modelConstruction

Ruth Walters

2025-02-12

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

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
```

```
library(corrplot)
```

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

```
## corrplot 0.95 loaded
```

```
library(tidyr)
```

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

```
library(cowplot)
library(ggpubr)
```

```
##
## Attaching package: 'ggpubr'
```

```
## The following object is masked from 'package:cowplot':
##
##     get_legend
```

```
theme_set(theme_bw())
theme_update(
  axis.title.x = element_blank(),
  axis.title.y = element_blank(),
  plot.title = element_text(size = 12, face = "italic")
)
```

From datacamp

[https://www.datacamp.com/tutorial/linear-regression-R][How to Do Linear Regression in R]

> "When a regression takes into account two or more predictors to create the linear regression, it's called multiple linear regression. In R, to add another coefficient, add the symbol"+" for every additional variable you want to add to the model.

Linear model: `lm([target] ~ [predictor], data = [data source])`

Data preparation

```
mobility <- read.csv("mobility-all.csv", header = TRUE, stringsAsFactors = TRUE)
```

Drop all non-quantitative rows

```
quals <- c("ID","Name", "State", "Latitude", "Longitude")

mobility <- mobility[,!(names(mobility) %in% quals)]
```

Drop low-quality columns

```
print(colSums(is.na(mobility)))
```

```
##                  Mobility                Population                     Urban
##                        12                         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
##                         0                         0
```

```
bad_cols <- c("Colleges","Tuition", "Graduation", "HS_dropout") # +100 NULL
mobility <- mobility[,!(names(mobility) %in% bad_cols)]
```

Drop remaining NULLS

```
before <- nrow(mobility)
mobility <- drop_na(mobility)
dropped <- before - nrow(mobility)

print("Data reduced by: ")
```

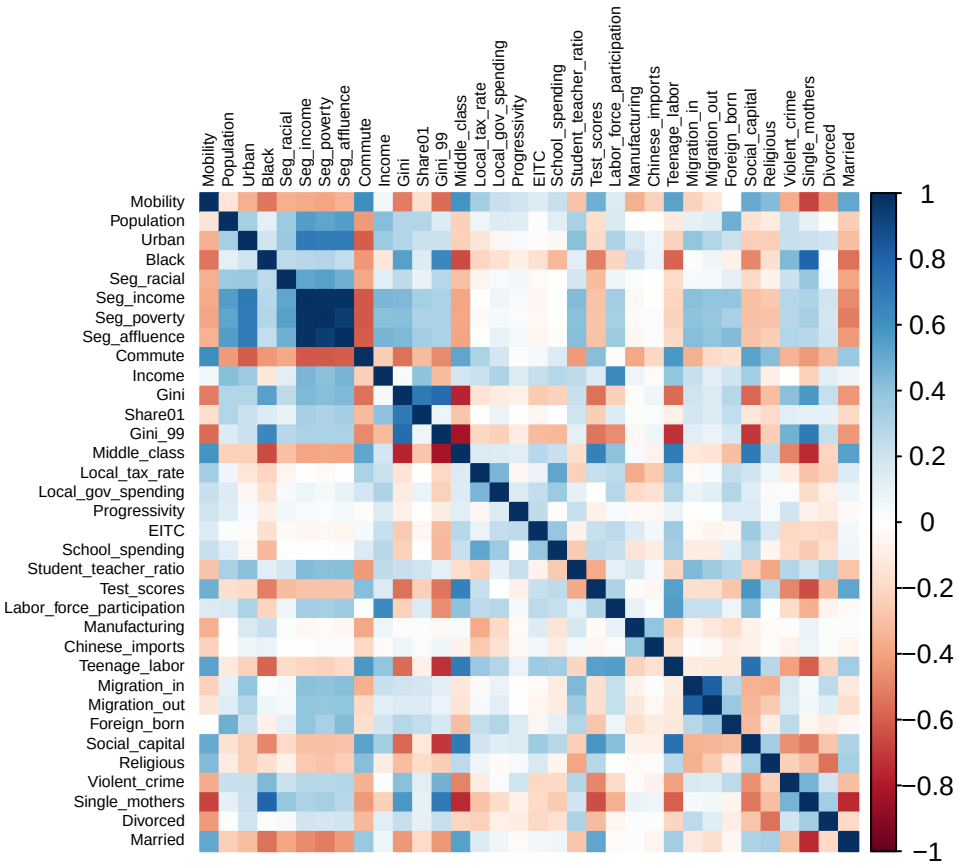
```
## [1] "Data reduced by: "
```

```
print((dropped/before))
```

```
## [1] 0.145749
```

Exploratory data analysis

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



Social determinants of mobility

Goal: explore potential social determinants of mobility

Selected variables: - segregation variables `Seg_racial`, `Seg_income`, `Seg_poverty`, and `Seg_affluence` - educational variables `School_spending`, `Student_teacher_ratio`, and `Test_scores` - family dynamic variables `Single_mothers`, `Divorced`

```
a <- ggplot(data = mobility, aes(x = Mobility, y = Seg_racial)) +
  geom_point(color = "cornflowerblue", alpha = .3) +
  #stat_smooth(method = "lm", formula = y ~ x, geom = "line", color = "darkorange") +
  stat_cor(label.x=.17, label.y=.5) +
  ggtitle("Race")
```

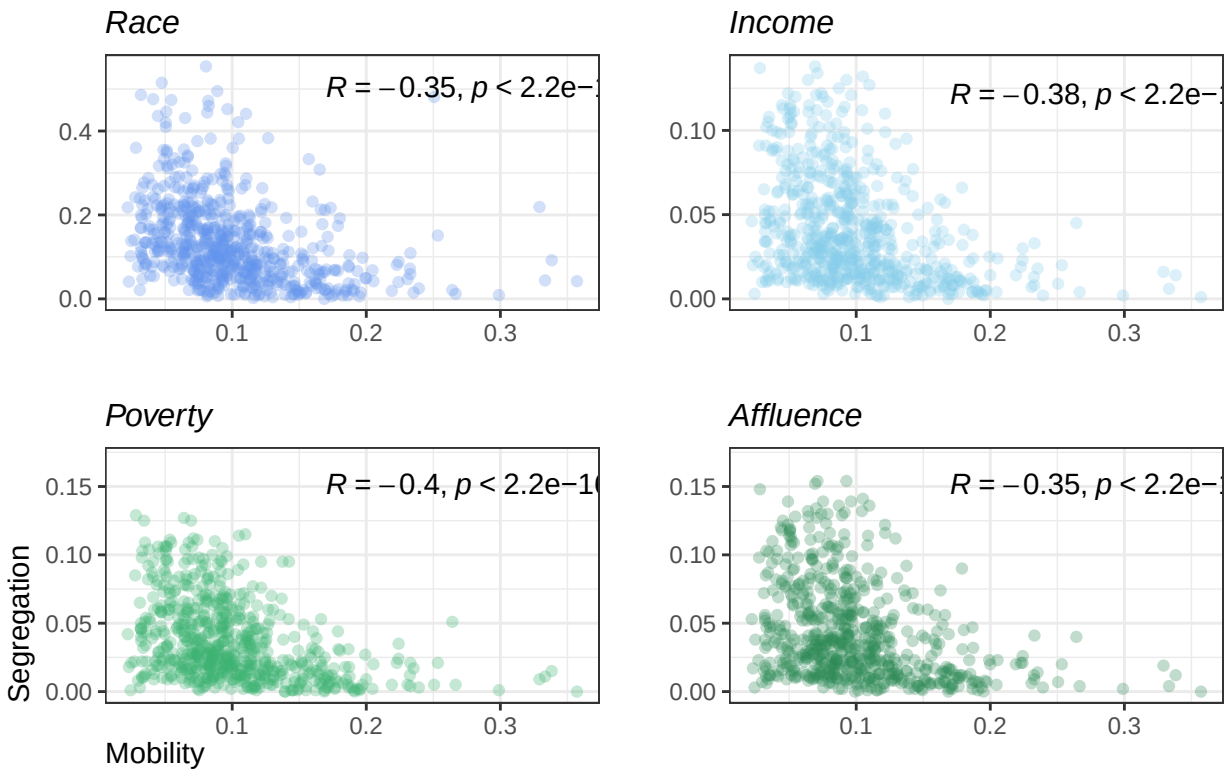
```r
b <- ggplot(data = mobility, aes(x = Mobility, y = Seg_income)) +
  geom_point(color = "skyblue", alpha = .3) +
  stat_cor(label.x=.17, label.y=.12) +
  ggtitle("Income")

c <- ggplot(data = mobility, aes(x = Mobility, y = Seg_poverty)) +
  geom_point(color = "mediumseagreen", alpha = .3) +
  stat_cor(label.x=.17, label.y=.15) +
  ylim(0,.17) +
  ggtitle("Poverty") +
  xlab("Mobility") +
  ylab("Segregation") +
  theme(axis.title.x = element_text(hjust = 0),
        axis.title.y = element_text(angle=90, hjust = 0, margin = margin(r = 5)))

d <- ggplot(data = mobility, aes(x = Mobility, y = Seg_affluence)) +
  geom_point(color = "seagreen", alpha = .3) +
  stat_cor(label.x=.17, label.y=.15) +
  ylim(0,.17) +
  ggtitle("Affluence")

plot_row <- plot_grid(a,b,c,d, align = "hv")

title <- ggdraw() +
  draw_label(
    "Segregation as a predictor of mobility",
    fontface = 'bold',
    x = 0,
    hjust = 0) +
  theme(plot.margin = margin(0, 0, 0, 7))
plot_grid(
  title, plot_row,
  ncol = 1,
  rel_heights = c(0.1, 1)
)
```



Segregation

Remove highly correlated variables:

```
mobility <- mobility[,!(names(mobility) %in% c("Seg_income", "Seg_affluence", "Gini_99", "Share01"))]
```

```
a <- ggplot(data = mobility, aes(x = Mobility, y = School_spending)) +
  geom_point(color = "cornflowerblue", alpha = .3) +
  #stat_smooth(method = "lm", formula = y ~ x, geom = "line", color = "darkorange") +
  stat_cor(label.x=.17, label.y=10.5) +
  ggtitle("School spending")

b <- ggplot(data = mobility, aes(x = Mobility, y = Student_teacher_ratio)) +
  geom_point(color = "skyblue", alpha = .3) +
  stat_cor(label.x=.17, label.y=22, label.size = 0.05) +
  ggtitle("Student teacher ratio")
```

Educational outcomes

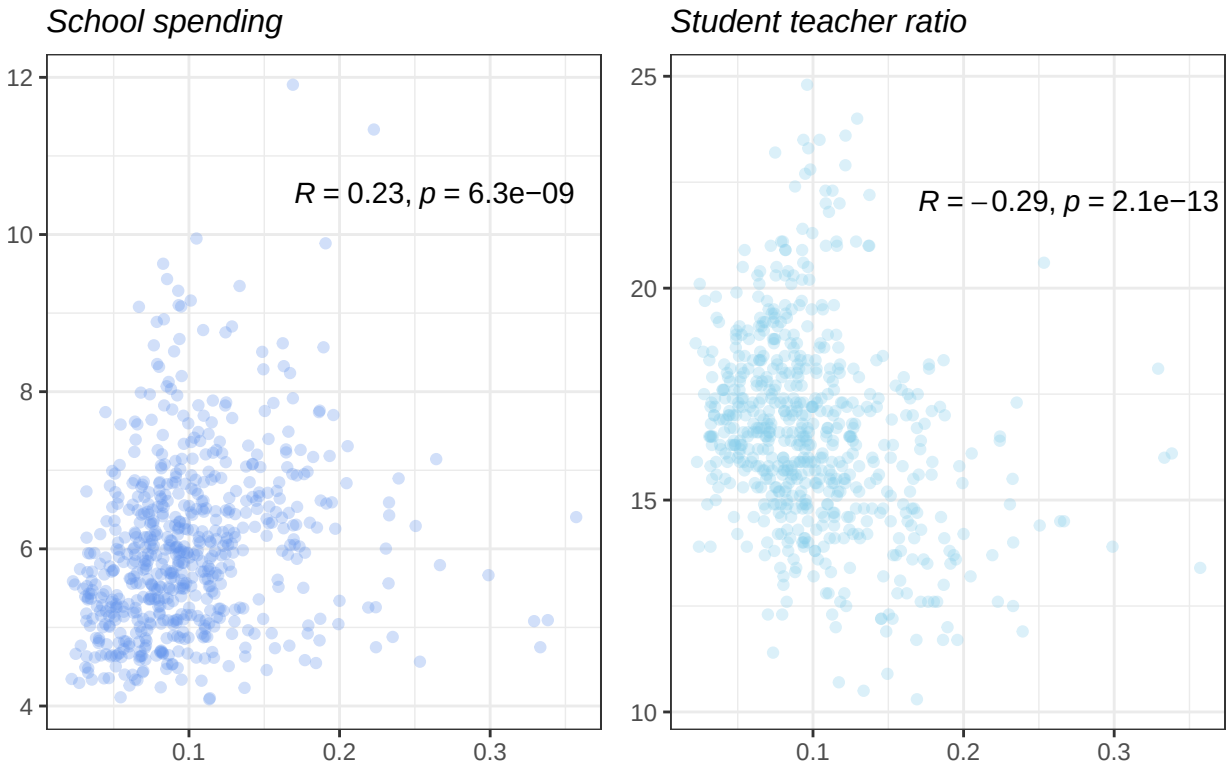
```
## Warning in stat_cor(label.x = 0.17, label.y = 22, label.size = 0.05): Ignoring
## unknown parameters: `label.size`
```

```
plot_row <- plot_grid(a,b, align = "hv")

title <- ggdraw() +
  draw_label(
    "Educational factors associated with mobility",
    fontface = 'bold',
    x = 0,
    hjust = 0) +
```

```
  theme(plot.margin = margin(0, 0, 0, 7))
plot_grid(
  title, plot_row,
  ncol = 1,
  rel_heights = c(0.1, 1)
)
```

Educational factors associated with mobility



Modeling

```
lm.1 <- lm(formula = Mobility ~ .,
   data = mobility)

summary(lm.1)
```

```
##
## Call:
## lm(formula = Mobility ~ ., data = mobility)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.074264 -0.014421 -0.001369  0.011195  0.156903
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.106e-01  5.377e-02   2.057 0.040132 *
## Population                  1.330e-09  1.472e-09   0.903 0.366644
## Urban                       1.476e-03  3.396e-03   0.434 0.664087
```

```
## Black                       5.556e-02  2.258e-02   2.460 0.014169 *
## Seg_racial                 -5.228e-02  1.455e-02  -3.594 0.000352 ***
## Seg_poverty                -7.910e-02  7.902e-02  -1.001 0.317188
## Commute                     5.874e-02  1.997e-02   2.941 0.003394 **
## Income                      7.401e-07  4.377e-07   1.691 0.091368 .
## Gini                       -4.191e-02  2.307e-02  -1.817 0.069760 .
## Middle_class                1.195e-01  3.337e-02   3.581 0.000370 ***
## Local_tax_rate              1.689e-01  1.667e-01   1.014 0.311159
## Local_gov_spending          1.218e-06  1.658e-06   0.735 0.462878
## Progressivity               5.210e-03  8.700e-04   5.989 3.64e-09 ***
## EITC                       -3.825e-04  3.518e-04  -1.087 0.277401
## School_spending             1.304e-03  1.504e-03   0.867 0.386389
## Student_teacher_ratio       5.096e-04  7.245e-04   0.703 0.482075
## Test_scores                -2.006e-04  2.223e-04  -0.903 0.367125
## Labor_force_participation -2.275e-02  3.674e-02  -0.619 0.536035
## Manufacturing              -1.673e-01  2.076e-02  -8.062 4.05e-15 ***
## Chinese_imports            -1.079e-03  7.214e-04  -1.496 0.135293
## Teenage_labor              -1.524e+00  1.722e+00  -0.885 0.376559
## Migration_in               -5.547e-01  2.418e-01  -2.294 0.022129 *
## Migration_out               1.403e-01  2.908e-01   0.482 0.629670
## Foreign_born                1.587e-02  3.879e-02   0.409 0.682595
## Social_capital             -3.105e-03  1.971e-03  -1.575 0.115727
## Religious                   4.650e-02  9.770e-03   4.759 2.44e-06 ***
## Violent_crime              -4.464e+00  1.316e+00  -3.392 0.000740 ***
## Single_mothers             -3.734e-01  6.936e-02  -5.384 1.04e-07 ***
## Divorced                   -2.025e-01  1.153e-01  -1.756 0.079667 .
## Married                    -4.519e-03  5.006e-02  -0.090 0.928101
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02662 on 603 degrees of freedom
## Multiple R-squared:  0.7173, Adjusted R-squared:  0.7037
## F-statistic: 52.77 on 29 and 603 DF,  p-value: < 2.2e-16
```

Extract highly correlated variables:

- Black
- Seg_racial
- Commute
- Gini
- Middle_class
- Progressivity
- Manufacturing
- Migration_in
- Religious
- Violent_crime
- Single mothers
- Divorced