homework_three

Yi (Chris) Chen

October 23, 2017

Homework Three

UNI:yc3356

Name: Yi Chen

Email:yc3356@columbia.edu (mailto:yc3356@columbia.edu)

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.

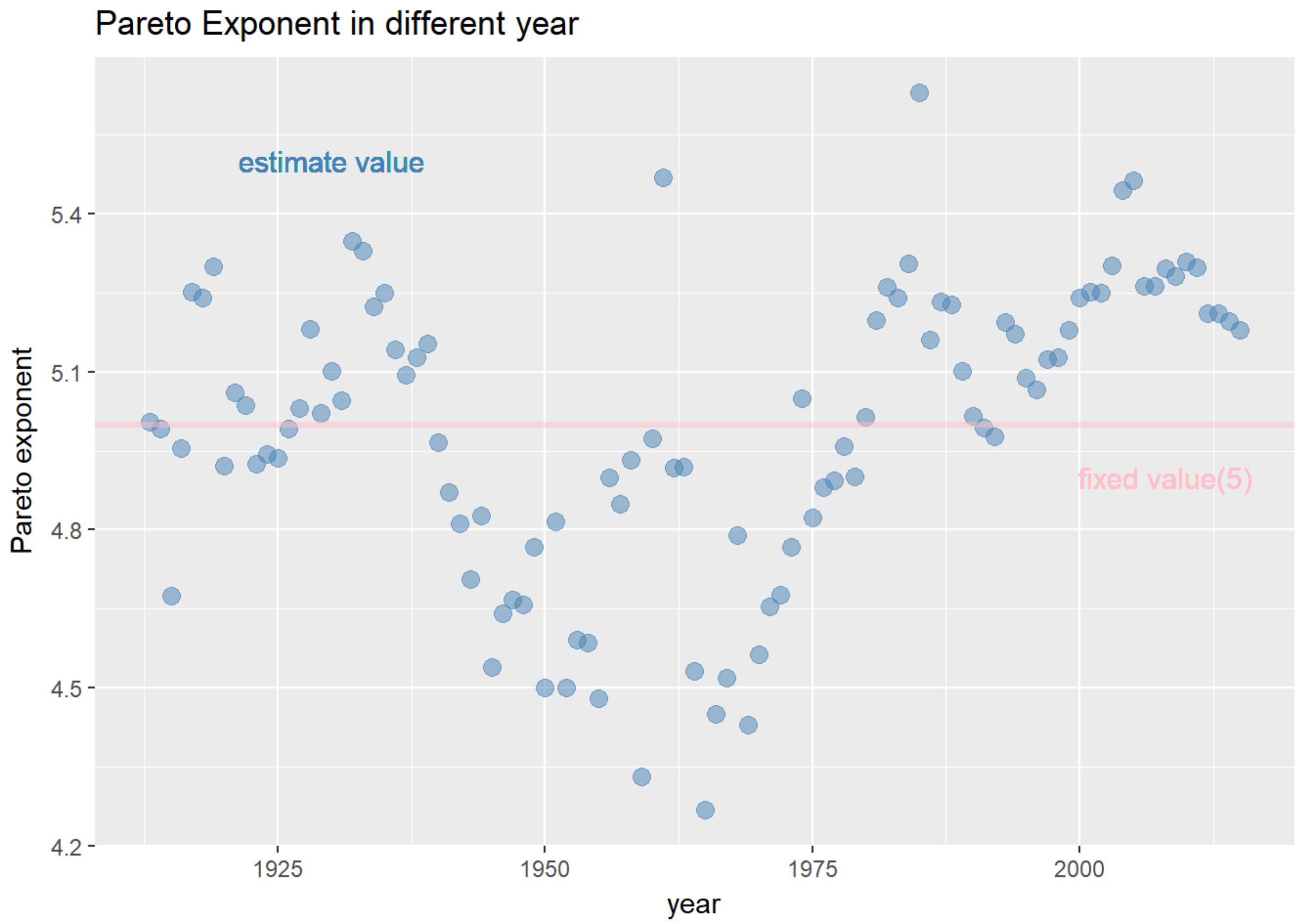
Part 1: Estimating a on US data

i.Write a function which takes P99.5, P99.9, and a, and calculates the lefthand 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?

```
#read the data
setwd("C:/Users/cheny/Desktop/study/statistical computing and intro to data science/homework/
homework three")
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



#function
problem_one <- function(P99.5,P99.9,P99){
        a <- 1- (log(10)/(log(P99/P99.9)))
        return((P99.5/P99.9)^(-a+1))
}

b <- problem_one(P99.5=Data$P99.5,P99.9=Data$P99.9,P99=Data$P99)


library(ggplot2)
```

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

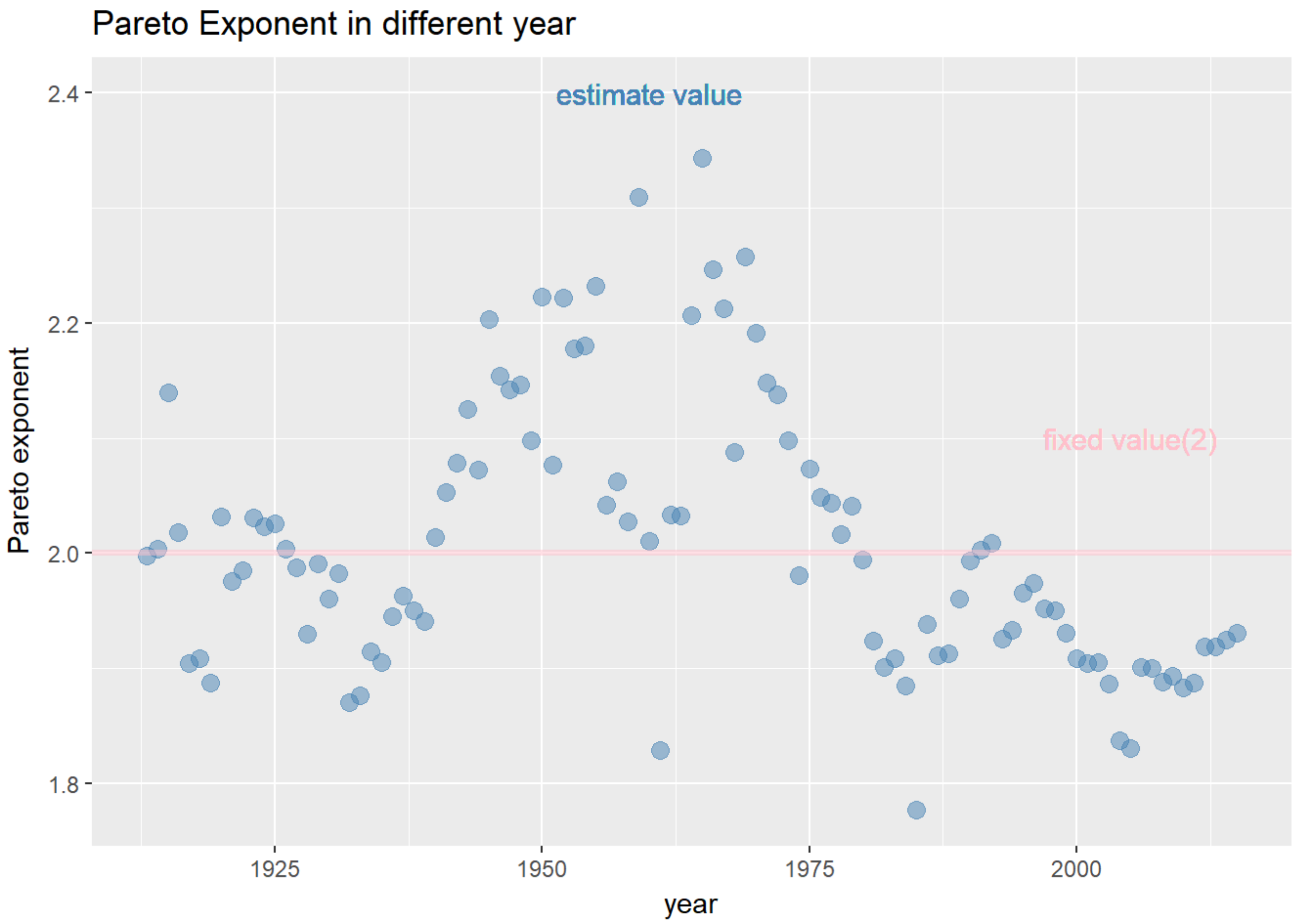
```
ggplot(data=Data) +
        labs(title='Pareto Exponent in different year',y='Pareto exponent',x='year') +
        geom_point(mapping = aes(x=year,y=b),size=3,col='steelblue',alpha=0.5)+
        geom_hline(yintercept = 5 , size=1.2, col='pink',alpha=0.5)+
        geom_text(mapping = aes(x=1930, y=5.5, label = 'estimate value'), size=4,col='steelbl
ue') +
        geom_text(mapping = aes(x=2008, y=4.9, label = 'fixed value(5)'), size=4,col='pink')
```



In general, estimate value is roughly round the fixed value 5.But it do has some outliers near 1960s.

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

```
problem_two <- function(P99.5,P99.9,P99){
        a <- 1- (log(10)/(log(P99/P99.9)))
        return((P99/P99.5)^(-a+1))
}

b <- problem_two(P99 = Data$P99 , P99.5 = Data$P99.5 ,P99.9 = Data$P99.9)


library(ggplot2)
ggplot(data=Data) +
        labs(title='Pareto Exponent in different year',y='Pareto exponent',x='year') +
        geom_point(mapping = aes(x=year,y=b),size=3,col='steelblue',alpha=0.5)+
        geom_hline(yintercept = 2 , size=1.2, col='pink',alpha=0.5)+
        geom_text(mapping = aes(x=1960, y=2.4, label = 'estimate value'), size=4,col='steelbl
ue') +
        geom_text(mapping = aes(x=2005, y=2.1, label = 'fixed value(2)'), size=4,col='pink')
```



In general, estimate value is roughly round the fixed value 5. But it do has some outliers near 1960s.

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 above. Check that when P99=1e6, P99.5=2e6, P99.9=1e7 and a = 2, your function returns 0.

- analysis : this is a **optimization problem** , the method we use is **Gradient Descent**

```
# as we can see from the plot before
percentile_ratio_discrepancies <- function(P99,P99.5,P99.9,a){
                return(((P99/P99.9)^(-a+1)-10)^2 +
                       ((P99.5/P99.9)^(-a+1)-5)^2 +
                       ((P99/P99.5)^(-a+1)-2)^2)
}

test <- percentile_ratio_discrepancies(P99 = 1e6,P99.5 = 2e6,P99.9 = 1e7,a=2)
test
```

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

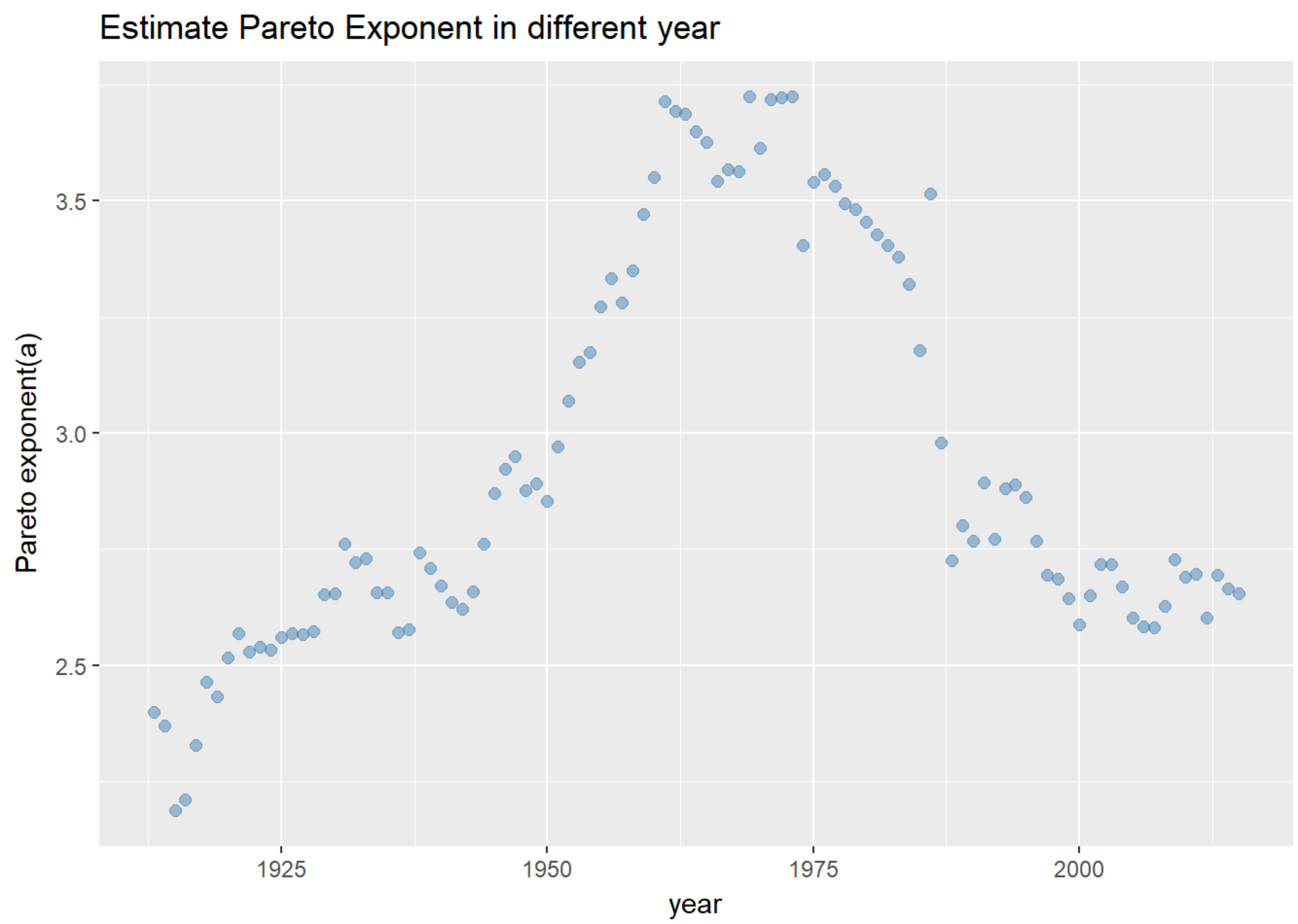
iv. Now we'd like to write a function, which takes as inputs the vectors P99, P99.5, P99.9, and estimates a. It should minimize the function percentile ratio discrepancies you wrote above.

```r
exponent.multi_ratios_est <- function(P99,P99.5,P99.9) {

        a <- 1- (log(10)/(log(P99/P99.9)))

        percentile_ratio_discrepancies <- function(x){
                return(((P99/P99.9)^(-x+1)-10)^2 +
                      ((P99.5/P99.9)^(-x+1)-5)^2 +
                      ((P99/P99.5)^(-x+1)-2)^2)
        }

        return(nlm(f = percentile_ratio_discrepancies,a))
}

result <- exponent.multi_ratios_est(P99=1e6,P99.5=2e6,P99.9=1e7)
cat('the estimate value of a is:',result$estimate)
```

```
## the estimate value of a is: 2
```

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.

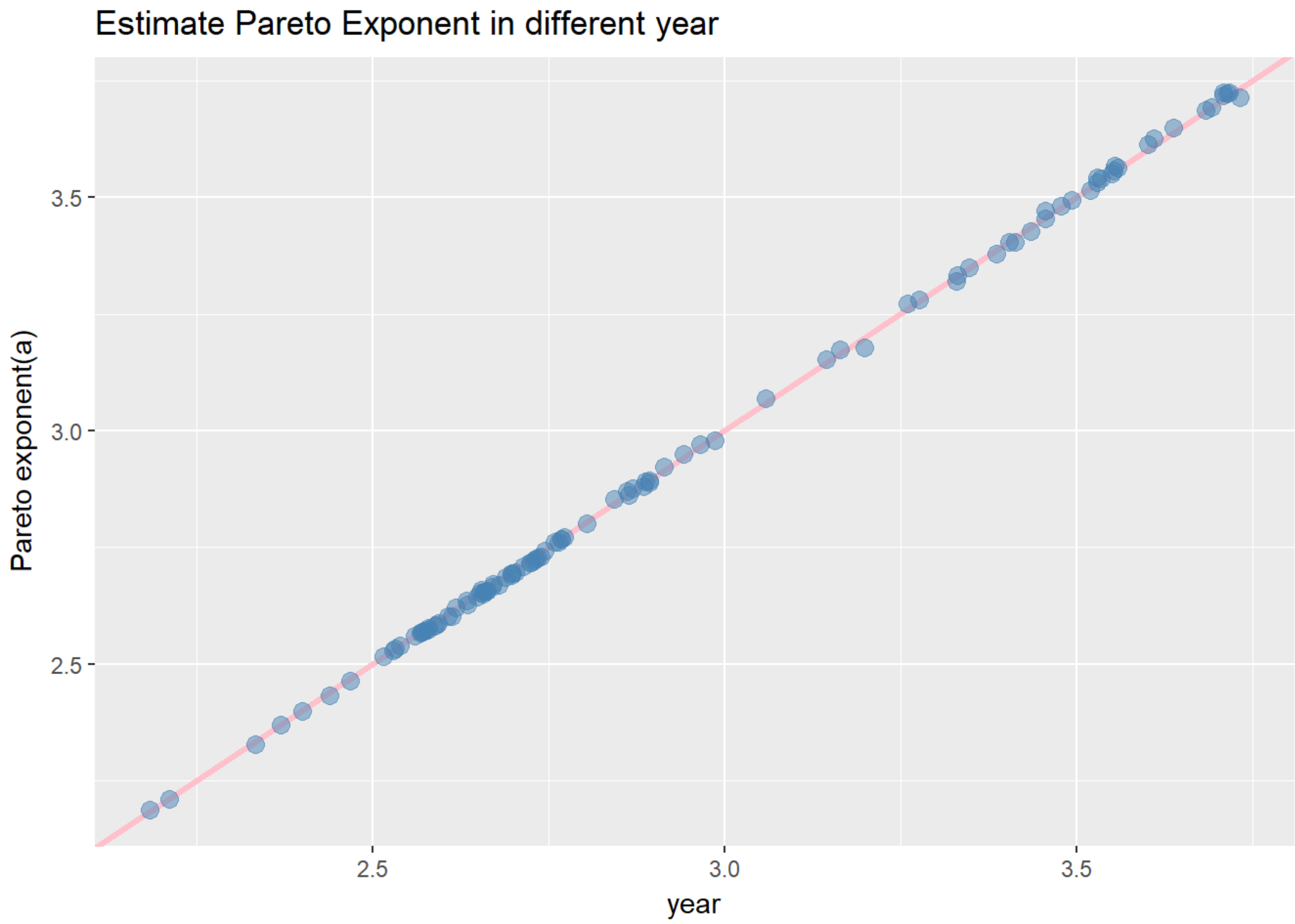
```r
problem_four <- function(data){

        estimates_of_a <- vector()

        for(i in 1:nrow(data)){
                estimate_P99 <- data$P99[i]
                estimate_p99.5 <- data$P99.5[i]
                estimate_P99.9 <- data$P99.9[i]
                result <- exponent.multi_ratios_est(P99=estimate_P99,P99.5=estimate_p99.5,P9
9.9=estimate_P99.9)$estimate
                estimates_of_a <- c(estimates_of_a,result)
        }

        return(estimates_of_a)

}

Data$Estimate_a <- problem_four(data = Data)

ggplot(data=Data,aes(year,Estimate_a)) +
        geom_point(size=2,col='steelblue',alpha=0.5) +
        labs(title='Estimate Pareto Exponent in different year',y='Pareto exponent(a)',x='yea
r')
```



vi. Use (1) to estimate a for the US for every year. Make a scatter-plot ofthese estimates against those from problem (v) using ggplot. If theyare 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?

```
problem_one_a <- 1- (log(10)/(log(Data$P99/Data$P99.9)))
problem_four_a <- problem_four(data = Data)

ggplot(data=Data) +
        geom_abline(x=2:4,y=2:4,color='pink',size=1.2) +
        geom_point(mapping = aes(x=problem_one_a,y=problem_four_a),size=3,col='steelblue',alp
ha=0.5) +
        labs(title='Estimate Pareto Exponent in different year',y='Pareto exponent(a)',x='yea
r')
```

```
## Warning: Ignoring unknown parameters: x, y
```




```
cor(problem_one_a,problem_four_a)
```

```
## [1] 0.9998843
```

In general, the two estimate fix with each other very well the correlation between them is 0.9. But they are not identical.

Part 2:Data for Other Countries

vii. Use your function from problem (v) to estimate a over time for each of them. Note that the size of the dataset is different for each of thesecountries, and there may be some NA values.

```
Data2 <- read.csv('wtid-homework.csv',header = TRUE)
library(dplyr)
```

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

```
##
## Attaching package: 'dplyr'
```

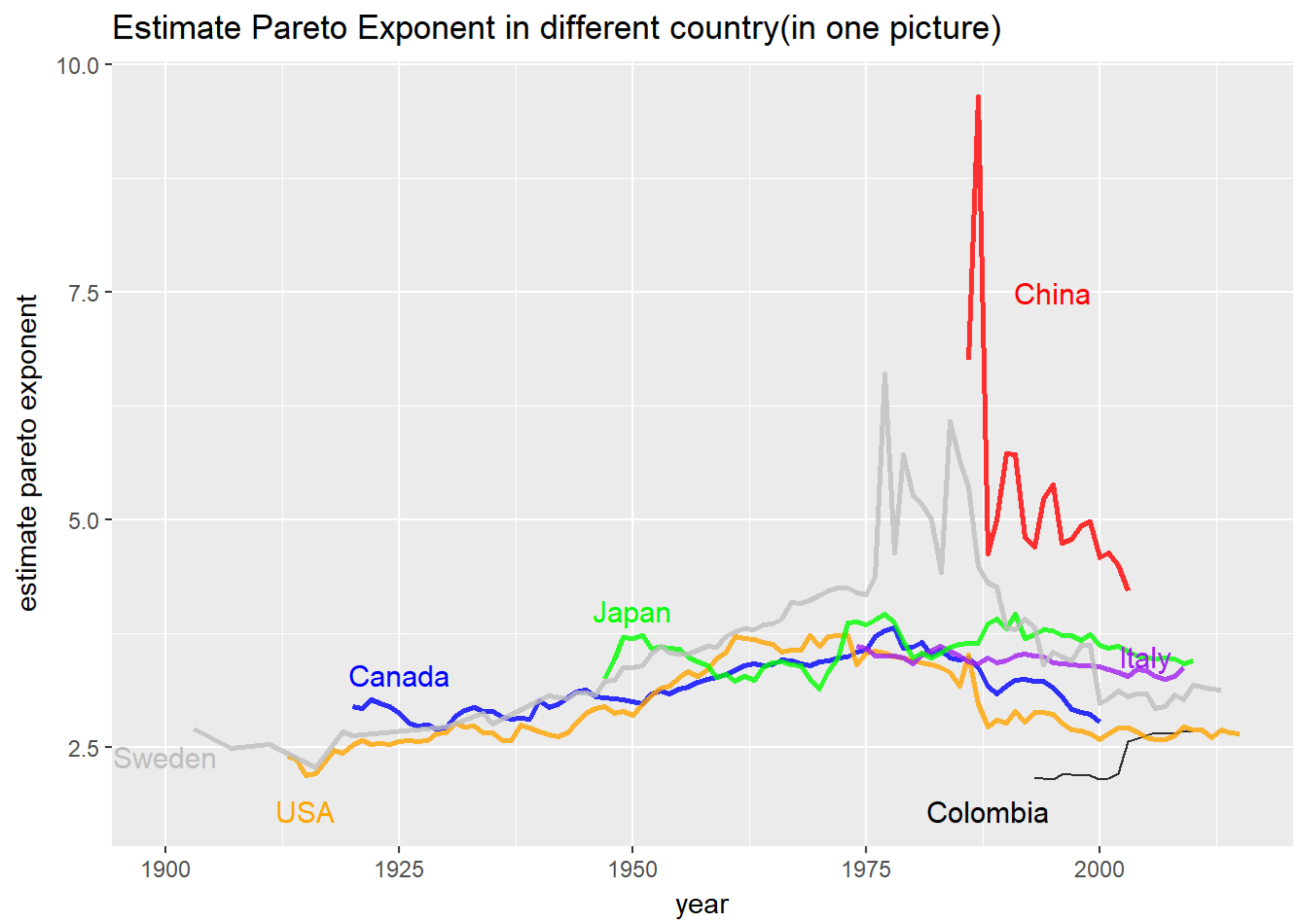
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

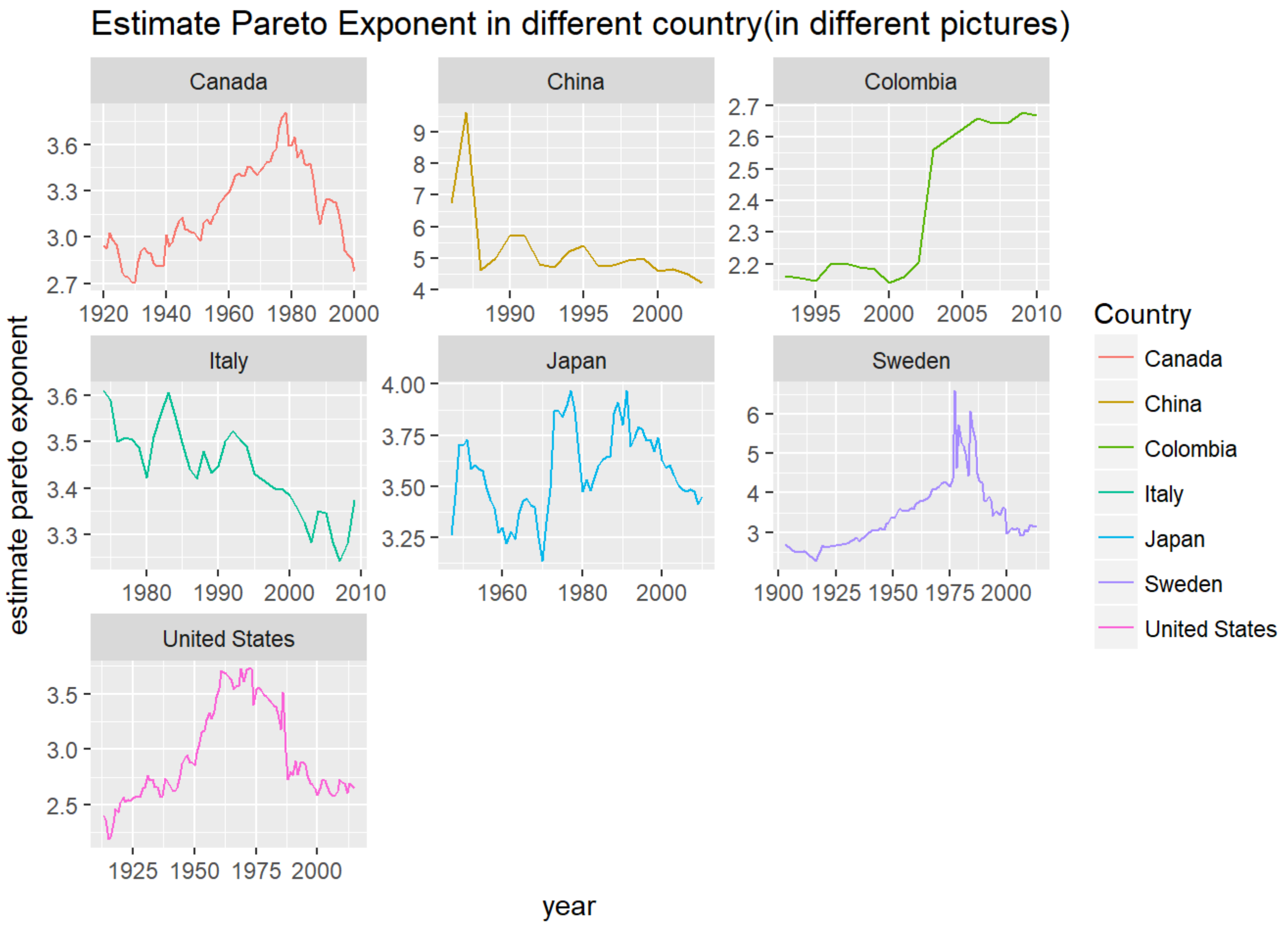
```
Canada <- filter(Data2,Country == 'Canada' & P99>0)
China <- filter(Data2,Country == 'China' & P99>0)
Colombia <- filter(Data2,Country == 'Colombia'& P99>0)
USA <- filter(Data2,Country == 'United States'& P99>0)
Italy <- filter(Data2,Country == 'Italy'& P99>0)
Japan <- filter(Data2,Country == 'Japan'& P99>0)
Sweden <- filter(Data2,Country == 'Sweden'& P99>0)

Canada$Estimate_a <- problem_four(data = Canada)
China$Estimate_a <- problem_four(data = China)
Colombia$Estimate_a <- problem_four(data = Colombia)
USA$Estimate_a <- problem_four(data = USA)
Italy$Estimate_a <- problem_four(data=Italy)
Japan$Estimate_a <- problem_four(data=Japan)
Sweden$Estimate_a <- problem_four(data=Sweden)
```

viii.Plot your estimates of a over time for all the countries using ggplot. Note that the years covered by the data are di erent for each country. You may either make multiple plots, or put all the series into one plot. Either way, make sure that the plots are clearly labeled.

```
# plot in one picture

library(ggplot2)
ggplot() +
        geom_line(aes(x=Canada$Year,y=Canada$Estimate_a),size=1,col='blue',alpha=0.8) +
        geom_line(mapping = aes(x=China$Year,y=China$Estimate_a),size=1,col='red',alpha=0.8)
+
        geom_line(mapping = aes(x=Colombia$Year,y=Colombia$Estimate_a),col='black',alpha=0.8)
 +
        geom_line(mapping = aes(x=USA$Year,y=USA$Estimate_a),size=1,col='orange',alpha=0.8) +
        geom_line(mapping = aes(x=Italy$Year,y=Italy$Estimate_a),size=1,col='purple',alpha=0.
8) +
        geom_line(mapping = aes(x=Japan$Year,y=Japan$Estimate_a),size=1,col='green',alpha=0.8
) +
        geom_line(mapping = aes(x=Sweden$Year,y=Sweden$Estimate_a),size=1,col='grey',alpha=0.
8) +
        labs(title='Estimate Pareto Exponent in different country(in one picture)',x='year',y
='estimate pareto exponent') +
        geom_text(mapping = aes(x=1900, y=2.4, label = 'Sweden'), size=4,col='grey') +
        geom_text(mapping = aes(x=1925, y=3.3, label = 'Canada'), size=4,col='blue') +
        geom_text(mapping = aes(x=1915, y=1.8, label = 'USA'), size=4,col='orange') +
        geom_text(mapping = aes(x=1950, y=4, label = 'Japan'), size=4,col='green') +
        geom_text(mapping = aes(x=1995, y=7.5, label = 'China'), size=4,col='red') +
        geom_text(mapping = aes(x=2005, y=3.5, label = 'Italy'), size=4,col='purple') +
        geom_text(mapping = aes(x=1988, y=1.8, label = 'Colombia'), size=4,col='black')
```

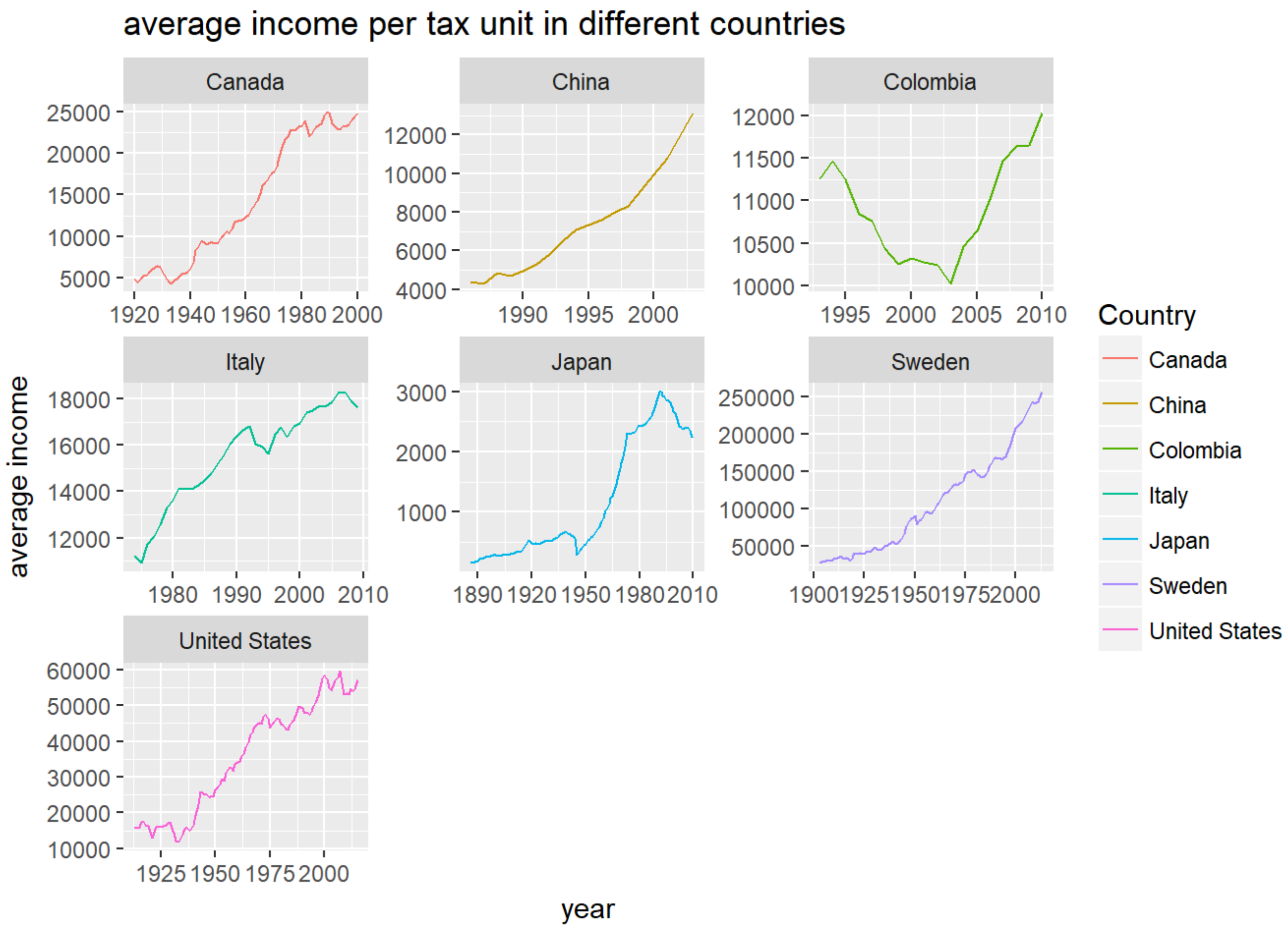



```
# plot in different picture
plot_data <- filter(Data2,P99>0)
plot_data$Estimate_a <- problem_four(data = plot_data)

ggplot(data = plot_data) +
        geom_line(mapping = aes(x=Year,y=Estimate_a,col=Country)) +
        facet_wrap(~ Country,nrow=3,scales = "free") +
        labs(title='Estimate Pareto Exponent in different country(in different pictures)',x=
'year',y='estimate pareto exponent')
```

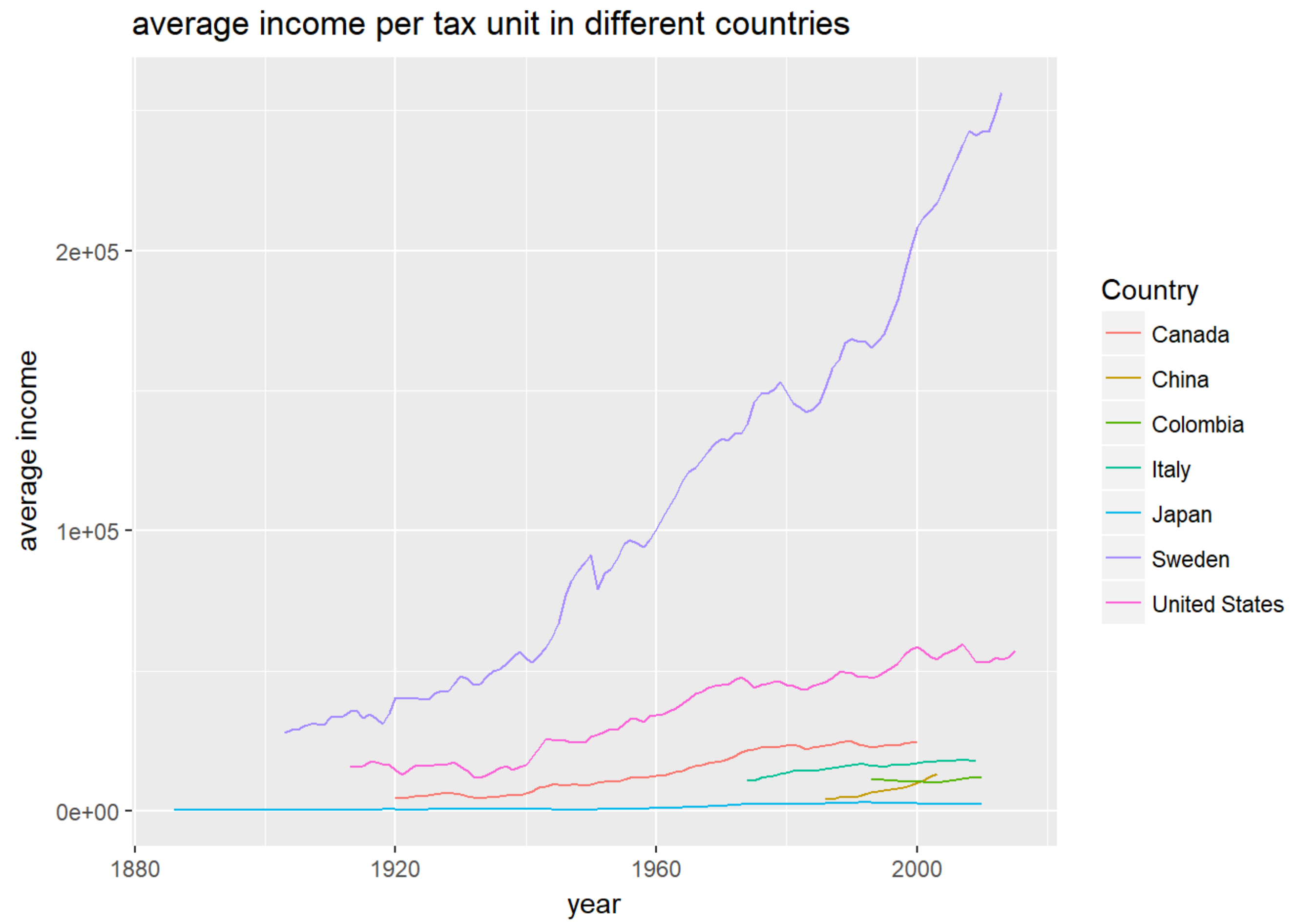


ix Plot the series of average income per unit" for the US and the countries against time in ggplot.

```
data_new <- filter(Data2,AverageIncome>0)

# in different plots
ggplot(data = data_new) +
        geom_line(mapping = aes(x=Year,y=AverageIncome,col=Country)) +
        facet_wrap(~ Country,nrow=3,scales = "free") +
        labs(title='average income per tax unit in different countries',x='year',y='average i
ncome')
```



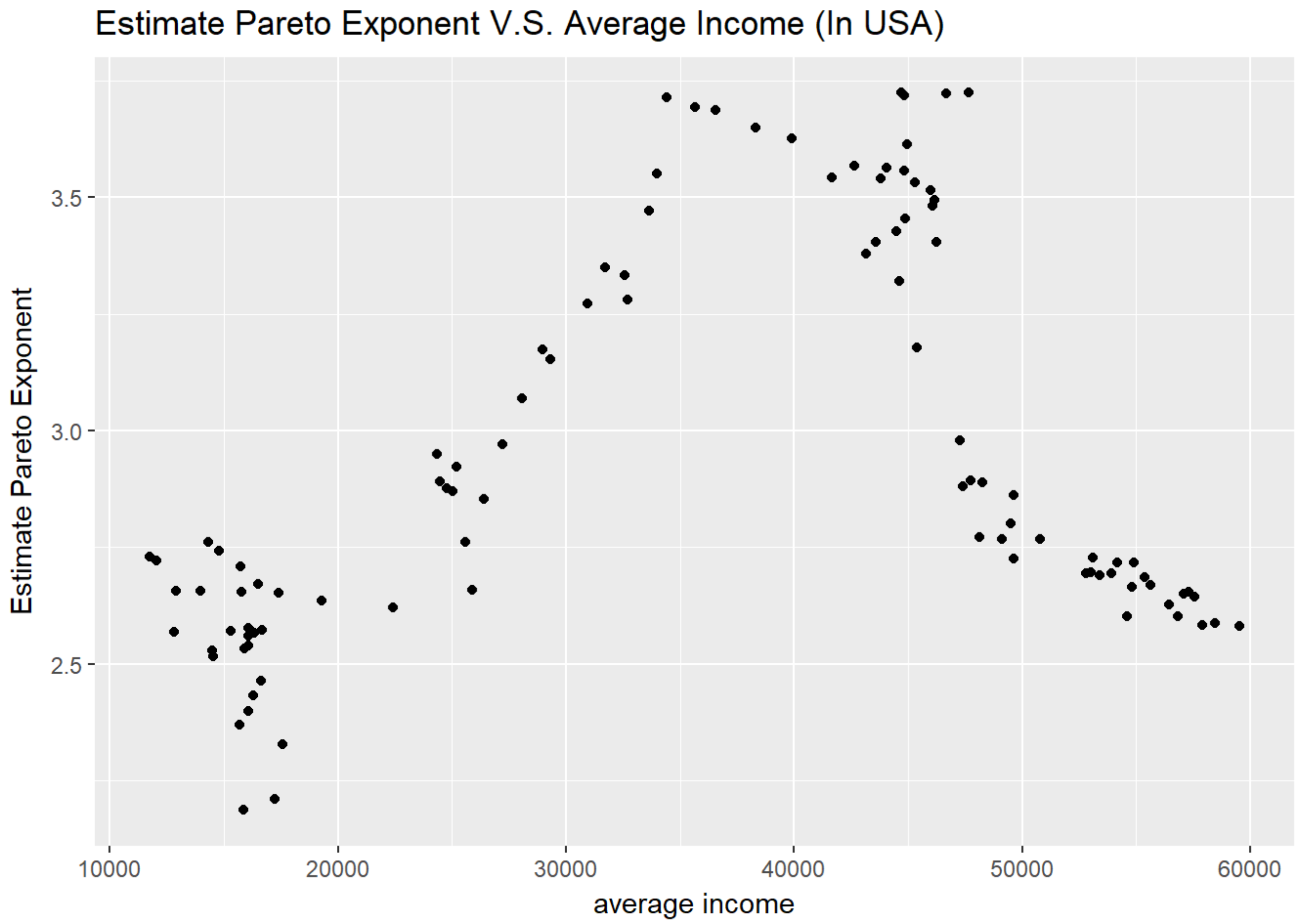

```
# in one plot

ggplot(data = data_new) +
        geom_line(mapping = aes(x=Year,y=AverageIncome,col=Country)) +
        labs(title='average income per tax unit in different countries',x='year',y='average i
ncome',size=1.5)
```



ix. Make a scatter-plot of your estimated exponents for the US against theaverage income for the US in ggplot. Qualitatively, can you say anything about the Kuznets curve? (Remember that smaller exponents indicate more income inequality.)

* in order to make the relation between average income and pareto exponent more obvious I make a picture with smooth line. *

```
ggplot(data = USA,aes(y= Estimate_a,x=AverageIncome)) +
        geom_point() +
        labs(title='Estimate Pareto Exponent V.S. Average Income (In USA)',y='Estimate Pareto
 Exponent',x='average income',size=1.5)
```



based on the plot, as we can see that first the extimate pareto exponent is increasing that means the income inequality in this country is decreasing. After the average income increasing, the income inequality start increasing. It seems to be opposite with the theorem.

X. 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 coe cients you get consistent with the hypothesis?

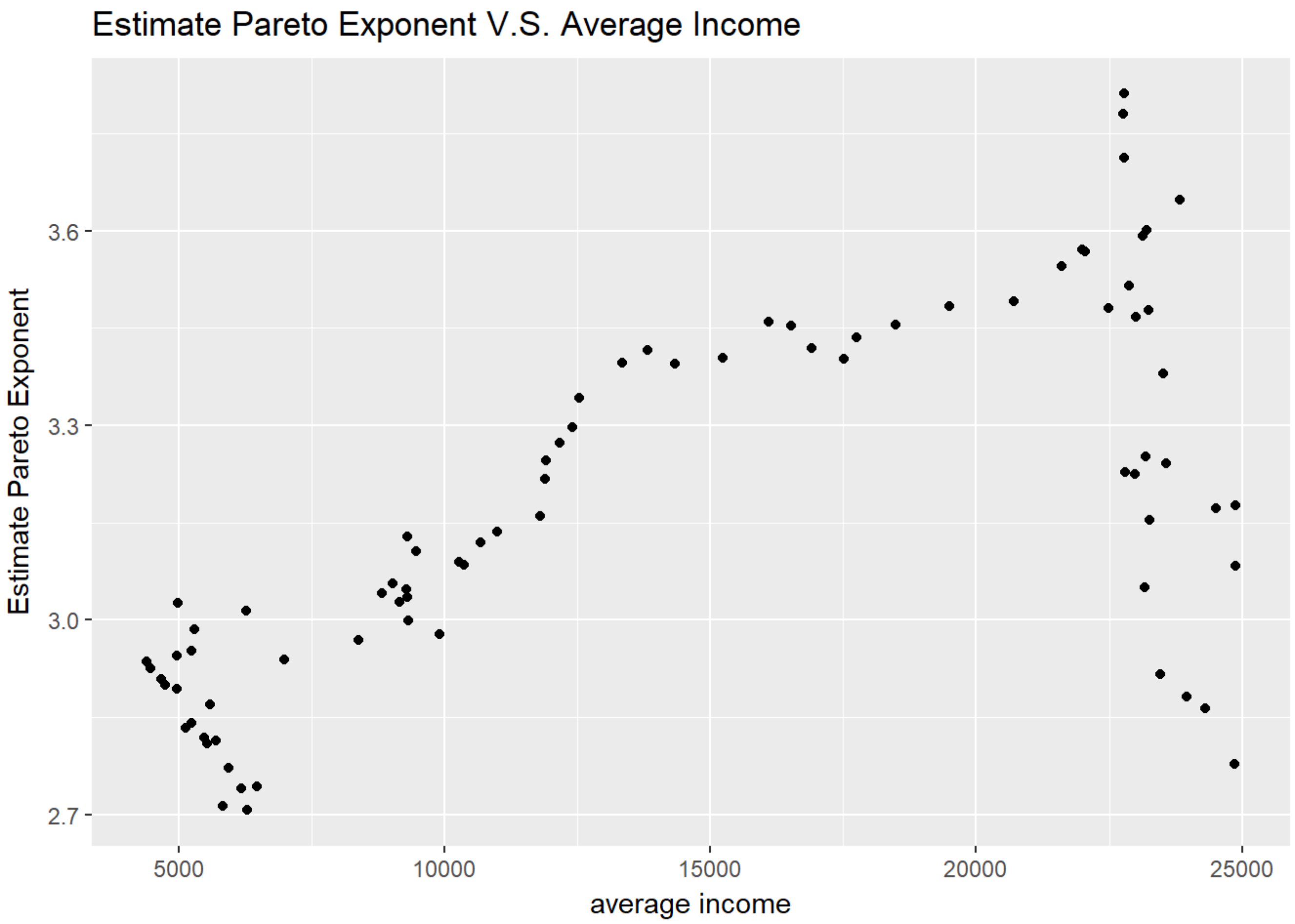
```
USA_result <- lm(USA$Estimate_a ~ USA$AverageIncome + I((USA$AverageIncome)^2))
USA_result
```

```
##
## Call:
## lm(formula = USA$Estimate_a ~ USA$AverageIncome + I((USA$AverageIncome)^2))
##
## Coefficients:
##              (Intercept)        USA$AverageIncome
##                8.230e-01                1.394e-04
## I((USA$AverageIncome)^2)
##               -1.891e-09
```

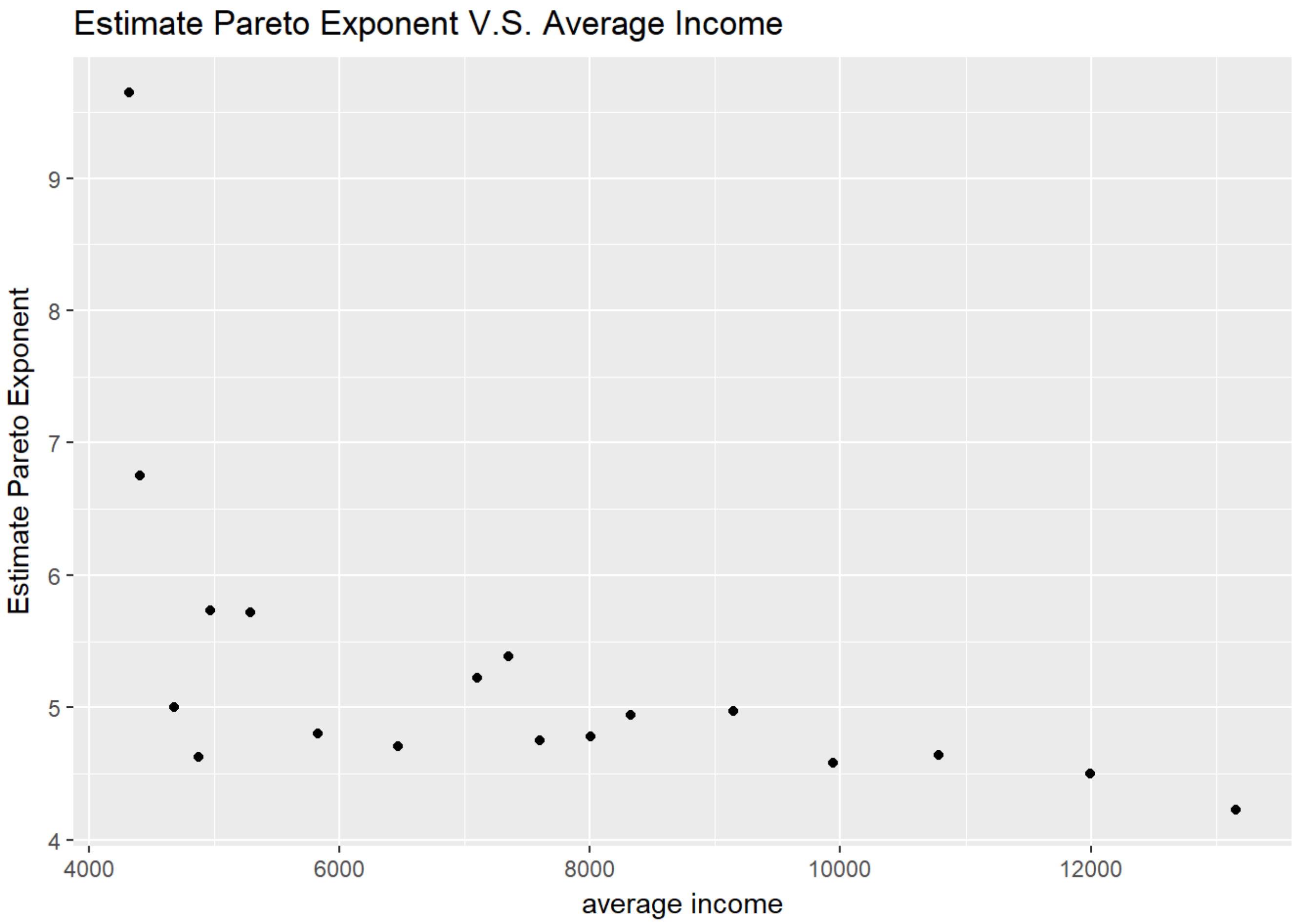
xii. Do a separate quadratic regression for each country. Which ones have estimates compatible with the hypothesis? Hint: Write a function to the model to the data for an arbitrary country.

* in order to make the relation between average income and pareto exponent more obvious I make a picture with smooth line. *

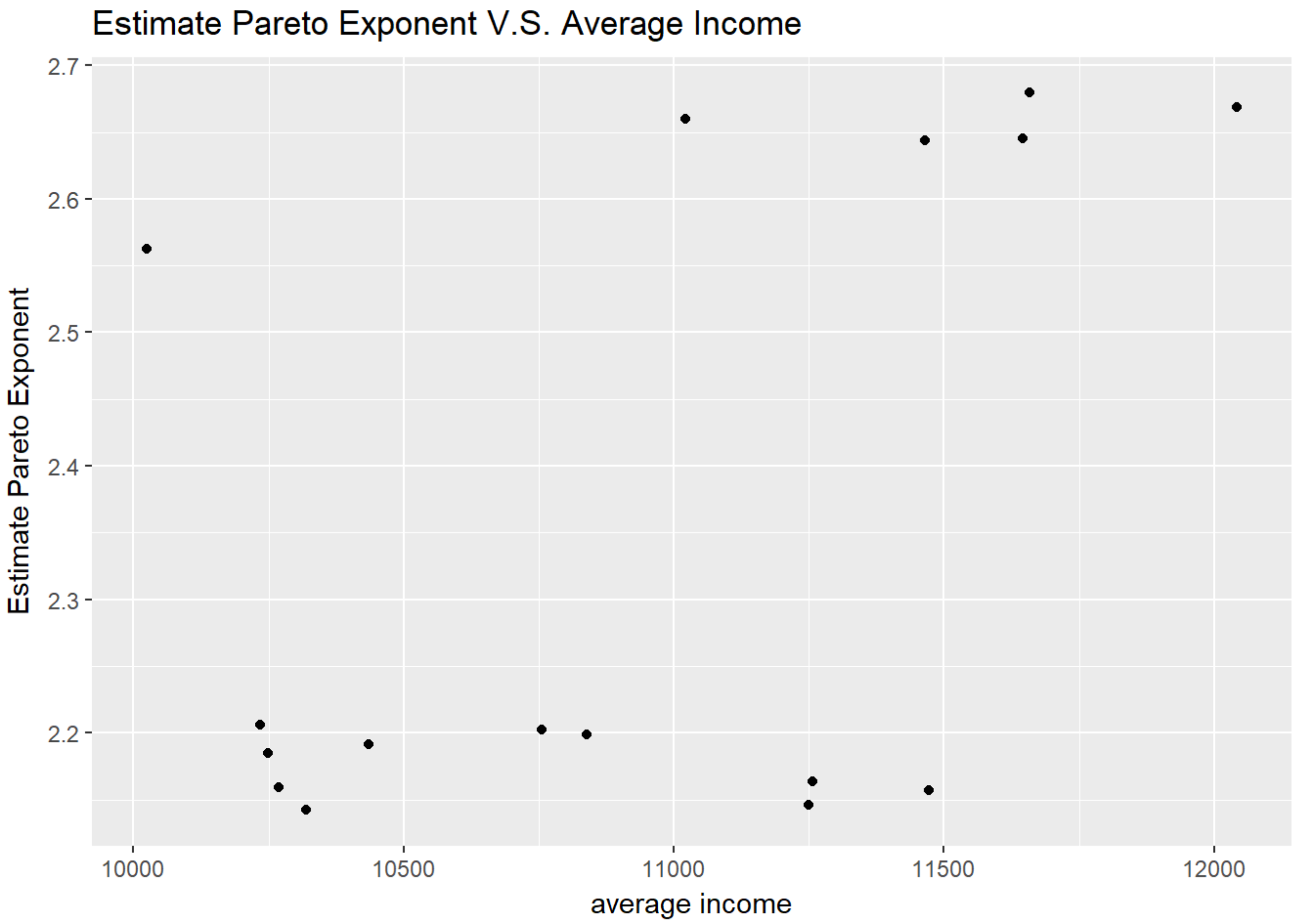
```
regression_function <- function(a,b){
        result <- lm(a ~ b + I(b^2))
}

# Canada
ggplot(data = Canada,aes(y= Estimate_a,x=AverageIncome)) +
        geom_point() +
        labs(title='Estimate Pareto Exponent V.S. Average Income',y='Estimate Pareto Exponen
t',x='average income',size=1.5)
```



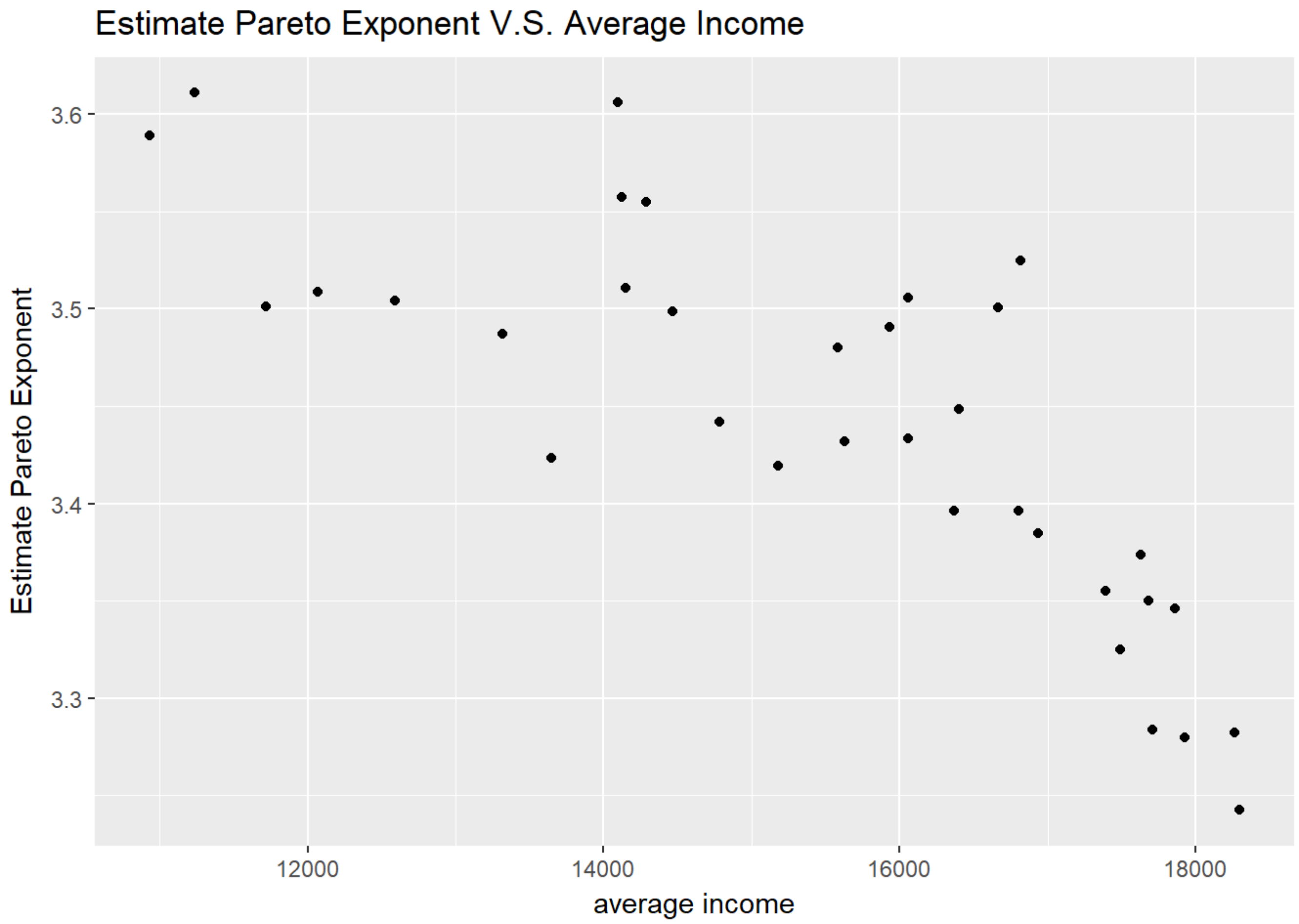

```
Canada_result <- regression_function(Canada$Estimate_a,Canada$AverageIncome)
#China
ggplot(data = China,aes(y= Estimate_a,x=AverageIncome)) +
        geom_point() +
        labs(title='Estimate Pareto Exponent V.S. Average Income',y='Estimate Pareto Exponen
t',x='average income',size=1.5)
```



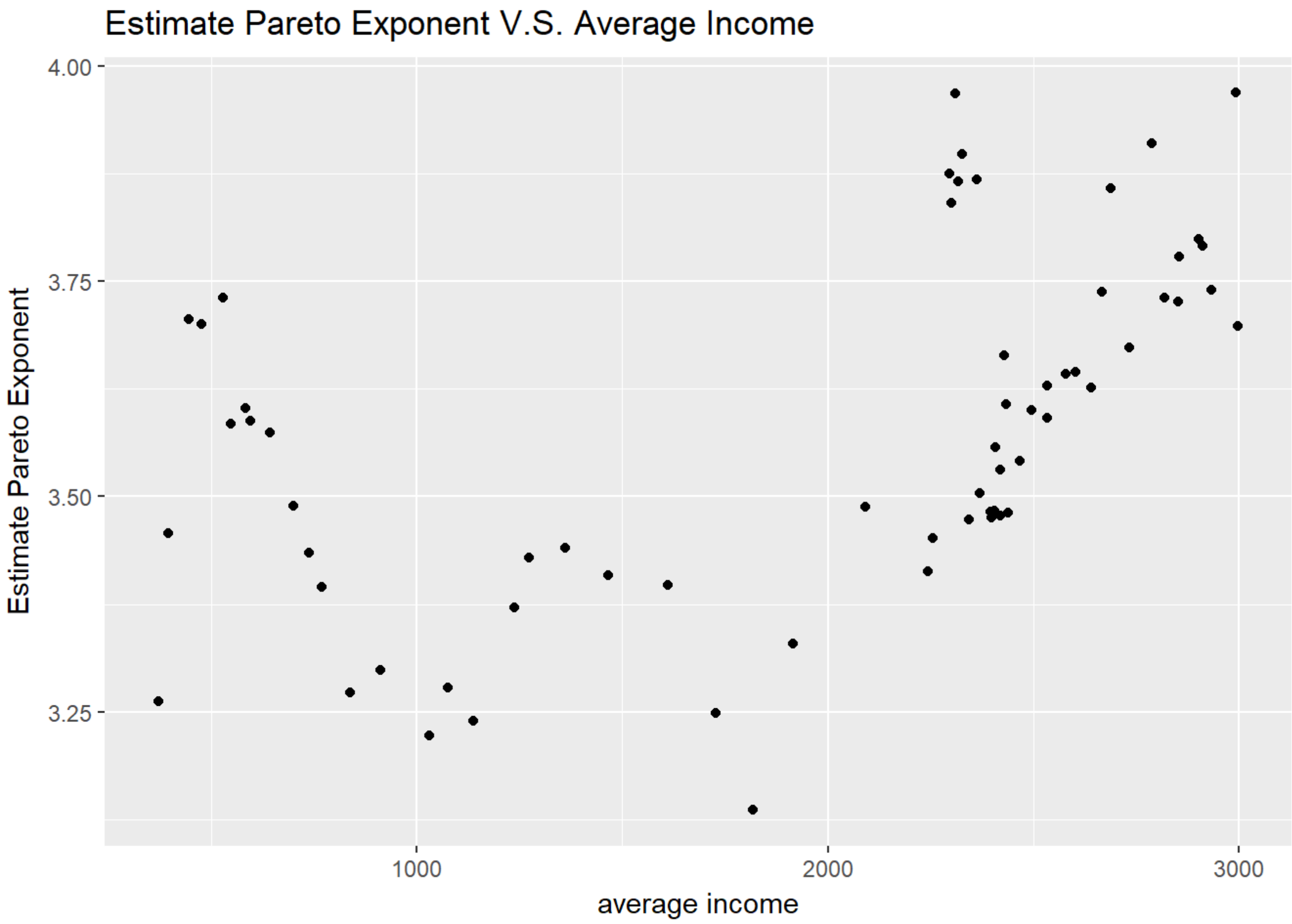

```
China_result <- regression_function(China$Estimate_a,China$AverageIncome)
#Colombia
ggplot(data = Colombia,aes(y= Estimate_a,x=AverageIncome)) +
        geom_point() +
        labs(title='Estimate Pareto Exponent V.S. Average Income',y='Estimate Pareto Exponen
t',x='average income',size=1.5)
```



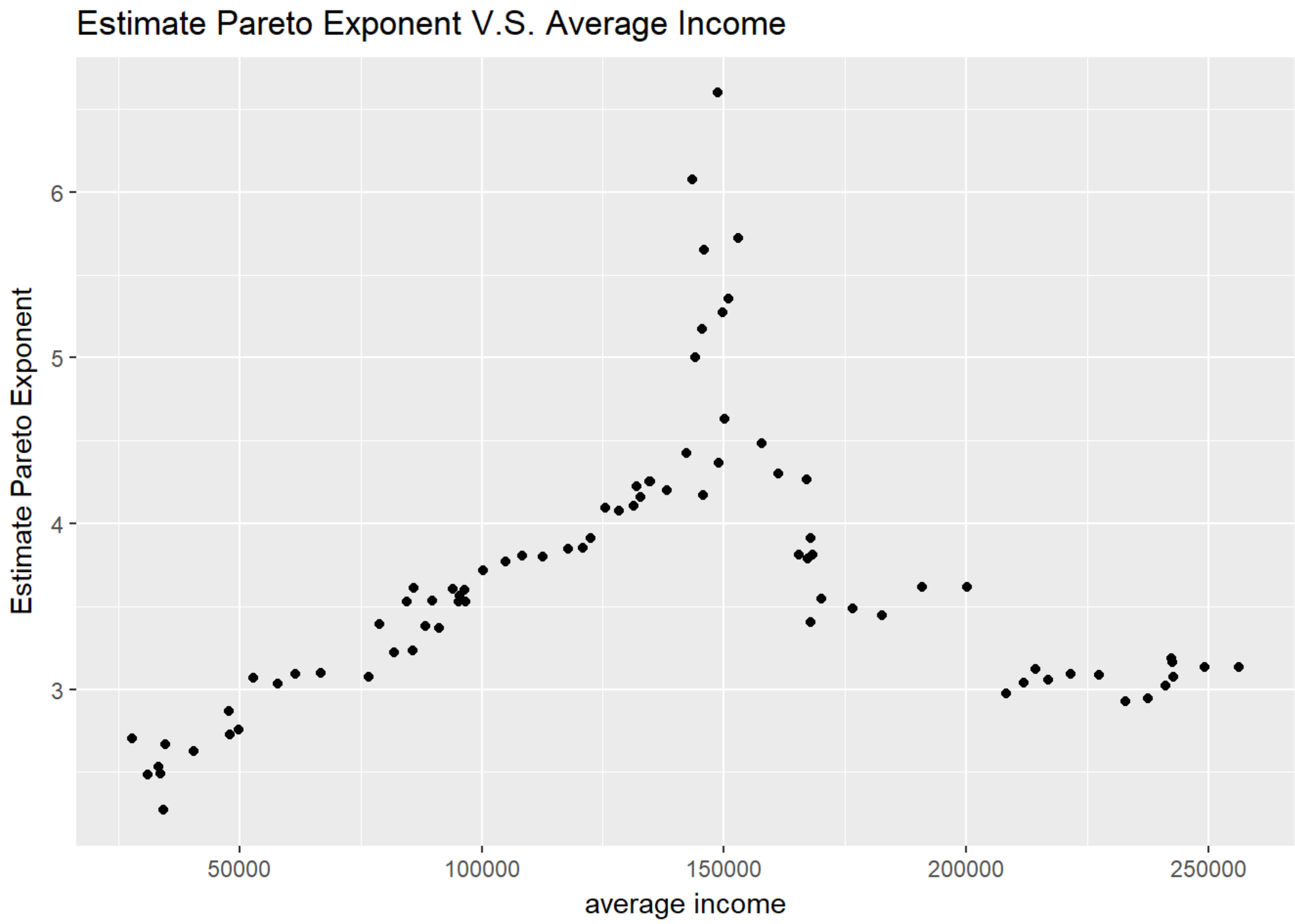

```
Colombia_result <- regression_function(Colombia$Estimate_a,Colombia$AverageIncome)
#Italy
ggplot(data = Italy,aes(y= Estimate_a,x=AverageIncome)) +
        geom_point() +
        labs(title='Estimate Pareto Exponent V.S. Average Income',y='Estimate Pareto Exponen
t',x='average income',size=1.5)
```




```
Italy_result <- regression_function(Italy$Estimate_a,Italy$AverageIncome)
#Japan
ggplot(data = Japan,aes(y= Estimate_a,x=AverageIncome)) +
        geom_point() +
        labs(title='Estimate Pareto Exponent V.S. Average Income',y='Estimate Pareto Exponen
t',x='average income',size=1.5)
```




```
Japan_result <- regression_function(Japan$Estimate_a,Japan$AverageIncome)
#Sweden
ggplot(data = Sweden,aes(y= Estimate_a,x=AverageIncome)) +
        geom_point() +
        labs(title='Estimate Pareto Exponent V.S. Average Income',y='Estimate Pareto Exponen
t',x='average income',size=1.5)
```




```
Sweden_result <- regression_function(Sweden$Estimate_a,Sweden$AverageIncome)
```

To sum up

It is obvious that if the estimate value of x^2 is positive, then we can conclude that the data from this country isa compatible with the hypothesis.

```
result <- data.frame('country'=c('Canada','China','Colombia','Italy','Japan','Sweden','USA'),
'the estimate of x^2'=c(Canada_result$coefficients[3],China_result$coefficients[3],Colombia_r
esult$coefficients[3],Italy_result$coefficients[3],Japan_result$coefficients[3],Sweden_result
$coefficients[3],USA_result$coefficients[3]))
result$whether_compatible_with_the_hypothesis <- result$the.estimate.of.x.2>0
result
```

```
##    country the.estimate.of.x.2 whether_compatible_with_the_hypothesis
## 1   Canada       -3.360837e-09                                  FALSE
## 2    China        5.257536e-08                                   TRUE
## 3 Colombia        2.867133e-07                                   TRUE
## 4    Italy       -6.591048e-09                                  FALSE
## 5    Japan        1.889447e-07                                   TRUE
## 6   Sweden       -1.496762e-10                                  FALSE
## 7      USA       -1.890556e-09                                  FALSE
```