Available online at www.sciencedirect.com
SCIENCE DIRECT®
Economics Letters 92 (2006) 256–262
economics letters
www.elsevier.com/locate/econbase

Is the Zipf law spurious in explaining city-size distributions?

Li Gan [a], Dong Li [b,*], Shunfeng Song [c]

[a] Department of Economics, Texas A & M University, College Station, TX 77843, USA
[b] Department of Economics, Kansas State University, Manhattan, KS 66506-4001, USA
[c] Department of Economics, University of Nevada, Reno, Reno, NV 89557-0207, USA

Received 27 January 2005; received in revised form 24 February 2006; accepted 6 March 2006
Available online 8 June 2006

Abstract

The Zipf law, which states that that the rank associated with some size S is proportional to S to some negative power, is a regularity observed in natural and social sciences. One popular application of the Zipf law is the relationship between city sizes and their ranks. This paper examines the rank–size relationship through Monte Carlo simulations and two examples. We show that a good fit (indicated by a high R^2 value) can be found for many statistical distributions. The Zipf law's good fit is a statistical phenomenon, and therefore, it does not require an economic theory that determines city-size distributions.

Keywords: Zipf law; Rank–size; Spurious

JEL classification: R1; C1

1. Introduction

One striking regularity observed by scientists and economists is the Zipf law. This law says that size distribution follows a power law: the rank associated with some size S is proportional to S to some negative power. Put differently, the rank–size relationship appears linear in the logarithms. A linear regression of log-rank on log-size yields a very high R^2. To a lesser degree, the coefficient of the log-size

* Corresponding author. Tel.: +1 785 532 4572; fax: +1 785 532 6919.
E-mail address: dongli@ksu.edu (D. Li).

0165-1765/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.econlet.2006.03.004

in the regression is often found to be close to 1. In this case, the Zipf law collapses into the so-called rank–size rule, implying that the second largest is half the size of the largest, the third largest is one-third the size of the largest, and so on.

The Zipf law has been applied in many fields. Shiode and Batty (2000) used this law to compare the statistical patterns of size and connectivity of global domains to the geographical distribution of the global population. Sinclair (2001) incorporated the Zipf law into his model for analyzing world income distribution. Li and Yang (2002) applied it to their study of how the single-gene classification ability decreases with the rank of the genes. Tachimori and Tahara (2002) found an inverse-power relationship between the rank order of diagnoses and the frequency of the appearance of these diagnoses.

Probably the most popular application of the Zipf law is in urban economics. Numerous studies have applied it to urban data in the United States and other countries. Previous studies have consistently shown a surprisingly high R^2-value when the Zipf law has been used to describe city-size distributions in many countries and at different times in the same country. For example, Rosen and Resnick (1980) used data from 44 countries and found that R^2-values were above 0.95 for 36 countries, with only Thailand having an R^2-value lower than 0.9 (0.83). Using 1990 data on 366 U.S. urbanized areas, Mills and Hamilton (1994) found an R^2-value of 0.99. Guerin-Pace (1995) used data on French cities from 1831 to 1982 and found that the R^2-value was always higher than 0.99 when a threshold of 2000 was used. Song and Zhang (2002) found an R^2-value of 0.92 for 665 Chinese cities in 1998. Given all these findings, it is not surprising that Krugman (1996, p. 40) claimed that "we are unused to seeing regularities this exact in economics—it is so exact that I find it spooky."

Compared with the regularity of high R^2-values in these studies, more variations in the coefficient of log-size can be found in the literature. In the urban literature, the coefficient is very close to 1 if the U.S. city data are used (Gabaix, 1999). However, in Rosen and Resnick (1980), the coefficient ranges from 0.809 for cities in Morocco and 1.963 for cities in Australia. In other studies, the R^2 is found to be high but the coefficient is not necessarily close to 1. For example, Shiode and Batty (2000) find that the coefficient is 2.91, while R^2 is 0.90 if the number of web pages in different countries is used, and the coefficient is 1.60 and the R^2 is 0.92 if total links in different countries is used.

Several studies have attempted to derive the Zipf law for city-size distributions. Gabaix (1999), for instance, theoretically proves that city-size distribution converges to a power law if all cities follow some proportional growth process (i.e., the growth of cities has the same mean and same variance). However, no consensus has been reached. Fujita et al. (1999, p. 219) stated that "the regularity of the urban size distribution poses a real puzzle, one that neither our approach nor the most plausible alternative approach to city sizes seems to answer." Krugman (1995, p. 44) called the rank–size rule "a major embarrassment for economic theory: one of the strongest statistical relationships we know, lacking any clear basis in theory."

Does the Zipf law really suggest an economic regularity? Is the Zipf law spurious in explaining city-size distributions? The economic literature has already indicated that a high degree of explanatory power between dependent and independent variables may not necessarily result from some kind of economic relationship, but rather through statistical properties of the series (e.g., Yule, 1926; Phillips, 1986). To answer the above questions, this study uses Monte Carlo simulations to examine rank–size relationships by running regressions on random numbers (and their ranks) generated from various probability distributions. Since the independent variable is randomly generated without reference to any economic

theory, a high R^2-value would suggest that the Zipf law is a statistical phenomenon and thus spurious in explaining city-size distributions.

The rest of this paper is organized as follows. In Section 2, we specify the regression model of the Zipf law and conduct a Monte Carlo study. In Section 3, we apply the Zipf law to U.S. and Chinese cities and perform Kolmogorov–Smirnov nonparametric tests to determine whether U.S. and Chinese cities follow the Pareto distribution. We present our conclusions in Section 4.

2. The model and simulation results

The Zipf law of city-size distribution can be written as

$$R_i = AS_i^{-\beta}, \tag{1}$$

where R_i is the rank of the ith city, S_i is this city's size, and A is a constant term. Accordingly, it is represented by the following regression model,

$$\log(R_i) = \alpha - \beta\log(S_i) + \varepsilon \tag{2}$$

where $\alpha = \log(A)$ and ε is the random error term.

Generally, in a regression model, the values of the dependent variable are not generated based on the values of the explanatory variable(s). They come with data, along with information on the explanatory variable(s). In other words, the values of the dependent variable are allowed to vary even when the values of all the independent variables are given, regardless of the existence of some relationship between the dependent and independent variables. In fact, this is the reason why we run a regression to determine whether or not the dependent variable is related to the independent variables.

The values of the dependent variable in the Zipf law, R or $\log(R)$, however, are not observed but are generated from S. With such a specification of the dependent variable, we expect that $\log(R)$ and $\log(S)$ are highly correlated, and a high degree of explanatory power is a statistical property of the series. Put differently, the Zipf law could be spurious and does not necessarily require a clear basis in economic theory that determines city-size distribution.

To show whether the Zipf law is spurious or not, and to prove whether or not its high degree of explanatory power is a statistical property of the series, in this paper, we use Monte Carlo simulations to examine the rank–size relationships by running regressions on random numbers (and their ranks) generated from various probability distributions. We conduct the simulations in three steps.

At Step 1, we randomly generate a sequence of random variables Y and a sequence of random variables S. We let Y be uniformly distributed in the interval (0,1), and let S come from various probability densities. We conduct a simple regression of Y on S.

At Step 2, we sort the sequences Y and S. We denote the sorted sequences as Y^s and S^s, respectively, and match the sort order of Y^s and the sort order of S^s. Then we conduct a regression of Y^s on S^s. To minimize the impact of outliers, we also conduct a regression of $\ln(Y^s)$ on $\ln(S^s)$ at this step.

At Step 3, we run a Zipf-Law type of regression on Eq. (2). We use the rank order of S^s, denoted as R, to run a regression on S^s itself. In particular, $R=1$ if it is the largest value of S^s, and $R=$nobs if it is the smallest value of S^s and the total number of draws in S^s is nobs. Note the rank order of S^s is the same as the rank order of Y^s. Here, we are interested in the coefficient of S^s, the R^2 of the regression, and finally, the improvement of R^2 from Step 2.

We generate the random variable Y based on the uniform distribution, and the random variable S from various probability distributions with different distribution parameters. The probability distributions we are considering here include normal, log-normal, negative exponential (Pareto), and gamma distributions. For each distributional setup, we carry out 1000 simulations with 100 observations for each replication.

Since the random number sequence Y and S is independently drawn, we should expect no relationship between the two sequences. The coefficient of S, denoted as β, is expected to be almost zero, and R^2 of the regression should be very low. However, we expect that the simple sorting at Step 2 should help to establish some statistical relationships between the two sequences. We are interested in determining if the coefficient of S^s is significant, and what the value of R^2 is. But note that any relationship which may emerge here comes from a simple sorting. The final step replaces Y by its rank ordering. Since Y and S are not generated based on any economic theory, high R^2 values for all probability distributions of S suggest that the high degree of explanatory power of the Zipf law is a pure statistical phenomenon. If this is the case, the Zipf law would be spurious in explaining city-size distribution, a strong statistical regularity that does not need an economic theory.

Table 1 presents the simulation results for the mean estimated values of β, and the R^2-values for various probability distributions in all three steps. As expected, the coefficient β and the R^2 values at Step 1 are all close to zero. Not surprisingly, sorting the two sequences does reveal some relationships between the two sequences. However, no clear pattern emerges at Step 2. For some densities, such as the Normal density or Gamma density, a simple sorting improves R^2 dramatically. For other densities, sorting does not improve R^2-values nearly as much.

More importantly, a rank–size regression (Zipf law) improves R^2-values for all density functions. In particular, we have three observations to make at Step 3. First and most importantly, high R^2-values are obtained in all densities reported in Table 1. Specifically, the results show an average R^2-value of 0.79 or higher for normal distributions, 0.84 or higher for log-normal distributions, 0.98 for negative exponential distributions, and 0.65 or higher for gamma distributions. Since the random numbers are not generated based on any economic theory, these results indicate that the high explanatory power of the Zipf law is a statistical phenomenon that results from the specification of the dependent variable. Therefore, the Zipf law could be spurious if it is used to explain city-size distributions. Second, the Zipf law performs best for Pareto distributions. Higher R^2-values for Pareto distributions, however, are expected. In fact, it can be shown statistically that the R^2-value asymptotically approaches 1 if an order series is independent and identically distributed according to a Pareto distribution (proof is available upon request). We also explore some distributions that are "almost" Pareto by using a mixture.[1] Each observation is drawn from two distributions with probabilities P and $(1-P)$, respectively. Our simulations show that a mixture of a

[1] We thank Kenneth A. Small for suggesting the mixture simulations.

Table 1
Simulation Results[a,b]

Model	Step 1		Step 2: Y^s=Sorted Y, S^s=Sorted S				Step 3: R=rank of S^s	
	$Y=\alpha+\beta S$		$Y^s=\alpha+\beta S^s$		$\ln(Y^s)=\alpha+\beta\ln(S^s)$		$\ln(R)=\alpha+\beta\ln(S^s)$	
	β	R^2	β	R^2	β	R^2	β	R^2
Distribution of S: Normal (μ,σ)								
(10,1)	−.001	.0106	.283	.943	.899	.807	8.22	.788
	(.030)	(.016)	(.023)	(.024)	(.130)	(.067)	(.73)	(.046)
(100,4)	−.0003	.0106	.0709	.943	.225	.807	21.17	.819
	(.007)	(.016)	(.006)	(.024)	(.032)	(.067)	(1.69)	(.042)
Distribution of S: Log-Normal (μ,σ)								
(0,1)	−.0004	.0099	.119	.570	.299	.306	.858	.838
	(.016)	(.013)	(.042)	(.138)	(.117)	(.094)	(.066)	(.040)
(0,5)	3.4e−8	.0104	1.44e−5	.073	3.0e−5	.027	.175	.850
	(1.3e−5)	(.011)	(3.6e−5)	(.043)	(7.6e−5)	(.017)	(.012)	(.035)
(0,0.5)	−.0016	.0101	.447	.817	1.243	.054	1.716	.838
	(.050)	(.015)	(.074)	(.083)	(.266)	(.089)	(.132)	(.039)
(2,1)	−5.3e−5	.0099	.0161	.570	.040	.306	.858	.838
	(.002)	(.013)	(.006)	(.138)	(.016)	(.094)	(.066)	(.40)
Distribution of S: Pareto								
(0.5,1)	1.9e−4	.0100	.023	.208	.052	.090	.947	.976
	(.006)	(.012)	(.025)	(.127)	(.059)	(.062)	(.136)	(.023)
(1,1)	9.9e−5	.0100	.0115	.208	.0260	.090	.947	.976
	(.003)	(.012)	(.013)	(.127)	(.030)	(.062)	(.136)	(.024)
(3,1)	3.3e−5	.0100	.0038	.208	.0087	.090	.947	.976
	(9.7e−4)	(.012)	(.0042)	(.127)	(.010)	(.062)	(.136)	(.023)
(1,0.5)	3.9e−6	.0099	1.8e−4	.084	3.9e−4	.032	.474	.976
	(1.2e−4)	(.011)	(4.3e−4)	(.059)	(9.3e−4)	(.024)	(.067)	(.023)
(1,3)	7.3e−4	.010	.324	.530	.800	.276	2.842	.976
	(.046)	(.014)	(.134)	(.145)	(.360)	(.094)	(.407)	(.023)
(3,4)	3.4e−4	.010	.184	.590	.460	.316	3.789	.976
	(.025)	(.014)	(.063)	(.132)	(.174)	(.091)	(.542)	(.023)
Distribution of S: Gamma								
(1,1)	−.0011	.0100	.259	.758	.675	.442	.596	.652
	(.030)	(.014)	(.044)	(.080)	(.142)	(.079)	(.082)	(.061)
(2,1)	−3.2e−4	.0099	.190	.841	.523	.546	.986	.710
	(.021)	(.013)	(.028)	(.067)	(.102)	(.083)	(.110)	(.056)
(3,1)	−2.6e−4	.0103	.158	.875	.446	.597	1.281	.737
	(.017)	(.014)	(.019)	(.054)	(.076)	(.081)	(.129)	(.050)

[a] Results are based on 1000 simulations with 100 observations for each replication.
[b] Standard deviations are in parentheses.

Pareto distribution with $P=0.85$ or more and another distribution performs as well as a "pure" Pareto distribution. Third, the coefficient, β, does not equal 1, although it is close to 1 for some distributions. Our simulation results show that the estimated values of β are sensitive to the parameters of distributions.

3. Empirical cases and tests: the city-size distribution in the United States and China

To illustrate the problem with the Zipf law, we apply the law to real data. We choose U.S. urbanized area data from 1990 and 2000 and Chinese city data from 1985 and 1999. The former is for a developed country; the latter is for a developing country. Both countries have many cities and, thus, are ideal subjects for city-size distribution analysis.

Table 2 shows the regression results. We reach two conclusions. First, for city-size distributions in the U.S., we have an R^2-value of 0.989 for both 1990 and 2000. For those distributions in China, we have an R^2-value of 0.857 for 1985 and 0.927 for 1999. Based on R^2 values, we conclude that the Zipf law fits well for both U.S. and Chinese city-size distributions. Second, although estimated values of β are close to 1 numerically (0.867–1.075), their corresponding standard deviations suggest that β is not equal to 1 statistically for all four cases. This finding indicates that the rank–size rule does not hold for U.S. and Chinese city-size distributions in the years studied in this paper.

Does a high R^2-value imply that city sizes follow a power law (Pareto distribution)? To answer this question, we perform the Kolmogorov–Smirnov nonparametric test. The Kolmogorov–Smirnov test checks the equality of distributions. By comparing the empirical distribution of the data with a given distribution, we are able to see if it is reasonable to conclude that the observations are drawn from the underlined distribution. In the U.S. cases, our results show that the 1990 data do not reject a Pareto distribution with $p=0.943$ but reject all other distributions (normal, log-normal, and gamma) with $p=0.000$, and the 2000 data do not reject a Pareto distribution with $p=0.649$ but reject all other distributions with $p=0.000$. In the Chinese cases, both the 1985 and 1999 data reject every distribution with $p=0.000$, suggesting that city-size distributions in China do not follow a Pareto distribution, even though the Zipf law performs very well on them. Therefore, the Zipf law's high degree of explanatory power does not necessarily imply that city sizes follow a power law.

4. Conclusions

In this paper, we have investigated the rank–size relationship by running regressions on random numbers generated from various probability distributions. Our results show a high R^2-value for most distributions. Specifically, we have obtained an average R^2-value of 0.79 or higher for normal distributions, 0.84 or higher for log-normal distributions, 0.98 for negative exponential distributions, and 0.65 or higher for gamma distributions. Since our independent variables are generated randomly without reference to any economic theory, and the dependent variables are generated according to the Zipf law after we sorted the random numbers, our results suggest that the high degree of explanatory power of the

Table 2
Regression results on city-size distribution in the United States and China

Nation	Year	α	β	R^2-value	Number of observations
United States	1990	15.717 (0.057)	0.895 (0.005)	0.989	396
United States	2000	15.681 (0.052)	0.879 (0.004)	0.989	452
China	1985	11.283 (0.149)	0.856 (0.020)	0.857	324
China	1999	13.684 (0.089)	1.075 (0.012)	0.927	667

Data sources: U.S. Bureau of Census and Urban Statistical Yearbook of China (1986, 2000).

Zipf law largely results from the variable and model specifications. Therefore, the Zipf law is a statistical phenomenon.

When we applied the Zipf law to U.S. and Chinese data, we found that it fits very well for city-size distributions in both countries. For U.S. city-size distributions, we obtained an R^2-value of 0.989 for 1990 and 2000. For Chinese city-size distributions, we had an R^2-value of 0.857 for 1985 and 0.927 for 1999. By performing the Kolmogorov–Smirnov nonparametric test, we could not reject the hypothesis that U.S. cities follow a Pareto distribution. However, we have strongly rejected the hypothesis that Chinese cities follow a Pareto distribution. This finding suggests that a high R^2-value does not necessarily imply that city sizes follow a power law.

The finding of a high R^2-value for randomly generated series indicates that the Zipf law does not need a basis in economic theory to show a high degree of explanatory power. Such statistical regularity exists in regression models in which the dependent variable is defined by ranks of the independent variable. The Zipf law's good fit is a statistical phenomenon; it does not require an economic theory that determines city-size distributions. Therefore, urban economists no longer need to consider the regularity of the urban size distribution as a puzzle and an embarrassment.

Acknowledgements

The authors thank Jianhua Huang, Kenneth A. Small and an anonymous referee for helpful comments. All remaining errors are ours.

References

Fujita, M., Krugman, P., Venables, A.J., 1999. The Spatial Economy. The MIT Press, Cambridge, MA.

Gabaix, X., 1999. Zipf's law for cities: an explanation. Quarterly Journal of Economics CXIV (3), 739–767.

Guerin-Pace, F., 1995. Rank–size distribution and the process of urban growth. Urban Studies 32 (3), 551–562.

Krugman, K., 1995. Development, Geography, and Economic Theory. The MIT Press, Cambridge, MA.

Krugman, K., 1996. The Self-Organizing Economy. Blackwell Publishing Inc., Malden, MA.

Li, W., Yang, Y., 2002. Zipf's law in importance of genes for cancer classification using microarray data. Theoretical Biology 219 (4), 539–551.

Mills, E.S., Hamilton, B.W., 1994. Urban Economics, 5th ed. Harper Collins College Publishers, New York.

Phillips, P.C.B., 1986. Understanding spurious regressions. Journal of Econometrics 33, 311–340.

Rosen, K.T., Resnick, M., 1980. The size distribution of cities: an examination of the Pareto law and primacy. Journal of Urban Economics 8, 165–186.

Shiode, N., Batty, M., 2000. Power Law Distributions in Real and Virtual Worlds (http://www.isoc.org/inet2000/cdproceedings/2a/2a_2.htm, also presented at INET 2000, Yokohama, Japan, 2000).

Sinclair, R., 2001. Examining the Growth Model's Implications: The World Income Distribution, Working Paper. Department of Economics, Syracuse University.

Song, S., Zhang, K.H., 2002. Urbanization and city size distribution in China. Urban Studies 39 (12), 2317–2327.

Tachimori, Y., Takashi, T., 2002. Clinical diagnoses following Zipf's Law. Fractals 10 (3), 341–351.

Yule, G.U., 1926. Why do we sometimes get nonsense correlations between time series? A study in sampling and the nature of time series. Journal of the Royal Statistical Society 89, 1–64.