STAT GR5206 Homework 5 [100 pts]
Due 8:00pm Monday, October 31st on Canvas

Your homework should be submitted on Canvas as an `R` Markdown file. **Please submit the knitted .pdf file** along with the .Rmd file. We will not (and cannot) accept any other formats. Please clearly label the questions in your responses and support your answers by textual explanations and the code you use to produce the result. Note that you cannot answer the questions by observing the data in the "Environment" section of `R`Studio or in Excel – you must use coded commands. We will print out your homeworks. Please do not waste paper by printing the dataset or any vector over, say, length 20.

Goals: writing functions to automate repetitive tasks and using them as larger parts of code, some practice with ggplot, working with data frames and manipulating data from one form to another.

This homework uses the World Top Incomes Database and the Pareto distribution, as in this week's lab.

The following notes are a repeat from the lab assignment:

On the exam we looked at a dataset containing information on America's richest people. In this lab we continue to look at the very rich by turning to a more systematic data source than Forbes magazine, the World Top Incomes Database (WTID) hosted by the Paris School of Economics [`http://topincomes.g-mond.parisschoolofeconomics.eu`]. 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

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

*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's total income is concentrated among the very richest people.*

The proportion of people whose income is at least x_{min} whose income is also at or above any level $w \geq x_{min}$ is thus

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

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.) WTID exports its data sets as `.xlsx` *spreadsheets. For this lab session, we have extracted the relevant data and saved it as* `wtid-report.csv`*.*

We show 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; for example, in the lab we use

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

Because of measurement error and sampling noise, we can't 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.)

Part 1: Data for the US

i. 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.

ii. Write a function, `exponent.multi_ratios_est`, which takes as inputs the vectors `P99`, `P99.5`, `P99.9`, and estimates a. It should minimize your the function `percentile_ratio_discrepancies` you wrote above. The starting value for the minimization should come from (4). There are many ways to do the minimization, one is using `nlm()` as in HW4. Check that when `P99=1e6`, `P99.5=2e6` and `P99.9=1e7`, your function returns an a of 2.

iii. 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 this, including loops.) Plot the estimates using `ggplot`; make sure the labels of the plot are appropriate.

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

Part 2: Data for Other Countries

Go to the World Top Incomes Database and obtain data files with `P99`, `P99.5`, and `P99.9` for Canada, China, Colombia, Germany, Italy, Japan, South Africa, and Sweden. These variables can be found under `Income -> Average income -> Fractiles income levels` variable selection and they are labeled, for example, as `P99 income threshold`. (Note there are subdivisions `-LAD`, `-married couples & single adults`, `-adults` but we'd like the original value). The WTID website also has data on the average income per "tax unit" (roughly, household) for the US and the other countries. While you're at it, obtain this information from the website as well (for all the countries listed above and the US). This is under `Average fiscal income`

`per tax unit` variable selection. It may be easiest to grab all variables for all countries, including the US, even though you already have the US threshold data.

Note: WTID exports data files in xls format, which can't be imported into `R` in the usual ways. There are two strategies for getting xls data into `R`. (1) Open the files in Excel and then save them as csv files (go to File, Save As and choose csv), after which we can read them in to `R` in the usual ways. Otherwise you can use `read.xls()` from the `gdata` package which tries very hard to work like `read.csv()`.

v. Use your function from problem (iii) to estimate a over time for each of them. Note that the size of the dataset is different for each of these countries, and there may be some NA values.

   Hint: You may find it helpful to create a separate dataframe for each country, but make sure they all have the same column names. You could also keep all the countries in a single dataframe with a `Country` variable.

vi. Plot your estimates of a over time for all the countries using `ggplot`. Note that the years covered by the data are different for each country. You may either make multiple plots, or put all the series into one plot. Either way, make sure that the plots are clearly labeled. I did this by using the `color` aesthetic on the categorical variable `Country`.

vii. Plot the series of average income per "tax unit" for the US and the countries against time in `ggplot`. Hint: You may find it helpful for all this information to be in the same data frames.

viii. The most influential hypothesis about how inequality is linked to economic growth is the "U-curve" hypothesis proposed by the great economist Simon Kuznets in the 1950s. According tho this idea, inequality rises during the early, industrializing phases of economic growth, but then declines as growth continues.

Make a scatter-plot of your estimated exponents for the US against the average income for the US in `ggplot`. Qualitatively, can you say anything about the Kuznets curve? (Remember that smaller exponents indicate more income inequality.)

ix. For a more quantitative check on the Kuznets hypothesis, use `lm()` to regress your estimated exponents on the average income, including a quadratic term for income. Are the coefficients you get consistent with the hypothesis? Hint: the following will regress y on both x and x^2:

```
lm(y ~ x + I(x^2))
```

x. Do a separate quadratic regression for each country. Which ones have estimates compatible with the hypothesis? Hint: Write a function to fit the model to the data for an arbitrary country.

(If we were doing a more rigorous check of the Kuznet hypothesis, we would want to control for other factors, and not just assume that a quadratic was the right functional form for the curve.)

Please submit the knitted .pdf file!