Homework 5 - Report

3170105743 李政达

2020/7/9

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

$$\begin{aligned}\left(\frac{P99}{P99.9}\right)^{-a+1} &= 10 && (1)\\ \left(\frac{P99.5}{P99.9}\right)^{-a+1} &= 5 && (2)\\ \left(\frac{P99}{P99.5}\right)^{-a+1} &= 2 && (3)\end{aligned}$$

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

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

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(a, P99, P99.5, P99.9) {
  n1 <- ((P99/P99.9) ^ (1-a) - 10) ^ 2
  n2 <- ((P99.5/P99.9) ^ (1-a) - 5) ^ 2
  n3 <- ((P99/P99.5) ^ (1-a) - 2) ^ 2
  return(n1 + n2 + n3)
}
percentile_ratio_discrepancies(2, 1e6, 2e6, 1e7)
```

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

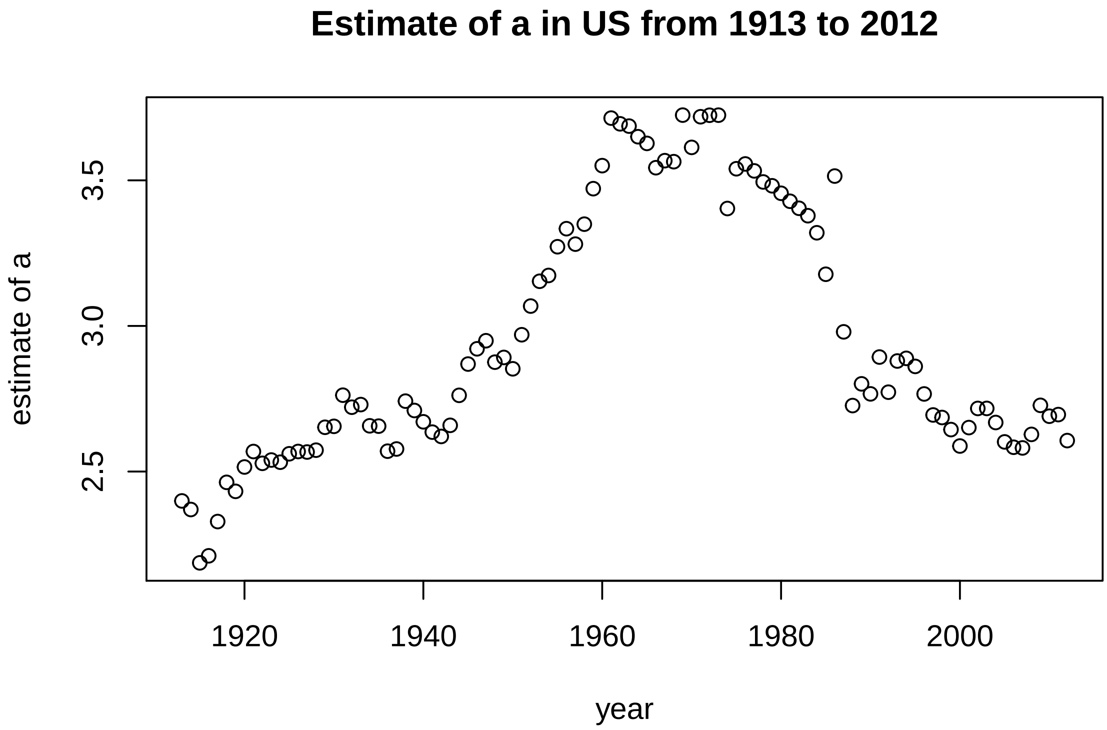
2. Write a function, `exponent.multi_ratios_est`, which takes as inputs `P99`, `P99.5`, `P99.9`, and estimates `a`. It should minimize your `percentile_ratio_discrepancies` function. The starting value for the minimization should come from (4). Check that when `P99=1e6`, `P99.5=2e6` and `P99.9=1e7`,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

```r
exponent.multi_ratios_est <- function(P99, P99.5, P99.9) {
  a <- 1 - log(10)/log(P99/P99.9)
  res <- nlm(percentile_ratio_discrepancies, a, P99, P99.5, P99.9)
  return(res$estimate)
}
exponent.multi_ratios_est(1e6, 2e6, 1e7)
```

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

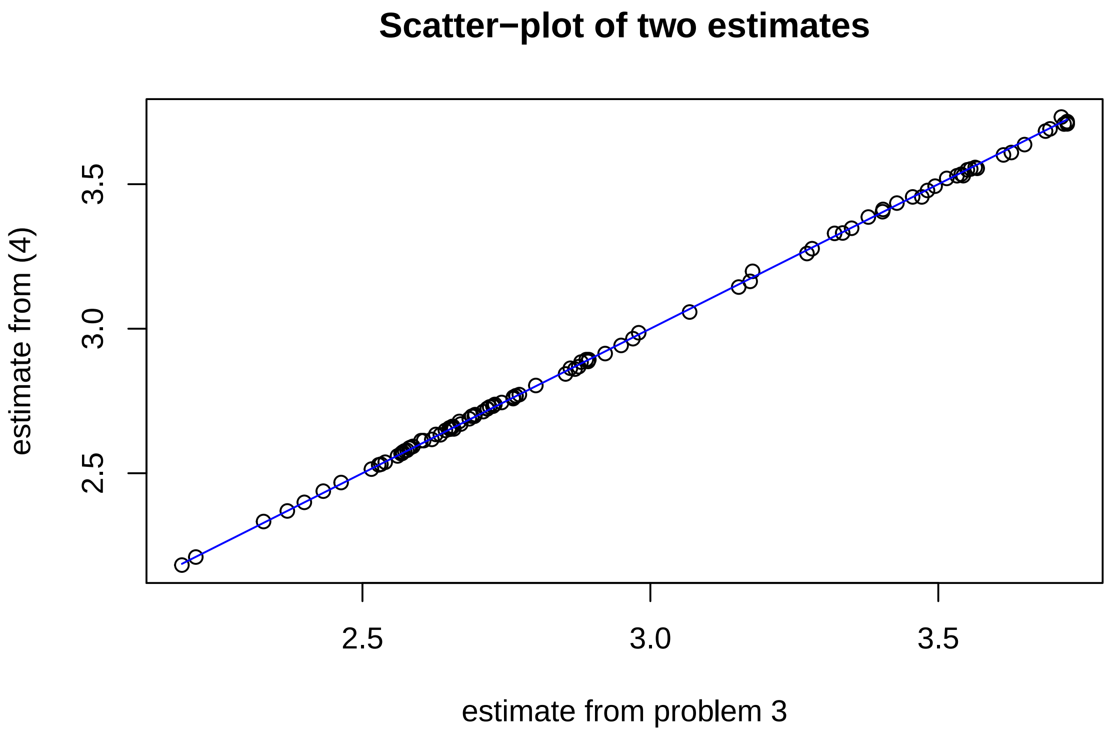
3. Write a function which uses `exponent.multi_ratios_est` to estimate *a* for the US for every year from 1913 to 2012. (There are many ways you could do thi, including loops.) Plot the estimates; make sure the labels of the plot are appropriate.

```
wtid <- read.csv("../data/wtid-report.csv")
wtid <- data.frame("year" = wtid$Year, "P99" = wtid$P99.income.threshold,
                   "p99.5" = wtid$P99.5.income.threshold,
                   "p99.9" = wtid$P99.9.income.threshold)
for (i in 1:dim(wtid)[1]) {
  wtid$a.est[i] <- exponent.multi_ratios_est(wtid$P99[i], wtid$p99.5[i],
                                             wtid$p99.9[i])
}
plot(wtid$a.est ~ wtid$year, xlab = "year", ylab = "estimate of a",
     main = "Estimate of a in US from 1913 to 2012")
```



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?

```
wtid$a.est2 <- 1 - log(10)/log(wtid$P99/wtid$p99.9)
plot(wtid$a.est2 ~ wtid$a.est,
     xlab = "estimate from problem 3", ylab = "estimate from (4)",
     main = "Scatter-plot of two estimates")
curve(x^1, add = TRUE, col = "blue")
```



From the scatter-plot we can find that the two estimates are not identical or completely independent. They distribute near the curve $y = x$, so we can conclude that they are similar but they are not identical.