# On Zipf's law and the bias of Zipf regressions

Christian Schluter

Empirical Economics
https://doi.org/10.1007/s00181-020-01879-3

# On Zipf's law and the bias of Zipf regressions

**Christian Schluter**[1,2]

Received: 6 July 2018 / Accepted: 30 April 2020


**Abstract**
City size distributions are not strictly Pareto, but upper tails are rather Pareto like (i.e. tails are regularly varying). We examine the properties of the tail exponent estimator obtained from ordinary least squares (OLS) rank size regressions (Zipf regressions for short), the most popular empirical strategy among urban economists. The estimator is then biased towards Zipf's law in the leading class of distributions. The Pareto quantile–quantile plot is shown to offer a simple diagnostic device to detect such distortions and should be used in conjunction with the regression residuals to select the anchor point of the OLS regression in a data-dependent manner. Applying these updated methods to some well-known data sets for the largest cities, Zipf's law is now rejected in several cases.

**Keywords** Rank size regression · Heavy tails · Extreme value index · Regular variation · Zipf's law · City size distributions

**JEL Classification** R12 · C13 · C14

## 1 Introduction

Zipf's law continues to fascinate economists. In urban economics, it concerns the largest city sizes and stipulates (in its strictest form) that the upper tail of the city size distribution not only decays like a power function, but also that the tail exponent equals unity. The most popular empirical strategy among urban economists is the estimation

**Electronic supplementary material** The online version of this article (https://doi.org/10.1007/s00181-020-01879-3) contains supplementary material, which is available to authorized users.

✉ Christian Schluter
christian.schluter@univ-amu.fr

1 Aix-Marseille Université, CNRS, EHESS, Centrale Marseille, Aix Marseille School of Economics (AMSE), 5 Boulevard Maurice Bourdet CS 50498, 13205 Marseille Cedex 01, France

2 Department of Economics, University of Southampton, Highfield, Southampton SO17 1BJ, UK

Published online: 06 June 2020

of the tail exponent by (variants of) an ordinary least squares (OLS) regression of log sizes on log ranks (a Zipf regression for short).[1] Since real-world (city) size distributions are not strictly Pareto but the upper tails are rather Pareto like (i.e. tails are regularly varying), such Zipf regressions suffer from asymptotic distortions. These distortions are rarely taken into account in applied work. In particular, it turns out that the Zipf regression estimator is biased *towards* Zipf's law in many situations, while the associated Pareto quantile–quantile (QQ) plot is concave like and becomes linear only eventually. This is of great practical relevance since practitioners usually select in a data-invariant manner the threshold point of the Zipf regression. This paper addresses these issues, by exploiting the relation between the Zipf regression and the Pareto QQ-plot, using methods that are new to urban economics.

To be more precise, consider the distribution function $F$ of positive independent and identically distributed city sizes that is regularly varying: for large $x$ and $\gamma \in (0, \infty)$

$$1 - F(x) = x^{-\frac{1}{\gamma}} l(x) \tag{1}$$

where $l$ is slowly varying at infinity.[2] Focussing instead on the largest city sizes, the tail quantile function $U(x) \equiv F^{-1}(1 - 1/x)$ gives an equivalent representation

$$U(x) = x^{\gamma} \tilde{l}(x)$$

where $F^{-1}$ denotes the generalised inverse and $\tilde{l}(x)$ is another slowly varying function. The parameter $\gamma$, usually referred to as extreme value index (and $1/\gamma$ as the tail exponent), is unknown and needs to be estimated.

In particular, $\gamma$ is the slope coefficient in the Pareto QQ-plot that Zipf regressions seek to estimate. To see this and the ensuing problems, reconsider the tail quantile function. As $x \to \infty$, $\log U(x) \sim \gamma \log(x)$. Replacing these population quantities with their empirical counterparts gives the Pareto QQ-plot. It follows that $\gamma$ is the *ultimate* slope of this plot. If the distribution were strictly Pareto,[3] this plot would be linear throughout. However, if the tail of the distribution varies regularly, the Pareto QQ-plot will become linear only *eventually*. In Appendix A.3.1 we show that, using the tail quantile function, the Pareto QQ-plot has a tendency to exhibit a concave-like curvature for leading parametric models. A slow decay in the nuisance functions $l(x)$ and $\tilde{l}(x)$ will then induce asymptotic distortions in the estimator of the slope coefficient in the Zipf regression. Below, this slow decay will be modelled formally by higher-order regular variation and quantified. In particular, building on asymptotic expansions

[1] For instance, the meta study of Nitsch (2005) cites 29 papers providing 515 estimates based on Zipf regressions, and a recent update in Cottineau (2016) 81 papers of which 23 are in core journals of regional science providing 1702 estimates. In comparative cross-country work, Soo (2005), for instance, runs rank size regressions for data on 75 countries, which are revisited in, for example, Nishiyama et al. (2008). We reconsider some of these in Sect. 3.

[2] That is, $l(tx)/l(x) = 1$ as $x \to \infty$. Recall that a (positive measurable) function $g$ is called regularly varying at infinity with index $\theta \in \mathbb{R}$ if $\lim_{x\to\infty} g(tx)/g(x) = t^{\theta}$ with $t > 0$. If $\theta = 0$, the function is said to be slowly varying. See, for example, Bingham et al. (1987).

[3] Strict Paretoness, $1 - F(x) = cx^{-1/\gamma}$, is usually assumed in formal statistical analyses in economics, see, for example, in Nishiyama et al. 2008 p. 696 equation (3), or Gabaix and Ibragimov 2011, p. 25 equation (2.1).

developed in Schluter (2018), we show that the OLS estimator *over*-estimates $\gamma$ in the leading class of distributions in which the nuisance function $l$ in model (1) converges to a constant at a polynomial rate. In this case Zipf regressions are biased *towards* Zipf's law. The Pareto QQ-plot therefore offers a simple diagnostic device to detect the presence of such distortions as it conveys important information about the behaviour of the Zipf regression estimator.

It is then shown how the threshold parameter (i.e. the $k$th upper-order statistic) for this Pareto QQ-plot and the OLS regression can now be selected in a data-dependent manner, using regression diagnostics based on the residuals of the OLS regression. The problem in common practice is that practitioners tend to select mechanically the number of observations to be included in the Zipf regression. As Gabaix and Ioannides (2004) observe "optimum cutoff techniques have not (..) been used in the context of the city size distribution". This choice determines the threshold, beyond which linearity is implicitly assumed. Such "blind" choice (i.e. without visual reference to the Pareto QQ-plot) then risks to fall within the curved, usually concave, part of the Pareto QQ-plot, thus distorting the estimator. For instance, it is common practice to select the top 1% of city sizes in complete census for all cities, or to consider only cities above 100,000 inhabitants (see, for example, Nitsch 2005, p. 95, or Giesen and Südekum 2011, p. 671, and reference therein), or using all observations in left-truncated data sets for the largest cities. The latter case is illustrated in Sect. 3, by revisiting the data and Zipf regressions reported in Soo (2005) and Nishiyama et al. (2008). When these proposed updated methods are applied to these well-known data sets for the largest cities, we detect some substantial differences to the results reported in the literature. Zipf's law (in the strictest sense with $\gamma = 1$) is now rejected in some of these cases and confirmed in others.

The empirical importance of this threshold selection in the presence of a Pareto QQ-plot that exhibit curvature is illustrated in Fig. 1 for administrative data for cities in Germany in the year 2000, using up to the largest 5000 cities. Panel (a) depicts the Pareto QQ-plot, and panel (b) plots the Zipf regression estimates $\hat{\gamma} = \hat{\gamma}(k)$ as a function of the $k$ upper-order statistics. The Pareto QQ-plot clearly depicts a concave-like curvature in the lower left part of the plot, which then leads to an *over-estimate* of $\gamma$. The larger $k$, the larger is the resulting distorted estimate $\hat{\gamma}(k)$. This curvature then explains the unexplained observation in, for example, Nitsch (2005, p. 94) or Gabaix and Ioannides (2004) that a larger number of observations tends to increase the estimate $\hat{\gamma}$ (i.e. in their notation reduce the estimate $1/\hat{\gamma}$).[4] In Appendix A.3.1, the curvature is examined parametrically using the tail quantile function. Below, we quantify these distortions and propose a method for choosing $k$ optimally.

This paper therefore makes a substantive contribution to the extensive literature on the city size distribution, surveyed in, for example, Gabaix and Ioannides (2004) and the meta-studies based on Zipf regressions (Nitsch 2005; Cottineau 2016) already mentioned. A recent applied literature extends this scope and estimates Zipf regressions for country size distributions (see, for example, Rose 2006) and considers the world city size distribution (see, for example, Luckstead and Devadoss 2014). Clarity

[4] For instance, Gonzales-Val (2010, Table 2) uses Zipf regressions to produce 165 estimates for US data, letting the truncation point for the year 2000 distribution range from the largest 100 to the largest 19,200 cities yielding point estimates that range from $\hat{\gamma}(100) = .76$ to $\hat{\gamma}(19,200) = 1.86$.

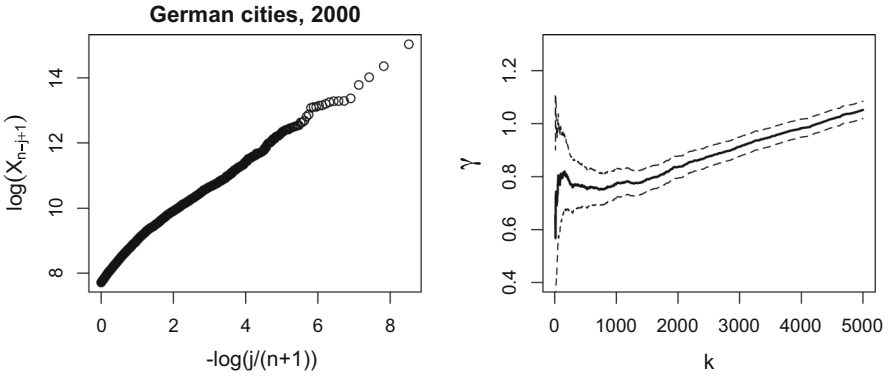



**Fig. 1** German cities: Pareto QQ-plot and the Zipf regression estimates $\hat{\gamma}(k)$. German cities in the year 2000. The data are described in Sect. 3.1. **a** Pareto QQ-plots using the 5000 largest cities. **b** Estimates $\hat{\gamma} = \hat{\gamma}(k)$ as a function of the $k$ upper-order statistics used in the Zipf regression (solid line) and associated pointwise 95% symmetric confidence intervals (dashed line). For the Zipf regression, see Eq. (3), and for the distributional theory, see Sect. 2.2

about the speed of tail decay for the largest cities is important. Firstly, the largest cities contain most of the population. For instance, using a cut-off of 100,000 people in the often used 2000 US census place data captures 63% of the population and 1% of places. 15% of all places contain 80% of the population. Secondly, the speed of tail decay informs about the underlying theoretical generative growth processes. For instance, Gibrat's classic model of i.i.d. proportional growth leads to a lognormal size distribution, while adding a lower reflecting barrier to geometric Brownian motion leads to a Pareto size distribution with unity exponent (used in Gabaix 1999b), and subordinating geometric Brownian motion can lead to the so-called double-Pareto-lognormal distribution (Reed 2002). See also Perline (2005). Debates about the speed of tail decay are ongoing and extend beyond urban economics into diverse fields in economics and the natural sciences, see, for example, Gabaix (2009) and Schluter and Trede (2019) for recent discussions.[5] In particular, Schluter and Trede (2019) propose a unifying statistical framework based on the classic Fisher–Tippett theorem and allied concepts of maximum domains of attraction. This reasoning gives rise to encompassing tests of whether the tail of the size distribution decays *faster* than any power function, i.e. tests of the so-called Gumbel–Gibrat hypothesis $\gamma = 0$ (which includes the case of the lognormal distribution). In the empirical applications to firm and city size data, the hypothesis that $\gamma$ be zero is robustly and clearly rejected in favour of $\gamma > 0$, the setting of model (1) and thus justifying the use of Zipf regressions.

In order to illustrate the debates and the problems of interpretation, Eeckhout (2004), for instance, using US Census Bureau data, states that "cities grow proportionately" and "it is shown that the size distribution of the entire sample is lognormal and not Pareto". However, using the same data, Levy (2009) observes that "[for the largest

[5] In the literature on exchange rates, finance, insurance and risk management [where often $1/\gamma \in (2, 4)$], Ibragimov et al. (2013, 2015) emphasise that heavy tails may lead to sub-optimal diversification in the value-at-risk framework, non-robustness of several economic and financial models, while finiteness of variances is crucial for the applicability of classical econometric approaches.

cities] the size distribution diverges dramatically and systematically from the lognormal distribution, and instead is much better described by a power law". This latter observation is reiterated in, for example, Ioannides and Skouras (2013) based on different methods (which is revisited below). While the literature beginning with the influential contribution of Eeckhout (2004) has the merit of considering the entire city size distribution, Schluter and Trede (2019) clarify that the analysis of the *largest* city sizes requires appropriate statistical techniques based on extreme value theory,[6] and that this task is distinct from fitting the main body of the size distribution. Moreover, the asymptotic distortions caused by the slowly varying nuisance function $l$ in model (1) render problematic fully parametric attempts in the applied literature that seek to test lognormality against strict Paretoness (see, for example, Malevergne et al. 2011, for a statistically sophisticated maximum-likelihood-based approach to discriminating between the tails of the two distributions).

A very recent literature in regional science seeks to combine the two distributional perspectives by smoothly pasting a strict Pareto tail to the main body of a lognormal size distribution. For instance, Ioannides and Skouras (2013) propose a maximum likelihood approach to estimate jointly the switching point and the distributional parameters. Fazio and Modica (2015) compare several other approaches to identifying the smoothly pasted switching point (and assess their performance in a simulation study when the data generating process is exactly Pareto-lognormal). These recent approaches address the question of how data from the entire city size distribution could be used. However, given the assumption of strict Paretoness in the upper tail, this approach inherits the asymptotic distortion discussed above caused by the confounding presence of the slowly varying function $l$ in model (1). This observation is numerically illustrated in Appendix A.3.3. The semi-parametric model (1) has the merit of avoiding the problems of fully specified distributions while imposing informative restrictions on the data of city sizes. Furthermore, the threshold points of the Zipf regression and the Pareto QQ-plot are determined below in a data-dependent manner.

The paper is organised as follows. In the next section we introduce the concept of higher-order regular variation that enables us to be precise about the decay of the nuisance function $l$ in model (1). We then recall the Pareto QQ-plot, relate it to the Zipf regression, recall the asymptotic theory for the OLS estimate of $\gamma$ and characterise the asymptotic distortions. In Sect. 2.4, we consider the choice of threshold. We illustrate the methods in several applications in Sect. 3. When these methods are applied to some well-known data sets for the largest cities, we detect some substantial differences to the results reported in the literature. Zipf's law is now rejected in some of these cases and confirmed in others.

[6] The empirical problem posed by lognormality is its *subexponentiality*: Although the speed of tail decay of the lognormal distribution is faster than that of power function class (1), it is slower than exponential and sufficiently slow to generate a tail that is commonly considered as "heavy", i.e. for both distributional classes we have $e^{\beta x}(1 - F(x)) = \infty$ for all $\beta > 0$ as $x \to \infty$. Some authors refer to the lognormal distribution as *rapidly varying*, since its index is $-\infty$ and $lim_{x\to\infty}[1 - F(tx)]/[1 - F(x)]$ equals $\infty$ if $0 < t < 1$ and equals 0 if $t > 1$, reserving the term regular variation to finite and nonzero indices (e.g. Bingham et al. 1987, definition 2.4.2). As discussed above, Schluter and Trede (2019) show that the appropriate test is that of the Gumbel–Gibrat hypothesis.

## 2 The Pareto QQ-plot and the rank size regression

### 2.1 Preliminaries: higher-order regular variation

The distributional theory for the Zipf regression estimator exploits modelling the slowly varying nuisance function $l$ in (1) as higher-order variation. Recalling the preceding discussion of the tail quantile function, it is immediate that model (1) has the equivalent (first-order regular variation) representation $\lim_{t\to\infty}[\log U(tx) - \log U(t)]/[a(t)/U(t)] = \log x$ for all $x > 0$ where $a$ is a positive norming function with the property $a(t)/U(t) \to \gamma$ (see, for example, Dekkers et al. 1989). The problem for estimating the extreme value index $\gamma$ is the behaviour of the slowly varying function $l$ in (1). It is therefore common practice in the extreme value literature to model such second-order behaviour by strengthening the first-order regular representation to second-order regular variation. Following de Haan and Stadtmüller (1996), we assume

$$\lim_{t\to\infty} \frac{\frac{\log U(tx)-\log U(t)}{a(t)/U(t)} - \log x}{A(t)} = H_{\gamma,\rho}(x) \tag{2}$$

for all $x > 0$, where $H_{\gamma>0,\rho<0}(x) = \frac{1}{\rho}(\frac{x^\rho-1}{\rho} - \log x)$ with $\rho < 0$. This parameter $\rho$ is the so-called second-order parameter of regular variation, and $A(t)$ is a rate function that is regularly varying with index $\rho$, with $A(t) \to 0$ as $t \to \infty$. As $\rho$ falls in magnitude, the nuisance part of $l$ in (1) decays more slowly. Most heavy-tailed distributions of interest satisfy representation (2). The Hall class of distributions (Hall 1982), which includes, for instance, the Burr, Student $t$, Fréchet, and Cauchy distributions, is but one example and considered explicitly in Appendix A.3, which illustrates the role of $\rho$, the concavity of the Pareto QQ-plot, and the induced substantial distortions of statistical inference.

### 2.2 The rank size regression estimator

We briefly recall the Pareto QQ-plot and the associated Zipf regression that yields an estimator of the tail index $\gamma$. Details are collected in Appendix A.1. Variants of this Zipf regression are discussed in Sect. 2.3.

The key insight is obtained from the tail quantile function: As $x \to \infty$, $\log U(x) \sim \gamma \log(x)$ in model (1). Replacing these population quantities with their empirical counterparts gives the Pareto QQ-plot whose *ultimate* slope is $\gamma$. To this end, let $X_{1,n} \leq \cdots \leq X_{n,n}$ denote the order statistics of $X_1, \ldots, X_n$, and consider the $k$ upper-order statistics. The Pareto QQ-plot becomes *ultimately* linear for a sufficiently high threshold $X_{n-k,n}$ where $k < n$. In Sect. 2.4, we consider how this threshold, which is usually ignored by practitioners in regional science, can be selected in a data-dependent manner.

The estimator of the slope coefficient in the Pareto QQ-plot is obtained by minimising with respect to $\gamma$ the least squares criterion of the Zipf regression of sizes on

ranks,[7]

$$\hat{\gamma} = \arg\min \sum_{j=1}^{k} \left( \log \frac{X_{n-j+1,n}}{X_{n-k,n}} - \gamma \log \frac{k+1}{j} \right)^2 \tag{3}$$

with $1 \leq j \leq k < n$. Schluter (2018) demonstrates that under assumption (2), as $k \to \infty$ and $k/n \to 0$, this estimator is weakly consistent, and if $\sqrt{k}A(n/k) \to 0$

$$\sqrt{k}(\hat{\gamma} - \gamma) \to^d N\left(0, \frac{5}{4}\gamma^2\right). \tag{4}$$

Asymptotically, the estimator is thus unbiased if $\sqrt{k}A(n/k) \to 0$. But if this decay is slow, the estimator will suffer from a higher-order distortion in finite samples given by

$$b_{k,n} \equiv \frac{1}{2}\frac{\gamma}{\rho}\frac{2-\rho}{(1-\rho)^2}A(n/k) \qquad (\gamma > 0, \rho < 0) \tag{5}$$

For instance, in the Hall class (see Appendix A.3 for details), the tail quantile function is $U(x) = cx^{\gamma}[1 + dx^{\rho} + o(x^{\rho})]$ so that $A(t) = (\rho^2/\gamma)dt^{\rho}$. The sign of the bias is therefore given by $-\text{sign}(d)$, and one can show that $d < 0$ for the nested Burr, Student $t$, Fréchet, and Cauchy distributions. It follows that $b_{k,n} > 0$, so $\gamma$ is *over-estimated*, and Zipf regressions are thus biased *towards* Zipf's law in models in which the nuisance function $l$ in model (1) converges to a constant at a polynomial rate. The empirical evidence presented in Sect. 3 is in line with this theory.

### 2.3 OLS regression variants in the literature

The literature contains several variants of regression (3). Usually, practitioners include the additional estimation of a regression constant: $\log X_{n-j+1,n}$ is regressed on a constant and $\log j$. Schultze and Steinebach (1996) prove weak consistency of the estimator in this setting. Kratz and Resnick (1996) also prove weak consistency, obtain the distributional theory for this alternative estimator, and show that its asymptotic variance is $2\gamma^2/k$, which exceeds the asymptotic variance of $\hat{\gamma}$ given in (4). Hence, this regression variant is less efficient (given the additional estimation of the regression constant) and the estimate exhibits excessive variability (which can be an issue for hypothesis testing, such as Zipf's law). Similar comments apply to the so-called dual regressions in which ranks are regressed on sizes (Nitsch 2005, refers to the two regressions types as the Lotka and Pareto forms). Shifting ranks, as examined formally in Gabaix and Ibragimov (2011) in the strict Pareto model, does not eliminate the

[7] Since the OLS estimator of the slope coefficient is not invariant to shifts in the data, it is conceivable that a purposefully chosen shift could yield an asymptotic refinement (Gabaix and Ibragimov 2011, demonstrate the optimality of a shift of ranks by 1/2 in the strict Pareto model, show the asymptotic normality of the estimator, and obtain the correct standard errors). Since it turns out that the higher-order distortion in model (1) remains intact, we ignore such a shift for the sake of notational simplicity.

asymptotic distortion in model (1) (Schluter 2018). Finally, we observe that some practitioners augment the OLS regression with a squared regressor in order to control directly the curvature of the QQ-plot (rather than selecting $k$). However, since the distributional theory for this augmented regression is currently unknown (not even in the strict Pareto model), statistical inference is not possible in this setting (Nishiyama et al. 2008 p. 703, make a similar observation).[8] Since Pareto-like tails lead to curved Pareto QQ-plots when the nuisance function $l$ in model (1) decays slowly (as illustrated in Fig. 6a), it is also not clear how significance tests for the squared regressor should be interpreted.

Many other estimators of $\gamma$ have been proposed in the statistical extreme value literature (see, for example, the textbook treatments in Embrechts et al. 1997, or Beirlant et al. 2004). The Hill estimator has received most attention, and its asymptotic normality has been studied in various settings (e.g. Hall 1982; Csörgő et al. 1985, or Haeusler and Teugels 1985). In particular, using a second-order condition similar to (2), de Haan and Peng (1998) show that if $\lim_{n\to\infty} \sqrt{k}A(n/k) = \lambda$, then $\sqrt{k}(\hat{\gamma}^{(\text{Hill})} - \gamma)$ follows asymptotically a normal law with mean $\lambda/(1-\rho)$ and variance $\gamma^2$.[9] We observe that the variance of the Hill estimator for a given $k$ is thus smaller than the variance of any of the rank size OLS estimators. However, the Hill estimator also suffers from asymptotic distortions, and requires, as the OLS estimator, the selection of the threshold level $k$. This problem is considered next.

### 2.4 The choice of the threshold *k*

The OLS regression (3) provides further diagnostics that can be used to select optimally the threshold level $k$ in a data-dependent manner. Specifically, the residuals enable us to estimate nonparametrically the asymptotic mean-squared error (AMSE), which, in view of the bias–variance trade-off implied by (4) and (5), is commonly used in the statistical literature as a selection criterion (e.g. Csörgő et al. 1985; Hall 1990, or Beirlant et al. 1996).

Following Beirlant et al. (1996), we observe that the expectation of the mean weighted theoretical squared deviation

$$\frac{1}{k}\sum_{j=1}^{k} w_{j,k} E\left(\log\left(\frac{X_{n-j+1,n}}{X_{n-k,n}}\right) - \gamma \log\left(\frac{k+1}{j}\right)\right)^2 \tag{6}$$

[8] In the Burr model, this augmented regression increases the positive distortions of the point estimates of $\gamma$ for high $k$. For instance, in a Monte Carlo with 1000 replications in the Burr model with $\gamma = 2/3$, sample sizes $n = 1000$ and $k = 500$, the mean point estimates of $\gamma$ were 1.59 when $\rho = -.5$, 1.17 when $\rho = -.75$, and 0.998 when $\rho = -1$. Evaluated at $k = n$, the distortions are even higher, exceeding the population value at least by a factor of 8; for instance, when $\rho = -.5$, the mean estimate is 10.89.

[9] In particular, de Haan and Peng (1998) obtain the asymptotic expansion for the Hill estimator $\hat{\gamma}^{(\text{Hill})} =^d \gamma + \frac{\gamma}{\sqrt{k}} Z_k + \frac{1}{1-\rho} A(n/k)(1 + o_p(1))$ where $Z_k = \sqrt{(k)}(k^{-1}\sum_{i=1}^{k} E_i - 1)$ and $E_i$ are i.i.d. standard exponential random variables. However, since their second-order assumption differs from (2), their bias expression is not directly comparable to (5).

**Table 1** Performance evidence for optimal $k$ selection: Burr distribution

| $\gamma$ | $\rho$ | $n$ | $\bar{k^*}$ | $k^*_{\text{Burr}}$ | $\hat{\gamma}(k^*)$ | $\hat{\gamma}(k^*) - b^{\text{Burr}}_{k^*,n}$ | $\hat{\gamma}(k^*) - \tilde{b}_{k^*,n}$ |
|---|---|---|---|---|---|---|---|
| 2/3 | −0.5 | 10,000 | 340 | 201 | 0.732 | 0.664 | 0.688 |
| 2/3 | −0.5 | 1000 | 118 | 64 | 0.805 | 0.678 | 0.741 |
| 2/3 | −0.75 | 10,000 | 604 | 444 | 0.698 | 0.661 | 0.664 |
| 2/3 | −0.75 | 1000 | 173 | 112 | 0.743 | 0.663 | 0.692 |
| 2/3 | −1 | 10,000 | 1114 | 765 | 0.692 | 0.664 | 0.664 |
| 2/3 | −1 | 1000 | 226 | 164 | 0.719 | 0.662 | 0.671 |

$\bar{k^*}$ is the mean value across 1000 Monte Carlo repetitions computed using the procedure described in Sect. 2.4. $k^*_{\text{Burr}}$ minimises the parametric AMSE in the Burr model given by $\text{Var}(\hat{\gamma}) + [b^{\text{Burr}}_{k,n}]^2$. $\tilde{b}_{k^*,n}$ is defined in Eq. (8) and we have set $\rho = -5$

equals, to first order,

$$c_k \text{Var}(\hat{\gamma}) + d_k(\rho) b^2_{k,n} \tag{7}$$

for some coefficients $c_k$ depending only on $k$, and $d_k(\rho)$ depending on $k$ and $\rho$ (see Appendix A.2 for details). The procedure then consists in applying two different weighting schemes $w^{(i)}_{j,k}$ $(i = 1, 2)$ in (6), estimating the corresponding two mean weighted theoretical deviations using the residuals of regression (3), and computing a linear combination thereof such that

$$\text{Var}(\hat{\gamma}) + b^2_{k,n}$$

obtains. We carry out this programme for weights $w^{(1)}_{j,k} \equiv 1$ and $w^{(2)}_{j,k} = j/(k+1)$ for a set of preselected values of $\rho$.[10]

Table 1 reports some performance evidence for this AMSE-based selection procedure in the Burr model parametrised as $1 - F_{(\gamma,\rho)}(x) = (1 + x^{-\rho/\gamma})^{1/\rho}$ with $\gamma = 2/3$, $\rho \in \{-0.5, -0.75, -1\}$, and $n \in \{1000, 10{,}000\}$. Appendix A.3 provides additional details for this model (e.g. the role of $\rho$ and the curvature of the Pareto QQ-plot). The higher-order distortion (5) becomes

$$b^{\text{Burr}}_{k,n} = \frac{1}{2}\gamma \frac{2-\rho}{(1-\rho)^2} \left(\frac{n}{k}\right)^{\rho} > 0.$$

Figure 2 illustrates further one such experiment. In panel (a) the theoretical AMSE, $\text{Var}(\hat{\gamma}) + [b^{\text{Burr}}_{k,n}]^2$, is plotted as well as a boxplot for the optimally selected $k^*$ in all 1000 Monte Carlo simulations. In panel (b) we examine one such random sample for which the selection procedure yielded $k^* = 126$ and depict the Pareto QQ-plot as well as the Zipf regression line with anchor $X_{n-k^*,n}$. In the table we report the mean

[10] Such preselection and the comparison of the resulting optimal $k^* = k^*(\rho)$ avoid the need to estimate $\rho$. It is well known in the extreme value literature that reliable estimation of second-order objects $\rho$ and $A$ is notoriously difficult.

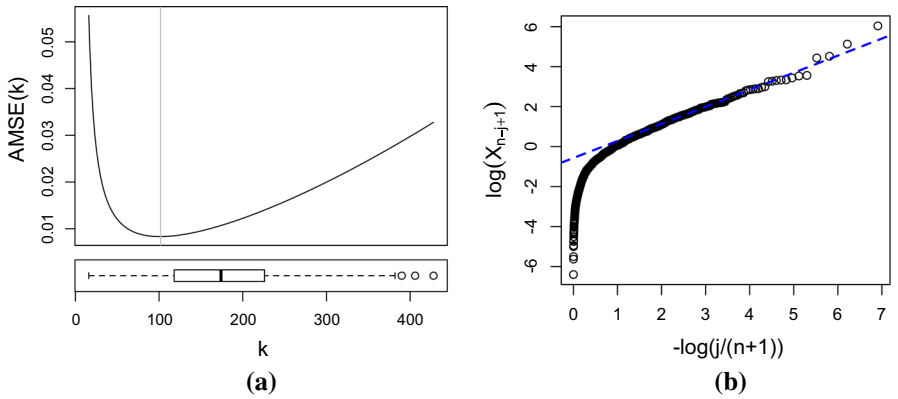



**Fig. 2** AMSE in the Burr model and selection of $k$. Burr model with $\gamma = 2/3$ and $\rho = -0.75$ and sample(s) of size $n = 1000$. **a** Parametric AMSE in the Burr model given by $\mathrm{Var}(\hat{\gamma}) + [b_{k,n}^{\mathrm{Burr}}]^2$. The theoretical $k^*$ is 112, depicted by the faint vertical line. The lower part of the figure shows the boxplot for the realised $k^*$ across all simulations for 1000 Monte Carlo repetitions. **b** For one random sample, Pareto QQ-plot, and Zipf regression line (dashed line) with slope $\hat{\gamma}(k^*) = .85$ and threshold $X_{n-k^*,n}$ where the selection procedure yielded $k^* = 126$

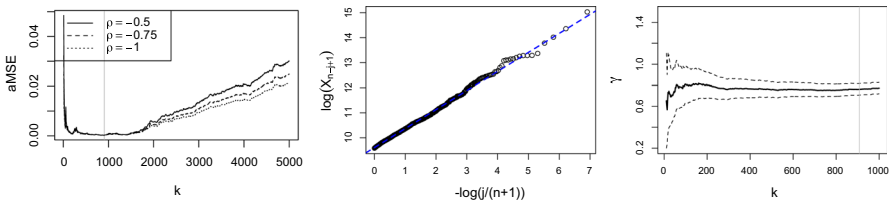
value $\bar{k}^*$. This mean has the correct order of magnitude. The tendency to exceed the theoretical optimal value $k^*_{\mathrm{Burr}}$ is explained by the asymmetry of the theoretical AMSE plot illustrated in the figure (which varies across the experiments since the squared bias increases at speed $k^{-\rho}$ whereas the variance does not depend on $\rho$). We also verify that the theoretical bias in the Burr model is a good guide for the actual distortions, by bias-correcting the estimate $\hat{\gamma}(k^*)$. The table shows that across all experiments the bias corrected estimate $\hat{\gamma}(k^*) - b_{k^*,n}^{\mathrm{Burr}}$ is very close to the population value $2/3$.

### 2.5 Bias correction and lower bounds analysis

By trading off asymptotic bias and variance, the resulting optimal estimate $\hat{\gamma}(k^*)$ still exhibits a bias. A simple pragmatic procedure is based on (6) with $w_{j,k} \equiv 1$, and yields a lower bound for $\gamma$ as follows. An estimate of the mean theoretical deviation is the mean of the squared residuals $k^{-1}\mathrm{SSR}_k$ of the rank size regression (3). All the measured deviation $k^{-1}\mathrm{SSR}_k$ is then ascribed to the bias,

$$\tilde{b}_{k,n}(\rho) = [k^{-1}\mathrm{SSR}_k / d_k(\rho)]^{1/2} \tag{8}$$

thereby defining a conservative bound $\hat{\gamma} - \tilde{b}_{k,n}(\rho)$. The sensitivity analysis then consists of examining this expression for a range of values of $\rho$. Table 1 reports the results of this exercise for the Burr case, setting $\rho = -.5$ as a conservative value, allowing, by Fig. 6a, for curvature in the Pareto QQ-plot. It turns out that the resulting estimates are very close to the population value of $\gamma$, improving on the estimate $\hat{\gamma}(k^*)$.




**(a)** **(b)** **(c)**

**Fig. 3** German cities: Pareto QQ-plot and the Zipf regression estimates $\hat{\gamma}(k)$. German cities in the year 2000. **a** Plot of the estimated AMSE as a function of $k$ for selected $\rho$. The minimiser is $k^* = 908$. **b** Pareto QQ-plots using the 1000 largest cities, and Zipf regression line with slope $\hat{\gamma}(k^*) = .761$ and threshold $X_{n-k^*,n}$. **c** Estimates $\hat{\gamma}(k)$ as a function of the $k$ upper-order statistics used in the Zipf regression (solid line) and associated pointwise 95% symmetric confidence intervals (dashed line), based on the distributional theory given in Eq. (4). The grey vertical line indicates $k^*$

## 3 Applications

We illustrate the methods in several applications to the upper tail of the size distribution of cities, focussing on the diagnostic Pareto QQ-plot, the positive distortions of the OLS estimator, and the selection of $k$.

### 3.1 The size distribution of cities in Germany

Our first empirical application concerns the size distribution of cities in Germany. We use first an administrative dataset for Germany for the year 2000, provided by the German Federal Statistical Office. These administrative data are highly accurate due to the legal obligation of citizens to register with the authorities. The unit of analysis is the "city", or more precisely the municipality or settlement ("Gemeinden"). Population sizes are as of December 31, and the year 2000 size distribution comprises 13,854 cities. Figure 3 depicts the results. In panel (a), we plot the estimated AMSE for several values of $\rho$. The minimisers closely agree, the estimated AMSE being minimised at $k^* = 908$. In panels (b) and (c) we revisit Fig. 1, now restricting the plots to the 1000 largest cities. In panel (b) we redraw the Pareto QQ-plot, as well as the regression line with slope $\hat{\gamma}(k^*) = .761$ and threshold $X_{n-k^*,n}$. In panel (c), we draw again the estimates $\hat{\gamma}(k)$ as a function of the $k$, as well as the pointwise 95% symmetric confidence intervals. The vertical line at $k^* = 908$ indicates the optimal choice of $k$, yielding the associated $\hat{\gamma}(k^*) = .761$. This value seems a very sensible choice, as the plot of $\hat{\gamma}(k)$ in the interval $[350, k^*]$ appears fairly flat, so the best choice in this interval is then such that the variance is minimised.[11] Returning to panel (b), the depicted regression line describes the Pareto QQ-plot well.

[11] Giesen and Südekum (2011, Figure 4) consider cities in West Germany. For the year 1997, and using only the 71 largest cities, their point estimate is $\tilde{\gamma}(71) = 1/1.23 = .81$ with 95% confidence interval [.61, 1.22], which is similar to the corresponding point in Fig. 3c.

**Table 2** Revisiting the cross-country OLS regression analysis

| Country | Year | $n_1$ | $\hat{\gamma}(n_1)$ | $k^*$ | $\hat{\gamma}(k^*)$ | $\hat{\gamma}(k^*) - \tilde{b}_{k^*,n}$ |
|---|---|---|---|---|---|---|
| Belgium | 2010 | 158 | 0.608 | 156 | 0.600 | 0.570 |
| Italy | 2011 | 144 | 0.678 | 108 | 0.721 | 0.658 |
| Netherlands | 2014 | 175 | 0.910 | 164 | 0.889 | 0.840 |
| Poland | 2011 | 234 | 0.909 | 132 | 0.817 | 0.781 |
| Russia | 2010 | 166 | 0.955 | 64 | 0.691 | 0.581 |
| Spain | 2011 | 145 | 0.835 | 124 | 0.754 | 0.700 |
| Sweden | 2010 | 128 | 0.800 | 46 | 0.626 | 0.545 |
| Switzerland | 2010 | 162 | 0.879 | 156 | 0.637 | 0.569 |
| Ukraine | 2014 | 215 | 1.087 | 190 | 1.054 | 0.973 |
| United Kingdom | 2011 | 202 | 0.737 | 138 | 0.696 | 0.604 |

Data obtained from http://citypopulation.de/. $n_1$ is the size of the left-truncated data set for the largest cities. $k^*$ minimises the estimated AMSE using the procedure described in Sect. 2.4. Standard errors, not reported but depicted in Fig. 4, can be easily computed using the distributional theory given in Eq. (4). $\tilde{b}_{k^*,n}$ is the conservative bias correction with $\rho = -0.5$ given by Eq. (8)

### 3.2 Cross-country analysis: cities

This illustration revisits and updates the cross-country comparative analysis of Soo (2005) and Nishiyama et al. (2008) using data for the largest cities from `citypopulation.de`.[12] These data sets are left-truncated, and we denote the resulting sample sizes by $n_1$. We consider the largest city sizes for European countries for which at least 100 observations are available. Practitioners use typically the complete data, thus computing (variants of) $\hat{\gamma}(n_1)$. The above theoretical analysis suggests that these are likely to be *over*-estimates (hence biased towards Zipf's law). The purpose of this illustration is to examine whether $k^* < n_1$, whether $\hat{\gamma}(k^*)$ differs from $\hat{\gamma}(n_1)$, and, if so, relate it to the curvature of the diagnostic Pareto QQ-plot. Finally, we perform the lower bounds analysis in order to gauge the magnitude of the potential distortion.

Table 2 reports the results. Although the data are for recent years, the sample sizes $n_1$ and estimates $\hat{\gamma}(n_1)$ are similar to those reported in Soo (2005) (where $1/\hat{\gamma}(n_1)$ is given). For the majority of countries considered, $k^*$ is substantially smaller than $n_1$, which then results in substantially smaller estimates of $\gamma$.[13] These positive distortions are thus in line with the statistical theory developed above.

In Fig. 4 we examine the diagnostic Pareto QQ-plot for four case in which we observe large differences. In panel (a), we depict the Swedish case. The plot reveals a pronounced initial curvature of the QQ-plot, and this significant departure from linearity explains the presence of positive distortions that increase as $k$ increases beyond

[12] Nishiyama et al. (2008) correctly point out that standard errors used in Soo (2005) are wrong, hence undermine his statistical inference. Their empirical analysis uses the correct variance, $2\gamma^2/k$, for their variant of the rank size regression as discussed in Sect. 2.3.

[13] We also observe that at $k^*$ the point estimates $\hat{\gamma}(k^*)$ are very similar to the Hill estimates $\hat{\gamma}^{(\text{Hill})}(k^*)$. For instance, for Sweden, Russia, Poland, and the UK, we obtain 0.608, 0.683, 0.818, and 0.681, respectively.

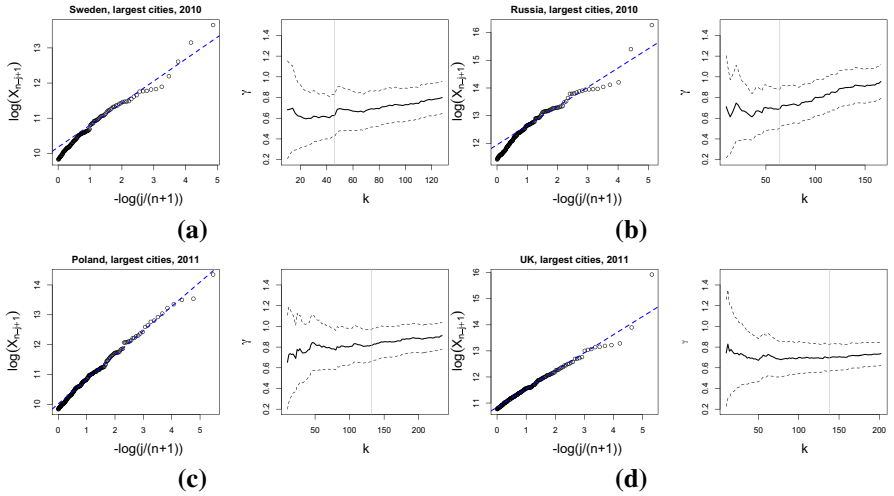



**Fig. 4** Diagnostic Pareto QQ-plot and the Zipf regression estimates $\hat{\gamma}(k)$. Pareto QQ-plots use the $n_1$ largest cities, and Zipf regression line with slope $\hat{\gamma}(k^*)$ and threshold $X_{n-k^*,n}$. Estimates $\hat{\gamma}(k)$ are depicted as a function of the $k$ upper-order statistics used in the Zipf regression (solid line) and associated pointwise 95% symmetric confidence intervals (dashed line),based on the distributional theory given in Eq. (4). The grey vertical line indicates $k^*$

$k^*$. This is further depicted in the accompanying plot of $\hat{\gamma}(k)$. Similar remarks apply to the case of Russia, depicted in panel (b), and Poland, depicted in panel (c). For the UK, the departure from linearity in the QQ-plot is very mild, thus explaining the small difference between $\hat{\gamma}(n_1)$ and $\hat{\gamma}(k^*)$. Turning briefly to Zipf's law, we also observe that the value of 1 lies above the pointwise 95% confidence interval at $k^*$ for Sweden, Russia, and the UK; thus, Zipf's law is rejected for these cases. Taking into account the likely distortion, Table 2 also reports the lower bound given by $\hat{\gamma}(k^*) - \tilde{b}_{k^*,n}$. A bias adjustment in the implied range then suggests that in all cases bar Ukraine, Zipf's law is rejected.

### 3.3 Two agglomerations: Japan and France

In our final illustration concerns two urban agglomerations. First, we revisit the Japanese Urban Employment (UEA) areas in the year 2000, based on commuting patterns, examined in Nishiyama et al. (2008). Table 3 reports the results, and Fig. 5 the diagnostic Pareto QQ-plot and the estimates $\hat{\gamma}(k)$. The point estimate using the complete data, $\hat{\gamma}(n_1)$, suggests a point estimate very close to the Zipf value 1 (almost identical to the value $1/.997$ reported in Nishiyama et al. 2008). But the diagnostic QQ-plot clearly shows an initial pronounced curvature inducing a substantial positive distortion. By contrast, the selection procedure yields $k^* = 70$, and a point estimate of 0.853. However, the estimated variability of the estimate is sufficiently large so that the Zipf value 1 still falls within the 95% confidence interval (even after accounting for its shift suggested by $\tilde{b}_{k^*,n}$). The same observations apply to the French agglomeration data for the year 2015. The selection procedure for $k^*$ substantially reduces the point

**Table 3** Agglomerations in Japan and France

| Country | Year | $n_1$ | $\hat{\gamma}(n_1)$ | $k^*$ | $\hat{\gamma}(k^*)$ | $\hat{\gamma}(k^*) - \tilde{b}_{k^*,n}$ |
|---|---|---|---|---|---|---|
| Japan | 2000 | 113 | 1.031 | 70 | 0.853 | 0.761 |
| France | 2015 | 2226 | 1.124 | 512 | 1.070 | 1.022 |

Data obtained from http://www.csis.u-tokyo.ac.jp/UEA/uea_code_e.htm (Japan) and http://citypopulation.de/ (France). $k^*$ minimises the estimated AMSE using the procedure described in Sect. 2.4. $\tilde{b}_{k^*,n}$ is the conservative bias correction given by Eq. (8)

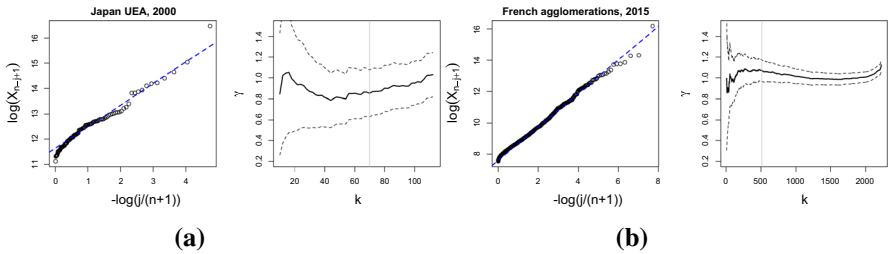



**Fig. 5** Pareto QQ-plot and the Zipf regression estimates $\hat{\gamma}(k)$: Agglomerations in Japan and France. As per Fig. 4

estimate compared to $\hat{\gamma}(n_1)$, but the associated variability is sufficiently large so that the Zipf value 1 is still contained in the confidence interval.

## 4 Conclusions

A Zipf regression is the most popular method for estimating the tail exponent of the city size distribution, and the established literature summarised in several meta-studies and surveys covers close to 100 articles which report thousands of estimates. The (deceptive) ease of computing such a regression has undoubtedly contributed to its popularity. However, the econometric challenges posed by regular-varying upper tails are often not well understood by practitioners: (i) the regression estimator suffers from asymptotic distortions (the bias being usually towards Zipf's law), and (ii) the choice of the threshold parameter, often made mechanically, has important consequences. Both issues have been addressed using techniques that focus on the tail quantile function and that exploit the link between the Zipf regression and the Pareto QQ-plot, a key insight being that this plot becomes linear only eventually and that $\gamma$ is its ultimate slope. The threshold parameter can now be selected in a data-dependent manner. These considerations and proposed methods are new to urban economics.

The relevance of these empirical methods is demonstrated by reconsidering some well-known data sets for the largest cities. While common practice in this established literature uses all available data points $n_1$, it has been shown that in several cases these threshold points belong to the curved part of the Pareto QQ-plot, leading to an over-estimation. By contrast, the proposed methods rectify this problem, yielding estimates

$\hat{\gamma}$ that are smaller than $\hat{\gamma}(n_1)$, sometimes substantially so. Zipf's law is now rejected in some of these cases and confirmed in others.

The formal analysis in this paper is based on the standard assumption made the urban literature that city sizes are independent and identically distributed random variables. All papers cited in footnote 1 and Sects. 1 and 2.3 adopt this assumption. In order to examine to which extent the theoretical predictions hold for dependent data, the Supplementary Material provides evidence for AR(1), MA(1) and GARCH(1,1) processes. Results in Hsing (1991) suggest that the current theory might be a reasonable guide if the dependence is sufficiently weak so that approximations to a normal law still hold. The Supplementary Material demonstrates that this is the case. In particular, in all experiments considered, the Pareto QQ-plots exhibit the concave-like curvature, and our method selects well the ultimate linear part of these QQ-plot.

**Acknowledgements** I thank the referees for their constructive comments that have helped to improve the paper. Financial support from ANR-DFG (Grant ANR-15-FRAL- 0007-01) and ANR-17-EURE-0020 is also gratefully acknowledged.

### Compliance with ethical standards

**Conflict of Interest** The author declares that he has no conflict of interest.

**Ethical approval** This article does not contain any studies with human participants or animals performed by the author.



## A Statistical Appendix

### A.1 The Pareto QQ-plot, the Zipf regression, and $\hat{\gamma}$

The Pareto QQ-plot has coordinates $(x, y) = (-\log(j/(n+1)), \log X_{n-j+1,n})_{j=1,\ldots,k}$. In model (1), this plot becomes *ultimately* linear for a sufficiently high threshold $X_{n-k,n}$ where $k < n$. The line through the threshold point $-\log((k+1)/(n+1)), \log X_{n-k,n})$ with slope $\gamma$ is thus given by

$$y = \log X_{n-k,n} + \gamma \left[ x + \log\left(\frac{k+1}{n+1}\right)\right] \qquad (1 \le j \le k < n).$$

The OLS estimator of the slope parameter in the Pareto QQ-plot is obtained by minimising the least squares criterion

$$\sum_{j=1}^{k}\left(\log\frac{X_{n-j+1,n}}{X_{n-k,n}}-\gamma\log\frac{k+1}{j}\right)^2 \qquad (1\le j\le k<n)$$

with respect to $\gamma$. The resulting OLS estimator is

$$\hat{\gamma}=\frac{\frac{1}{k}\sum_{j=1}^{k}\log\left(\frac{k+1}{j}\right)\left[\log X_{n-j+1,n}-\log X_{n-k,n}\right]}{\frac{1}{k}\sum_{j=1}^{k}\left[\log\frac{k+1}{j}\right]^2}.$$

### A.2 The choice of *k*: details

The mean weighted theoretical squared deviation

$$\frac{1}{k}\sum_{j=1}^{k}w_{j,k}E\left(\log\left(\frac{X_{n-j+1,n}}{X_{n-k,n}}\right)-\gamma\log\left(\frac{k+1}{j}\right)\right)^2$$

equals, to first order, $c_k\mathrm{Var}(\hat{\gamma})+d_k(\rho)b_{k,n}^2$ for some coefficients $c_k$ depending only on $k$, and $d_k(\rho)$ depending on $k$ and $\rho$. For model (2) Schluter (2018) shows that these coefficients are $d_k(\rho)=(\frac{1}{2}\frac{2-\rho}{(1-\rho)^2})^{-2}\tilde{d}_k(\rho)$ and $c_k=(4/5)\tilde{c}_k$ with

$$\tilde{d}_k(\rho)=\frac{1}{k}\sum_{j=1}^{k}w_{j,k}\left(\frac{(j/(k+1))^{-\rho}-1}{\rho}\right)^2$$

$$\tilde{c}_k=\sum_{j=1}^{k}w_{j,k}\left(\sum_{l=1}^{k-j+1}\left(\frac{1}{k-l+1}\right)^2+\left(\sum_{l=1}^{k-j+1}\frac{1}{k-l+1}-\log\left(\frac{k+1}{j}\right)\right)^2\right).$$

### A.3 Example: the Hall (1982) class of distributions

The Hall class of distributions (Hall 1982) satisfies the second-order representation (2) and possesses a nuisance function $l$ in model (1) that converges to a constant at a polynomial rate. This class contains, among others, the Burr, Student $t$, Fréchet, and Cauchy distributions.

In particular, the distribution function of the Hall class is given by

$$1-F(x)=ax^{-1/\gamma}[1+bx^{\beta}+o(x^{\beta})]$$

for large $x$ with $\gamma,a>0$, $b\in R$, $\beta<0$, and tail quantile function

$$U(x)=cx^{\gamma}[1+dx^{\rho}+o(x^{\rho})].$$

It satisfies the second-order representation (2) with $\rho = \gamma\beta < 0$, and rate function $A(t) = (\rho^2/\gamma)dt^\rho$.[14]

The Burr distribution, parametrised here as

$$1 - F_{(\gamma,\rho)}(x) = (1 + x^{-\rho/\gamma})^{1/\rho},$$

is a member of the Hall class with parameters $\gamma$ and $\rho < 0$, $c = 1$ and $d = \gamma/\rho < 0$. Its tail quantile function can be expanded as

$$U(x) = x^\gamma[1 + (\gamma/\rho)x^\rho + o(x^\rho)].$$

Other members, for instance, are (ii) the Student $t_\delta$ distribution with $\delta$ degrees of freedom: with $\gamma = 1/\delta$, $\rho = -2/\delta$, $d = \gamma BC^{-2\gamma}$, $B = -.5\delta^2(\delta + 1)/(\delta + 2)$, and $C = \Gamma((\delta + 1)/2)\delta^{(\delta-1)/2}/(\delta\pi)^{1/2}\Gamma(\delta/2)$ (valid for $\delta > 2$); (iii) the Fréchet distribution $F_\gamma(x) = \exp(-x^{-1/\gamma})$ with $\rho = -1$, $c = 1$, and $d = -.5\gamma$; and (iv) the Cauchy distribution with $\gamma = 1$, $\rho = -2$, $c = 1/\pi$, and $d = -.5\pi^2$.

### A.3.1 Concavity of the Pareto QQ-plot

For distribution model (1) the Pareto QQ-plot becomes linear only eventually and exhibits typically a concave-like curvature. This can be easily verified using the population analogue, i.e. the tail quantile function.[15]

In particular, in the Burr case, $\log U(x) \approx \gamma \log(x) + (\gamma/\rho)x^\rho$ for large $x$, and it follows immediately that

$$\frac{\partial \log U(x)}{\partial \log(x)} = \gamma + \gamma x^\rho > 0 \quad \text{and} \quad \frac{\partial^2 \log U(x)}{\partial \log(x)^2} = \gamma\rho x^\rho \leq 0$$

$\log U(x) \sim \gamma \log(x)$ only as $x \to \infty$. Thus, the presence of the nuisance function $l$ in model (1) augments the slope and induces concavity when $x$ is not sufficiently large, leading then to an over-estimation of $\gamma$ by the Zipf regression. As the second-order parameter $\rho$ decreases in magnitude, and the nuisance function $l$ decays more slowly, the Pareto QQ-plot becomes steeper, $\frac{\partial}{\partial|\rho|}\frac{\partial \log U(x)}{\partial \log(x)} < 0$, and the distortion increases. These properties are illustrated in Sect. A.3.3.

More generally, a similar calculation for the general Hall class reveals that the signs of the first and second derivatives of $\log U(x)$ are given by sign(-$d$) and sign($d$) respectively, recalling that $d < 0$ for the Burr, Student t, Fréchet, and Cauchy distributions.

[14] Gabaix and Ibragimov (2011) provide numerical evidence to show that the log-log rank size regressions with shifted ranks performs well in Hall's model with $\rho = -1$.

[15] This calculation is valid, of course, for any parametric model. It is of interest, for instance, to consider the lognormal distribution. Schluter and Trede (2019) show that in this limiting case $\gamma = 0$ and that its tail quantile function satisfies, to first order, $\log(U(x)) = \sigma\sqrt{2}[\log x]^{1/2}$. Then $\partial \log U(x)/\partial \log(x) = (\sigma/\sqrt{2})[\log x]^{-1/2} \downarrow 0$. The Pareto QQ-plot can thus be seen to be concave, and its eventual slope is 0.

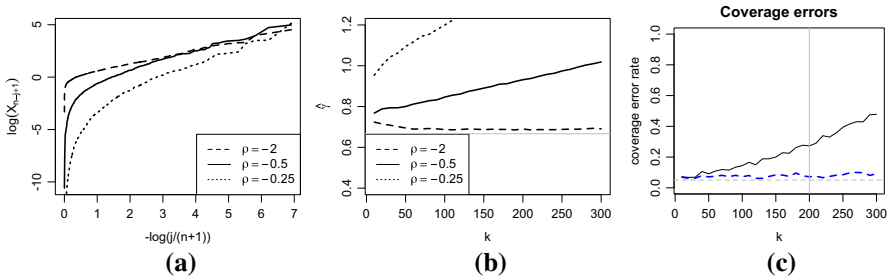



**Fig. 6** The Burr distribution: Pareto QQ-plot and the Zipf regression estimate $\hat{\gamma}$. Notes. Based on the Burr distribution with $\gamma = 2/3$, and $\rho \in \{-2, -.5, -.25\}$. **a** Pareto QQ-plots for 3 random samples drawn from the Burr distribution. Sample size is 1000. To aid comparison across cases, the points of each QQ-plot have been connected and rendered as lines. **b** Mean of estimates $\hat{\gamma}$ across 1000 Monte Carlo simulations for given $\rho$, drawing samples of size 1000 in each iteration. The faint horizontal line is the population value $\gamma = 2/3$. **c** For $\rho = -.5$, coverage error rate of the 95% symmetric confidence intervals (solid line),based on the distributional theory given in Eq. (4), and bias corrected by the theoretical $b_{k,n}^{\text{Burr}}$ (dashed line)

### A.3.2 The sign of the higher-order bias of $\hat{\gamma}$

In the Hall class, it then follows that the sign of the higher-order bias of the slope estimator $\hat{\gamma}$, $b_{k,n}$ stated in Eq. (5), is given by $-\text{sign}(d)$.

For many members of the Hall class such as the Burr, Student $t$, Fréchet, and Cauchy distributions, the above results imply $d < 0$, so $b_{k,n} > 0$. For instance, $d = \gamma/\rho < 0$ in the Burr case. Moreover, $\partial b_{k,n}^{Burr}/\partial|\rho| < 0$, so that the smaller the magnitude of $\rho$, the larger is the distortion.

We conclude that $\gamma$ is *over-estimated*, so that Zipf regressions are biased *towards* Zipf's law in this settings.

The consequences are illustrated quantitatively next.

### A.3.3 Numerical illustrations

We illustrate the second-order behaviour specifically for the Burr distribution. Figure 6 illustrates the role of $\rho$ for the Pareto QQ-plot, and the resulting estimates $\hat{\gamma}(k)$ when $\gamma = 2/3$. In line with the theoretical discussion of Sect. A.3.1, we observe that the smaller the magnitude of $\rho$, the greater the initial concave-like curvature and steepness of the Pareto QQ-plot, and the larger the induced positive distortions of the OLS estimator of its slope coefficient. This is also consistent with the theoretical bias $b_{k,n}^{\text{Burr}}$ discussed above. The qualitative results are similar to those for the real-world distribution depicted in Fig. 1.

In panel (c) of Fig. 6 we illustrate the consequences of the distortions for statistical inference for the case $\rho = -.5$, by plotting the empirical coverage error rates of the usual 95% symmetric confidence intervals. The higher-order distortions lead to undermining inference because of the considerable size distortions. For instance, at $k = 200$, the empirical coverage error rate is 30% for a nominal 5% rate. Shifting the estimate by the theoretical bias $b_{k,n}^{\text{Burr}}$ reduces the coverage error rate to 7%.

**Table 4** Performance evidence for the switching model: Burr distribution

| $\rho$ | $N = 10{,}000$ | | $N = 1000$ | |
|---|---|---|---|---|
| | Mean $\hat{\gamma}$ | SD($\hat{\gamma}$) | Mean $\hat{\gamma}$ | SD($\hat{\gamma}$) |
| $-2$ | 0.684 | (.0138) | 0.729 | (.430) |
| $-.75$ | 0.854 | (.0334) | 0.876 | (.084) |
| $-.5$ | 1.000 | (.0317) | 0.977 | (.077) |

Burr model with fixed $\gamma = 2/3$ and variable $\rho$ and sample of sizes $N$. The Monte Carlo is based on $R = 1000$ repetitions. The switching model is as proposed in Ioannides and Skouras (2013) and set out in the main text. Estimation of all parameters is by maximum likelihood. Reported is only the estimate of $\gamma$

Finally, we illustrate how recent switching models that are designed to fit the entire city size distribution inherit the bias problem caused by the strict Pareto assumption. In particular, we examine the performance of the switching model of Ioannides and Skouras (2013) that smoothly pastes a strict Pareto tail to a lognormal body. Beyond cut-off $\tau$, the density model is proportional to $a \times x^{-1/\gamma-1}$ where the parameter-dependent scaling factor $a$ ensures that the density is continuous at the cut-off $\tau$. The parameters of the model (location and scale of the lognormal body, cut-off $\tau$ and $\gamma$) are estimated by maximum likelihood.[16] For this Monte Carlo illustration, we use the Burr model above with parameters $\gamma = 2/3$ and varying $\rho$, repeat the experiment $R = 1000$ times, and draw in each iteration a sample of size $N = 10{,}000$ or $N = 1000$. The maximum likelihood procedure, it turns out, correctly dismisses the lognormal body by invariably estimating a very low cut-off point $\tau$. However, $\gamma$ is over-estimated and the distortion increases as $\rho$ falls in magnitude (and the nuisance function $l$ in model (1) decays more slowly), as predicted by our statistical theory. Table 4 reports the results. In particular, for samples of size $N = 10{,}000$, the distortion increases as $\rho$ changes from - 2 to - 0.5, the mean value of $\hat{\gamma}$ being 1.0003 with mean standard deviation 0.0317 for $\rho = -.5$. Drawing smaller samples has the predicted effect of increasing variability.

## References

Beirlant J, Vynckier P, Teugels JL (1996) Tail index estimation, Pareto quantile plots, and regression diagnostics. J Am Stat Assoc 9(436):1659–1667

Beirlant J, Goegebeur Y, Segers J, Teugels J (2004) Statistics of extremes, wiley series in probability and statistics. Wiley, Chichester

Bingham NH, Goldie CM, Teugels JL (1987) Regular variation encyclodepedia of mathematics and its applications. Cambridge University Press, Cambridge

Cottineau C (2016) (Re)producing knowledge about city size distributions. UCL working paper

Csörgő S, Deheuvels P, Mason D (1985) Kernel estimates of the tail index of a distribution. Ann Stat 13(3):1050–1077

[16] The implementation of the maximum likelihood estimation procedure allows a close replication of the empirical results reported in Ioannides and Skouras (2013, Table 1, p. 22) for the US year 2000 Census Bureau data. We obtain point estimates of 7.26 and 1.73 for the location and scale parameters of the lognormal distribution, a cut-off estimate $\hat{\tau}$ of 60203, and $1/\hat{\gamma} = 1/.802 = 1.247$. The reported estimates are 7.26, 1.73, 60290, and 1.25, respectively.

de Haan L, Ferreira A (2006) Extreme value theory. Springer, New York
de Haan L, Peng L (1998) Comparison of tail index estimators. Stat Neerl 52:60–70
de Haan L, Stadtmüller U (1996) Generalized regular variation of second order. J Aust Math Soc (Ser A) 61:381–395
Dekkers ALM, Einmahl JHJ, de Haan L (1989) A moment estimator for the index of an extreme-value distribution. Ann Stat 17:1833–1855
Draisma G, de Haan L, Peng L, Pereira TT (1999) A bootstrap-based method to achieve optimality in estimating the extreme-value index. Extremes 2(4):367–404
Eeckhout J (2004) Gibrat's law for (all) cities. Am Econ Rev 94(5):1429–1451
Embrechts P, Kluppelberg C, Mikosch T (1997) Modelling extremal events. Springer, Berlin
Fazio G, Modica M (2015) Pareto or log-normal? Best fit and truncation in the distribution of all cities. J Reg Sci 55(5):736–756
Gabaix X (1999b) Zipf 's law for cities: an explanation. Quart J Econ 114(3):739–767
Gabaix X (2009) Power laws in economics and finance. Annu Rev Econ 1:255–293
Gabaix X, Ibragimov R (2011) Rank-1/2: a simple way to improve the OLS estimation of tail exponents. J Bus Econ Stat 29(1):24–39
Gabaix X, Ioannides Y (2004) The evolution of city size distributions. In: Henderson J, Thisse J-F (eds) Handbook of regional and urban economics, volume 4: cities and geography. Elsevier, Amsterdam
Giesen K, Südekum J (2011) Zipf's law for cities in the regions and the country. J Econ Geogr 11:667–686
Gonzalez-Val R (2010) The evolution of U.S. city size distribution from a long-term perspective (1900–2000). J Reg Sci 50(5):952–972
Haeusler E, Teugels JL (1985) On asymptotic normality of Hill's estimator for the exponent of regular variation. Ann Stat 13(2):743–756
Hall P (1982) On some simple estimate of an exponent of regular variation. J R Stat Soc Ser B 44:37–42
Hall P (1990) Using the Bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. J Multivar Anal 32:177–203
Hsing T (1991) On tail index estimation using dependent data. Ann Stat 19:1547–1569
Ibragimov M, Ibragimov R, Kattuman P (2013) Emerging markets and heavy tails. J Bank Finance 7:2546–2559
Ibragimov M, Ibragimov R, Walden J (2015) Heavy-tailed distributions and robustness in economics and finance. Springer, Berlin
Ioannides Y, Skouras S (2013) US city size distributions: Robustly Pareto, but only in the tail. J Urban Econ 73:18–29
Kratz M, Resnick SI (1996) The QQ-estimator and heavy tails. Commun Stat Stoch Models 12:699–724
Levy M (2009) Gibrat's law for (all) cities: comment. Am Econ Rev 99(4):1672–1675
Luckstead J, Devadoss S (2014) Do the world's largest cities follow Zipf's and Gibrat's laws? Econ Lett 125:182–186
Malevergne Y, Pisarenko V, Sornette D (2011) Testing the Pareto against the lognormal distributions with the uniformly most powerful unbiased test applied to the distribution of cities. Phys Rev E 83:036111-1–036111-11
Nishiyama Y, Osada S, Sato Y (2008) OLS estimation and the *t* test revisited in rank-size rule regression. J Reg Sci 48(4):691–715
Nitsch V (2005) Zipf Zipped. J Urban Econ 57:86–100
Perline R (2005) Strong, weak and false inverse power laws. Stat Sci 20(1):68–88
Reed WJ (2002) On the rank-size distribution for human settlements. J Reg Sci 42:1–17
Rose AK (2006) Cities and countries. J Money Credit Banking 38(8):2225–2246
Schluter C (2018) Top incomes, heavy tails, and rank-size regressions. Econometrics 6:10
Schluter C, Trede M (2019) Size distributions reconsidered. Econ Rev 38(6):695–710
Schultze J, Steinebach J (1996) On least squares estimates of an exponential tail coefficient. Stat Decis 14:353–372
Soo KT (2005) Zipf's law for cities: a cross country investigation. Reg Sci Urban Econ 35:239–263