TEAM_PROJECT

Bryan Gutierrez

2025-02-19

```
# read the csv
data <- read.csv("/Users/bryangutierrez/Downloads/mobility-all.csv", stringsAsFactors = FALSE)
```

Research Question 1

```
# I asked chatgbt how to see what variables are good to see economic mobility it recomended doing a cor

# Load necessary library
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
# Select only numeric variables
numeric_data <- select(data, where(is.numeric))

# Compute correlation with Mobility
cor_results <- cor(numeric_data, use = "complete.obs")["Mobility",]

# Sort correlations in descending order
cor_results <- sort(cor_results, decreasing = TRUE)

# see top correlated variables
cor_results
```

```
##                  Mobility             Middle_class                  Commute
##                1.00000000               0.67078130               0.63653119
##            Social_capital            Teenage_labor             Test_scores
##                0.58536224               0.57605507               0.56895549
##                        ID                 Latitude                  Married
```

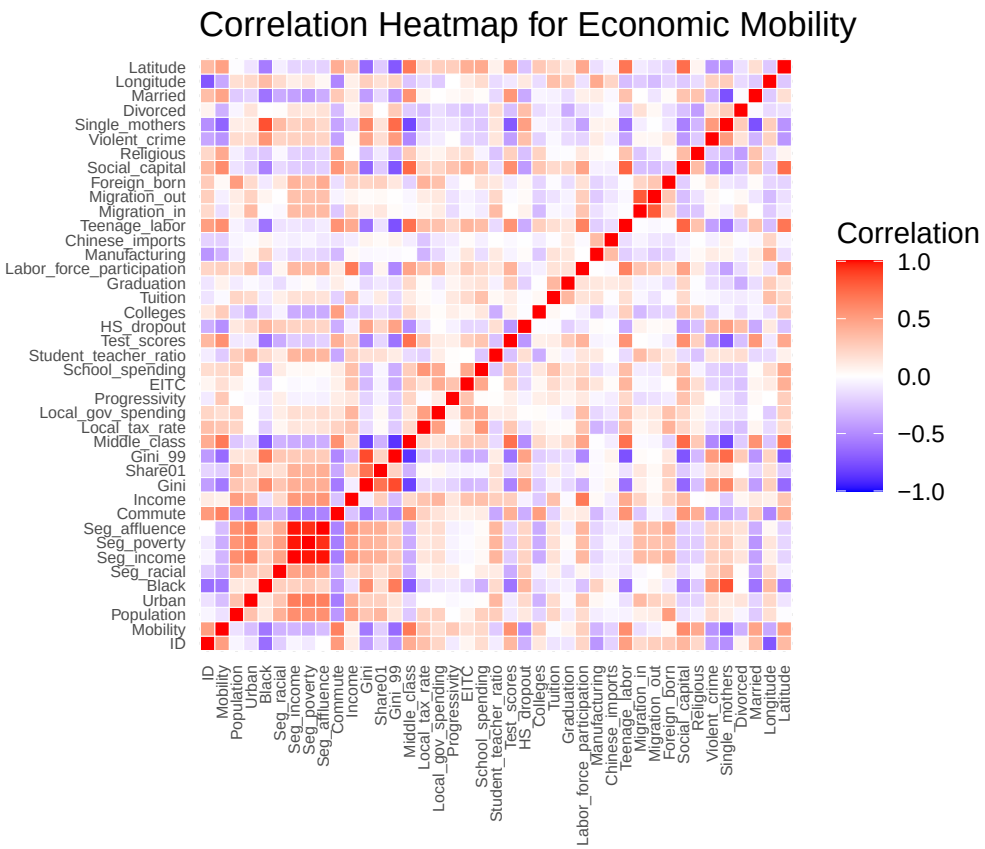
```
##               0.50135361               0.48667517               0.46996380
##                Religious           Local_tax_rate            Progressivity
##               0.44443928               0.30257849               0.28565059
##                 Colleges Labor_force_participation          School_spending
##               0.26397424               0.24801586               0.19973234
##       Local_gov_spending                     EITC                   Income
##               0.17458741               0.16576087               0.07473668
##               Graduation            Foreign_born                  Tuition
##               0.06964537               0.03113484              -0.02723655
##            Migration_out               Population             Migration_in
##              -0.06831542              -0.12555744              -0.14047598
##          Chinese_imports                  Share01    Student_teacher_ratio
##              -0.19900270              -0.21347536              -0.22411268
##                    Urban            Seg_affluence               Seg_income
##              -0.27911945              -0.28453759              -0.31810940
##            Manufacturing              Seg_racial                 Divorced
##              -0.32418869              -0.32910817              -0.33761641
##              Seg_poverty                Longitude            Violent_crime
##              -0.35267409              -0.37542674              -0.46313976
##               HS_dropout                     Gini                    Black
##              -0.48116338              -0.58172968              -0.58815180
##                  Gini_99           Single_mothers
##              -0.63925281              -0.67123794
```

```r
library(tidyverse)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths
```

```r
# I did this chart ecause i feel like it offers another view and we can see multicolinarity paterns wer

# Select numeric columns
numeric_data <- select(data, where(is.numeric))

# Calculate correlation matrix
cor_matrix <- cor(numeric_data, use = "complete.obs")

# Melt the correlation matrix
melted_cor <- melt(cor_matrix)

# Plot the heatmap with improved formatting
ggplot(data = melted_cor, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1, 1), space = "Lab",
                       name = "Correlation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1, size = 6),
```

```
        axis.text.y = element_text(size = 6)) +
  coord_fixed() +
  labs(title = "Correlation Heatmap for Economic Mobility",
       x = NULL, y = NULL)
```



Important Positive Correlations- Middle_class (0.67) and Commute (0.63) suggest that higher middle-class representation and longer commutes may be linked to higher mobility.

Important Negative Correlations- Single_mothers (-0.67) and Gini_99 (-0.64) suggest that inequality and single-parent households might limit mobility.

```
#Regressional Model Multivariate
numeric_columns <- sapply(data, is.numeric)
data_numeric <- data[,numeric_columns]

full_model <- lm(Mobility ~ .,
                 data = data_numeric)

summary(full_model)
```

```
##
## Call:
## lm(formula = Mobility ~ ., data = data_numeric)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
```

```
## -0.080348 -0.010973 -0.000703  0.009491  0.131817
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.773e-01  7.214e-02   2.458 0.014440 *
## ID                          3.005e-07  2.323e-07   1.293 0.196631
## Population                  1.348e-09  2.176e-09   0.620 0.535860
## Urban                       1.137e-03  3.527e-03   0.322 0.747394
## Black                       9.645e-02  2.610e-02   3.696 0.000252 ***
## Seg_racial                 -4.667e-02  1.658e-02  -2.815 0.005136 **
## Seg_income                  1.085e+00  8.307e-01   1.306 0.192360
## Seg_poverty                -8.740e-01  4.468e-01  -1.956 0.051197 .
## Seg_affluence              -3.339e-01  4.162e-01  -0.802 0.422814
## Commute                     7.246e-02  2.560e-02   2.831 0.004895 **
## Income                      2.227e-07  6.013e-07   0.370 0.711317
## Gini                        2.933e+00  2.885e+00   1.016 0.310103
## Share01                    -2.935e-02  2.886e-02  -1.017 0.309905
## Gini_99                    -3.029e+00  2.886e+00  -1.050 0.294517
## Middle_class                8.817e-02  4.263e-02   2.068 0.039278 *
## Local_tax_rate              1.211e-01  2.379e-01   0.509 0.610859
## Local_gov_spending          1.532e-06  2.790e-06   0.549 0.583164
## Progressivity               5.927e-03  1.146e-03   5.171 3.78e-07 ***
## EITC                       -6.157e-04  4.093e-04  -1.504 0.133418
## School_spending            -1.475e-03  2.070e-03  -0.713 0.476579
## Student_teacher_ratio      -2.743e-04  1.035e-03  -0.265 0.791245
## Test_scores                 4.349e-04  2.763e-04   1.574 0.116235
## HS_dropout                 -1.763e-01  7.765e-02  -2.271 0.023738 *
## Colleges                   -9.634e-02  7.246e-02  -1.330 0.184463
## Tuition                    -3.631e-08  3.999e-07  -0.091 0.927698
## Graduation                 -1.400e-02  1.263e-02  -1.108 0.268434
## Labor_force_participation -5.610e-02  4.855e-02  -1.156 0.248570
## Manufacturing              -1.715e-01  2.528e-02  -6.785 4.52e-11 ***
## Chinese_imports            -8.389e-04  6.986e-04  -1.201 0.230566
## Teenage_labor              -2.481e+00  1.945e+00  -1.275 0.202981
## Migration_in               -1.065e-01  2.764e-01  -0.385 0.700252
## Migration_out              -5.456e-01  3.380e-01  -1.614 0.107384
## Foreign_born                9.734e-02  5.035e-02   1.933 0.053930 .
## Social_capital             -2.061e-03  2.428e-03  -0.849 0.396567
## Religious                   6.048e-02  1.156e-02   5.230 2.81e-07 ***
## Violent_crime              -3.098e+00  1.482e+00  -2.091 0.037227 *
## Single_mothers             -3.574e-01  8.363e-02  -4.274 2.44e-05 ***
## Divorced                    7.426e-02  1.416e-01   0.524 0.600434
## Married                    -8.635e-02  6.704e-02  -1.288 0.198520
## Longitude                   2.869e-04  2.449e-04   1.171 0.242197
## Latitude                    1.444e-03  5.309e-04   2.719 0.006844 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02238 on 377 degrees of freedom
##   (323 observations deleted due to missingness)
## Multiple R-squared:  0.767,  Adjusted R-squared:  0.7423
## F-statistic: 31.02 on 40 and 377 DF,  p-value: < 2.2e-16
```

The variables that stood out to use as significant predictors for economic mobility.

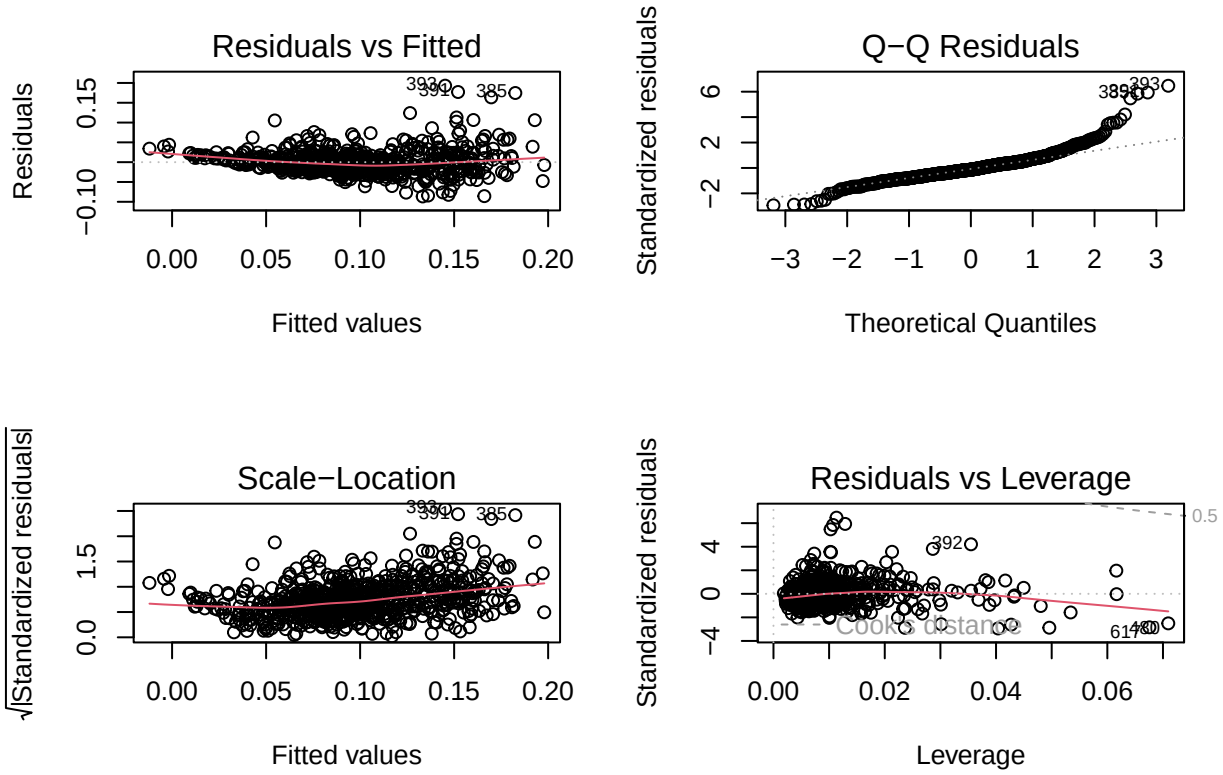
Positively- Commute(p= 0.0049), Middle_class(p = 0.0393), Progressivity(p = 3.78e-07), Religious(p=2.81e-07)

Negatively-Black (p = .0003), Seg_racial (p = .0051), HS_dropout(0.0237), Manufacturing(p=4.52e-11), Violent_crime(p = .0372)

```
# important variables that are imoortant with coralations and the regression

refined_model1 <- lm(Mobility ~ Middle_class + Commute + Single_mothers +
                     Seg_racial + Local_tax_rate + Religious, data = data)

summary(refined_model1)
```

```
##
## Call:
## lm(formula = Mobility ~ Middle_class + Commute + Single_mothers +
##     Seg_racial + Local_tax_rate + Religious, data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.086495 -0.016772 -0.003965  0.012092  0.193100
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.079871   0.018545   4.307 1.89e-05 ***
## Middle_class    0.035620   0.022689   1.570   0.1169
## Commute         0.094268   0.011107   8.487  < 2e-16 ***
## Single_mothers -0.400510   0.034944 -11.461  < 2e-16 ***
## Seg_racial     -0.023980   0.013107  -1.830   0.0677 .
## Local_tax_rate  0.609813   0.131620   4.633 4.29e-06 ***
## Religious       0.049978   0.007991   6.254 6.94e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03003 on 701 degrees of freedom
##   (33 observations deleted due to missingness)
## Multiple R-squared:  0.6127, Adjusted R-squared:  0.6094
## F-statistic: 184.8 on 6 and 701 DF,  p-value: < 2.2e-16
```

```
# Plotting residual diagnostics to check the models details
par(mfrow=c(2,2))  # 2x2 grid
plot(refined_model1)
```




```r
# Create an empty data frame to store results
bivariate_results <- data.frame(Variable = character(),
                                Estimate = numeric(),
                                P_Value = numeric(),
                                R_Squared = numeric(),
                                stringsAsFactors = FALSE)

# Loop through each predictor for bivariate regression
for (var in predictors) {
  formula <- as.formula(paste("Mobility ~", var))
  model <- lm(formula, data = data_numeric)
  model_summary <- summary(model)

  # Append results to the data frame
  bivariate_results <- rbind(bivariate_results, data.frame(
    Variable = var,
    Estimate = coef(model_summary)[2, "Estimate"],
    P_Value = coef(model_summary)[2, "Pr(>|t|)"],
    R_Squared = model_summary$r.squared
  ))
}

# Display results as a formatted table
library(knitr)
kable(bivariate_results, caption = "Bivariate Regression Results")
```

Table 1: Bivariate Regression Results

Variable	Estimate	P_Value	R_Squared
ID	0.0000022	0.0000000	0.2182191
Population	0.0000000	0.0002608	0.0181844
Urban	-0.0375720	0.0000000	0.1259701
Black	-0.2168483	0.0000000	0.2560629
Seg_racial	-0.1835019	0.0000000	0.1204213
Seg_income	-0.6300985	0.0000000	0.1439030
Seg_poverty	-0.7109757	0.0000000	0.1551999
Seg_affluence	-0.5331022	0.0000000	0.1270247
Commute	0.2218695	0.0000000	0.3488484
Income	0.0000003	0.3623496	0.0011414
Gini	-0.3449282	0.0000000	0.2765501
Share01	-0.0017182	0.0000013	0.0326314
Gini_99	-0.4949509	0.0000000	0.3314996
Middle_class	0.3538504	0.0000000	0.3358499
Local_tax_rate	1.8082600	0.0000000	0.1155664
Local_gov_spending	0.0000105	0.0000002	0.0360734
Progressivity	0.0068255	0.0000002	0.0360707
EITC	0.0016335	0.0012163	0.0143020
School_spending	0.0114643	0.0000000	0.0621277
Student_teacher_ratio	-0.0074103	0.0000000	0.1090234
Test_scores	0.0026132	0.0000000	0.2038473
HS_dropout	-1.0447010	0.0000000	0.2110610
Colleges	0.4871396	0.0000000	0.0659575
Tuition	-0.0000007	0.1461628	0.0036502
Graduation	0.0188566	0.1333785	0.0038865
Labor_force_participation	0.1375827	0.0000271	0.0239450
Manufacturing	-0.2295334	0.0000000	0.1306179
Chinese_imports	-0.0062172	0.0000000	0.0441453
Teenage_labor	18.4264531	0.0000000	0.2873508
Migration_in	-1.3005781	0.0000000	0.0684527
Migration_out	-1.0700546	0.0000233	0.0246057
Foreign_born	-0.0147418	0.7036504	0.0001992
Social_capital	0.0210186	0.0000000	0.2591744
Religious	0.1411580	0.0000000	0.1918278
Violent_crime	-10.0505704	0.0000000	0.0793012
Single_mothers	-0.6886385	0.0000000	0.4704386
Divorced	-1.2725612	0.0000000	0.1818879
Married	0.5684711	0.0000000	0.2535977
Longitude	-0.0011080	0.0000000	0.0985870
Latitude	0.0031297	0.0000000	0.1284254

Research Question 2

Variable selection process

To investigate the extent to which measures of better education predict higher levels of economic mobility, I chose variables revolving around performance indicators and resource factors in regards to education.

Educational Performance Indicators: Test_scores, HS_dropout, Graduation **School Resource Fac-**

tors: School_spending, Student_teacher_ratio

```
model <- lm(Mobility ~ Test_scores + HS_dropout +
              Graduation + School_spending +
              Student_teacher_ratio, data = data)
```

```
# Cleaning up model by getting rid of outliers
cooksd <- cooks.distance(model)
influential_points <- which(cooksd > (4/length(cooksd)))
data_clean <- data[-influential_points, ]
model_clean <- lm(Mobility ~ Test_scores + HS_dropout + Graduation +
                  School_spending + Student_teacher_ratio, data = data_clean)
summary(model_clean)
```

```
##
## Call:
## lm(formula = Mobility ~ Test_scores + HS_dropout + Graduation +
##     School_spending + Student_teacher_ratio, data = data_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.06453 -0.02270 -0.00575  0.01209  0.20242
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             0.0957844  0.0198886   4.816 2.02e-06 ***
## Test_scores             0.0024851  0.0002676   9.286  < 2e-16 ***
## HS_dropout             -0.5689862  0.0977069  -5.823 1.12e-08 ***
## Graduation             -0.0346678  0.0130907  -2.648  0.00838 **
## School_spending         0.0024514  0.0017026   1.440  0.15064
## Student_teacher_ratio -0.0009866  0.0009446  -1.044  0.29686
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03583 on 436 degrees of freedom
##   (273 observations deleted due to missingness)
## Multiple R-squared:  0.3455, Adjusted R-squared:  0.338
## F-statistic: 46.04 on 5 and 436 DF,  p-value: < 2.2e-16
```

Variable Interpretation

Test_scores (0.0025, $p < 0.001$): Higher test scores are positively associated with greater economic mobility. **HS_dropout (-0.5690, $p < 0.001$):** Higher high school dropout rates significantly decrease economic mobility. **Graduation (-0.0347, $p = 0.0084$):** Higher graduation rates negatively correlate with mobility, suggesting that other factors may be influencing this relationship. **School_spending (0.0025, $p = 0.1506$):** While increased school spending appears to have a positive effect, it is not statistically significant at the 5% level. **Student_teacher_ratio (-0.0010, $p = 0.2969$):** The student-to-teacher ratio does not significantly impact economic mobility,

Model Diagnostics and Selection

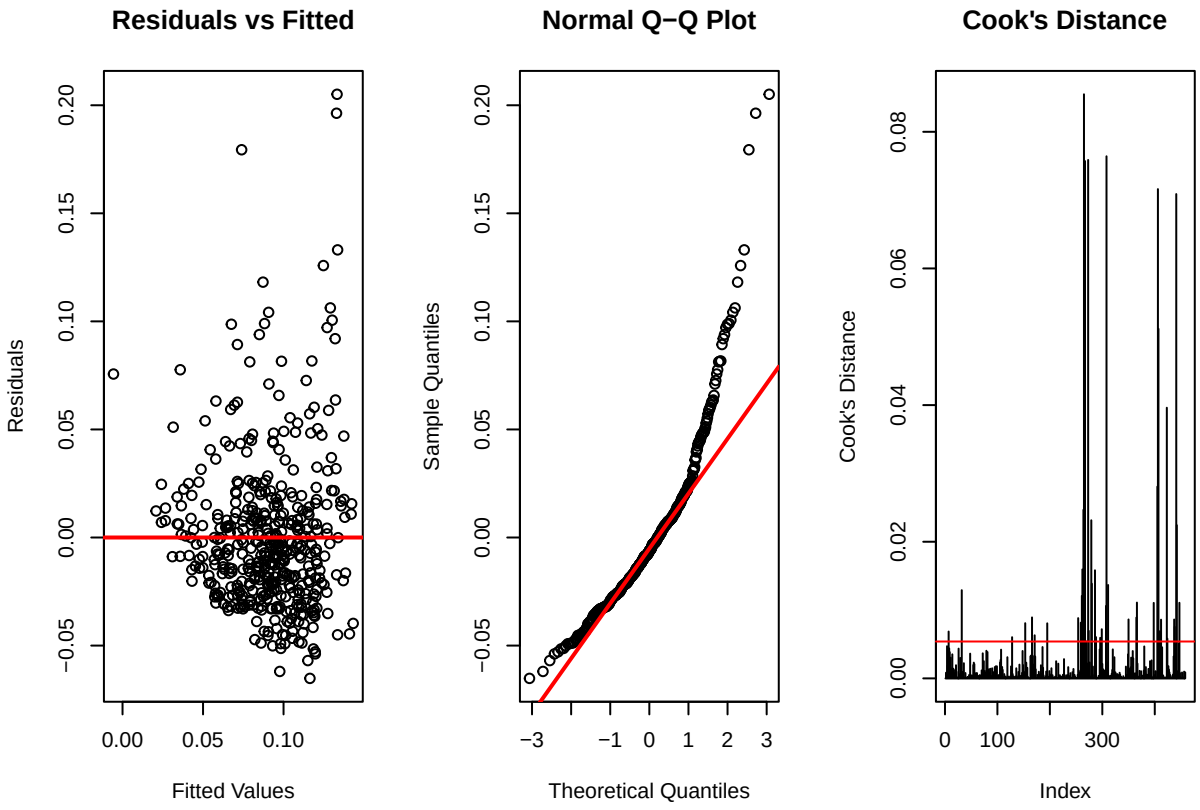
```
# Model without Student_teacher_ratio
reduced_model <- lm(Mobility ~ Test_scores + HS_dropout + Graduation + School_spending, data = data_clea
```

Test_scores (0.0025, p < 0.001): Higher test scores are positively associated with greater economic mobility. **HS_dropout (-0.5690, p < 0.001):** Higher high school dropout rates significantly decrease economic mobility. **Graduation (-0.0347, p = 0.0084):** Higher graduation rates negatively correlate with mobility, suggesting that other factors may be influencing this relationship. **School_spending (0.0025, p = 0.1506):** While increased school spending appears to have a positive effect, it is not statistically significant at the 5% level. **Student_teacher_ratio (-0.0010, p = 0.2969):** The student-to-teacher ratio does not significantly impact economic mobility,

Model Diagnostics and Selection

```
# Model without Student_teacher_ratio
reduced_model <- lm(Mobility ~ Test_scores + HS_dropout + Graduation + School_spending, data = data_clea
```

```
par(mfrow = c(1, 3))

# Residuals vs. Fitted Plot
plot(reduced_model$fitted.values, resid(reduced_model),
     main = "Residuals vs Fitted",
     xlab = "Fitted Values", ylab = "Residuals")
abline(h = 0, col = "red", lwd = 2)

# Q-Q Plot
qqnorm(resid(reduced_model))
qqline(resid(reduced_model), col = "red", lwd = 2)

# Cook's Distance
plot(cooksd, type = "h", main = "Cook's Distance", ylab = "Cook's Distance")
abline(h = 4/(nrow(data)), col = "red")
```



Research Question 3

Variable Selection Process

To investigate the extent to which measures of integration across social groups predict economic mobility, I selected a set of variables that capture different dimensions of social integration and demographic characteristics. I grouped them info two groups:

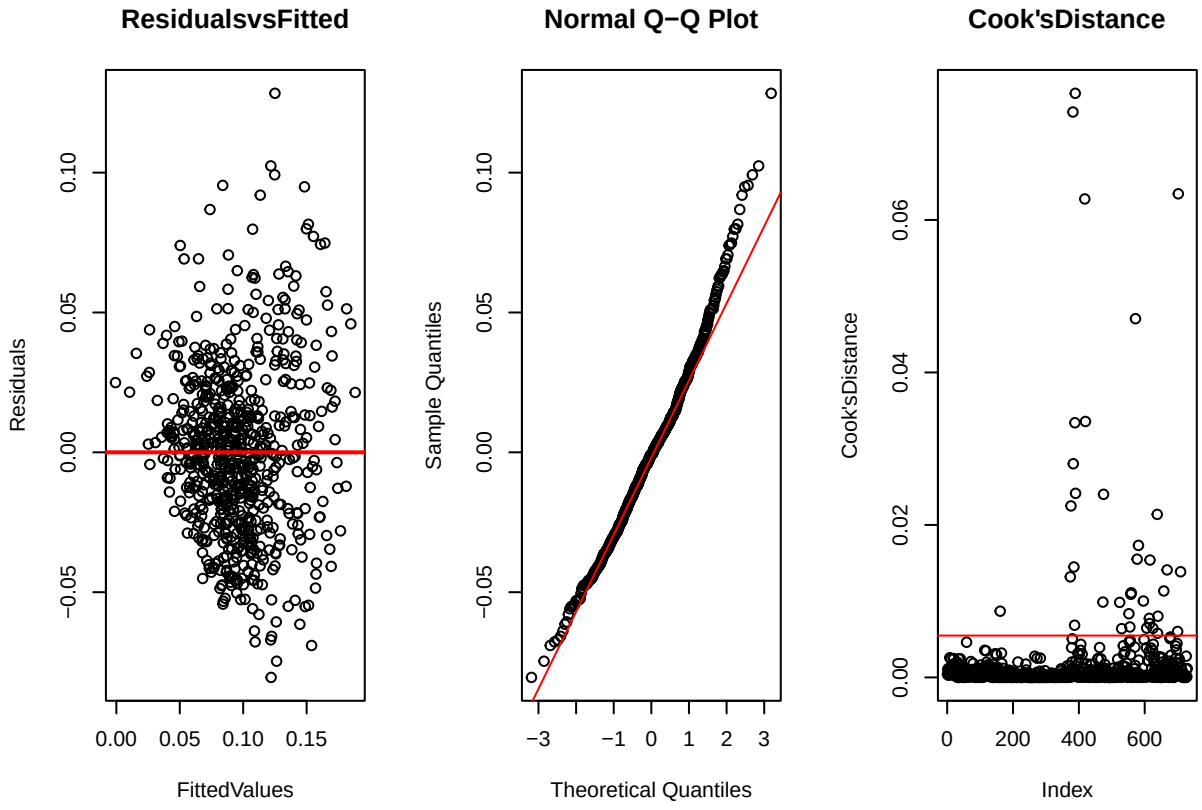
- **Segregation Measures** : Seg_racial, Seg_poverty, and Seg_affluence
- **Social and Demographic Factors** : Married, Divorced, Foreign_born, and Religious

```
# Create Variables for Social Groups
social_groups <- c("Seg_racial", "Seg_poverty", "Seg_affluence", "Mobility", "Married", "Divorced", "R
data <- data[social_groups] %>% na.omit()
```

```
##
## Call:
## lm(formula = Mobility ~ ., data = data_clean)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.080465 -0.020463 -0.000876  0.016714  0.128351
##
```

```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.041376   0.021128  -1.958 0.050595 .
## Seg_racial    -0.110122   0.015029  -7.327 6.64e-13 ***
## Seg_poverty   -0.023435   0.137671  -0.170 0.864884
## Seg_affluence -0.026678   0.111770  -0.239 0.811423
## Married        0.340929   0.031023  10.990  < 2e-16 ***
## Divorced      -0.738534   0.084099  -8.782  < 2e-16 ***
## Religious      0.047426   0.009283   5.109 4.21e-07 ***
## Foreign_born   0.110176   0.029285   3.762 0.000183 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03031 on 684 degrees of freedom
## Multiple R-squared:  0.5205, Adjusted R-squared:  0.5156
## F-statistic: 106.1 on 7 and 684 DF,  p-value: < 2.2e-16
```

Model Diagnostics



1. Residuals vs. Fitted Plot: This plot displays the residuals against the fitted values from the regression model. This model is assumed valid because the plots are scattered randomly.

2. Normal Q-Q Plot: This plot checks whether the residuals follow a normal distribution. It compares the observed residuals to a theoretical normal distribution. Becuase the points follow the red line, the residuals are normally distributed

3. Cook's Distance Plot: This plot identifies influential data points in the regression model. Cook's distance measures how much a single observation affects the model. In this model, I removed all outliers seen on the plot.

Research Question 4

```
# read the csv
data <- read.csv("/Users/bryangutierrez/Downloads/mobility-all.csv", stringsAsFactors = FALSE)
```

Variable Selection Process

To investigate variables which can be directly affected by government policy predict economic mobility, we looked into the following:

- **Tax Policy & Government Spending** (Local_tax_rate), (Progressivity)
- **Education Policy** (HS_dropout)
- **Public Safety & Social Programs** (Violent_crime), (Seg_racial)
- **Community & Infrastructure** (Single_mothers), (Commute), (Religious)

```
economic_model <- lm(Mobility ~ Local_tax_rate + Progressivity + HS_dropout + Violent_crime +

        Seg_racial + Single_mothers + Commute + Religious, data = data)

summary(economic_model)
```

```
##
## Call:
## lm(formula = Mobility ~ Local_tax_rate + Progressivity + HS_dropout +
##     Violent_crime + Seg_racial + Single_mothers + Commute + Religious,
##     data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.085179 -0.016918 -0.004293  0.012534  0.164877
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.089971   0.010017   8.982  < 2e-16 ***
## Local_tax_rate  0.898424   0.149955   5.991 3.74e-09 ***
## Progressivity   0.009056   0.001026   8.829  < 2e-16 ***
## HS_dropout     -0.173341   0.070172  -2.470 0.013802 *
## Violent_crime  -3.158314   0.939981  -3.360 0.000833 ***
```

```
## Seg_racial     -0.040280   0.014416  -2.794 0.005385 **
## Single_mothers -0.340976   0.030832 -11.059  < 2e-16 ***
## Commute         0.102731   0.011113   9.244  < 2e-16 ***
## Religious       0.029825   0.008997   3.315 0.000976 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02957 on 556 degrees of freedom
##   (176 observations deleted due to missingness)
## Multiple R-squared:  0.6684, Adjusted R-squared:  0.6636
## F-statistic: 140.1 on 8 and 556 DF,  p-value: < 2.2e-16
```