Lab 5 Solutions

Yi Chen, yc3356

October 20, 2017

Instructions

Before you leave lab today make sure that you upload a .pdf file to the canvas page (this should have a .pdf extension). This should be the PDF output after you have knitted the file, we don't need the .Rmd file (don't upload the one with the .Rmd extension). The file you upload to the Canvas page should be updated with commands you provide to answer each of the questions below. You can edit this file directly to produce your final solutions. Note, however, in the file you upload you should the above header to have the date, your name, and your UNI. Similarly, when you save the file you should replace **UNI** with your actualy UNI.

Background

In this lab we look at dataset containing information on the world's richest people from the World Top Incomes Database (WTID) hosted by the Paris School of Economics [http://wid.world]. 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} and 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 dataframe containing only the year, the "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. Rename the columns of your new dataframe as `Year`, `P99`, `P99.5`, `P99.9`. What was P99 in 1993? P99.5 in 1942? You must identify these using your code rather than looking up the values manually.

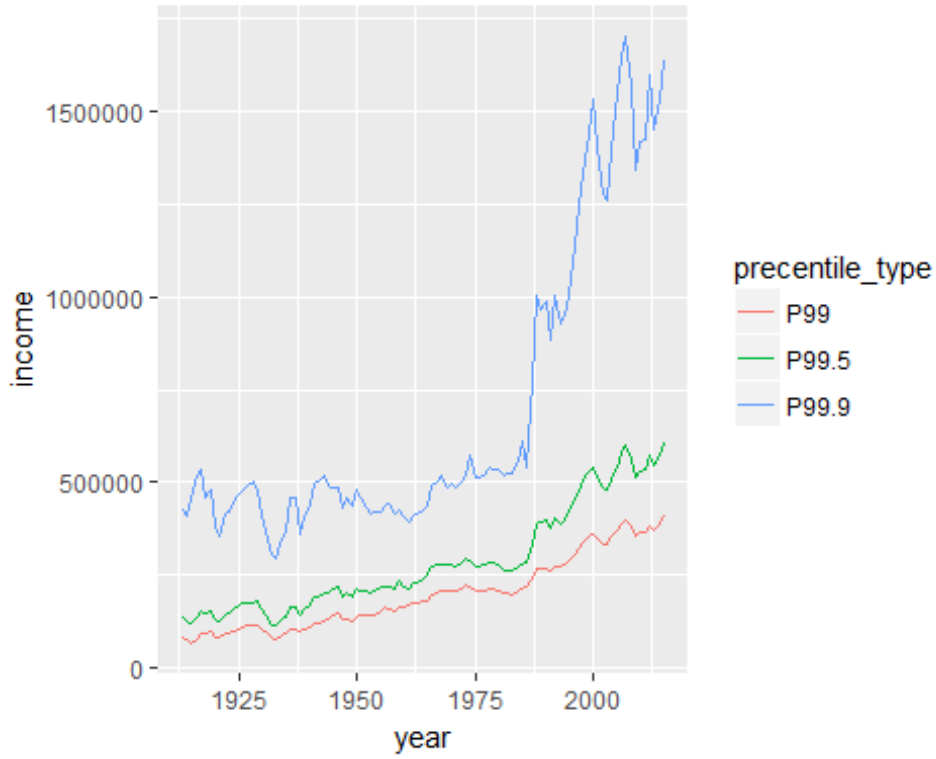
```
setwd("C:/Users/cheny/Desktop/study/statistical computing and intro to data
science/lab/lab five")

Data <- read.csv('wtid-report.csv',header = TRUE)
Data <- Data[,-1] # only take the col that needed
colnames(Data) <- c('year','P99','P99.5','P99.9') # rename the col name

#P99 in 1993
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

P99_1993 <- filter(Data,Data$year==1993) %>% select(P99)
cat('P99 in 1993 is:',as.numeric(P99_1993))

## P99 in 1993 is: 273534.9

#P99.5 in 1942
library(dplyr)
P99.5_1942 <- filter(Data,Data$year==1942) %>% select(P99.5)
cat('P99 in 1993 is:',as.numeric(P99.5_1942))

## P99 in 1993 is: 189140.6
```

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)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.2
```

```
library(reshape2)
plot_data <- melt(Data,id.vars = 'year')
colnames(plot_data) <- c('year','precentile_type','income')
ggplot(data=plot_data,aes(x=year,y=income)) +
        geom_line(aes(color=precentile_type))
```



3. It can be shown from the earlier equations that one can estimate the exponent by the formula

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

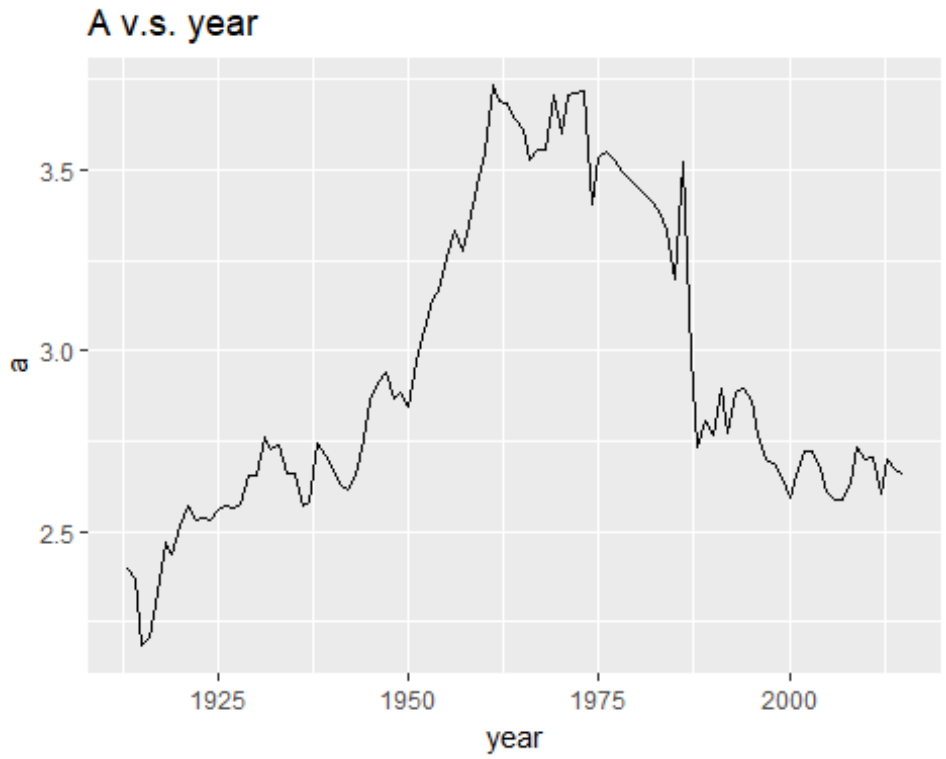
```
exponent.est_ratio <- function(p1=Data$P99,p2=Data$P99.9){
        a <- 1- (log(10)/(log(p1/p2)))
        return(a)
}


exponent.est_ratio(1e6,1e7)
```

```
## [1] 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. Think about whether these results look reasonable. (Remember that smaller exponents mean more income inequality.)

```
result <- exponent.est_ratio()
plot_data2 <- data.frame(result,Data$year)
ggplot(data=plot_data2,aes(x=Data.year,y=result)) +
        geom_line()+
        labs(title = "A v.s. year", x = "year", y = "a")
```



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