Estimating UK earnings distributions using composite Log-Normal-Generalized-Pareto-Distributions and coarse-grained fractile data

Lukas Kikuchi, Autonomy Data Unit

October 5, 2020

Abstract

We present a method by which cumulative distribution functions (CDFs) of UK earnings within industry sectors are estimated using composite Log-Normal and Generalized-Pareto-Distributions, and data from the Annual Survey of Hours and Earnings (ASHE) table 4.

1 Introduction

It is well known that wealth, income and earnings distributions take on remarkably different statistical character in their tail-ends. Whilst the majority of the distributions are well-approximated as Log-Normal or Gamma, at the tail-ends they fail to capture the large concentration of wealth and earnings in amongst the richest [4, 5, 6].

Various methods have been proposed to rectify this issue, from fitting Log-Normal-Pareto distributions to micro-data [1], improving such fits using rich-lists [9], and fitting a Generalized Pareto Curves directly on fine-grained fractile data [2]. The method we propose is inspired by the lack of neither freely available earnings microdata, nor the fine-grained fractile data used by Blanchet et al [2].

We use data from the Annual Survey of Hours and Earnings (ASHE) table 4 [8], which provides earnings data for each industry as classified under the 2007 UK Standard Industrial Classification (UK SIC). For each industry, the data provides 11 income fractiles $0 \leq p_1 < \cdots < p_{11} < 1$, and their corresponding quantiles $q_1 < \cdots < q_{11}$, as well as the mean earnings $\bar{x}$ in each industrial sector. As $p_{11} = 90\%$, we have that the largest fracticle is well below any threshold we would imagine a Pareto distribution to take over.

Although the data provided by ASHE is coarse, this is ameliorated by the fact that the sample size is comparatively large with respect to other similar

surveys. For example, the sample size of ASHE is approximately 1% of the population of the UK, whilst the Labour Force Survey's is 0.1% [7]. Due to this fact we assume, with some confidence [5], that the mean earnings $\bar{x}$ is close to its true value.

Given that we only have limited fractile data (that only covers what we would assume to be the non-Pareto part of the distribution), but with a mean earnings that we hold to reliable, our method is as follows:

1. Fit a Log-Normal distribution onto $\{p_i\}$ and $\{q_i\}$.
2. To capture the distribution at the tail-end, we construct a composite Log-Normal-GDP (Generalized-Pareto-Distribution). We use the fit from the first step to remove some free parameters, and we furter impose normality, continuity and differentiability to remove all but one of the free parameters θ.
3. Find the value of θ by imposing that the mean matches the mean reported in ASHE.

The next section will go through these steps in furthe detail.

2 Method

We assume that the distribution takes the form of a composite Log-Normal-GDP (Generalized Pareto Distribution):

$$f(x; r, s, \mu, \sigma^2, \xi, \tau, \theta) = r f_{\text{LN}}(x;\ \mu, \sigma^2) + s f_{\text{GDP}}(x;\ \xi, \tau, \theta) \quad (1) \tag{1}$$

where

$$\begin{aligned} f_{\text{LN}}(x;\ \mu, \sigma^2) &= \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log(x)-\mu}{\sigma}\right)^2} \mathbb{I}_{\{\eta \le x \le \theta\}} \\ f_{\text{GDP}}(x;\ \xi, \tau, \theta) &= \frac{1}{\tau}\left(1 + \xi\frac{x-\theta}{\tau}\right)^{-\left(\frac{1}{\xi}+1\right)} \mathbb{I}_{\{\theta < x\}}. \end{aligned} \tag{2}$$

The Log-Normal distribution is truncated from both ends, the truncation at the lower end is to take into account the fact that there is a minimum wage. η is the minimum annual earnings, and θ is the earnings beyond which the distribution will be described by the GPD.

Note that r and s are not independent of each other, as they must together be chosen so that (1) is normalized. The parameterisation was chosen so as to suggest that (1) is a linear combination of the Log-Normal and the GDP, but in reality this is not the case. Due to the specific order with which we fit the two parts of the distribution (the Log-Normal first, and then the GDP), we keep these two parameters separate for the sake of conveniance. Once we solve s for r, we could interpret r as a mixing weight [1].

The corresponding CDF is

$$F(x; r, s, \mu, \sigma^2, \xi, \tau, \theta) = \begin{cases} rF_{\text{LN}}(x;\ \mu, \sigma^2) & \eta \leq x \leq \theta \\ r(F_{\text{LN}}(\theta;\ \mu, \sigma^2) - F_{\text{LN}}(\eta;\ \mu, \sigma^2)) & \\ \quad + sF_{\text{GDP}}(x;\ \xi, \tau, \theta) & \theta < x \end{cases} \tag{3}$$

To satisfy normality, continuity and C^1-differentiability, we must have

$$\text{Normality: } s = 1 - r(F_{\text{LN}}(\theta) - F_{\text{LN}}(\eta)) \tag{4}$$

$$\text{Continuity: } \tau = \frac{s}{rf_{\text{LN}}(\theta)} (2) \tag{5}$$

$$\text{Differentiability: } \xi = e^{-\frac{(\mu - \log\theta)^2}{2\sigma^2}} \frac{r\tau^2(\log\theta + \sigma^2 - \mu)}{\sqrt{2\pi} s\theta^2\sigma^3} - 1 \tag{6}$$

Furthermore, we must have that the expectation is equal to the mean earning $\bar{x}$ reported in ASHE:

$$\int_0^\infty dx\ xf(x) = \bar{x} \tag{7}$$

Evaluating the LHS, we get

$$rG_{\text{LN}}(\theta) + s\left(\theta + \frac{\tau}{1-\xi}\right) = \bar{x} \tag{8}$$

where

$$G_{\text{LN}}(x) = \int_\eta^x dx'\ x' f_{\text{LN}}(x') = \left[-\frac{1}{2} e^{\mu + \sigma^2/2} \text{erf}\left(\frac{\mu + \sigma^2 - \log x}{\sqrt{2}\sigma}\right)\right]_\eta^x \tag{9}$$

and

$$\theta + \frac{\tau}{1-\xi} = \int_\theta^\infty dx'\ f_{\text{GPD}}(x') \tag{10}$$

We use equations (4-6) and (8) to constrain the parameter space to only three parameters (r, μ, σ). This means that we only to concern ourselves with the Log-Normal part of (1)

$$rf_{\text{LN}}(x;\ \mu, \sigma^2) \tag{11}$$

when we fit onto the fractile data, after which we compute the GDP part of the distribution by satisfying the constraint equation.

With the necessary mathematical groundwork completed, the fit is now straightforward to carry through:

1. Fit the Log-Normal part of the CDF $rF_{\text{LN}}(x;\ \mu, \sigma^2)$ onto the ASHE data. In other words minimise the residual

   $$|rF_{\text{LN}}(q_i;\ \mu, \sigma^2) - p_i|^2$$

   using standard `scipy` fitting algorithms, finding estimates (r^*, μ^*, σ^*). ASHE provides coefficients of variation (CV) for q_i. In other words, we

have errors only for the dependent as opposed to the independent variables. For this reason we eschew the standard least-squares fit in favour of Orthogonal Distance Regression [3], implemented in `scipy`.odr, which supports regression in the dependent variables.

2. Use the constraints (4-6) to compute the parameters (s^*, τ^*, ξ^*).

3. Find the θ^* such that the mean of the distribution reproduces the reported mean $\bar{x}$ in ASHE. In other words, solve for θ^* in (8).

3 Acknowledgements

Many thanks for the helpful advice, references and discussions with Rob Calvert Jump and Rafael Wildauer.

References

[1] Marco Bee. Estimation of the lognormal-pareto distribution using probability weighted moments and maximum likelihood. *Communications in Statistics-Simulation and Computation*, 44(8):2040–2060, 2015.

[2] Thomas Blanchet, Juliette Fournier, and Thomas Piketty. Generalized pareto curves: theory and applications. 2017.

[3] Philip J Brown, Wayne A Fuller, et al. *Statistical analysis of measurement error models and applications: Proceedings of the AMS-IMS-SIAM joint summer research conference held June 10-16, 1989, with support from the National Science Foundation and the US Army Research Office*, volume 112. American Mathematical Soc., 1990.

[4] Fabio Clementi and Michele Giammatteo. The pareto law and the distribution of labour income in italy. *Labour Economics: PLUS Empirical Studies*, pages 119–146, 2010.

[5] Paul Eckerstorfer, Johannes Halak, Jakob Kapeller, Bernhard Schütz, Florian Springholz, and Rafael Wildauer. Correcting for the missing rich: An application to wealth survey data. *Review of Income and Wealth*, 62(4):605–627, 2016.

[6] Stephen P Jenkins. Pareto models, top incomes and recent trends in UK income inequality. *Economica*, 84(334):261–289, 2017.

[7] Catrin Ormerod and Felix Ritchie. Linking ASHE and LFS: can the main earnings sources be reconciled? *Economic & Labour Market Review*, 1(3), 2007.

[8] Roger Smith. Earnings and hours worked, industry by two-digit SIC: ASHE table 4. 2019.

[9] Philip Vermeulen. How fat is the top tail of the wealth distribution? *Review of Income and Wealth*, 64(2):357–387, 2018.