POLS 207

Problem Set 5*

Villaseñor-Derbez, J.C.

Problem 1: 2SLS

a) Show results from (1) the first stage regression, (2) the reduced form regression, and (3) the 2SLS estimation using the following two specifications:

$$log(PGP_i^{1995}) = \beta_0 + \beta_1 avexpr_i + \epsilon_i$$

$$avexpr_i = \gamma_0 + \gamma_1 logem4 + \mu_i$$

And

$$log(PGP_i^{1995}) = \beta_0 + \beta_1 avexpr_i + \beta_2 latabst + \epsilon_i$$

$$avexpr_i = \gamma_0 + \gamma_1 logem4 + \gamma_2 latabst + \mu_i$$

```
# Load packages
suppressPackageStartupMessages({
  library(here)
  library(stargazer)
  library(foreign)
  library(AER)
  library(tidyverse)
})
```

```
# Load the data
arj <- read.dta(file = here("ps5", "arj.dta"))
```

```
# First stage
first_simple <- lm(avexpr ~ logem4, data = arj)
first_lat <- lm(avexpr ~ logem4 + lat_abst, data = arj)

# Reduced form
reduced_simple <- lm(logpgp95 ~ logem4, data = arj)
reduced_lat <- lm(logpgp95 ~ logem4 + lat_abst, data = arj)

# Two-stage using IVreg
two_stage_simple <- ivreg(logpgp95 ~ avexpr | logem4, data = arj)
two_stage_lat <- ivreg(logpgp95 ~ avexpr + lat_abst | logem4 + lat_abst, data = arj)
```

```
stargazer(list(first_simple, first_lat),
          single.row = T,
          header = F,
          title = "Coefficients for first-stage regression.")
```

*Available on GitHub: https://github.com/jcvdav/POLS207/blob/master/ps5/

Table 1: Coefficients for first-stage regression.

	Dependent variable:	
	avexpr	
	(1)	(2)
logem4	−0.607*** (0.127)	−0.510*** (0.141)
lat_abst		2.002 (1.337)
Constant	9.341*** (0.611)	8.529*** (0.812)
Observations	64	64
R^2	0.270	0.296
Adjusted R^2	0.258	0.273
Residual Std. Error	1.265 (df = 62)	1.252 (df = 61)
F Statistic	22.947*** (df = 1; 62)	12.824*** (df = 2; 61)
Note:		*p<0.1; **p<0.05; ***p<0.01

```
stargazer(list(reduced_simple, reduced_lat),
          single.row = T,
          header = F,
          title = "Coefficients for reduced form regression.")
```

Table 2: Coefficients for reduced form regression.

	Dependent variable:	
	logpgp95	
	(1)	(2)
logem4	−0.573*** (0.076)	−0.508*** (0.084)
lat_abst		1.346* (0.800)
Constant	10.731*** (0.367)	10.185*** (0.486)
Observations	64	64
R^2	0.477	0.500
Adjusted R^2	0.469	0.484
Residual Std. Error	0.760 (df = 62)	0.749 (df = 61)
F Statistic	56.603*** (df = 1; 62)	30.551*** (df = 2; 61)
Note:		*p<0.1; **p<0.05; ***p<0.01

```
stargazer(list(two_stage_simple, two_stage_lat),
          single.row = T,
          header = F,
          title = "Coefficients for 2SLS.")
```

Table 3: Coefficients for 2SLS.

	Dependent variable:	
	logpgp95	
	(1)	(2)
avexpr	0.944^{***} (0.157)	0.996^{***} (0.222)
lat_abst		-0.647 (1.335)
Constant	1.910^{*} (1.027)	1.692 (1.293)
Observations	64	64
R^2	0.187	0.102
Adjusted R^2	0.174	0.073
Residual Std. Error	0.948 (df = 62)	1.005 (df = 61)
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 4: Two-stage least squares regression table

		(1) no covariates	(2) including latitude
First stage (dep: `avexpr`):	`logem4`	-0.6067782	-0.5102681
	`lat_abst`		2.0017746
Reduced form (dep: `logpgp95`):	`logem4`	-0.5729682	-0.508076
	`lat_abst`		1.3459679
2SLS (dep: `logpgp95`):	`avexpr`	0.9442794	0.995704
	`lat_abst`		-0.6558067

Regress `logpgp95`, `avexpr`, and `logem4` on `lat_abst` ("partialling out" the effect of latitude) and re-do the 2SLS estimation using the residuals. Do you get the same result as in Column 2 in the previous question? (Don't worry about the standard errors – actually they are quite close to the right ones.)

The regression table below shows the OLS estimates for each stage, as well as the IV regression. Using the residuals after removing the effect of latitude, we obtain the same coefficients as the second column in the previous question.

```
res_logpgp95 <- lm(logpgp95 ~ lat_abst, data = arj)$residuals
res_avexpr <- lm(avexpr ~ lat_abst, data = arj)$residuals
res_logem4 <- lm(logem4 ~ lat_abst, data = arj)$residuals
```

```
# First stage
res_first_simple <- lm(res_avexpr ~ res_logem4)

# Reduced form
res_reduced_simple <- lm(res_logpgp95 ~ res_logem4)

# Two-stage using IVreg
res_two_stage_simple <- ivreg(res_logpgp95 ~ res_avexpr | res_logem4)
```

```
stargazer(... = list(res_first_simple, res_reduced_simple, res_two_stage_simple),
          single.row = T,
          header = F,
          title = "Two-stage regression on the residuals of each variable on latabst.")
```

Table 5: Two-stage regression on the residuals of each variable on latabst.

	Dependent variable:		
	res_avexpr	res_logpgp95	
	OLS	*OLS*	*instrumental variable*
	(1)	(2)	(3)
res_logem4	−0.510*** (0.140)	−0.508*** (0.084)	
res_avexpr			0.996*** (0.220)
Constant	−0.000 (0.155)	0.000 (0.093)	0.000 (0.125)
Observations	64	64	64
R^2	0.177	0.373	−0.127
Adjusted R^2	0.163	0.363	−0.145
Residual Std. Error (df = 62)	1.242	0.743	0.996
F Statistic (df = 1; 62)	13.308***	36.838***	
Note:			*p<0.1; **p<0.05; ***p<0.01

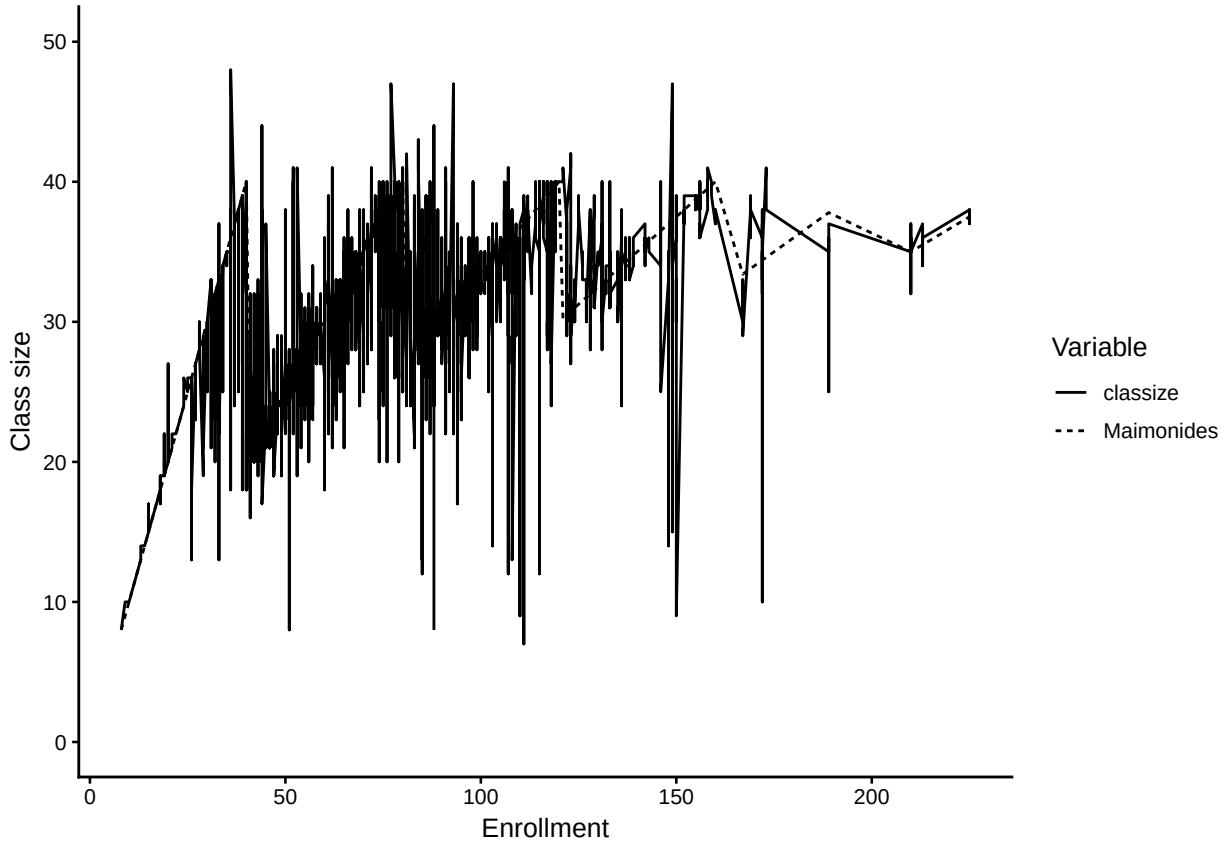
Problem 2: Fuzzy IV

```
# Load data
angrist_lavy <- read.dta(here("ps5", "angrist_lavy.dta"))
```

a) Say you were to run a regression of reading scores on class sizes. Would this provide a valid estimate of the causal effect of class size? Why or why not?

b) Use the data and plot the actual average class size (solid line) and the class size implied by Maimonides rule (dashed line) against enrollment count. That is, replicate Figure 1 of the paper for the fourth grade. What do the results imply about the determinants of class size? (you may find the `floor()` function useful).

```
angrist_lavy %>%
  mutate(Maimonides = angrist_lavy$enrollment / (floor((angrist_lavy$enrollment - 1) / 40) + 1)) %>%
  select(enrollment, classize, Maimonides) %>%
  gather(variable, value, -enrollment) %>%
  ggplot(aes(x = enrollment, y = value, linetype = variable)) +
  geom_line() +
  startR::ggtheme_plot() +
  lims(y = c(0, 50)) +
  labs(x = "Enrollment", y = "Class size") +
  guides(linetype = guide_legend(title = "Variable"))
```

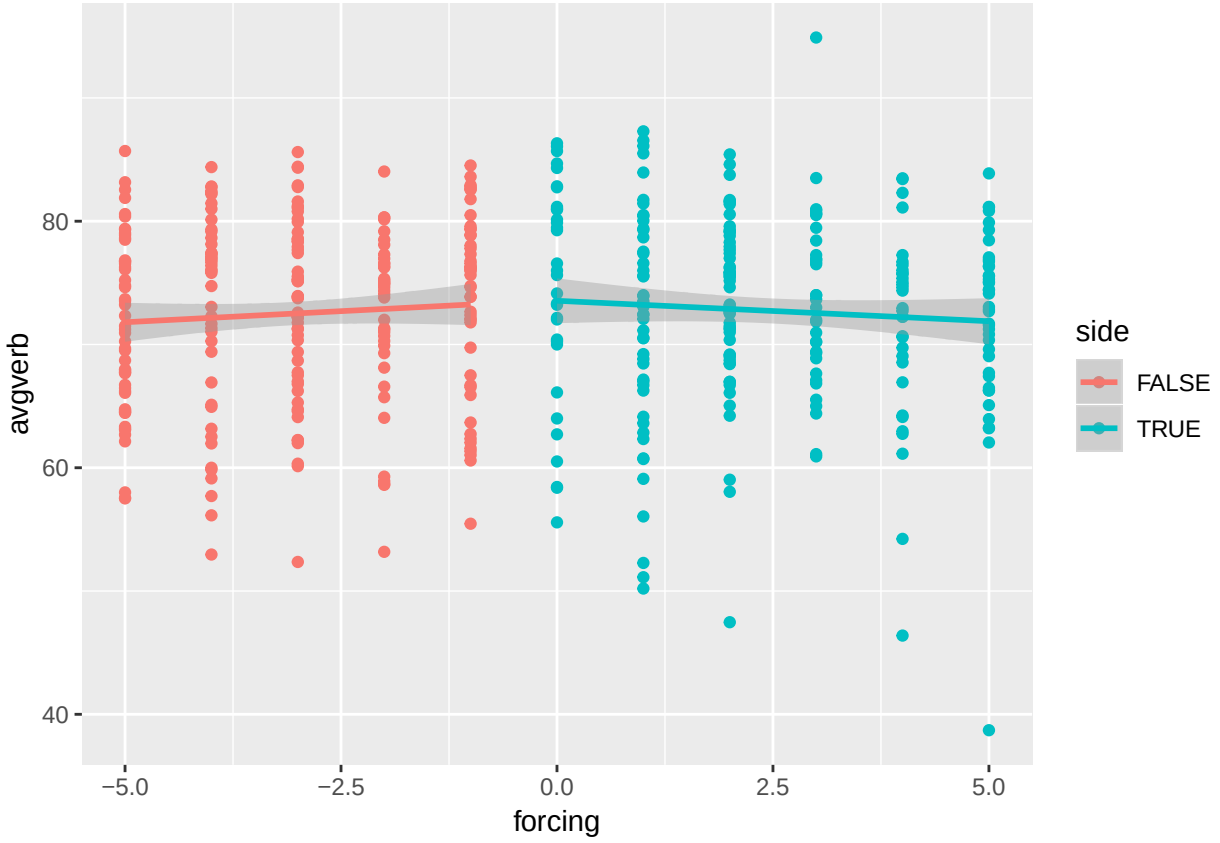



```
# Define discontinuity points to keep
disc_vec <- c(36:46, 76:86, 116:126, 156:166, 196:206)

# Define forcing variable
angrist_lavy_disc <- angrist_lavy %>%
```

```
  filter(enrollment %in% disc_vec) %>%
  rowwise() %>%
  mutate(disc_point = case_when(between(enrollment, 36, 46) ~ 41,
                                between(enrollment, 76, 86) ~ 81,
                                between(enrollment, 116, 126) ~ 121,
                                between(enrollment, 156, 166) ~ 161,
                                T ~ 201),
         forcing = enrollment - disc_point,
         side = forcing >= 0)
```

```
ggplot(data = angrist_lavy_disc,
       mapping = aes(x = forcing, y = avgverb, color = side)) +
  geom_point() +
  geom_smooth(method = "lm")
```




```
# Effect of discontinuities on class size
class_model <- lm(classize ~ forcing, data = angrist_lavy_disc)

# Effect of disc on reading comprehension scores
score_model <- lm(avgverb ~ forcing + side, data = angrist_lavy_disc)
```

```
stargazer(list(class_model, score_model),
          single.row = T,
          header = F,
          title = "Effect of the discontinuities on class size and average reading comprehension scores
```

Table 6: Effect of the discontinuities on class size and average reading comprehension scores.

	Dependent variable:	
	classize	avgverb
	(1)	(2)
forcing	-1.101^{***} (0.092)	-0.038 (0.228)
side		0.449 (1.446)
Constant	30.920^{***} (0.292)	72.377^{***} (0.862)
Observations	482	482
R^2	0.231	0.0003
Adjusted R^2	0.229	-0.004
Residual Std. Error	6.381 (df = 480)	7.726 (df = 479)
F Statistic	143.826^{***} (df = 1; 480)	0.071 (df = 2; 479)
Note:		*p<0.1; **p<0.05; ***p<0.01

Probelm 3: Bootstrapping

```
set.seed(43)

vector1 <- rnorm(500, mean = 7, sd = 3)

vector2 <- rnorm(500, mean = 5, sd = 2)
```

```
boot <- function(vec1, vec2, rep) {
  n <- length(vec1)
  mean_vec <- numeric(length = n)

  for (i in 1:rep) {

    vec1_i <- sample(x = vec1, size = n, replace = T)
    vec2_i <- sample(x = vec2, size = n, replace = T)

    mean_i <- mean(vec1_i) - mean(vec2_i)

    mean_vec[i] <- mean_i
  }

  return(mean_vec)
}


a <- boot(vector1, vector2, 10000)
```