Lab 7: Division and Power in the 1%

July 13, 2020

Agenda: Importing and manipulating data; writing functions to estimate parameters; writing functions to check model fit.

We continue to look at the very rich by turning to a more systematic data source than Forbes magazine, the World Top Incomes Database hosted by the Paris School of Economics [https://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.

$$f(x) = \frac{(a-1)}{x_{\min}} \left(\frac{x}{x_{\min}}\right)^{-a} \tag{1}$$

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 total income is concentrated among the very richest people. The proportion of people whose income is at least xmin whose income is also at or above any level $w \geq x_{\min}$ is thus

$$\Pr(x \geq w) = \int_w^\infty f(x)dx = \int_w^\infty \frac{(a-1)}{x_{\min}} \left(\frac{x}{x_{\min}}\right)^{-a} dx = \left(\frac{w}{x_{\min}}\right)^{-a+1} \tag{2}$$

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.) For this lab session, we have extracted the relevant data and saved it as `wtid-report.csv`.

Part I

1. Open the file and make a new variable containing only the year, `P99`,`P99.5` and `P99.9` variables; these are the income levels which put one at the 99th, 99.5th, and 99.9th, percentile of income. What was `P99` in 1972? `P99.5` in 1942? `P99.9` in 1922? You must identify these using your code rather than looking up the values manually. (You may want to modify the column names to make some of them shorter.)

```
setwd("D:/github/Rlab/")
wtid.report <- read.csv("data/wtid-report.csv")
wtid.report_select = wtid.report%>%select(Year,P99.income.threshold ,P99.5.income.threshold,P99.9.incom
print("P99 in 1972:")
```

```
## [1] "P99 in 1972:"
```

```
wtid.report_select$P99.income.threshold[which(wtid.report_select$Year==1972)]
```

```
## [1] 209076.6
```

```
print("P99.5 in 1942:")
```

```
## [1] "P99.5 in 1942:"
```

```
wtid.report_select$P99.5.income.threshold[which(wtid.report_select$Year==1942)]
```

```
## [1] 183217
```

```
print("P99.9 in 1922:")
```

```
## [1] "P99.9 in 1922:"
```

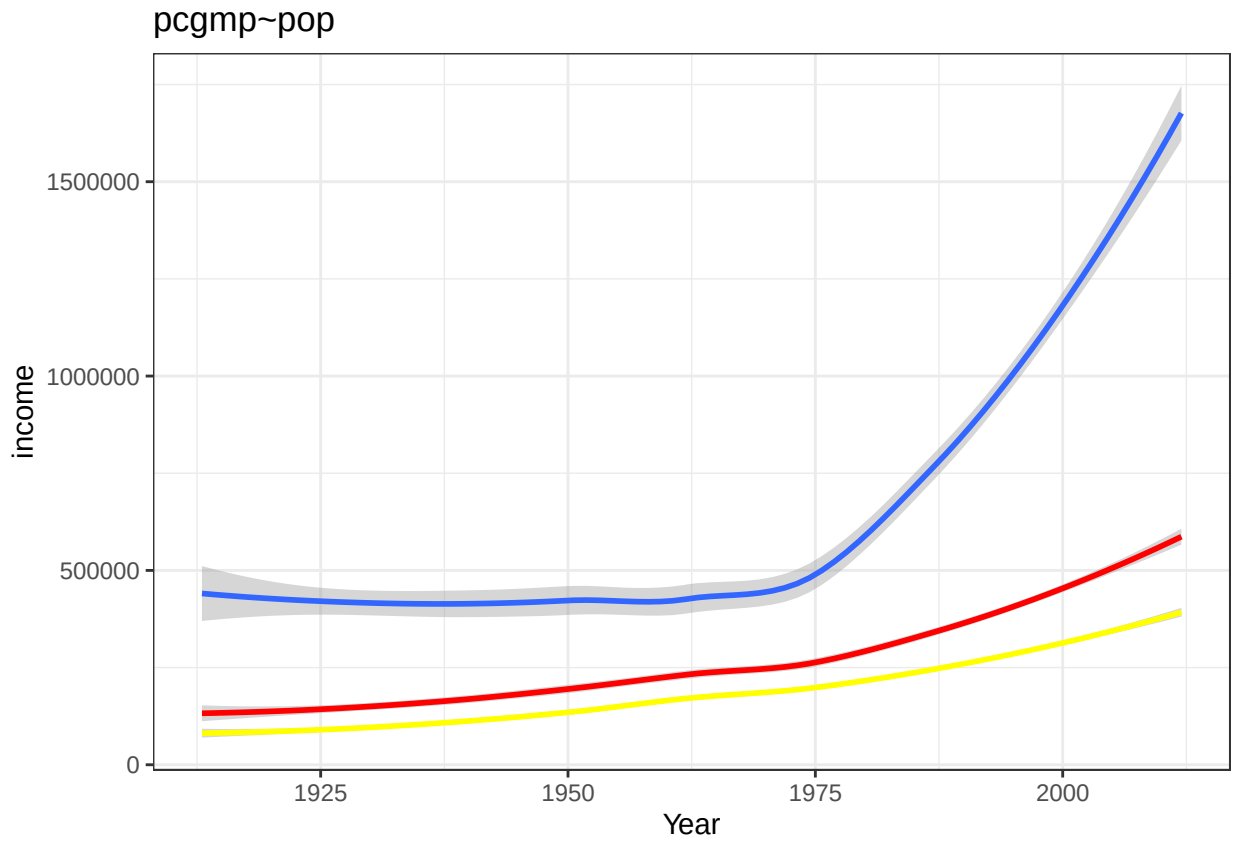
```
wtid.report_select$P99.9.income.threshold[which(wtid.report_select$Year==1922)]
```

```
## [1] 400214.2
```

2. Plot the three percentile levels against time. Make sure the axes are labeled appropriately, and in particular that the horizontal axis is labeled with years between 1913 and 2012, not just numbers from 1 to 100.

```
ggplot(data = wtid.report_select) +
  geom_smooth(aes(x = Year, y = P99.income.threshold),colour = 'yellow') +
  geom_smooth(aes(x = Year, y = P99.5.income.threshold),colour='red')+
  geom_smooth(aes(x = Year, y = P99.9.income.threshold))+
  labs(title = "pcgmp~pop",    y = "income",    x = "Year") +
  theme_bw()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



3. One can show from the earlier equations that one can estimate the exponent by the formula

$$a = 1 - \frac{\log 10}{\log P99/P99.9} \tag{3}$$

Write a function, `exponent.est_ratio()` which takes in values for `P99` and `P99.9`, and returns the value of a implied by ((3)). Check that if `P99=1e6` and `P99.9=1e7`, your function returns an a of 2.

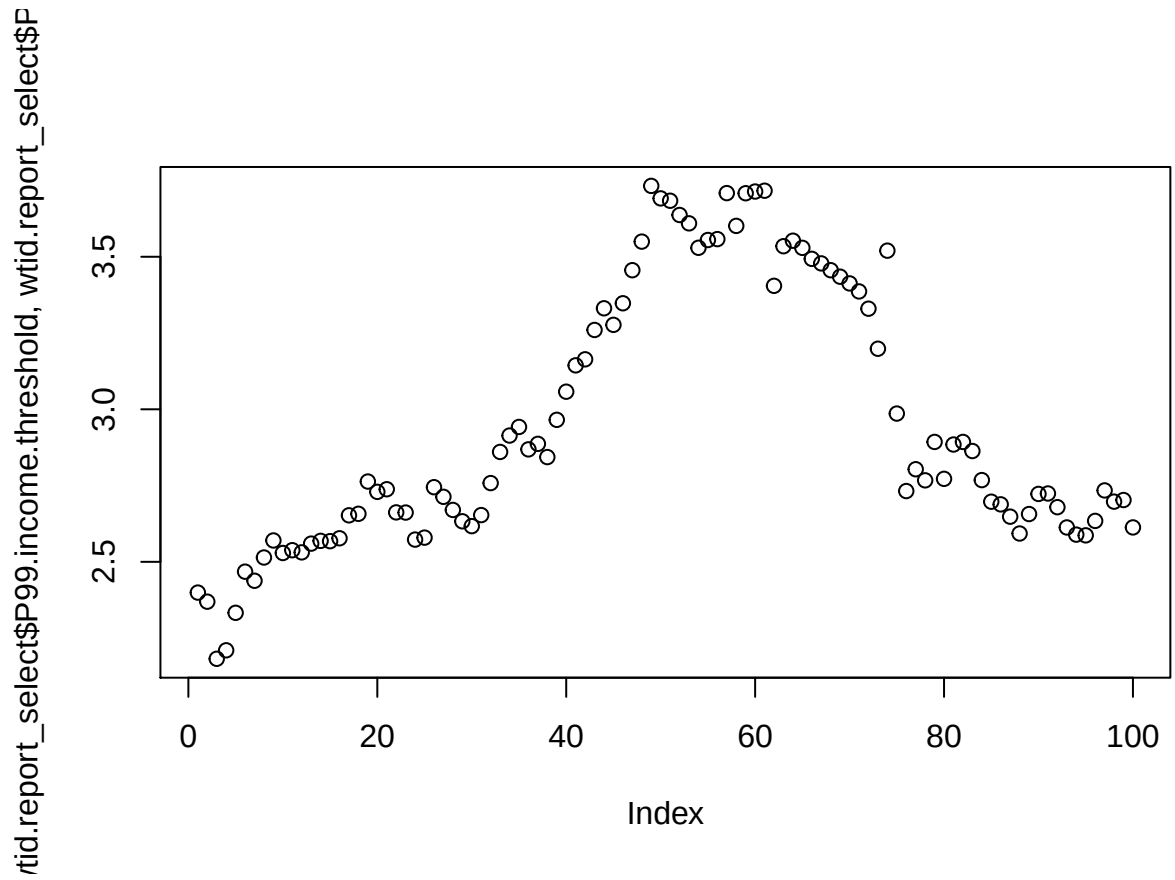
```
exponent.est_ratio<-function(p99,p99.9){
  a = 1 - log(10)/log(p99/p99.9)
  return(a)
}
exponent.est_ratio(1e6,1e7)
```

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

Part II

4. Estimate a for each year in the data set, using your `exponent.est_ratio()` function. If the function was written properly, you should not need to use a loop. Plot your estimate of a over time. Do the results look reasonable? (Remember that smaller exponents mean more income inequality.)

```
plot(exponent.est_ratio(wtid.report_select$P99.income.threshold,wtid.report_select$P99.9.income.thresho
```



It looks reasonable, because the ratio between P99 and P99.9 first getting bigger and then getting smaller from 2.

5. There were 160,681 households in the top 0.1% in the US in 2012. Using your estimated value of a for 2012, calculate approximately how many households had an income of over $50 million.

```
a_2012 = exponent.est_ratio(wtid.report_select$P99.income.threshold[which(wtid.report_select$Year==2012)
```

6. The logic leading to ((3)) 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 the data and your estimates of the exponent. Add a horizontal line with vertical coordinate 5. How good is the fit?

7. By parallel reasoning, we should have $(P90/P95)^{-a+1} = 2$. Repeat the previous step with this formula. How would you describe this fit compared to the previous ones? (Note: the formula in (3) is not the best way to estimate a, but it is one of the simplest.)