The Effect of Education Level on Income Inequality

Names + Studentnumbers

2025-06-20

#Set up your environment

setwd("~/")

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.4 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.4   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(renv)

##   
## Attaching package: 'renv'  
##   
## The following object is masked from 'package:purrr':  
##   
## modify  
##   
## The following objects are masked from 'package:stats':  
##   
## embed, update  
##   
## The following objects are masked from 'package:utils':  
##   
## history, upgrade  
##   
## The following objects are masked from 'package:base':  
##   
## autoload, load, remove

library(dplyr)

packageVersion("tidyverse")

## [1] '2.0.0'

packageVersion("renv")

## [1] '1.1.4'

packageVersion("dplyr")

## [1] '1.1.4'

#Title Page Carlijn Calori, Leah Delikát, Fadhil Dhafir, Sten Groen, Julia Koeleman, Anne Schrama, Marie-Louise Stevens, Sophia Zentgraf Tutorial Group 1 J.F. Fitzgerald

#Part 1 - Identification of the Social Problem

##1.1 Describe the Social Problem Income inequality in the USA is a social problem because it limits equal opportunities, increases social tensions, and harms health and democracy.The Pew Research Center warns that the middle class is shrinking rapidly, while the rich grow richer.In addition, The OECD ranks the U.S. among the most unequal developed nations. These two studies emphasize the importance of this problem ABCDEF.

<https://www.pewresearch.org/social-trends/2020/01/09/trends-in-income-and-wealth-inequality/>? <https://www.oecd.org/en/publications/inequality-and-poverty-in-the-united-states_5k46957cwv8q-en.html>

#Part 2 - Data Sourcing

##2.1 Load in Data

library(readxl)  
Distribution <- read\_excel("data/Frank\_WID\_2020.xls", sheet = 3)  
Thresholds <- read\_excel("data/Frank\_WID\_2020.xls", sheet = 4)  
Education <- read\_excel("data/tabn012.xls", col\_types = c("text", "skip", "text", "skip", "skip", "skip", "skip", "skip", "skip", "skip", "skip", "skip", "skip", "skip", "text", "skip", "skip", "skip", "skip", "skip", "skip", "skip", "skip", "skip", "skip", "skip", "text", "skip", "skip", "skip", "skip", "skip", "skip","skip", "skip", "skip", "skip", "skip", "text", "skip", "skip", "skip", "skip", "skip", "skip", "skip", "skip", "skip", "skip"))

## New names:  
## • `` -> `...2`  
## • `` -> `...3`  
## • `` -> `...4`  
## • `` -> `...5`

##2.2 Clean the data sets

Distribution <- Distribution %>%  
 filter(Year %in% 1990:2000,   
 State != c("United States", "District of Columbia"))  
  
Thresholds <- Thresholds %>%  
 filter(Year %in% 1990:2000,   
 State != c("United States", "District of Columbia"))  
  
Education <- Education[ -c(1:13, 14, 20, 24, 26, 32, 38, 44, 50, 56, 62, 68, 75:79), ]  
  
names(Education)[1:5] <- c("State", "Highschool1990", "Highschool2000", "Bachelor1990", "Bachelor2000")  
  
Education <- Education %>%   
 mutate(State = State %>%  
 str\_replace\_all("…+", "") %>%  
 str\_replace\_all("\\.+", "") %>%  
 str\_trim()) # to delete all the dots after the state name

##2.3 Reshape the Education dataset to the same form as the others so you have a variable Year

df1990 <- data.frame(State = Education$State, Year = 1990, Highschool = Education$Highschool1990, Bachelor = Education$Bachelor1990, stringsAsFactors = FALSE)  
  
df2000 <- data.frame(State = Education$State,Year = 2000,Highschool = Education$Highschool2000,Bachelor = Education$Bachelor2000,stringsAsFactors = FALSE)  
  
Education <- rbind(df1990, df2000)  
  
Education <- Education %>%  
 mutate(across(3:4, ~ as.numeric(.)))

##2.4 To merge the datasets

Allmerged <- Distribution %>%  
 inner\_join(Thresholds, by = c("State", "Year")) %>%  
 full\_join(Education, by = c("State", "Year"))  
  
Allmerged <- Allmerged %>%  
 select(-st.x, -st.y) %>%   
 mutate(across(where(is.numeric), ~ round(.x, 2)))

##2.5 Short summary of dataset

head(Allmerged)

## # A tibble: 6 × 16  
## Year State Top10\_adj Top5\_adj Top1\_adj Top05\_adj Top01\_adj Top001\_adj y90  
## <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1990 Alaba… 39.9 26.0 12.4 9.1 4.74 1.78 1.02e5  
## 2 1990 Alaska 34.0 22.7 11.2 8.74 4.93 2.15 1.56e5  
## 3 1990 Arizo… 38.9 26.5 12.3 9.25 4.56 1.64 1.11e5  
## 4 1990 Arkan… 38.2 25.6 12.4 8.84 4.72 1.8 9.18e4  
## 5 1990 Calif… 42.8 30.6 16.6 12.5 6.82 2.84 1.31e5  
## 6 1990 Color… 36.4 25.4 12.0 9.16 4.74 1.82 1.20e5  
## # ℹ 7 more variables: y95 <dbl>, y99 <dbl>, y995 <dbl>, y999 <dbl>,  
## # y9999 <dbl>, Highschool <dbl>, Bachelor <dbl>

This is our merged data set. In the first column you can see what year the data refers to. Next to it is the column with the State\_ID. We put the states in alphabetical order and then started numbering from 1 to 100 (example: Alabama is numbered 1 in 1990 and 51 (1+50) in 2000).

The next 5 columns show the percentages of total income that a group earns. For example, the first column shows that the top 10 percent of big earners, earn 39.9% of total income, from the state. This is the same for the 4 other columns, but here the groups get smaller and smaller.

Columns 9 through 14 show how much you have to earn to belong to that percentage of big earners. For example, in 1990 you had to earn more than $101,762.24 to be among the top 10 big earners in Alabama. Did you want to be among the top 5 in 1990? Then you had to earn more than $131,506.30.

The last two columns show level of education. These variables are in percent. So for example, in 1990, 66.9% of the population in Alabama had a high school diploma or higher and 15.7% had a bachelor’s degree or higher.

<<<<<<< HEAD

#Part 3 - Quantifying

##3.1 Generating necessary variables

To measure inequality: top 10% divided by bottom 90%.Not an ideal measuring, but we don’t have more data on the distribution if income.

Allmerged <- Allmerged %>%  
 mutate(Top10\_vs\_bottom90 = round(Top10\_adj/(100-Top10\_adj), 2))

##3.2 Spatial variation of change in income inequality

# Make sure these packages are installed and loaded  
library(dplyr)  
library(tidyr)  
library(maps)

##   
## Attaching package: 'maps'

## The following object is masked from 'package:purrr':  
##   
## map

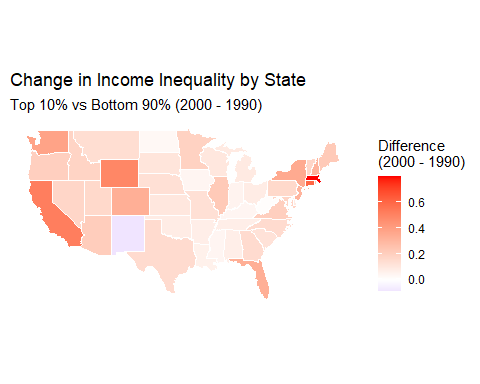
library(ggplot2)  
  
diff\_data <- Allmerged %>%  
 filter(Year %in% c(1990, 2000)) %>%  
 select(State, Year, Top10\_vs\_bottom90) %>%  
 pivot\_wider(names\_from = Year, values\_from = Top10\_vs\_bottom90) %>%  
 mutate(  
 # Extract only the real state name before any dots or extra characters  
 state\_lower = tolower(gsub("^([a-zA-Z ]+).\*", "\\1", State)),  
 state\_lower = trimws(state\_lower),  
 diff = `2000` - `1990`  
 ) %>%  
 filter(state\_lower != "district of columbia")  
  
us\_states <- map\_data("state")  
map\_plot\_data <- left\_join(us\_states, diff\_data, by = c("region" = "state\_lower"))  
  
# Check: should now match!  
print(setdiff(sort(unique(us\_states$region)), sort(unique(diff\_data$state\_lower))))

## [1] "district of columbia"

print(setdiff(sort(unique(diff\_data$state\_lower)), sort(unique(us\_states$region))))

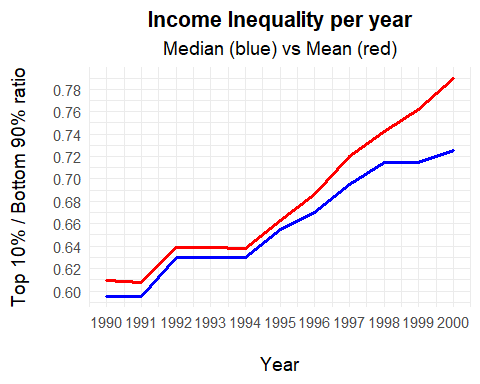
## [1] "alaska" "hawaii"

ggplot(map\_plot\_data, aes(long, lat, group = group, fill = diff)) +  
 geom\_polygon(color = "white") +  
 coord\_fixed(1.3) +  
 scale\_fill\_gradient2(  
 low = "blue", mid = "white", high = "red", midpoint = 0,  
 name = "Difference\n(2000 - 1990)"  
 ) +  
 labs(  
 title = "Change in Income Inequality by State",  
 subtitle = "Top 10% vs Bottom 90% (2000 - 1990)"  
 ) +  
 theme\_minimal() +  
 theme(  
 axis.text = element\_blank(),  
 axis.title = element\_blank(),  
 panel.grid = element\_blank()  
 )



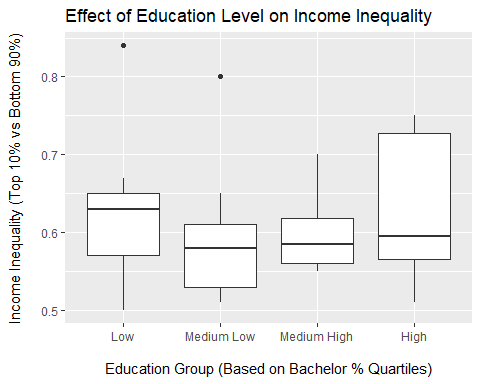
##3.3 Temporal visualization of income inequality

df\_summary <- Allmerged %>%  
 group\_by(Year) %>%  
 summarise(median\_ineq = median(Top10\_vs\_bottom90),  
 mean\_ineq = mean(Top10\_vs\_bottom90))  
  
ggplot(df\_summary, aes(x = Year)) +  
 geom\_line(aes(y = median\_ineq), color = "blue",   
 linewidth = 1.2,   
 lineend = "round") +  
 geom\_line(aes(y = mean\_ineq),   
 color = "red",  
 linewidth = 1.2,  
 lineend = "round") +  
 labs(title = "Income Inequality per year",   
 subtitle = "Median (blue) vs Mean (red)",  
 x = "\nYear",   
 y = "Top 10% / Bottom 90% ratio\n") +  
 scale\_x\_continuous(limits = c(1990, 2000), breaks = 1990:2000) +  
 scale\_y\_continuous(breaks = seq(0.5, 0.8, by = 0.02)) +  
 theme\_minimal(base\_size = 14) +  
 theme(plot.title = element\_text(hjust = 0.5, face = "bold"),  
 plot.subtitle = element\_text(hjust = 0.5))



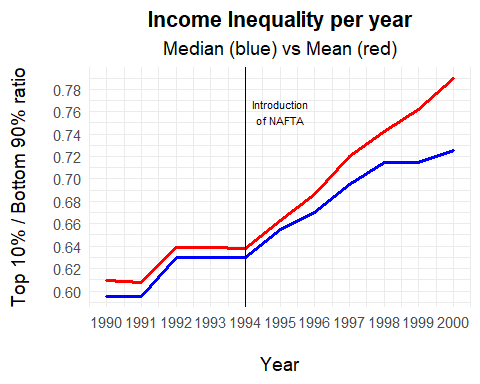
##3.4 To visualize income inequality across different percentages of bachelor’s degree, we divide them into quartiles.

library(ggplot2)  
  
#Create quartile-based education groups per year  
Allmerged <- Allmerged %>%  
 group\_by(Year) %>%  
 mutate(Quartile = ntile(Bachelor, 4),  
 Bachelor\_level = case\_when(  
 Quartile == 1 ~ "Low",  
 Quartile == 2 ~ "Medium Low",  
 Quartile == 3 ~ "Medium High",  
 Quartile == 4 ~ "High"  
 ))  
Allmerged <- Allmerged %>%  
 mutate(Quartile = NULL)  
  
# Make sure Education is a factor with the right order  
Allmerged$Bachelor\_level <- factor(  
 Allmerged$Bachelor\_level,  
 levels = c("Low", "Medium Low", "Medium High", "High")  
)  
  
#We can use indexing or create new dataframes for 1990 and 2000 to create boxplots per year  
library(ggplot2)  
  
ggplot(data = Allmerged[Allmerged$Year == 1990, ], aes(x = Bachelor\_level, y = Top10\_vs\_bottom90)) +  
 geom\_boxplot() +  
 labs(  
 x = "\nEducation Group (Based on Bachelor % Quartiles)",  
 y = "Income Inequality (Top 10% vs Bottom 90%)\n",  
 title = "Effect of Education Level on Income Inequality"  
 )



##3.5 Event analysis

df\_summary <- Allmerged %>%  
 group\_by(Year) %>%  
 summarise(median\_ineq = median(Top10\_vs\_bottom90),  
 mean\_ineq = mean(Top10\_vs\_bottom90))  
  
ggplot(df\_summary, aes(x = Year)) +  
 geom\_line(aes(y = median\_ineq), color = "blue",   
 linewidth = 1.2,   
 lineend = "round") +  
 geom\_line(aes(y = mean\_ineq),   
 color = "red",  
 linewidth = 1.2,  
 lineend = "round") +  
 labs(title = "Income Inequality per year",   
 subtitle = "Median (blue) vs Mean (red)",  
 x = "\nYear",   
 y = "Top 10% / Bottom 90% ratio\n") +  
 scale\_x\_continuous(limits = c(1990, 2000), breaks = 1990:2000) +  
 scale\_y\_continuous(breaks = seq(0.5, 0.8, by = 0.02)) +  
 theme\_minimal(base\_size = 14) +  
 theme(plot.title = element\_text(hjust = 0.5, face = "bold"),  
 plot.subtitle = element\_text(hjust = 0.5)) +  
 geom\_vline(xintercept = 1994) +  
 annotate("text", x = 1995, y = 0.76, size = 3, label = "Introduction\nof NAFTA")



#4 Discussion

##4.1 Discuss our findings

#5 Reproductibility

##5.2 Reference list Pew Research Center. (2020, January 9). Trends in U.S. income and wealth inequality. Pew Research Center. Retrieved from <https://www.pewresearch.org/social-trends/2020/01/09/trends-in-income-and-wealth-inequality/>

Denk, O., Hagemann, R., Lenain, P., & Somma, V. (2013, May 27). Inequality and poverty in the United States: Public policies for inclusive growth (OECD Working Paper No. 1052). OECD. Retrieved from <https://www.oecd.org/en/publications/inequality-and-poverty-in-the-united-states_5k46957cwv8q-en.html>