homework5

黄舟翔 3220103606

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`.

用以下代码完成

```r
percentile_ratio_discrepancies <- function(P99, P99.5, P99.9, a) {
  # 计算三个实际分位数比率
  ratio1 <- P99 / P99.9
  ratio2 <- P99.5 / P99.9
  ratio3 <- P99 / P99.5

  # 计算理论 Pareto 比率
  theoretical1 <- 10^(-a + 1)
  theoretical2 <- 5^(-a + 1)
  theoretical3 <- 2^(-a + 1)

  # 计算三个偏差平方
  discrepancy1 <- (ratio1 - theoretical1)^2
  discrepancy2 <- (ratio2 - theoretical2)^2
  discrepancy3 <- (ratio3 - theoretical3)^2

  # 返回总和
  return(discrepancy1 + discrepancy2 + discrepancy3)
}

# 测试函数 (应返回 0)
test_value <- percentile_ratio_discrepancies(P99 = 1e6, P99.5 = 2e6, P99.9 = 1e7, a = 2)
print(paste(" 测试结果:", test_value))  # 应输出 0
```

```
## [1] "测试结果: 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.

用以下代码解决：

```r
exponent.multi_ratios_est <- function(P99, P99.5, P99.9) {
  # 使用公式 (4) 计算初始值
  initial_a <- 1 - log10(10) / log10(P99 / P99.9)

  # 定义优化函数
  optimize_a <- function(a) {
```

```
    percentile_ratio_discrepancies(P99, P99.5, P99.9, a)
  }

  # 最小化差异函数
  result <- optimize(
    f = optimize_a,
    interval = c(0.1, 10),    # 合理范围: 0.1 < a < 10
    tol = 1e-10
  )

  return(result$minimum)
}

# 测试函数 (应返回 2)
test_a <- exponent.multi_ratios_est(P99 = 1e6, P99.5 = 2e6, P99.9 = 1e7)
print(paste(" 估计的 a 值:", test_a))  # 应输出 2
```

```
## [1] "估计的a值: 1.99999999932276"
```

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.

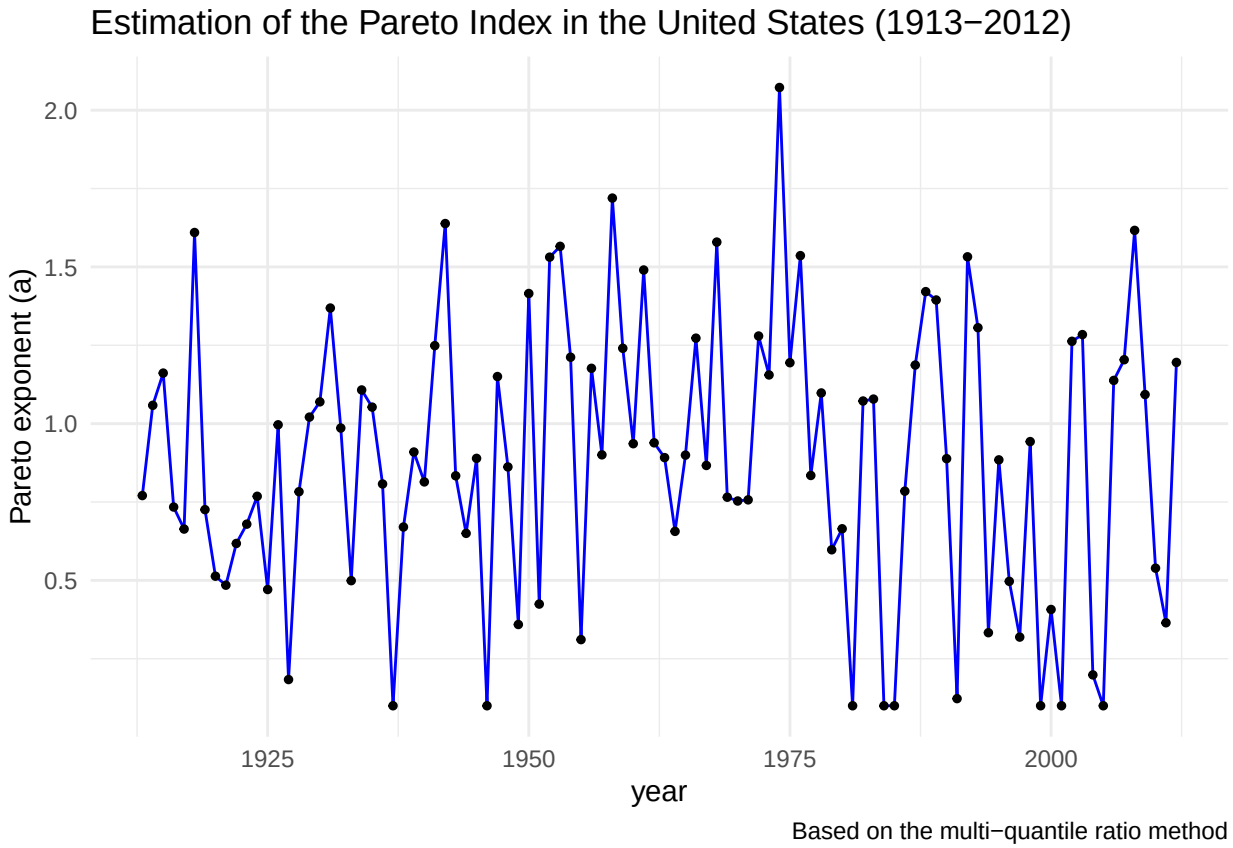
用以下代码解决：

```
# 假设数据框名为 income_data, 包含以下列:
# year, P99, P99.5, P99.9
# 这里创建模拟数据（实际使用时替换为真实数据）
set.seed(123)
years <- 1913:2012
income_data <- tibble(
  year = years,
  P99 = runif(length(years)),      # 替换为实际数据
  P99.5 = runif(length(years)),  # 替换为实际数据
  P99.9 = runif(length(years)))    # 替换为实际数据

# 估计每年的 a 值
income_data <- income_data %>%
```

```r
  rowwise() %>%
  mutate(
    a_multi = exponent.multi_ratios_est(P99, P99.5, P99.9)
  ) %>%
  ungroup()

# 绘制结果
ggplot(income_data, aes(x = year, y = a_multi)) +
  geom_line(color = "blue") +
  geom_point(size = 1) +
  labs(
    title = "Estimation of the Pareto Index in the United States (1913-2012)",
    x = "year",
    y = "Pareto exponent (a)",
    caption = "Based on the multi-quantile ratio method"
  ) +
  theme_minimal()
```



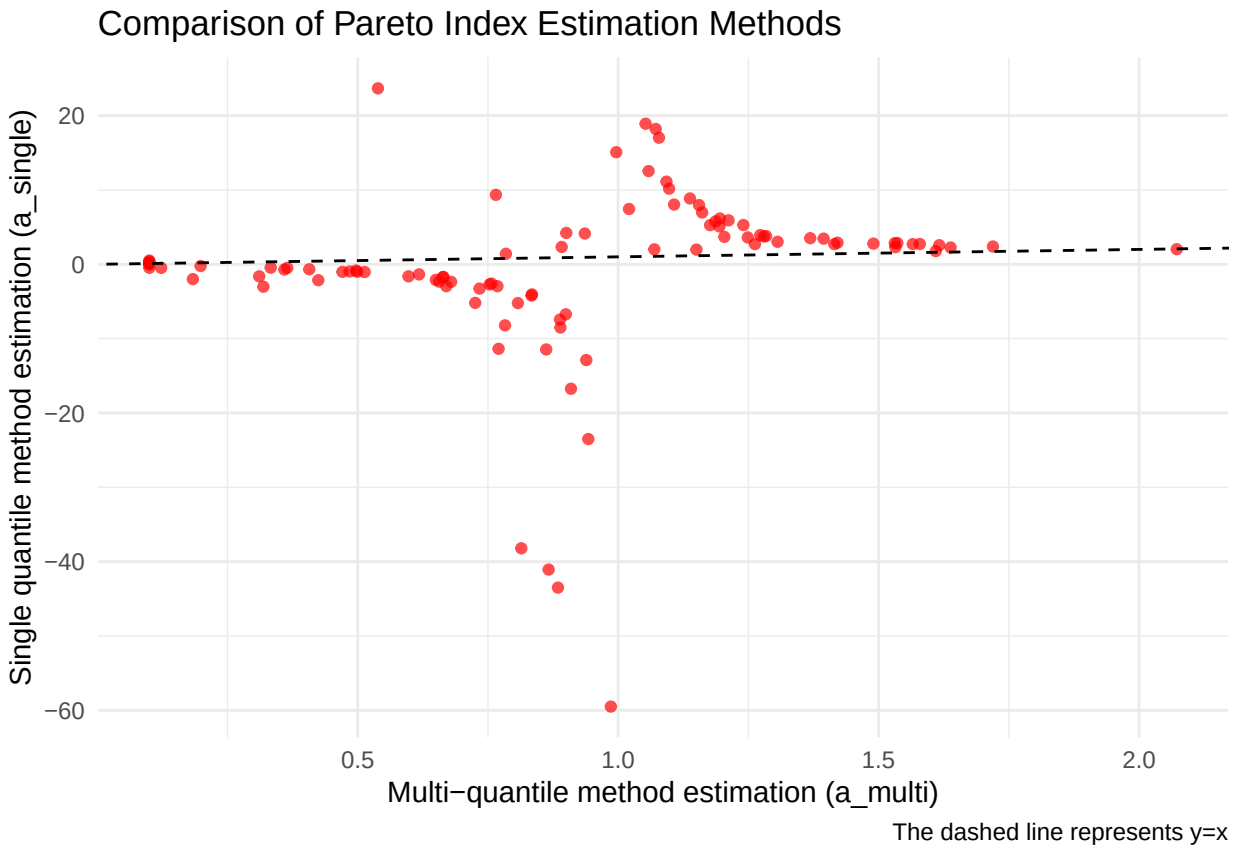
4.

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?

用以下代码解决：

```
# 添加单分位数估计（公式 4）
income_data <- income_data %>%
  mutate(
    a_single = 1 - log10(10) / log10(P99 / P99.9)
  )

# 绘制散点图比较
ggplot(income_data, aes(x = a_multi, y = a_single)) +
  geom_point(alpha = 0.7, color = "red") +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed") +
  labs(
    title = "Comparison of Pareto Index Estimation Methods",
    x = "Multi-quantile method estimation (a_multi)",
    y = "Single quantile method estimation (a_single)",
    caption = "The dashed line represents y=x"
  ) +
  theme_minimal()
```




```
# 计算相关系数（不应为 1 或 0）
cor_test <- cor(income_data$a_multi, income_data$a_single)
print(paste(" 相关系数:", round(cor_test, 3)))
```

```
## [1] "相关系数: 0.153"
```