Project1

Ruth Walters

2025-02-05

```r
# Import dependencies
library(ggplot2)
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
```

```r
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.4.1

## corrplot 0.95 loaded
```

```r
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.4.1
```

```r
theme_set(theme_bw())

# Import data
mobility <- read.csv("mobility-all.csv", header = TRUE)
```

Introduction

Write four to five sentences introducing the research problem and describing specific research hypotheses. Cite any information sources in parentheses or foot- or end- notes.

Research questions:

1. Which variables are the most important variables for predicting economic mobility?
2. To what extent do measures of better education predict higher levels of economic mobility?
3. To what extent do measures of integration across social groups predict economic mobility?
4. To what extent do variables which can be directly affected by government policy predict economic mobility?

Exploratory data analysis

Visually and numerically investigate which variables seem associated with economic mobility? - Examine the (predictor and response) variables univariately and multivariately. You will likely not be able to include all of *the plots, think carefully about which ones would be good to include. - Are there any variables that you would consider transforming based on the plots?*

```r
# Check which columns have NA/null values
print(colSums(is.na(mobility)))
```

```
##                        ID                      Name                  Mobility
##                         0                         0                        12
##                     State                Population                     Urban
##                         0                         0                         0
##                     Black                Seg_racial                Seg_income
##                         0                         0                         0
##               Seg_poverty             Seg_affluence                   Commute
##                         0                         0                         0
##                    Income                      Gini                   Share01
##                         0                         0                        32
##                   Gini_99              Middle_class            Local_tax_rate
##                        32                        32                         1
##        Local_gov_spending             Progressivity                      EITC
##                         2                         0                         0
##           School_spending     Student_teacher_ratio               Test_scores
##                        10                        30                        36
##                HS_dropout                  Colleges                   Tuition
##                       148                       157                       161
##                Graduation Labor_force_participation             Manufacturing
##                       160                         0                         0
##           Chinese_imports             Teenage_labor              Migration_in
##                        19                        32                        17
##             Migration_out              Foreign_born            Social_capital
##                        17                         0                        19
##                 Religious             Violent_crime            Single_mothers
##                         0                        27                         0
##                  Divorced                   Married                 Longitude
##                         0                         0                         0
##                  Latitude
##                         0
```

```r
# for (i in colSums(is.na(mobility))) {
#   if (i != 0) {print(i)}
# }
```

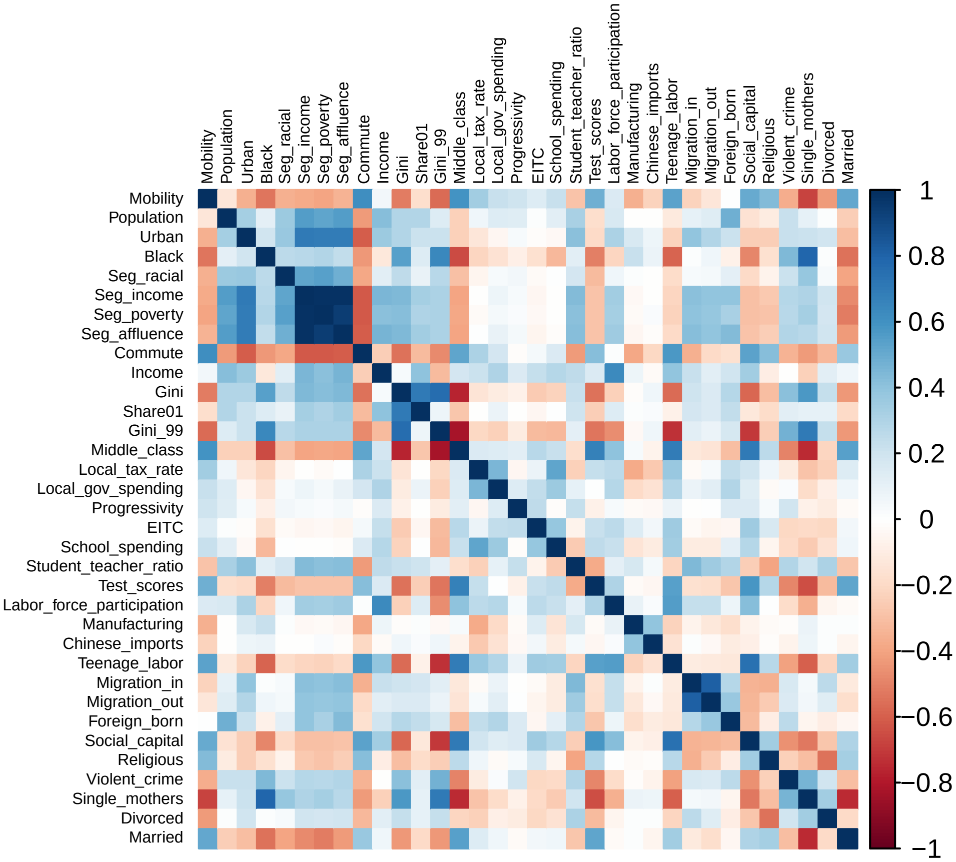
```r
# Drop columns with >100 NA values
mobility <- mobility[,!(names(mobility) %in% c("Colleges","Tuition", "Graduation", "HS_dropout"))]

# Drop rows with NAs
mobility <- drop_na(mobility)
```

```r
# Recheck NA
for (i in colSums(is.na(mobility))) {
  if (as.numeric(i) != 0) {print(i)}
}
```

```r
mobility_numeric <- mobility[,!(names(mobility) %in% c("ID","Name", "State", "Latitude", "Longitude"))]

corrplot(cor(mobility_numeric),
         tl.col = "black",
         tl.cex = .5,
         method = 'color')
```



Mobility appears to be highly positively correlated with the cluster of variables that measure segregation

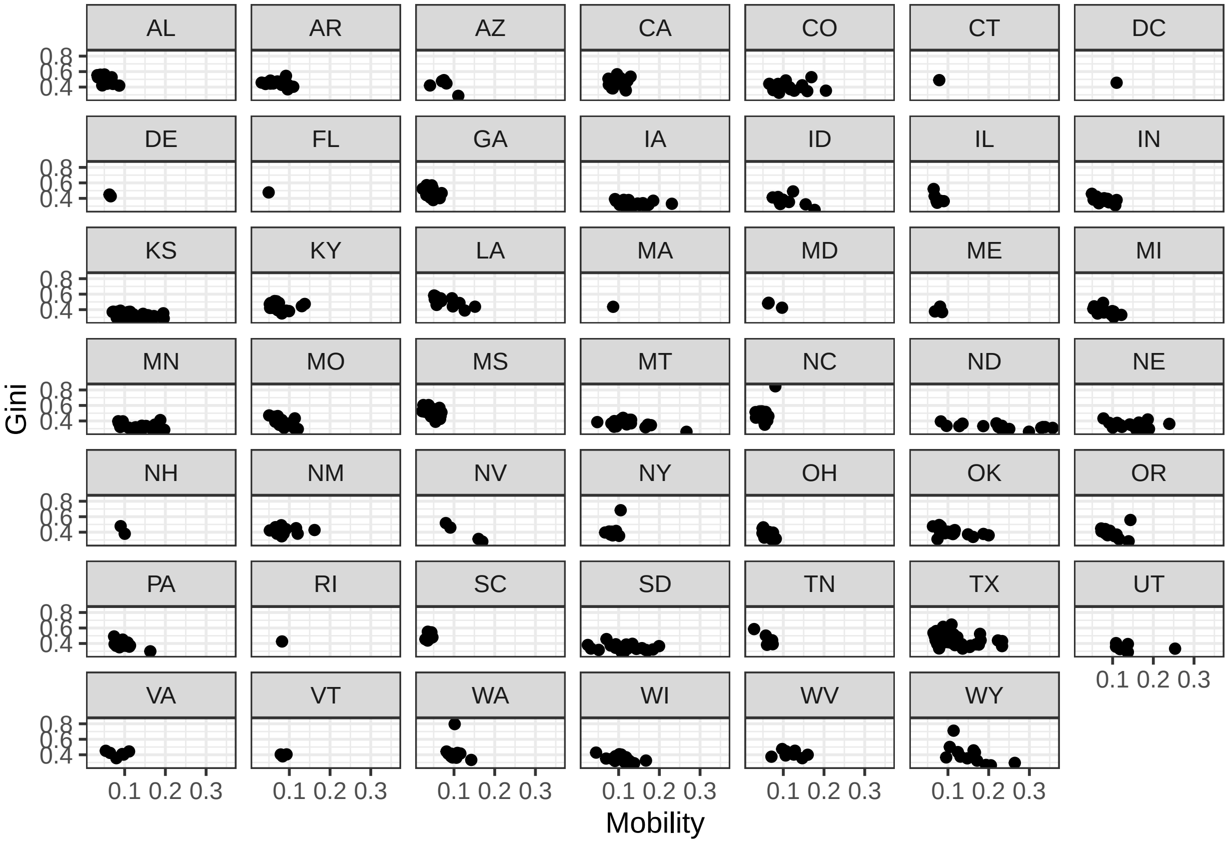
We can further identify three clusters of highly correlated variables:

- measures of segregation (`seg_racial`, `seg_income`, and `seg_affluence`)
- measures of the Gini index (`Gini`, `Share01`, `Gini_99` and `middle_class`)
- measures of migration (`migration_in` and `migration_out`)

```
# Drop highly correlated variables
mobility <- mobility[,!(names(mobility) %in% c("Seg_income","Seg_affluence", "Share01", "Gini_99"))]

mobility_numeric <- mobility[,!(names(mobility) %in% c("ID","Name", "State", "Latitude", "Longitude"))]
```

```
ggplot(mobility, aes(Mobility, Gini)) +
  geom_point() +
  facet_wrap(~State)
```



```
northeast <- c("northeast", "CT", "ME", "MA", "NH", "NJ", "NY", "PA", "RI", "VT")
southeast <- c("AL", "AR", "FL", "GA", "KY", "LA", "MS", "NC", "SC", "TN", "VA", "WV")
midwest <- c("IL", "IN", "IA", "KS", "MI", "MN", "MO", "NE", "ND", "OH", "SD", "WI")
southwest <- c("AZ", "NM", "OK", "TX")
west <- c("AK", "CA", "CO", "HI", "ID", "MT", "NV", "OR", "UT", "WA", "WY")

us_regions <- c(northeast, southeast, southwest, midwest, west)

count <- 1

for (i in mobility$State) {
  for (j in us_regions) {
    if (i %in% j) {
      mobility$region[count] <- j[1]
    }
  }
  count = count + 1
}
```

Model selection

Initial modeling

Start by building a multivariate linear regression using the covariates to predict mobility variable. Address the specific questions of above when building the model. Be sure to justify the choices you made in building this initial model

Diagnostics

- Are the basic assumptions met for your multivariate linear regression model? Why or why not?
- What transformations do you choose (if any)? Why?
- Are there any outliers in your sample overly influencing your model? Identify any outlier candidates and decide whether or not to remove them. Give details.
- Do you exclude any variables? Why? All exclusions/inclusions must be justified

Final model selection

Model results

Create a table that summarizes your final model (coefficients, standard errors, confidence intervals, p-values). Provide interpretations of all your coefficients in the con- text of the problem. Be sure to address the specific questions of the client (above).

Discussion

What are your conclusions? Identify a few key findings, and discuss, with reference to the supporting evidence. Can you come up with explanations for the patterns you have found? Suggestions or recommendations for the client? How could your analysis be improved? (6–8 sentences)