Homework 3 Solutions

Cynthia Rush (cgr2130)

October 30, 2017

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 (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`.

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

Part 1: Estimating a on US data

```r
wtid <- read.csv("wtid-report.csv", as.is = TRUE)
wtid <- wtid[, c("Year", "P99.income.threshold", "P99.5.income.threshold", "P99.9.income.threshold")]
names(wtid) <- c("Year", "P99", "P99.5", "P99.9")

exponent.est_ratio <- function(p99, p999) {
  return(1 - log(10)/(log(p99/p999)))
}

ahat <- exponent.est_ratio(wtid$P99, wtid$P99.9)
library(ggplot2)
```

(i.) In lab we use the fact that we can estimate the exponent using the following formula, which we refer to as Equation (1):

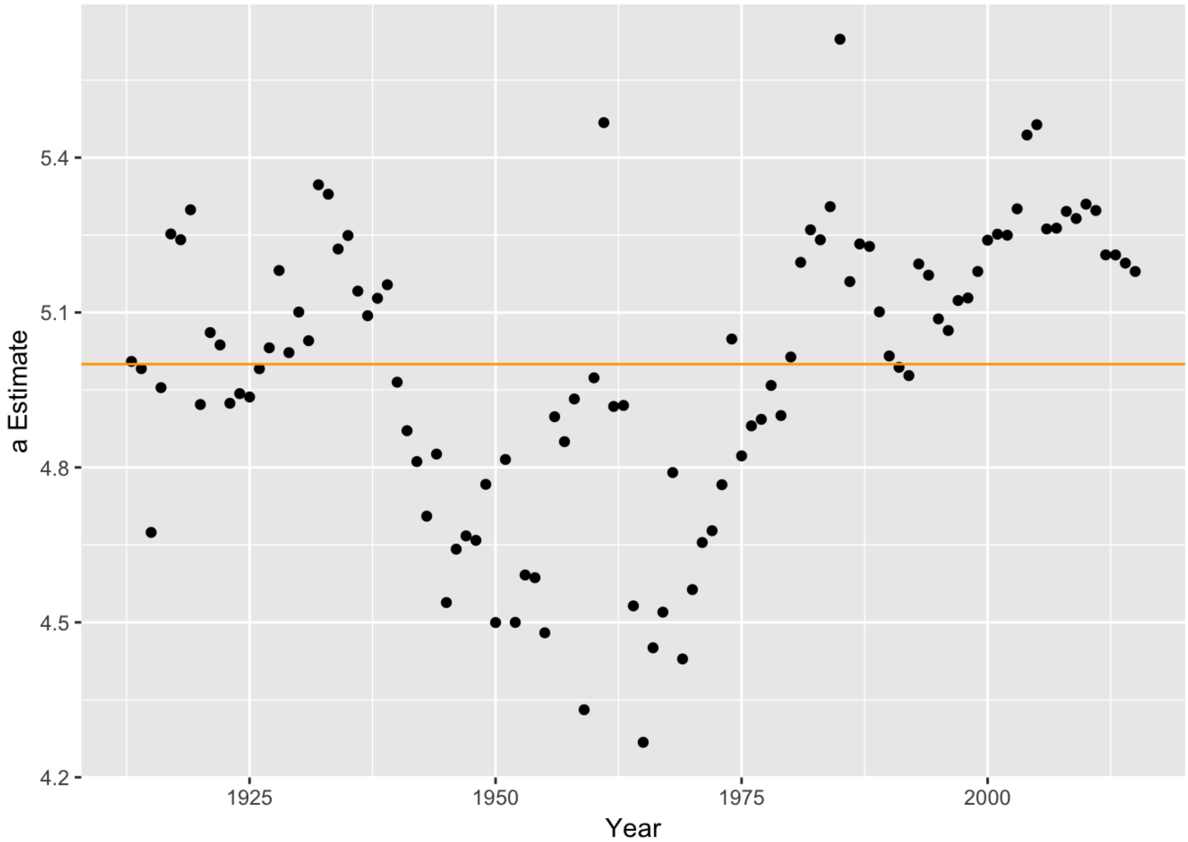
$$a = 1 - \frac{\log 10}{\log(\frac{P99}{P99.9})}.$$

The logic leading to Equation (1) also implies that

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

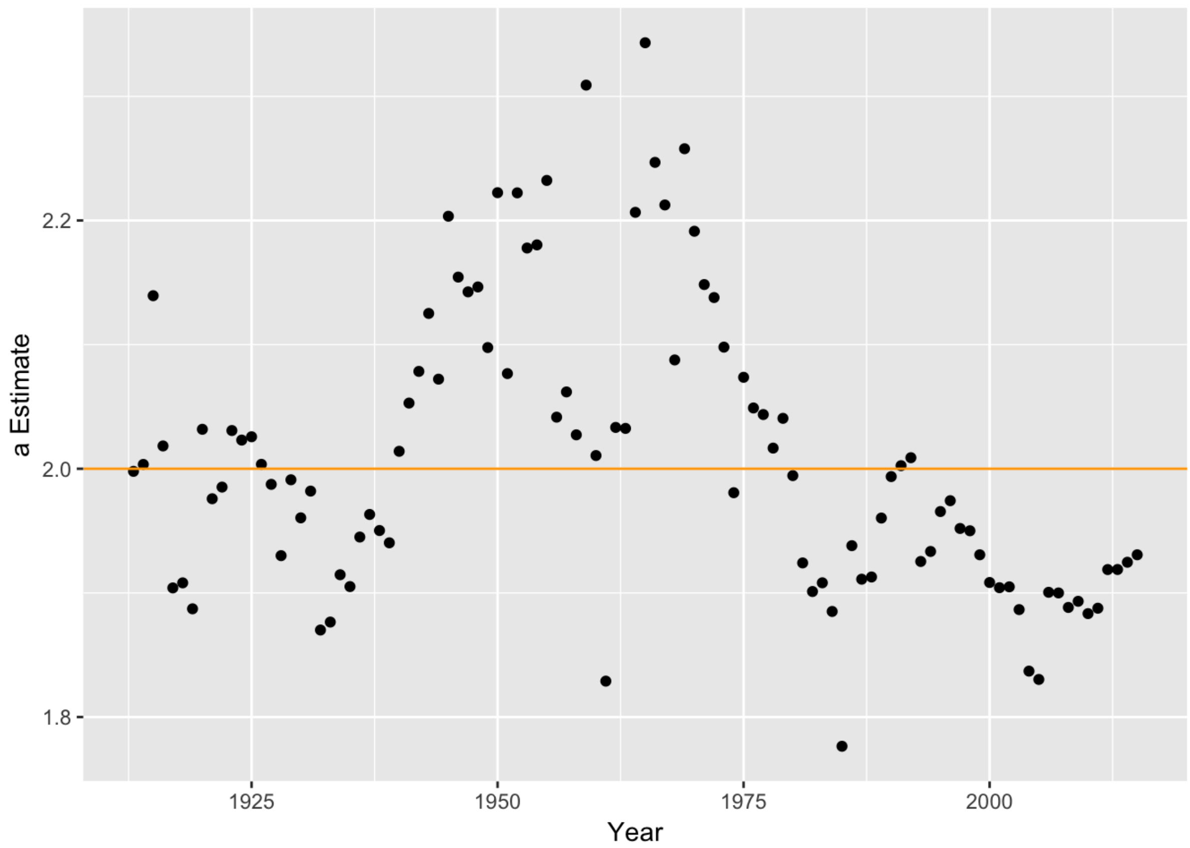
Write a function which takes `P99.5`, `P99.9`, and a, and calculates the left-hand side of that equation. Plot the values for each year using `ggplot`, using the data and your estimates of the exponent. Add a horizontal line with vertical coordinate 5. How good is the fit?

```
exponent.est_ratio2 <- function(p995, p999, a) {
  return((p995/p999)^(-a+1))
}
ests <- exponent.est_ratio2(wtid$P99.5, wtid$P99.9, ahat)
ggplot(data = wtid) +
  geom_point(mapping = aes(x = Year, y = ests)) +
  geom_abline(intercept = 5, slope = 0, color = "orange") +
  labs(x = "Year", y = "a Estimate")
```



(ii.) 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 one?

```
exponent.est_ratio3 <- function(p99, p995, a) {
  return((p99/p995)^(-a+1))
}
ests2 <- exponent.est_ratio3(wtid$P99, wtid$P99.5, ahat)
ggplot(data = wtid) +
  geom_point(mapping = aes(x = Year, y = ests2)) +
  geom_abline(intercept = 2, slope = 0, color = "orange") +
  labs(x = "Year", y = "a Estimate")
```



(iii.) We have shown that if the upper tail of the income distribution followed a perfect Pareto distribution, then

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

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

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

We could estimate the Pareto exponent by solving any one of these equations for a; we did this in lab and in the previous two questions. Because of measurement error and sampling noise, we can't find one value of a which will work for all three equations - . Generally, trying to make all three equations come close to balancing gives a better estimate of a than just solving one of them. (This is analogous to finding the slope and intercept of a regression line by trying to come close to all the points in a scatterplot, and not just running a line through two of them.)

We will therefore estimate a by minimizing

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

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.

```
percentile_ratio_discrepancies <- function(a, p99, p995, p999) {
  return(((p99/p999)^(-a+1) - 10)^2 + ((p995/p999)^(-a+1) - 5)^2 + ((p99/p995)^(-a+1) - 2)^2)
}
percentile_ratio_discrepancies(2, 1e6, 2e6, 1e7)
```

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

(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.

Recall that in class we used gradient descent to minimize the mean squared error (MSE) of a model fit depending on a parameter β to data. For the gradient descent algorithm, we approximated the derivative of the MSE, and adjusted our estimate of β by an amount proportional (and opposite) to that approximation. We stopped the algorithm when the derivative became small (assuming, then, that we were near a minimum). For this homework we will use a built-in `R` optimization function to do the minimization, which essentially does a fancier version of what we did in class.

`R` has several built-in functions for optimization, and one of the simplest, which we use today, is `nlm()`, or non-linear minimization. `nlm()` takes two required arguments, a function to minimize and an initial value for the parameter.

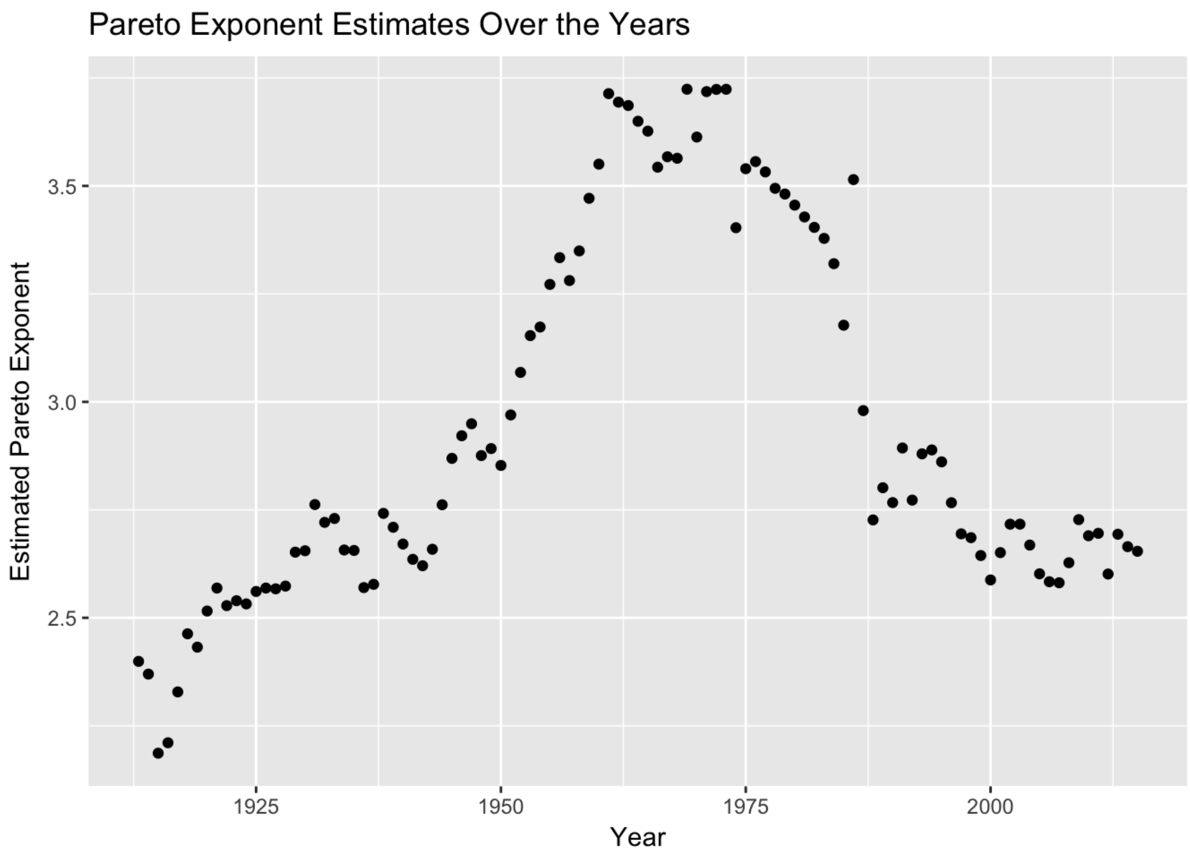
For this problem write a function, `exponent.multi_ratios_est`, which takes as inputs the vectors `P99`, `P99.5`, `P99.9`, and estimates a by minimizing the function `percentile_ratio_discrepancies`. The initial value for the minimization should come from the estimate of a given by Equation (1). Check that when `P99=1e6`, `P99.5=2e6` and `P99.9=1e7`, your function returns an a of 2.

```
exponent.multi_ratios_est <- function(p99, p995, p999) {
  a0  <- 1 - (log(10))/(log(p99/p999))
  min <- nlm(percentile_ratio_discrepancies, a0, p99, p995, p999)$estimate
  return(min)
}
exponent.multi_ratios_est(1e6, 2e6, 1e7)
```

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

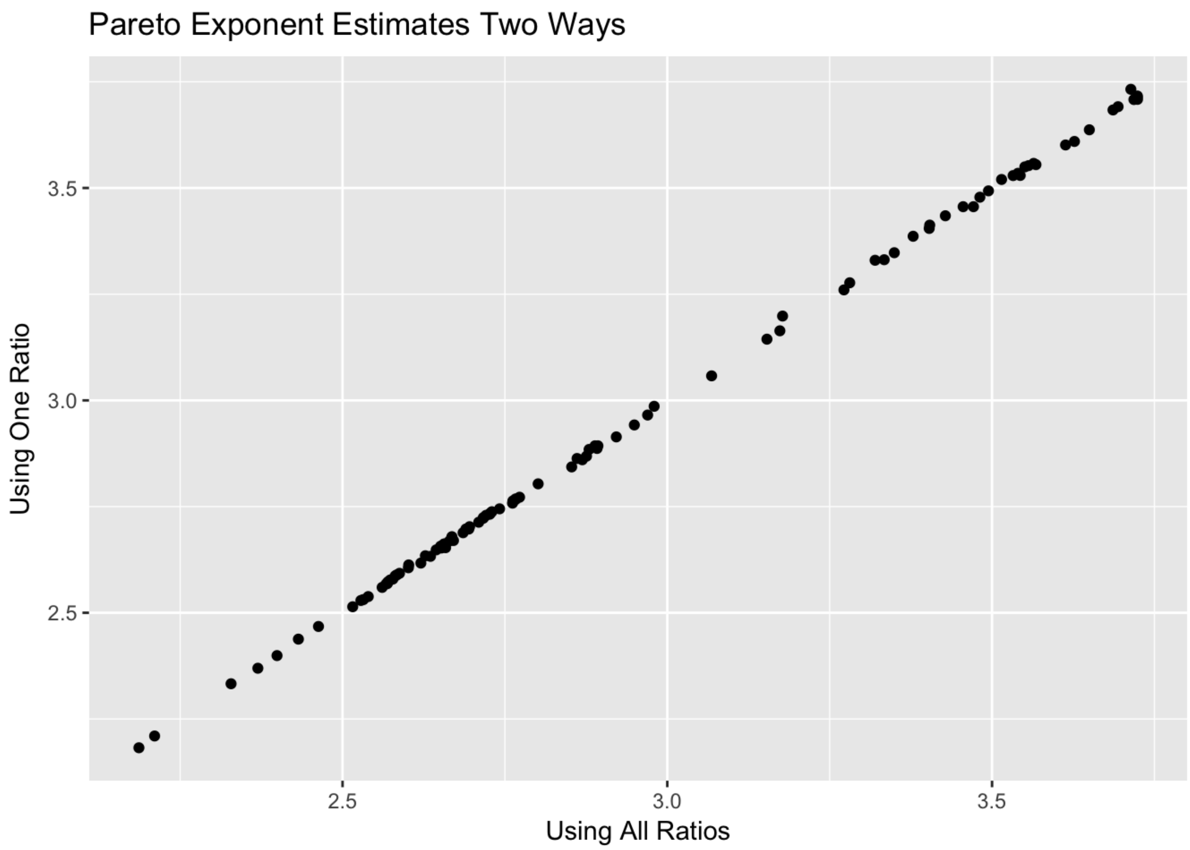
(v.) Write a function which uses `exponent.multi_ratios_est` to estimate a for the US for every year from 1913 to 2012. (There are many ways you could do this, including loops.) Plot the estimates using `ggplot`; make sure the labels of the plot are appropriate.

```
multi_ratios_allyears <- function(country_data) {

  # country_data should be a dataframe with the variables "Year", "P99", "P99.5", and "P99.9" where each
 row corresponds to one year.

  n <- nrow(country_data)
  estimates <- rep(NA, n)
  names(estimates) <- country_data$Year
  for (i in 1:n) {
    if (all(is.na(country_data[i, c("P99", "P99.5", "P99.9")]))) {
      estimates[i] <- NA
    } else {
      estimates[i] <- exponent.multi_ratios_est(country_data$P99[i], country_data$P99.5[i], country_data$
P99.9[i])
    }
  }
  return(estimates)
}
USestimates <- multi_ratios_allyears(wtid)

ggplot(data = wtid) +
  geom_point(mapping = aes(x = Year, y = USestimates)) +
  labs(title = "Pareto Exponent Estimates Over the Years", x = "Year", y = "Estimated Pareto Exponent")
```



(vi.) Use Equation (1) to estimate a for the US for every year. Make a scatter-plot of these estimates against those from problem (v.) using `gplot` . 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?

```
single_ratio  <- 1 - (log(10))/(log(wtid$P99/wtid$P99.9))
names(single_ratio) <- wtid$Year

ggplot(data = wtid) +
  geom_point(mapping = aes(x = USestimates, y = single_ratio)) +
  labs(title = "Pareto Exponent Estimates Two Ways", x = "Using All Ratios", y = "Using One Ratio")
```



(vii.) 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-report.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.

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 these countries, and there may be some NA values.

```
wtid2 <- read.csv("wtid-homework.csv", as.is = TRUE)
country_num <- length(unique(wtid2$Country))

for(i in 1:country_num) {
  these_rows <- wtid2$Country == unique(wtid2$Country)[i]
  this_data  <- wtid2[these_rows, c("Year", "P99", "P99.5", "P99.9")]
  estimates  <- multi_ratios_allyears(this_data)
  wtid2$Estimate[these_rows] <- estimates
}
```

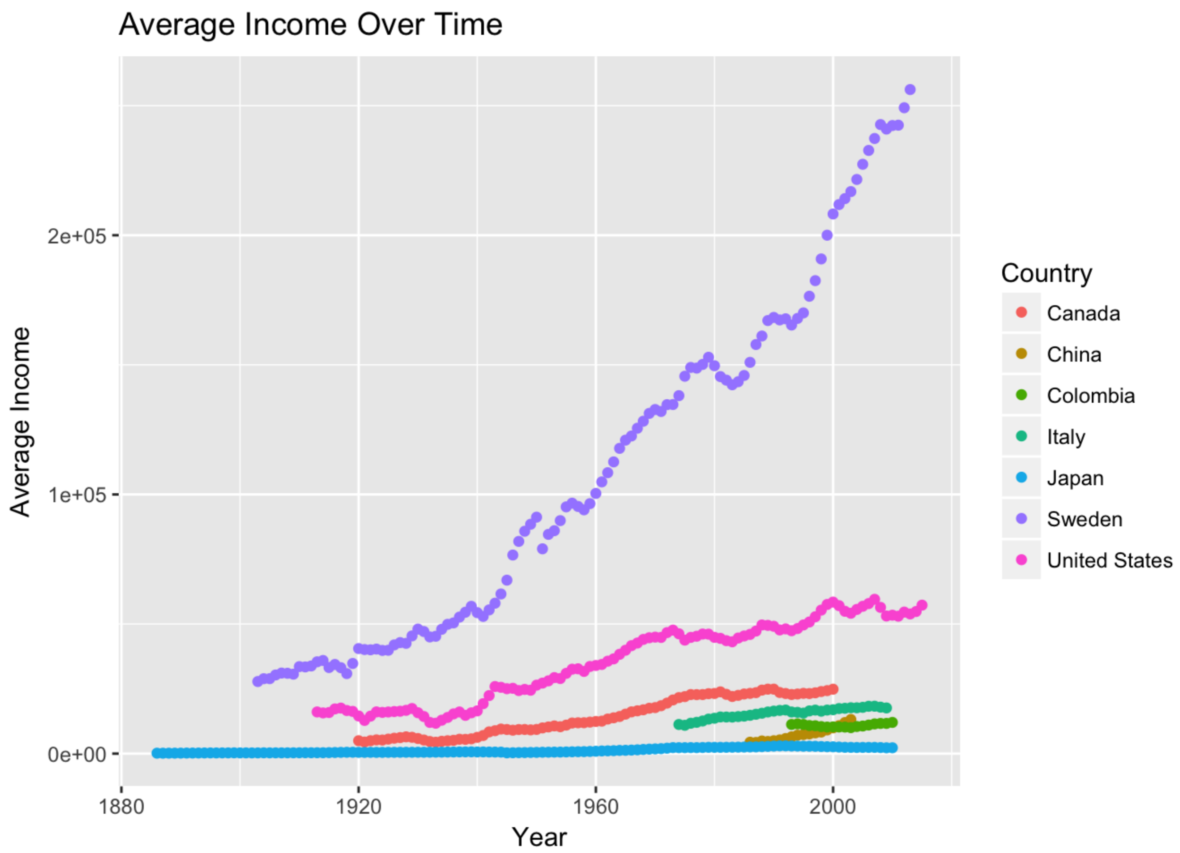
(viii.) Plot your estimates of a over time for all the countries using `ggplot`. Note that the years covered by the data are different for each country. You may either make multiple plots, or put all the series into one plot. Either way, make sure that the plots are clearly labeled.

```
ggplot(data = wtid2) +
  geom_point(mapping = aes(x = Year, y = Estimate, color = Country)) +
  labs(title = "Pareto Exponent Estimate", y = "Estimate", x = "Year")
```

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

```
ggplot(data = wtid2) +
  geom_point(mapping = aes(x = Year, y = AverageIncome, color = Country)) +
  labs(title = "Average Income Over Time", y = "Average Income", x = "Year")
```

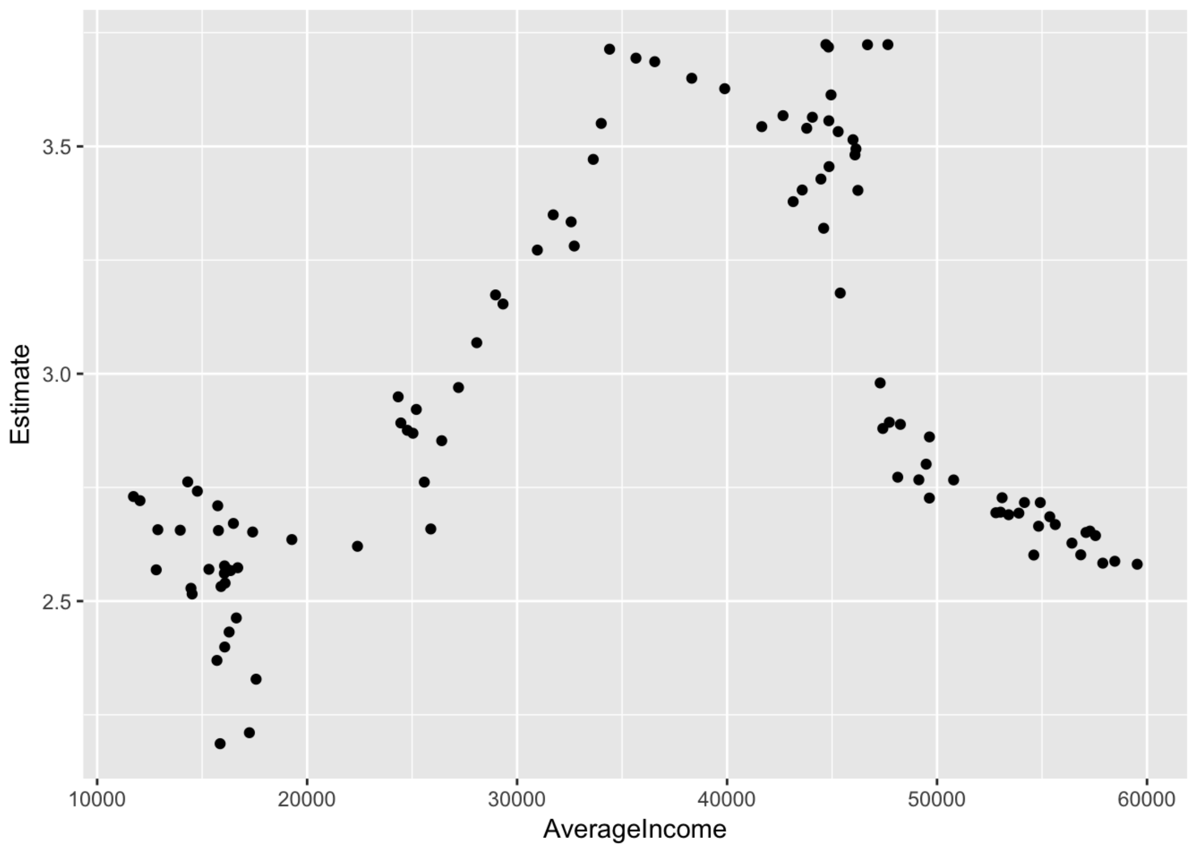
```
## Warning: Removed 10 rows containing missing values (geom_point).
```



(x.) The most influential hypothesis about how inequality is linked to economic growth is the “U-curve” hypothesis proposed by the great economist Simon Kuznets in the 1950s. According tho this idea, inequality rises during the early, industrializing phases of economic growth, but then declines as growth continues.

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? (Remember that smaller exponents indicate more income inequality.)

```
ggplot(data = wtid2[wtid2$Country == "United States", ]) +
  geom_point(mapping = aes(x = AverageIncome, y = Estimate))
```



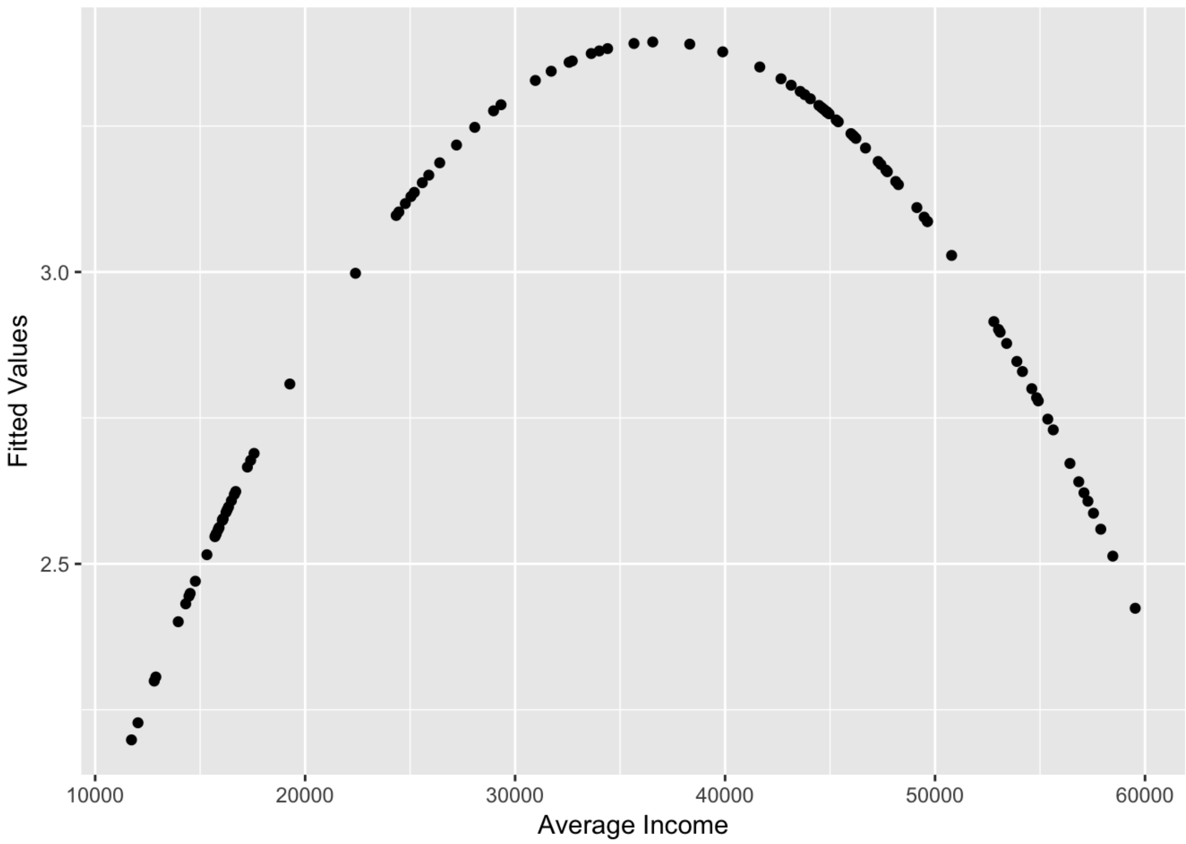
Our scatterplot doesn't seem to support this hypothosis. Smaller a coefficients mean more inequality, so our scatterplot shows inequality high, then decreasing for a while but now rising again – the opposite of the "U-curve" hypothesis.

(xi.) For a more quantitative check on the Kuznets hypothesis, use `lm()` to regress your estimated exponents on the average income for the US, including a quadratic term for income. Are the coefficients you get consistent with the hypothesis? Hint: $lm(y \tilde{}\ x + I(x^2))$ will regress y on both x and x^2.

```
lm0 <- lm(Estimate ~ AverageIncome + I(AverageIncome^2), data = wtid2[wtid2$Country == "United States", ]
)
summary(lm0)
```

```
##
## Call:
## lm(formula = Estimate ~ AverageIncome + I(AverageIncome^2), data = wtid2[wtid2$Country ==
##     "United States", ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50724 -0.18364 -0.02531  0.18689  0.54918
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         8.230e-01  1.515e-01   5.432 3.93e-07 ***
## AverageIncome       1.394e-04  1.015e-05  13.740  < 2e-16 ***
## I(AverageIncome^2) -1.891e-09  1.451e-10 -13.027  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2466 on 100 degrees of freedom
## Multiple R-squared:  0.6679, Adjusted R-squared:  0.6612
## F-statistic: 100.5 on 2 and 100 DF,  p-value: < 2.2e-16
```

```
ggplot(data = wtid2[wtid2$Country == "United States", ]) +
  geom_point(mapping = aes(x = AverageIncome, y = fitted(lm0))) +
  labs(main = "United States", x = "Average Income", y = "Fitted Values")
```



Our coefficients are not consistent with the hypothesis. Since the coefficient in front of the quadratic term is negative, the model is a parabola opening downwards, not upwards, like a ‘U’ would.

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

```r
reg_fit <- function(data, country) {
  # data is a data frame with the variables Country, Estimate, and Average Income.  This function returns
 the linear regression coefficients and plots the fitted values against the average income.

  if (all(is.na(data$Estimate[data$Country == country]))) {
    return(rep(NA, 3))
  } else {
    lm0 <- lm(Estimate ~ AverageIncome + I(AverageIncome^2), data = data[data$Country == country, ])
    summary(lm0)
    return(lm0$coef)
  }
}

country_num <- length(unique(wtid2$Country))

coefs <- matrix(NA, ncol = 3, nrow = country_num)
colnames(coefs) <- c("Intercept", "AverageIncome", "AverageIncome2")
rownames(coefs) <- unique(wtid2$Country)

for (i in 1:country_num) {
  coefs[i, ] <- reg_fit(wtid2, unique(wtid2$Country)[i])
}
coefs
```

```
##                Intercept AverageIncome AverageIncome2
## Canada         2.2660536  1.240966e-04  -3.360837e-09
## China         10.3978092 -1.126763e-03   5.257536e-08
## Colombia      34.6124009 -6.095234e-03   2.867133e-07
## Italy          2.5824163  1.594300e-04  -6.591048e-09
## Japan          3.7291069 -5.136191e-04   1.889447e-07
## Sweden         1.0123533  4.414454e-05  -1.496762e-10
## United States  0.8230049  1.394435e-04  -1.890556e-09
```

From the coefficients output, the relationship between the coefficient estimates and the average income is modeled by a “U” shape for China, Colombia, and Japan, but not for the others.

(If we were doing a more rigorous check of the Kuznet hypothesis, we would want to control for other factors, and not just assume that a quadratic was the right functional form for the curve.)