41903 Pset 2

Andrew McKinley, Lauren Mostrom, Pietro Ramella, Francisco Ruela, and Bohan Yang

April 21, 2022

Question 1

Carefully evaluate the following comment. "I'm interested in estimating a model for the quantity of demand for widgets as a function of price, and I have some variables that tell me about the costs of producing widgets that I know are unrelated to demand shocks. I've estimated my model using OLS and 2SLS and have found that OLS produces better within sample forecasts and that the OLS forecasts are also better within a holdout sample I set aside before estimating the model. I also know that if a model doesn't forecast well, it's not a good model; so I'm going to use the OLS estimates to gauge the possible effects on quantity sold of implementing a 10% *increase in price."* First and foremost we know that trying to run traditional OLS models to estimate demand run into simultaneity. In fact this is quite literally the motivating example used throughout Chapter 3 of Hayashi. If we describe the economy as a system of equations we immediately notice that we have two *simultaneous* equations as a function of price (noting that in market equilibrium $q_i^d = q_i^s = q_i$):

$$q_i = \alpha_0 + \alpha_1 p_1 + u_i \quad \text{(demand)}$$
$$q_i = \beta_0 + \beta_1 p_1 + v_i \quad \text{(supply)}$$

If you are interested in q_i or p_i and solve the above equations to isolate the variable, you quickly see that they are functions of both errors and in OLS we will have a biased estimate.

The motivation for 2SLS is that we have a consistent estimator, even in cases of endogeniety. With a valid, strong instrument, as suggested by the statement, 2SLS will be the more "correct" model. If the instrument is weak, then we can make an argument that in sum cases OLS is the more precise estimate (though this still concedes a falsehood in the statement)

Finally, forecasting isn't the end all be all of good models. At the end of the day if a variable is random and has high variance a perfect the TRUTH will be unforecastable, so even if the model is the "truth" we will have bad forecastability, but a complete understanding of the underlying mechanisms of the event/phenomenon we are interested.

Question 2

(a)

We consider the following model:

$$\log(bwght) = \beta_0 + \beta_1 male + \beta_2 parity + \beta_3 \log(faminc) + \beta_4 packs + u$$

We might expect *packs* to be correlated with u for a number of reasons. One example could be age cohort effects: older mothers are more likely to give birth to underweight babies, and may have grown into adulthood at a time when the dangers of smoking were not as well understood or communicated in schools. Another reason may be underlying health conditions of the mother, such as stress and anxiety, eating disorders, and conditions that cause chronic pain that could induce the mother to smoke more as a self-medicating mechanism and would reduce the birthweight of the baby. This could result in omitted variable bias, where β_4 would overstate the effect of smoking on baby birthweight because it would pick up the effects of the underlying health condition(s) correlated both with birthweight and smoking.

(b)

For cigarette prices to be a suitable instrument we need for it to be significantly correlated with *packs* (relevance) but not with u (exclusion). We would expect the relevance condition to hold, but it may be weak. From basic laws of demand we would expect high cigarette prices to reduce cigarette consumption, but because of the addictive qualities of nicotine we may also worry that demand for cigarettes is relatively inelastic. The exclusion restriction is also somewhat dubious because state and local policies that raise the prices of cigarettes may also be correlated with the quality of healthcare access in those states, which could be correlated with women's consumption of cigarettes to self-treat underlying health conditions, as discussed in (a).

(c)

```
# Load data BWGHT.raw into R
#setwd("C:/Users/17036/OneDrive/Documents/GitHub/metrics3-zombie-boards/Psets/2/PS2Data")
BWGHT <- read.delim("PS2Data/BWGHT_RAW.txt", header=FALSE)
colnames(BWGHT) <- c("faminc","cigtax","cigprice","bwght","fatheduc","motheduc","parity",
                     "male","white","cigs","lbwght","bwghtlbs","packs","lfaminc")


# select variables
lbwght <- BWGHT$lbwght # dependent variable: log(bwght)
male <- BWGHT$male # independent variable: male
parity <- BWGHT$parity # independent variable: parity
lfaminc <- BWGHT$lfaminc # independent variable: log(faminc)
packs <- BWGHT$packs # independent variable: packs

# (i) OLS
reg_ols <- lm(lbwght ~ male + parity + lfaminc + packs)
coef_ols <- coef(reg_ols)
HCV.coef_ols <- vcovHC(reg_ols, type = 'HC')
coef_ols.se <- sqrt(diag(HCV.coef_ols)) # White standard errors
output_ols <- cbind(coef_ols,coef_ols.se)

# (ii) 2SLS
cigprice <- BWGHT$cigprice # IV: cigprice
# Note that 2SLS should be estimated in one step to get correct standard errors
reg_2sls <- ivreg(lbwght ~ packs+male+parity+lfaminc|cigprice+male+parity+lfaminc)
```

```
coef_2sls <- coef(reg_2sls)
HCV.coef_2sls <- vcovHC(reg_2sls, type = 'HC')
coef_2sls.se <- sqrt(diag(HCV.coef_2sls)) # White standard errors
output_2sls <- cbind(coef_2sls,coef_2sls.se)
```

The results shown in Table 1 show that the OLS estimate for the coefficient on *packs* is negative and statistically significant, which is consistent with what we expect about smoking reducing babies' birthweights. However the 2SLS estimate is very imprecise (SE=1.0863), and the point estimate is large (0.7971) and positive. The interpretation of this would be that an additional pack of cigarettes consumed by the mother *increases* the birthweight of her baby by 80%, which simply cannot be true. The discrepancy between these results suggests we should reconsider whether *cigprice* is a valid instrument for *packs*.

(d)

```
# First stage: regress X (endogenous variable) on Z (instrumental variable)
stage1 <- lm(packs ~ male + parity + lfaminc + cigprice)
# predetermined regressors should be included in the list of instruments
coef_stage1 <- coef(stage1)
HCV.coef_stage1 <- vcovHC(stage1, type = 'HC')
coef_stage1.se <- sqrt(diag(HCV.coef_stage1)) # White standard errors
output_stage1 <- cbind(coef_stage1,coef_stage1.se)
```

We estimate the reduced form for *packs* as follows:

$$\hat{packs} = \gamma_0 + \gamma_1 male + \gamma_2 parity + \gamma_3 \log(faminc) + \gamma_4 cigprice$$

As shown in Table 2, the coefficient on *cigprice* is very small and not at all significant. For *cigprice* to be a valid instrument for *packs*, even if we were able to tell a good story about why *cigprice* should be uncorrelated with u (which is of course untestable), we would still need to be confident that the matrix ZX' is very far from zero. However since the estimate for γ_4 is very close to zero and nowhere near being statistically significant even at the 5% level, *cigprice* does not satisfy the relevance condition and cannot be taken as a valid instrument.

The results in Table 2 help explain why in part (c) the 2SLS results were so far off from the OLS results. This is because the instrument and covariates are very poor predictors of *packs*, so it makes sense that the point estimate on packs from 2SLS was so imprecise that it was basically indistinguishable from zero (despite the point estimate itself being large and positive). This is also a good example of a time when a point estimate is much less informative (and, in fact, misleading), and reporting a confidence interval would be much more useful to the reader.

```
# OLS vs 2SLS  Table
texreg(list(reg_ols, reg_2sls), digits=4, caption.above=TRUE)
```

```
# Reduced Form Table
texreg(stage1, digits=4, caption.above=TRUE)
```

Table 1: Statistical models

	Model 1	Model 2
(Intercept)	4.6756***	4.4679***
	(0.0219)	(0.2588)
male	0.0262**	0.0298
	(0.0101)	(0.0178)
parity	0.0147**	−0.0012
	(0.0057)	(0.0219)
lfaminc	0.0180**	0.0636
	(0.0056)	(0.0570)
packs	−0.0837***	0.7971
	(0.0171)	(1.0863)
R^2	0.0350	−1.8118
Adj. R^2	0.0322	−1.8199
Num. obs.	1388	1388

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 2: Statistical models

	Model 1
(Intercept)	0.1374
	(0.1040)
male	−0.0047
	(0.0159)
parity	0.0181*
	(0.0089)
lfaminc	−0.0526***
	(0.0087)
cigprice	0.0008
	(0.0008)
R^2	0.0305
Adj. R^2	0.0276
Num. obs.	1388

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Question 3

```
# Prepare the data
df <- read.table("PS2Data/CARD.raw", quote="\"", comment.char="")

name_string <- "id   nearc2    nearc4    educ      age       fatheduc  motheduc
weight   momdad14  sinmom14  step14    reg661    reg662    reg663    reg664
reg665 reg666    reg667    reg668    reg669    south66   black     smsa   south
smsa66    wage      enroll    KWW       IQ        married   libcrd14  exper
lwage     expersq   "

name_string <- gsub("[\r\n]", " ", name_string)
name_string <- strsplit(name_string, " ")

name_vec <- vector()
for (i in name_string[[1]]) {
  if (i != "1" & i != ""){
    name_vec <- append(name_vec, i)
  }
}
colnames(df) <- name_vec
```

(a)

In the table below, the *iid* and *robust* standard errors are quite similar (*robust* standard errors are usually a little larger), so the inference (ttest and pvalue) is also quite similar.

The *iid* standard error assumes that the variance of the error term is constant and does not depend on independent variables. However, the *robust* standard error does not assume this, so it can work under both homoskedasticity and heteroskedasticity.

```
# Homoskedastic
homo <- feols(lwage ~ educ + exper + expersq + black + south + smsa + smsa66 +
              reg661 + reg662 + reg663 + reg664 + reg665 + reg666 + reg667 + reg668,
              se = "iid", df)

# Heteroskedastic
hetero <- feols(lwage ~ educ + exper + expersq + black + south + smsa + smsa66 +
                reg661 + reg662 + reg663 + reg664 + reg665 + reg666 + reg667 + reg668,
                se = "hc1", df)

out <- etable(homo, hetero, tex = TRUE, se.row = TRUE)
```

```
knitr::asis_output(c("\\begin{center}", out, "\\end{center}"))
```

Dependent Variable:	lwage	
Model:	(1)	(2)
Variables		
(Intercept)	4.739***	4.739***
	(0.0715)	(0.0746)
educ	0.0747***	0.0747***
	(0.0035)	(0.0036)
exper	0.0848***	0.0848***
	(0.0066)	(0.0068)
expersq	-0.0023***	-0.0023***
	(0.0003)	(0.0003)
black	-0.1990***	-0.1990***
	(0.0182)	(0.0182)
south	-0.1480***	-0.1480***
	(0.0260)	(0.0280)
smsa	0.1364***	0.1364***
	(0.0201)	(0.0192)
smsa66	0.0262	0.0262
	(0.0194)	(0.0186)
reg661	-0.1186***	-0.1186***
	(0.0388)	(0.0388)
reg662	-0.0222	-0.0222
	(0.0283)	(0.0299)
reg663	0.0260	0.0260
	(0.0274)	(0.0285)
reg664	-0.0635*	-0.0635*
	(0.0357)	(0.0368)
reg665	0.0095	0.0095
	(0.0361)	(0.0387)
reg666	0.0220	0.0220
	(0.0401)	(0.0411)
reg667	-0.0006	-0.0006
	(0.0394)	(0.0415)
reg668	-0.1750***	-0.1750***
	(0.0463)	(0.0470)
Fit statistics		
Standard-Errors	IID	Heteroskedasticity-robust
Observations	3,010	3,010
R^2	0.29984	0.29984
Adjusted R^2	0.29633	0.29633

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

(b)

There exists a practically and statistically significant partial correlation between *educ* and *nearc*4: the coefficient of *nearc*4 is 0.32, which means indivisuals near 4 yr college have additional 0.32 years of schooling. It's significant under 1% level with both *iid* and *robust* standard errors.

For some variables, the *robust* standard error is a little larger, and for other variables, the *robust* standard error is a little smaller. But they are still quite similar and the inference (ttest and pvalue) gives same conclusions. For *near*4, the *robust* standard error is a little smaller, but the coefficient is both significant under 1% level.

```
reduce_homo <- feols(educ ~ nearc4 + exper + expersq + black + south + smsa + smsa66 +
                     reg661 + reg662 + reg663 + reg664 + reg665 + reg666 + reg667 + reg668,
                     se = "iid", df)

reduce_hetero <- feols(educ ~ nearc4 + exper + expersq + black + south + smsa + smsa66 +
                      reg661 + reg662 + reg663 + reg664 + reg665 + reg666 + reg667 + reg668,
                      se = "hc1", df)

out <- etable(reduce_homo, reduce_hetero, tex = TRUE, se.row = TRUE)
```

```
knitr::asis_output(c("\\begin{center}", out, "\\end{center}"))
```

Dependent Variable:	educ	
Model:	(1)	(2)
Variables		
(Intercept)	16.85***	16.85***
	(0.2111)	(0.1866)
nearc4	0.3199***	0.3199***
	(0.0879)	(0.0851)
exper	-0.4125***	-0.4125***
	(0.0337)	(0.0321)
expersq	0.0009	0.0009
	(0.0016)	(0.0017)
black	-0.9355***	-0.9355***
	(0.0937)	(0.0925)
south	-0.0516	-0.0516
	(0.1354)	(0.1420)
smsa	0.4022***	0.4022***
	(0.1048)	(0.1112)
smsa66	0.0255	0.0255
	(0.1058)	(0.1106)
reg661	-0.2103	-0.2103
	(0.2025)	(0.1994)
reg662	-0.2889**	-0.2889*
	(0.1473)	(0.1513)
reg663	-0.2382*	-0.2382*
	(0.1426)	(0.1431)
reg664	-0.0931	-0.0931
	(0.1860)	(0.1799)
reg665	-0.4829**	-0.4829**
	(0.1882)	(0.1951)
reg666	-0.5131**	-0.5131**
	(0.2096)	(0.2090)
reg667	-0.4271**	-0.4271**
	(0.2056)	(0.2110)
reg668	0.3136	0.3136
	(0.2417)	(0.2338)
Fit statistics		
Standard-Errors	IID	Heteroskedasticity-robust
Observations	3,010	3,010
R^2	0.47712	0.47712
Adjusted R^2	0.47450	0.47450

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

(c)

The IV estimate has wider CI and the lower bound is closer to 0, hence it's more conservative. However, the estimates from both IV and OLS are significant on 95% level since the lower bounds are both larger than 0.

```
iv_c <- feols(lwage ~ exper + expersq + black + south + smsa + smsa66 +
              reg661 + reg662 + reg663 + reg664 + reg665 + reg666 + reg667 + reg668 |
              educ ~ nearc4,
              se = "hc1", df)

out <- etable(iv_c, hetero, tex = TRUE,
              coefstat = c("confint"),
              headers=list("IV" = 1, "OLS" = 1))
```

```
knitr::asis_output(c("\\begin{center}", out, "\\end{center}"))
```

Dependent Variable:	lwage	
	IV	OLS
Model:	(1)	(2)
Variables		
(Intercept)	3.774***	4.739***
	[1.970; 5.578]	[4.593; 4.886]
educ	0.1315**	0.0747***
	[0.0253; 0.2377]	[0.0675; 0.0818]
exper	0.1083***	0.0848***
	[0.0624; 0.1542]	[0.0716; 0.0981]
expersq	-0.0023***	-0.0023***
	[-0.0030; -0.0017]	[-0.0029; -0.0017]
black	-0.1468***	-0.1990***
	[-0.2497; -0.0438]	[-0.2346; -0.1634]
south	-0.1447***	-0.1480***
	[-0.2018; -0.0875]	[-0.2029; -0.0930]
smsa	0.1118***	0.1364***
	[0.0507; 0.1729]	[0.0987; 0.1741]
smsa66	0.0185	0.0262
	[-0.0218; 0.0588]	[-0.0102; 0.0627]
reg661	-0.1078***	-0.1186***
	[-0.1884; -0.0273]	[-0.1946; -0.0425]
reg662	-0.0070	-0.0222
	[-0.0733; 0.0592]	[-0.0809; 0.0365]
reg663	0.0404	0.0260
	[-0.0235; 0.1044]	[-0.0299; 0.0818]
reg664	-0.0579	-0.0635*
	[-0.1350; 0.0192]	[-0.1357; 0.0087]
reg665	0.0385	0.0095
	[-0.0588; 0.1357]	[-0.0664; 0.0853]
reg666	0.0551	0.0220
	[-0.0474; 0.1576]	[-0.0586; 0.1025]
reg667	0.0268	-0.0006
	[-0.0717; 0.1253]	[-0.0820; 0.0808]
reg668	-0.1909***	-0.1750***
	[-0.2905; -0.0912]	[-0.2671; -0.0829]
Fit statistics		
Observations	3,010	3,010
R^2	0.23817	0.29984
Adjusted R^2	0.23435	0.29633

Heteroskedasticity-robust co-variance matrix, 95% confidence intervals in brackets
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

(d)

The table below presents estimation results from the reduced form. After adding *nearc*2, the coefficient of *nearc*4 is even a littler larger and the *se* is essentially the same. However, the coefficient of *nearc*2 is way smaller than *nearc*4 and not significant. Therefore, *nearc*4 is more strongly related to *educ* than *nearc*2. After adding *nearc*2, the adjusted R square also increased a little - the independent variables can jointly explain more variation of *educ*.

```
reduce_d <- feols(educ ~ nearc2 + nearc4 + exper + expersq + black + south + smsa + smsa66 +
                         reg661 + reg662 + reg663 + reg664 + reg665 + reg666 + reg667 + reg668,
                         se = "hc1", df)

out <- etable(reduce_d, reduce_hetero, tex = TRUE, se.row = TRUE)
```

```
knitr::asis_output(c("\\begin{center}", out, "\\end{center}"))
```

Dependent Variable:	educ	
Model:	(1)	(2)
Variables		
(Intercept)	16.77***	16.85***
	(0.1940)	(0.1866)
nearc2	0.1230	
	(0.0776)	
nearc4	0.3206***	0.3199***
	(0.0850)	(0.0851)
exper	-0.4123***	-0.4125***
	(0.0320)	(0.0321)
expersq	0.0008	0.0009
	(0.0017)	(0.0017)
black	-0.9452***	-0.9355***
	(0.0925)	(0.0925)
south	-0.0419	-0.0516
	(0.1417)	(0.1420)
smsa	0.4014***	0.4022***
	(0.1113)	(0.1112)
smsa66	7.82×10^{-5}	0.0255
	(0.1118)	(0.1106)
reg661	-0.1688	-0.2103
	(0.2009)	(0.1994)
reg662	-0.2690*	-0.2889*
	(0.1524)	(0.1513)
reg663	-0.1902	-0.2382*
	(0.1468)	(0.1431)
reg664	-0.0377	-0.0931
	(0.1828)	(0.1799)
reg665	-0.4371**	-0.4829**
	(0.1979)	(0.1951)
reg666	-0.5022**	-0.5131**
	(0.2095)	(0.2090)
reg667	-0.3775*	-0.4271**
	(0.2144)	(0.2110)
reg668	0.3820	0.3136
	(0.2381)	(0.2338)
Fit statistics		
Standard-Errors	Heteroskedasticity-robust	
Observations	3,010	3,010
R^2	0.47756	0.47712
Adjusted R^2	0.47476	0.47450

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

After using two IV, the coefficient increases to 0.16, larger than former IV result (0.13) and the se is even smaller, so it does indicates a stronger relationship. The estimate is larger and we are more confident that it's significantly different from 0.

```
iv_d <- feols(lwage ~ exper + expersq + black + south + smsa + smsa66 +
              reg661 + reg662 + reg663 + reg664 + reg665 + reg666 + reg667 + reg668 |
              educ ~ nearc2 + nearc4,
              se = "hc1", df)

out <- etable(iv_d, iv_c, tex = TRUE,
              headers=list("#IV:2" = 1, "#IV:1" = 1))
```

```
knitr::asis_output(c("\\begin{center}", out, "\\end{center}"))
```

Dependent Variable:	lwage	
	#IV:2	#IV:1
Model:	(1)	(2)
Variables		
(Intercept)	3.340***	3.774***
	(0.8933)	(0.9199)
educ	0.1571***	0.1315**
	(0.0525)	(0.0541)
exper	0.1188***	0.1083***
	(0.0230)	(0.0234)
expersq	-0.0024***	-0.0023***
	(0.0004)	(0.0003)
black	-0.1233**	-0.1468***
	(0.0516)	(0.0525)
south	-0.1432***	-0.1447***
	(0.0303)	(0.0291)
smsa	0.1008***	0.1118***
	(0.0314)	(0.0311)
smsa66	0.0151	0.0185
	(0.0212)	(0.0206)
reg661	-0.1030**	-0.1078***
	(0.0427)	(0.0411)
reg662	-0.0002	-0.0070
	(0.0346)	(0.0338)
reg663	0.0470	0.0404
	(0.0336)	(0.0326)
reg664	-0.0554	-0.0579
	(0.0410)	(0.0393)
reg665	0.0515	0.0385
	(0.0508)	(0.0496)
reg666	0.0700	0.0551
	(0.0536)	(0.0523)
reg667	0.0391	0.0268
	(0.0516)	(0.0502)
reg668	-0.1980***	-0.1909***
	(0.0524)	(0.0508)
Fit statistics		
Observations	3,010	3,010
R^2	0.17020	0.23817
Adjusted R^2	0.16605	0.23435

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

(e)

IQ is significantly correlated with $nearc4$. Intuitively, IQ is also correlated with $education$, so $near4$ might not be a valid IV for $education$ since it does not satisfy the exclusion assumption. The previous IV estimations might be biased.

```
df$IQ <- as.numeric(df$IQ)
ols_e <- feols(IQ ~ nearc4, se = "hc1", df)
out <- etable(ols_e, tex = TRUE)
```

```
knitr::asis_output(c("\\begin{center}", out, "\\end{center}"))
```

Dependent Variable:	IQ
Model:	(1)
Variables	
(Intercept)	100.6***
	(0.6331)
nearc4	2.596***
	(0.7495)
Fit statistics	
Observations	2,061
R^2	0.00586
Adjusted R^2	0.00537

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

(f)

After controlling for *region* variables, the coefficient of *nearc*4 is not significant anymore. Therefore, after controlling for these *region* variables, *nearc*4 can serve as a valid IV for *education* since we can exclude the bias led by IQ.

```
df$IQ <- as.numeric(df$IQ)
ols_f <- feols(IQ ~ nearc4 + smsa66 + reg661 + reg662 + reg669, se = "hc1", df)
out <- etable(ols_f, ols_e, tex = TRUE)
```

```
knitr::asis_output(c("\\begin{center}", out, "\\end{center}"))
```

Dependent Variable:	IQ	
Model:	(1)	(2)
Variables		
(Intercept)	99.38***	100.6***
	(0.7135)	(0.6331)
nearc4	0.8681	2.596***
	(0.8183)	(0.7495)
smsa66	1.355*	
	(0.7904)	
reg661	4.768***	
	(1.423)	
reg662	5.808***	
	(0.8679)	
reg669	1.845	
	(1.142)	
Fit statistics		
Observations	2,061	2,061
R^2	0.03019	0.00586
Adjusted R^2	0.02783	0.00537

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*