Instrumental Variable Analysis

Maxwell Snodgrass

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.

1 Set up and opening the data

Code:

```
#Load libraries

library(haven)
library(dplyr)
```

```
##
## Attaching package: '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 Statistics Tables.

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

```
library(lfe)
```

```
## Loading required package: Matrix
```

```
library(ggplot2)
```

```
#Load data

aej<- read_dta("aej_maindata.dta")
```

```
#Is it dataframe type?

is.data.frame(aej)
```

```
## [1] TRUE
```

1.1 The dataset contains many variables, some of which we do not need for this analysis. Below is a table of the relevant variables and their description.

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

Code:

```
aej_final <- aej %>% 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 First we will look at some basic information about our data. We want to know the total number of observations and the observation type. Each observation represents city characteristics. Each city is its own observation and attached to it are many different characteristics, including geography statistics, demographics by year, poverty information, and others.

Code:

```
nrow(aej_final)
```

```
## [1] 121
```

2.2 Now we will use stargazer to create a summary statistics table about the most relevant variables. We have chosen the dissimilarity index for 1990 (dism1990), the railroad division index (herf), the length of the railroad track (lenper), and the poverty rates of blacks and whites. We will use the dism1990 as our explanatory variable, lenper and herf as our instruments, and the poverty rate as our outcome variables.

Code:

```
aej_summary <- aej %>% select(dism1990,herf,lenper,povrate_w,povrate_b)
```

```
aej_summary = as.data.frame(aej_summary)
```

```
stargazer(aej_summary, type = "latex", title = "Summary Statistics",
          summary.stat = c("n", "mean", "sd", "min", "p25", "median", "p75", "max"))
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Thu, Mar 13, 2025 - 8:13:51 PM

Table 2: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
dism1990	121	0.569	0.135	0.329	0.457	0.574	0.673	0.873
herf	121	0.723	0.141	0.238	0.638	0.742	0.830	0.987
lenper	121	0.001	0.001	0.0002	0.0004	0.001	0.001	0.013
povrate_w	121	0.095	0.035	0.035	0.069	0.085	0.114	0.216
povrate_b	121	0.264	0.080	0.093	0.209	0.264	0.313	0.504

3 Reduced Form:

3.1 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. First, to find the reduced form, we will regress these two outcome variables on segregation in 1990, our explanatory variable, and interpret your results.

Code:

```
reg1<-felm(povrate_w~dism1990,data=aej_final)
reg2<-felm(povrate_b~dism1990,data=aej_final)
```

```
stargazer(reg1,reg2, type = "latex", se=list(reg1$rse,reg2$rse),
          header=FALSE, title="White vs Black Poverty Regressed on
          Segregation (1990)",single.row =TRUE
)
```

Table 3: White vs Black Poverty Regressed on Segregation (1990)

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

3.2 The table above shows that a one standard deviation increase in the segregation index is associated with a (0.14 * (-0.073) = -0.0102), one percentage point decrease in white poverty and a (0.14 * (0.182) = 0.025), 2.5 percentage point increase in black poverty. Both are statistically significant at the 95% level.

3.3 You cannot give a causal interpretation of this regression because there are omitted variables that affect the outcome variable and the explanatory variable. Some of these omitted variables include political corruption, legacy of a manufacturing economy, and historical laws that may have affected both the poverty rate and segregation index.

4 Validity of the instrument:

4.1 The two conditions that are necessary for a valid instrument would be succeeding the first stage and the exclusion restriction.

4.2 First we will run the first stage of the two stage least squares regression. This stage regresses our explanatory variable on our instrumental variables. It is used to test the validity of our instrumental variable on our predictor variable. If this regression yields insignificant results, then our instrument fails the first stage and should not be used. Instrumental variables are required to be correlated with the predictor variable.

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

Code:

```
reg3<-felm(dism1990~herf+lenper,data=aej_final)
```

```
stargazer(reg3, type = "latex", se=list(reg3$rse),
          header=FALSE, title="Track Characteristics and Segregation",single.row =TRUE
)
```

Table 4: Track Characteristics and Segregation

	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

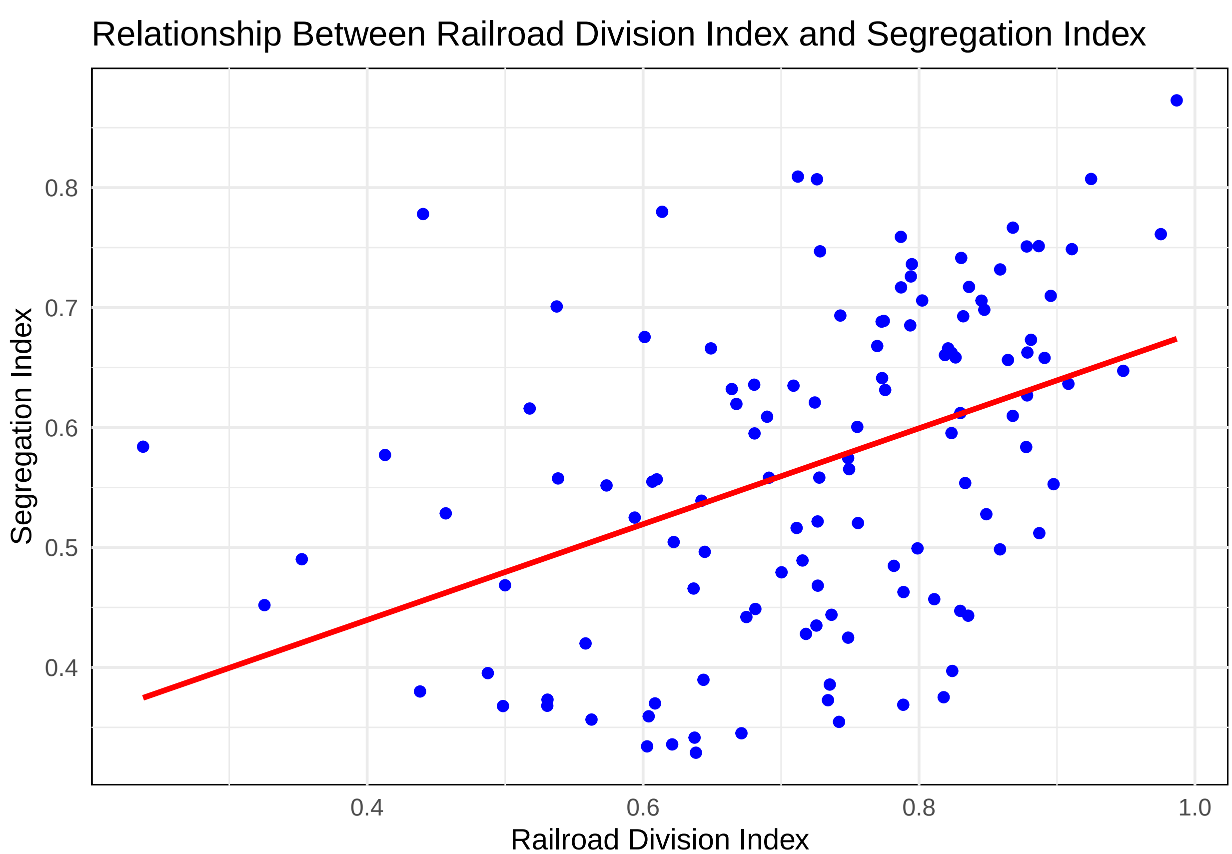
4.3 The table above shows a one standard deviation increase in the RDI indicates a (0.14*0.357=0.049) 5 percentage point increase in the segregation index (1990). A one unit increase in track length per square km indicates an 18.514 increase in the segregation index 1990. Both variables are significant at the 90% level.

4.4 We can use scatterplots to showcase the correlated relationship between our instruments and our explanatory variable.

Code:

```
ggplot(aej_final, aes(x = herf, y = dism1990)) +
  geom_point(color = "blue") +
  theme_minimal() +geom_smooth(method = "lm", se = FALSE, color = "red")+
  labs(title = "Relationship Between Railroad Division Index and Segregation Index",
       x = "Railroad Division Index",
       y = "Segregation Index")+
  theme(panel.background = element_rect(fill = "white"))
```

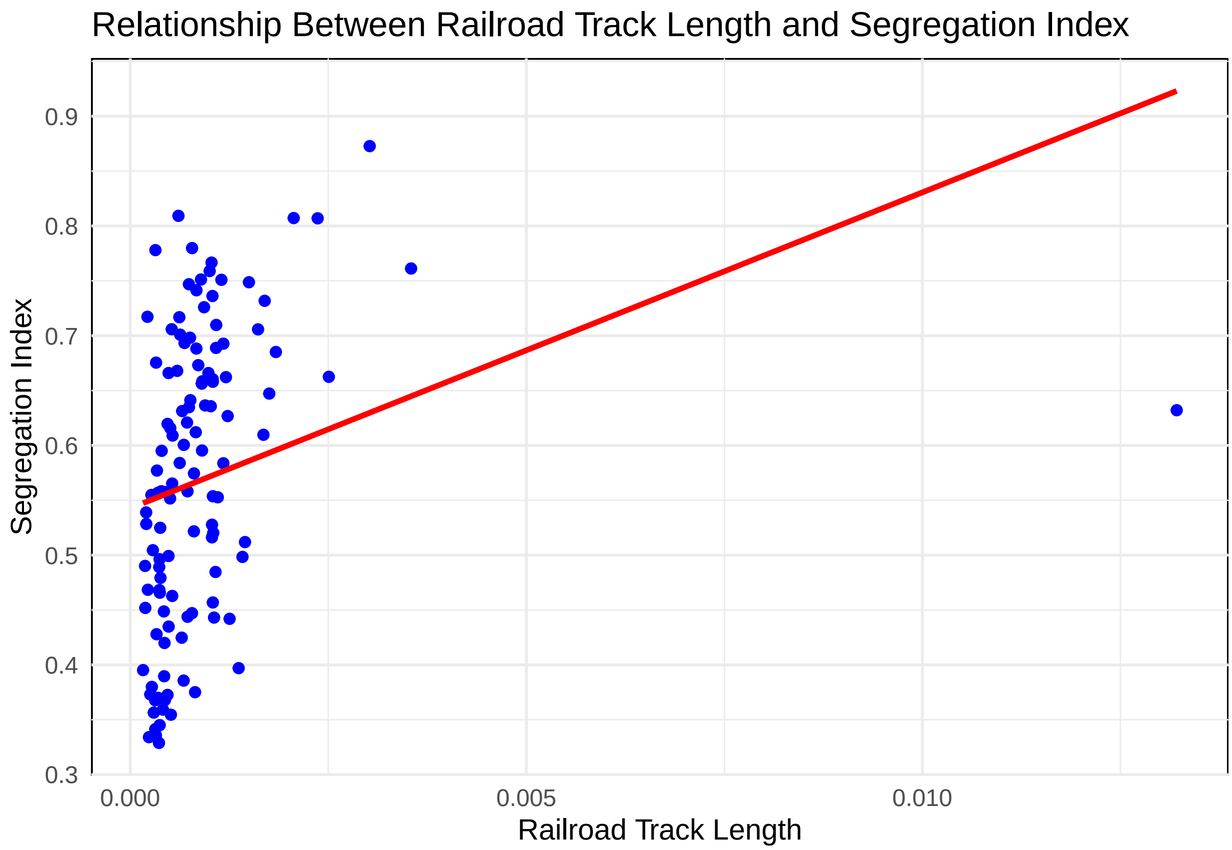
```
## `geom_smooth()` using formula = 'y ~ x'
```



Code:

```
ggplot(aej_final, aes(x = lenper, y = dism1990)) +
  geom_point(color = "blue") +
  theme_minimal() +geom_smooth(method = "lm", se = FALSE, color = "red")+
  labs(title = "Relationship Between Railroad Track Length and Segregation Index",
       x = "Railroad Track Length",
       y = "Segregation Index")+
  theme(panel.background = element_rect(fill = "white"))
```

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



4.5 Continuing with the analysis of our instrument, we can also directly calculate the correlation between it and the explanatory variable.

Code:

```
cor(aej_final$herf,aej_final$dism1990)
```

```
## [1] 0.417972
```

4.6 When determining if an instrument is weak, we must look at the correlation between the instrumental variable and the predictor variable. The correlation here is .42. This indicates a moderate positive correlation between the variables. Since the correlation is not strong, this could be a potential issue. A weak instrument would mean that our estimates for our predictor variable will not be as accurate and may be biased by an unknown variable. Nonetheless, a moderate correlation is still acceptable to determine causality. The F-statistic is also used to detect the strength of an instrument. The F-statistic for the first stage model is greater than 10, the benchmark used to detect weak instruments so the weak instrument problem likely does not apply here.

4.7 We continue our analysis of the instrumental variable by looking at the exclusion restriction. This is the relationship of our instrument with other explanatory variables. A good instrument should not be correlated with any of these variables. We will regress the following characteristics on the RDI and track length: `area1910` `count1910`, `black1910`, `incseg`, `lfp1920`.

Code:

```
reg4<-felm(area1910~herf+lenper,data=aej_final)
reg5<-felm(count1910~herf+lenper,data=aej_final)
reg20<-felm(black1910~herf+lenper,data=aej_final)
reg21<-felm(incseg~herf+lenper,data=aej_final)
reg22<-felm(lfp1920~herf+lenper,data=aej_final)
```

```
stargazer(reg4,reg5, type = "latex", se=list(reg4$rse,reg5$rse),
          header=FALSE, omit.stat=c( "ser"),
          title="Track Characteristics and Segregation",single.row =TRUE
)
```

Table 5: Track Characteristics and Segregation

	Dependent variable:	
	area1910	count1910
	(1)	(2)
herf	−3,992.637 (11,986.490)	665.751 (1,362.964)
lenper	−574,401.000 (553,669.000)	75,553.190 (134,814.900)
Constant	18,409.570** (8,612.320)	976.876 (927.189)
Observations	58	121
R^2	0.007	0.006
Adjusted R^2	−0.029	−0.011
Note:		*p<0.1; **p<0.05; ***p<0.01

```
stargazer(reg20,reg21,reg22, type = "latex", se=list(reg20$rse,reg21$rse,
                                                      reg22$rse),
          header=FALSE, omit.stat=c( "ser"),
          title="Track Characteristics and Segregation Part 2",single.row =TRUE
)
```

4.8 Our output shows that there is an insiginificant relationship between the instruments and these other variables. The exclusion restriction is fulfilled and we can continue.

4.9 I believe the instrument is valid since the railroads were built for manufacturing purposes and do not have a relationship with omitted variables such as population percentages. Railroads were built before the Great migration and unlike highways, they were not built to intentionally divide neighborhoods.

Table 6: Track Characteristics and Segregation Part 2

	Dependent variable:		
	black1910	incseg	lfp1920
	(1)	(2)	(3)
herf	−0.001 (0.010)	0.032 (0.032)	0.028 (0.024)
lenper	9.236*** (0.650)	−2.504 (1.626)	−3.427** (1.500)
Constant	0.007 (0.007)	0.196*** (0.025)	0.401*** (0.018)
Observations	121	69	121
R^2	0.290	0.028	0.015
Adjusted R^2	0.278	−0.001	−0.002
Note:			*p<0.1; **p<0.05; ***p<0.01

5 Regression Analysis

5.1 Now we will use regression analysis to estimate the causal effect of segregation on the poverty rate of blacks and whites. First, we will compare two regressions. One that simply uses OLS between segregration and poverty and one that uses the instrumental variables.

Code:

```
reg1<-felm(povrate_w~dism1990,data=aej_final)
reg2<-felm(povrate_b~dism1990,data=aej_final)

iv_model1 <- felm(povrate_w ~ lenper | 0 | (dism1990 ~ herf), data = aej_final)
iv_model2 <- felm(povrate_b ~ lenper | 0 | (dism1990 ~ herf), data = aej_final)
```

```
stargazer(reg1,reg2,iv_model1,iv_model2, type = "latex",
          se=list(reg1$rse,reg2$rse,iv_model1$rse,
                                                        iv_model2$rse),
          header=FALSE, omit.stat=c( "ser"),
          title="Effects of Segregation on Poverty",single.row =TRUE
)
```

5.2 We can see how the use of the RDI instrument changes the estimated coefficients. The stronger effect suggests that OLS might have underestimated the impact of segregation on poverty rates due to endogeneity.

Table 7: Effects of Segregation on Poverty

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

5.3 Before analyzing the output, we must check the reduced form which represents the instrumental variables direct relationship with the target variable.

Answer:

$$povrate_i = \beta_0 + \beta_1 RDI_i + \beta_2 tracklength_i + \eta$$

5.4 For the two poverty rates, we will estimate the reduced form on all the cities and illustrate the reduced form relationships graphically.

Code:

```r
reduced1<-felm(povrate_w~herf,data=aej_final)
reduced2<-felm(povrate_b~herf,data=aej_final)
```

```r
stargazer(reduced1,reduced2, type = "latex", se=list(reduced1$rse,reduced2$rse),
          header=FALSE, omit.stat=c( "ser"),
          title="Effects of Segregation on Poverty",single.row =TRUE
)
```

Table 8: Effects of Segregation on Poverty

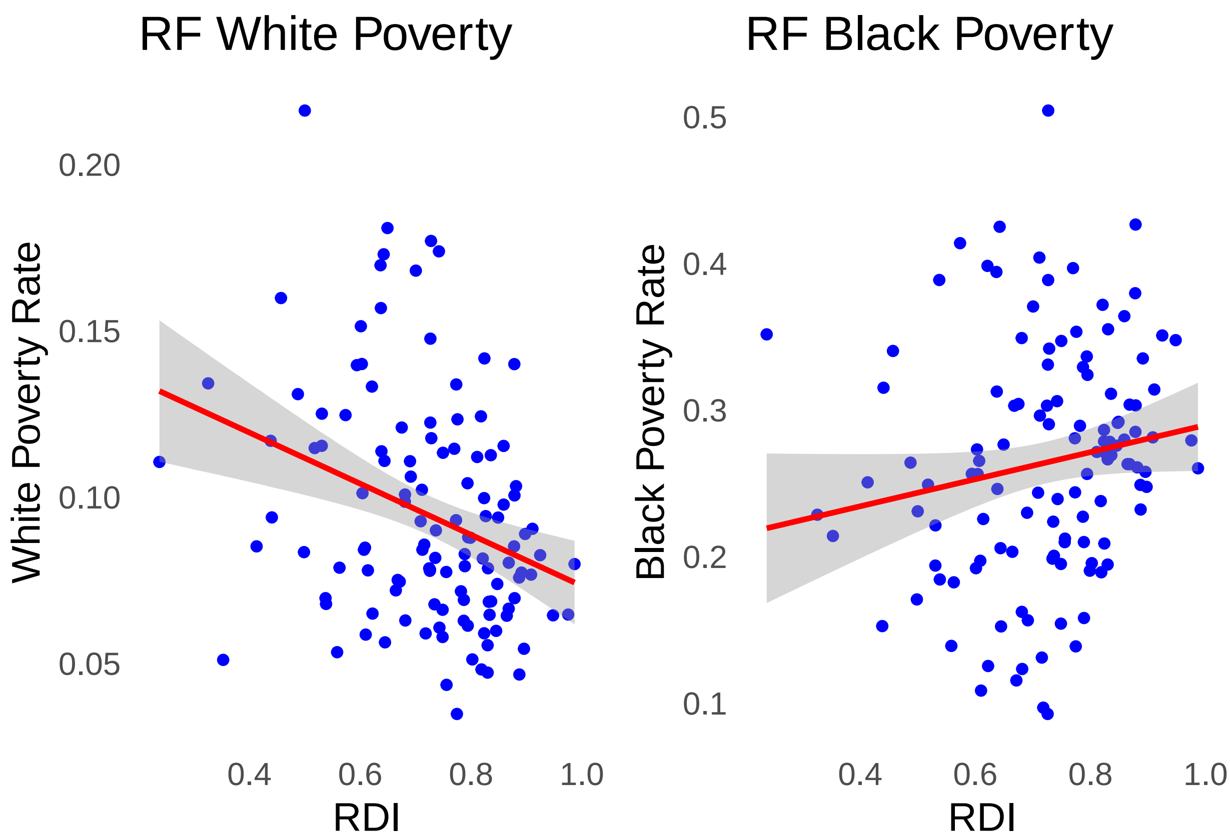
	Dependent variable:	
	povrate_w	povrate_b
	(1)	(2)
herf	-0.077^{***} (0.022)	0.092^{**} (0.046)
Constant	0.150^{***} (0.017)	0.197^{***} (0.036)
Observations	121	121
R^2	0.099	0.027
Adjusted R^2	0.092	0.019
Note:		*p<0.1; **p<0.05; ***p<0.01

```r
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
plot1 <- ggplot(aej_final, aes(x = herf, y = povrate_w)) +
  geom_point(color = "blue") +                # Blue points
  geom_smooth(method = "lm", color = "red") +  # Line of best fit in red
  labs(title = "RF White Poverty",y = "White Poverty Rate",x="RDI") +                            # Lab
  theme_minimal(base_size = 15) +              # White background
  theme(panel.grid = element_blank())

# Create the second scatter plot for y2
plot2 <- ggplot(aej_final, aes(x = herf, y = povrate_b)) +
  geom_point(color = "blue") +                # Blue points
  geom_smooth(method = "lm", color = "red") +  # Line of best fit in red
  labs(title = "RF Black Poverty",y = "Black Poverty Rate",x="RDI") +                            # Lab
  theme_minimal(base_size = 15) +              # White background
  theme(panel.grid = element_blank())

# Arrange the two plots side by side
grid.arrange(plot1, plot2, ncol = 2)
```

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



5.5 **We see a positive relationship with the railroad index and black poverty rate and a negative correlation between railroad index and white poverty rate.**

5.6 Now we generate a table with six columns that check whether the main results are robust to adding additional controls for city characteristics.

Code:

```
robust1w<-felm(povrate_w ~ lenper + pctbk1990 | 0 | (dism1990 ~ herf), data = aej_final)
robust2w<-felm(povrate_w ~ lenper + manshr | 0 | (dism1990 ~ herf), data = aej_final)
robust3w<-felm(povrate_w ~ lenper + incseg| 0 | (dism1990 ~ herf), data = aej_final)

robust1b<-felm(povrate_b ~ lenper + pctbk1990 | 0 | (dism1990 ~ herf), data = aej_final)
robust2b<-felm(povrate_b ~ lenper + manshr | 0 | (dism1990 ~ herf), data = aej_final)
robust3b<-felm(povrate_b ~ lenper + incseg| 0 | (dism1990 ~ herf), data = aej_final)
```

```
stargazer(robust1w,robust2w,robust3w,robust1b,robust2b,robust3b,
          type = "latex", se=list(robust1w$rse,robust2w$rse,
                                                                                                    robust3w$rse,rob
                            robust3b$rse),
          header=FALSE, omit.stat=c( "ser"),
          title="Robustness Checks",single.row =TRUE
)
```

Table 9: Robustness Checks

	Dependent variable:					
	povrate_w			povrate_b		
	(1)	(2)	(3)	(4)	(5)	
lenper	−0.479 (1.801)	2.007 (3.506)	−0.120 (0.568)	−2.331 (2.402)	−4.772 (5.669)	−5.62
pctbk1990	0.211 (0.153)			−0.478* (0.246)		
manshr		0.108 (0.125)			−0.013 (0.231)	
incseg			0.179* (0.097)			−0.
‘dism1990(fit)‘	−0.241** (0.097)	−0.272** (0.124)	−0.107** (0.053)	0.360** (0.141)	0.219 (0.195)	0.478
Constant	0.219*** (0.048)	0.230*** (0.049)	0.112*** (0.039)	0.091 (0.068)	0.149** (0.074)	0.0
Observations	121	111	69	121	111	
R^2	−0.254	−0.319	−0.035	0.108	0.065	
Adjusted R^2	−0.286	−0.356	−0.083	0.085	0.038	
Note:					*p<0.1; **p<0.0	

5.7 Most of these control variables are correlated with the racial poverty rates besides a small significant relationship between incseg and white poverty rates. The true effect of dism1990 does not change that much in each regression and stays relatively close to our estimates of -.196 for whites and .258 for blacks. The betas for dism1990 in all regressions contain our estimates within their 95% confidence intervals.

5.8 Now we will estimate the first stage regression and use the estimates to generate the predicted values for the explanatory variable for all the observations.

Code:

```r
first_stage <- felm(data=aej_final,dism1990~herf)
summary(first_stage)
```

```
##
## Call:
##    felm(formula = dism1990 ~ herf, data = aej_final)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.23135 -0.10322  0.00791  0.08834  0.32227
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.27970    0.05866   4.768 5.32e-06 ***
## herf         0.39954    0.07961   5.019 1.84e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1234 on 119 degrees of freedom
## Multiple R-squared(full model): 0.1747   Adjusted R-squared: 0.1678
## Multiple R-squared(proj model): 0.1747   Adjusted R-squared: 0.1678
## F-statistic(full model):25.19 on 1 and 119 DF, p-value: 1.84e-06
## F-statistic(proj model): 25.19 on 1 and 119 DF, p-value: 1.84e-06
```

```r
hatgamma0<-first_stage$coefficients[1]
hatgamma1<-first_stage$coefficients[2]

aej_final$prediction<-hatgamma0+hatgamma1*aej_final$herf
```

5.9 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 we run the second stage by regressing our outcome variable on the predicted values generated above and the relevant controls (lenper).

5.10 If the regression coefficient for predicted values turns out to be statistically significant, then we can say that there is a causal effect of the segregation index on racial poverty rate.

Code:

```
second_stagew<-felm(data=aej_final,povrate_w~prediction+lenper)
second_stageb<-felm(data=aej_final,povrate_b~prediction+lenper)
```

```
stargazer(second_stageb,second_stagew, type = "latex", se=list(second_stageb$rse,second_stagew$rse),
          header=FALSE, omit.stat=c( "ser"), title="Second Stage Regressions",single.row =TRUE
)
```

Table 10: Second Stage Regressions

	Dependent variable:	
	povrate_b	povrate_w
	(1)	(2)
prediction	0.231* (0.120)	−0.175*** (0.054)
lenper	0.004 (4.398)	−3.022*** (1.011)
Constant	0.133* (0.069)	0.197*** (0.032)
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

5.11 The regression coefficients are significant and this indicates a causal relationship between segregation and racial poverty rate.