Reproducing Reasearch: Causal Impact of Segregation on Poverty Rates

This project 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.

Keep the following variables in the final dataset.

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.

Setup

Code:

```
library(dplyr)
library(stargazer)
library(lfe)
library(ggplot2)
library(haven)
library(huxtable)
library(kableExtra)
library(stringr)
library(AER)
basic<- read_dta('C:\\Users\\roryq\\Downloads\\aej.dta')
```

```
df<-basic %>% select(dism1990,herf,lenper,povrate_w,povrate_b,area1910,count1910,ethiso10
                     ,ethseg10,black1910,passpc,black1920,lfp1920,incseg,pctbk1990
                     ,manshr,pop1990, name)
```

Data description:

Each observation is a city. The row observation contains information in multiple years, ie one city is a row and that row has multiple columns that cover data from different years.

Report summary statistics of the following variables in the dataset:"dism1990", "herf", "lenper", "povrate_w", "povrate_b".

Code:

```
s <- df %>% select(dism1990, herf, lenper, povrate_w, povrate_b)
s <- s %>%
  mutate(across(everything(), ~ as.numeric(as.character(.))))

summary_stats <- s %>%
  summarise(
    Mean = colMeans(., na.rm = TRUE),
    SD = sapply(., sd, na.rm = TRUE),
    Min = sapply(., min, na.rm = TRUE),
    Max = sapply(., max, na.rm = TRUE)
  )

summary_stats_with_rownames <- cbind(Row = c("dism1990", "herf", "lenper", "povrate_w",
  "povrate_b"), summary_stats)

# Convert the updated dataframe with row names to a huxtable
kable_summary <- knitr::kable(summary_stats_with_rownames, format = "latex", booktabs = TRUE,
  col.names = c("Variable", "Mean", "SD", "Min", "Max"))

# Apply kableExtra functions for styling
kable_summary %>%
  kable_styling(latex_options = c("striped", "hold_position")) %>%
  column_spec(1, bold = TRUE) %>%
  column_spec(2:ncol(summary_stats_with_rownames), width = "3cm") %>%
  row_spec(0, bold = TRUE, color = "white", background = "#2c3e50")
```

Variable	Mean	SD	Min	Max
dism1990	0.5686943	0.1352301	0.3288715	0.8727629
herf	0.7233133	0.1414675	0.2375684	0.9867913
lenper	0.0009014	0.0012602	0.0001622	0.0132102
povrate_w	0.0945234	0.0345104	0.0347802	0.2161613
povrate_b	0.2641034	0.0796990	0.0925764	0.5042186

‘

Reduced Form:

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.

Code:

```
model_b<-felm(povrate_b~dism1990,data=df)
model_w<-felm(povrate_w~dism1990,data=df)

stargazer(model_w,model_b,
          se = list(model_w$rse, model_b$rse),  # Robust standard errors
          type = "latex",                       # Output LaTeX table
          dep.var.labels = c( "Poverty Rate (White)","Poverty Rate (Black)"),
          covariate.labels = c("Segregation (1990)"),
           omit.stat = c("f", "ser", "adj.rsq","rsq","n"))
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Fri, Nov 15, 2024 - 9:17:53 PM

Table 2:

	Dependent variable:	
	Poverty Rate (White)	Poverty Rate (Black)
	(1)	(2)
Segregation (1990)	-0.073^{***}	0.182^{***}
	(0.019)	(0.045)
Constant	0.136^{***}	0.161^{***}
	(0.012)	(0.029)
Note:		*p<0.1; **p<0.05; ***p<0.01

Answer:

For an increase of one standard deviation of segregation, the poverty rate of white people decreases by approximately 1.02%.

For an increase of one standard deviation of segregation, the poverty rate of black people increases by approximately 2.5%.

Explain the problem with giving a causal interpretation to the estimates we just produced. Give examples of specific confounds that might make a causal interpretation of our result problematic.

Answer:

We can not interpret this as causal because there are a lot of omitted variables that could also affect poverty rate and is related to the segregation of the city. For example city size could be an omitted variable where larger cities have more infrastructure and are therefore more segregated, but also have higher poverty rates because of the wealth inequality with high poverty in low class neighborhoods.

Another possible confounder could be education and policy. There have long been policy the excludes black people from attaining education and the same opportunities as their white counterparts. Therefore the higher segregation is correlated with less education/opportunity for black people which contributes to the income inequality and poverty rates, because of the strong relationship between education and income.

Validity of the instrument:

Estimate the following regression and interpret it's coefficients,

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

Code:

```
model_I <- felm(dism1990 ~ herf+ lenper, data = df)

model.sum <- summary(model_I)


stargazer(model_I,
          se = list(model_I$rse),  # Robust standard errors
          type = "latex",                       # Output LaTeX table
          dep.var.labels = c( "Segregation (1990)"),
          covariate.labels = c("RDI", "Track length per square km"),
           omit.stat = c("f", "ser", "adj.rsq","rsq","n"))
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Fri, Nov 15, 2024 - 9:17:54 PM

Table 3:

	Dependent variable:
	Segregation (1990)
RDI	0.357***
	(0.088)
Track length per square km	18.514*
	(10.731)
Constant	0.294***
	(0.064)
Note:	*p<0.1; **p<0.05; ***p<0.01

Answer:

For an increase of one standard deviation of RDI, segregation 1990 increases by approximately 5%.

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

Answer:

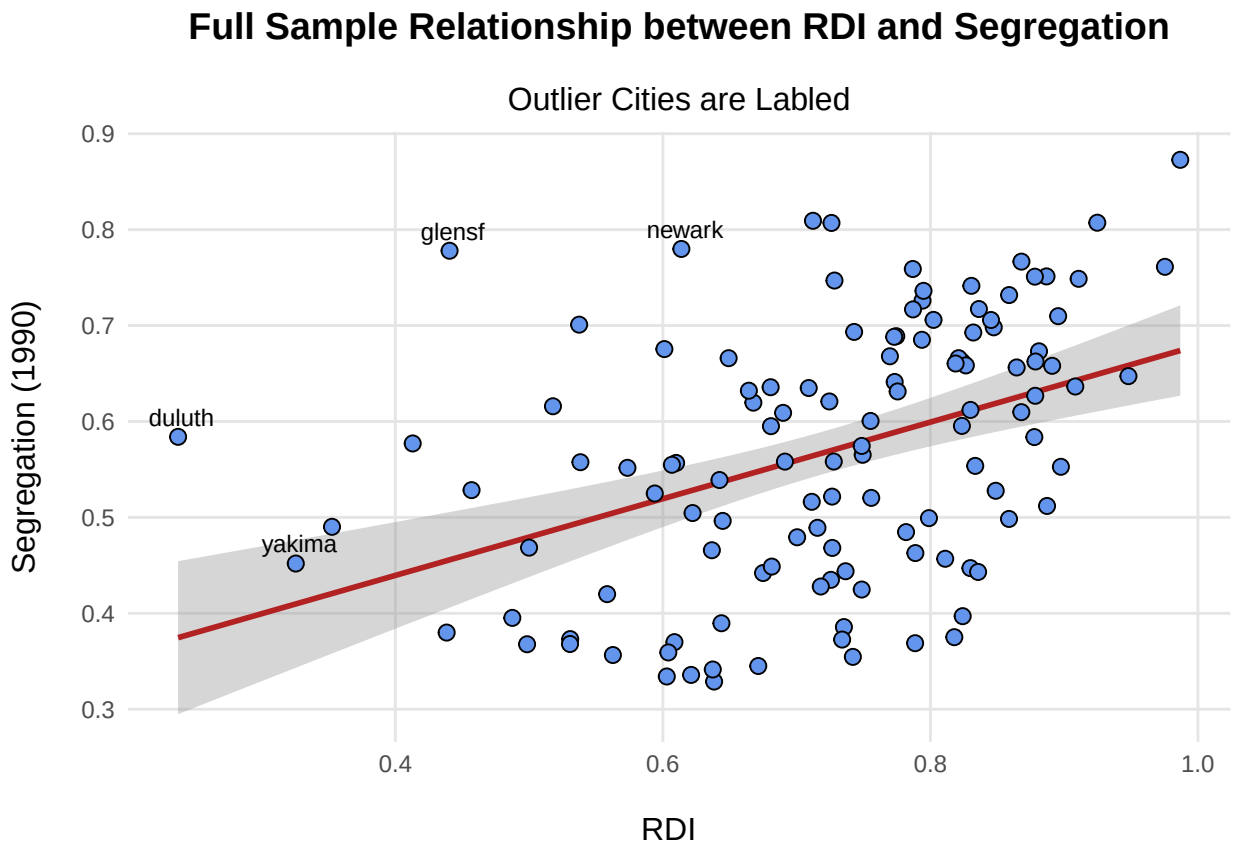
This regression is the first stage, it 'cleans' the variation in the explanatory variable to allow for causal inference, and also establishes that there is a relationship between the instrument and explanatory variable.

Illustrate the relationship between the RDI and segregation graphically.

Code:

```
model <- lm(dism1990 ~ herf, data = df)

# Calculate residuals
df$residuals <- residuals(model)

# Calculate the IQR for the 'herf' variable (explanatory variable)
Q1 <- quantile(df$herf, 0.25)
Q3 <- quantile(df$herf, 0.75)
IQR_value <- Q3 - Q1

# Define the lower and upper bounds for IQR outliers
lower_bound_iqr <- Q1 - 1.5 * IQR_value
upper_bound_iqr <- Q3 + 1.5 * IQR_value

# Identify IQR outliers (based on herf variable)
df$outlier_iqr <- df$herf < lower_bound_iqr | df$herf > upper_bound_iqr

# Identify residual outliers (based on the model residuals)
outlier_threshold_residuals <- 2 * sd(df$residuals)  # You can adjust this threshold
df$outlier_residuals <- abs(df$residuals) > outlier_threshold_residuals

# Combine both outlier flags (IQR and residuals)
df$outlier_combined <- df$outlier_iqr | df$outlier_residuals
```

```
ggplot(df, aes(x=herf,y=dism1990))+ geom_point() + # Scatter plot points
  labs(title = "Full Sample Relationship between RDI and Segregation",
       x = "RDI",
       y = "Segregation (1990)",
       subtitle = "Outlier Cities are Labled") +
  theme_minimal()+
  geom_smooth(method = "lm", se = TRUE, color = "firebrick")+
  geom_point(color='black', fill='cornflowerblue', shape=21, size=2.5)+
  theme(plot.title = element_text(hjust = 0.49,margin = margin(b = 15),size=14))+
    theme(
    plot.title = element_text(hjust = 0.5,face = "bold"),  # Center the title
    panel.grid.major = element_line(color = "gray90"),
    panel.grid.minor = element_blank(),
    panel.border = element_blank(),
    axis.title.x = element_text(margin = margin(t = 15),hjust = 0.49, size=11.5),
    axis.title.y = element_text(margin = margin(r = 15),hjust = 0.5,size=11.5),
    plot.subtitle = element_text(size = 11.5, color = "black", hjust = 0.5)
  )+

    geom_text(aes(label = ifelse(outlier_combined, str_sub(name, 1, nchar(name) - 2), "")),
            vjust = -0.65, hjust = 0.45, size = 3, color = "black")
```



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

Answer:

The F stat value of the model is 14.9827197. This indicates that there is a not weak first stage. The F stat larger than 10 indicates that this is a strong instrument.

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?

Code and Answer:

```
model_A<- felm(area1910~herf+lenper,data=df)
model_B<- felm(count1910~herf+lenper,data=df)
model_C<- felm(black1910~herf+lenper,data=df)
model_D<- felm(incseg~herf+lenper,data=df)
model_E<- felm(lfp1920~herf+lenper,data=df)

stargazer(model_A,model_B,model_C,model_D,model_E,omit.stat=c( "ser"),
          se = list(model_A$rse, model_B$rse,model_C$rse,model_D$rse,model_E$rse),
          type = "latex",
          covariate.labels = c("RDI", "Track length per square km"))
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Fri, Nov 15, 2024 - 9:17:56 PM

Table 4:

	Dependent variable:				
	area1910	count1910	black1910	incseg	lfp1920
	(1)	(2)	(3)	(4)	(5)
RDI	−3,992.637	665.751	−0.001	0.032	0.028
	(11,986.490)	(1,362.964)	(0.010)	(0.032)	(0.024)
Track length per square km	−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

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:

Condition 1, there has to be a relationship between the control and the instrument (preferrably more than just a weak relationship, if you want a strong instrument). Condition 2 is the exclusion restriction, the instrument shouldnt be correlated with any other omitted variables that affect the dependent variable as well.

Since the tracks were laid down previously to the city being drawn, and the tracks were randomly laid out according to ease of transportation. Therefore the tracks randomness by coming before the urbanization ensures the exclusionary restriction. From the first stage regression we can also see that there is a relationship between the instrument and segregation, fulfilling condition 1.

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

Answer:

Yes, because of the above conditions above that are explained, and I dont think there are any violations of the exclusion restriction that would invalidate the instrument.

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?

These are the exact results reported in row 2 of columns 1-4 in table 2.

Code and Answer:

```
# Model without IV
model_WP<- felm(povrate_w~dism1990,data=df)
model_BP<-felm(povrate_b~dism1990,data=df)


# Model with IV
model_WI<- felm(povrate_w ~ lenper |0| (dism1990~herf+lenper), data = df)


model_BI<-felm(povrate_b ~ lenper |0| (dism1990~herf+lenper), data = df)
# Table output
stargazer(model_WP,model_BP,model_WI,model_BI,se = list(model_WP$rse, model_BP$rse,model_WI$rse
                                                        ,model_BI$rse),  # Robust standard errors
          type = "latex",
          covariate.labels = c("Dissimilarity"),omit.stat=c( "ser"))
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Fri, Nov 15, 2024 - 9:17:56 PM

Table 5:

	Dependent variable:			
	povrate_w	povrate_b	povrate_w	povrate_b
	(1)	(2)	(3)	(4)
Dissimilarity	−0.073***	0.182***		
	(0.019)	(0.045)		
lenper			0.602	−4.780
			(1.970)	(3.067)
‘dism1990(fit)‘			−0.196***	0.258**
			(0.065)	(0.108)
Constant	0.136***	0.161***	0.205***	0.121**
	(0.012)	(0.029)	(0.037)	(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

The use of IV increases the coefficients (and robust standard errors) compared to the original regression.

What is the reduced form equation?

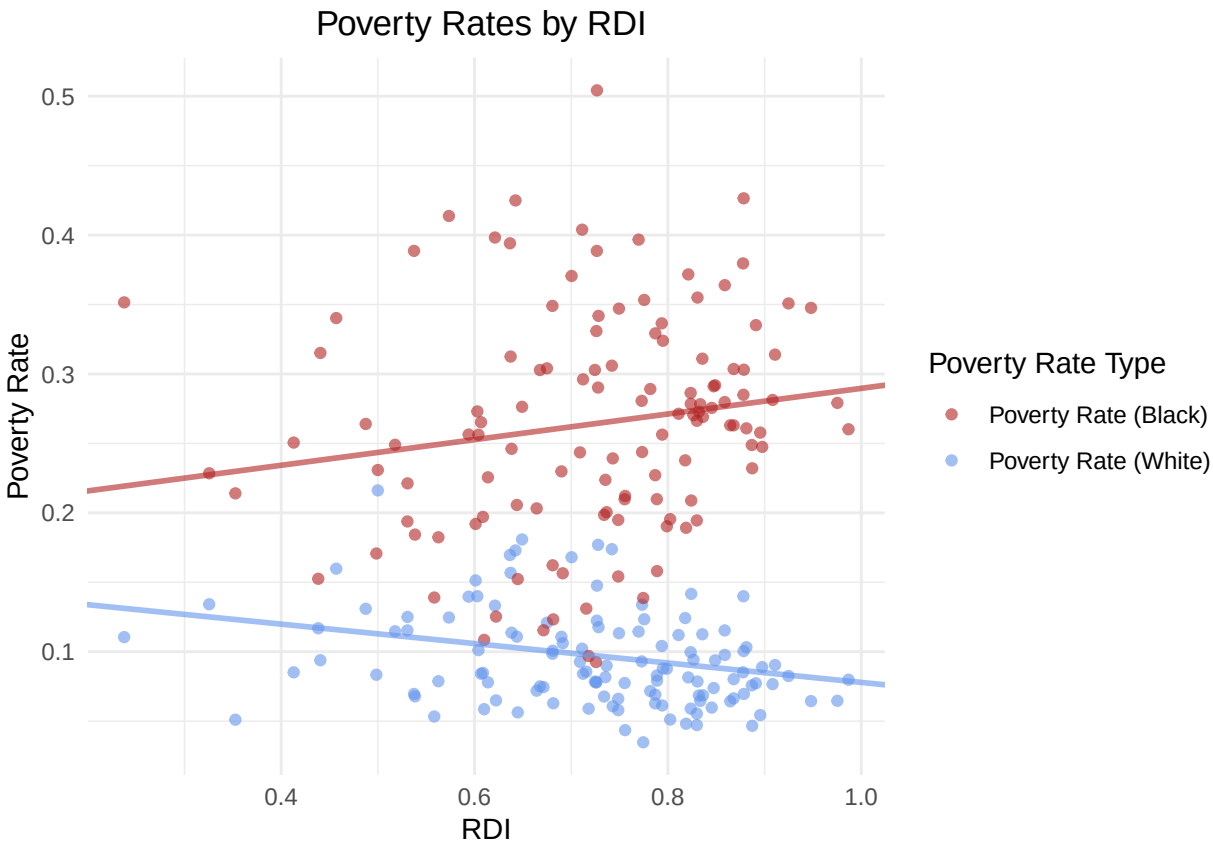
Answer:

$$\text{Reduced Form: } Y = \beta_0 + \beta_1 RDI_i + \beta_2 lenper + \epsilon_i$$

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

Code:

```
model_WIR<- felm(povrate_w~herf+lenper,data=df)
model_BIR<-felm(povrate_b~herf+lenper,data=df)
df$povrate<-(df$povrate_w+df$povrate_b)/2



ggplot(df, aes(x = herf)) +
  # Scatter plot for povrate_w (white)
  geom_point(aes(y = povrate_w, color = "Poverty Rate (White)"), alpha = 0.6) +

  # Scatter plot for povrate_b (black)
  geom_point(aes(y = povrate_b, color = "Poverty Rate (Black)"), alpha = 0.6) +
   # Regression line for model_WIR (White)
  geom_abline(slope = coef(model_WIR)["herf"], intercept = coef(model_WIR)["(Intercept)"],
              color = "cornflowerblue", size = 1, alpha = 0.6) +

  # Regression line for model_BIR (Black)
  geom_abline(slope = coef(model_BIR)["herf"], intercept = coef(model_BIR)["(Intercept)"],
              color = "firebrick", size = 1, alpha = 0.6)  +
  labs(
    title = "Poverty Rates by RDI ",
    x = "RDI",
    y = "Poverty Rate"
  ) +
  scale_color_manual(values = c("Poverty Rate (White)" = "cornflowerblue", "Poverty Rate (Black)" = "fi
  theme_minimal() +
  guides(color = guide_legend(title = "Poverty Rate Type")) +  # Customize the legend title
  theme(
    plot.title = element_text(hjust = 0.5)  # Center the title
  )
```


Poverty Rates by RDI
Poverty Rate
0.5
0.4
0.3
0.2
0.1
0.4
0.6
0.8
1.0
RDI
Poverty Rate Type
Poverty Rate (Black)
Poverty Rate (White)

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?

You can choose the controls you want.

Code:

```
# baseline
model_WI<- felm(povrate_w ~ lenper |0| (dism1990~herf+lenper), data = df)
model_BI<-felm(povrate_b ~ lenper |0| (dism1990~herf+lenper), data = df)

# Add first control
m1<-felm(povrate_w ~ lenper+pop1990 |0| (dism1990~herf+lenper+pop1990), data = df)
m1b<-felm(povrate_b ~ lenper +pop1990 |0| (dism1990~herf+lenper+pop1990), data = df)

# Add second control
m2w<-felm(povrate_w ~ lenper+black1910  |0| (dism1990~herf+lenper+black1910), data = df)

m2b<-felm(povrate_b ~lenper+ black1910 |0| (dism1990~ herf+lenper+black1910), data = df)

stargazer(model_WI, model_BI,m1,m1b,m2w,m2b, omit.stat=c("ser"))
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Fri, Nov 15, 2024 - 9:17:58 PM

Table 6:

	Dependent variable:					
	povrate_w	povrate_b	povrate_w	povrate_b	povrate_w	povrate_b
	(1)	(2)	(3)	(4)	(5)	(6)
lenper	0.602	−4.780	0.665	−4.908	1.669	−2.170
	(3.368)	(6.940)	(3.421)	(6.921)	(3.959)	(8.151)
pop1990			0.000*	−0.000*		
			(0.000)	(0.000)		
black1910					−0.115	−0.282
					(0.191)	(0.393)
‘dism1990(fit)‘	−0.196***	0.258*	−0.212***	0.291*	−0.196***	0.258*
	(0.070)	(0.144)	(0.075)	(0.151)	(0.070)	(0.144)
Constant	0.205***	0.121	0.210***	0.111	0.206***	0.123
	(0.038)	(0.078)	(0.040)	(0.080)	(0.038)	(0.079)
Observations	121	121	121	121	121	121
R^2	−0.150	0.084	−0.172	0.101	−0.147	0.088
Adjusted R^2	−0.170	0.068	−0.202	0.078	−0.177	0.065
Note:						*p<0.1; **p<0.05; ***p<0.01

Answer:

Each regression has a similar effect of segragation on poverty rate (in white and black regressions). the coefficients of these regressions for dism1990, are all similar to eachother and with the confidence interval of the baseline regression (and all are statistically significant).

After adding these controls it appears that the original model without the controls is robust because the new regression β_{1i} is within the original confidence interval.

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

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

Code:

```
# Generate the predicted values from the first-stage regression
df$pred <- predict(lm(dism1990 ~ herf + lenper, data = df))
first<- felm(dism1990 ~ herf + lenper, data = df)
```

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 our estimates from this regression to those generated earlier. How do they compare?

Using robust standard errors

Code:

```
new_w<- felm(povrate_w~pred+lenper,data=df)
new_b<- felm(povrate_b~pred+lenper,data=df)
original_w<- felm(povrate_w ~ lenper |0| (dism1990~herf+lenper), data = df)
original_b<- felm(povrate_b ~ lenper |0| (dism1990~herf+lenper), data = df)

stargazer(original_w,new_w,original_b,new_b,se = list( original_w$rse,new_w$rse,original_b$rse,
                                                       new_b$rse), type="latex",
  column.labels = c("Original", "New", "Original", "New"),
  dep.var.labels = c("Poverty Rate White",
                     "Poverty Rate Black","Poverty Rate White", "Poverty Rate Black"),
  omit.stat = c("ser"),
  title = "Additional Controls")
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Fri, Nov 15, 2024 - 9:17:58 PM

Table 7: Additional Controls

	Dependent variable:			
	Poverty Rate White		Poverty Rate Black	
	Original	New	Original	New
	(1)	(2)	(3)	(4)
pred		-0.196^{***}		0.258^{*}
		(0.060)		(0.134)
lenper	0.602	0.602	-4.780	-4.780
	(1.970)	(1.516)	(3.067)	(5.578)
‘dism1990(fit)‘	-0.196^{***}		0.258^{**}	
	(0.065)		(0.108)	
Constant	0.205^{***}	0.205^{***}	0.121^{**}	0.121
	(0.037)	(0.034)	(0.061)	(0.075)
Observations	121	121	121	121
R^2	-0.150	0.111	0.084	0.027
Adjusted R^2	-0.170	0.096	0.068	0.010
Note:			*p<0.1; **p<0.05; ***p<0.01	

Answer:

predicting the segregation (dism1990) first and then regressing those predictions on the poverty lead to the exact same results as out initial instrumental regression with felm as seen in the above table.

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

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:

```
b<- coef(model_BIR)["herf"]/coef(first)["herf"]
w<- coef(model_WIR)["herf"]/coef(first)["herf"]
```

The quotient of the coefficients for black people is 0.2583901 and the quotient of the coefficients for white people is -0.1957495. These match the second stage regression coefficients for β_1 from the previous regression.