Income Inequality and the contribution of Geography

Charles Liebenberg

30/08/2024

Purpose

Community level mental illness, violence, imprisonment, teenage births and drug abuse are conditions of a society that rhyme[1] with inequality (Wilkinson and Pickett, 2009). The form of inequality that is perhaps most on the public consciousness in 2024 is economic inequality. This paper aims to quantify economic inequality in taxpayers from Australia and measure the contribution of granular geography.

Background

Income inequality refers to the distribution and spread of income in a population. For this analysis, this refers to the concentration of income amongst taxpayers in Australia.

To model income inequality, distributional assumptions were required due to the absence of individual level data. There is a body of research that indicates log-normal and pareto distributions represent a good fit for income distributions given their right skewness (i.e. the long tail of large incomes) and their domain (Clementi, F, 2005).

Economists commonly theorise that the bottom 97-99% of a population's income distribution follows a log-normal distribution, with the remaining following a Pareto distribution (Clementi, F, 2005). Figure 1 demonstrates a log-log plot of US incomes from 2019 with a log-normal distribution fit to the data, where a good fit is evident for the middle incomes.

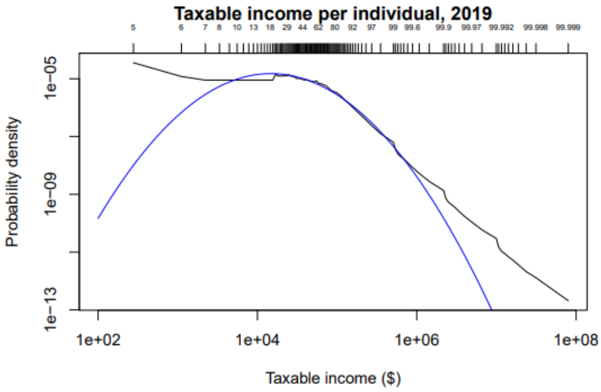

Figure 1: Log-normal log-log plot against US Taxable Income.

As shown, the fit seems a good match for the incomes visually over most of the distribution of income, with deviance coming from the tails. For this reason, this analysis will use a Pareto distribution for the modelling of incomes that implies the probability of a person earning above a certain threshold is proportional to some power of that threshold. At the right tail of the distribution (large incomes), the Pareto distribution is a much better fit than the log-normal distribution. This is as a result of the increased skewness of the Pareto distribution representing a better fit for the long tail of income.

Methodology

To model income distributions at a granular location level, public income and demographic data was sourced from the ATO and the ABS.

At a geography level, the data available for analysis amounted to summary statistics (excluding variance/standard deviation of income) and the Gini coefficient. This is insufficient for

[1] Wilkinson and Picket found that these factors are all negatively correlated to inequality.

maximum likelihood methods which require individual level data. Therefore, a modified method-of-moments approach was taken.

The Gini coefficient is independent of scale. Subsequently any distribution that can be expressed in terms of a shape and scale parameter can calculate the Gini coefficient explicitly as a function of shape. This analysis used this result as well as the expected value method of moments to estimate shape and scale for the Log-normal and Pareto distributions by SA2 geography (refer to Appendix 1 for calculation details).

After parameter estimation was performed at a granular geographical level, the distributional assumptions can be used to make inferences about income inequality. The first tool used for the analysis of inequality by location was the Lorenz Curve.

The Lorenz curve describes cumulative x against cumulative y, normalized to 100%. In the case of income distributions, it describes the cumulative share of income owned by the cumulative share of people.

If an income distribution is described by a statistical continuous distribution, then the Lorenz curve can be described by the cumulative distribution function. This results in defined functions for calculating the Lorenz curve for both log-normal and Pareto distributions (Irwin R., Hautus, M 2015). Appendix 2 describes the methodology for the calculation of Lorenz curves using the given distributional assumptions.

Finally, in order to determine the contribution of granular location to overall economic inequality, Lerman and Yitzhaki decomposition was used (Lerman, R., Yitzhaki S. 1985). This method asserts that the marginal impact of a given factor can be defined using the factor's own Gini coefficient and its share of income and population. Resultingly, an overall Gini can be decomposed to its "within-group" inequality, which determines a given locations specific inequality amongst all the people who live there, and a "between-group" inequality, which determines the inequality between difference locations and hence the contribution of granular geography. Appendix 3 describes the calculation method for this decomposition.

Results

It was modelled that SA-2 level geography contributed to 22.4% of overall income inequality (or 0.108 $G_{between}$ of an Australia-wide Gini coefficient of 0.483). It was found that inequality by area varies significantly, with the maximum G_{within} observed in the Sydney Eastern suburbs with 0.499 and the minimum observed in North Adelaide of 0.288. It is hypothesized that a large proportion of this area specific income inequality is driven by wealth inequality, which remains an area for investigation. This result represents significantly larger inequality than as measured by H. Miller and L. Dixie in their state-level analysis that found a 2% contribution at the state level. This implies that most geography-driven inequality exists at a more granular level than Australian state or territory.

Appendix 1: Parameter Estimation

Log-normal distribution:

$$Income \sim Lognormal(\mu, \sigma^2)$$

Estimating shape:

$$Gini\ Coefficient = \text{erf}\,(\frac{\sigma}{2})$$

$$erf^{-1}(Gini) = \frac{\sigma}{2}$$

$$\sigma = 2 * erf^{-1}(Gini)$$

Estimating scale:

$$E[X] = e^{\mu+\frac{1}{2}\sigma^2}$$

$$\ln\,(E[X]) = \mu + \frac{1}{2}\sigma^2$$

$$\mu = \ln(E[X]) - \frac{1}{2}\sigma^2$$

Pareto Distribution

$$Income\ {\sim}Pareto(\alpha, \beta)$$

Estimating shape:

if $\alpha < 1$ then distribution of income is perfect equality

If $\alpha \geq 1$:

$$Gini\ Coefficient = \frac{1}{2\alpha - 1}$$

$$2\alpha - 1 = \frac{1}{Gini}$$

$$\alpha = \frac{1}{2}(Gini + 1)$$

Estimating scale:

$$E[X] = \frac{\alpha\beta}{\alpha - 1}$$

$$\frac{E[X]}{\beta} = \frac{\alpha}{\alpha - 1}$$

$$\beta = \frac{E[X](\alpha - 1)}{\alpha}$$

Appendix 2: Estimation of Lorenz curve

Lorenz curve for the Pareto Distribution

For the Pareto distribution, it can be shown that[2]:

$$L(F) = 1 - F(x)^{1-\frac{1}{a}}$$

where F refers to the CDF of the Pareto distribution:

$$F(x) = 1 - (\frac{\beta}{x})^{\alpha}$$

To get the Lorenz curve as a function of x instead, the inverse of the CDF can be substituted:

$$L(x) = 1 - x(F)^{1-\frac{1}{a}}$$

where x(F) refers to the inverse of the CDF for the Pareto distribution:

$$x(F) = \frac{\beta}{(1 - F(x))^{\frac{1}{a}}}$$

Lorenz curve for the Log-normal distribution

It has been shown[2] that for a log-normal distribution:

$$L(x) = \varphi(\varphi^{-1}(x) - \sigma)$$

where φ and φ^{-1} refer to the CDF and inverse CDF for the standard normal distribution respectively.

Non-parametric model

Economic distributional data is commonly reported at an aggregated level. It is for this reason that researchers Sitthiyot T and Holasut K proposed a non-parametric model (i.e. no distributional assumption) for estimating a Lorenz curve that requires only three summary statistics: the Gini coefficient, the share of income that the top 10% holds, and the share of income that the bottom 10% holds. This works because the Gini coefficient has useful information about the middle of a distribution, and the bottom/top 10% income share ratio holds information about the tails of the distribution.

Specifically, the functional form for calculation of the Lorenz function assumed was a weighted average of the Pareto distribution and the exponential function:

$$y(x) = (1-k) * x^{P} + k * \left(1 - (1-x)^{\frac{1}{P}}\right)$$

The parameters k and P can both be expressed purely as a function of the Gini coefficient and the ratio of the income share of the top/bottom 10% of the income distribution. In addition, the research showed that its R^2 across a wide range of income distributions was the largest when compared to commonly used approximations for the Lorenz curve.[3]

[2] https://www.stat.cmu.edu/~cshalizi/ineq/21/lectures/03/lecture-03.pdf

[3] https://www.nature.com/articles/s41599-021-00948-x

Appendix 3: Decomposition of the Gini Coefficient

An overall Gini coefficient can be decomposed into within-group and between-group components, often referred to as the Lerman and Yitzhaki decomposition (1985).

Within-group Inequality

The within-group Gini coefficient is calculated by aggregating the Gini coefficients of each subgroup (here, geographical locations), weighted by their population share and mean income relative to the overall mean income.

$$G_{within} = \sum_i p_i \frac{u_i}{u} G_i$$

where:

G_i is the Gini coefficient of location i

u_i is the mean income of location i

p_i is the population share of location i

Between-group Inequality

The between-group Gini coefficient measures inequality due to differences in mean incomes between subgroups.

$$G_{between} = \frac{1}{2u} \sum_i \sum_j p_i p_j \left| u_i - u_j \right|$$

where:

p represents the population share of a location

u represents the mean income of a location

i and j represent each pair possible

Appendix 4: Data sources

This analysis uses publicly available ABS data for all calculations performed as below.

Historical data for income and wealth

https://www.abs.gov.au/statistics/economy/finance/household-income-and-wealth-australia/latest-release

Geographical data for income

https://www.abs.gov.au/statistics/labour/earnings-and-working-conditions/personal-income-australia/latest-release

References

1. **Clementi, F., Gallegati, M** (2005). Pareto's Law of Income Distribution: Evidence for Grermany, the United Kingdom, and the United States. *WPA Working Papers*. Available at: http://ideas.repec.org/p/wpa/wuwpmi/0505006.html (Accessed: 28 July 2024).

2. **Irwin R., Hautus, M** (2015). " Lognormal Lorenz and normal receiver operating characteristic curves as mirror images ", *Royal Society Open Science*, Available at: https://royalsocietypublishing.org/doi/10.1098/rsos.140280 (Accessed: 28 July 2024).

3. **Lerman, R., Yitzhaki S.** (1985). " Income Inequality Effects by Income Source: A New Approach and Applications to the United States", *The Review of Economic Studies*, 151-156 (6 pages). Available at: https://www.jstor.org/stable/1928447 (Accessed: 28 July 2024).

4. **Carnegie Mellon University** (2021). Modelling Income and Wealth Distributions. Available at: https://www.stat.cmu.edu/~cshalizi/ineq/21/lectures/03/lecture-03.pdf (Accessed: 28 July 2024).

5. **Sitthiyot, T** (2021). "**A simple method for estimating the Lorenz curve**", *Humanities and Social Sciences Communications*, Available at: https://www.nature.com/articles/s41599-021-00948-x (Accessed: 28 July 2024).

6. H. Miller and L Dixie (2023). " **Not a level playing field** ", Available at: https://www.actuaries.asn.au/Library/Miscellaneous/2023/230501NOTALEVEL.pdf (Accessed: 28 July 2024).