Nota, Fungisai; Song, Shunfeng

Article

Further analysis of the Zipf's law: Does the rank-size rule really exist?

Journal of Urban Management

Provided in Cooperation with:
Chinese Association of Urban Management (CAUM), Taipei

Suggested Citation: Nota, Fungisai; Song, Shunfeng (2012) : Further analysis of the Zipf's law: Does the rank-size rule really exist?, Journal of Urban Management, ISSN 2226-5856, Elsevier, Amsterdam, Vol. 1, Iss. 2, pp. 19-31, https://doi.org/10.1016/S2226-5856(18)30058-X

This Version is available at:
https://hdl.handle.net/10419/194395

Further Analysis of the Zipf's Law: Does the Rank-Size Rule Really Exist?

Fungisai Nota
Department of Business Administration and Economics, Wartburg College, Waverly, IA 50677, U. S. A.; Tel: 1-319-693-9084; Email: fungisai.nota@wartburg.edu

Shunfeng Song
Department of Economics, University of Nevada, Reno, NV 89557, School of Economics, Zhejiang University, Hangzhou, Zhejiang, China; Tel: 1-775-785-6860; Email: song@unr.edu

ABSTRACT. Zipf's law has two striking regularities: excellent fit and an exponent close to 1.0. When the exponent equals 1.0, Zipf's law collapses into the rank-size rule. This paper alters the sample size, the truncation point, and the mix of cities in the sample to analyze the Zipf exponent. Our results demonstrate that the exponent is close to 1.0 only for a number of selected sub-samples. Small samples of large cities provide higher values, while samples of small cities produce lower values. Using the estimated values of the exponent derived from the rolling sample method revealed elasticity in the exponent with regard to sample size. Our results also suggest that the rank-size rule should be interpreted with caution. Although it is well-known and commonly used, the rank-size rule may be more of a statistical phenomenon than an economic regularity.

KEYWORDS. ***Zipf's law, rank-size rule, rolling sample method***

1. INTRODUCTION

Zipf's law describes an empirical regularity observed in both the natural and social sciences (e.g., Zipf, 1949; Shiode and Batty, 2000; Sinclair, 2001; Li and Yang, 2002; Tachimori and Tahara, 2002). It states that the rank associated with a given size *S* is inversely proportional to *S* at a given power. If this power is equal to one, Zipf's law collapses into what is commonly called the rank-size rule. This implies that in the case of cities, the second largest city is one-half the size of the first and the third largest city is one-third the size of the largest and so on. In cases where the power is greater than one, Zipf's law suggests that the second largest city is more than half as large as the largest city and the third largest city is more than a third as large as the largest city, and so on. Conversely, an exponent of less than one would suggest that the second largest city is less than half the size of the largest city, and so on. Linearizing the relationship between rank and size using a log transformation can facilitate the estimation of the negative exponent.

One of the striking characteristics of Zipf's law is its excellent fit. Numerous empirical studies have shown that a linear regression of log-rank on log-size generates an excellent fit (very high R2-value). Using data from 44 countries, Rosen and Resnick (1980) found that R2-values were above 0.95 for 36 countries and only Thailand had an R2-value below 0.9 (0.83). Mills and Hamilton (1994) obtained an R2-value of 0.99 using 1990 data on 366 urbanized areas in the U.S. Song and Zhang (2002) obtained an R2-value of 0.91 for 665 Chinese cities in 1998. This astonishing regularity led Krugman (1995, p.44) to claim that the rank-size rule is "a major embarrassment for economic theory: one of the strongest statistical phenomenon we know, lacking any clear basis in theory." Fujita et al. (1999, p. 219) stated "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."

The second notable observation is related to Zipf's coefficient. In studies related to urban development, the coefficient is very close to 1.0, thus the rank-size rule holds. Gabaix (1999a, b) argued that the rank-size rule is theoretically a natural result of urban growth independent of the initial size of the city. Fujita et al. (1999) suggested that the rank-size rule does indeed approximate the long-run spatial distribution of a mature spatial system. Among the 44 countries empirically studied by Rosen and Resnick (1980), the estimated coefficient ranged from 0.809 for cities in Morocco to 1.963 for cities in Australia. Nitsche (2005) analyzed 515 estimates from 29 studies related to the rank-size relationship and found that two-thirds of the estimated coefficients were between 0.80 and 1.20, with a median estimate of 1.09. This implies that city-size distributions tend to be more even than what is suggested by the rank-size rule.

Some have used economic theory to explain why Zipf's law holds, others cited Gibrat's law, and still others counted it as a purely statistical phenomenon. The economic explanation relies

on a delicate balance between transportation costs, positive and negative externalities, and differences in productivity (Gabaix, 1999b; Fujita et al., 1999). However, this approach poses a number of inherent difficulties. For example, it is difficult to imagine how radically different economies, such as those of the U.S., China, and India, could maintain the same delicate balance of forces over time. Gabaix (1999a) proved that Zipf's law can be derived from Gibrat's law, as a matter of steady state distribution. "The existence of power law can be thought as due to a simple principle: the scale invariance. Because the growth process is the same at all scales, the final distribution process should be scale invariant. This forces it to be a power law" Gabaix (1999a, p. 744). A more recent explanation was provided by L. Gan et al. (2006), who proved that a high R2-value exists because the dependent variable (rank) is generated from the independent variable (size). Using randomly generated data as well as data from China and the U.S., they concluded that Zipf's law does not need a basis in economic theory to show a high degree of explanatory power. In other words, Zipf's law is a statistical phenomenon. This explanation is supported by the fact that Zipf's law holds in many other cases, such as firm size and web server domains.

Despite these attempts at an explanation, the tendency of Zipf coefficient to stay so close to 1.0 remains puzzling. Is it an economic regularity or statistical phenomenon? In an attempt to solve this puzzle, this study uses data from Chinese cities (1985 and 1999) and U.S. urbanized areas (1980, 1990 and 2000) to identify the factors driving the distribution of the estimated coefficient. We selected the U.S. and China for two reasons. First, both countries have many cities, and a larger number of cities allows for more rigorous statistical analysis. Second, they have very different economic systems, which could help to indicate whether the Zipf coefficient is sensitive to economic factors.

Gaining a better understanding of the Zipf exponent is no trivial matter. The validity of the rank-size rule hinges on this exponent having a value close to 1. If the value of this exponent fluctuated very far from 1, the rule could be put in jeopardy and support the assertion by Gan et al. (2006) that Zipf's law is merely a statistical phenomenon, rather than an economic regularity. Using rolling sample methods with and without replacement as well as Monte Carlo simulation, this study determined how sample size affects the Zipf exponent. To the best of our knowledge, this is the first study to provide a systematic examination of the relationship between the Zipf exponent and the size of samples used in regression.

In the following section, we outline the methodologies used in this analysis, including the rolling sample method and the random rolling sample with replacement. The third section provides our results. The final section summarizes the empirical results and discusses their economic significance. In short, this paper is an attempt to determine the impact of sample size, truncation point, and the mix of cities on the estimated exponent of Zipf's law.

2. THE MODEL AND METHODOLOGY

Before outlining the model, it might be helpful to visualize Zipf's law. In accordance with the example provided by Gabaix (1999a), we can take a country such as the U.S. and order the cities by population: No. 1, New York; No. 2, Los Angeles, and so on. We then draw a graph in which the y-axis is the log of rank and the x-axis is the log of population. This provides a straight line, as shown in Fig. 1, with a slope close to -1.

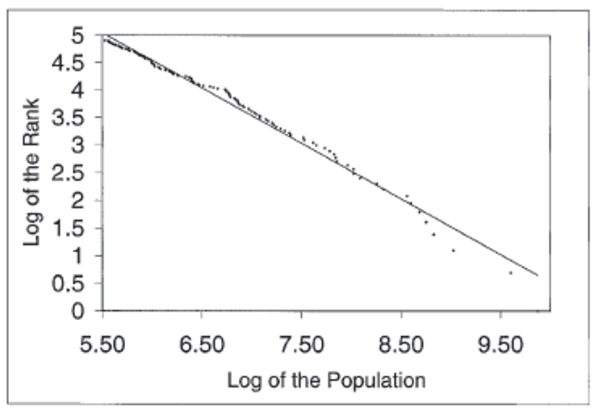

Figure 1. Log Size versus Log Rank for the 140 Largest U.S. Cities in 1990

The most common approach to estimating the slope in Fig. 1 is as follows:

$$\log(R_i) = \alpha - \beta \log(S_i) + \varepsilon \quad (1)$$

where R_i is the rank of the i^{th} city, S_i is the size of the city, and α is a constant term. Equation (1) is a log transformation of the regular Zipf law, which is generally expressed as Eq. (2):

$$R_i = AS_i^{-\beta} \quad (2)$$

Although Eq. (1) is a popular means of estimating the Zipf exponent, Gabaix and Ibragimov (2011) showed that the model produces bias when using small samples. Specifically, they determined that the standard error related to the Zipf exponent is not OLS standard error, but rather asymptotically $(2/\bar{n}_i)^{0.5}\hat{\beta}$, where $\bar{n}_i$ is the corresponding sample size. They further demonstrated that a shift of 1/2 for the rank is optimal to correct this bias. Therefore, it is preferable to make these estimations using Eq. (3), rather than Eq. (1):

$$\log(R_i - 0.5) = \alpha - \beta \log(S_i) + \varepsilon \quad (3)$$

This corrected version is known as the rank minus half rule. Throughout this paper, we provide results using both calculation methods and comment on the differences between them.

The main purpose of this research is to investigate how sample size influences the Zipf

exponent. The first method we employ is the rolling sample method. This involves estimating exponent coefficient β using OLS, and repeating the estimation process using a moving truncation point. The start point of each sub-sample is fixed at the largest city and the truncation point moves down by one city at a time, thereby increasing the sub-sample size by one each time. For example, the full sample size of urbanized areas in the U.S for 1990 is 396. These urbanized areas are ordered from the largest urbanized area, New York (16,044,012 persons), to the smallest urbanized area, Brunswick, GA (50,066 persons). The first sub-sample size of regressions (1) and (3) is $\bar{n}_i$, the 10 largest cities for example. Thus, the second sub-sample is $\bar{n}_2 = \bar{n}_1 + 1$, comprising the 11 largest cities. The third sub-sample size is $\bar{n}_3 = \bar{n}_2 + 1$, or the 12 largest cities; and so on. This process is continued until the last sub-sample is equal to the full sample size of 396.

The advantage of the rolling sample method is that it constructs the distribution of the Zipf exponent with respect to sample size; an approach commonly used in the literature. Nearly all previous studies on Zipf's law have listed the largest cities, such as the top 50 cities in Rosen and Resnick (1980), the top 140 cities in Gabaix (1999a), and full U.S. and China samples in Gan et al. (2006). We believe that adopting rolling samples, including large sample sizes as well as smaller ones, will provide more intuitive results. This approach will also help to reveal how the Zipf exponent responds to changes in sample size when smaller cities are added one by one. However, one disadvantage of the rolling sample method is its inability to untangle the pure effect of sample size on the Zipf exponent. This is because the rolling sample method simultaneously captures the truncation point effect and sample size effect as well as variations in the size of cities.

To more precisely examine the relationship between sample size and the estimated Zipf exponent, this study employed the random rolling sample method with replacement. We first determined the sample size $\tilde{n}_1$ ($\tilde{n}_1$=5 in this research). From the full sample, we then randomly selected $\tilde{n}_1$ cities. The selected cities were then ranked and the Zipf exponent was estimated using Eqs. (1) and (3). We then repeated this process 100 times to obtain an average estimated Zipf exponent for $\tilde{n}_1$. We increased the sample sizes one by one ($\tilde{n}_2 = \tilde{n}_1$+1, and so on) and repeated the above four steps for each new sample size. All samples were selected independently and randomly, and 100 regressions were run for each sample size; therefore, we were able to show how sample size affects the Zipf exponent.

We also employed Monte Carlo simulation to verify the robustness of our results from the random rolling sample method described above. For the Monte Carlo simulation, we randomly generated 1000 numbers from a normal distribution, instead of using actual data related to urbanized areas in the U.S. We then applied the random rolling sample method and examined the relationship between the average estimated Zipf exponent and sample size.

3. EMPIRICAL RESULTS

Table 1 shows the full-sample results related to Zipf's law. As expected, all cities in both countries (in all years) provided high R^2-values (0.857 to 0.989). It is interesting to note that the estimated coefficients ($\hat{\beta}$'s) are slightly higher for the OLS bias corrected model than the original uncorrected version. Thus, we conclude that the uncorrected Zipf's law (Eq. 1) has a downward bias on the estimated coefficient.

Table 1. Zipf's Law Regression Results Using City Size Data from China and the U.S.

Nation	Year	$\hat{\beta}$	OLS Bias Corrected $\hat{\beta}$	R^2 (from unadjusted)	Sample Size
U.S.	1980	0.91	0.925	0.989	366
U.S.	1990	0.895	0.913	0.989	396
U.S.	2000	0.875	0.895	0.989	452
China	1985	0.856	0.875	0.857	324
China	1999	1.075	1.09	0.927	667

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

3.1 Results from the Rolling Sample Method

Figures 2-6 present the results from the rolling sample method, revealing several interesting findings. First, a negative relationship exists between the estimated Zipf exponent and sample size, for both countries in all sample years. This implies that the estimated Zipf exponent depends on the size of samples used in the regression. Small samples of large cities yield higher coefficients than large samples that include smaller cities. In fact, Figs. 2-4 even suggest that the estimated exponent follows a lognormal distribution with respect to sample size. This is tested in Section 3.4.

Second, all solid curves are located above the dashed curves, indicating that the unadjusted regression model has a downward bias on the estimated coefficient, particularly for small samples. This is consistent with the findings of Gabaix and Ibragimov (2011).

Third, the estimated Zipf exponent is generally higher for Chinese cities than for urbanized areas in the U.S., even though the sample size is larger for China. This suggests that cities in China are more evenly distributed than those in the U.S. One explanation may be China's long-time policy of strict control of the large-sized cities, reasonable development of the medium-sized cities, and aggressive development of the small-sized cities.

Fourth, the rank-size rule (i.e., β=1) holds only for a selected range of sample sizes. For urbanized areas in the U.S, the 95% confidence interval includes β=1 when using a sample size of between 180 and 205 in 1980, between 140 and 195 in 1990; and between 140 and 205 in 2000. For Chinese cities, the rank-size rule holds only when the sample includes between 315 and 320 cities in 1985; all estimated Zipf exponents are statistically greater than 1.0 in 1999. This suggests that the well-known rank-size rule for city-size distribution is not a uniform phenomenon.

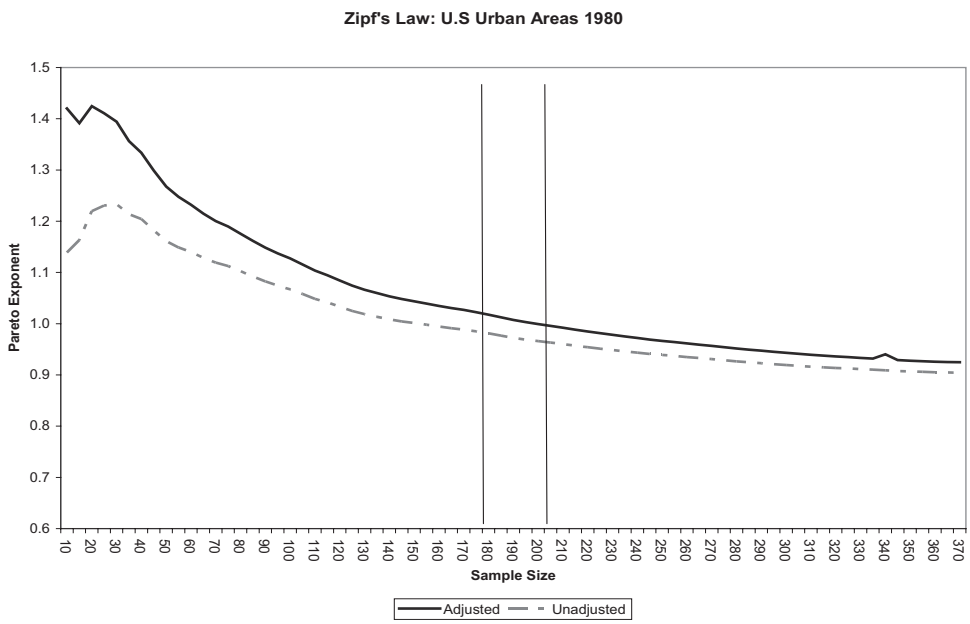

Figure 2. Zipf's Law: Urbanized Areas in the U.S in 1980

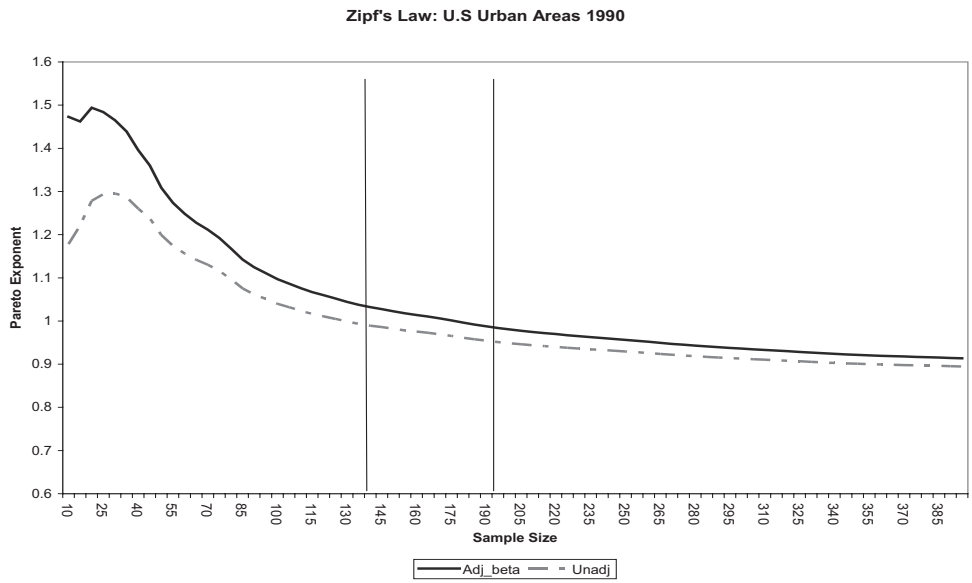

Figure 3. Zipf's Law: Urbanized Areas in the U.S in 1990

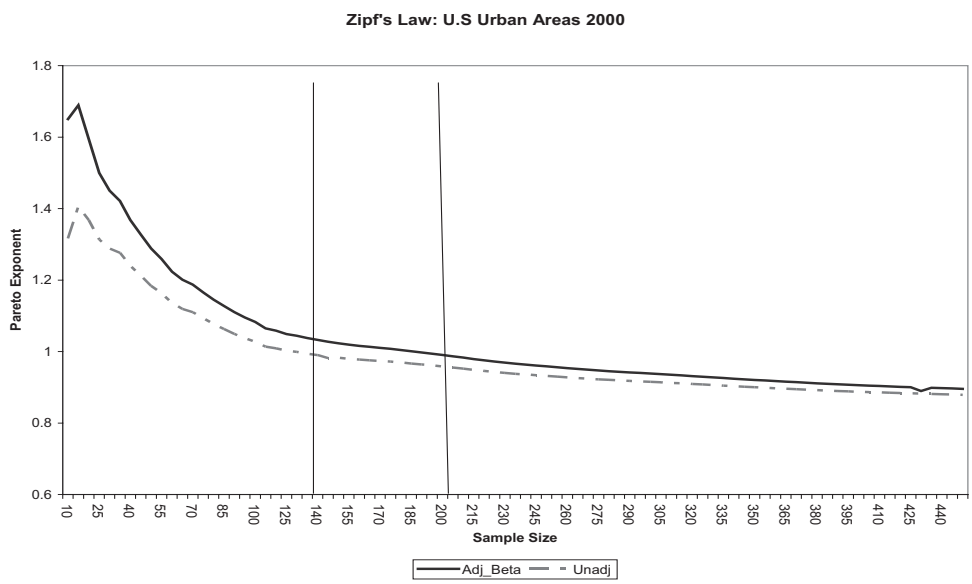

Figure 4. Zipf's Law: Urbanized Areas in the U.S in 2000

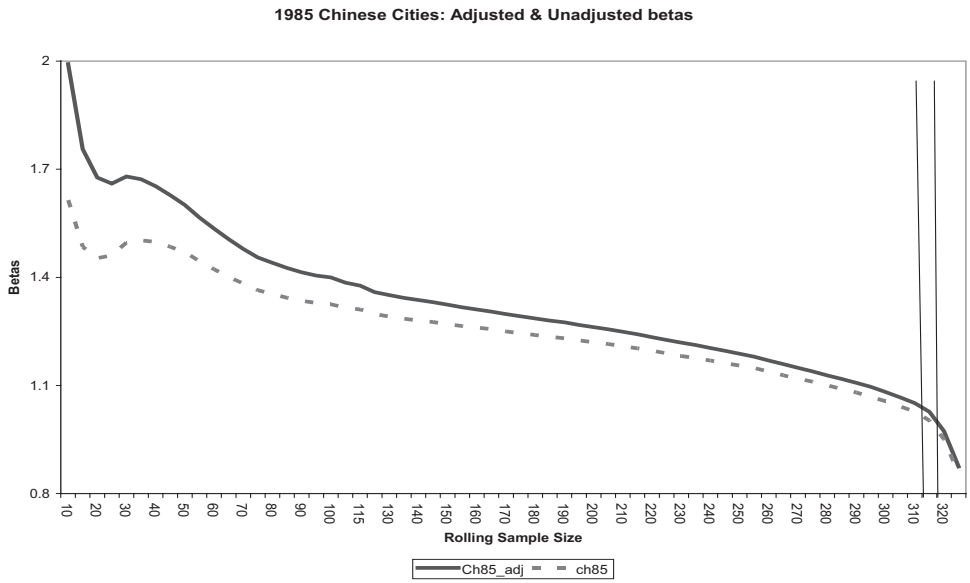

Figure 5. Zipf's Law: Chinese Cities in 1985

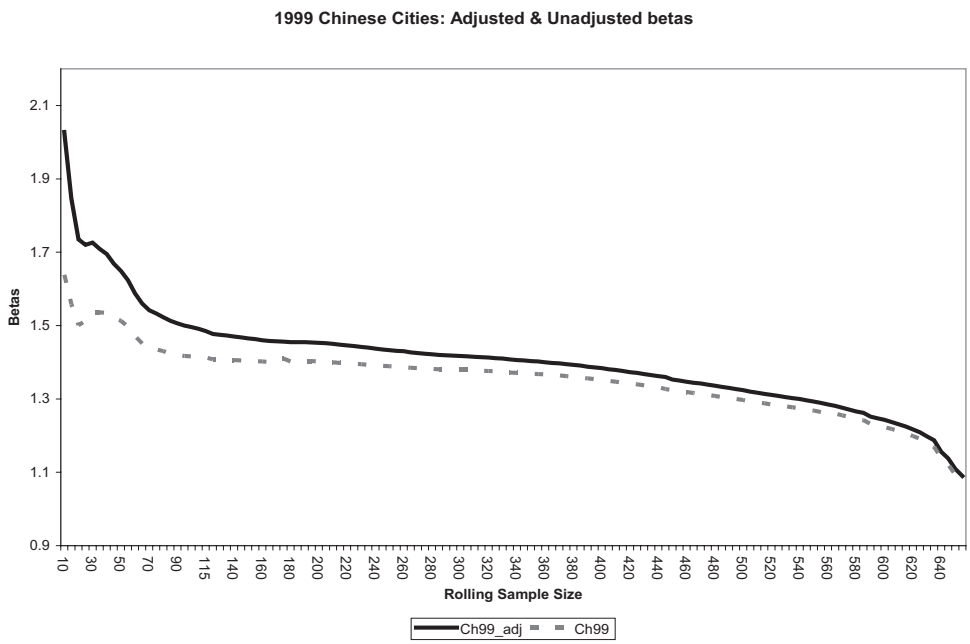

Figure 6. Zipf's Law: Chinese Cities in 1999

3.2 Results from Random Rolling Sample Method

As we discussed in the methodology section, the rolling sample method begins with a sample that includes the largest cities, increasing the sample size by adding one smaller city with each iteration. Thus, the truncation point effect, sample size effect, and variations in the size of cities are captured simultaneously. The purpose of the random rolling sample method is to single out the influence of sample size on the Zipf exponent. Through replacement and random sampling, variations in the size of cities become random, thereby avoiding bias for samples that include only large cities. Replacement and random sampling enable random changes in the truncation point, which eliminates the truncation point bias inherent in the rolling sample method.

Figure 7 presents the average estimated Zipf exponent using 100 regressions for each sample size. Both curves show that the estimated Zipf exponent is quite sensitive to sample size when they are relative small; however, it becomes stable for larger samples. For example, when examining urbanized areas in the U.S. in 1990, the estimated Zipf exponent decreases dramatically with sample size until 50, whereupon the effect of sample size disappears (i.e., the estimated coefficient remains nearly constant). In 2000, the estimated Zipf exponent decreased dramatically before the sample size reached 65. This again suggests that the Zipf exponent decreases with sample size, particularly when the sample size is relatively small.

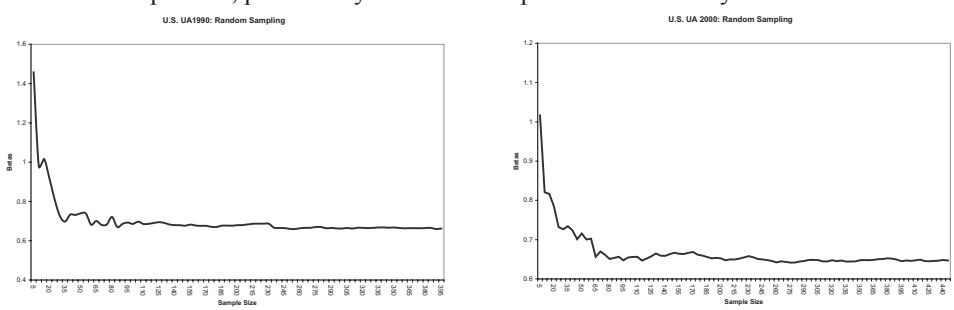

Figure 7. Results from Random Rolling Sampling

3.3 Simulation Results Using Random Sampling with Replacement

We conducted a Monte Carlo simulation using 1000 numbers with normal distribution to verify the robustness of the above findings. We then used the random rolling method with replacement, as in Section 3.1. Figure 8 presents the average estimated Zipf exponent from 100 regressions for each sample size. Interestingly, we still captured the effect of small sample sizes on the estimated Zipf exponent. The estimated Zipf exponent decreased dramatically before the sample size reached 35, continued decreasing until the sample sized reached 200, and then stabilized. This

confirms our earlier conclusion that the Zipf exponent decreases with sample size, particularly when samples are relatively small.

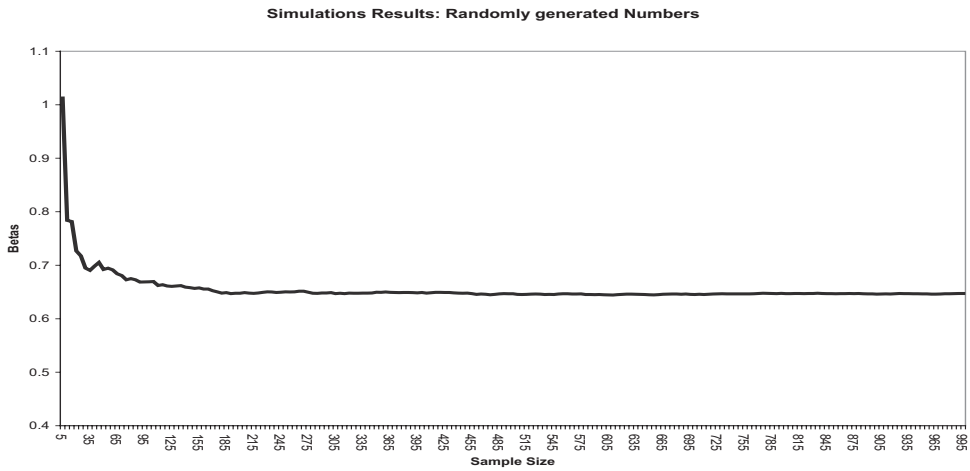

Figure 8. Simulation Results Using Random Sampling with Replacement

3.4 Further Evidence on the Relationship between Sample Size and the Zipf Exponent

As mentioned in Section 3.1, Figs. 2-4 suggest a lognormal relationship between sample size and the Zipf exponent. This section outlines a regression we ran to empirically determine the relationship between the estimated exponent ($\hat{\beta}$) and sample size (SS). This analysis involves running the following equation, the results of which are presented in Table 2.

$$\log(\hat{\beta}_i) = \alpha - \delta \log(SS_i) + \varepsilon \quad (4)$$

Table 2. Elasticity of Estimated Zipf Exponent with Respect to Sample Size

Nation	Year	$\hat{\delta}$	OLS Bias Corrected $\hat{\delta}$	R^2 (from adjusted)	Number of observations
U.S.	1980	-0.10***	-0.15***	0.98	355
U.S.	1990	-0.11***	-0.16***	0.96	385
U.S.	2000	-0.13***	-0.17***	0.97	441

***: significant at 1%

The numbers in Table 2 are the number of estimated exponents ($\hat{\beta}$'s) obtained using the rolling sample method (see Section 3.1). Using data from urbanized areas in the U.S., Eq. 4 yields very high R2-values (0.96 or higher) and highly significant results. Based on the OLS bias-corrected model (Eq. 3, column 4 in Table 2), the elasticity of the estimated exponent with respect to sample sizes are -0.15, -0.16, and -0.17 for 1980, 1990, and 2000, respectively. An increase of approximately one percent in the number of urban areas used in regression causes a 0.16 percent reduction in the value of the estimated exponent (for the adjusted model). The unadjusted model (Eq. 2, column 3 in Table 2) shows a smaller decrease in the value of the estimated exponent as the size of the sample increases; which explains why the unadjusted model converges with the adjusted model in Figs. 2-4.

These results are important. If the value of the estimated exponent is influenced significantly by sample size, we cannot expect the value of this exponent to remain close to 1.0 in all cases. Therefore, the validity of the rank-size rule depends largely on the size of the sample. In other words, the rank-size rule is not necessarily an economic regularity and may in fact be a statistical phenomenon.

4. CONCLUSIONS

This paper examined the validity of the rank-size rule according to the estimated Zipf exponent. Using a rolling sample technique, we proved that small samples of large cities tend to generate higher values for the estimated exponent compared to samples dominated by smaller cities. We demonstrated that the rank-size rule holds only for a number of selected sub-samples. Among the U.S. samples, the estimated Zipf exponent remains close to 1.0 for between only 180 and 205 cities (1980), between 140 and 195 cities (1990), and between 140 and 205 cities (2000). Among the Chinese cities, the estimated Zipf exponent is close to 1.0 only for sub-samples containing between 315 and 320 cities (1985) and never approached 1.0 for the 1999 data. Empirical evidence from a random rolling sample method and the results of a Monte Carlo simulation confirm that the Zipf exponent is negatively related to the size of samples used in regression.

The double log regression model for the estimated Zipf exponents and sample sizes yielded high R2-values and significant results. It revealed elasticity in the estimated exponent with respect to sample size. For urbanized areas in the U.S in 1980, 1990, and 2000, an increase of approximately one percent in the number of urban areas used in regression would cause a 0.16 percent reduction in the value of the estimated exponent. This statistically determines how the Zipf exponent responds (negatively) to changes in sample size. It also suggests that the rank-size rule should be interpreted with caution. In other words, this well-known and commonly used rank-size rule may be more of a statistical phenomenon than an economic regularity.

REFERENCES

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

Gabaix, X. (1999a). Zipf's Law for Cities: An Explanation. *The Quarterly Journal of Economics, 144*(3), 739-67.

Gabaix, X. (1999b). Zipf's Law and the Growth of Cities. *American Economic Review, 89*(2), 129-132.

Gabaix, X., Ibragimov, R. (2011). Rank – ½: A Simple Way to Improve the OLS Estimation of Tail Exponents. *Journal of Business Economic and Statistics, 29*(1), 24-39.

Gan, L., D. Li, S. Song. (2006). Is the Zipf's Law Spurious in Explaining City-Size Distributions? *Economic Letters, 92*(2), 256-62.

Krugman, K. (1995). *Development, Geography, and Economic Theory.* Cambridge, MA: MIT Press.

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.* New York, NY: Harper Collins College Publishers.

Nitsche, V. (2005). Zipf Zipped. *Journal of Urban Economics, 57*(1), 86-100.

Rosen, K., Resnick, M. (1980). The Size Distribution of Cities: An Explanation of the Pareto Law and Primacy. *Journal of Urban Economics, 8*(2), 165-186.

Shiode, N., Batty, M. (2000). *Power Law Distribution in Real and Virtual Worlds*. CASA Working Paper 19.

Sinclair, R. (2001). *Examining the Growth Model's Implications: The World Income Distribution.* Syracuse University Working Paper.

Song, S., Zhang, K. H. (2002). Urbanization and City Size Distribution in China. *Urban Studies, 39*(12), 2317-27.

Tachimori, Y., Takashi, T. (2002). Clinical Diagnoses Following Zipf's Law. *Fractals, 10*(3), 341-351.

Zipf, G. (1949). *Human Behavior and the Principle of Last Effort.* Cambridge, MA: Addison Wesley Press.