Applied Micro PSET 4

Matthew Zhao working with Cedric Elkouh

2023-01-29

```
library(tidyverse)
library(haven)
library(stargazer)
library(xtable)
library(lmtest)
library(sandwich)
library(broom)
library(whitestrap)
library(recipes)
library(ivreg)
library(car)
options(scipen=999)
```

Question 1

a)

We interpret β_0 as the approximate expected job quality for an individual who has not taken an applied microeconometrics class and does not drink on a heavy basis. We interpret β_1 as the approximate expected change in job quality for an individual who takes an applied microeconometrics course holding their level of drinking fixed. We interpret β_2 as the approximate expected change in job quality for an individual who has heavy alcohol consumption on a regular basis and holding if they have taken an applied microeconometrics class fixed.

b)

We would suspect that drinking is not truthfully reported and that students may under- or over-report drinking habits i.e. report as heavy drinker even if they're not or report as not heavy drinker even if they are. If drinking is measured with error, particularly if that error is uncorrelated with the true value of the drinking variable (and as a result correlated with the observed value of drinking), we may not be able to get an unbiased estimate for the effect of drinking on job quality. $\hat{\beta}_2$ will be unbiased if instead the measurement error is uncorrelated with the observed measure of drinking, given that MLR.1-4 hold.

c)

A violation of MLR.4 would cause $\hat{\beta}_1$ to be biased and thus not accurately estimate β_1. In particular, we may be led to believe that students who take an applied microeconometrics course are more motivated than those who do not after holding drinking habits fixed As a result, MLR.4 would be violated.

d)

There are two assumptions that need to be fulfilled for econ_cost to be a valid instrument for econ. First is that econ_cost must affect the outcome (job quality) only through econ, specifically econ_cost must be independent of the error term (exclusion/exogeneity). The other is that econ_cost must be correlated with econ (relevance).

It is easy to see how econ_cost affects econ by reducing the cost of taking the course since this would naturally prompt more people to take the class. However, it is potentially more difficult to show that econ_cost is independent of the error term. Here we can plausibly say that because econ_cost is an indicator for attending a *randomly* chosen college that offers taking the econometrics class at a lower cost, exogeneity is satisfied.

e)

$$econ = \pi_0 + \pi_1 econ_{cost} + \nu \tag{1}$$

We can test instrument relevance by running this regression (econ~econ_cost) and examining the significance of π_1 via its t-stat and corresponding p-value as well as examining the F-stat to see if the regression coefficients are jointly significant.

f)

The IV estimator $\tilde{\beta}_1$ of β_1 solves the endogeneity bias issue with our regression because we introduce a proxy for the endogenous variable econ that by assumption is uncorrelated with the error term (which was the original issue with econ). As a result, by estimating $\tilde{\beta}_1$, we are able to isolate the causal effect of econ on job quality that is solely due to the econ variable.

g)

We cannot test if econ_cost is exogenous in the structural model since the full composition of the error term is unknown, and since the exogeneity assumption states that $Cov(econ\_cost, U) = 0$, we are unable to directly test this. However, we can show the extent to which econ_cost is exogenous using any observables that we have in our dataset and in a sense perform a balance test to show that the variable is plausibly exogenous.

If econ_cost is not truly exogenous in the structural model, then we will get a biased estimate of β_1. Specifically, we can express the IV estimate of β_1 as $\hat{\beta}_{IV} = \beta + \dfrac{\hat{Cov}(Z,U)}{\hat{Cov}(X,Z)}$. It is apparent from this expression that if exogeneity is violated, that is $Cov(Z,U) \neq 0$, $\hat{\beta}_{IV} \neq \beta$.

h)

If econ_cost is a weak instrument, this means that the correlation betwene econ_cost and econ is small resulting in a weak first-stage. Since the probability limit of the IV estimator $\underset{n\to\infty}{\text{plim}}\ \tilde{\beta}_1 = \beta_1 + \dfrac{Cov(Z,U)}{Cov(X,Z)}$, even if $Cov(X,Z)$ is small, as long as exogeneity holds, the IV estimator will remain consistent. However, since $Var(\tilde{\beta}_1) = \dfrac{\sigma^2}{SST_X \rho^2_{X,Z}}$, if the instrument is weak, the variance of the IV estimator will be large. Furthermore, the bias of 2SLS can be expressed as $\mathbb{E}[\hat{\beta}^{2SLS} - \beta] \approx \dfrac{\sigma_{\epsilon U}}{\sigma^2_\epsilon} \dfrac{1}{F+1}$. If the instrument is weak, F which is the population analogue of the F-stat for the joint significance of the instruments in the first-stage,

will be small. This causes the bias of 2SLS to approach that of OLS, resulting in the IV estimator giving a biased estimate of β_1.

If econ_cost is both a weak instrument and exogeneity is violated, we can see from the first equation above that $\tilde{\beta}_1$ will no longer be a consistent estimator for β_1 since $Cov(X, Z)$ is small while $Cov(Z, U) \neq 0$, meaning that the additional term will not go to zero and $\tilde{\beta}_1$ will not converge in probability to β_1. In this case, we would prefer the OLS estimator $\hat{\beta}_1$ to the IV estimator if $\frac{\rho_{Z,U}}{\rho_{Z,X}} > \rho_{X,U}$ since in this case the bias for the OLS estimator would be smaller than that of the TSLS IV estimator.

Problem 2

a)

The authors use an instrumental variables strategy to determine the causal effect of compulsory schooling on earnings. Specifically, they use quarter of birth as an instrument for compulsory schooling. Quarter of birth is plausibly random (hence exogenous), since we believe that quarter of birth is likely not related to personal attributes other than age at school entry. The instrument is related to compulsory schooling via school start age policies and compulsory school attendance law, whereby students born in the first quarter of the year tend to attend school for a fewer number of years due to reaching the minimum legal dropout age earlier in their educational careers.

b)

One potential violation is that quarter of birth is correlated with family background in some way e.g. parents who are more educated/wealthy tend to have children towards the end of the year. Another could be that students who are born earlier in the year tend enter school at an older age and so are more likely to have greater physical abilities than their peers. This could have induced them to leave school earlier to take on blue collar jobs since they may not believe that further education would benefit them.

c)

Here the authors are trying to show that quarter of birth is a relevant instrument for years of schooling. Specifically, in the first two panels up to high school graduate, they are trying to show that the relationship they previously described between quarter of birth and schooling holds, whereby those born in the first quarter tend to drop out of school earlier and thus have fewer years of education than those in other quarters due to the structure of compulsory schooling laws.

In the other panels, the authors seek to show that this relationship between quarter of birth and educational attainment becomes insignificant due to compulsory schooling laws no longer being binding past high school, thus indicating that the differences in education from the first two panels is generated by this quarter of birth effect.

The first two panels show highly significant coefficients for the quarter dummies and joint F-test, indicating that there could be a relationship. However, in attempting to show that this relationship fades when compulsory schooling laws are no longer binding, the authors find and acknowledge a significant effect of first quarter birth on college attainment, indicating that there could be systematic differences between those born in the first quarter and those born in other quarters.

d)

We could if other developed nations had similar institutional structures relating to compulsory schooling laws that are based on birth dates. This would generate the exogenous variation in educational attainment

that the authors exploited for their instrumental variables strategy and allow us to apply this identification strategy in other settings.

e)

This could potentially be a problem since the instruments used by the authors may be weak i.e. quarter of birth has little ability to explain the variation in schooling. We may be led to believe that this is the case because the TSLS estimates of $beta_1$ appear to vary wildly for different specifications and at times are extremely large, indicating that $Cov(X, Z)$ may be small. While the authors attempt to address this relevance assumption in Table II by showing that compulsory schooling laws have a statistically significant effect on enrollment rates, we are unable to conclusively say they have shown relevance of the instruments.

f)

The primary difference between the specifications in Table V and Table VII is that Table VII includes state fixed effects as well as higher level interactions. While this improves the precision of the TSLS estimates, it also removes some of the variation in the instruments which likely further weakens our ability to satisfy the relevance assumption. Additionally, adding more weak instruments likely biases the TSLS estimate towards that of OLS, as well as potentially violating instrument exogeneity.

Given that the estimates in Table V face the issues of weak instruments and overidentification, it is likely that these estimates are already biased and do not recover the causal effect of interest. If exogeneity is still plausible, then the estimate is at least consistent. However, with Table VII there is potentially a greater chance that instrument exogeneity is violated since there are 180 instruments. These estimates also face the same dual issue of overidentification and weak instruments and are likely extremely biased to the point of being the same as the OLS estimates.

g)

Even if the IV estimates in the paper are internally valid, it is unlikely that the estimates are generalizable to the average effect of education on earnings in the population as a whole. One reason is that education in high school is simply different than education provided during other periods of life, so at most this would describe the average effect of education on earnings specifically for additional years of high school education. Furthermore, there are likely heterogeneous treatment effects of education on earnings in the broader population making it difficult to generalize the results from the paper since they only consider homogeneous treatment effects for a narrow subset of the population and type of education.

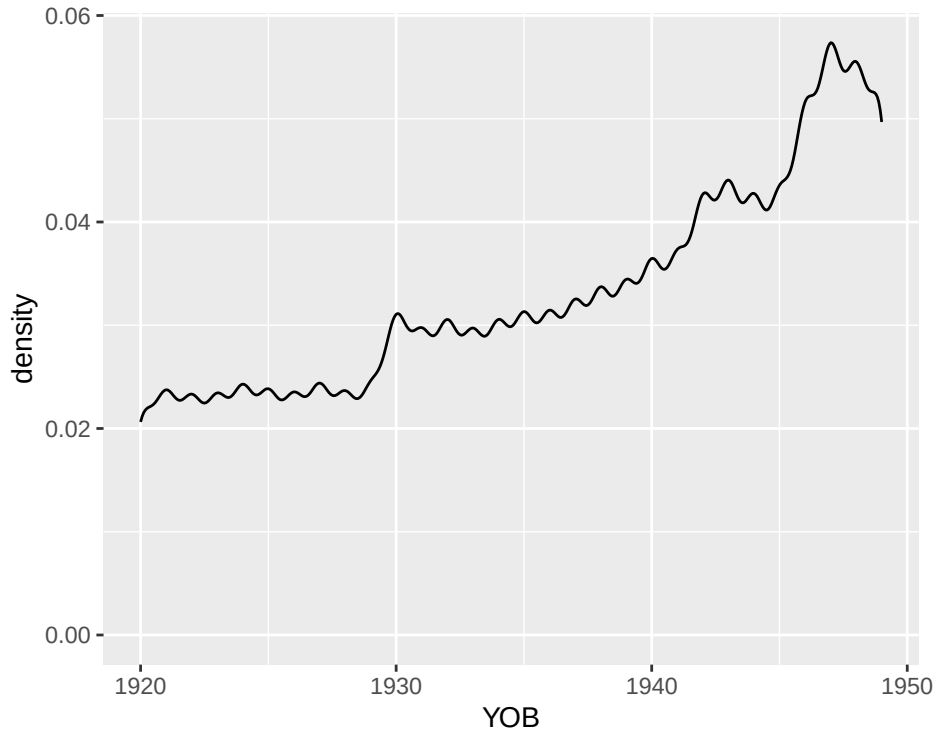
h)

```
df <- read_dta('data/CENSUS7080.dta')
```

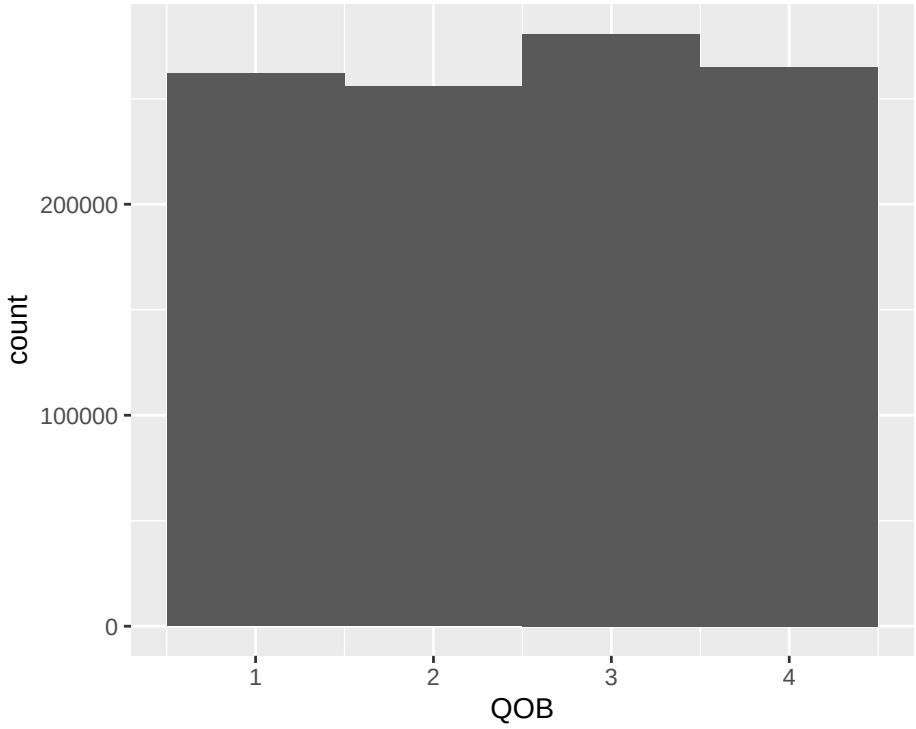
```
tab4_df <- df %>%
  filter((YOB >= 1920) & (YOB < 1930) & (CENSUS == 70)) %>%
  mutate(yob = as.factor(YOB),
         qob = as.factor(QOB)) %>%
  select(!c(YOB,QOB))
first_stage <- recipe(EDUC ~ ., data = tab4_df)

tab4_df <- first_stage %>%
  step_dummy(yob,qob,one_hot = T) %>%
```

```
  step_interact(terms = ~ starts_with("qob"):starts_with("yob")) %>%
  prep(training = tab4_df) %>%
  bake(tab4_df) %>%
  mutate(age_square = AGEQ^2)
```

```
ggplot(data = df, aes(x=YOB)) +
  geom_density()
```




```
ggplot(data = df, aes(x=QOB)) +
  geom_histogram(bins=length(unique(df$QOB)))
```




```
edu_sum <- df %>%
  group_by(QOB) %>%
  summarise(mean_edu = mean(EDUC),.groups='keep')
edu_sum
```

```
## # A tibble: 4 x 2
## # Groups:   QOB [4]
##     QOB mean_edu
##   <dbl>    <dbl>
## 1     1     12.7
## 2     2     12.8
## 3     3     12.9
## 4     4     12.9
```

i)

```
yobxq_dummies <- colnames(tab4_df)[str_detect(colnames(tab4_df),'_yob_')]
year_dummies <- setdiff(colnames(tab4_df)[str_detect(colnames(tab4_df),'yob')],yobxq_dummies)
reg_dummies <- c('ENOCENT','ESOCENT','MIDATL','MT','NEWENG','SOATL','WNOCENT','WSOCENT')

tab4_tsls <- tab4_df

#table 4 regressions
tab4_col1 <- lm(paste('LWKLYWGE ~ EDUC + ',paste(year_dummies,collapse='+')), data=tab4_df)

tab4_tsls$EDUC <- lm(paste('EDUC ~',paste(yobxq_dummies,collapse='+'),'+',paste(year_dummies,collapse='
tab4_col2 <- lm(paste('LWKLYWGE ~ EDUC + ',paste(year_dummies,collapse='+')), data=tab4_tsls)
```

```
tab4_col3 <- lm(paste('LWKLYWGE ~ EDUC + AGEQ + age_square + ' ,paste(year_dummies,collapse='+')), data=

tab4_tsls$EDUC <- lm(paste('EDUC ~ AGEQ + age_square + ',paste(yobxq_dummies,collapse='+'),'+',paste(yea
tab4_col4 <- lm(paste('LWKLYWGE ~ EDUC + AGEQ + age_square + ' ,paste(year_dummies,collapse='+')), data=

tab4_col5 <- lm(paste('LWKLYWGE ~ EDUC + RACE + SMSA + MARRIED + ',paste(year_dummies,collapse='+'),'+'

tab4_tsls$EDUC <- lm(paste('EDUC ~ RACE + SMSA + MARRIED + AGEQ + age_square + ',paste(yobxq_dummies,col
tab4_col6 <- lm(paste('LWKLYWGE ~ EDUC + RACE + SMSA + MARRIED + ',paste(year_dummies,collapse='+'),'+'

tab4_col7 <- lm(paste('LWKLYWGE ~ EDUC + RACE + SMSA + MARRIED + AGEQ + age_square + ',paste(year_dummie

tab4_tsls$EDUC <- lm(paste('EDUC ~ RACE + SMSA + MARRIED + AGEQ + age_square + ',paste(yobxq_dummies,col
tab4_col8 <- lm(paste('LWKLYWGE ~ EDUC + RACE + SMSA + MARRIED + AGEQ + age_square + ',paste(year_dummie
```

```
# table 4 columns 1-4
models <- list(tab4_col1,tab4_col2,tab4_col3,tab4_col4)
stargazer(models,
          type='text',digits=4, column.sep.width = "-15pt",
          keep = 1:4, keep.stat = c('chi2'),
          dep.var.labels.include = F,
          title='Table IV (1-4)')
```

```
##
## Table IV (1-4)
## ================================================
##                       Dependent variable:
##             ---------------------------------
##                (1)       (2)       (3)       (4)
## ------------------------------------------------
## EDUC       0.0802*** 0.0769*** 0.0802*** 0.1310***
##            (0.0004)  (0.0165)  (0.0004)  (0.0352)
##
## AGEQ                            0.1446**   0.1409*
##                                 (0.0676)  (0.0743)
##
## age_square                      -0.0015**  -0.0014
##                                 (0.0007)  (0.0008)
##
## yob_X1920
##
##
## ================================================
## ================================================
## Note:                  *p<0.1; **p<0.05; ***p<0.01
```

```
# table 4 columns 5-8
models <- list(tab4_col5,tab4_col6,tab4_col7,tab4_col8)
stargazer(models,
          type='text',digits=4,column.sep.width = "-15pt",
          dep.var.labels.include = F,
          keep = 1:7, keep.stat = c('chi2'),
          title='Table IV (5-8)')
```

```
##
## Table IV (5-8)
## ================================================
##                         Dependent variable:
##             ---------------------------------------
##                (1)        (2)        (3)        (4)
## ------------------------------------------------
## EDUC       0.0701***  0.0705***  0.0701***  0.1007***
##             (0.0004)   (0.0169)   (0.0004)   (0.0354)
##
## RACE       -0.2980*** -0.2827*** -0.2980*** -0.2271***
##             (0.0043)   (0.0429)   (0.0043)   (0.0822)
##
## SMSA       -0.1343*** -0.1259*** -0.1343*** -0.1163***
##             (0.0026)   (0.0122)   (0.0026)   (0.0210)
##
## MARRIED    0.2928***  0.2954***  0.2928***  0.2804***
##             (0.0037)   (0.0074)   (0.0037)   (0.0150)
##
## AGEQ                               0.1162*    0.1170*
##                                    (0.0652)   (0.0701)
##
## age_square                         -0.0013*   -0.0012
##                                    (0.0007)   (0.0008)
##
## yob_X1920
##
##
## ================================================
## ================================================
## Note:                      *p<0.1; **p<0.05; ***p<0.01
```

```
tab5_df <- df %>%
  filter((YOB >= 1930) & (YOB < 1940) & (CENSUS == 80)) %>%
  mutate(yob = as.factor(YOB),
         qob = as.factor(QOB)) %>%
  select(!c(YOB,QOB))
first_stage <- recipe(EDUC ~ ., data = tab5_df)

tab5_df <- first_stage %>%
  step_dummy(yob,qob,one_hot = T) %>%
  step_interact(terms = ~ starts_with("qob"):starts_with("yob")) %>%
  prep(training = tab5_df) %>%
  bake(tab5_df) %>%
  mutate(age_square = AGEQ^2)

yobxq_dummies <- colnames(tab5_df)[str_detect(colnames(tab5_df),'_yob_')]
year_dummies <- setdiff(colnames(tab5_df)[str_detect(colnames(tab5_df),'yob')],yobxq_dummies)
reg_dummies <- c('ENOCENT','ESOCENT','MIDATL','MT','NEWENG','SOATL','WNOCENT','WSOCENT')

tab5_tsls <- tab5_df

#table 5 regressions
tab5_col1 <- lm(paste('LWKLYWGE ~ EDUC + ',paste(year_dummies,collapse='+')), data=tab5_df)
```

```
tab5_tsls$EDUC <- lm(paste('EDUC ~',paste(yobxq_dummies,collapse='+'),'+',paste(year_dummies,collapse='+
tab5_col2 <- lm(paste('LWKLYWGE ~ EDUC + ',paste(year_dummies,collapse='+')), data=tab5_tsls)

tab5_col3 <- lm(paste('LWKLYWGE ~ EDUC + AGEQ + age_square + ' ,paste(year_dummies,collapse='+')), data=

tab5_tsls$EDUC <- lm(paste('EDUC ~ AGEQ + age_square + ',paste(yobxq_dummies,collapse='+'),'+',paste(yea
tab5_col4 <- lm(paste('LWKLYWGE ~ EDUC + AGEQ + age_square + ' ,paste(year_dummies,collapse='+')), data=

tab5_col5 <- lm(paste('LWKLYWGE ~ EDUC + RACE + SMSA + MARRIED + ',paste(year_dummies,collapse='+'),'+'

tab5_tsls$EDUC <- lm(paste('EDUC ~ RACE + SMSA + MARRIED + AGEQ + age_square + ',paste(yobxq_dummies,col
tab5_col6 <- lm(paste('LWKLYWGE ~ EDUC + RACE + SMSA + MARRIED + ',paste(year_dummies,collapse='+'),'+'

tab5_col7 <- lm(paste('LWKLYWGE ~ EDUC + RACE + SMSA + MARRIED + AGEQ + age_square + ',paste(year_dummie

tab5_tsls$EDUC <- lm(paste('EDUC ~ RACE + SMSA + MARRIED + AGEQ + age_square + ',paste(yobxq_dummies,col
tab5_col8 <- lm(paste('LWKLYWGE ~ EDUC + RACE + SMSA + MARRIED + AGEQ + age_square + ',paste(year_dummie
```

```
# table 5 columns 1-4
models <- list(tab5_col1,tab5_col2,tab5_col3,tab5_col4)
stargazer(models,
          type='text',digits=4,column.sep.width = "-15pt",
          dep.var.labels.include = F,
          keep = 1:4, keep.stat = c('chi2'),
          title='Table V (1-4)')
```

```
##
## Table V (1-4)
## ================================================
##                      Dependent variable:
##             ------------------------------------
##                (1)       (2)       (3)      (4)
## ------------------------------------------------
## EDUC       0.0711*** 0.0891*** 0.0711*** 0.0754**
##            (0.0003)  (0.0171)  (0.0003)  (0.0308)
##
## AGEQ                             -3.0694  -3.2185
##                                 (2.6767)  (3.0376)
##
## age_square                       0.0008    0.0008
##                                 (0.0007)  (0.0008)
##
## yob_X1930
##
##
## ================================================
## ================================================
## Note:                  *p<0.1; **p<0.05; ***p<0.01
```

```
# table 5 columns 5-8
models <- list(tab5_col5,tab5_col6,tab5_col7,tab5_col8)
stargazer(models,
```

```
          type='text',digits=4,column.sep.width = "-15pt",
          dep.var.labels.include = F,
          keep = 1:7, keep.stat = c('chi2'),
          title='Table V (5-8)')
```

```
##
## Table V (5-8)
## ========================================================
##                          Dependent variable:
##             ---------------------------------------------
##                (1)         (2)         (3)         (4)
## --------------------------------------------------------
## EDUC        0.0632***   0.0808***   0.0632***    0.0597*
##              (0.0003)    (0.0172)    (0.0003)    (0.0305)
##
## RACE        -0.2575***  -0.2190***  -0.2575***  -0.2631***
##              (0.0040)    (0.0297)    (0.0040)    (0.0481)
##
## SMSA        -0.1763***  -0.1482***  -0.1763***  -0.1800***
##              (0.0029)    (0.0203)    (0.0029)    (0.0321)
##
## MARRIED     0.2479***   0.2494***   0.2479***   0.2487***
##              (0.0032)    (0.0043)    (0.0032)    (0.0076)
##
## AGEQ                                 -3.0028     -2.8871
##                                      (2.6040)    (2.9099)
##
## age_square                            0.0008      0.0007
##                                      (0.0007)    (0.0007)
##
## yob_X1930
##
##
## ========================================================
## ========================================================
## Note:                        *p<0.1; **p<0.05; ***p<0.01
```

```
tab6_df <- df %>%
  filter((YOB >= 1940) & (YOB < 1950) & (CENSUS == 80)) %>%
  mutate(yob = as.factor(YOB),
         qob = as.factor(QOB)) %>%
  select(!c(YOB,QOB))
first_stage <- recipe(EDUC ~ ., data = tab6_df)

tab6_df <- first_stage %>%
  step_dummy(yob,qob,one_hot = T) %>%
  step_interact(terms = ~ starts_with("qob"):starts_with("yob")) %>%
  prep(training = tab6_df) %>%
  bake(tab6_df) %>%
  mutate(age_square = AGEQ^2)

yobxq_dummies <- colnames(tab6_df)[str_detect(colnames(tab6_df),'_yob_')]
year_dummies <- setdiff(colnames(tab6_df)[str_detect(colnames(tab6_df),'yob')],yobxq_dummies)
```

```
reg_dummies <- c('ENOCENT','ESOCENT','MIDATL','MT','NEWENG','SOATL','WNOCENT','WSOCENT')

tab6_tsls <- tab6_df

#table 6 regressions
tab6_col1 <- lm(paste('LWKLYWGE ~ EDUC + ',paste(year_dummies,collapse='+')), data=tab6_df)

tab6_tsls$EDUC <- lm(paste('EDUC ~',paste(yobxq_dummies,collapse='+'),'+',paste(year_dummies,collapse='+
tab6_col2 <- lm(paste('LWKLYWGE ~ EDUC + ',paste(year_dummies,collapse='+')), data=tab6_tsls)

tab6_col3 <- lm(paste('LWKLYWGE ~ EDUC + AGEQ + age_square + ' ,paste(year_dummies,collapse='+')), data=

tab6_tsls$EDUC <- lm(paste('EDUC ~ AGEQ + age_square + ',paste(yobxq_dummies,collapse='+'),'+',paste(yea
tab6_col4 <- lm(paste('LWKLYWGE ~ EDUC + AGEQ + age_square + ' ,paste(year_dummies,collapse='+')), data=

tab6_col5 <- lm(paste('LWKLYWGE ~ EDUC + RACE + SMSA + MARRIED + ',paste(year_dummies,collapse='+'),'+'

tab6_tsls$EDUC <- lm(paste('EDUC ~ RACE + SMSA + MARRIED + AGEQ + age_square + ',paste(yobxq_dummies,col
tab6_col6 <- lm(paste('LWKLYWGE ~ EDUC + RACE + SMSA + MARRIED + ',paste(year_dummies,collapse='+'),'+'

tab6_col7 <- lm(paste('LWKLYWGE ~ EDUC + RACE + SMSA + MARRIED + AGEQ + age_square + ',paste(year_dummie

tab6_tsls$EDUC <- lm(paste('EDUC ~ RACE + SMSA + MARRIED + AGEQ + age_square + ',paste(yobxq_dummies,col
tab6_col8 <- lm(paste('LWKLYWGE ~ EDUC + RACE + SMSA + MARRIED + AGEQ + age_square + ',paste(year_dummie
```

```
# table 6 columns 1-4
models <- list(tab6_col1,tab6_col2,tab6_col3,tab6_col4)
stargazer(models,
          type='text',digits=4,column.sep.width = "-15pt",
          dep.var.labels.include = F,
          keep = 1:4, keep.stat = c('chi2'),
          title='Table VI (1-4)')
```

```
##
## Table VI (1-4)
## ================================================
##                       Dependent variable:
##             ---------------------------------------
##                (1)       (2)       (3)        (4)
## ------------------------------------------------
## EDUC       0.0573*** 0.0553*** 0.0573***  0.0981***
##            (0.0003)  (0.0143)   (0.0003)  (0.0227)
##
## AGEQ                            9.0735***  5.8887**
##                                  (2.1638)  (2.8615)
##
## age_square                      -0.0023*** -0.0015**
##                                  (0.0006)  (0.0007)
##
## yob_X1940
##
##
## ================================================
```

```
## ================================================
## Note:                  *p<0.1; **p<0.05; ***p<0.01
```

```
# table 6 columns 5-8
models <- list(tab6_col5,tab6_col6,tab6_col7,tab6_col8)
stargazer(models,
          type='text',digits=4,column.sep.width = "-15pt",
          dep.var.labels.include = F,
          keep = 1:7, keep.stat = c('chi2'),
          title='Table VI (5-8)')
```

```
##
## Table VI (5-8)
## ================================================
##                        Dependent variable:
##            ----------------------------------
##               (1)        (2)        (3)        (4)
## ------------------------------------------------
## EDUC       0.0520***  0.0430***  0.0521***  0.0809***
##             (0.0003)   (0.0149)   (0.0003)   (0.0244)
##
## RACE       -0.2107*** -0.2161*** -0.2108*** -0.1749***
##             (0.0032)   (0.0208)   (0.0032)   (0.0305)
##
## SMSA       -0.1418*** -0.1469*** -0.1419*** -0.1154***
##             (0.0023)   (0.0150)   (0.0023)   (0.0225)
##
## MARRIED    0.2445***  0.2469***  0.2444***  0.2451***
##             (0.0022)   (0.0026)   (0.0022)   (0.0023)
##
## AGEQ                              7.5438***   5.4491*
##                                    (2.1081)   (2.8043)
##
## age_square                        -0.0019***  -0.0014*
##                                    (0.0005)   (0.0007)
##
## yob_X1940
##
##
## ================================================
## ================================================
## Note:                    *p<0.1; **p<0.05; ***p<0.01
```

We are able to fairly closely replicate the results from the paper in that the coefficient estimates for β_1 are relatively similar, with the standard errors somewhat different. Comparing the OLS and TSLS estimates, we see that the TSLS estimates are generally larger than those of the OLS ones, potentially indicating a weak first stage. Furthermore, we see that the coefficient for TSLS varies significantly more than for OLS across specifications and samples, which could be related.

j)

```
tab4_yobxq_dummies <- colnames(tab4_df)[str_detect(colnames(tab4_df),'_yob_')]
tab5_yobxq_dummies <- colnames(tab5_df)[str_detect(colnames(tab5_df),'_yob_')]
tab6_yobxq_dummies <- colnames(tab6_df)[str_detect(colnames(tab6_df),'_yob_')]

tab4_year_dummies <- setdiff(colnames(tab4_df)[str_detect(colnames(tab4_df),'yob')],tab4_yobxq_dummies)
tab5_year_dummies <- setdiff(colnames(tab5_df)[str_detect(colnames(tab5_df),'yob')],tab5_yobxq_dummies)
tab6_year_dummies <- setdiff(colnames(tab6_df)[str_detect(colnames(tab6_df),'yob')],tab6_yobxq_dummies)

tab4fstg2 <- lm(paste('EDUC ~',paste(tab4_yobxq_dummies,collapse='+'),'+',paste(tab4_year_dummies,colla
tab4fstg4 <- lm(paste('EDUC ~ AGEQ + age_square + ',paste(tab4_yobxq_dummies,collapse='+'),'+',paste(tal
tab4fstg6 <- lm(paste('EDUC ~ RACE + SMSA + MARRIED + AGEQ + age_square + ',paste(tab4_yobxq_dummies,co
tab4fstg8 <- lm(paste('EDUC ~ RACE + SMSA + MARRIED + AGEQ + age_square + ',paste(tab4_yobxq_dummies,co


tab5_fstg2 <- lm(paste('EDUC ~',paste(tab5_yobxq_dummies,collapse='+'),'+',paste(tab5_year_dummies,colla
tab5_fstg4 <- lm(paste('EDUC ~ AGEQ + age_square + ',paste(tab5_yobxq_dummies,collapse='+'),'+',paste(ta
tab5_fstg6 <- lm(paste('EDUC ~ RACE + SMSA + MARRIED + AGEQ + age_square + ',paste(tab5_yobxq_dummies,c
tab5_fstg8 <- lm(paste('EDUC ~ RACE + SMSA + MARRIED + AGEQ + age_square + ',paste(tab5_yobxq_dummies,c

tab6_fstg2 <- lm(paste('EDUC ~',paste(tab6_yobxq_dummies,collapse='+'),'+',paste(tab6_year_dummies,colla
tab6_fstg4 <- lm(paste('EDUC ~ AGEQ + age_square + ',paste(tab6_yobxq_dummies,collapse='+'),'+',paste(ta
tab6_fstg6 <- lm(paste('EDUC ~ RACE + SMSA + MARRIED + AGEQ + age_square + ',paste(tab6_yobxq_dummies,c
tab6_fstg8 <- lm(paste('EDUC ~ RACE + SMSA + MARRIED + AGEQ + age_square + ',paste(tab6_yobxq_dummies,c
```

```
models <- list(tab4fstg2,tab4fstg4,tab4fstg6,tab4fstg8)

stargazer(models,
          type='latex',digits=2,column.sep.width = "0pt",
          dep.var.labels.include = F,
          keep = 1:2, keep.stat = c('f'),
          title='First Stage for Table IV')
```

Table 1: First Stage for Table IV

	Dependent variable:			
	(1)	(2)	(3)	(4)
RACE			-2.52^{***}	-2.32^{***}
			(0.02)	(0.02)
SMSA				
F Statistic	14.85^{***} (df = 39)	14.85^{***} (df = 39)	336.30^{***} (df = 42)	371.35^{***} (df = 50)
Note:				*p<0.1; **p<0.05; ***p<0.01

```
models <- list(tab5_fstg2,tab5_fstg4,tab5_fstg6,tab5_fstg8)

stargazer(models,
          type='latex',digits=2,column.sep.width = "0pt",
```

```
          dep.var.labels.include = F,
          keep = 1:2, keep.stat = c('f'),
          title='First Stage for Table V')
```

Table 2: First Stage for Table V

	Dependent variable:			
	(1)	(2)	(3)	(4)
RACE			−1.71***	−1.57***
			(0.02)	(0.02)
SMSA				
F Statistic	27.90*** (df = 39)	27.21*** (df = 40)	333.62*** (df = 43)	398.38*** (df = 51)
Note:				*p<0.1; **p<0.05; ***p<0.01

```
models <- list(tab6_fstg2,tab6_fstg4,tab6_fstg6,tab6_fstg8)

stargazer(models,
          type='latex',digits=2,column.sep.width = "0pt",
          dep.var.labels.include = F,
          keep = 1:2, keep.stat = c('f'),
          title='First Stage for Table VI')
```

Table 3: First Stage for Table VI

	Dependent variable:			
	(1)	(2)	(3)	(4)
RACE			−1.38***	−1.24***
			(0.02)	(0.02)
SMSA				
F Statistic	77.36*** (df = 39)	75.44*** (df = 40)	442.76*** (df = 43)	507.20*** (df = 51)
Note:				*p<0.1; **p<0.05; ***p<0.01

From these first stage regressions, we see that the first stage regression is somewhat weak for the first two specifications of Table IV (just above 10) which could be a sign of a weak instrument issue. We do not see these with Tables V and VI where the joint F-tests for the first two specifications of both have F stat over 10, but we cannot rule out the possibility of weak instruments.

Problem 3

a)

Currently our plan for paper extension is to determine the elasticity of demand for coal in the United States since the paper we are replicating does this in China. Typically when estimating demand elasticities, an instrumental variables strategy is used due to the simultaneity problem of price. As such, we plan to use the size of coal seams as an instrument for coal price since the size of a coal seam naturally impacts the supply of coal without any impact on the demand curve. Since the size of coal seams is random (determined by nature), it satisfies exogeneity. Furthermore, since it directly impacts the supply of coal, it will have an impact on coal prices, thus satisfying the relevance assumption.