STAT GR5206 HW3_mjs2364

Mathieu Sauterey - UNI: MJS2364

26 octobre 2017

Goals : writing functions to automate repetitive tasks and using them as larger parts of code, some practice with ggplot, working with data frames and manipulating data from one form to another.

This homework uses the World Top Incomes Database and the Pareto distribution, as in this week's lab. The following notes are a repeat from the lab assignment: 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.

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.

Part1. Estimating a on US data

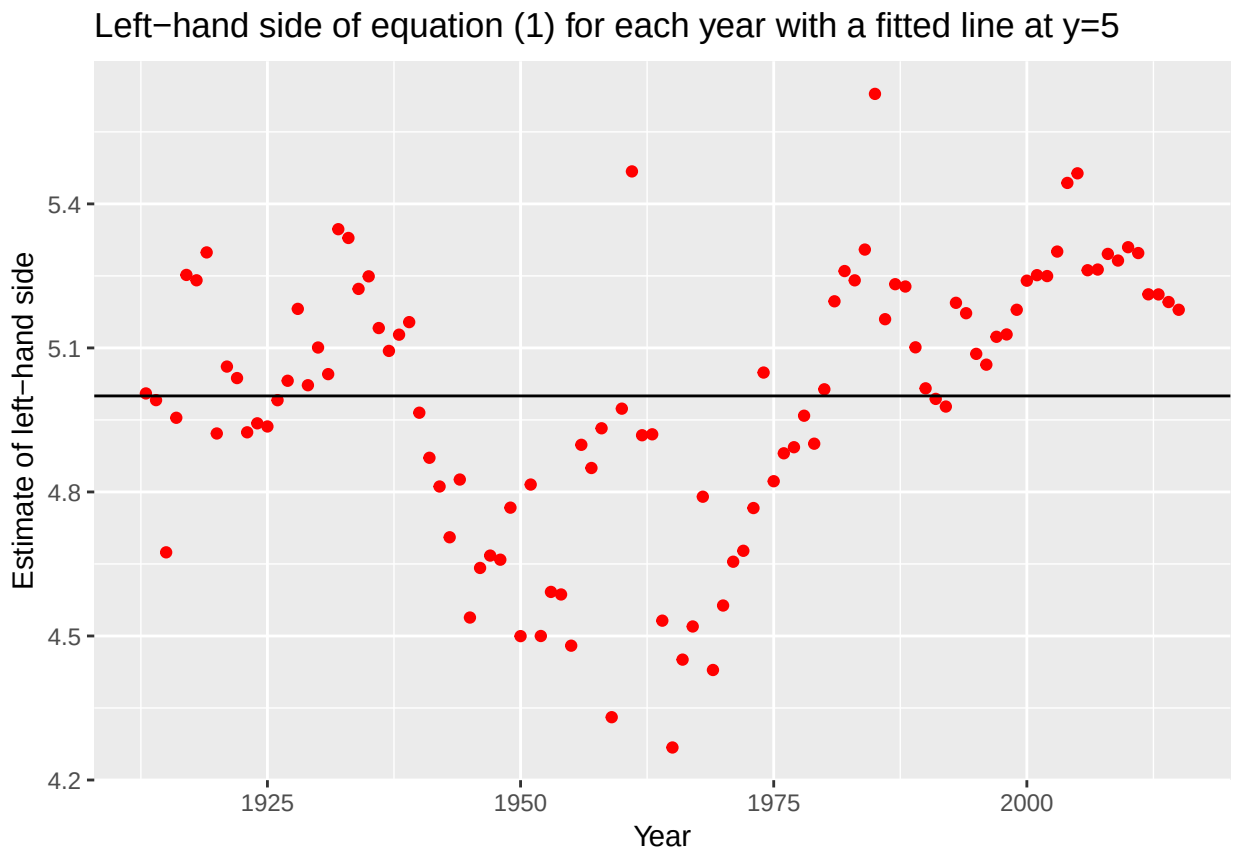
$$(\frac{P99.5}{P99.9})^{-a+1} = 5$$

Equation (1)

i. 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 from lab (using the exponent.est ratio()). Add a horizontal line with vertical coordinate 5. How good is the fit?

```
#Loads the ggplot package
library(ggplot2)

#First we load the US wealth data
wealth <- read.csv("wtid-report.csv", as.is = TRUE)

#Then we select only the columns of interest(Year, P99,P99.5,P99.9)
cols<-c("Year", "P99.income.threshold","P99.5.income.threshold","P99.9.income.threshold")
new_wealth <- wealth[colnames(wealth) %in% cols]

#We also shorten their name
colnames(new_wealth) <- c("Year", "P99","P99.5","P99.9")

#This function estimates exponent "a" using P99 and P99.9
exponent.est_ratio <- function(P99,P99.9){
  a <- rep(NA,length(P99))
  a <- 1-(log(10)/log(P99/P99.9))
  return(a)
}
```

```r
#We load each wealth column into a vector
P99 <- new_wealth$P99
P99.5 <- new_wealth$P99.5
P99.9 <- new_wealth$P99.9

# We create a vector containing estimates of the exponent for all US data
exponent <- exponent.est_ratio(P99,P99.9)

#This function calculates the left-hand side of equation (1). It should be equal to 5
left_hand <- function(P99.5, P99.9, a){
  est_5 <- rep(NA,length(P99))
  est_5 <- (P99.5/P99.9)^(-a+1)
  return(est_5)}

#The following returns a vector of the left-hand side of eq (1) for all US data
vec_est_5 <- left_hand(P99.5, P99.9, exponent)

#This graphs the left-hand side vector against time, and adds a horizontal line y=5
ggplot(data = new_wealth)+
  geom_point(mapping = aes(x = Year, y = vec_est_5), color="red")+
  geom_hline(yintercept = 5)+
  labs(title = "Left-hand side of equation (1) for each year with a fitted line at y=5",
       x = "Year", y = "Estimate of left-hand side")
```



The horizontal line y = 5 very poorly fits the data. The fitted line is linear while the data display a curvilinear

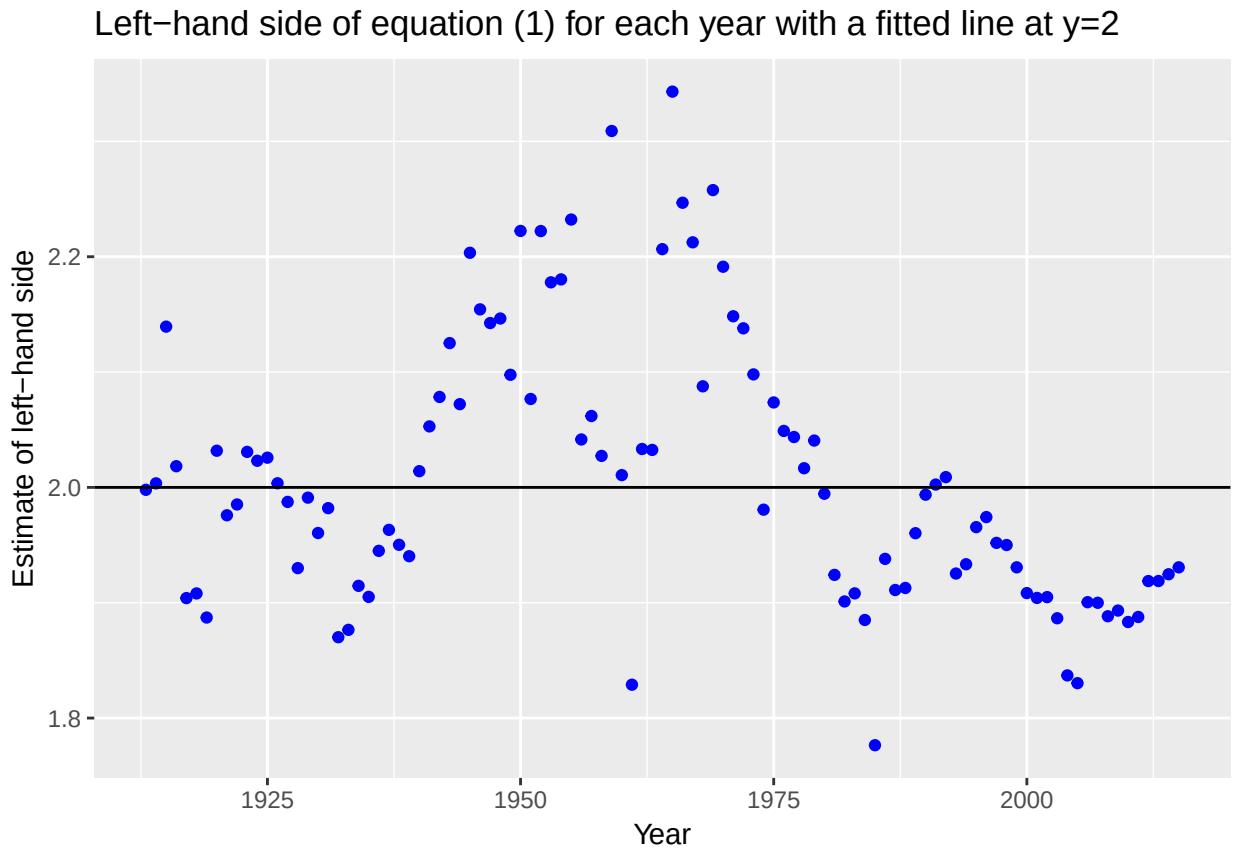
pattern, therefore this fit has a very large bias.

ii. Repeat the previous step with the formula below. How would you describe this fit compared to the previous one?

$$(\frac{P99.5}{P99.9})^{-a+1} = 2$$

Equation (2)

```
#This function calculates the left-hand side of equation (2). It should be equal to 2
left_hand2 <- function(P99, P99.5, a){
  est_2 <- rep(NA,length(P99))
  est_2 <- (P99/P99.5)^(-a+1)
  return(est_2)}

#The following returns a vector of the left-hand side of eq (2) for each year in the US
vec_est_2 <- left_hand(P99, P99.5, exponent)

#This graphs the left-hand side vector against time, and adds a horizontal line y=2
ggplot(data = new_wealth)+
  geom_point(mapping = aes(x = Year, y = vec_est_2),color="blue")+
  geom_hline(yintercept = 2)+
  labs(title = "Left-hand side of equation (1) for each year with a fitted line at y=2",
       x = "Year", y = "Estimate of left-hand side")
```



Similar to the previous step, the horizontal line y = 2 very poorly fits the data. The fitted line is linear while the data display a curvilinear pattern, therefore this fit has a very large bias too.

iii. Write a function, percentile_ratio_discrepancies, which takes as inputs P99, P99.5, P99.9 and a, and returns the value of the expression below Check that when P99=1e6, P99.5=2e6, P99.9=1e7 and a = 2, your function returns 0.

$$(\frac{P99}{P99.9})^{-a+1} = 10$$

Equation (2)

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

Equation (3)

$$(\frac{P99}{P99.5})^{-a+1} = 2$$

Equation (4)

$$((\frac{P99}{P99.9})^{-a+1} - 10)^2 + ((\frac{P99.5}{P99.9})^{-a+1} - 5)^2 + ((\frac{P99}{P99.5})^{-a+1} - 2)^2$$

Equation (5)

```
#The function takes the wealth percentiles as inputs and returns the cost equation (5)
 percentile_ratio_discrepancies <- function(P99,P99.5,P99.9,a){
   total <- ((P99/P99.9)^(-a+1)-10)^2+((P99.5/P99.9)^(-a+1)-5)^2+((P99/P99.5)^(-a+1)-2)^2
   return(total)}

#Test for the function. Should return 0.
percentile_ratio_discrepancies(1e6, 2e6, 1e7,2)
```

```
## [1] 0
```

The percentile_ratio_discrepancies function correctly returns 0 for the test case.

iv. Now we'd like to write a function, exponent.multi_ ratios_est, which takes as inputs the vectors P99, P99.5, P99.9, and estimates a. It should minimize the function percentile ratio discrepancies you wrote above. Check that when P99=1e6, P99.5=2e6 and P99.9=1e7, your function returns an a of 2.

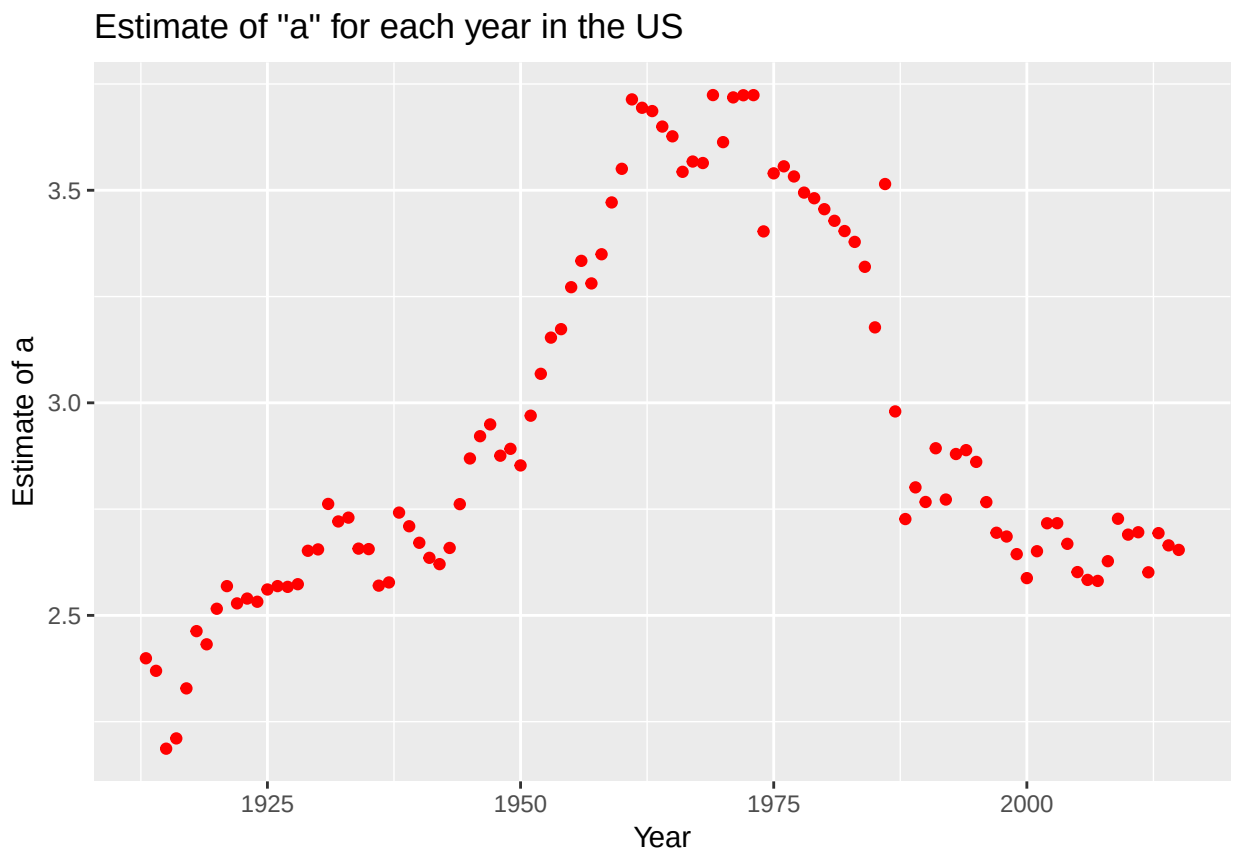
```
#The function takes the wealth percentiles as inputs and returns the best estimate of "a"
exponent.multi_ratios_est <- function(P99A,P99.5A,P99.9A){
  a <- exponent.est_ratio(P99A,P99.9A)
  nlm(percentile_ratio_discrepancies, a, P99=P99A, P99.5=P99.5A, P99.9=P99.9A)$estimate}

#Test for the function. Should return a=2.
exponent.multi_ratios_est(1e6, 2e6,1e7)
```

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

The function exponent.multi_ ratios_est correctly returns 2 for the test case.

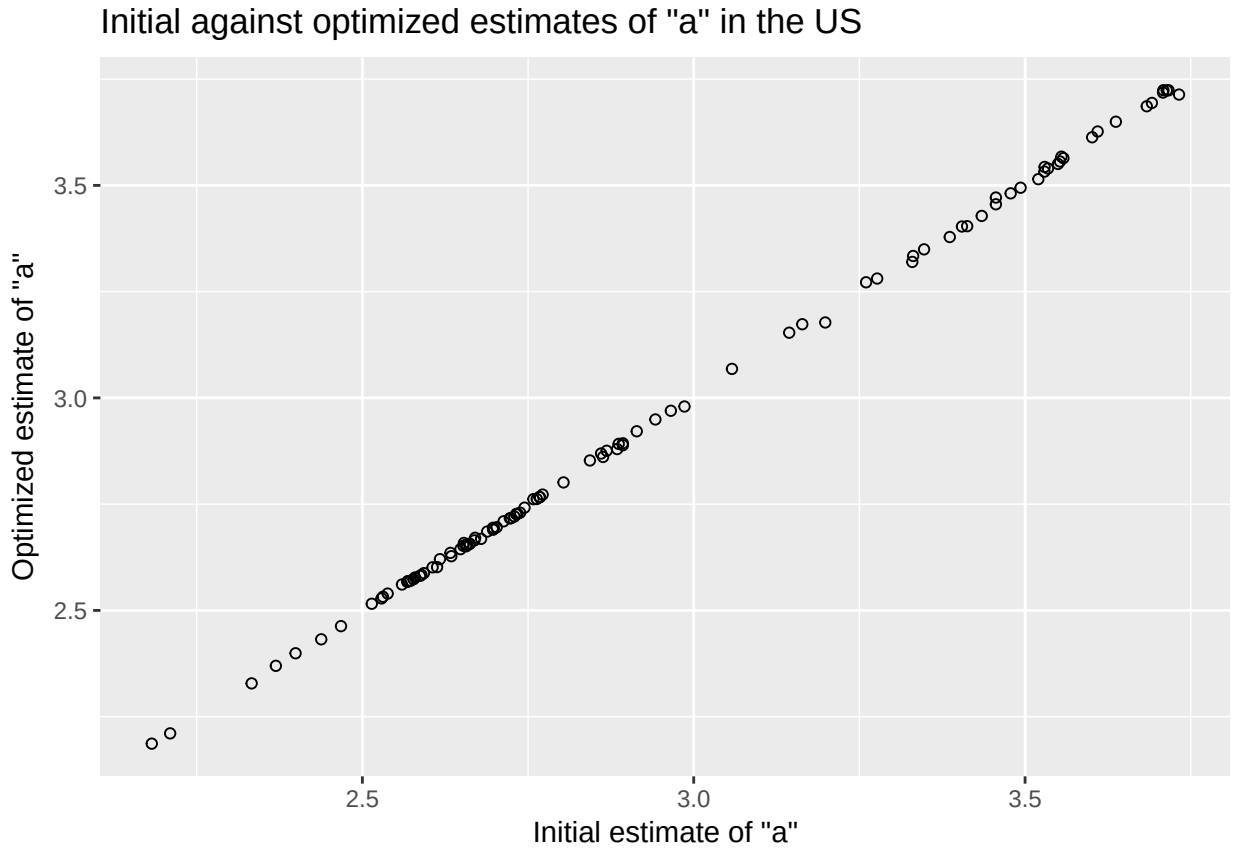
v. Write a function which uses exponent.multi_ratios_est to estimate a for the US for every year from 1913 to 2015. (There are many ways you could do this, including loops.) Plot the estimates using ggplot; make sure the labels of the plot are appropriate.

```
#The function returns a vector of best estimates of "a" for each year in the US
a_Years <- function(P99, P99.5, P99.9){
    vec_a <- rep(NA,length(P99))

    for (i in 1:length(P99)){
      if (anyNA(c(P99[i], P99.5[i], P99.9[i]))){vec_a[i]=NA}
      else{vec_a[i] <- exponent.multi_ratios_est(P99[i], P99.5[i], P99.9[i])}}

    return(unlist(vec_a))}

#This returns a vector of best estimates of "a" for each year in the US
US_a <- a_Years(P99, P99.5, P99.9)

#This plots the vector of best estimates of "a" against each year in the US
ggplot(data = new_wealth)+
  geom_point(mapping = aes(x = Year, y = US_a), color="red")+
  labs(title = "Estimate of \"a\" for each year in the US", x = "Year",
       y = "Estimate of a")
```



vi. Use (1) to estimate a for the US for every year. Make a scatter-plot of these estimates against those from problem (v) using ggplot. If they are identical or completely independent, something is wrong with at least one part of your code. Otherwise, can you say anything about how the two estimates compare?

```
#Scatterplot of the initial estimates against best estimates of \"a\"

ggplot(data = new_wealth) +
    geom_point(mapping = aes(x = exponent, y = US_a), shape=1) +
    labs(title = "Initial against optimized estimates of \"a\" in the US",
         x = "Initial estimate of \"a\"", y ="Optimized estimate of \"a\"")
```



We can see that the line above is almost straight with slope=1 , which means that the optimized estimate of "a" slightly differs from the initial estimate. We expected to observe this result because the optimization should slightly adjust "a" to minimize the cost function given as equation (5).

Part 2: Data for Other Countries

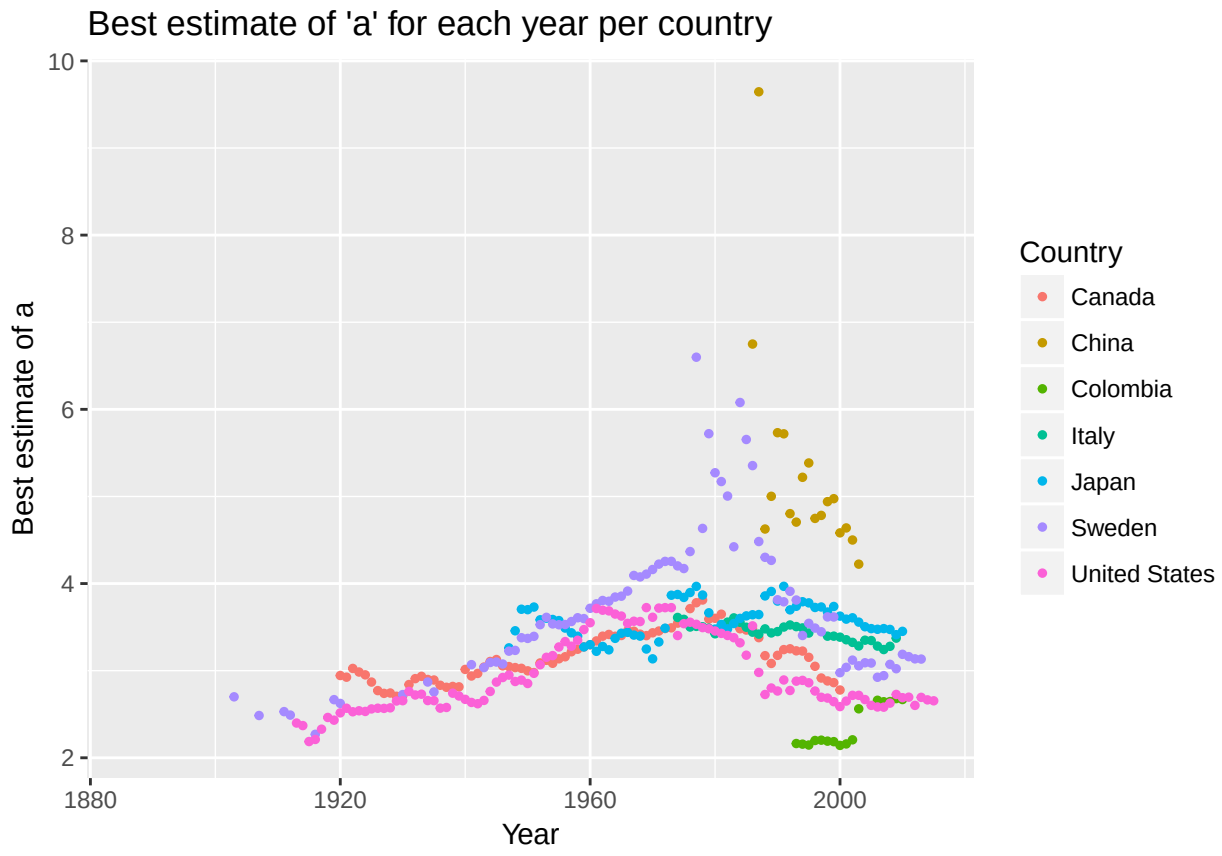
We're now going to look at this same data for some other countries: Canada, China, Colombia, India, Italy, Japan, and Sweden. This data is in the file wtid-homework.csv. The WTID website also has data on the average income per "tax unit" (roughly, household) for the US and the other countries. This info is in stored in the AverageIncome column.

vii. Use your function from problem (v) to estimate a over time for each of country.

```
#First we load the worldwide wealth data
nations <- read.csv("wtid-homework.csv", as.is = TRUE)

#We load each wealth column into a vector
P99W <- nations$P99
P99.5W <- nations$P99.5
P99.9W <- nations$P99.9

#This returns a vector of best estimates of "a" for each year in the world
World_a <- a_Years(P99W, P99.5W, P99.9W)
```

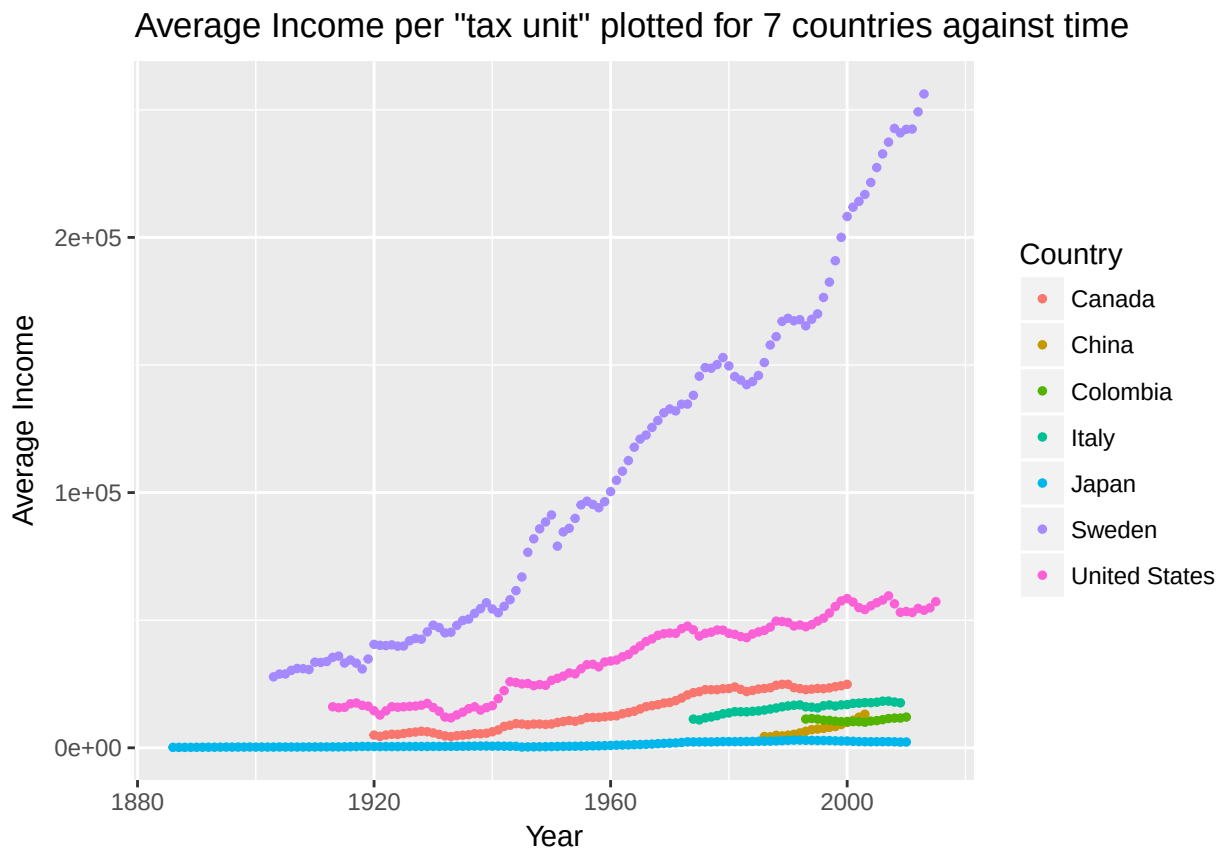
vii. Plot your estimates of a over time for all the countries using ggplot.

```
#This plots the vector of best estimates of "a" against each year in the world
ggplot(data = nations)+
  geom_point(mapping = aes(x = Year, y = World_a, color=Country), size=1, na.rm = TRUE)+
  labs(title = "Best estimate of 'a' for each year per country ",x = "Year",
       y = "Best estimate of a")
```



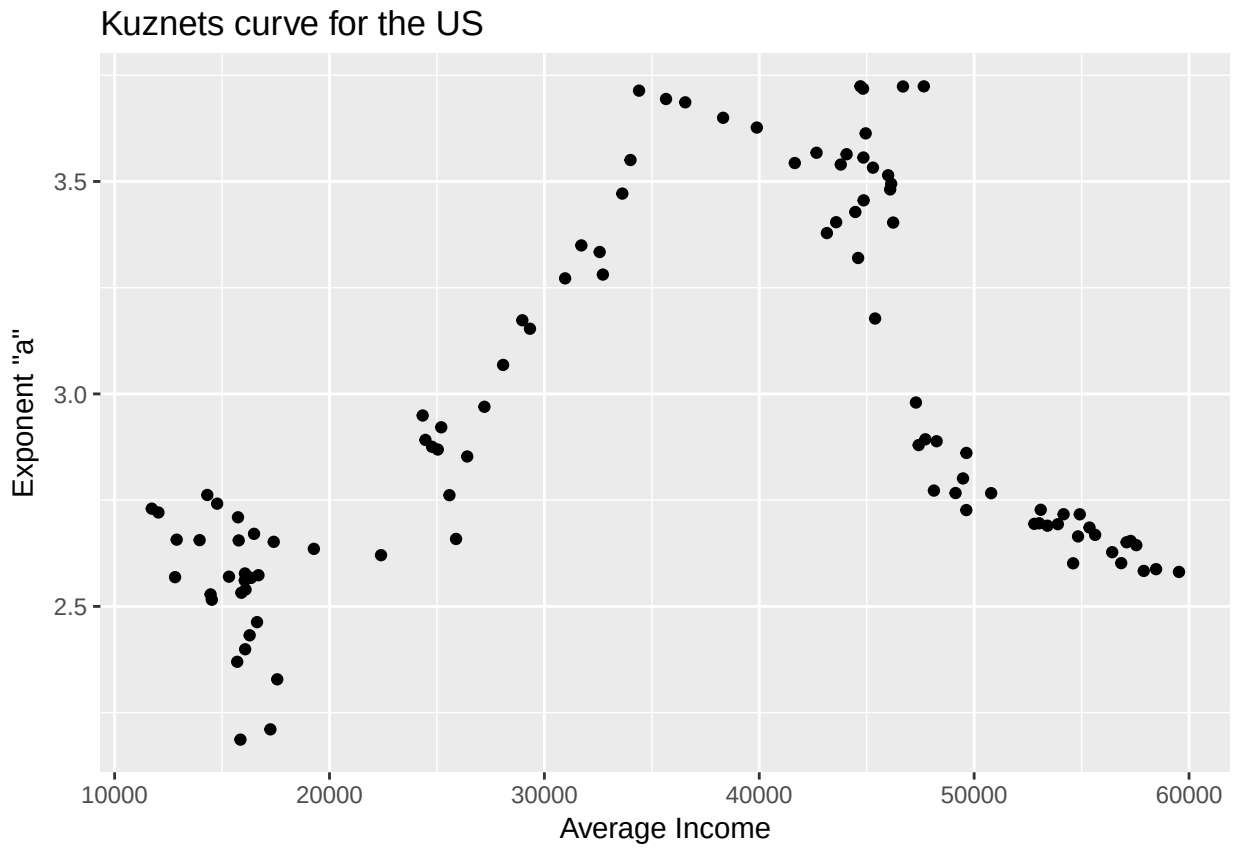
ix. Plot the series of average income per "tax unit" for the US and the countries against time in ggplot.

```
ggplot(data = nations)+
  geom_point(mapping = aes(x = Year, y = AverageIncome, color=Country), size=1,
       na.rm = TRUE)+
  labs(title = "Average Income per \"tax unit\" plotted for 7 countries against time",
       x = "Year", y = "Average Income")
```



x. Make a scatter-plot of your estimated exponents for the US against the average income for the US in ggplot. Qualitatively, can you say anything about the Kuznets curve?

```
US_AveIncome <- nations[nations$Country=="United States","AverageIncome"]

ggplot(data = new_wealth) + geom_point(mapping = aes(x = US_AveIncome, y = US_a))+
    labs(title = "Kuznets curve for the US", x = "Average Income", y = "Exponent \"a\"")
```



We know that a lower exponent "a" indicates more income inequality. According to the economist Kuznets, income inequality rises during the early stages of economic growth but then declines as growth continues. On the plot above we observe that exponent "a" first increases, reaches a turning point and then decreases with increasing income. In other words, income inequality first decreases, reaches a turning point and then increases as average income grows. This result is the exact opposite of Kuznets' prediction.

xi. For a more quantitative check on the Kuznets hypothesis, use lm() to regress your estimated exponents on the average income, including a quadratic term for income. Are the coefficients you get consistent with the hypothesis? Hint: the following will regress y on both x and x2:

```
lm(US_a ~ US_AveIncome + I(US_AveIncome^2))$coefficients
```

```
##       (Intercept)      US_AveIncome I(US_AveIncome^2)
##      8.230049e-01      1.394435e-04     -1.890556e-09
```

The coefficient of the x^2 term in our quadratic equation is negative. This means that the function has an inverted "U-curve" appearance. We just demonstrated in the previous question x. that this shape was not consistent with Kuznets' hyptohesis because it entails the inequality first decreases and then increases over time. Therefore the coefficients are also not consistent with the hypothesis.

xii. Do a separate quadratic regression for each country. Which ones have estimates compatible with the hypothesis?

```
lm_country <- function(Country){

  P99C <- nations$P99[nations$Country==Country]
  P99.5C <- nations$P99.5[nations$Country==Country]
  P99.9C <- nations$P99.9[nations$Country==Country]

  Country_AveIncome <- nations[nations$Country==Country,"AverageIncome"]
  Country_a <- a_Years(P99C, P99.5C, P99.9C)

  lm(Country_a ~ Country_AveIncome + I(Country_AveIncome^2))$coefficients
}

lm_country("Japan")
```

```
##            (Intercept)      Country_AveIncome I(Country_AveIncome^2)
##           3.729107e+00          -5.136191e-04           1.889447e-07
```

```
lm_country("Italy")
```

```
##            (Intercept)      Country_AveIncome I(Country_AveIncome^2)
##           2.582416e+00           1.594300e-04          -6.591048e-09
```

```
lm_country("Sweden")
```

```
##            (Intercept)      Country_AveIncome I(Country_AveIncome^2)
##           1.012353e+00           4.414454e-05          -1.496762e-10
```

```
lm_country("China")
```

```
##            (Intercept)      Country_AveIncome I(Country_AveIncome^2)
##           1.039781e+01          -1.126763e-03           5.257536e-08
```

```
lm_country("Colombia")
```

```
##            (Intercept)      Country_AveIncome I(Country_AveIncome^2)
##           3.461240e+01          -6.095234e-03           2.867133e-07
```

```
lm_country("Canada")
```

```
##            (Intercept)      Country_AveIncome I(Country_AveIncome^2)
##           2.266054e+00           1.240966e-04          -3.360837e-09
```

From the responses to our previous questions we can infer that a quadratic regression will be compatible with the hypothesis if the fitted line has a positive coefficient in front of the x^2 term. This entails that the fitted line has a "U-curve" appearance where inequality first increase (because inequality is invertly correlated with "a") and then decreases. The only countries which have estimates compatible with the hypothesis are Japan, China and Colombia.