Problem Set 3: Instrumental Variables

Claire Duquennois

NAME: ROHAN KRISHNAN

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.

Finding the data

I have downloaded Ananat's `aej_maindata.dta` file and made it available in the RCloud assignment workspace. I downloaded this data from the AER's website which links you to the ICPSR's data repository. Anyone can sign in to get access to the replication data files. These include the typical files in a replication folder: several datasets, several .do files (which is a STATA command file), and text files with the data descriptions which tell you about the different variables included in the dataset.

1 Set up and opening the data

1.1 Question: Load the have, dplyr, stargazer, lfe and ggplot2 packages and the data contained in the `aej_maindata.dta` file. Make sure it is stored as a data frame.

Code:

```
#Load necessary libraries
library(haven)
library(dplyr)
library(stargazer)
library(lfe)
library(ggplot2)

#Load aej_maindata.dta
maindata <- read_dta("~/Downloads/aej_maindata.dta")

#Check if maindata is a data frame
is(maindata)
```

```
## [1] "tbl_df"     "tbl"        "data.frame" "list"       "oldClass"
## [6] "vector"
```

```
#Convert maindata to data frame
maindata <- as.data.frame(maindata)

#Check if maindata is a data frame
is(maindata)
```

```
## [1] "data.frame" "list"       "oldClass"   "vector"
```

1.2 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
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:

```
#Filter data
md2 <- maindata %>%
  select(dism1990, herf, lenper, povrate_w, povrate_b, area1910,
         count1910, ethseg10, ethiso10, black1910, passpc, black1920,
         lfp1920, incseg, pctbk1990, manshr, pop1990)
```

2 Data description:

2.1 Question: How many observations are contained in the data. What is the level of an observation?

Answer:

There are 121 observations in the data. Each observation is on the city level.

2.2 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:

```
#Display summary statistics
stargazer(md2[,c("dism1990", "herf", "lenper", "povrate_w", "povrate_b")],
          type = "latex", header = FALSE, no.space = TRUE,
          title = "Summary Statistics of Relevant Variables")
```

Table 2: Summary Statistics of Relevant Variables

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

3 Reduced Form:

3.1 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.

Hint 1: These exact results are reported in the second row of columns 1 and 2 of table 2.

Hint 2: Since the units of the explanatory variable are strange, it is helpful to interpret the effect in terms of standard deviations. So instead of interpreting a one unit change in `dism1990`, interpret a one standard deviation (0.14) change in `dism1990`.

Code:

```
#Generate regressions
reg_b <- felm(povrate_b ~ dism1990, data = md2)
reg_w <- felm(povrate_w ~ dism1990, data = md2)

#Display regression summaries
stargazer(reg_b, reg_w, type = "latex", se = list(reg_b$rse, reg_w$rse),
          header = FALSE, no.space = TRUE,
          title = "Naive Regression of Segregation on Poverty Rate")
```

Table 3: Naive Regression of Segregation on Poverty Rate

	Dependent variable:	
	povrate_b	povrate_w
	(1)	(2)
dism1990	0.182***	−0.073***
	(0.045)	(0.019)
Constant	0.161***	0.136***
	(0.029)	(0.012)
Observations	121	121
R^2	0.095	0.081
Adjusted R^2	0.088	0.074
Residual Std. Error (df = 119)	0.076	0.033
Note:	*p<0.1; **p<0.05; ***p<0.01	

Answer:

A one standard deviation (0.14) increase in a city's dissimilarity index is statistically significantly correlated with a 0.025 (2.5%) *increase* in the poverty rate among black residents in 1990. Meanwhile, a one standard deviation (0.14) increase in a city's dissimilarity's index is statistically significantly correlated with a 0.010 (1%) *decrease* in the poverty rate among white residents in 1990.

3.2 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:

The naive regression above has serious issues with the conditional independence assumption. In order to give a causal interpretation to a regression, the treatment (in this case segregation) must be plausibly random given the controlling variables in the regression. The regression above has no control variables and it is impossible to say that segregation was randomly assigned across cities without any controls. For example, the level or corruption, ratio of black to white residents, location of the city, population and many other factors are important ommitted variables that would have definitely affected the level of segregation of a city.

4 Validity of the instrument:

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

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

Hint 1: These exact results are reported in the first column of the top panel of table 1.

Hint 2: Since the units of the explanatory variable are strange, it is helpful to interpret the effect in terms of standard deviations. So instead of interpreting a one unit change in `herf`, interpret a one standard deviation (0.14) change in `herf`.

Code:

```
#Generate regression
reg_dism <- felm(dism1990 ~ herf + lenper, data = md2)

#Display regression summary table
stargazer(reg_dism, type = "latex", header = FALSE,
          no.space = TRUE, se = list(reg_dism$rse),
          title = "RDI and Track Length's Effect on Dissimilarity Index"
          )
```

Table 4: RDI and Track Length's Effect on Dissimilarity Index

	Dependent variable:
	dism1990
herf	0.357***
	(0.088)
lenper	18.514*
	(10.731)
Constant	0.294***
	(0.064)
Observations	121
R^2	0.203
Adjusted R^2	0.189
Residual Std. Error	0.122 (df = 118)
Note:	*p<0.1; **p<0.05; ***p<0.01

Answer:

A one standard deviation (0.14) increase in a city's railroad division index is statistically significantly correlated with a 0.050 *increase* in the city's dissimilarity index score in 1990. Meanwhile, a one square kilometer increase in a city's railroad track's length is statistically significantly (at the 0.01 level) correlated with a 18 point *decrease* in the city's dissimilarity index score in 1990.

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

Answer:

This regression is referred to as the *first stage*. It is important in the context of instrumental variables because it establishes that there is a relationship between our endogenous variable of interest (segregation) and our chosen instrument(s) (RDI and track length).

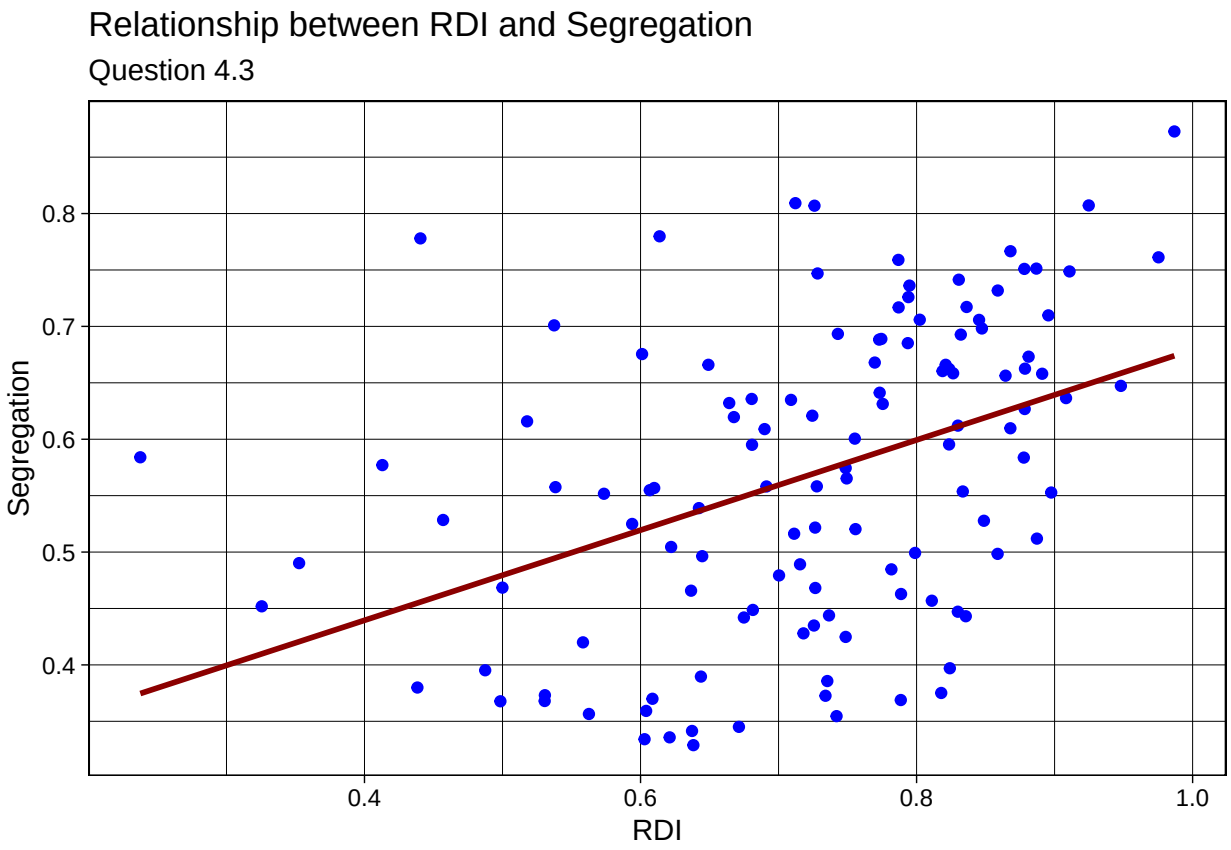
4.3 Question: Illustrate the relationship between the RDI and segregation graphically.

Hint: See figure 3.

Code:

```
#Create graph comparing RDI and segregation
md2 %>%
  ggplot(aes(x = herf, y = dism1990)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", se = FALSE, color = "dark red") +
  labs(x = "RDI", y = "Segregation", title = "Relationship between RDI and Segregation", subtitle = "Qu
  theme_linedraw()
```

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



4.4 Question: Is there a concern that this might be a weak instrument? Why would this be a problem? Hint: check the fstat.

Answer:

```
#Recover f-statistic from regression
summary(reg_dism)$F.fstat[c(1,4)]
```

```
##            F            p
## 1.498272e+01 1.590675e-06
```

From the output above, the f-statistic is 14.98 with a p-value much less than 0.001. The typical threshold for a strong instrument is an f-statistic of 10, so we can safely say that RDI and track length are strong instruments for segregation. A weak instrument would be a problem because the estimate for our instrument variable is as follows: $\hat{\beta_{iv}} = \beta + \frac{cov(z,v)}{cov(z,x_1)}$. If we have a weak instrument, then $cov(z, x_1)$ is very small. Thus, the

4.5 Question: Regress the following cith characteristics on the RDI and track length: `area1910` `count1910`, `black1910`, `incseg`, `lfp1920`. Present your results and interpret your findings. Why do these results matter for answering our question of interest?

Hint: In stargazer, add the option `omit.stat=c( "ser")` to remove the residual standard errors from the table footer so that the table fits the width of a page.

Code and Answer:

```
#Generate exclusion restriction regressions
reg1 <- felm(area1910 ~ herf + lenper, data = md2)
reg2 <- felm(count1910 ~ herf + lenper, data = md2)
reg3 <- felm(black1910 ~ herf + lenper, data = md2)
reg4 <- felm(incseg ~ herf + lenper, data = md2)
reg5 <- felm(lfp1920 ~ herf + lenper, data = md2)

#Display regression summaries in stargazer
stargazer(reg1, reg2, reg3, reg4, reg5, se = list(reg1$rse, reg2$rse, reg3$rse, reg4$rse, reg5$rse),
          header = FALSE, type = "latex", no.space = TRUE,
          omit.stat = c("ser"), title = "Exclusion Restriction")
```

Table 5: Exclusion Restriction

	Dependent variable:				
	area1910	count1910	black1910	incseg	lfp1920
	(1)	(2)	(3)	(4)	(5)
herf	−3,992.637	665.751	−0.001	0.032	0.028
	(11,986.490)	(1,362.964)	(0.010)	(0.032)	(0.024)
lenper	−574,401.000	75,553.190	9.236***	−2.504	−3.427**
	(553,669.000)	(134,814.900)	(0.650)	(1.626)	(1.500)
Constant	18,409.570**	976.876	0.007	0.196***	0.401***
	(8,612.320)	(927.189)	(0.007)	(0.025)	(0.018)
Observations	58	121	121	69	121
R^2	0.007	0.006	0.290	0.028	0.015
Adjusted R^2	−0.029	−0.011	0.278	−0.001	−0.002
Note:					*p<0.1; **p<0.05; ***p<0.01

4.6 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:

The two conditions necessary for a valid instrument are a (preferably) strong first stage (proved above) and the satisfaction of the exclusion restriction. To satisfy the exclusion restriction, we must be able to argue that our instrument is not correlated at all with any of the omitted variables that are confounding our original endogenous variable. We have shown that we have a strong first stage by analyzing the first stage f-statistics in the question above. For the exclusion restriction, the above regressions show the relationship between our two instruments (*herf* and *lenper*) with several of the included control variables that are suspected to have a selection effect on the level of segregation across cities. As shown in the tables, none of the variables have a statistically significant relationship with either instrument except for *black1910* (percent black in 1910) with track length and *lfp1920* and track length, though the effect is quite small. Since we have a strong first stage, a minor violation of the exclusion restriction should not blow up our model's error. In her paper, Ananat also provides extensive historical and contextual evidence and reasoning as to why a city's RDI and track length were not correlated with any of the omitted variables.

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

Answer:

I believe the instrument is valid because it clearly has a strong first stage and is almost entirely consistent in adhering to the exclusion restriction. Outside of the calculations within this assignment, I also believe that Ananat provides extensive and convincing evidence that none of her omitted variables had a relationship with her instruments (a.k.a railroad construction was purely dependent on economic considerations).

4.8 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?

Hint: these will be the exact results reported in row 2 of columns 1-4 in table 2.

Code and Answer:

```
#Generate regressions
reg_olsb <- felm(povrate_b ~ dism1990, data = md2)
reg_olsw <- felm(povrate_w ~ dism1990, data = md2)
reg_insb <- felm(povrate_b ~ lenper|0|
                   (dism1990 ~ herf), data = md2)
reg_insw <- felm(povrate_w ~ lenper |0|
                   (dism1990 ~ herf), data = md2)

#Display regression summaries in table
stargazer(reg_olsb, reg_olsw, reg_insb, reg_insw,
          se = list(reg_olsb$rse, reg_olsw$rse, reg_insb$rse, reg_insw$rse),
          header = FALSE, type = "latex", no.space = TRUE,
          omit.stat = c("ser"), title = "OLS vs IV Regressions")
```

Table 6: OLS vs IV Regressions

	Dependent variable:			
	povrate_b	povrate_w	povrate_b	povrate_w
	(1)	(2)	(3)	(4)
dism1990	0.182***	−0.073***		
	(0.045)	(0.019)		
lenper			−4.780	0.602
			(3.067)	(1.970)
‘dism1990(fit)‘			0.258**	−0.196***
			(0.108)	(0.065)
Constant	0.161***	0.136***	0.121**	0.205***
	(0.029)	(0.012)	(0.061)	(0.037)
Observations	121	121	121	121
R^2	0.095	0.081	0.084	−0.150
Adjusted R^2	0.088	0.074	0.068	−0.170
Note:				*p<0.1; **p<0.05; ***p<0.01

4.9 Question: What is the reduced form equation?

Answer:

The reduced form equation is $povertyrate = RDI + tracklength$. The track length variable is an important control to allow us to control for the length of tracks by city. In practice, we would make a regression for the white and black poverty rate variables separately.

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

Code:

```
#Generate reduced form regressions
reg_reduced1 <- felm(povrate_b ~ herf + lenper, data = md2)
reg_reduced2 <- felm(povrate_w ~ herf + lenper, data = md2)

#Display results in stargazer
stargazer(reg_reduced1, reg_reduced2, se = list(reg_reduced1$rse, reg_reduced2$rse),
          header = FALSE, no.space = TRUE, type = "latex",
          omit.stat = c("ser"),
          title = "Reduced Form")
```

Table 7: Reduced Form

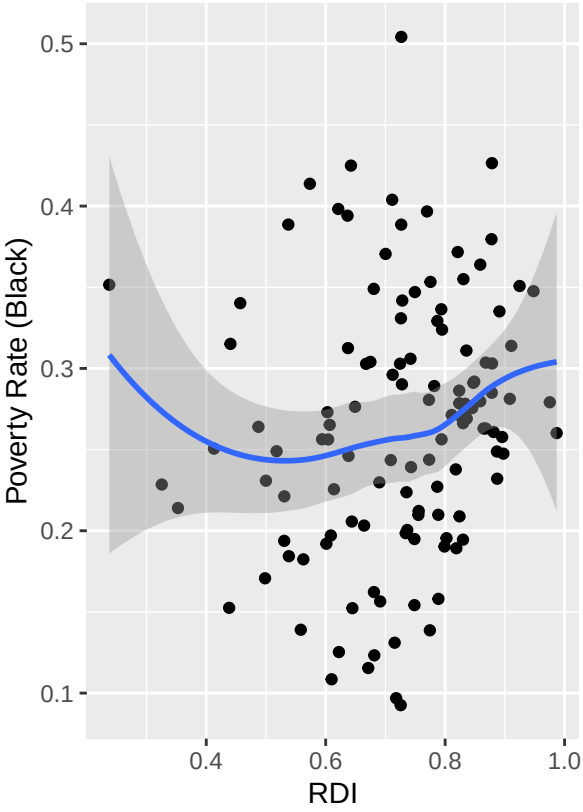
	Dependent variable:	
	povrate_b	povrate_w
	(1)	(2)
herf	0.092^{*}	-0.070^{***}
	(0.048)	(0.021)
lenper	0.004	-3.022^{***}
	(4.398)	(1.011)
Constant	0.197^{***}	0.148^{***}
	(0.036)	(0.017)
Observations	121	121
R^2	0.027	0.111
Adjusted R^2	0.010	0.096
Note:	*p<0.1; **p<0.05; ***p<0.01	

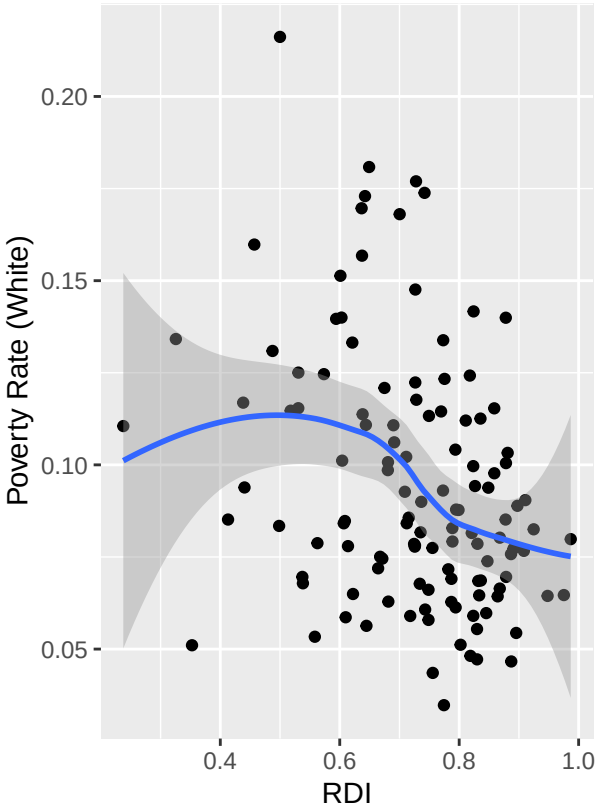
```
#Load patchwork library
library(patchwork)

#Create graphs
g1 <- md2 %>%
  ggplot(aes(x = herf,
             y = povrate_b)) +
  geom_point() +
  geom_smooth() +
  labs(x = "RDI",
       y = "Poverty Rate (Black)")
g2 <- md2 %>%
  ggplot(aes(x = herf,
             y = povrate_w)) +
  geom_point() +
  geom_smooth() +
  labs(x = "RDI",
       y = "Poverty Rate (White)")

#Display graphs side by side
g1 + g2
```

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





4.11 Question: Generate a table with six columns that check whether the main results are robust to adding additional controls for city characteristics. What do you conclude? (2 pages)

Hint: In stargazer, add the option `omit.stat=c( "ser")` to remove the residual standard errors from the table footer so that the table fits in a page.

Code:

```
#Generate regressions
reg_1 <- felm(povrate_b ~ lenper |0|
                (dism1990 ~ herf), data = md2)
reg_2 <- felm(povrate_w ~ lenper |0|
                (dism1990 ~ herf), data = md2)
reg_3 <- felm(povrate_b ~ lenper + area1910 |0|
                (dism1990 ~ herf), data = md2)
reg_4 <- felm(povrate_w ~ lenper + area1910 |0|
                (dism1990 ~ herf), data = md2)
reg_5 <- felm(povrate_b ~ lenper + area1910 + black1910
              |0|(dism1990 ~ herf), data = md2)
reg_6 <- felm(povrate_w ~ lenper + area1910 + black1910
              |0|(dism1990 ~ herf), data = md2)

#Display regressions
stargazer(reg_1, reg_2, reg_3, reg_4, reg_5, reg_6,
          se = list(reg_1$rse, reg_2$rse, reg3$rse, reg_4$rse, reg_5$rse, reg_6$rse),
          type = "latex", header = FALSE, no.space = TRUE, omit.stat = c("ser"),
          title = "Testing Robustness to Controls for IV Regressions")
```

Table 8: Testing Robustness to Controls for IV Regressions

	Dependent variable:					
	povrate_b	povrate_w	povrate_b	povrate_w	povrate_b	povrate_w
	(1)	(2)	(3)	(4)	(5)	(6)
lenper	−4.780	0.602	−6.992***	−1.190	−3.307	0.315
	(3.067)	(1.970)	(0.650)	(0.746)	(4.924)	(1.681)
area1910			−0.00000	0.00000**	−0.00000	0.00000**
				(0.00000)	(0.00000)	(0.00000)
black1910					−0.393	−0.161
					(0.471)	(0.149)
‘dism1990(fit)‘	0.258**	−0.196***	0.372	−0.159	0.379*	−0.156
	(0.108)	(0.065)		(0.144)	(0.209)	(0.144)
Constant	0.121**	0.205***	0.062***	0.185*	0.060	0.184*
	(0.061)	(0.037)	(0.007)	(0.095)	(0.135)	(0.095)
Observations	121	121	58	58	58	58
R^2	0.084	−0.150	−0.196	0.045	−0.193	0.059
Adjusted R^2	0.068	−0.170	−0.262	−0.008	−0.283	−0.012
Note:						*p<0.1; **p<0.05; ***p<0.01

Answer:

The instruments magnitude appears to slightly change as you add control variables to the regression. Its significance also decreases. The models for *povrate_w* had a smaller change in the magnitude of their

coefficients, indicating they may be slightly more robust to the addition of controls. However, in general, it appears that both models *are* robust to the addition of controls as the coefficients remain within 2 standard errors of the main estimation.

5 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 ($\hat{\beta}_{IV} = \hat{\beta}_{2SLS}$).

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

Code:

```
#Estimate first stage regression
reg_fs <- lm(dism1990 ~ herf + lenper, data = md2)

#Generate predicted poverty rates
clean_dism1990 <- predict(reg_fs, md2)
```

5.2 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:

```
#Add cleaned variable to data
md2$clean_dism1990 <- clean_dism1990

#Run second stage regression
reg_ss <- lm(povrate_b ~ clean_dism1990 + lenper, data = md2)
reg_ss2 <- lm(povrate_w ~ clean_dism1990 + lenper, data = md2)

stargazer(reg_insb, reg_insw, reg_ss, reg_ss2,
          se = list(reg_insb$rse, reg_insw$rse, reg_ss$rse, reg_ss2$rse),
          header = FALSE, no.space = TRUE, omit.stat = c("ser"),
          title = "2SLS vs IV")
```

Table 9: 2SLS vs IV

	Dependent variable:			
	povrate_b	povrate_w	povrate_b	povrate_w
	felm	*felm*	*OLS*	*OLS*
	(1)	(2)	(3)	(4)
clean_dism1990			0.258*	−0.196***
			(0.148)	(0.061)
lenper	−4.780	0.602	−4.780	0.602
	(3.067)	(1.970)	(7.153)	(2.961)
‘dism1990(fit)‘	0.258**	−0.196***		
	(0.108)	(0.065)		
Constant	0.121**	0.205***	0.121	0.205***
	(0.061)	(0.037)	(0.081)	(0.033)
Observations	121	121	121	121
R^2	0.084	−0.150	0.027	0.111
Adjusted R^2	0.068	−0.170	0.010	0.096
F Statistic (df = 2; 118)			1.629	7.336***
Note:				*p<0.1; **p<0.05; ***p<0.01

Answer:

The coefficients are exactly the same. This indicates that our 2SLS regression is equivalent to our IV regression. There is a slight difference in the standard errors due to the different procedures used to calculate each regression.

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

6.1 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:

For *povrate_w* we get $\frac{0.092}{0.357} = 0.258$ and for *povrate_b* we get $\frac{-0.070}{0.357} = -0.196$ which exactly matches the estimates form our main result.

7 Submission instructions:

1) Knit your assignment in PDF.
2) Make sure you have ONE question and answer per page unless a question spans two pages where noted (this allows gradescope to easily find your answers). It should be 24 pages.
3) Upload your assignment PDF to gradescope.