90)
local k = `k' + 1

qui sum OLS_se, detail
```

```
matrix Results[`k',1] = r(mean)
matrix Results[`k',2] = r(sd)
matrix Results[`k',3] = r(p10)
matrix Results[`k',4] = r(p50)
matrix Results[`k',5] = r(p90)
local k = `k' + 1

qui sum OLS_rej, detail
matrix Results[`k',1] = r(mean)
matrix Results[`k',2] = r(sd)
matrix Results[`k',3] = r(p10)
matrix Results[`k',4] = r(p50)
matrix Results[`k',5] = r(p90)
local k = `k' + 1

qui sum TSLS_b, detail
matrix Results[`k',1] = r(mean)
matrix Results[`k',2] = r(sd)
matrix Results[`k',3] = r(p10)
matrix Results[`k',4] = r(p50)
matrix Results[`k',5] = r(p90)
local k = `k' + 1

qui sum TSLS_se, detail
matrix Results[`k',1] = r(mean)
matrix Results[`k',2] = r(sd)
matrix Results[`k',3] = r(p10)
matrix Results[`k',4] = r(p50)
matrix Results[`k',5] = r(p90)
local k = `k' + 1

qui sum TSLS_rej, detail
matrix Results[`k',1] = r(mean)
matrix Results[`k',2] = r(sd)
matrix Results[`k',3] = r(p10)
matrix Results[`k',4] = r(p50)
matrix Results[`k',5] = r(p90)
local k = `k' + 1

qui sum TSLS_F, detail
matrix Results[`k',1] = r(mean)
matrix Results[`k',2] = r(sd)
matrix Results[`k',3] = r(p10)
matrix Results[`k',4] = r(p50)
matrix Results[`k',5] = r(p90)
local k = `k' + 1

mat2txt, matrix(Results) saving(q2_4.txt) format(%9.4f) replace

*******************************************************************************
*** Question 3: Weak Instrument - Empirical Study
*******************************************************************************
* Q3 setup
clear all
```

```
use Angrist_Krueger

*******************************************************************************
* Q3.1: Angrist and Krueger (1991)

* variables for use in regressions
local ols_short = "educ non_white SMSA married"
local iv_short  = "non_white SMSA married"
local ols_long  = "educ non_white SMSA married age_q age_sq"
local iv_long   = "non_white SMSA married age_q age_sq"

* note going out of order from PS to match with A&K(1991)

* OLS 1 (these SEs match the table, so I don't think they use robust SEs)
reg l_w_wage `ols_short' i.region i.YoB_ld

* output for LaTeX
outreg2 using q3_1.tex, stats(coef se) keep(`ols_short') noaster dec(4) replace ///
        addtext(9 Year-of-birth dummies, Yes, 8 Region-of-residence dummies, Yes) ///
        ctitle(OLS 1) nor2 noobs

* 2SLS 1
ivregress2 2sls l_w_wage `iv_short' i.region i.YoB_ld ///
        (educ = i.YoB_ld##i.QoB)

* output for LaTeX
outreg2 using q3_1.tex, stats(coef se) keep(`ols_short') noaster dec(4) append ///
        addtext(9 Year-of-birth dummies, Yes, 8 Region-of-residence dummies, Yes) ///
        ctitle(2SLS 1) nor2 noobs

* OLS 2
reg l_w_wage `ols_long' i.region i.YoB_ld

* output for LaTeX
outreg2 using q3_1.tex, stats(coef se) keep(`ols_long') noaster dec(4) append ///
        addtext(9 Year-of-birth dummies, Yes, 8 Region-of-residence dummies, Yes) ///
        ctitle(OLS 2) nor2 noobs

* 2SLS 2
ivregress2 2sls l_w_wage `iv_long' i.region i.YoB_ld ///
    (educ = i.YoB_ld##i.QoB)

* output for LaTeX
outreg2 using q3_1.tex, stats(coef se) keep(`ols_long') noaster dec(4) append ///
        addtext(9 Year-of-birth dummies, Yes, 8 Region-of-residence dummies, Yes) ///
        ctitle(2SLS 2) nor2 noobs

*******************************************************************************
* Q3.2: Bound, Jaeger, and Baker (1995)

* program defined in the problem set appendix for quicker IV estimation
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

* permute QoB 500 times for each model, save results
permute QoB TSLS_1_b = r(beta), reps(500) seed(22) saving(q3_model1, replace): ///
    IV_quick QoB, model(1)

permute QoB TSLS_2_b = r(beta), reps(500) seed(22) saving(q3_model2, replace): ///
    IV_quick QoB, model(2)

* summarize results necessary for table in LaTeX
clear all
use q3_model1
sum TSLS_1_b

clear all
use q3_model2
sum TSLS_2_b

*******************************************************************************
log close
*******************************************************************************
```