Problem Set 3: Instrumental Variables Key

Claire Duquennois

Name: Xiang Li

1. Empirical Analysis using Data from Ananat (2011, AEJ:AE)

This exercise uses data from Elizabeth Ananat's paper, "The Wrong Side(s) of the Tracks: The Causal Effects of Racial Segregation on Urban Poverty and Inequality," published in the *American Economic Journal: Applied Economics* in 2011. This paper studies how segregation has affected population characteristics and income disparity in US cities using the layout of railroad tracks as an instrumental variable.

2. Finding the data

The data can be found by following the link on the AEJ: Applied Economics' website which will take you to the ICPSR's data repository. You will need to sign in to get access to the data files. Once logged in, you will find the set of files that are typically included in a replication file. These include several datasets, several .do files (which is a STATA command file). For this assignment we will be using the`aej_maindata.dta` file.

3. Set up and opening the data

3.1 Question: Load any packages you will need and the data contained in the `aej_maindata.dta` file. How many observations are contained in the data. What is the level of an observation?

Code and Answer:

```
library("haven")
library("dplyr")
```

```
##
## 载入程辑包：'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library("stargazer")

##
## Please cite as:

##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Sta
tistics Tables.

##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer

library("lfe")

## 载入需要的程辑包：Matrix

## Warning: 程辑包'Matrix'是用 R 版本 4.3.2 来建造的

maindata <- read_dta("aej_maindata.dta")
nrow(maindata)

## [1] 121
```

Answer: There are 121 observations are contained in the data, it's a city level of an observation.

4. Data Description

4.1 Question:The dataset contains many variables, some of which are not used in this exercise. Keep the following variables in the final dataset (Hint: use the `select` function in `dplyr`).

Name	Description
dism1990	1990 dissimilarity index
herf	RDI (Railroad division index)
lenper	Track length per square km
povrate_w	White poverty rate 1990
povrate_b	Black poverty rate 1990
area1910	Physical area in 1910 (1000 sq. miles)
count1910	Population in 1910 (1000s)
ethseg10	Ethnic Dissimilariy index in 1910
ethiso10	Ethnic isolation index in 1910
black1910	Percent Black in 1910

Name	Description
passpc	Street cars per capita 1915
black1920	Percent Black 1920
lfp1920	Labor Force Participation 1920
incseg	Income segregation 1990
pctbk1990	Percent Black 1990
manshr	Share employed in manufacturing 1990
pop1990	Population in 1990

You can find the detailed description of each variable in the original paper.

Code:

```
dataselect <- select(maindata, dism1990, herf, lenper, povrate_w, povrate_b, a
rea1910, count1910, ethseg10, ethiso10, black1910, passpc, black1920, lfp1920,
 incseg, pctbk1990, manshr, pop1990)
```

5. Summary Statistics:

5.1 Question: Report summary statistics of the following variables in the dataset:"dism1990", "herf", "lenper", "povrate_w", "povrate_b". Present these summary statistics in a formatted table, you can use `stargazer` or other packages.

Code:

```
data <- as.data.frame(dataselect)
selected_vars <- c("dism1990", "herf", "lenper", "povrate_w", "povrate_b")
stargazer(data[, selected_vars], type = "text", digits=3, title = "Summary Sta
tistics")
```

```
##
## Summary Statistics
## ========================================
## Statistic  N  Mean  St. Dev.  Min    Max
## ----------------------------------------
## dism1990  121 0.569  0.135   0.329  0.873
## herf      121 0.723  0.141   0.238  0.987
## lenper    121 0.001  0.001   0.0002 0.013
## povrate_w 121 0.095  0.035   0.035  0.216
## povrate_b 121 0.264  0.080   0.093  0.504
## ----------------------------------------
```

6. Reduced Form:

6.1 Question: We are interested in understanding how segregation affects population characteristics and income disparity in US cities. We will focus on two outcome variables: the poverty rate for blacks and whites. Regress these two outcome variables on segregation in 1990, our explanatory variable, and interpret your results. Report robust standard errors. Make sure you specify the units of measurement in your interpretation.

Code:

```
#install.packages("lmtest", "sandwich")
library(lmtest)

## Warning: 程辑包'lmtest'是用 R 版本 4.3.2 来建造的

## 载入需要的程辑包：zoo

##
## 载入程辑包：'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

##
## 载入程辑包：'lmtest'

## The following object is masked from 'package:lfe':
##
##     waldtest

library(sandwich)

#Regress the two outcome variables on segregation in 1990

model.black <- lm(povrate_b ~ dism1990, data = data)
model.white <- lm(povrate_w ~ dism1990, data = data)

stargazer(model.black, model.white, header = FALSE, type = "text", se = list(m
odel.black$rse, model.white$rse), title = "Effect of Poverty Rate for Blacks a
nd Whites on Segregation")

##
## Effect of Poverty Rate for Blacks and Whites on Segregation
## ============================================================
##                                 Dependent variable:
##                             ---------------------------
##                               povrate_b      povrate_w
```

```
##                                   (1)           (2)
## ------------------------------------------------------------
## dism1990                        0.182***      -0.073***
##                                (0.051)       (0.022)
##
## Constant                        0.161***      0.136***
##                                (0.030)       (0.013)
##
## ------------------------------------------------------------
## Observations                      121           121
## R2                               0.095         0.081
## Adjusted R2                      0.088         0.074
## Residual Std. Error (df = 119)    0.076         0.033
## F Statistic (df = 1; 119)       12.511***     10.538***
## ============================================================
## Note:                          *p<0.1; **p<0.05; ***p<0.01
```

```
# Report robust standard errors

robust_black <- coeftest(model.black, vcov = vcovHC)
robust_white <- coeftest(model.white, vcov = vcovHC)

stargazer(robust_black, robust_white, header = FALSE, type = "text", title = "
Robust Standard Errors")
```

```
##
## Robust Standard Errors
## ====================================
##             Dependent variable:
##         ---------------------------
##
##              (1)           (2)
## ------------------------------------
## dism1990    0.182***      -0.073***
##            (0.046)       (0.020)
##
## Constant    0.161***      0.136***
##            (0.029)       (0.012)
##
## ====================================
## ====================================
## Note:     *p<0.1; **p<0.05; ***p<0.01
```

Answer: The segregation in 1990 and the poverty rate for blacks is positive related, which means with 1 unit increasing of segregation, the poverty rate for blacks is expected to increase by 0.182 percentage points. The segregation in 1990 and the poverty rate for whites is negative related, which means with 1 unit increasing of segregation, the poverty rate for whites is expected to decrease by 0.073 percentage points.

6.2 Question: Explain the problem with giving a causal interpretation to the estimates you just produced. Give examples of specific factors that might make a causal interpretation of your result problematic.

Answer: There might be other variables that we don't observe, such as public policy or schooling years, that affect the poverty rate, also high poverty rate could lead to increasing of segregation.

7. Validity of the instrument:

7.1 Question: Estimate the following regression and interpret it's coefficients,

$$dism1990_i = \beta_0 + \beta_1 RDI_i + \beta_2 tracklength_i + \epsilon.$$

Code:

```
reg.dism <- felm(dism1990 ~ herf + lenper, data)
stargazer(reg.dism, header = FALSE, type = "text", title = "Effect of RDI and
Tracklength on Segregation")
```

```
##
## Effect of RDI and Tracklength on Segregation
## ================================================
##                        Dependent variable:
##                    ----------------------------
##                              dism1990
## ------------------------------------------------
## herf                          0.357***
##                               (0.081)
##
## lenper                        18.514**
##                               (9.126)
##
## Constant                      0.294***
##                               (0.058)
##
## ------------------------------------------------
## Observations                    121
## R2                             0.203
## Adjusted R2                    0.189
## Residual Std. Error      0.122 (df = 118)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

Answer: B_1 = 0.357: This indicates the estimated change in dism1990 for a one-unit increase in RDI, holding other variables constant, which means with 1 unit increasing of RDI, the dism1990 is expected to increase by 0.357 percentage points.

B_2 = 18.514: This indicates the estimated change in dism1990 for a one-unit increase in tracklength, holding other variables constant, which means with 1 unit increasing of tracklength, the dism1990 is expected to increase by 18.514 percentage points.

7.2 Question: Re-estimate the specification above using the `scale()` command around the variables you wish to standardize in the regression. What do you notice?

Code:

```
## Standardize the tracklength in the regression
reg.sd <- felm(dism1990 ~ herf + scale(lenper), data)
stargazer(reg.sd, header = FALSE, type = "text", title = "Re-estimate")
```

```
##
## Re-estimate
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                               dism1990
## ------------------------------------------------
## herf                          0.357***
##                               (0.081)
##
## scale(lenper)                  0.023**
##                               (0.012)
##
## Constant                      0.310***
##                               (0.060)
##
## ------------------------------------------------
## Observations                    121
## R2                             0.203
## Adjusted R2                    0.189
## Residual Std. Error      0.122 (df = 118)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

Answer: The coefficient of tracklength decreased from 18.524 to 0.023, which looks more similar to the coefficient of RDI, this can help us to see the influence easier.

7.3 Question: In the context of instrumental variables, what is this regression referred to as and why is it important?

Answer: This referred to first stage, it is important because it can show explanatory power explain in some way to segregation, RDI to be a good instrument.

7.4 Question: Illustrate the relationship between the RDI and segregation graphically.

Code:

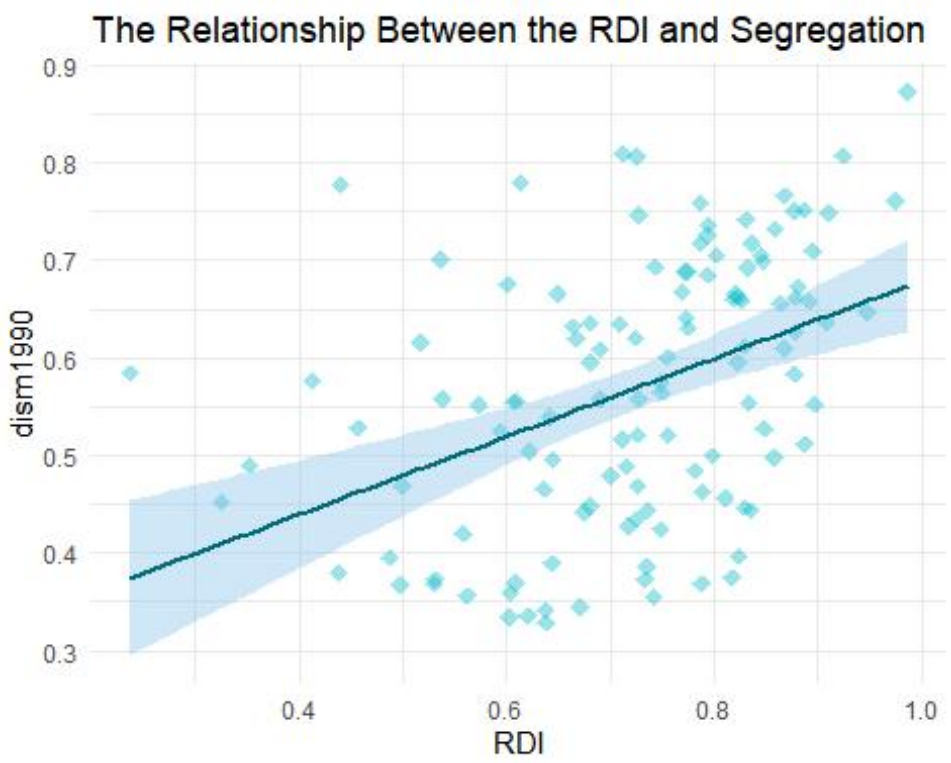
```
library(ggplot2)

## Warning: 程辑包'ggplot2'是用 R 版本 4.3.2 来建造的

plotted <- ggplot(data = data, aes(x = herf, y = dism1990, color = variable))
+
  geom_point(aes(y = dism1990,col = "dism1990"), color = "#00bdcd", alpha = 0.
4, shape = 18, size = 3) +
  geom_smooth(method = "lm", aes(y = dism1990, col = "dism1990"), color = "#00
6b7b", fill = "#88c4e8") +
  theme_minimal() +
  labs(title = "The Relationship Between the RDI and Segregation" , face = "bo
ld", x = "RDI", y = "dism1990", color = "Regressand Variable")

plotted

## `geom_smooth()` using formula = 'y ~ x'
```



7.5 Question: Is there a concern that this might be a weak instrument? Why would this be a problem?

Answer:

```
summary(reg.sd)

##
## Call:
##    felm(formula = dism1990 ~ herf + scale(lenper), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.22582 -0.10817  0.00864  0.08031  0.32113
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.31025    0.05984   5.185 9.05e-07 ***
## herf           0.35731    0.08130   4.395 2.43e-05 ***
## scale(lenper)  0.02333    0.01150   2.029   0.0447 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1218 on 118 degrees of freedom
## Multiple R-squared(full model): 0.2025   Adjusted R-squared: 0.189
```

```
## Multiple R-squared(proj model): 0.2025   Adjusted R-squared: 0.189
## F-statistic(full model):14.98 on 2 and 118 DF, p-value: 1.591e-06
## F-statistic(proj model): 14.98 on 2 and 118 DF, p-value: 1.591e-06
```

We can see the F-statistic is 14.98 from above, which is larger than 10. This means that it is not a weak instrument.

7.6 Question: Select a number of relevant city characteristics in the data to regress on the RDI and track length. Present your results and interpret your findings. Why do these results matter for answering our question of interest?

Code and Answer:

```
reg.1 <- felm(area1910 ~ herf + lenper, data)
reg.2 <- felm(count1910 ~ herf + lenper, data)
reg.3 <- felm(black1910 ~ herf + lenper, data)
reg.4 <- felm(ethseg10 ~ herf + lenper, data)
reg.5 <- felm(ethiso10 ~ herf + lenper, data)


stargazer(reg.1, reg.2, reg.3, reg.4, reg.5, header = FALSE, type = "text",
          title = "Impact of City Characteristics on RDI",
          se = list(reg.1$rse, reg.2$rse, reg.4$rse, reg.5$rse))

##
## Impact of City Characteristics on RDI
## ================================================================================
===================================
##                                                    Dependent variable:

##                     ---------------------------------------------------------
---------------------------------
##                            area1910           count1910          black1910
   ethseg10       ethiso10
##                               (1)                (2)                (3)
    (4)            (5)
## --------------------------------------------------------------------------------
-----------------------------------
## herf                      -3,992.637           665.751            -0.001
    0.076          0.027
##                          (11,986.490)        (1,362.964)         (0.185)
    (0.070)        (0.086)
##

## lenper                    -574,401.000         75,553.190          9.236
    15.343         -12.439
##                          (553,669.000)       (134,814.900)        (53.248)
    (17.288)       (19.930)
```

```
##

## Constant                18,409.570**           976.876             0.007
    0.238***          0.048
##                        (8,612.320)           (927.189)           (0.121)
    (0.051)         (0.054)
##

## ------------------------------------------------------------------------
-----------------------------------
## Observations                  58                  121                 121
     49              49
## R2                         0.007               0.006              0.290
    0.014           0.009
## Adjusted R2                -0.029              -0.011              0.278
    -0.029          -0.034
## Residual Std. Error 15,050.340 (df = 55) 1,903.415 (df = 118) 0.018 (df = 1
18) 0.184 (df = 46) 0.075 (df = 46)
## ========================================================================
==================================
## Note:
  *p<0.1; **p<0.05; ***p<0.01
```

All of the coefficients of variables are not significant, which means that the exclusion restriction holds.

7.7 Question: What are the two conditions necessary for a valid instrument? What evidence do you have that the RDI meet these conditions? Be specific in supporting this claim.

Answer: Cov(z, x_1) != 0 (the first stage) Cov(z, v) = 0 (the exclusion restriction) y:poverty rate x_1: dism1990(segregation) z: herf(RDI) v: city characteristics

The first stage quite strong, very significant, which means RDI it related with segregation, while not related to other characteristic, which means the exclusion restriction holds.

7.8 Question: Do you believe the instrument is valid? Why/why not?

Answer: Maybe, however, if there are something elso can influence the poverty rate, such as the price of train or the places accesible by railwae, this might be not valid.

7.9 Question: Generate a table that estimates the effect of segregation on the poverty rate for blacks and whites by OLS and then using the RDI instrument. Make sure you report robust standard errors. How does the use of the RDI instrument change the estimated coefficients?

Code and Answer:

```
model.black.ols <- felm(povrate_b ~ dism1990, data = data)
model.white.ols <- felm(povrate_w ~ dism1990, data = data)

model.black.RDI <- felm(povrate_b ~ lenper|0|(dism1990 ~ herf), data = data)
model.white.RDI <- felm(povrate_w ~ lenper|0|(dism1990 ~ herf), data = data)

stargazer(model.black.ols, model.white.ols, model.black.RDI, model.white.RDI,

          header = FALSE, type = "text",
          title = "Effect of Segregation on the Poverty Rate for Blacks and Whi
tes (OLS and IV)",
          se = list(model.black.ols$rse, model.white.ols$rse,model.black.RDI$r
se, model.white.RDI$rse))

##
## Effect of Segregation on the Poverty Rate for Blacks and Whites (OLS and IV)
## ================================================================================
=============
##                                          Dependent variable:

##                     ---------------------------------------------------------
-----------
##                        povrate_b       povrate_w       povrate_b       povr
ate_w
##                          (1)             (2)             (3)             (4)

## ---------------------------------------------------------------------------
-------------
## dism1990                 0.182***       -0.073***

##                         (0.045)         (0.019)

##

## lenper                                                   -4.780          0.60
2
##                                                         (3.067)         (1.97
0)
##

## `dism1990(fit)`                                          0.258**        -0.19
```

```
6***
##                                                    (0.108)          (0.06
5)
##

## Constant                0.161***         0.136***         0.121**          0.2
05***
##                        (0.029)          (0.012)          (0.061)          (0.0
37)
##

## ------------------------------------------------------------------------
-------------
## Observations              121              121              121              12
1
## R2                      0.095            0.081            0.084            -0.
150
## Adjusted R2              0.088            0.074            0.068            -0.
170
## Residual Std. Error 0.076 (df = 119) 0.033 (df = 119) 0.077 (df = 118) 0.03
7 (df = 118)
## ========================================================================
=============
## Note:                                                  *p<0.1; **p<0.05; *
**p<0.01
```

The use of RDI instrument can remove the bias and address endogeneity concerns.

7.10 Question: What is the reduced form equation?

Answer:

$$ povrateblack_i = \pi_i0 + \pi_1 \times herf_i + \eta\\ povratewhite_i = \gamma_i0 + \gamma_1 \times herf_i + \eta $$

7.11 Question: (2 pages) For the two poverty rates, estimate the reduced form on all the cities and illustrate the reduced form relationships graphically.

Code:

```r
model.black.rf <- felm(povrate_b ~ herf, data = data)
model.white.rf <- felm(povrate_w ~ herf, data = data)

stargazer(model.black.rf, model.white.rf, header = FALSE, type = "text", se =
list(model.black.rf$rse, model.white.rf$rse), title = "Reduced Form")
```

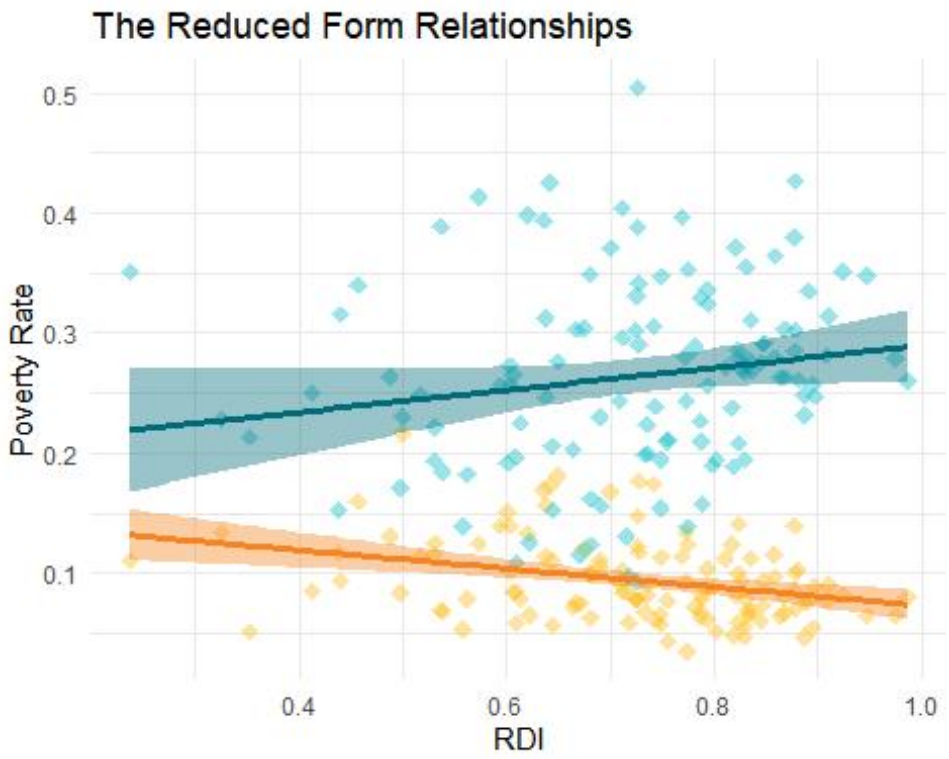
```
##
## Reduced Form
## ===============================================================
##                                  Dependent variable:
##                              ---------------------------
##                                 povrate_b      povrate_w
##                                    (1)            (2)
## ---------------------------------------------------------------
## herf                              0.092**       -0.077***
##                                  (0.046)        (0.022)
##
## Constant                          0.197***      0.150***
##                                  (0.036)        (0.017)
##
## ---------------------------------------------------------------
## Observations                        121            121
## R2                                 0.027          0.099
## Adjusted R2                        0.019          0.092
## Residual Std. Error (df = 119)     0.079          0.033
## ===============================================================
## Note:                            *p<0.1; **p<0.05; ***p<0.01
```

```
plotted_po <- ggplot(data = data, aes(x = herf, y = value, color = variable))
+
  geom_point(aes(y = povrate_w, col = "povrate_w"), color = "#ffbc14", alpha =
 0.4, shape = 18, size = 3) +
  geom_point(aes(y = povrate_b, col = "povrate_b"), color = "#00bdcd", alpha =
 0.4, shape = 18, size = 3) +
  geom_smooth(method = "lm", aes(y = povrate_w, col = "povrate_w"), color = "#
f88421", fill = "#f88421", size = 1.3) +
  geom_smooth(method = "lm", aes(y = povrate_b, col = "povrate_b"), color = "#
006b7b", fill = "#006b7b", size = 1.3) +
  theme_minimal() +
  labs(title = "The Reduced Form Relationships", face = "bold",
       x = "RDI",
       y = "Poverty Rate",
       color = "Regressand Variable") +
  guides(color = guide_legend(title = "Regressand Variable"))
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
plotted_po
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```




```
##Sorry, I have tried many methods to display the legend, but none of them hav
e been successful.
```

7.12 Question: Generate a table with at least six estimations that checks whether the main results are robust to adding additional controls for city characteristics. What do you conclude?

Code:

```
reg.71 <- felm(povrate_w ~ lenper|0|(dism1990 ~ herf + lenper), data)

## Warning in chol.default(mat, pivot = TRUE, tol = tol): the matrix is either
## rank-deficient or not positive definite

reg.72 <- felm(povrate_w ~ lenper + pctbk1990|0|(dism1990 ~ herf +lenper + pct
bk1990),data)

## Warning in chol.default(mat, pivot = TRUE, tol = tol): the matrix is either
## rank-deficient or not positive definite

reg.73 <- felm(povrate_w ~ lenper + pctbk1990 + lfp1920|0|(dism1990 ~ herf +le
nper + pctbk1990 +lfp1920),data)

## Warning in chol.default(mat, pivot = TRUE, tol = tol): the matrix is either
## rank-deficient or not positive definite
```

```
reg.74 <- felm(povrate_b ~ lenper|0|(dism1990 ~ herf + lenper), data)

## Warning in chol.default(mat, pivot = TRUE, tol = tol): the matrix is either
## rank-deficient or not positive definite

reg.75 <- felm(povrate_b ~ lenper + pctbk1990|0|(dism1990 ~ herf +lenper + pct
bk1990),data)

## Warning in chol.default(mat, pivot = TRUE, tol = tol): the matrix is either
## rank-deficient or not positive definite

reg.76 <- felm(povrate_b ~ lenper + pctbk1990 + lfp1920|0|(dism1990 ~ herf +le
nper + pctbk1990 +lfp1920),data)

## Warning in chol.default(mat, pivot = TRUE, tol = tol): the matrix is either
## rank-deficient or not positive definite

stargazer(reg.71, reg.72, reg.73,reg.74, reg.75, reg.76, header = FALSE, type
= "text", se = list(reg.71$rse, reg.72$rse, reg.73$rse,reg.74$rse, reg.75$rse,
 reg.76$rse))

##
## ===============================================================================
================================================
##                                                          Dependent variable:

##                    ------------------------------------------------------------
-----------------------------------------
##                                         povrate_w
          povrate_b
##                           (1)             (2)             (3)             (4)
            (5)             (6)
## -------------------------------------------------------------------------------
-----------------------------------------
## lenper                   0.602           -0.479          -0.918          -4.
780          -2.331          -1.586
##                         (1.970)         (1.801)         (1.550)         (3.0
67)         (2.402)         (2.655)
##

## pctbk1990                                 0.211           0.180
            -0.478*         -0.426*
##                                          (0.153)         (0.136)
           (0.246)         (0.230)
##

## lfp1920                                                   -0.122
                           0.207
##                                                         (0.105)
                          (0.208)
##
```

```
## `dism1990(fit)`        -0.196***         -0.241**         -0.224**          0.
258**          0.360**           0.333**
##                        (0.065)          (0.097)          (0.088)          (0.1
08)          (0.141)           (0.132)
##

## Constant                0.205***         0.219***         0.263***          0.1
21**           0.091             0.016
##                        (0.037)          (0.048)          (0.073)          (0.0
61)          (0.068)           (0.120)
##

## ------------------------------------------------------------------------
-------------------------------------------
## Observations              121              121              121              12
1              121               121
## R2                      -0.150           -0.254           -0.172           0.0
84            0.108             0.130
## Adjusted R2             -0.170           -0.286           -0.212           0.
068            0.085             0.100
## Residual Std. Error 0.037 (df = 118) 0.039 (df = 117) 0.038 (df = 116) 0.07
7 (df = 118) 0.076 (df = 117) 0.076 (df = 116)
## ========================================================================
=============================================
## Note:
             *p<0.1; **p<0.05; ***p<0.01
```

Answer: Controls do not have a large effect on the results

8. Why Two Stage least squares?

Because the estimates in this paper only feature one endogenous regressor and one instrument, it is an excellent example with which to illustrate build intuition and see what the instrumental variables regressor is actually doing because in this scenario the IV estimator is exactly equal to the two stage least squares estimator ($\widehat{\beta}_{IV} = \widehat{\beta}_{2SLS}$).

8.1 Question: Estimate the first stage regression and use your estimates to generate the predicted values for the explanatory variable for all the observations.

Code:

```
first_stage <- lm(dism1990 ~ herf + lenper, data = data)
stargazer(first_stage, header = FALSE, type = "text", se = list(first_stage$rs
e))
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                               dism1990
## ------------------------------------------------
## herf                          0.357***
##                               (0.081)
##
## lenper                        18.514**
##                               (9.126)
##
## Constant                      0.294***
##                               (0.058)
##
## ------------------------------------------------
## Observations                    121
## R2                             0.203
## Adjusted R2                    0.189
## Residual Std. Error      0.122 (df = 118)
## F Statistic          14.983*** (df = 2; 118)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

```
data$predicted_dism1990 <- predict(first_stage)
```

8.2 Question: If our instrument is valid, the step above "removed" the "bad" endogenous variation from the predicted explanatory variable, keeping only the exogenous variation that is generated by the instrument. Now run the second stage by regressing our outcome variable on the predicted values generated above and the relevant controls. Compare your estimates from this regression to those generated earlier. How do they compare?

Code:

```
#model.black.RDI <- felm(povrate_b ~ lenper|0|(dism1990 ~ herf), data = data)
#model.white.RDI <- felm(povrate_w ~ lenper|0|(dism1990 ~ herf), data = data)
model.black.pd <- felm(povrate_b ~ predicted_dism1990 + lenper, data = data)
model.white.pd <- felm(povrate_w ~ predicted_dism1990 + lenper, data = data)

stargazer(model.black.pd, model.white.pd, header = FALSE, type = "text", se =
list(model.black.pd$rse, model.white.pd$rse))
```

```
##
## ============================================================
##                                  Dependent variable:
##                              ---------------------------
```

```
##                                 povrate_b     povrate_w
##                                    (1)           (2)
## ------------------------------------------------------------
## predicted_dism1990                 0.258*      -0.196***
##                                   (0.134)       (0.060)
##
## lenper                             -4.780        0.602
##                                   (5.578)       (1.516)
##
## Constant                           0.121        0.205***
##                                   (0.075)       (0.034)
##
## ------------------------------------------------------------
## Observations                         121           121
## R2                                  0.027         0.111
## Adjusted R2                         0.010         0.096
## Residual Std. Error (df = 118)      0.079         0.033
## ============================================================
## Note:                           *p<0.1; **p<0.05; ***p<0.01
```

Answer: They are pretty similar.

9. Yet another IV trick: Taking the "Good" variation and scaling it

9.1 Question: Take the coefficient from you reduced form estimate and divide it by your first stage estimate. How does this value compare your earlier estimate for the main result?

Answer: Reduced form for blacks : 0.092 Reduced form for whites : -0.077 Divided by first stage: 0.357

We get 0.258 for blacks and -0.196 for whites, which is the same with what we got in 7.9 and 8.2.

10. Submission instructions:

- Make sure the final version of your assignment is knit in pdf format and uploaded to gradescope. Make sure you have one question response per page (unless otherwise indicated) so that question positions align with the template in gradescope.The final PDF should be 22 pages long.