# 41903 Pset 2 Q2

Lauren Mostrom

April 18, 2022

## 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 $\beta_4$ 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)

```
library(dplyr)
library(ggplot2)
library(lubridate)
library(data.table)
library(zoo)
library(tidyr)
library(ggpubr)
library(stargazer)
library(sandwich)
library(ivpack) # for IV
library(MASS) # for simulation from a multivariate normal distribution
library(texreg)
```

```r
# Load data BWGHT.raw into R
setwd("C:/Users/17036/OneDrive/Documents/GitHub/metrics3-zombie-boards/Psets/2/PS2Data")
BWGHT <- read.delim("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)**

```r
# 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 $\gamma_4$ 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, model.names=c("OLS","2SLS"))
```

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

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 |

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