View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Research Papers in Economics

UNR Joint Economics Working Paper Series
Working Paper No. 08-005

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

Fungisai Nota and Shunfeng Song

Department of Economics /030
University of Nevada, Reno
Reno, NV 89557-0207
(775) 784-6850 │ Fax (775) 784-4728
email: song@unr.nevada.edu

September, 2008

Abstract

The widely-used Zipf law has two striking regularities: excellent fit and close-to-one exponent. When the exponent equals to one, the Zipf law collapses into the rank-size rule. This paper further analyzes the Zipf exponent. By changing the sample size, the truncation point, and the mix of cities in the sample, we found that the exponent is close to one only for some selected sub-samples. Using the values of estimated exponent from the rolling sample method, we obtained an elasticity of the exponent with respect to sample size.

JEL Classification: C1, R1

Keywords: Zipf law; Rank-size rule; Rolling sample method

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

Fungisai Nota[1] and Shunfeng Song[2]

Abstract:

The widely-used Zipf law has two striking regularities: excellent fit and close-to-one exponent. When the exponent equals to one, the Zipf law collapses into the rank-size rule. This paper further analyzes the Zipf exponent. By changing the sample size, the truncation point, and the mix of cities in the sample, we found that the exponent is close to one only for some selected sub-samples. Using the values of estimated exponent from the rolling sample method, we obtained an elasticity of the exponent with respect to sample size.

JEL classification: C1; R1

Keywords: Zipf law; Rank-size rule; Rolling sample method

[1] Nota is an Assistant Professor of Economics at the Wartburg College, Iowa, USA. Email: fungisai.nota@wartburg.edu

[2] Song is a Professor of Economics at the University of Nevada, Reno, NV 89557, USA and an adjunct research fellow at Center for Research of Private Economy, Zhejiang University, China. Email: song@unr.edu

1. Introduction:

Zipf law states that the rank associated with some size *S* is proportional to *S* to some negative power (Zipf, 1949). It has two striking observations. One is its excellent fit. Numerous empirical studies have shown that a linear regression of log-rank on log-size generates an excellent fit (high R^2-value). 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). This astonishing regularity led Krugman (1995, p.44) to say that the rank-size rule is "a major embarrassment for economic theory: one of the strongest statistical phenomena we know, lacking any clear basis in theory." Fujita et al. (1999, p. 219) stated "the regularity of the urban size distribution posses a real puzzle, one that neither our approach nor the most plausible alternative approach to city sizes seems to answer." The other striking observation is about the Zipf coefficient. For the 44 countries studied by Rosen and Resnick (1980), the estimated coefficient ranges from 0.809 for cities in Morocco to 1.963 for cities in Australia. Nitsche (2005) analyzed 515 estimates from 29 studies of the rank-size relationship and found that two-third of the estimated coefficients are between 0.80 and 1.20.

Several studies have attempted to explain why Zipf law holds. Gabaix (1999a, 1999b) proved that the Zipf law derives from the Gibrat law, where the Gibrat law states that the growth process is independent of size. Gan et al. (2006) concluded that the Zipf law is a statistical phenomenon rather than an economic regularity. However, the striking observation of Zipf coefficient close to 1 remains a puzzle. Is it an economic regularity or a statistical phenomenon? This paper attempts to solve this puzzle.

The next section outlines the methodologies used in this analysis, the rolling sample method and the random sampling method with replacement. The third section provides the results. The final section summarizes the empirical results and discusses their economic significance. Succinctly, this paper seeks to find the impact of sample size, truncation point and the mix of cities on the estimated exponent of the Zipf law.

2. The model and methodology

Zipf law is commonly expressed in the following form:

$$R_i = AS_i^{-\beta} \qquad [1]$$

where R_i is the rank of the ith city, S_i is the city's size and β is the exponent coefficient. With a log transformation, it estimates β as follows:

$$\log(R_i) = \alpha - \beta \log(S_i) + \varepsilon \qquad [2]$$

Several studies have noted that estimating Equation [2] yields an OLS bias through the standard errors. To correct this bias, Gabaix and Ibragimov (2006) offered the following version that gives unbiased standard errors of $(2/\bar{n}_i)^{0.5}\hat{\beta}$, where $\bar{n}_i$ is the corresponding sub-sample size:

$$\log(R_i - 0.5) = \alpha - \beta \log(S_i) + \varepsilon \qquad [3]$$

This corrected version is known as the rank-minus-half rule. Throughout this analysis, we will provide results from both versions and comment on the differences that exist between them.

The first method we use in this paper is the rolling sample method. We estimate the exponent coefficient β using OLS and repeat 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 every time, thereby increasing the sub-sample size by one each time. For example, the full sample size of U.S. urbanized areas for 1990 is 396. These urbanized areas are ordered decreasingly from the largest urbanized area of New York to the smallest one of

Brunswick, GA. The first sub-sample size is $\bar{n}_1$, the 10 largest cities for example; then the second sub-sample is $\bar{n}_2 = \bar{n}_1 + 1$, the 11 largest cities, and so on. We continue this process until the last sub-sample becomes the full sample of 396. The advantage of this methodology is to capture the coefficient variation as both the sample size and truncation point change.

The second method, random sampling with replacement, separates some of the simultaneous effects captured with the rolling sample. The rolling sample provides the gross variation in the estimated coefficient as three factors change simultaneously (sample size, truncation point and the variation in city sizes). To untangle these effects, we use our original data for each year as a pool to select from. We then randomly select the first sub-sample, 10 random cities for example, and rank them up. The second sub-sample is independent of the first sub-sample. However, it contains one more city, and so on. We continue this process until the last sub-sample becomes the full sample. Since this is a random process, we run regression 100 times for each sample size and get 100 estimated coefficients. We average these series and obtain the distribution of the coefficient with respect to sample size.

The third method is to further test the effect of sample size on the distribution of the estimated coefficient. For this, we randomly generate 1000 numbers from a normal distribution. We then apply the random sampling technique and repeat the process we did above. After 100 iterations, we average the series of $\hat{\beta}$'s and obtain the distribution of the coefficient with respect to sample size.

3. Results

Table 1 shows the full-sample results of Zipf law. Not surprisingly, we obtained very high R^2-values. Comparing the estimated coefficients between OLS bias corrected and uncorrected models, we conclude that the uncorrected Zipf law has a downward bias.

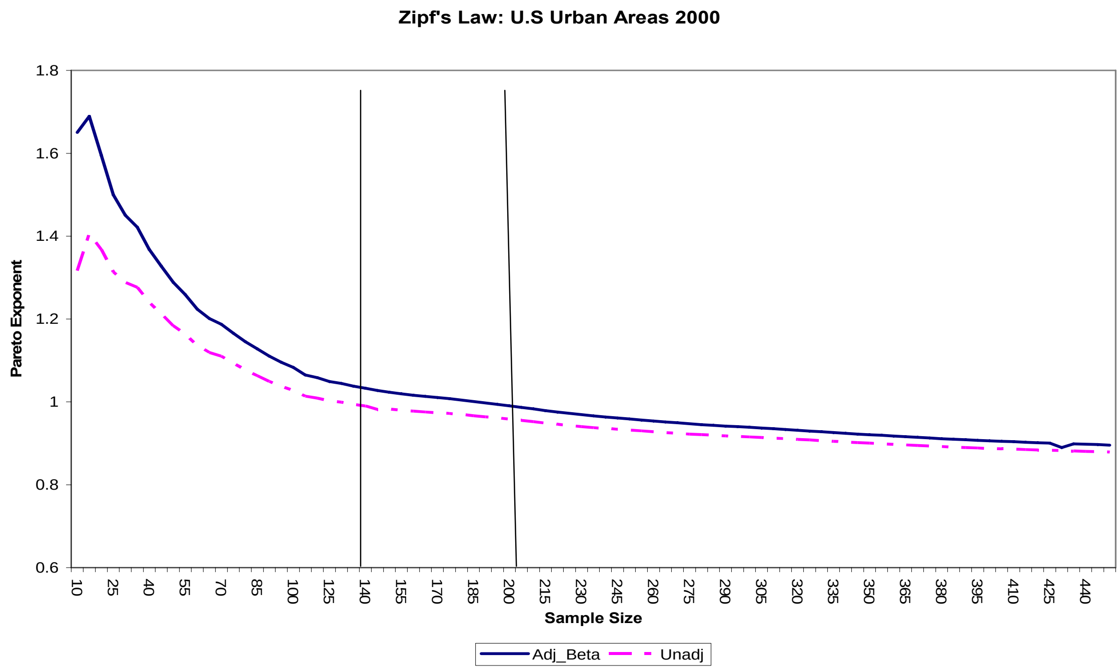
Table 1: Regression results on Zipf's law using data on US urbanized areas

Year	$\hat{\beta}$	**OLS Bias Corrected** $\hat{\beta}$	$\mathbf{R^2}$ **(from unadjusted)**	**Sample Size**
1980	0.91	0.925	0.989	366
1990	0.895	0.913	0.989	396
2000	0.875	0.895	0.989	452

Data Sources: U.S Bureau of Census (2000).

The rolling sample results show a negative relationship between the estimated coefficient and sample size. This implies that small samples of big cities yield higher coefficients than large samples that also include smaller cities. Does the rank-size rule exist? We note that the rank-size rule only holds for certain sub-samples where the 95% confidence interval includes 1. Specifically, for the 1980 cities, rank-size rule holds only for sub-samples between 180 and 205; for 1990, 140 to 195; and for 2000, 140 to 205. This finding suggests that the rank-size rule (i.e., β=1) does not holds for either large cities or the larger samples with more small cities. Figure 1 shows the distribution of estimated coefficients with respect to sample size for 2000.

Figure 1



Interestingly, the graph suggests a lognormal distribution. To confirm this observation, we run the following regression between the estimated exponent ($\hat{\beta}$) and the sample size (*SS*):

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

Table 2 presents the results, with observations being the number of estimated exponents ($\hat{\beta}$'s) obtained from the rolling sample method. Surprisingly, the lognormal regression yields a very high R^2-value, indicating a strong statistical relationship between estimated coefficient and sample size. For the OLS-bias-corrected model, Table 2 shows that a one percent increase in the sample size would lead to a 0.15 percent or more decrease in the value of the estimated exponent. The uncorrected model shows a smaller elasticity of estimated exponent with respect to sample size, and this explains why the uncorrected model converges with the corrected model in Figure 1. These results are important, because they prove that the validity of the rank-size rule largely depends on the sample size used in a study. In other words, the rank-size is not an economic regularity but a statistical phenomenon.

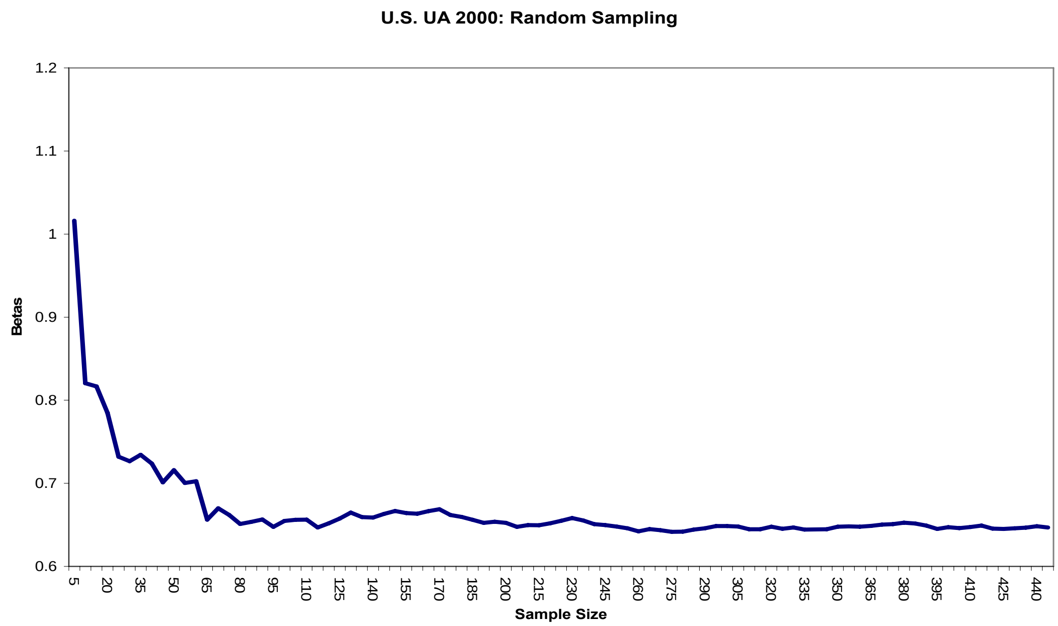
Table 2: The relationship between the estimated Zipf exponent and the sample size

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

***: significant at 1%

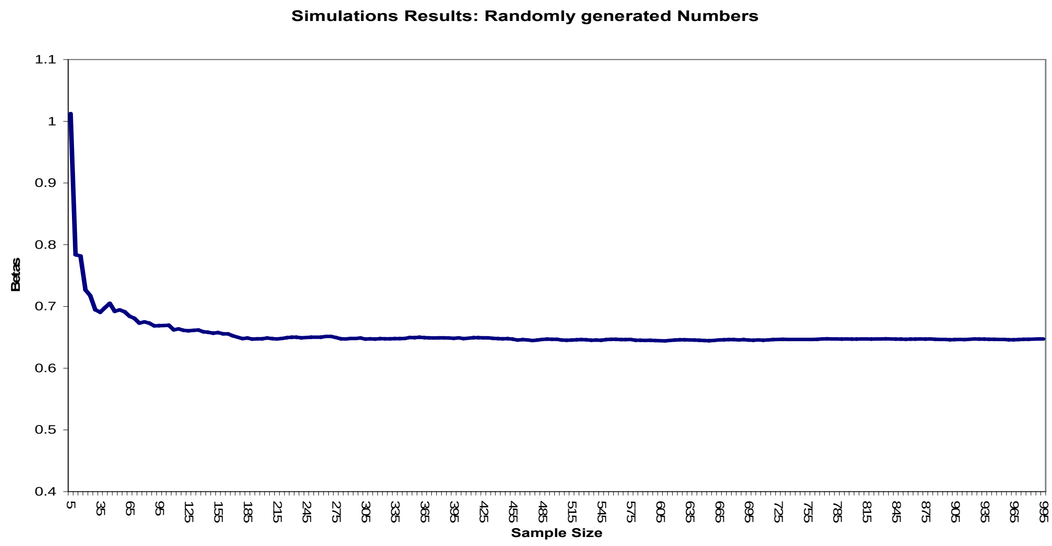
As we discussed in the methodology section, a dilemma exists with the rolling sample technique because it captures the joint effect of truncation point, sample size, and the assortment of cities in the sample. Using the random sampling with replacement technique while increasing the sub-sample, we capture an assortment of cities that can include all sizes from the beginning. This eliminates the bias due to large cities in the first sub-samples. By randomly sampling each time, the truncation point also randomly changes. This eliminates the systematically changing truncation point bias inherent in the rolling sample technique. Figure 2 presents the distribution of estimated coefficients based on the random sampling method for 2000. It shows that sample size alone has an upward bias mainly for sub-samples below 100. For sample sizes greater than 100, the effect of sample size disappears as we increase the sample size. i.e., the estimated coefficient stays almost constant.

Figure 2



To further test the effect of sample size on the distribution of the estimated coefficient, we randomly generate 1000 numbers from a normal distribution. We then apply the random sampling technique and repeat the process we did above. After 100 iterations, we average the series of $\hat{\beta}$'s and show the results in the graph below. Surprisingly, we still capture the effect of very small sample sizes below 100. Figure 3 confirms the upward bias of samples less than 100. For sample size greater than 100, the sample size has little influence on the value of estimated coefficient.

Figure 3



4. Conclusions

This paper has examined the validity of the rank-size rule based on estimated Zipf exponent. Using the rolling sample technique, we proved that small samples with large cities tend to generate high values of the estimated coefficient compared to samples dominated with small cities. The rank-size rule holds only for some selected sub-samples. We also observed the upward bias of the estimated coefficient when we used random sampling with replacement technique and got random samples from a normal distribution.

The double log regression model of estimated exponents and sample sizes yielded a very high R^2-value. It also produced an elasticity of the estimated exponent with respect to sample size, with a one percent increase in the sample size leading to about 0.15 percent or more decrease in the value of the estimated exponent. Therefore, we conclude that the Zipf exponent depends on the sample size used in a study and the rank-size rule does not hold in general. In other words, the rank-size is not an economic regularity but a statistical phenomenon.

References:

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

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

Gabaix, X., 1999b. Zipf's law and the growth of cities. American Economic Review, Vol. 89 (2), 129—132.

Gabaix, X., Ibragimov, R., 2006. Rank – ½: A simple way to improve the OLS estimation of tail exponents. Working Paper.

Gan, L., Li, D., Song, S., 2006. Is the Zipf's law spurious in explaining city-size distributions? Economic Letters 92, 256—262.

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

Nitsche, V., 2005. Zipf zipped. Journal of Urban Economics 57, 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, 165-186.

Zipf, G., 1949. Human behavior and the principle of last effort. Cambridge, MA: Addison Wesley Press.