Homework 5: Pareto and Kuznets on the Grand Tour

2025 暑假短学期《数学实践》

任心然, 3220103492

2025/06/30

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`.

```r
percentile_ratio_discrepancies <- function(P99, P99.5, P99.9, a) {
  term1 <- ( (P99 / P99.9)^(-a + 1) - 10 )^2
  term2 <- ( (P99.5 / P99.9)^(-a + 1) - 5 )^2
  term3 <- ( (P99 / P99.5)^(-a + 1) - 2 )^2
  return(term1 + term2 + term3)
}
```

```r
# Test case
P99 <- 1e6
P99.5 <- 2e6
P99.9 <- 1e7
a <- 2
result <- percentile_ratio_discrepancies(P99, P99.5, P99.9, a)
print(result)  # Expected output: 0
```

```
## [1] 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.

```r
exponent.multi_ratios_est <- function(P99, P99.5, P99.9) {
  # Initial guess from (4)
  a_init <- 1 - log(10) / log(P99 / P99.9)

  # Minimize the discrepancy function
  result <- optimize(
    f = function(a) percentile_ratio_discrepancies(P99, P99.5, P99.9, a),
    interval = c(a_init - 2, a_init + 2)  # Search around initial guess
  )

  return(result$minimum)
}
```

```r
# Test case
P99 <- 1e6
P99.5 <- 2e6
P99.9 <- 1e7
a_estimated <- exponent.multi_ratios_est(P99, P99.5, P99.9)
print(a_estimated)  # Expected output: 2
```

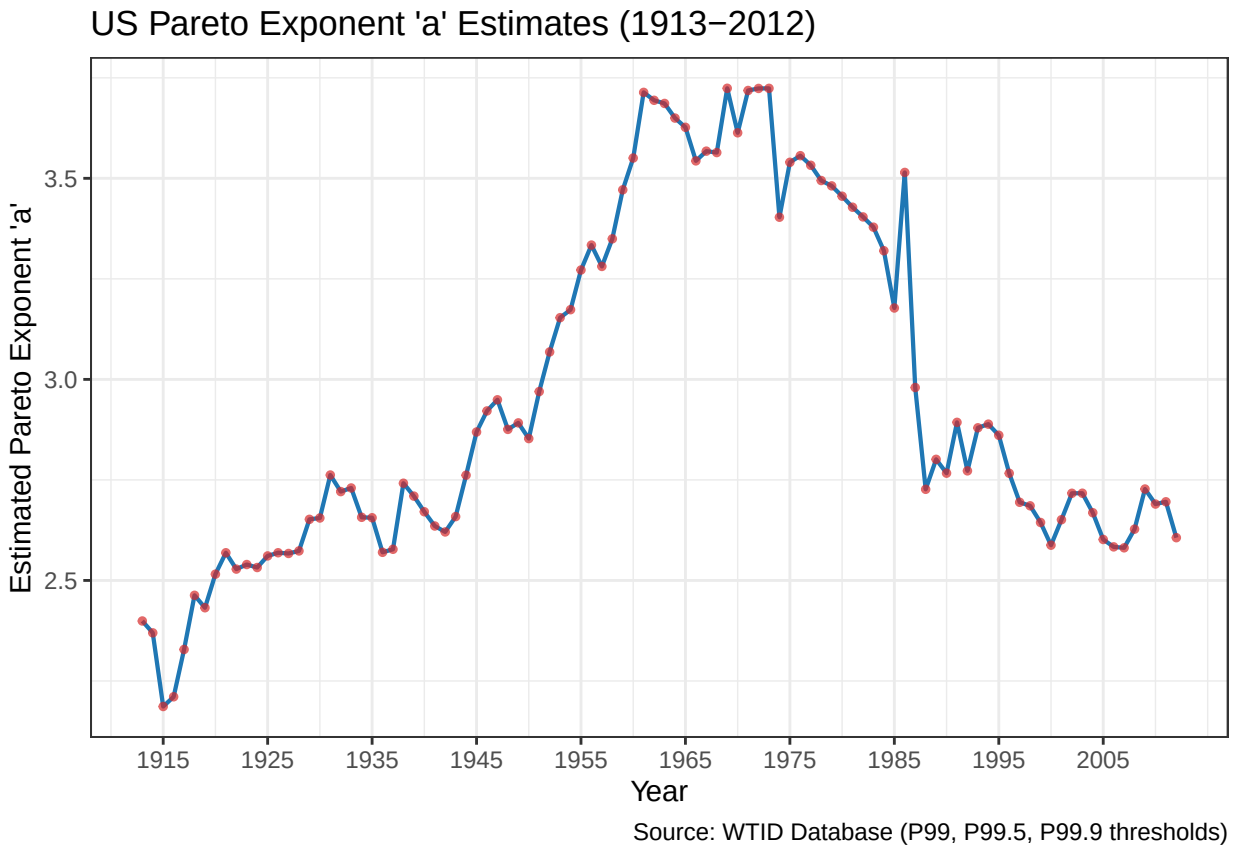
```
## [1] 2
```

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.

```
# Load Data and Required Libraries
library(tidyverse)
data <- read.csv("data/wtid-report.csv")
```

```
# Filter US Data (1913-2012) and Clean NA Values
us_data <- data |>
  filter(Country == "United States", Year >= 1913, Year <= 2012) |>
  select(Year,
         P99 = P99.income.threshold,
         P99.5 = P99.5.income.threshold,
         P99.9 = P99.9.income.threshold) |>
  na.omit()  # Remove rows with missing values

# Function to Estimate a for All Years
estimate_a_time_series <- function(data) {
  years <- data$Year
  a_estimates <- numeric(length(years))

  for (i in seq_along(years)) {
    a_estimates[i] <- exponent.multi_ratios_est(
      P99 = data$P99[i],
      P99.5 = data$P99.5[i],
      P99.9 = data$P99.9[i]
    )
  }

  return(data.frame(year = years, a = a_estimates))
}

# Calculate `a` for each year
results <- estimate_a_time_series(us_data)

# Plot the Results
ggplot(results, aes(x = year, y = a)) +
  geom_line(color = "#1F77B4", linewidth = 0.75) +
  geom_point(color = "#D62728", size = 1, alpha = 0.7) +
  labs(
    title = "US Pareto Exponent 'a' Estimates (1913-2012)",
```

```
    x = "Year",
    y = "Estimated Pareto Exponent 'a'",
    caption = "Source: WTID Database (P99, P99.5, P99.9 thresholds)"
  ) +
  theme_bw() +
  scale_x_continuous(breaks = seq(1915, 2010, by = 10))  # Improve x-axis labels
```

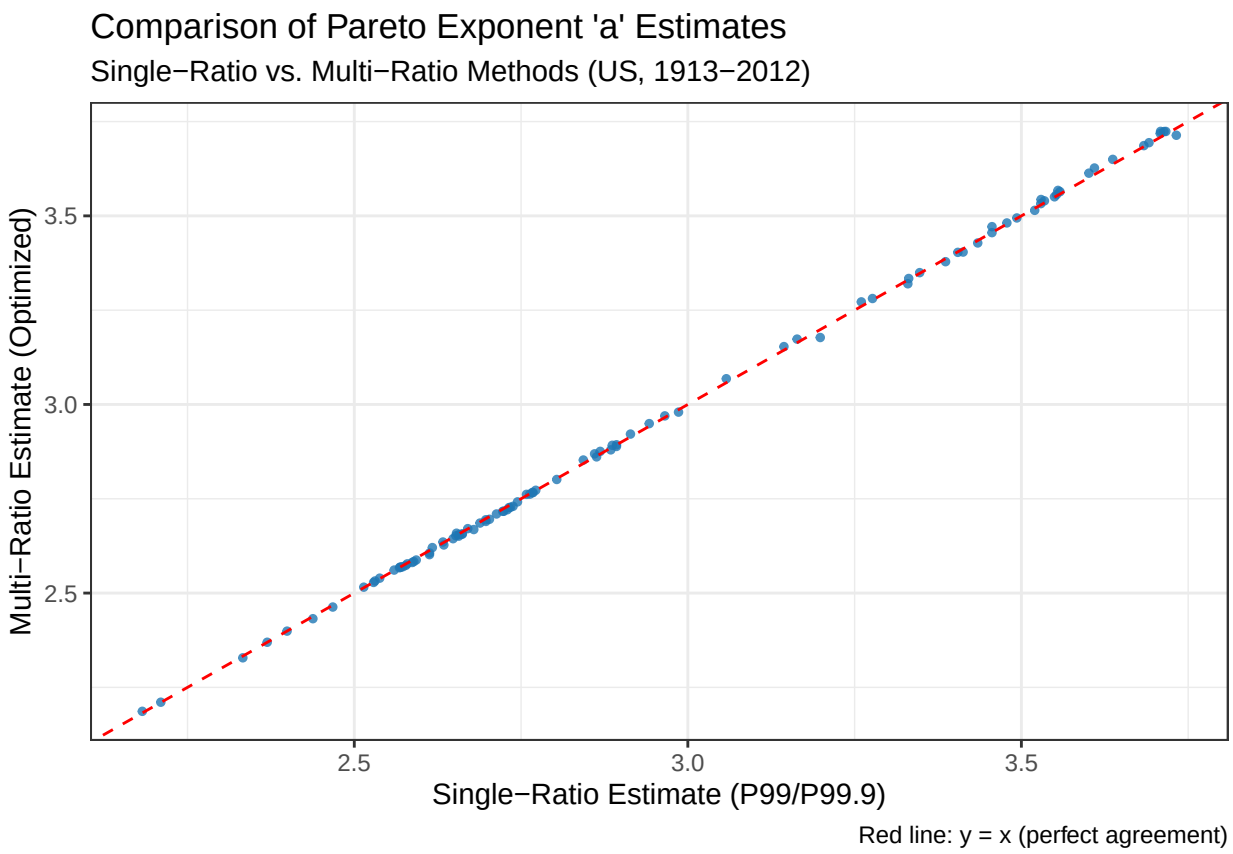


4. Use (4) to estimate a for the US for every year. Make a scatter-plot of these estimates against those from problem 3. If they are identical or completely independent, something is wrong with at least one part of your code. Otherwise, can you say anything about how the two estimates compare?

```
# Function to estimate `a` from P99/P99.9 ratio only
estimate_a_single_ratio <- function(P99, P99.9) {
  1 - log(10) / log(P99 / P99.9)
}

# Apply to each year
us_data <- us_data |>
  mutate(a_single = estimate_a_single_ratio(P99, P99.9))
```

```r
# Merge with Multi-Ratio Results (Problem 3)
combined_results <- results |> # From Problem 3
  left_join(us_data |> select(Year, a_single), by = c("year" = "Year"))
```

```r
# Scatter Plot Comparison
ggplot(combined_results, aes(x = a_single, y = a)) +
  geom_point(color = "#1F77B4", size = 1, alpha = 0.8) +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "red") +
  labs(
    title = "Comparison of Pareto Exponent 'a' Estimates",
    subtitle = "Single-Ratio vs. Multi-Ratio Methods (US, 1913-2012)",
    x = "Single-Ratio Estimate (P99/P99.9)",
    y = "Multi-Ratio Estimate (Optimized)",
    caption = "Red line: y = x (perfect agreement)"
  ) +
  theme_bw()
```




```r
# Quantitative Comparison: calculate correlation and mean absolute difference
correlation <- cor(combined_results$a_single, combined_results$a)
mean_diff <- mean(abs(combined_results$a_single - combined_results$a))
```

```r
cat(sprintf(
  "Correlation: %.3f\nMean Absolute Difference: %.3f",
  correlation, mean_diff
))
```

```
## Correlation: 1.000
## Mean Absolute Difference: 0.006
```

- **Conclusion:** The strong correlation between the two estimates indicates that the initial value of a (derived from the single-ratio method) is already very close to the optimal solution found through multi-ratio minimization. This suggests:
  - The single-ratio estimate provides an excellent starting point for optimization;
  - The additional constraints from multiple percentile ratios make only marginal adjustments;
  - The income distribution's tail behavior is remarkably consistent with Pareto distribution assumptions.
- This alignment demonstrates that while the multi-ratio method is theoretically more rigorous, the single-ratio method yields nearly identical results for this particular dataset, serving as both an efficient initial estimate and reliable standalone measure.