Sociol 90z Final Project

Research Lab: Inequality

Angie Shin

5/7/2021

Introduction

The National Longitudinal Surveys (NLS), sponsored by the U.S. Bureau of Labor Statistics, are nationally representative surveys that follow the same sample of individuals from specific birth cohorts over time on tens of thousands of different variables. The National Longitudinal Survey of Youth in 1997 (NLSY97) Cohort is a longitudinal project that follows the lives of a sample of American youth born between 1980-84; 8,984 respondents were ages 12-17 when first interviewed in 1997. This ongoing cohort has been surveyed 18 times to date—originally surveyed annually until 2011, and now interviewed biennially. The most recent facilitation of the NLSY took place in 2017. For this project, I will be exploring 2,096 filtered observations of variables such as race, parent education background, and parent income and poverty level to determine whether or not they affect individuals finishing college.

Literature Review

The paper “Parental Education Better Helps White than Black Families Escape Poverty: National Survey of Children’s Health” by Shervin Assari investigates racial variation in the effects of highest education of parents on family’s ability to scale poverty, defined as the household’s income-to-needs ratio using a nationally representative sample of American families with children. Contextualized with how systemic racism encompasses residential segregation, low quality of education, low paying jobs, discrimination in the labor market, and extra costs of upward social mobility for minorities such that POC families face more challenges for leveraging their education to escape poverty, the paper hypothesizes that consequently, higher education of parents in the household was associated with lower risk of poverty. Race, however, interacted with parental education attainment on household-income-to-needs ratio, indicating smaller effects for Black compared to White families. Lower number of parents and higher number of children in Black families did not explain such racial disparities. The economic gain of parental education on helping family escape poverty is smaller for Black than White families, and this is not as a result of a lower parent-to-child ratio in Black households. Assari also includes recommendations that policies should specifically address structural barriers in the lives of all minorities to minimize the diminished return of SES resources across racial minority groups, enhance quality of education and reduce the extra cost of upward social mobility for racial minorities, and employ multilevel action plans to avoid difficulties with eliminating the existing economic gap between racial groups.

The working paper “The Great Escape: Intergenerational Mobility in the United States Since 1940” by Nathaniel Hilger develops a method to estimate intergenerational mobility in education on large cross-sectional surveys and apply the method to U.S. census data with an institutional focus. The method overcomes the problem that most children cannot be linked to parents by ages of school completion, and thereby allows for estimation of final educational outcomes by parental income and education. The new methodology in conjunction with multiple additional datasets yields several important new historical facts.

Findings included that educational intergenerational mobility increased significantly after 1940 (1911-14 birth cohorts) before stabilizing and then declining after 1980 (1951-54 birth cohorts), post-1940 educational intergenerational mobility gains plausibly increased aggregate annual earnings growth by 0.25 percentage points over the 1940-70 period, and such gains were particularly large in the South for both whites and blacks, implying larger gains for blacks nationally due to their greater geographic concentration in the South. Hilger contextualizes these results with the increase in relative educational intergenerational mobility after 1940 stemming from greater high school enrollment, rather than college enrollment, invalidating the GI Bills, the Civil Rights Movement, school desegregation, the black high school movement, and the Great Migration as accounted for in such gains. In particular, Hilger notes some discrepancies between the working paper and the paper "Is the United States Still a Land of Opportunity? Recent Trends in Intergenerational Mobility" by Chetty et al. in 2014, in that for income-based intergenerational mobility across places in the 2000s education correlated with black population shares, income inequality, and educational quality, even conditional on state and year fixed effects; however, unlike Chetty et al., Hilger found a robust positive association of state income levels with education, which he attributes to the much larger forces of modernization accompanying economic growth after 1940, particularly in the South where lower-SES voters gained political power and K-12 public school input gaps narrowed dramatically.

Other readings were already reviewed in class.

Method

```
knitr::opts_chunk$set(echo = T)
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(glmnet)
```

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

```
## Loaded glmnet 4.1-1
```

```
library(estimatr)
library(infer)
library(ggforce)
```

```
## Loading required package: ggplot2
```

```
library(shinydashboard)
```

```
##
## Attaching package: 'shinydashboard'
## The following object is masked from 'package:graphics':
##
##     box
```

```r
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```r
library(haven)
library(Hmisc)
```

```
## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```r
library(tidymodels)
```

```
## -- Attaching packages -------------------------------------- tidymodels 0.1.2 --

## v broom     0.7.5      v rsample   0.0.9
## v dials     0.0.9      v tibble    3.1.0
## v dplyr     1.0.5      v tidyr     1.1.3
## v modeldata 0.1.0      v tune      0.1.3
## v parsnip   0.1.5      v workflows 0.2.2
## v purrr     0.3.4      v yardstick 0.0.7
## v recipes   0.1.15

## -- Conflicts ----------------------------------------- tidymodels_conflicts() --
## x purrr::discard()     masks scales::discard()
## x tidyr::expand()      masks Matrix::expand()
## x dplyr::filter()      masks stats::filter()
## x dplyr::lag()         masks stats::lag()
## x tidyr::pack()        masks Matrix::pack()
## x dplyr::src()         masks Hmisc::src()
## x recipes::step()      masks stats::step()
## x dplyr::summarize()   masks Hmisc::summarize()
## x parsnip::translate() masks Hmisc::translate()
## x tidyr::unpack()      masks Matrix::unpack()
## x recipes::update()    masks Matrix::update(), stats::update()
```

```r
library(broom)
library(mice)
```

```
##
## Attaching package: ’mice’

## The following object is masked from ’package:stats’:
##
##     filter

## The following objects are masked from ’package:base’:
##
##     cbind, rbind
```

```r
library(margins)
library(janitor)
```

```
##
## Attaching package: ’janitor’

## The following objects are masked from ’package:stats’:
##
##     chisq.test, fisher.test
```

```r
library(readxl)
library(rpart.plot)
```

```
## Loading required package: rpart

##
## Attaching package: ’rpart’

## The following object is masked from ’package:dials’:
##
##     prune
```

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v readr   1.4.0     v forcats 0.5.1
## v stringr 1.4.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x readr::col_factor() masks scales::col_factor()
## x purrr::discard()    masks scales::discard()
## x tidyr::expand()     masks Matrix::expand()
## x mice::filter()      masks dplyr::filter(), stats::filter()
## x stringr::fixed()    masks recipes::fixed()
## x dplyr::lag()        masks stats::lag()
## x tidyr::pack()       masks Matrix::pack()
## x readr::spec()       masks yardstick::spec()
## x dplyr::src()        masks Hmisc::src()
## x dplyr::summarize()  masks Hmisc::summarize()
## x tidyr::unpack()     masks Matrix::unpack()
```

```
final <- read_csv("final.csv",
                  # to avoid the cols specification warning message
                  col_types = cols(
  R0000100 = col_double(),
  R0532200 = col_double(),
  R0536300 = col_double(),
  R0538600 = col_double(),
  R0538700 = col_double(),
  R1201300 = col_double(),
  R1204500 = col_double(),
  R1204900 = col_double(),
  R1236201 = col_double(),
  R1302400 = col_double(),
  R1302500 = col_double(),
  U1990100 = col_double(),
  U1990700 = col_double()
)) %>%
  # finalize variable selection
  select(id = R0000100,
         wgt = R1236201,
         parents = R0532200, # filter for both biological parents
         gender = R0536300,
         race = R0538700,
         origin = R1201300,
         ed = U1990700, # highest degree attained
         edy = U1990100, # highest grade of school completed
         edf = R1302400, # biological father's highest grade of school completed
         edm = R1302500, # biological mother's highest grade of school completed
         pr = R1204900, # ratio of pinc to poverty level
         pinc = R1204500,
         hisp = R0538600) %>%
  # add some parameters
  filter(parents == 1, ## there are non biological parent versions of edf and edm, so this will only in
         edy != 95, ## 95 == school experiences were ungraded in the codebook
         edf != 95,
         edm != 95) %>%
  # recode for survey-wide non-response procedures
  mutate_all(~ ifelse(.x %in% c(-1, -2, -3, -4, -5), NA, .x)) %>%
  mutate(gender = ifelse(gender == "Male", 0, 1), # binarify gender
         white = ifelse(race == 1, 1, 0),
         black = ifelse(race == 2, 1, 0),
         indig = ifelse(race == 3, 1, 0),
         aapi = ifelse(race == 4, 1, 0),
         race = case_when(race == 1 ~ "White",
                          race == 2 ~ "Black or African American",
                          race == 3 ~ "American Indian, Eskimo, or Aleut",
                          race == 4 ~ "Asian or Pacific Islander",
                          hisp == 1 ~ "Hispanic"),
         race = ifelse(is.na(race), "Other", race),
         origin = case_when(origin == 1 ~ 0,
                            origin == 2 ~ 1), # country of birth; originally a citizenship question bas
         edy = case_when(ed == 0 & is.na(edy) ~ 0,
                         ed %in% c(1, 2) & is.na(edy) ~ 12,
                         ed == 3 & is.na(edy) ~ 14,
                         ed == 4 & is.na(edy) ~ 16,
                         ed == 5 & is.na(edy) ~ 18,
                         ed %in% c(6, 7) & is.na(edy) ~ 20,
                         TRUE ~ edy),
         ed = factor(ed, levels = 0:7,
                     labels = c("None",
                                "GED",
                                "HS",
                                "AA",
                                "BA or BS",
                                "MA or MS",
                                "PhD",
                                "DDS, JD, or MD")),
         college = ifelse(edy >= 16, 1, 0),
         pinc_rank = 100 * wtd.rank(x = pinc,
                                    weights = wgt,
                                    normwt = T)/sum(!is.na(pinc)), # code from classwork
         pr = pr/100,
         pr_group = as.integer(substr(pr, 1, 1)) + 1) %>%
  select(-parents) %>% # no longer needed now that filter and edf/edm are properly configured
  drop_na(origin, ed, edf, edm, pinc, pr) # needed for modeling
```

Analysis

```
glimpse(final)
```

```
## Rows: 2,096
## Columns: 19
## $ id        <dbl> 14, 25, 34, 36, 56, 58, 62, 63, 97, 102, 104, 105, 107, 108,~
## $ wgt       <dbl> 277820, 99208, 318843, 348666, 279754, 104863, 113580, 10137~
## $ gender    <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ race      <chr> "Hispanic", "Hispanic", "White", "White", "White", "Black or~
## $ origin    <dbl> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, ~
## $ ed        <fct> "GED", "GED", "MA or MS", "BA or BS", "PhD", "BA or BS", "MA~
## $ edy       <dbl> 12, 12, 18, 16, 20, 15, 18, 18, 18, 18, 16, 14, 20, 20, 18, ~
## $ edf       <dbl> 12, 12, 9, 16, 16, 12, 14, 14, 12, 12, 16, 16, 19, 19, 12, 1~
## $ edm       <dbl> 12, 12, 16, 15, 18, 12, 14, 14, 12, 12, 13, 13, 16, 16, 12, ~
## $ pr        <dbl> 0.00, 0.00, 0.60, 2.55, 6.96, 1.60, 9.43, 9.43, 2.50, 4.27, ~
## $ pinc      <dbl> 0, 0, 9800, 40500, 110700, 25500, 150000, 150000, 41100, 800~
## $ hisp      <dbl> 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ white     <dbl> 0, 0, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ black     <dbl> 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ indig     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ aapi      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ college   <dbl> 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, ~
## $ pinc_rank <dbl> 0.3790101, 0.3790101, 2.8640379, 30.2934689, 89.8144026, 13.~
## $ pr_group  <dbl> 1, 1, 1, 3, 7, 2, 10, 10, 3, 5, 4, 4, 7, 7, 4, 4, 7, 5, 5, 4~
```

```r
summary(final$edf)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   12.00   13.00   13.34   16.00   20.00
```

```r
summary(final$edm)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   12.00   13.00   13.22   15.00   20.00
```

```r
summary(final$pinc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0   35094   53040   61050   75500  246474
```

```r
summary(final$pinc_rank)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.379  22.948  47.423  47.900  71.900  98.131
```

```r
summary(final$pr)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.988   3.100   3.684   4.475  16.270
```

```r
summary(complete(mice(final, m = 1)))
```

```
##
##  iter imp variable
##   1   1
##   2   1
##   3   1
##   4   1
##   5   1
```

```
## Warning: Number of logged events: 2
```

```
##        id             wgt              gender      race
##  Min.   :  14   Min.   : 76071   Min.   :1   Length:2096
##  1st Qu.:2325   1st Qu.:223145   1st Qu.:1   Class :character
##  Median :4314   Median :276222   Median :1   Mode  :character
##  Mean   :4249   Mean   :246272   Mean   :1
##  3rd Qu.:5947   3rd Qu.:297684   3rd Qu.:1
##  Max.   :9016   Max.   :456751   Max.   :1
```

```
##
##      origin               ed           edy             edf       
##  Min.   :0.00000   AA      :713   Min.   : 3.00   Min.   : 2.00  
##  1st Qu.:0.00000   MA or MS:657   1st Qu.:14.00   1st Qu.:12.00  
##  Median :0.00000   BA or BS:276   Median :16.00   Median :13.00  
##  Mean   :0.03578   PhD     :258   Mean   :16.07   Mean   :13.34  
##  3rd Qu.:0.00000   HS      : 93   3rd Qu.:18.00   3rd Qu.:16.00  
##  Max.   :1.00000   GED     : 78   Max.   :20.00   Max.   :20.00  
##                    (Other) : 21                                  
##       edm              pr              pinc             hisp       
##  Min.   : 1.00   Min.   : 0.000   Min.   :     0   Min.   :0.0000  
##  1st Qu.:12.00   1st Qu.: 1.988   1st Qu.: 35094   1st Qu.:0.0000  
##  Median :13.00   Median : 3.100   Median : 53040   Median :0.0000  
##  Mean   :13.22   Mean   : 3.684   Mean   : 61050   Mean   :0.1384  
##  3rd Qu.:15.00   3rd Qu.: 4.475   3rd Qu.: 75500   3rd Qu.:0.0000  
##  Max.   :20.00   Max.   :16.270   Max.   :246474   Max.   :1.0000  
##                                                                    
##      white            black            indig               aapi        
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000000   Min.   :0.00000  
##  1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.000000   1st Qu.:0.00000  
##  Median :1.0000   Median :0.0000   Median :0.000000   Median :0.00000  
##  Mean   :0.7629   Mean   :0.1579   Mean   :0.004294   Mean   :0.01002  
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.000000   3rd Qu.:0.00000  
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.000000   Max.   :1.00000  
##                                                                          
##     college         pinc_rank         pr_group    
##  Min.   :0.0000   Min.   : 0.379   Min.   : 1.00  
##  1st Qu.:0.0000   1st Qu.:22.948   1st Qu.: 2.00  
##  Median :1.0000   Median :47.423   Median : 3.00  
##  Mean   :0.5596   Mean   :47.900   Mean   : 3.76  
##  3rd Qu.:1.0000   3rd Qu.:71.900   3rd Qu.: 5.00  
##  Max.   :1.0000   Max.   :98.131   Max.   :10.00  
##
```

```
# set formula
fm <- college ~ edf + edm + pinc + race + pr

# set linear model
linear_mod <- lm(fm,
                 data = final)
linear_mod
```

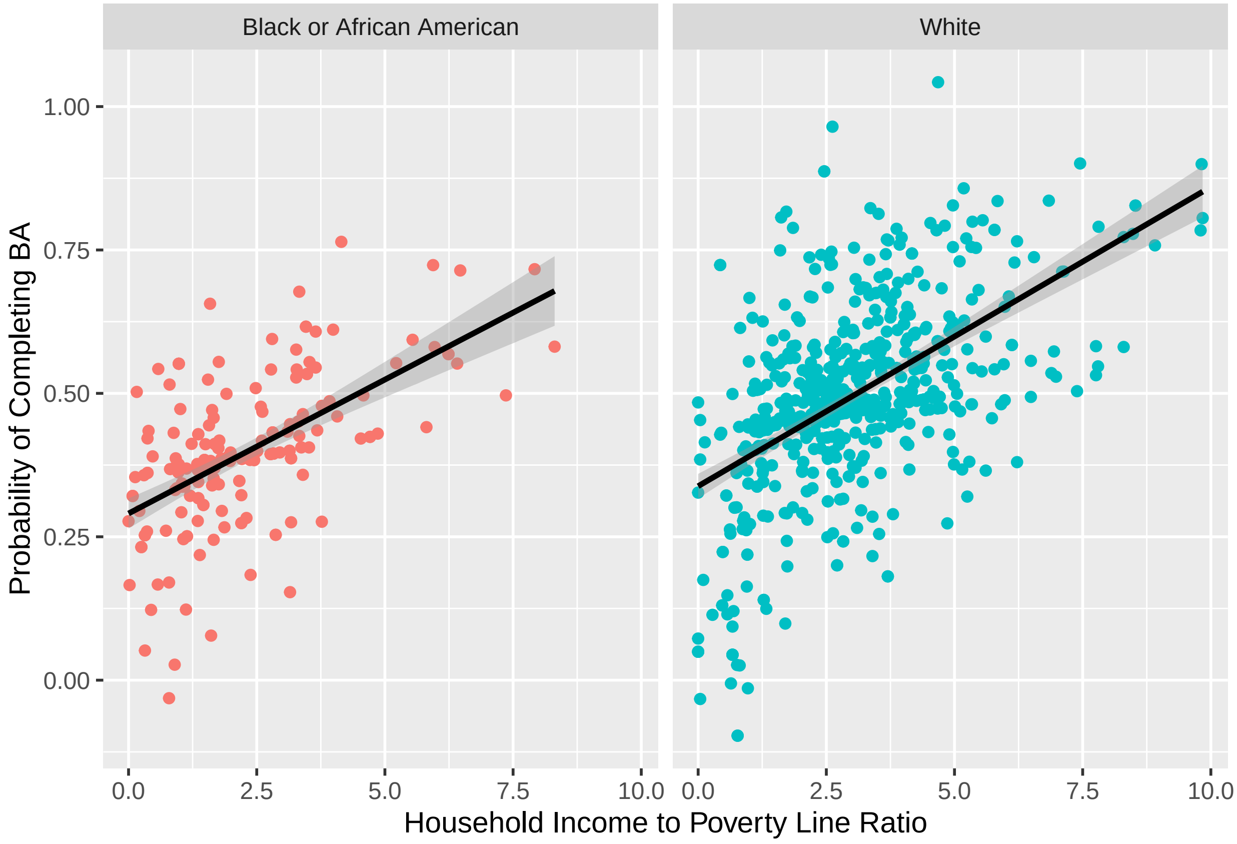
```
##
## Call:
## lm(formula = fm, data = final)
##
## Coefficients:
##                    (Intercept)                            edf  
##                     -0.3950502                      0.0371714  
##                            edm                           pinc  
##                      0.0314440                      0.0000013  
## raceAsian or Pacific Islander  raceBlack or African American  
##                      0.2651893                     -0.0769029  
##                   raceHispanic                      raceOther  
##                    -0.0344453                        0.0012376
##                     raceWhite                               pr
##                    -0.0069968                       -0.0052543
```

```
# set linear fit
linear_fit <- augment(linear_mod,
                      se_fit = TRUE,
                      type.predict = "response")
linear_fit
```

```
## # A tibble: 2,096 x 13
##    college   edf   edm   pinc race      pr .fitted .se.fit .resid    .hat .sigma
##      <dbl> <dbl> <dbl>  <dbl> <chr>  <dbl>   <dbl>   <dbl>  <dbl>   <dbl>  <dbl>
##  1       0    12    12      0 Hispa~  0      0.394  0.0445 -0.394 9.60e-3  0.454
##  2       0    12    12      0 Hispa~  0      0.394  0.0445 -0.394 9.60e-3  0.454
##  3       1     9    16   9800 White   0.6    0.445  0.0326  0.555 5.15e-3  0.454
##  4       1    16    15  40500 White   2.55   0.704  0.0181  0.296 1.60e-3  0.454
##  5       1    16    18 110700 White   6.96   0.866  0.0218  0.134 2.31e-3  0.454
##  6       0    12    12  25500 Black~  1.6    0.376  0.0256 -0.376 3.18e-3  0.454
##  7       1    14    14 150000 Black~  9.43   0.634  0.0359  0.366 6.27e-3  0.454
##  8       1    14    14 150000 Black~  9.43   0.634  0.0359  0.366 6.27e-3  0.454
##  9       1    12    12  41100 White   2.5    0.462  0.0135  0.538 8.83e-4  0.454
## 10       1    12    12  80000 White   4.27   0.503  0.0169  0.497 1.38e-3  0.454
## # ... with 2,086 more rows, and 2 more variables: .cooksd <dbl>,
## #   .std.resid <dbl>
```

```
# plot
linear_fit %>%
  filter(college < 1,
         race %in% c("White", "Black or African American"),
         pr < 16) %>%
  ggplot(aes(x = pr,
             y = .fitted)) +
  geom_point(aes(color = race)) +
  geom_smooth(method = "lm",
              color = "black") +
  scale_x_continuous("Household Income to Poverty Line Ratio") +
  scale_y_continuous("Probability of Completing BA") +
  theme(legend.position = "none") +
  facet_wrap(~ race)
```

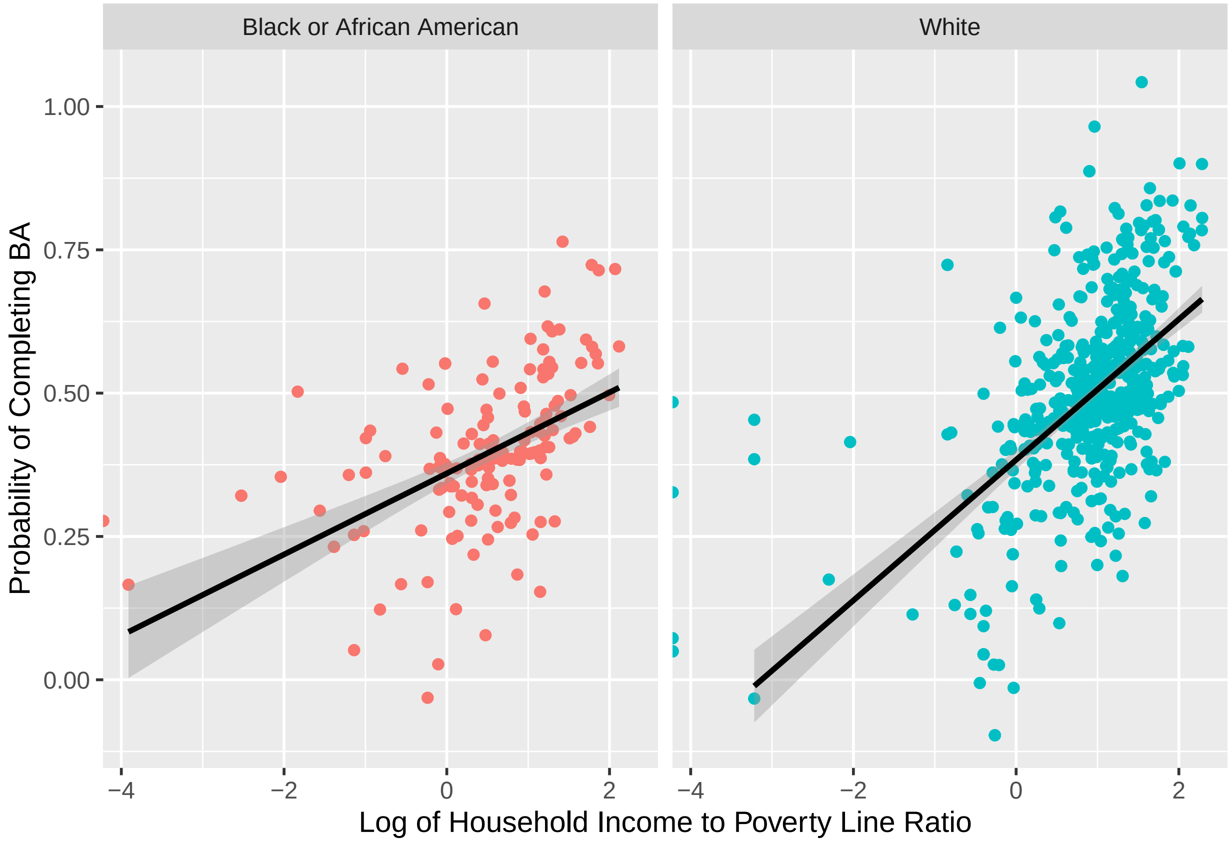
```
## `geom_smooth()` using formula 'y ~ x'
```



```r
# plot
linear_fit %>%
  filter(college < 1,
         race %in% c("White", "Black or African American"),
         pr < 16) %>%
  ggplot(aes(x = log(pr),
             y = .fitted)) +
  geom_point(aes(color = race)) +
  geom_smooth(method = "lm",
              color = "black") +
  scale_x_continuous("Log of Household Income to Poverty Line Ratio") +
  scale_y_continuous("Probability of Completing BA") +
  theme(legend.position = "none") +
  facet_wrap(~ race)
```

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

```
## Warning: Removed 7 rows containing non-finite values (stat_smooth).
```



```
# set logit model
logit_mod <- glm(fm, family = binomial("logit"), data = final)

tidy(logit_mod)
```

```
## # A tibble: 10 x 5
##    term                            estimate  std.error statistic  p.value
##    <chr>                              <dbl>      <dbl>     <dbl>    <dbl>
##  1 (Intercept)                   -4.59       0.760        -6.04   1.50e- 9
##  2 edf                            0.186      0.0224        8.31   9.58e-17
##  3 edm                            0.158      0.0253        6.26   3.80e-10
##  4 pinc                           0.00000694 0.00000441    1.57   1.16e- 1
##  5 raceAsian or Pacific Islander  1.83       0.958         1.91   5.56e- 2
##  6 raceBlack or African American -0.307      0.696        -0.441  6.59e- 1
##  7 raceHispanic                  -0.163      0.719        -0.227  8.20e- 1
##  8 raceOther                     -0.0267     0.848        -0.0315 9.75e- 1
##  9 raceWhite                     -0.0148     0.688        -0.0215 9.83e- 1
## 10 pr                            -0.0141     0.0646       -0.218  8.27e- 1
```

```
# one unit increase in X is associated with a beta unit change in log(p/(1-p))
# average marginal effect (AME): average effect in the prob scale

margins_summary(logit_mod)
```

```
##                            factor      AME      SE        z       p    lower   upper
##                             edf  0.0379 0.0043  8.8399 0.0000  0.0295 0.0463
##                             edm  0.0323 0.0050  6.4796 0.0000  0.0225 0.0420
##                            pinc  0.0000 0.0000  1.5769 0.1148 -0.0000 0.0000
##                              pr -0.0029 0.0132 -0.2182 0.8273 -0.0287 0.0229
##  raceAsian or Pacific Islander  0.2944 0.1565  1.8815 0.0599 -0.0123 0.6010
##  raceBlack or African American -0.0635 0.1427 -0.4452 0.6561 -0.3432 0.2161
##                    raceHispanic -0.0337 0.1475 -0.2287 0.8191 -0.3227 0.2553
##                       raceOther -0.0055 0.1739 -0.0315 0.9749 -0.3463 0.3353
##                       raceWhite -0.0030 0.1410 -0.0215 0.9828 -0.2794 0.2733
```

```
logit_fit <- augment(logit_mod, se_fit = TRUE,
                     type.predict = "response")
```

```
# set factored formula
ffm <- factor(college) ~ edf + edm + pinc + race + pr

# set logistic model
logistic_mod <- logistic_reg() %>%
  set_engine("glm")

logistic_fit <- fit(logistic_mod,
                    ffm,
                    data = final)
logistic_fit
```

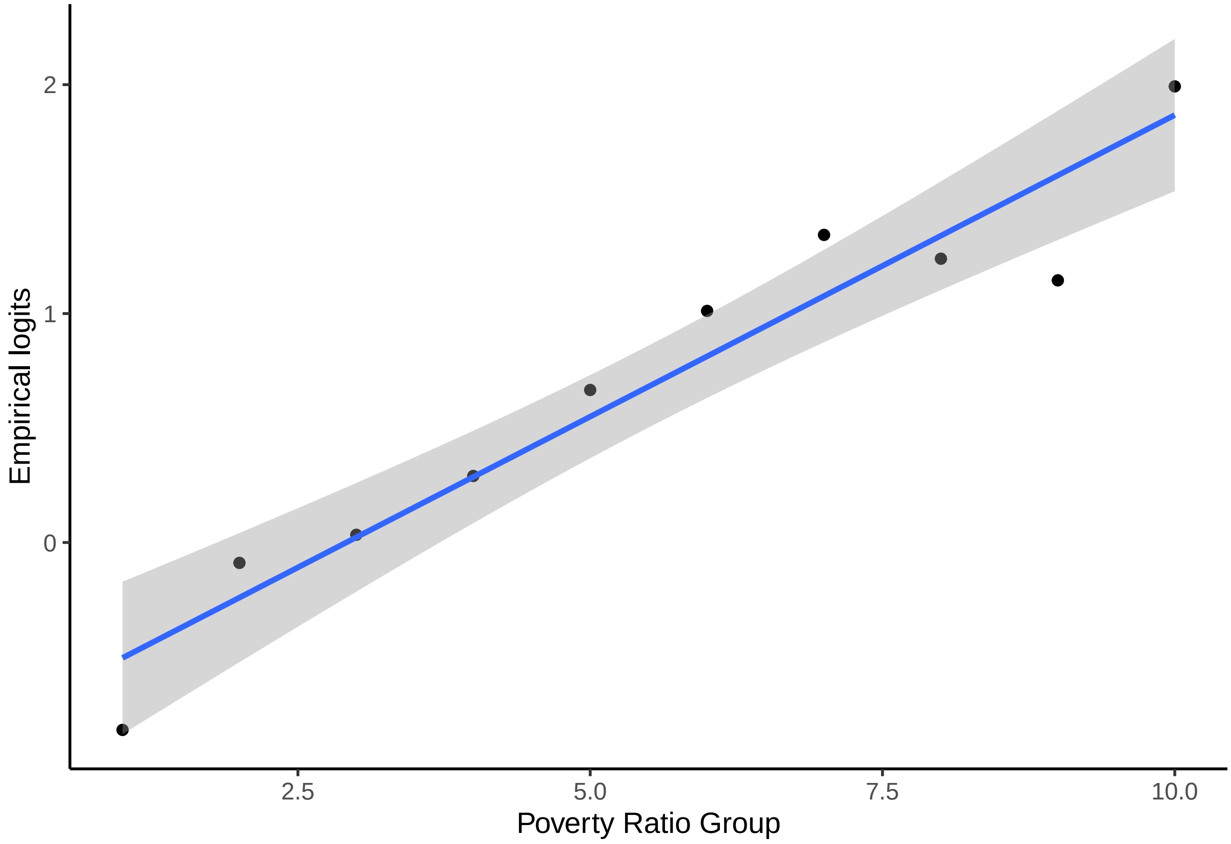
```
## parsnip model object
##
## Fit time:  15ms
##
## Call:  stats::glm(formula = factor(college) ~ edf + edm + pinc + race +
##     pr, family = stats::binomial, data = data)
##
## Coefficients:
##                    (Intercept)                            edf
##                     -4.593e+00                      1.861e-01
##                            edm                           pinc
##                      1.584e-01                      6.936e-06
## raceAsian or Pacific Islander  raceBlack or African American
##                      1.834e+00                     -3.069e-01
##                   raceHispanic                      raceOther
##                     -1.634e-01                     -2.670e-02
##                      raceWhite                             pr
##                     -1.479e-02                     -1.408e-02
##
## Degrees of Freedom: 2095 Total (i.e. Null);  2086 Residual
## Null Deviance:       2876
## Residual Deviance: 2478  AIC: 2498
```

```
logistic_fit %>%
  tidy(conf.int = TRUE) %>%
  select(term, estimate, conf.low, conf.high)
```

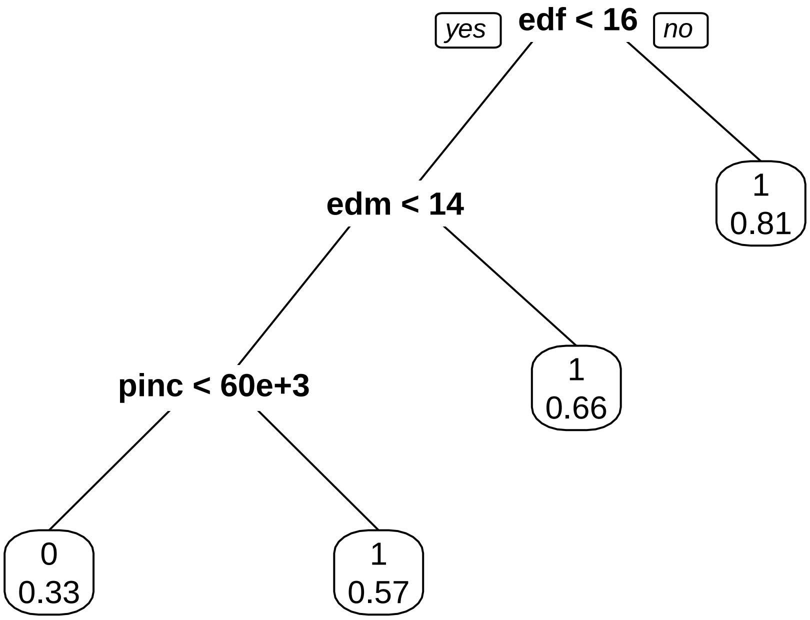
```
## # A tibble: 10 x 4
##    term                             estimate    conf.low  conf.high
##    <chr>                               <dbl>       <dbl>      <dbl>
##  1 (Intercept)                   -4.59       -6.10       -3.05
##  2 edf                            0.186       0.143       0.231
##  3 edm                            0.158       0.109       0.208
##  4 pinc                           0.00000694 -0.00000188  0.0000155
##  5 raceAsian or Pacific Islander  1.83       -0.0244      3.81
##  6 raceBlack or African American -0.307      -1.74        1.07
##  7 raceHispanic                  -0.163      -1.64        1.26
##  8 raceOther                     -0.0267     -1.73        1.66
##  9 raceWhite                     -0.0148     -1.44        1.35
## 10 pr                            -0.0141     -0.137       0.117
```

```
final %>%
  group_by(pr_group) %>%
  summarise(pct_success = sum(college)/n()) %>%
  mutate(emp_logit = qlogis(pct_success)) %>%
  ggplot(aes(pr_group, emp_logit)) +
  geom_point() +
  geom_smooth(method = lm) +
  theme_classic() +
  labs(
    x = "Poverty Ratio Group",
    y = "Empirical logits")
```

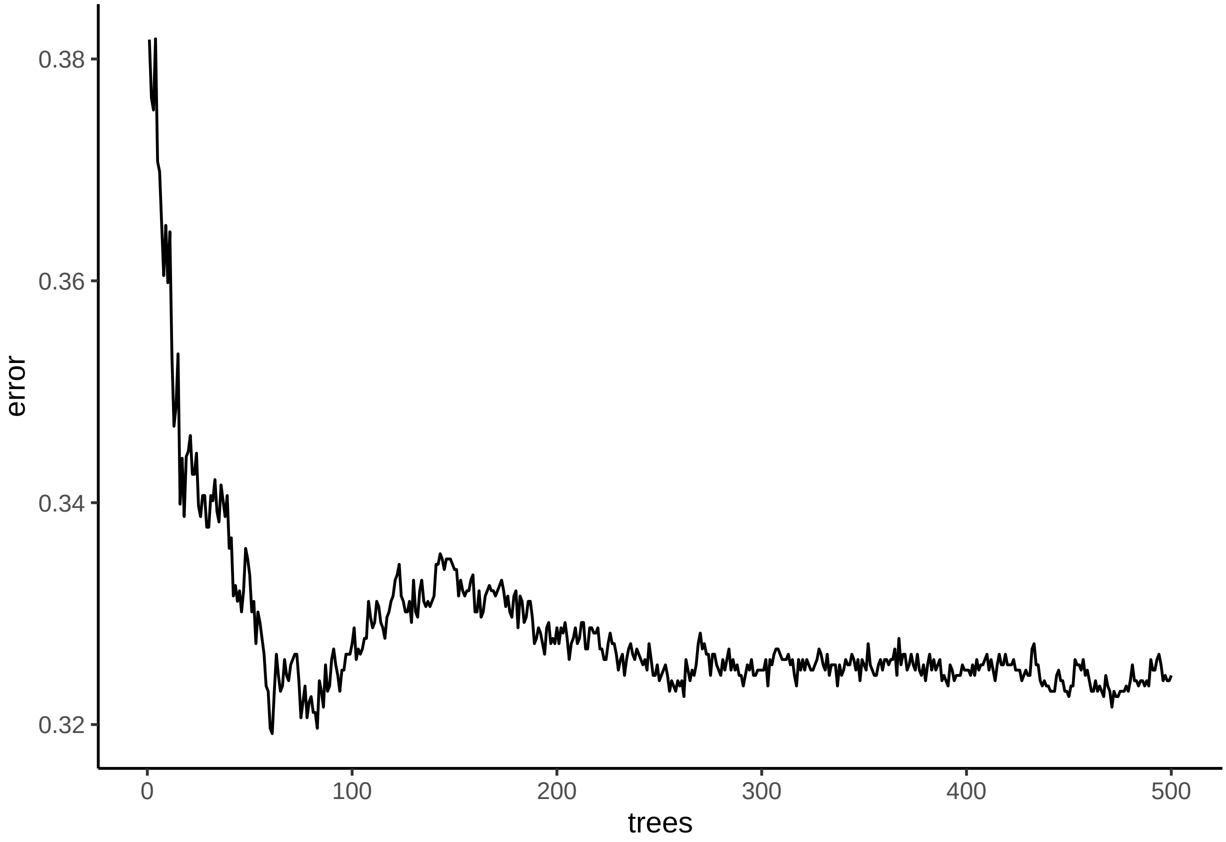
```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



```r
# set tree model
tree_mod <- decision_tree() %>%
  set_engine("rpart",
             model = TRUE) %>%
  set_mode("classification")

# set tree fit
house_region_tree <- fit(tree_mod, ffm, data = final)

# plot
house_region_tree$fit %>%
  prp(extra = 6, varlen = 0, faclen = 0)
```



```r
# set forest model
forest_mod <- rand_forest() %>%
  set_engine("randomForest") %>%
  set_mode("classification")

# set forest fit
house_forest <- fit(forest_mod, ffm, data = final)

# plot
tibble(error = house_forest$fit$err.rate[, "OOB"],
       trees = 1:500) %>%
  ggplot(aes(x = trees, y = error)) +
  geom_line() +
  theme_classic()
```



Results

Ultimately, the filtered data utilized for this research question yielded two consistent variables for being correlated to college graduation: the education backgrounds of biological mothers and fathers of individuals. Across linear, log linear, logit, logistic regression, and decision tree and randomForest models, the education backgrounds of biological mothers and fathers consistently produced confidence intervals that indicate statistical significance. Surprisingly, race, parent income, and poverty group did not get incorporated into the models above as impactful enough factors. I attribute this to two limitations: (1) needing a more complex model that can accurately gauge the exact relationship these three variables have on college completion, such as a confounding scenario in which the education backgrounds of the individuals' parents also depend on race, parent income and poverty group (one posit is that the lower the parent income and poverty group is for the individual, the more likely the parents themselves were also in that bracket when achieving their highest grade of schooling measurable by the NLSY97); and (2) constraints found in the dataset itself, with over 75% of its survey respondents being white. This is bolstered by the log linear relationship plotted above in the first code chunk, in which the Black proportion of the data (the second largest racial demographic after white respondents) has a significantly smaller slope for its regression line than that of the white respondents. Another unusual observation is the tertiary decision tier for the randomForest plot, in which they add parent income as a significant threshold; as a caveat, the error margin for this tree model does plateau at around 0.33.

Conclusion

The relationship between college completion and parent background may be too cloudy of a question for American data, as college degrees are becoming more and more normalized and data on parent background becomes more and more clustered in coverage towards survey methodology that favors white respondents. For future study, college entrance among immigrant people of color may be a more specific research question to tackle, with more nuanced covariates to observe such as whether gentrification in residential areas affect high school entrance, how secondary schooling is funded such that different resources terrace individuals into different likelihoods of college entrance, how clear distinctions of college type can get beyond the aggregated college degree type, and how resources and their accessibility outside of high school preparation for college affects college entrance.

References

Assari, S. (2018). Parental education better helps white than black families escape poverty: National survey of Children’s health. Economies, 6(2).

Bailey, M.J. and Dynarski, S.M. (2011). Gains and gaps: Changing inequality in US college entry and completion, No.17633. National Bureau of Economic Research.

Duncan, O.D. (1968). Inheritance of race or inheritance of poverty?. On understanding poverty, pp.85-110.

Hilger, N. (2015). The Great Escape: Intergenerational Mobility in the United States Since 1940, No.21217. National Bureau of Economic Research.

Reardon, S.F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. Whither opportunity, pp.91-116.