Lab 4

Bangda Sun, bs2996

October 25, 2016

Instructions

We'd like you to knit this lab as a `.pdf` file today. From now on, we'll hand in homeworks as `.pdf`s so we'll use lab today as an opportunity to practice this. Include output for each question in its own individual code chunk and don't print out any vector that has more than 20 elements.

Objectives: Importing and manipulating data; writing functions to estimate parameters; writing functions to check model fit.

Background

On the exam we looked at a dataset containing information on America's richest people. In this lab we continue to look at the very rich by turning to a more systematic data source than Forbes magazine, the World Top Incomes Database (WTID) hosted by the Paris School of Economics [http://topincomes.g-mond.parisschoolofeconomics.eu]. This is derived from income tax reports, and compiles information about the very highest incomes in various countries over time, trying as hard as possible to produce numbers that are comparable across time and space. For most countries in most time periods, the upper end of the income distribution roughly follows a Pareto distribution, with probability density function

$$f(x) = \frac{(a-1)}{x_{min}} \left(\frac{x}{x_{min}}\right)^{-a}$$

for incomes $X \geq x_{min}$. (Typically, x_{min} is large enough that only the richest 3%-4% of the population falls above it.) As the *Pareto exponent*, a, gets smaller, the distribution of income becomes more unequal, that is, more of the population's total income is concentrated among the very richest people.

The proportion of people whose income is at least x_{min} whose income is also at or above any level $w \geq x_{min}$ is thus

$$\mathbf{Pr}(X \geq w) = \int_w^\infty f(x)dx = \int_w^\infty \frac{(a-1)}{x_{min}} \left(\frac{x}{x_{min}}\right)^{-a} dx = \left(\frac{w}{x_{min}}\right)^{-a+1}.$$

We will use this to estimate how income inequality changed in the US over the last hundred years or so. (Whether the trends are good or bad or a mix is beyond our scope here.) WTID exports its data sets as `.xlsx` spreadsheets. For this lab session, we have extracted the relevant data and saved it as `wtid-report.csv`.

Part 1

1. Open the file and make a new variable containing only the year, "P99", "P99.5" and "P99.9" variables; these are the income levels which put someone at the 99th, 99.5th, and 99.9th, percentile of income. What was P99 in 1972? P99.5 in 1942? P99.9 in 1922? You must identify these using your code rather than looking up the values manually. (You may want to modify the column names to make some of them shorter.)

```
setwd("C://Users//Bangda//Desktop//GR5206 Materials//GR5206 Lab4")
wtiddata <- read.csv("wtid-report.csv", header = TRUE)
head(wtiddata)
```

```
##         Country Year P99.income.threshold P99.5.income.threshold
## 1 United States 1913             82677.22               135583.5
## 2 United States 1914             76405.62               126910.5
## 3 United States 1915             64409.44               122555.7
## 4 United States 1916             77289.78               138102.3
## 5 United States 1917             95326.69               154537.8
## 6 United States 1918             95202.66               147850.1
##   P99.9.income.threshold
## 1               428630.4
## 2               410528.7
## 3               451668.3
## 4               518327.4
## 5               536356.5
## 6               457045.0
```

```
# create a new variable
wtiddata2 <- wtiddata[,-1]
names(wtiddata2) <- c("year", "P99", "P99.5", "P99.9")
# P99 in 1972
wtiddata2[wtiddata2$year == 1972, "P99"]
```

```
## [1] 215836.3
```

```
# P99.5 in 1942
wtiddata2[wtiddata2$year == 1942, "P99.5"]
```

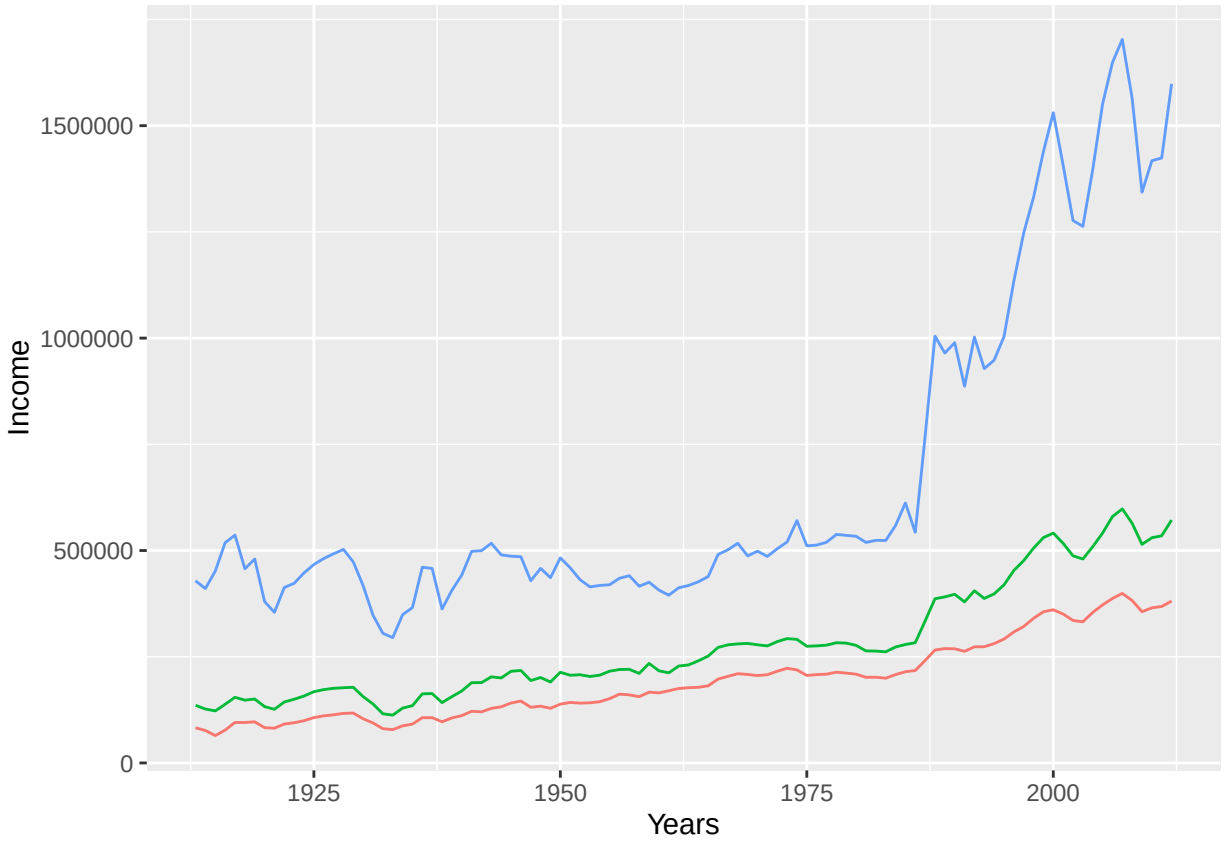
```
## [1] 189140.6
```

```
# P99.9 in 1922
wtiddata2[wtiddata2$year == 1922, "P99.9"]
```

```
## [1] 413153.5
```

2. Plot the three percentile levels against time using `ggplot`. Make sure the axes are labeled appropriately, and in particular that the horizontal axis is labeled with years between 1913 and 2012, not just numbers from 1 to 100. Remember `library(ggplot2)`. In my plot I used multiple layers of `geom_line` and didn't include a lengend (but plotted the years in different colors).

```
library(ggplot2)
ggplot(data = wtiddata2[wtiddata2$year <= 2012, ],
       aes(x = year, y = value, color = variable)) +
  geom_line(mapping = aes(y = P99, col = "P99"), show.legend = FALSE) +
  geom_line(mapping = aes(y = P99.5, col = "P99.5"), show.legend = FALSE) +
  geom_line(mapping = aes(y = P99.9, col = "P99.9"), show.legend = FALSE) +
  labs(x = "Years", y = "Income")
```



3. One can show from the earlier equations that one can estimate the exponent by the formula

$$a = 1 - \frac{\log 10}{\log\left(\frac{P99}{P99.9}\right)} \tag{1}$$

Write a function, `exponent.est_ratio()` which takes in values for `P99` and `P99.9`, and returns the value of a implied by (1). Check that if `P99=1e6` and `P99.9=1e7`, your function returns an a of 2.

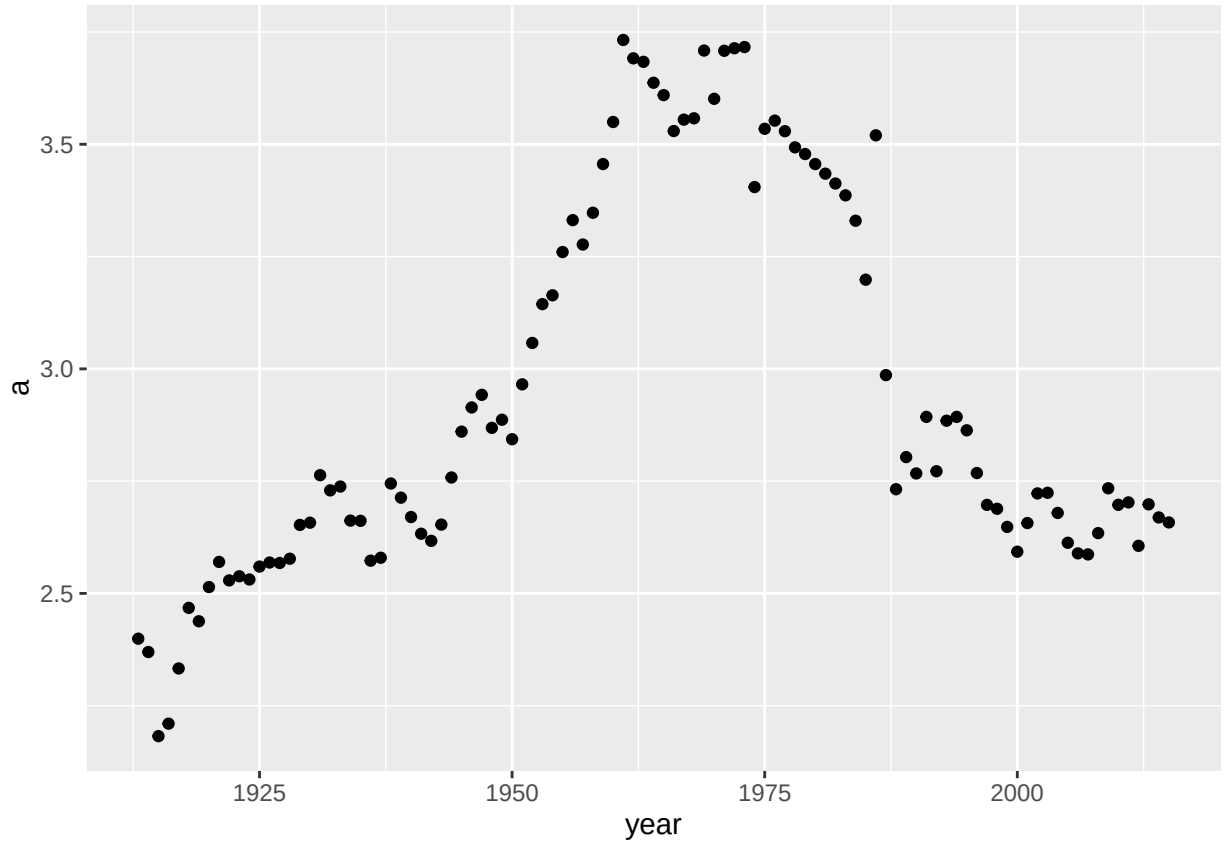
```
exponent.est_ratio <- function(P99, P99.9){
  a <- 1 - (log(10)) / (log(P99 / P99.9))
  return(a)
}
# Check the return
exponent.est_ratio(P99 = 1e6, P99.9 = 1e7)
```

```
## [1] 2
```

Part 2

4. Estimate a for each year in the data set, using your `exponent.est_ratio()` function. If the function was written properly, you should not need to use a loop. Plot your estimate of a over time using `ggplot`. Do the results look reasonable? (Remember that smaller exponents mean more income inequality.)

```
a <- exponent.est_ratio(P99 = wtiddata2$P99, P99.9 = wtiddata2$P99.9)
ggplot(data = data.frame(year = wtiddata2$year, a = a)) +
  geom_point(mapping = aes(x = year, y = a))
```



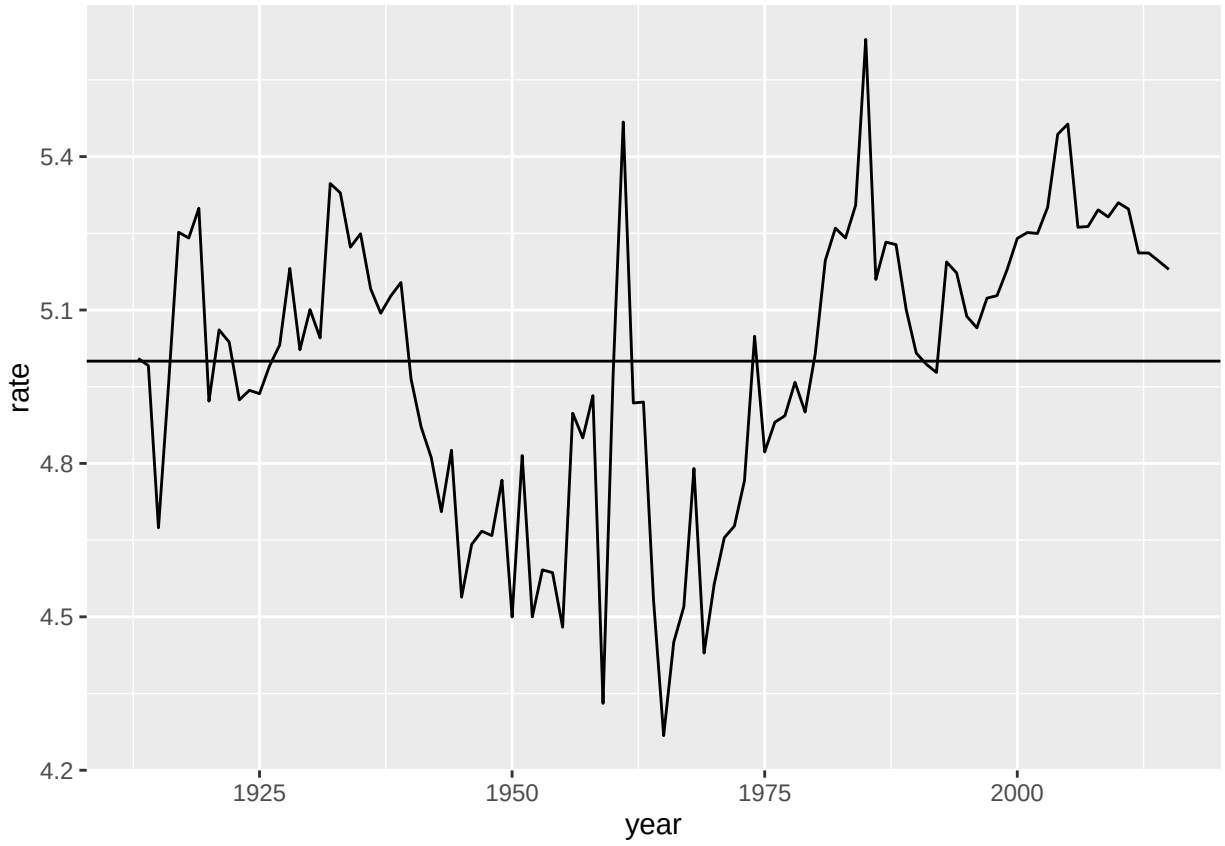
The result is reasonable. We can refer to some history reasearch materials and find that US income inequality has grown siginificantly since the early 1970s after several decades of stability. (from Wikipedia)

5. The logic leading to (1) also implies that

$$\left(\frac{P99.5}{P99.9}\right)^{-a+1} = 5$$

Write a function which takes `P99.5`, `P99.9`, and a, and calculates the left-hand side of that equation. Plot the values for each year using `ggplot`, using the data and your estimates of the exponent. Add a horizontal line with vertical coordinate 5. How good is the fit?

```
ratio.95 <- function(P99.5, P99.9, a){
   ratio <- (P99.5 / P99.9)^(-a + 1)
   return(ratio)
}
leftside <- ratio.95(P99.5 = wtiddata2$P99.5, P99.9 = wtiddata2$P99.9, a = a)
ggplot(data.frame(year = wtiddata2$year, rate = leftside)) +
  geom_line(mapping = aes(x = year, y = rate)) +
  geom_hline(yintercept = 5)
```



As we can see, the data has some trend, the horizontal line is not so good. We can calculate the MSE

```
MSE1 <- mean((leftside - 5)^2); MSE1
```

```
## [1] 0.07693592
```

6. By parallel reasoning, we should have $(P99/P99.5)^{-a+1} = 2$. Repeat the previous step with this formula. How would you describe this fit compared to the previous ones?

(Note: the formula in (1) is not the best way to estimate a, but it is one of the simplest.)

```
ratio.05 <- function(P99, P99.5, a){
   ratio <- (P99 / P99.5)^(-a + 1)
   return(ratio)
}
leftside2 <- ratio.05(P99 = wtiddata2$P99, P99.5 = wtiddata2$P99.5, a = a)
ggplot(data.frame(year = wtiddata2$year, rate = leftside2)) +
  geom_line(mapping = aes(x = year, y = rate)) +
  geom_hline(yintercept = 2)
```

Calculate the MSE again

```
MSE2 <- mean((leftside2 - 2)^2); MSE2
```

```
## [1] 0.0134066
```

Compare two MSEs we can find the fitness of the later one is better.