Homework 5: Pareto and Kuznets on the Grand Tour

We continue working with the World Top Incomes Database [https://wid.world], and the Pareto distribution, as in the lab. We also continue to practice working with data frames, manipulating data from one format to another, and writing functions to automate repetitive tasks.

We saw in the lab that if the upper tail of the income distribution followed a perfect Pareto distribution, then

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

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

$$\left(\frac{P99}{P99.5}\right)^{-a+1} = 2 \tag{3}$$

We could estimate the Pareto exponent by solving any one of these equations for a; in lab we used

$$a = 1 - \frac{\log 10}{\log\left(P99/P99.9\right)} \ , \tag{4}$$

Because of measurement error and sampling noise, we can't find find one value of a which will work for all three equations (1)–(3). 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.)

1. We 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(P99, P99.5, P99.9, a) {
  ratio1 <- P99 / P99.9
  ratio2 <- P99.5 / P99.9
  ratio3 <- P99 / P99.5

  term1 <- (ratio1^(-a + 1) - 10)^2
  term2 <- (ratio2^(-a + 1) - 5)^2
  term3 <- (ratio3^(-a + 1) - 2)^2
  term1 + term2 + term3
}

test_value <- percentile_ratio_discrepancies(P99 = 1e6, P99.5 = 2e6, P99.9 = 1e7, a = 2)
cat("Verify test values:", test_value, "(should be 0)\n")
```

```
## Verify test values: 0 (should be 0)
```

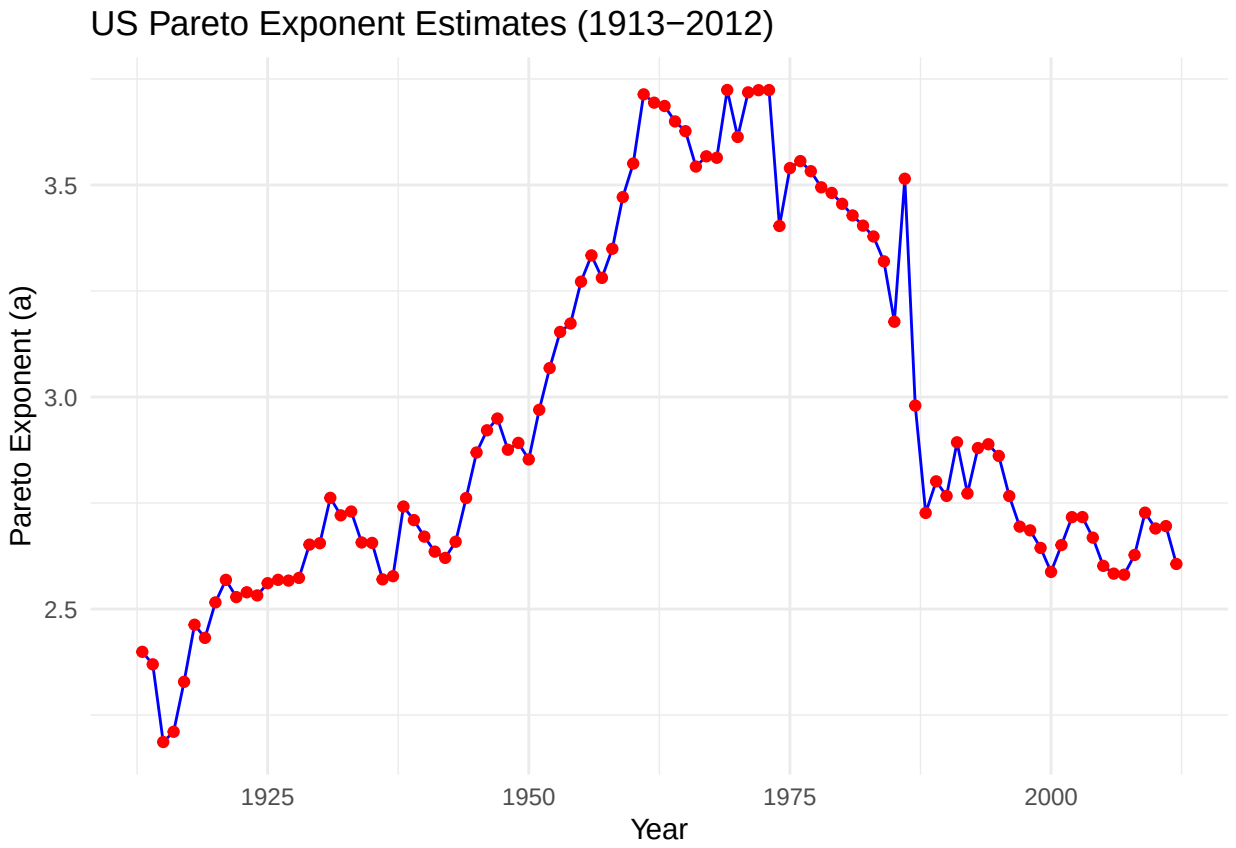
2. Write a function, `exponent.multi_ratios_est`, which takes as inputs `P99`, `P99.5`, `P99.9`, and estimates `a`. It should minimize your `percentile_ratio_discrepancies` function. The starting value for the minimization should come from (4). Check that when `P99=1e6`, `P99.5=2e6` and `P99.9=1e7`, your function returns an `a` of 2.

```
exponent.multi_ratios_est <- function(P99, P99.5, P99.9) {
  initial_a <- 1 - log(10) / log(P99 / P99.9)
  objective_fn <- function(a) {
    percentile_ratio_discrepancies(P99, P99.5, P99.9, a)
  }
  result <- optimize(
    f = objective_fn,
    interval = c(initial_a - 2, initial_a + 2)
  )
  result$minimum
}

test_a <- exponent.multi_ratios_est(P99 = 1e6, P99.5 = 2e6, P99.9 = 1e7)
cat("Verify test estimates:", test_a, "(should be 2)\n")
```

```
## Verify test estimates: 2 (should be 2)
```

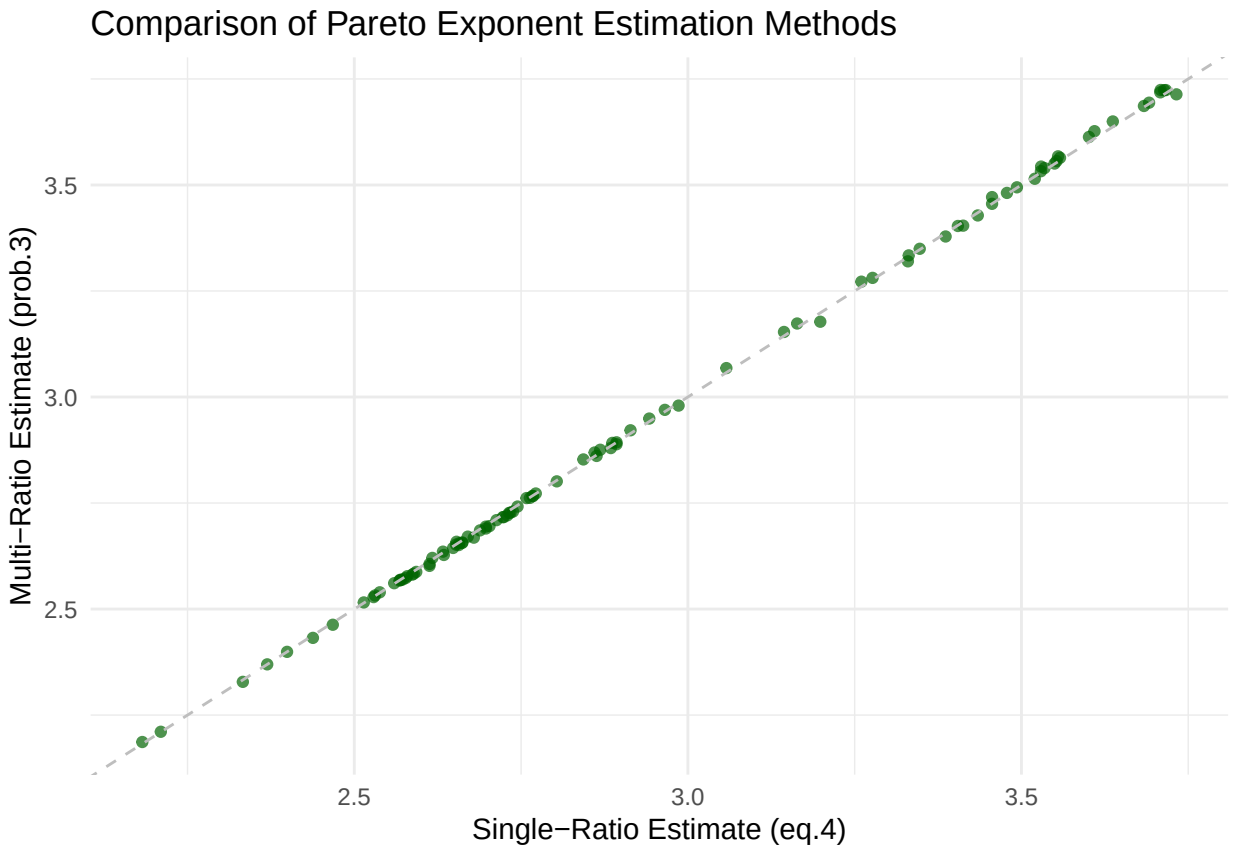
3. Write a function which uses `exponent.multi_ratios_est` to estimate a for the US for every year from 1913 to 2012. (There are many ways you could do thi, including loops.) Plot the estimates; make sure the labels of the plot are appropriate.

```
income_data <- read.csv("data/wtid-report.csv")

income_data$multi_ratio_a <- mapply(
  exponent.multi_ratios_est,
  P99 = income_data$P99.income.threshold,
  P99.5 = income_data$P99.5.income.threshold,
  P99.9 = income_data$P99.9.income.threshold
)

library(ggplot2)
ggplot(income_data, aes(x = Year, y = multi_ratio_a)) +
  geom_line(color = "blue") +
  geom_point(color = "red") +
  labs(title = "US Pareto Exponent Estimates (1913-2012)",
       x = "Year", y = "Pareto Exponent (a)") +
  theme_minimal()
```



4. Use (4) to estimate a for the US for every year. Make a scatter-plot of these estimates against those from problem 3. 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?

```
income_data$single_ratio_a <- 1 - log(10) / log(income_data$P99.income.threshold / income_data$P99.9.in

ggplot(income_data, aes(x = single_ratio_a, y = multi_ratio_a)) +
  geom_point(alpha = 0.7, color = "darkgreen") +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "gray") +
  labs(title = "Comparison of Pareto Exponent Estimation Methods",
       x = "Single-Ratio Estimate (eq.4)",
       y = "Multi-Ratio Estimate (prob.3)") +
  theme_minimal()
```




```
correlation <- cor(income_data$single_ratio_a, income_data$multi_ratio_a)
cat("Correlation coefficient between the two estimation methods:", round(correlation, 3), "\n")
```

```
## Correlation coefficient between the two estimation methods: 1
```

```
mean_diff <- mean(income_data$multi_ratio_a - income_data$single_ratio_a)
cat("Multi-equation estimation averages higher than single-equation estimation:", round(mean_diff, 3),
```

```
## Multi-equation estimation averages higher than single-equation estimation: 0
```