Lab 6

Christine Chong cc4190

June 14, 2017

Instructions

Please knit this lab as a `.pdf` file. 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

We consider a dataset containing information about the world's richest people. The dataset is taken from 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?

```
wtid <-read.csv("wtid-report.csv", header = TRUE, sep = ",")
year1 <- 1972
year2 <- 1942
year3 <- 1922
P99<-wtid$P99.income.threshold[wtid$Year == year1]
P99.5<-wtid$P99.5.income.threshold[wtid$Year == year2]
P99.9<-wtid$P99.9.income.threshold[wtid$Year == year3]
P99
```

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

```
P99.5
```

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

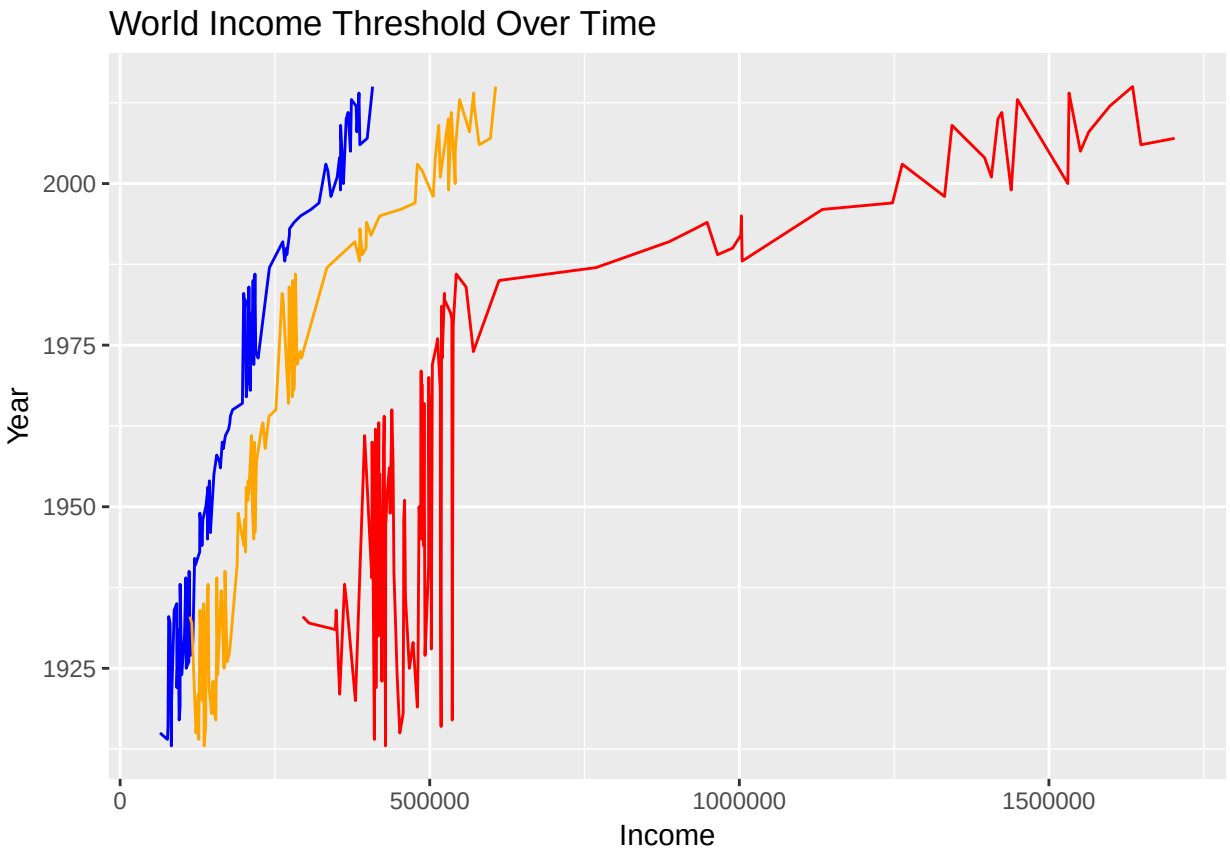
```
P99.9
```

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

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.)

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=wtid)+
  geom_line(mapping = aes(x=P99.income.threshold, y = Year), color = "blue")+
  geom_line(mapping = aes(x=P99.5.income.threshold, y = Year), color = "orange")+
  geom_line(mapping = aes(x=P99.9.income.threshold, y = Year), color = "red")+
  labs(title = "World Income Threshold Over Time", x="Income", y ="Year")
```



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){
  max = length(P99)
  a <- c()
  for(i in 1:max){
    a[i] <- 1- (log (10)/log(P99[i]/P99.9[i]))
  }
  return (a)
}
exponent.est_ratio(1e+06, 1e+07)
```

```
## [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.)

```
all99 <- wtid$P99.income.threshold
all99.9  <- wtid$P99.9.income.threshold
wtid$a<- exponent.est_ratio(all99, all99.9)
head(wtid$a, 5)
```

```
## [1] 2.399194 2.369454 2.182215 2.209948 2.332909
```

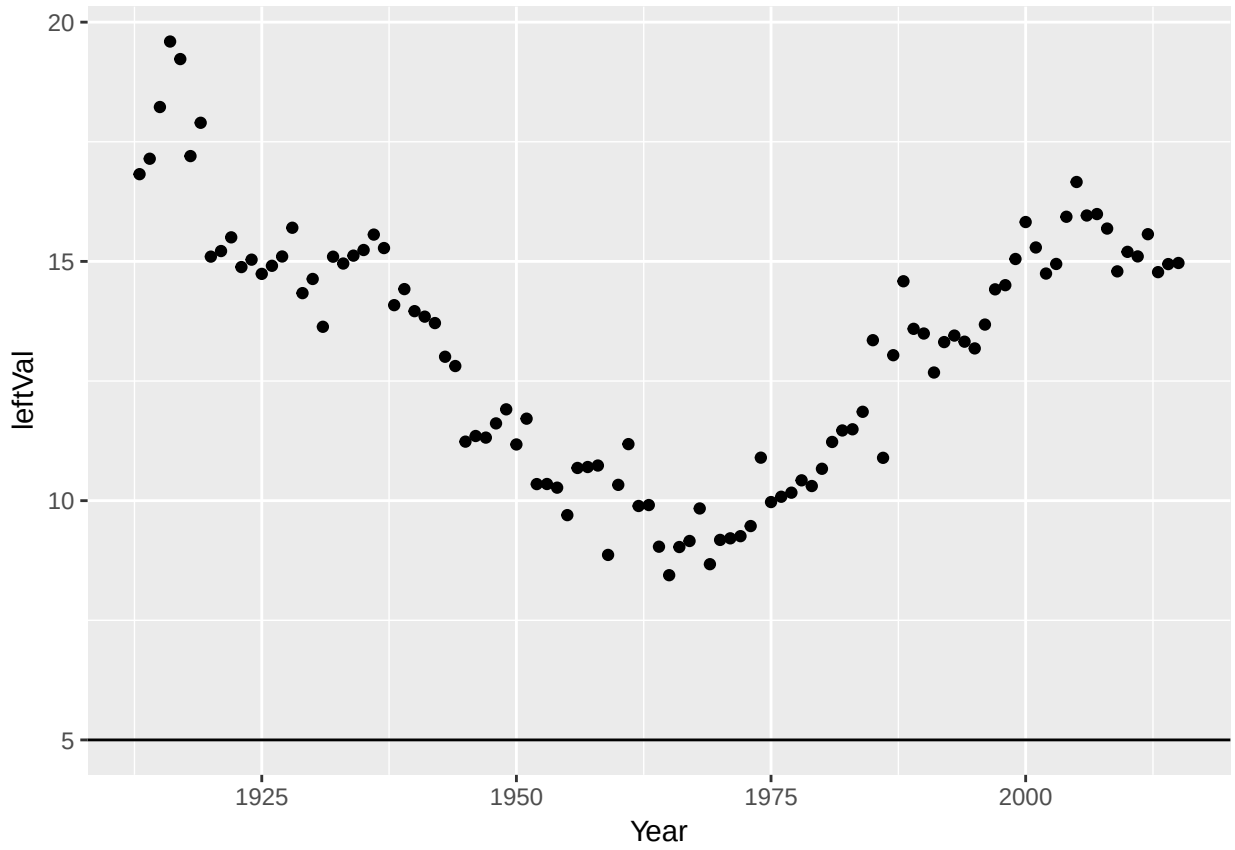
The results look reasonable. They are close to what we got earlier. And it seems like even though there are fluctuations, it does not differ from the range of 2-3.

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?

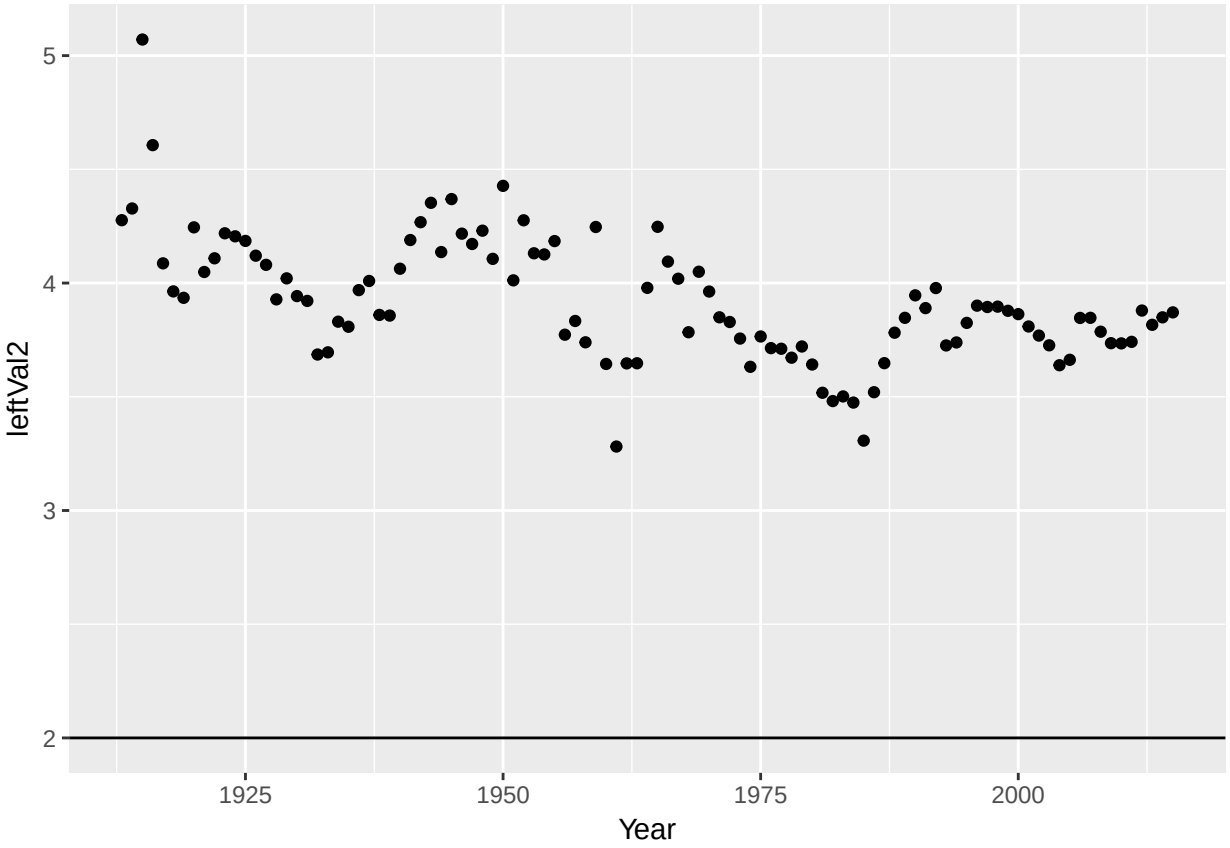
```
leftSide<- function(LP99.5, LP99.9, LallA ){
  Lmax = length(LP99.5)
  result <- c()
  for(k in 1:Lmax){
    result[k] <- (LP99.5[k]/LP99.9[k])^-LallA[k]+1
  }
  return (result)
}
all99.5 <- wtid$P99.5.income.threshold
allA <- wtid$a
wtid$leftVal<- leftSide(all99.5, all99.9, allA )

ggplot(data=wtid) +
  geom_point(mapping = aes(x= Year , y= leftVal))+
  geom_hline(yintercept = 5)
```



It seems to not fit at all. It comes close near 1950-1975. But that is about it

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?

```
wtid$leftVal2 <- leftSide(all99, all99.5, allA)
ggplot(data=wtid) +
  geom_point(mapping = aes(x= Year , y= leftVal2))+
  geom_hline(yintercept = 2)
```



This Value seems to have a better fit than the other value. (it is only 1.5 difference compared to 3+ difference) (Note: the formula in (1) is not the best way to estimate a, but it is one of the simplest.)