Homework 5: Pareto and Kuznets on the Grand Tour

36-350

Due at 11:59 pm on Thursday, 2 October 2014

We continue working with the World Top Incomes Database [http://topincomes.g-mond.parisschoolofeconomics.eu], 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

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

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

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

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`.

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.

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.

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?

5. Go to the World Top Incomes Database and obtain data files with `P99`, `P99.5`, and `P99.9` for Canada, Colombia, Germany, Italy, Japan, South Africa, and Sweden. Use your function from problem 3 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.
   Note: WTID exports data files in `xls` format; you will have to either read them in, or convert them to `csv` and read them in.
   Hint: You may find it helpful to create a separate dataframe for each country, but make sure they all have the same column names.

6. Plot your estimates of a over time for all the countries. 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.

7. The WTID website also has data on the average income per "tax unit" (roughly, household) for the US and the countries in problem 4. Obtain this data, load it into R, and plot the series for each country.
   Hint: You may find it helpful to add this information as new columns to existing data frames.

8. 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. Qualitatively, can you say anything about the Kuznets curve? (Remember that smaller exponents indicate more income inequality.)

9. 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: `lm(y ~ x +I(x^2))` will regress y on both x and x^2.

10. 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. Take 36-401 and 36-402 to learn more.)