Homework 2: More Problems about Having More Money

36-313, Fall 2022

Due at 6 pm on Thursday, 15 September 2022

In the last homework, we looked at the "body" of the income distribution. In this homework, we'll shift our focus in two ways, by looking at *wealth* rather than *income*, and by focusing our attention on the "right tail", those who have the largest amounts of wealth.

As discussed in lecture, the body of the wealth and income distribution in market economies is approximately log-normal[1], but the right tail is "heavier" than a log-normal can explain. The first person to seriously investigate the shape of the right tail of wealth and income distributions was the economist Vilfredo Pareto[2]. Back in the 1890s, he noticed that these distributions approximately follow **power-laws**, with probability density

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

for incomes $x \geq x_{\min}$. (Typically, $x_{\min}$ is large enough that only the richest 3–4% of the population falls above it.) These distributions are now often called **Pareto distributions** in his honor. As the Pareto exponent α gets smaller, the distribution of income becomes *more* unequal, in the sense that more of the population's total income is concentrated among the population's richest members (the Gini index goes up as α goes down).

In this assignment, we'll work with data from the World Inequality Database [https://wid.world/]. This uses the tax systems of multiple countries to collect information on the distribution of wealth and income over multiple countries over time, with an emphasis on the rich. While there is an R package that gives access to the database [https://github.com/WIDworld/wid-r-tool], it's a little complicated to install and use, so refer to the data file on the class website, [http://www.stat.cmu.edu/~cshalizi/ineq/22/hw/02/uswealth.csv].

1. (5) *The complementary CDF* The cumulative distribution function we're used to, say $F(w)$, shows the probability that a random variable X is $\leq w$ for any given level w, $\mathbb{P}(X \leq w)$. There **complementary CDF**, or **CCDF**, or **survival function**, is $\mathbb{P}(X \geq w)$. Show that the CCDF of a Pareto distribution takes a particularly simple form:

$$\mathbb{P}(X \geq w | X \geq x_{\min}) = \left(\frac{w}{x_{\min}}\right)^{-\alpha+1} \tag{2}$$

   Hint: Integrate!

2. *Quantiles* Let's write w_p for the level of wealth so large that only p of the population has that much money, $\mathbb{P}(X \geq w_p) = p$.

   a. (5) Using Q1, show that

$$p = \mathbb{P}(X \geq x_{\min})(w_p/x_{\min})^{-\alpha+1} \tag{3}$$

[1] More exactly, the distribution of the part of the population with positive net worth is roughly log-normal. There are typically many people with a neworth which is zero or even negative. (This is a bit different from income, which is at least zero, and positive for almost everyone. [Or at least that's the way it is in reality, as opposed to the tax code.])

[2] If you've taken an economics course, you've probably heard "Pareto improvements" (a change that makes at least one person better off without harming anyone) and "Pareto optima" (a situation which can't be changed without making at least one person worse off, i.e., a situation where no change is a Pareto improvement); same Pareto.

b. (4) Using Q2a, show that

$$\frac{p}{q} = \left(\frac{w_p}{w_q}\right)^{-\alpha+1} \tag{4}$$

c. (4) Using Q2b, show that

$$\alpha = 1 - \frac{\log(p/q)}{\log(w_p/w_q)} \tag{5}$$

d. (5) Suppose we know the 99th and 99.9th percentile of the wealth distribution. Explain, using Q2c, how we can use these to estimate α.

3. *Examining the data*

a. (4) The P_10 variable shows the wealth of the 10th percentile of the population. This should be 0 for every year, if you've loaded the data properly. Explain what this means.

b. (4) Make a plot showing the wealth of the 25th, 50th and 75th percentiles over time. Describe this (roughly) in words.

c. (4) Make a plot showing the wealth of the 50th, 75th, 90th, 99th, 99.9th and 99.99th percentiles. Again, describe it, roughly, in words.

d. (4) Re-do the previous plot using a logarithmic scale for the vertical axis. Describe it in words.

e. (5) How does the plot you just made relate to the properties of the Pareto distribution you worked out in Q2?

4. *Looking at one year's distribution* In this question, we'll focus on the year 2012, just to keep things simple. In particular, we'll look at the complementary CDF for the wealth distribution in this year.

a. (4) Using the P_50 and P_75 variables, create a plot with two points, showing 0.5 and 0.25 on the vertical axis, and the values of the P_50 and P_75 variables on the horizontal axis.

b. (4) Using all the quantiles from P_50 upwards, create a plot of the complementary CDF. Describe its shape in words. *Hint*: Build on Q4a.

c. (4) Re-do the plot from the previous question, but use logarithmic scales on both the vertical and horizontal axes. Describe the shape in words.

5. *Connecting to Pareto*[3]

a. (4) What shape should the previous plot have, if the right tail of the wealth distribution follows a Pareto distribution? (Hint: Q1.)

b. (5) What α do you estimate for 2012, using Q2d?

c. (5) Using your estimated α, and the observed 99th percentile, you can calculate values for all the higher percentiles. Do so, and add them to the plot from Q4c. (Make sure they're distinct, visually, from the observed values.) Do these seem like decent matches? *Hint*: Q2b.

d. (5) Repeat Q5c for the lower percentiles (again taking the 99th percentile as given). About where do the Pareto model's predictions stop matching the data? Explain why this is a reasonable guess for $x_{\min}$.

6. *Pareto over time*

[3] The way this problem has you estimate α and $x_{\min}$ is rather crude, but it's simple and fast. A more accurate, and perhaps also more stable, estimation method for α would be to use a few different wealth ratios, from multiple percentiles, and try to find the single α which comes closest to fitting all of them at once. This sort of "nonlinear least squares" is a very old and common extension of the linear least squares we use to fit linear regression models, and you can do it in R with the **nlm()** function, but setting that up right involves more coding than I want you to do. If you have individual-level data, even more efficient estimation methods exist; see the paper by Clauset et al. in this week's optional reading.

a. (5) Using Q2d, estimate α for every year in the data set. (Make sure your code's estimate for 2012 matches what you got in Q5b.) Make a plot of estimated αs over time.

b. (4) In what periods was wealth become more concentrated / unequal, and in what periods was wealth become more equally distributed? Can you think of any events which might help explain these changes?

c. (5) Pareto thought his exponent was probably a constant everywhere and always, around 2.5, but he also knew he didn't have enough data to be very sure about that. Does your analysis in this assignment tend to support or undermine this conjecture of Pareto's? Explain.

7. *Timing on Pareto* (1) How long, roughly, did you spend on this assignment?

Presentation rubric (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision.

Extra credit (5): In Q4, you made a 2D plot of the wealth distribution in a single year. In this question, make a 3D plot which shows how the wealth distribution has changed over time. Use any software for 3D plotting you like, but make sure the result is easy to read.