Power Law in COVID-19 Cases in China

BEHZOD B. AHUNDJANOV[†], SHERZOD B. AKHUNDJANOV[‡], and BOTIR B. OKHUNJANOV[§]

[†]*Department of Economics, Dickinson College, Carlisle, PA*
[‡]*Corresponding author. Department of Applied Economics, Utah State University, 4835 Old Main Hill, Logan, UT 84322-4835. E-mail: sherzod.akhundjanov@usu.edu*
[§]*School of Economic Sciences, Washington State University, Pullman, WA*

Abstract

The novel coronavirus (COVID-19) was first identified in China in December 2019. Within a short period of time, the infectious disease has spread far and wide. This study focuses on the distribution of COVID-19 confirmed cases in China—the original epicenter of the outbreak. We show that the upper tail of COVID-19 cases in Chinese cities is well described by a power law distribution, with exponent less than one, and that a random proportionate growth model predicated by Gibrat's law is a plausible explanation for the emergence of the observed power law behavior. This finding is significant because it implies that COVID-19 cases in China is heavy-tailed and disperse, that a few cities account for a disproportionate share of COVID-19 cases, and that the distribution has no finite mean or variance. The power-law distributedness has implications for effective planning and policy design as well as efficient use of government resources.

Keywords: COVID-19; Gibrat's law; heavy-tailedness; Pareto distribution; Power law.

JEL: C13, C46, I1.

Declarations:

Funding: Not applicable

Conflicts of interest/Competing interests: Not applicable

Availability of data and material: Data will be made available upon publication

Code availability: Code will be made available upon publication

1 Introduction

The coronavirus disease 2019 (COVID-19) was first discovered in Wuhan region of China in December 2019 (Zhu et al., 2020). The contagious disease quickly spread within China, despite unprecedented and aggressive containment measures, and crossed the borders reaching every corner of the world within a short period of time, with the World Health Organization (WHO) declaring COVID-19 outbreak a global pandemic on March 11, 2020 (Cucinotta and Vanelli, 2020). This study focuses on the distribution and growth dynamics of COVID-19 cases in China—the original epicenter of the outbreak. The presence of Chinese cities with very large number of COVID-19 confirmed cases, the very wide dispersion in COVID-19 cases across Chinese cities, and the effect of COVID-19 pandemic on the economy and welfare make it crucial for researchers and policymakers to better understand COVID-19 distribution for effective planning and policy design as well as efficient use of government resources.

In this paper, we demonstrate that the right tail of the distribution of COVID-19 confirmed cases in Chinese cities is well-characterized by a power law (Pareto) distribution, meaning that the probability that a number of COVID-19 cases is more than x is roughly proportional to $x^{-\gamma}$, i.e., $\text{Prob}(X > x) \sim x^{-\gamma}$, where γ is the power law (Pareto) exponent.[1] The estimated power law exponent is $\gamma < 1$, meaning the fitted power law distribution has no finite moments, including mean and variance. The power law fit is robust to a range of estimation methods and goodness-of-fit tests, and the distribution parsimoniously describes the heavy tail of the data. Power law distributions are characterized by heavy tails, which make the likelihood of extreme (upper-tail) events more typical. So, in case of COVID-19, this implies an extremely large number of cases becomes more likely, which is actually true for China, where few cities had extremely large number of cases (Han et al., 2020).

Power laws are extraordinarily ubiquitous in the social and natural sciences, having been confirmed for the distributions of income and wealth (Pareto, 1896; Champernowne, 1953; Wold and Whittle, 1957; Singh and Maddala, 1976; Klass et al., 2006; Toda, 2012), consumption (Toda, 2017; Toda and Walsh, 2015), firm size (Stanley et al., 1995; Axtell,

[1]For a detailed review of power laws, see Reed (2001), Newman (2005), Sornette (2006) and Gabaix (2009, 2016). The power law distribution with exponent $\gamma \simeq 1$ generates Zipf's law (Zipf, 1949).

2001; Luttmer, 2007), farmland (Akhundjanov and Chamberlain, 2019), city size (Krugman, 1996; Gabaix, 1999; Ioannides and Overman, 2003; Devadoss et al., 2016), natural gas and oil production (Balthrop, 2016), carbon dioxide (CO_2) emissions (Akhundjanov et al., 2017), frequency of words (Zipf, 1949; Irmay, 1997), among others. Our paper finds the existence of a power law in epidemiology as well. The omnipresence of power laws is partly explained by the fact they are preserved over an extensive array of mathematical transformations (Gabaix, 2009).

An interesting aspect of power law distribution is that it is the macro-level steady-state phenomenon that, in theory, can arise from a micro-level random proportionate growth process, known as Gibrat's law (Gibrat, 1931),[2] whereby each unit's (e.g., city's) growth rate is drawn randomly and independently of its current size.[3] Given power law and Gibrat's law often go hand-in-hand, Gibrat's law has also been extensively documented in the social and natural sciences.[4] The robust fit of power law to cross-sectional distribution of COVID-19 cases in Chinese cities potentially provides macro-level evidence for random proportionate growth posited by Gibrat's law. However, it is well known that power laws can similarly be obtained from other models and systems (Barabási and Albert, 1999; Carlson and Doyle, 1999; Mitzenmacher, 2004; Newman, 2005; Gabaix, 2016). Therefore, we formally test for random proportionate growth at micro-level by analyzing growth rates of COVID-19 cases in Chinese cities. Our empirical analysis provides support for Gibrat's law of proportionate growth, which, in turn, offers a plausible explanation for the emergence of a power law behavior in the data.

[2]For a detailed review of Gibrat's law, see Sutton (1997).

[3]Gibrat's law alone is not sufficient to give rise to a power law. In fact, it leads to the lognormal distribution as shown by Gibrat (1931); though many examples used by Gibrat (1931) have recently been shown to actually follow a Pareto-type distribution rather than the lognormal (Akhundjanov and Toda, 2020). Nonetheless, Gibrat's law can generate a power law with an auxiliary assumption (Gabaix, 1999). Section 4.1 elaborates on a link between Gibrat's law and power law.

[4]In particular, Gibrat's law has been shown to explain the growth process of consumption (Battistin et al., 2009), firms (Luttmer, 2007), farms (Clark et al., 1992), trucking industry (Balthrop, 2020), cities (Ioannides and Overman, 2003; Eeckhout, 2004; González-Val, 2010; Luckstead and Devadoss, 2014), countries (Rose, 2006; González-Val and Sanso-Navarro, 2010), bird population (Keitt and Stanley, 1998), among others.

There are a number of practical implications of the power law fit. First, given the estimated Pareto exponent is less than one ($\gamma < 1$), the distribution is heavy-tailed and so disperse that observations near the mean account for little of the cumulative distribution of COVID-19 cases. This implies talking about the average number of COVID-19 cases is inconsequential as it no longer represents the majority of cases. In fact, even though it is possible to compute sample mean and variance for the observed data, these moments are generally non-convergent. In this case, quantile analysis or order statistics would be more appropriate. Second, the heavy upper-tail of the distribution (also, the confirmation of Gibrat's law) is suggestive of concentration of COVID-19 cases in China, with the total cases essentially determined by a few cities that bore the brunt of the outbreak, which is true in case of China (Han et al., 2020). This has implications for more effective epidemiological planning and policy design as by focusing on and directing appropriate amount of resources to those epicenters total cases of infection can be greatly slashed and the spread of an outbreak potentially contained. Finally, on a more technical note, a heavy upper-tail of COVID-19 cases in Chinese cities has implications for empirical research. Specifically, it shows that thin-tailed distributions (e.g., the normal) or moderately heavy-tailed distributions (e.g., the lognormal), which are often 'go-to' distributions in empirical research, are inappropriate for COVID-19 cases in Chinese cities as such distributions dismiss extremely large number of cases as an improbable observation. On the other hand, a power-law distribution is able to capture the heavy upper-tail of the data, which it does so parsimoniously, outperforming a number of competing distributions.

The literature in this area is thin, but gradually forming. In the concurrent work, Beare and Toda (2020), studying the distribution of COVID-19 confirmed cases for US counties, find that the upper-tail of this distribution follows a power law, with Pareto exponent close to 1. Similarly, Blasius (2020), examining the distribution of COVID-19 confirmed cases and deaths for US counties, concludes that both distributions exhibit a power-law behavior. Our paper contributes to this nascent line of literature by exploring the distribution as well as underlying growth dynamics of COVID-19 confirmed cases in China—the origin of the outbreak. A distinctive feature of our study is that COVID-19 cases in China affords us to capture the entire life cycle of the pandemic (at least in its first wave): outbreak detection,

spread, peak, and decline to zero daily cases. In contrast, the analyses of Beare and Toda (2020) and Blasius (2020) are based on data sets that were largely evolving at the time, as both the United States of America and a whole host of other countries are still battling to contain the spread of the virus to this date. Thus, the results of the above studies are likely subject to change with newer data.

The remainder of the paper is structured as follows. Section 2 introduces the data for COVID-19 cases in China. Section 3 presents the methods and findings for power law analysis. Section 4 provides the methods and results for Gibrat's law analysis. Section 5 includes some concluding remarks.

2 Data

Daily data on the cumulative number of COVID-19 confirmed cases for Chinese cities comes from Harvard Dataverse (China Data Lab, 2020). The dataset includes 339 cities in China and covers periods between January 15, 2020, and May 23, 2020, which are determined by the data source. Our main (power-law) analysis focuses on COVID-19 cases as of May 23, 2020, the latest data on cumulative cases, while an auxiliary analysis potentially explaining the emergence of a power law behavior uses the data between January 15, 2020, and May 23, 2020. A power law analysis is data intensive, with Clauset et al. (2009) recommending a minimum of 50 observations for reliable analysis. This condition is well-satisfied here, including for the upper tail (see Section 3.3).

Figure 1 shows the evolution of empirical distribution of COVID-19 cases in Chinese cities over select dates. It is apparent that the distribution has been right-skewed, with heavier right tail, and it has been gradually sliding rightward, which reflects increasing number of COVID-19 cases across Chinese cities over time.

3 Power law analysis

In this section, we study the distribution of the cumulative number of COVID-19 confirmed cases for Chinese cities. We first present the methodology for power law analysis, followed by estimation results.

3.1 Power-law parameter estimation

Suppose X is a random variable whose data generating process is a continuous power law (Pareto) distribution. The corresponding probability distribution function (PDF) is specified as

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

where x is an outcome of X for $x \in \mathbb{R}_+$, where $\mathbb{R}_+ = \{x \in \mathbb{R} | x > 0\}$, $x_{\min}$ is the threshold beyond which (i.e., $x \geq x_{\min}$) power-law behavior sets in, and α is the power-law (Pareto) exponent, a parameter of interest. The mth non-central moment for the power law distribution is given by

$$\langle x^m \rangle = \int_{x_{\min}}^{\infty} x^m f(x) dx = \left(\frac{\alpha - 1}{\alpha - 1 - m} \right) x_{\min}^m, \quad \forall \alpha > m + 1. \tag{2}$$

Hence only the first $\lfloor \alpha - 1 \rfloor$ moments exist for $m < \alpha - 1$. Although higher-order moments can be calculated for any finite sample, these estimates do not asymptotically converge to any particular value. Given the sample $x_1, \ldots, x_n$, the joint log-likelihood function can be written as

$$\begin{aligned} \ln \mathcal{L}(\alpha; x_1, \ldots, x_n) &= \sum_{i=1}^{n} \left[\ln(\alpha - 1) - \ln x_{\min} - \alpha \ln \frac{x_i}{x_{\min}} \right] \\ &= n \ln(\alpha - 1) - n \ln x_{\min} - \alpha \sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}}. \end{aligned} \tag{3}$$

First-order condition yields the maximum likelihood estimate (MLE) of

$$\alpha^{MLE} = 1 + n \left(\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}} \right)^{-1} \tag{4}$$

with the standard error (SE) of the estimate given by

$$SE(\alpha^{MLE}) = \sqrt{n} \left(\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}} \right)^{-1} = \frac{\alpha^{MLE} - 1}{\sqrt{n}}. \tag{5}$$

It is standard to report the counter-cumulative parameter $\gamma = \alpha^{MLE} - 1$, known as the Hill estimator (Hill, 1975), instead of (4). The Hill estimator is obtained from (4), after a small-sample adjustment, and takes the following form

$$\gamma^{Hill} = \frac{n-2}{\sum_{i=1}^{n-1}(\ln x_i - \ln x_{\min})} \tag{6}$$

with the standard error of the estimate given by

$$SE(\gamma^{Hill}) = \frac{\gamma^{Hill}}{\sqrt{n-3}}. \tag{7}$$

The power-law fit to data is depicted by plotting the counter- (complimentary-) cumulative distribution function (CDF) on doubly logarithmic axes. The counter-CDF of a power law is specified as

$$\text{Prob}(X > x) = \left(\frac{x}{x_{\min}}\right)^{-\alpha+1} = \frac{k}{x^{\gamma}}, \tag{8}$$

where $k = x_{\min}^{\alpha-1}$ is a constant. Taking the log of both sides of (8) yields a linear relationship between log counter-cumulative probability (i.e., $\ln \text{Prob}(X > x)$) and log data (i.e., $\ln x$), with the counter-cumulative parameter $-\gamma$ being the slope of the line.

An alternative approach to estimate the counter-cumulative parameter γ is through a regression-based technique. Specifically, estimate the following regression equation with ordinary least squares (OLS)

$$\ln(rank_i) = \phi - \gamma^{OLS} \ln x_i + \varepsilon_i, \tag{9}$$

where $rank_i$ is observation i's rank in the distribution, ϕ is the intercept term, γ^{OLS} is the parameter of interest, and ε_i is the idiosyncratic disturbance term. Equation 9 also shows that a power law distributed process appears approximately linear on a log-log plot of $rank_i$ against x_i, with slope of $-\gamma^{OLS}$. The asymptotic standard error for γ^{OLS} is given by (Gabaix

and Ibragimov, 2011)

$$SE(\gamma^{OLS}) = \frac{\gamma^{OLS}}{\sqrt{n/2}}. \tag{10}$$

An important consideration in power law analysis is the specification of the threshold parameter $x_{\min}$, beyond which power-law behavior takes hold. There are several approaches proposed in the literature in this regard. For instance, one strand of literature suggests to select $x_{\min}$ arbitrarily at either the 95% quantile of the data (Gabaix, 2009) or the point where empirical PDF or CDF roughly straightens out on a log-log plot. Clearly, both of these approaches are rather subjective and thus suffer from a certain degree of uncertainty about whether they are able to capture the true starting point of power-law behavior. In fact, Perline (2005), exploring the empirical consequences of this concern, shows that sufficiently truncated Gumbel-type distributions (e.g., the lognormal) can also produce a linear pattern on a log-log plot, hence imitating the power law distribution. Consequently, we adopt a more systematic, data-driven procedure proposed by another strand of literature (Clauset et al., 2009) to select $x_{\min}$. This approach essentially treats each observation in the sample as a potential candidate for $x_{\min}$ and selects the best candidate based on the minimization of the Kolmogorov-Smirnov (KS) goodness-of-fit statistic, which is given by

$$KS = \max_{x \geq x_{\min}} |E(x) - \hat{F}(x)|, \tag{11}$$

where $E(x)$ is the empirical CDF and $\hat{F}(x)$ is the estimated power-law CDF. The optimal $x_{\min}$ minimizes the distance between the empirical CDF and the estimated power-law CDF. The computational algorithm takes the following form:

Step 1: Set $x_{\min} = x_1$;

Step 2: Perform power-law parameter estimation using $x \geq x_{\min}$;

Step 3: Compute the KS statistic in (11);

Step 4: Repeat steps 1-4 for all x_i for $i = 1, \ldots, n$;

Step 5: Select $x_{\min}$ with the lowest KS statistic.

3.2 The goodness of fit tests

Power-law analysis is accompanied by a series of diagnostic tests given significant parameter estimates alone do not provide sufficient evidence in favor of power-law fit to data. In order to guard against potential misspecification issues, one needs to conduct a goodness-of-fit test and compare the power-law fit to data with those of alternative distributions.

Gabaix and Ibragimov (2011) proposed 'rank - 1/2' test to verify the goodness-of-fit of power law distribution. Let x^* be defined as

$$x^* = \frac{\text{Cov}[(\ln x_i)^2, \ln x_i]}{2\text{Var}(\ln x_i)}. \tag{12}$$

Then, regress bias-adjusted log rank against the log data and a quadratic deviation term, as in

$$\ln\left(rank_i - \frac{1}{2}\right) = \phi + \zeta \ln x_i + q(\ln x_i - x^*)^2 + \varepsilon_i. \tag{13}$$

The goodness-of-fit statistic is specified as q/ζ^2. The null hypothesis of power-law distributedness is rejected if $q/\zeta^2 > 1.95(2n)^{-1/2}$, where the latter term is the goodness-of-fit threshold.

Further, Clauset et al. (2009) suggest comparisons of power-law fit with those of other, competing, heavy-tailed distributions, such as the lognormal and exponential. Accordingly, we fit these alternative distributions to the data by MLE and provide visual comparisons of the distributions' fits on a doubly logarithmic plot as detailed above. We also implement the likelihood ratio test of Clauset et al. (2009) for a more formal comparison. The likelihood ratio statistic is specified as

$$\mathcal{R} = \sum_{i=1}^{n} \left[\ln \hat{f}_1(x_i) - \ln \hat{f}_2(x_i)\right], \tag{14}$$

where $\hat{f}_1(x_i)$ and $\hat{f}_2(x_i)$ are the probabilities predicted by power law and an alternative distribution, respectively. If the likelihood ratio statistic is positive, it indicates the power law distribution fits the data more closely. If it is negative, then an alternative distribution

yields a better fit.[5]

3.3 Application

The methods discussed in Section 3 are applied to the cumulative number of COVID-19 confirmed cases in Chinese cities (x) as of May 23, 2020. The results from power law analysis are provided in Tables 1-2 and Figure 2. As noted earlier, the requirement placed on sample size for credible power law analysis is a minimum of 50 observations (Clauset et al., 2009). This condition is well-satisfied here as the upper-tail sample ($x > x_{\min}$) contains 151 observations. The Hill and OLS estimates of the counter-cumulative parameter γ are around 0.80 and highly statistically significant. Given $m < 0.80$, the moments of the fitted power law distribution (including mean and variance) are generally non-convergent. The goodness-of-fit test of Gabaix and Ibragimov (2011) suggests we fail to reject the null hypothesis of power-law distributedness, which provides strong evidence in favor of power law fit to COVID-19 cases in China.

Figure 2 depicts the power law and competing heavy-tailed distributions' fits to the data. It is apparent that the power law distribution generally fits the data better than the rivals, particularly in the lower to mid quantiles of the upper tail, where the observed data forms a distinct linear pattern. The power law slightly overestimates the frequency of the largest cases in the extreme upper tail (after log confirmed cases of about 7.8), where the distribution decays relatively slowly. Now, there are more flexible forms of the Pareto distribution—often with an extra parameter and/or of mixture form—that allow the extreme upper-tail probabilities of the distribution to decay more quickly, and they have repeatedly been shown to improve upon the benchmark power law distribution in fitting empirical data (see, for instance, Patel and Schoenberg, 2011). The main goal of the present study is to examine whether a power law in general approximates COVID-19 cases in China, and not the investigation of various (modified) distributions within the power law family.

The fits of competing distributions—the lognormal and exponential—noticeably deviate from the empirical data throughout the domain. The likelihood ratio tests in Table 2 provide

[5]For other properties of the likelihood ratio statistic, including the derivation of the corresponding p-value, see Clauset et al. (2009).

formal evidence in this regard. As is evident from large positive likelihood ratio statistics, the power law distribution significantly outperforms both the lognormal and exponential distribution in fitting COVID-19 cases in China, which is in line with our observations from Figure 2. We reject both the lognormal and exponential as an adequate specification for COVID-19 cases. In summary, our estimation results and diagnostic tests provide strong evidence that the COVID-19 cases in Chinese cities can be well characterized by the power law (Pareto) distribution.

4 Gibrat's law as a plausible explanation for power law behavior

In this section, we explore whether a growth model involving Gibrat's law (Gibrat, 1931) can potentially explain the emergence of the observed power-law behavior in COVID-19 cases in China. We focus on Gibrat's law specifically granted a random multiplicative growth (with a caveat) is the prevalent attribute of models explaining the genesis of power laws (Gabaix, 1999, 2009).

4.1 A link between power law and Gibrat's law

There are different mechanisms proposed in the literature, including the Yule process (Willis and Yule, 1922; Yule, 1925) and random growth models with geometrically distributed age distribution (Wold and Whittle, 1957; Reed, 2001; Toda, 2014; Beare and Toda, 2017), that can generate power laws.[6] In what follows, we describe a simple of such mechanisms.

Suppose S_{it} is the size of a stochastic process of interest for unit i at time t. For instance, COVID-19 cases in city i up to day t. According to Gibrat's law, the size of the process (at least in the upper tail) exhibits random multiplicative growth, evolving as

$$S_{it+1} = \mu_{it+1} S_{it} \tag{15}$$

over time, where μ_{it+1} is independently and identically distributed (i.i.d.) random variable with an associated PDF of $f(\mu)$. Hence, random growth factor $\mu_{it+1} = S_{it+1}/S_{it}$ is independent of the current size S_{it}, which is commonly known as Gibrat's law of proportionate

[6]For a detailed review, see Newman (2005).

growth. Gibrat's law alone does not generate power law but, instead, gives rise to the log-normal distribution for the size of the process (see Section 4.2 for details), which was noted by Gibrat (1931) himself early on. Later, Gabaix (1999) showed that power law can arise from Gibrat's law with an additional assumption, a sketch of which we provide below.

Let $G_t(s) = \text{Prob}(S_t > s)$ be the counter-CDF of S_t. Substituting (15) into the counter-CDF, the equation of motion for $G_{t+1}(s)$ boils down to

$$\begin{aligned} G_{t+1}(s) &= \text{Prob}(S_{t+1} > s) = \text{Prob}(\mu_{t+1} S_{it} > s) = \text{Prob}\left(S_{it} > \frac{s}{\mu_{t+1}}\right) \\ &= \int_0^\infty G_t\left(\frac{s}{\mu}\right) f(\mu) d\mu. \end{aligned} \tag{16}$$

If there is a steady state process $G_t = G$, then

$$G(s) = \int_0^\infty G\left(\frac{s}{\mu}\right) f(\mu) d\mu. \tag{17}$$

The mechanism ensuring that power law distribution is the (only) suitable steady state distribution in (17) is if S_t has lower reflecting barrier $S_{\min}$, i.e., the minimal size of the process, such that $S_t > S_{\min}$ (Gabaix, 1999, Proposition 1). In this case, $G(s) = \frac{k}{x^\gamma}$, from (8). Thus, Gibrat's law combined with a lower bound on S_t can plausibly yield power law distribution.

4.2 Testing for Gibrat's law

For empirical purposes, we consider a continuous time representation of Gibrat's law, given by geometric Brownian motion

$$dS_{it} = gS_{it}dt + \nu S_{it} dB_{it}, \tag{18}$$

where g is the expected growth rate, $\nu > 0$ is the volatility, and B_{it} is a standard Brownian motion that is i.i.d. across cross-sectional units. Applying Itô's lemma to (18) yields

$$d \ln S_{it} = (g - \nu^2/2)dt + \nu dB_{it}, \tag{19}$$

meaning the cross-sectional distribution of S_{it}, with the initial size of S_{i0}, is lognormal

$$\ln(S_{it}/S_{i0}) \sim N[(g - \nu^2/2)t, \nu^2 t]. \tag{20}$$

Equation (19) (along with Proposition 1 in Gabaix (1999)) suggests that growth rates under Gibrat's law can be described by a random walk process of the form (Sutton, 1997; Eeckhout, 2004; Gabaix, 2009)

$$\ln S_{it} = \ln S_{it-1} + \zeta_{it}. \tag{21}$$

Setting the random growth component $\zeta_{it} = \phi_i + \xi_{it}$, where ϕ_i is the effect of unit-wide factors and ξ_{it} is an i.i.d. random effect, produces a random walk with drift. A standard method to test for Gibrat's law is through estimation of the following cross-sectional regression equation

$$\ln S_{it} = \phi + \rho \ln S_{it-1} + \xi_{it}. \tag{22}$$

In (22), ρ is the parameter of interest, with $\rho \simeq 1$ providing statistical evidence that the growth process of S_t adheres to Gibrat's law.

An alternative approach for testing for Gibrat's law of proportionate growth is through estimation of the cross-sectional regression equation of the form (Beare and Toda, 2020)

$$\Delta \ln S_{it+1} = \beta_{0t} + \beta_{1t} \ln S_{it} + \beta_{2t} \Delta \ln S_{it} + \beta_3 D_{it} + e_{it}, \tag{23}$$

where Δ is the difference operator, $\Delta \ln S_{it+1}$ is the COVID-19 growth rate in city i between day t and $t+1$, $\Delta \ln S_{it}$ is the COVID-19 growth rate in city i between day $t-1$ and t, D_{it} is the number of days between day t and the day of the first COVID-19 case in city i, and e_{it} is an i.i.d. error term. The parameters of interest are $\beta_{1t}, \beta_{2t}, \beta_{3t}$, with $\beta_{1t} \simeq 0, \beta_{2t} \simeq 0, \beta_{3t} \simeq 0$ providing empirical evidence for the presence of Gibrat's law. The distinctive feature of equation (23) is the inclusion of age distribution—days since outbreak for each city—in addition to the growth rate. Obtaining age distribution has traditionally been cumbersome in power law analysis (e.g., of cities). Fortunately, our data conveniently affords us this

variable as we observe the entire timeline of the evolution of COVID-19 across Chinese cities.

4.3 Application

We apply the methods discussed in Section 4.2 to each day between January 23, 2020, and February 25, 2020 (inclusive). The reason for starting from January 23 is because at least 30 cities had a positive number of cumulative cases ($S_{it} > 0$) starting from January 23 (see Figure 3). The reason for stopping at February 25 is because COVID-19 dynamics in China had largely stabilized by February 25 ($S_{it+1} \simeq S_{it}$), with a small to zero number of new daily cases after February 25, which left the distribution of cumulative cases after February 25 virtually unaffected (see Figure 1). This will also become apparent from our findings below.

Figure 4 shows the estimation results for ρ_t in equation (22) for $t = \text{Jan } 23, \ldots, \text{Feb } 25$. Clearly, the estimates of ρ_t are statistically indistinguishable from unity ($\rho_t \simeq 1$), which confirms the random growth model predicated by Gibrat's law. The 95% confidence interval shrinks moving left to right, which can be attributed to two factors. First, it reflects increasing sample size (i.e., increasing number of cities with confirmed cases), at least until February 8, when most Chinese cities had reported a positive number of cases (Figure 3). Second, the thinning of the confidence interval can also be attributed to the stabilization of COVID-19 situation in China, which saw a rapid decline in new daily cases starting from mid-February, with the daily change (growth rate) approaching to zero.

Figure 5 reports the estimation results for $\beta_{0t}, \beta_{1t}, \beta_{2t}, \beta_{3t}$ in equation (23) for $t = \text{Jan } 23, \ldots, \text{Feb } 25$. Panels (b)-(d) contain the estimates for $\beta_{1t}, \beta_{2t}, \beta_{3t}$, which are of main interest here. It is apparent that these estimates are largely equal to zero or close to zero ($\beta_{1t} \simeq 0, \beta_{2t} \simeq 0, \beta_{3t} \simeq 0$), which indicates the growth rate between days t and $t+1$ does not depend on the number of cases on day t, nor on the growth rate between days $t-1$ and t, nor on the number of days since the first confirmed case. This also provides evidence for the presence of Gibrat's law for COVID-19 cases in Chinese cities. The estimates of β_{0t} in panel (a) show that the expected growth rate of confirmed cases declined over the study period, with some fluctuations, and eventually approached to zero around February 9, which is consistent with the observed data.

In light of the discussion in Section 4.1, the confirmation of Gibrat's law for COVID-19 cases in Chinese cities provides a plausible explanation for the emergence of power law behavior shown for the data.

5 Conclusion

The dynamics of the novel coronavirus pandemic are complex and affected by a plethora of factors, which are yet to be fully understood. In spite of the apparent chaotic evolution of the pandemic, surprising regularities can still be observed in the size distribution and growth process of COVID-19 cases. In this paper, we examined the distribution of the novel coronavirus cases in China—the original epicenter of the ongoing pandemic. We presented empirical evidence for a power law distribution for the upper tail of the number of COVID-19 cases in Chinese cities. The power law fit is robust to different estimation methods, passes rigorous diagnostic tests, and fits the data better than a number of rivaling distributions. The implications of the power law fit are that the number of COVID-19 cases in Chinese cities is heavy-tailed and disperse, so that average number of COVID-19 cases is problematic to talk about; that COVID-19 cases are concentrated within a few cities that account for a disproportionately large amount of infections; and that mean and variance are generally not finite. Admittedly, there may always be a distribution that fits the data better than a power law granted there are virtually an infinite number of distributions. What we showed here is that the power law distribution is able to capture the upper tail of the data, which it do so parsimoniously, and better than a couple 'go-to' distributions. In addition, given that the data is not lognormally distributed, we reject Gibrat's law of random proportionate growth in its standard form. However, the nuanced version of Gibrat's law (Gabaix, 1999), as discussed in Section 4.1, is demonstrated to be a plausible mechanism for the emergence of power law behavior in COVID-19 cases in China.

References

Akhundjanov, S. B. and Chamberlain, L. (2019). The Power-Law Distribution of Agricultural Land Size. *Journal of Applied Statistics*, 46(16):3044–3056.

Akhundjanov, S. B., Devadoss, S., and Luckstead, J. (2017). Size Distribution of National CO_2 Emissions. *Energy Economics*, 66:182–193.

Akhundjanov, S. B. and Toda, A. A. (2020). Is Gibrat's 'Economic Inequality' Lognormal? *Empirical Economics*, pages 1–21.

Axtell, R. L. (2001). Zipf Distribution of US Firm Sizes. *Science*, 293(5536):1818–1820.

Balthrop, A. (2016). Power Laws in Oil and Natural Gas Production. *Empirical Economics*, 51(4):1521–1539.

Balthrop, A. (2020). Gibrat's Law in the Trucking Industry. *Empirical Economics*, pages 1–16.

Barabási, A.-L. and Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512.

Battistin, E., Blundell, R., and Lewbel, A. (2009). Why is Consumption more Log Normal than Income? Gibrat's Law Revisited. *Journal of Political Economy*, 117(6):1140–1154.

Beare, B. K. and Toda, A. A. (2017). Geometrically Stopped Markovian Random Growth Processes and Pareto Tails. *arXiv preprint arXiv:1712.01431*.

Beare, B. K. and Toda, A. A. (2020). On the Emergence of a Power Law in the Distribution of COVID-19 Cases. *arXiv preprint arXiv:2004.12772*.

Blasius, B. (2020). Power-law Distribution in the Number of Confirmed COVID-19 Cases. *arXiv preprint arXiv:2004.00940*.

Carlson, J. M. and Doyle, J. (1999). Highly Optimized Tolerance: A Mechanism for Power Laws in Designed Systems. *Physical Review E*, 60(2):1412.

Champernowne, D. G. (1953). A Model of Income Distribution. *The Economic Journal*, 63(250):318–351.

China Data Lab (2020). China COVID-19 Daily Cases with Basemap. Harvard Dataverse, Available at: https://doi.org/10.7910/DVN/MR5IJN.

Clark, J. S., Fulton, M., and Brown, D. J. (1992). Gibrat's Law and Farm Growth in Canada. *Canadian Journal of Agricultural Economics*, 40(1):55–70.

Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-Law Distributions in Empirical Data. *SIAM Review*, 51(4):661–703.

Cucinotta, D. and Vanelli, M. (2020). WHO Declares COVID-19 a Pandemic. *Acta Bio-Medica: Atenei Parmensis*, 91(1):157–160.

Devadoss, S., Luckstead, J., Danforth, D., and Akhundjanov, S. (2016). The Power Law Distribution for Lower Tail Cities in India. *Physica A: Statistical Mechanics and its Applications*, 442:193–196.

Eeckhout, J. (2004). Gibrat's Law for (All) Cities. *American Economic Review*, 94(5):1429–1451.

Gabaix, X. (1999). Zipf's Law for Cities: An Explanation. *The Quarterly Journal of Economics*, 114(3):739–767.

Gabaix, X. (2009). Power Laws in Economics and Finance. *Annual Review of Economics*, 1(1):255–294.

Gabaix, X. (2016). Power Laws in Economics: An Introduction. *Journal of Economic Perspectives*, 30(1):185–206.

Gabaix, X. and Ibragimov, R. (2011). Rank - 1/2: A Simple Way to Improve the OLS Estimation of Tail Exponents. *Journal of Business & Economic Statistics*, 29(1):24–39.

Gibrat, R. (1931). *Les Inegalites Economiques*. Librairie du Recueil Sirey, Paris.

González-Val, R. (2010). The Evolution of US City Size Distribution from a Long-Term Perspective (1900–2000). *Journal of Regional Science*, 50(5):952–972.

González-Val, R. and Sanso-Navarro, M. (2010). Gibrat's law for countries. *Journal of Population Economics*, 23(4):1371–1389.

Han, Y., Liu, Y., Zhou, L., Chen, E., Liu, P., Pan, X., and Lu, Y. (2020). Epidemiological Assessment of Imported Coronavirus Disease 2019 (COVID-19) Cases in the Most Affected City Outside of Hubei Province, Wenzhou, China. *JAMA Network Open*, 3(4):e206785–e206785.

Hill, B. M. (1975). A Simple General Approach to Inference about the Tail of a Distribution. *The Annals of Statistics*, 3(5):1163–1174.

Ioannides, Y. M. and Overman, H. G. (2003). Zipf's Law for Cities: An Empirical Examination. *Regional Science and Urban Economics*, 33(2):127–137.

Irmay, S. (1997). The Relationship between Zipf's Law and the Distribution of First Digits. *Journal of Applied Statistics*, 24(4):383–394.

Keitt, T. H. and Stanley, H. E. (1998). Dynamics of North American Breeding Bird Populations. *Nature*, 393(6682):257–260.

Klass, O. S., Biham, O., Levy, M., Malcai, O., and Solomon, S. (2006). The Forbes 400 and the Pareto Wealth Distribution. *Economics Letters*, 90(2):290–295.

Krugman, P. (1996). *The Self-Organizing Economy*. Blackwell, Cambridge, MA.

Luckstead, J. and Devadoss, S. (2014). Do the World's Largest Cities Follow Zipf's and Gibrat's Laws? *Economics Letters*, 125(2):182–186.

Luttmer, E. G. J. (2007). Selection, Growth, and the Size Distribution of Firms. *The Quarterly Journal of Economics*, 122(3):1103–1144.

Mitzenmacher, M. (2004). A Brief History of Generative Models for Power Law and Lognormal Distributions. *Internet Mathematics*, 1(2):226–251.

Newman, M. E. J. (2005). Power Laws, Pareto Distributions and Zipf's Law. *Contemporary Physics*, 46(5):323–351.

Pareto, V. (1896). Cours d'économie politique professé al'université de lausanne. Vol. I.

Patel, R. D. and Schoenberg, F. P. (2011). A Graphical Test for Local Self-Similarity in Univariate Data. *Journal of Applied Statistics*, 38(11):2547–2562.

Perline, R. (2005). Strong, Weak and False Inverse Power Laws. *Statistical Science*, 20(1):68–88.

Reed, W. J. (2001). The Pareto, Zipf and Other Power Laws. *Economics Letters*, 74(1):15–19.

Rose, A. K. (2006). Cities and Countries. *Journal of Money, Credit, and Banking*, 38(8):2225–2246.

Singh, S. K. and Maddala, G. S. (1976). A Function for Size Distribution of Incomes. *Econometrica*, 44(5):963–970.

Sornette, D. (2006). *Critical Phenomena in Natural Sciences*. Springer, New York.

Stanley, M. H. R., Buldyrev, S. V., Havlin, S., Mantegna, R. N., Salinger, M. A., and Stanley, H. E. (1995). Zipf Plots and the Size Distribution of Firms. *Economics Letters*, 49(4):453–457.

Sutton, J. (1997). Gibrat's Legacy. *Journal of Economic Literature*, 35(1):40–59.

Toda, A. A. (2012). The Double Power Law in Income Distribution: Explanations and Evidence. *Journal of Economic Behavior & Organization*, 84(1):364–381.

Toda, A. A. (2014). Incomplete Market Dynamics and Cross-Ssectional Distributions. *Journal of Economic Theory*, 154:310–348.

Toda, A. A. (2017). A Note on the Size Distribution of Consumption: More Double Pareto than Lognormal. *Macroeconomic Dynamics*, 21(6):1508–1518.

Toda, A. A. and Walsh, K. (2015). The Double Power Law in Consumption and Implications for Testing Euler Equations. *Journal of Political Economy*, 123(5):1177–1200.

Willis, J. C. and Yule, G. U. (1922). Some Statistics of Evolution and Geographical Distribution in Plants and Animals, and their Significance.

Wold, H. O. A. and Whittle, P. (1957). A Model Rxplaining the Pareto Distribution of Wealth. *Econometrica*, 25(4):591–595.

Yule, G. U. (1925). A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis. *Philosophical Transactions of the Royal Society of London. Series B*, 213(402-410):21–87.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., et al. (2020). A Novel Coronavirus from Patients with Pneumonia in China, 2019. *The New England Journal of Medicine*, 382:727–733.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, MA.

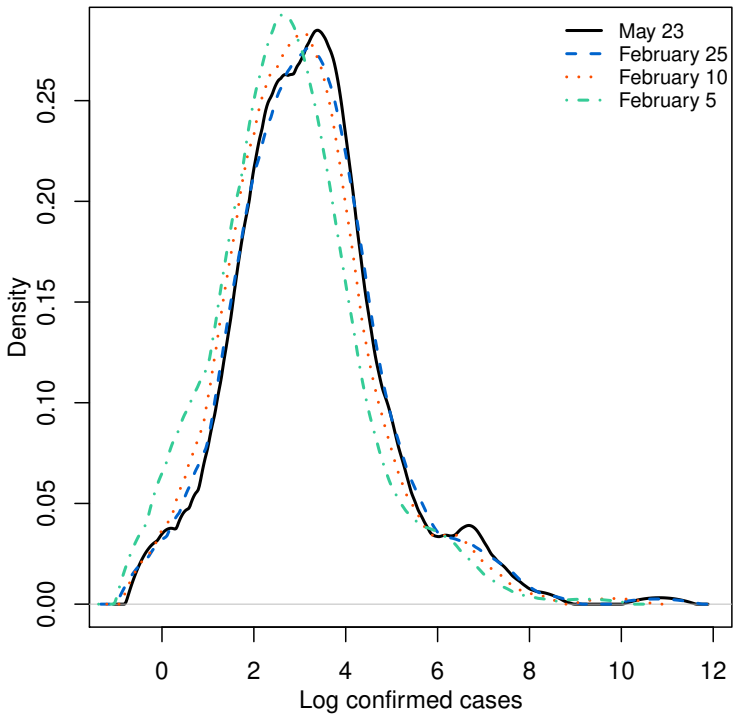


Figure 1: The empirical distribution of cumulative number of COVID-19 confirmed cases for Chinese cities. The empirical distribution is obtained using kernel density with Epanechnikov kernel and the smoothing bandwidth based on unbiased cross-validation method.

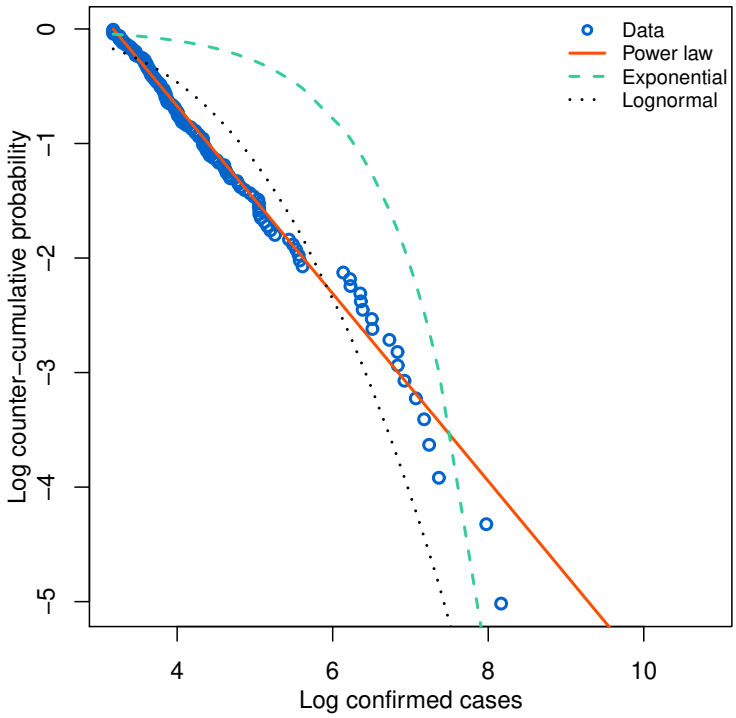


Figure 2: Plot of empirical and fitted log counter-cumulative probability and log COVID-19 confirmed cases. Estimation is based on upper-tail observations $x > x_{\min}$ as of May 23, 2020, where $x_{\min}$ is determined based on the minimization of the KS statistic. Clauset et al. (2009) recommend to have at least 50 observations for accurate power law analysis, a condition well-satisfied here.

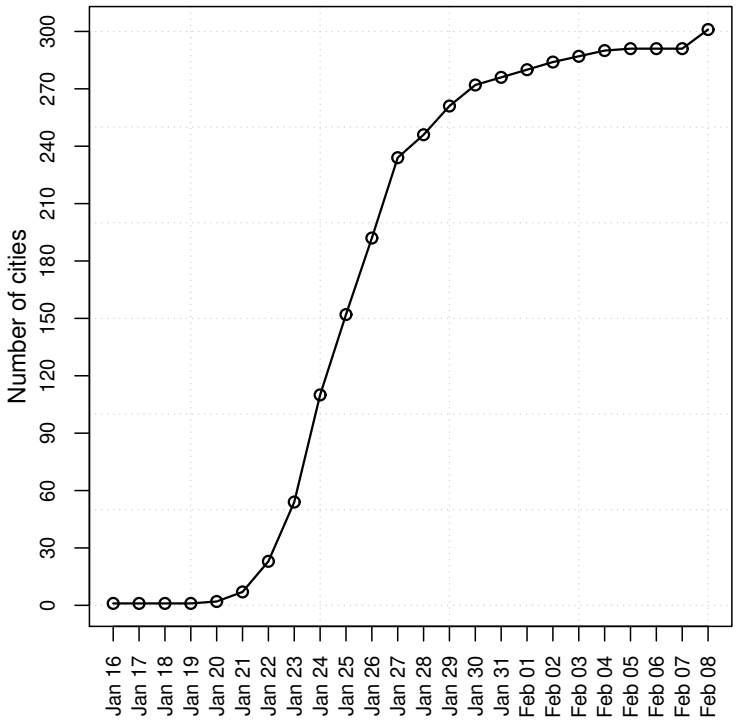


Figure 3: The number of Chinese cities with confirmed COVID-19 cases over time. By February 8, 2020, most Chinese cities had reported a positive number of cases.

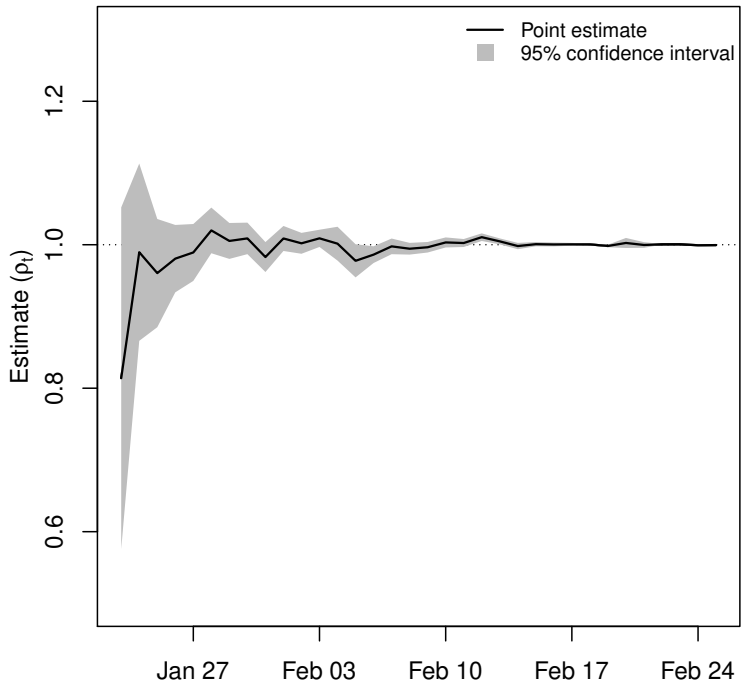
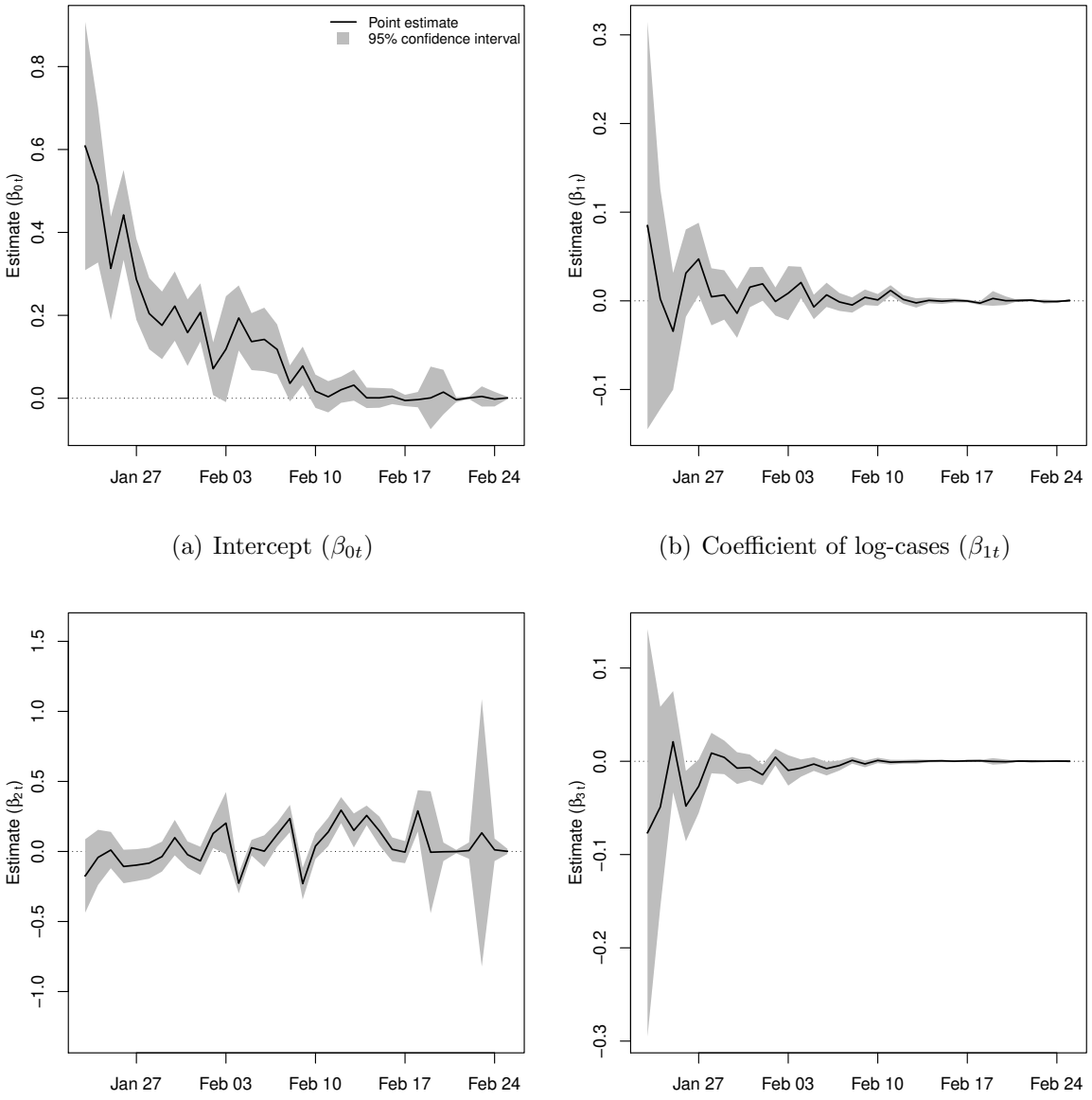


Figure 4: Estimates of ρ_t in equation (22) between January 23, 2020, and February 25, 2020, with 95% confidence bands. $\rho_t \simeq 1$ provides empirical evidence for Gibrat's law.



(a) Intercept (β_{0t})

(b) Coefficient of log-cases (β_{1t})

(c) Coefficient of log-growth rate of cases (β_{2t})

(d) Coefficient of days since first case (β_{3t})

Figure 5: Estimates of β_{0t}, β_{1t}, β_{2t}, β_{3t} in equation (23) between January 23, 2020, and February 25, 2020, with 95% confidence bands. The parameters of interest are $\beta_{1t}, \beta_{2t}, \beta_{3t}$, with $\beta_{1t} \simeq 0, \beta_{2t} \simeq 0, \beta_{3t} \simeq 0$ providing empirical evidence for Gibrat's law.

Table 1: Power law parameter estimates and goodness-of-fit test.

	Estimate	Standard Error
γ^{Hill}	0.808	(0.066)
γ^{OLS}	0.762	(0.008)
$x_{\min}$	24	
Observations ($x > x_{\min}$)	151	
Observations (total)	339	
The Gabaix and Ibragimov goodness-of-fit test		
Goodness of fit test statistic	0.013	
Goodness of fit threshold	0.112	

Note: Estimation is based on upper-tail observations $x > x_{\min}$ as of May 23, 2020, where $x_{\min}$ is determined based on the minimization of the KS statistic. For the Gabaix and Ibragimov (2011) test, the null hypothesis that COVID-19 confirmed cases is distributed according to a power law is rejected if test statistic $>$ threshold. Clauset et al. (2009) recommend to have at least 50 observations for accurate power law analysis, a condition well-satisfied here.

Table 2: Likelihood ratio tests of competing distributions.

	Likelihood ratio statistic	P-value
Power law vs. exponential	252.176	0.002
Power law vs. lognormal	329.306	0.000

Note: Estimation is based on upper-tail observations $x > x_{\min}$ as of May 23, 2020, where $x_{\min}$ is determined based on the minimization of the KS statistic. A positive value of the likelihood ratio statistic indicates that the power law is the better fitting distribution. A negative value indicates the alternative distribution fits the data more closely. P-values are calculated using the methods detailed in Clauset et al. (2009). The null hypothesis is there is no significant differences in likelihoods of the distributions tested.