Estimation and Testing for Rank Size Rule Regression under Pareto Distribution

Y. Nishiyama[a] , **S. Osada**[a] **and K. Morimune**[b]

[a] *Kyoto Institute of economic Research, Kyoto University, Kyoto 606-8501, Japan*

[b] *Graduate School of Economics, Kyoto University, Kyoto 606-8501, Japan*

Abstract: Letting $S_{(i)}$ be the *i*-th largest city in a country, it is often observed that $\log S_{(i)} \approx \alpha_0 + \alpha_1 \log i$ for some $\alpha_0 > 0$ and $\alpha_1 < 0$. It is called rank size rule when $\alpha_1 = -1$. This relationship has been examined by means of ordinary least squares estimation and *t* test in the literature. However, since $S_{(i)}$ is heteroskedastic and autocorrelated, *t* statistics do not have standard distribution. Indeed we show $t \to^p \infty$ as the sample size increases. The purpose of this paper is to obtain statistical properties of OLS estimator of the rank size rule regression and distribution of *t* statistics under Pareto distribution, and further to propose more efficient estimation procedures in two ways. Firstly, we improve efficiency by adjusting the heteroskedasticity and autocorrelation by GLS method. Another source of efficiency gain is to exclude some large variance observations. It seems GLS attains the Cramer-Rao lower bound for α_1.

Keywords: Rank size rule; Zipf law; Pareto distribution; City size

1. INTRODUCTION

After pioneering work on city size distribution by Auerbach [1913] and Zipf [1949], many researchers have investigated a wide range of settlement systems. Zipf's main result called Zipf law is the following. Let *S* denote a random variable representing city size measured by its population, then for large *x*,

$$P(S \geq x) = A / x$$

for some *A*>0 or Pareto distribution with unit exponent. This is closely related to so-called rank size rule of city size data. Let S_i, $i = 1,...,n$ be population of cities in a country, and $S_{(i)}$ be its order statistics satisfying $S_{(1)} \geq ... \geq S_{(n)}$ then we often observe that

$$\log S_{(i)} \approx \alpha_0 + \alpha_1 \log i, \quad i = 1,2,...,n, \qquad (1.1)$$

where $\alpha_0 > 0$ and $\alpha_1 < 0$. This relationship is called rank size rule when $\alpha_1 = -1$. When Zipf law holds, rank size rule follows approximately. Regarding (1.1) as a regression model, many researchers have estimated α_0 and α_1 by ordinary least squares (OLS) method and implemented *t* test for $\alpha_1 = -1$.

One of the most important papers in this field is Rosen and Resnick [1980]. They examined city size distribution of the 50 largest administrative urban areas in 44 countries and they concluded validity of the urban rank-size rule appears to be an open question. Soo [2002] also made an international comparison using updated data of 73 countries.

Many researchers, including the above mentioned ones, have studied (1.1) based on the OLS estimation and *t* test. But since the dependent variable there does not satisfy the standard conditions of OLS regression, we cannot evaluate the results. The purpose of this paper is to derive the exact and approximate properties of the OLS estimator and *t* test statistics for the rank size rule null, or $\alpha_1 = -1$. We obtain the bias and variance of the estimator assuming S_i are independently and identically distributed (iid). Further we show *t* statistic does not have *t* distribution unlike standard classical linear regression theory because $S_{(i)}$ are in fact autocorrelated and heteroskedastic

under the iid assumption. Since Zipf suggested, it is often assumed that S_i have Pareto distribution. Under this assumption, we can show

$$E[\log S_{(i)}] = \alpha_0 + \alpha_1 \log i$$

does not strictly hold for $\forall \alpha_0, \alpha_1$ in small samples, but it does approximately only for large *n* and *i*.

The following section shows exact and approximate expressions for $E[\log S_{(i)}]$ and $V[\log S_{(i)}]$ then derive the bias and variance of the OLS estimator for (1.1). Then we present Monte Carlo results on the distribution of *t* value for the estimator which is far away from *t* distribution. We further show *t* explodes asymptotically, indicating *t* test is never applicable to test the null of $\alpha_1 = -1$. Section 3 proposes more efficient estimators, while Section 4 gives empirical results from Japanese city size data of Metropolitan Employment Area (MEA). Section 5 is conclusions.

2. OLS ESTIMATION OF THE RANK SIZE RULE REGRESSION

2. 1 Bias and Variance of the OLS Estimator

We state some results on the properties of the estimator without proofs in the sequel. Assume $S_i, i = 1,..., n$ are iid from a Pareto distribution function $F_S(x) = 1 - x^{-\beta}$ $(\beta > 0, x \geq 1)$. Then the following lemma holds.

LEMMA 1 Letting $\{S_{(1)},..., S_{(n)}\}$ be the order statistics of $S_i, i = 1,..., n$ satisfying $S_{(1)} \geq ... \geq S_{(n)}$,

(a) $E[\log S_{(i)}] = \beta^{-1} \sum_{k=1}^{n-i+1} (n-k+1)^{-1}$

(b) $V[\log S_{(i)}] = \beta^{-2} \sum_{k=1}^{n-i+1} (n-k+1)^{-2}$

(c) $Cov[\log S_{(i)}, \log S_{(j)}] = Var[\log S_{(j)}], \ (i < j)$.

This lemma straightforwardly yields the following proposition.

PROPOSITION 1

$$E[\log S_{(i)}] = \frac{\log n}{\beta} - \frac{\log i}{\beta} + O(\frac{1}{n} + \frac{1}{i})$$

as $n \to \infty, i \to \infty$.

Proposition 1 implies that approximation of (1.1) is justified when *n* and *i* are large. Based on Proposition 1 and Lemma 1, we can obtain the exact expectation and variance of the OLS estimator $\hat{\alpha}_1$ for α_1 as follows.

PROPOSITION 2

$$E(\hat{\alpha}_1) = \frac{n\sum \log i(n^{-1} + ... + i^{-1}) - n\sum \log i}{\beta\{n\sum \log^2 i - (\sum \log i)^2\}}.$$

$$V(\hat{\alpha}_1) = \frac{\sum_{j=1}^{n} (\frac{n}{j} \sum_{k=1}^{j} \log k - \sum \log i)^2}{\beta^2 \{n\sum \log^2 i - (\sum \log i)^2\}^2}.$$

We suppress the expressions for those of $\hat{\alpha}_0$ or the OLS estimate for α_0 because it is of less importance and interest. From this proposition, we derive the asymptotic expressions of bias and variance:

$$E(\hat{\alpha}_1) - (-\frac{1}{\beta}) = \frac{C_n}{\beta}, \ C_n = -\frac{\log n}{4n} + o(\frac{\log n}{n}),$$

$$V(\hat{\alpha}_1) = \frac{D_n}{\beta^2}, \ D_n = \frac{2}{n} + \frac{\log^3 n}{3n^2} + O(\frac{\log^2 n}{n^2}).$$

Letting $B_n = E(\hat{\alpha}_1) - (-\frac{1}{\beta})$, Figure 1 draws $B_n \times \beta = C_n$. The bias decays as the sample size increases. Figure 2 shows $V(\hat{\alpha}_1) \times \beta^2 = D_n$ which also decreases with *n*.

Figure 1. Bias of OLS estimator

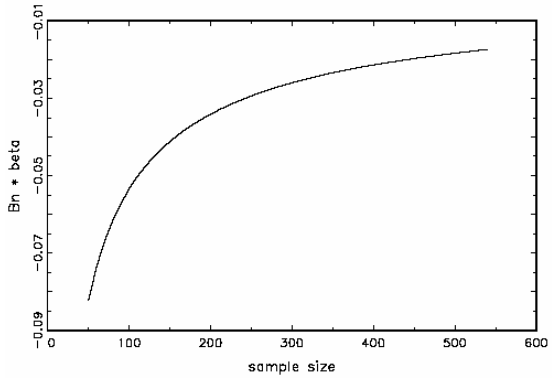


Figure 2. Variance of OLS estimator

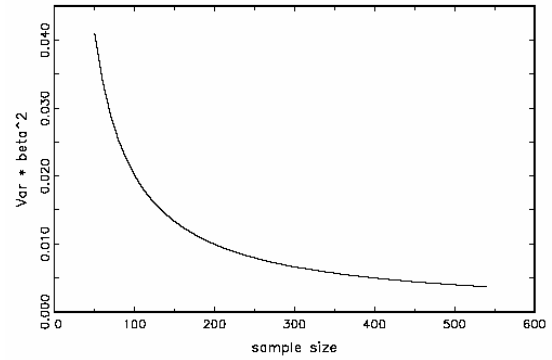


Table 1 tabulates these values for some *n*. The above results directly suggest a simple bias correction of the following form:

$$\bar{\alpha}_1 = \frac{n\sum \log^2 i - (\sum \log i)^2}{n\sum \log i(n^{-1} + ... + i^{-1} - 1)} \hat{\alpha}_1.$$

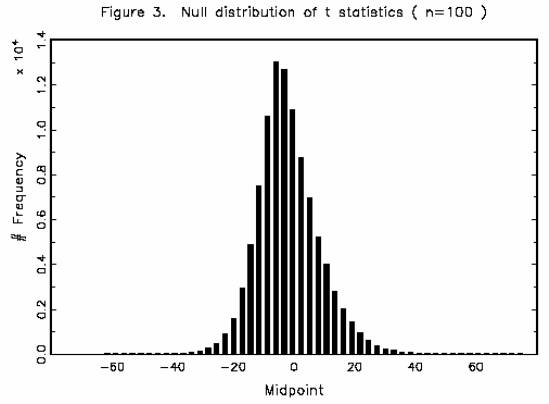
We have two remarks regarding this estimator. Firstly the multiplicative constant on the right depends only on *n*, free from any unknown quantities, and thus it is easy and feasible. Secondly, this method not only eliminates the bias but also reduces the variance because the multiplicative constant is smaller than unity.

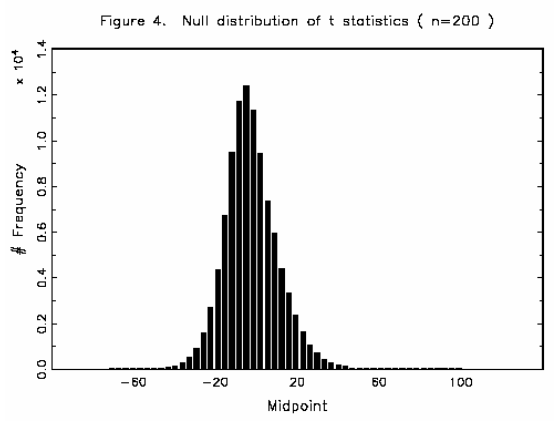
Table 1. Values of C_n and D_n.

n	C_n	D_n
50	-0.0822	0.0410
100	-0.0534	0.0201
200	-0.0341	0.0099
300	-0.0183	0.0040

2.2 The distribution of *t* statistics

In testing the significance of coefficients of linear regression models, we implement *t* test. In the present case, because $\log S_{(i)}$, the dependent variables, are not only normally distributed but also heteroskedastic and autocorrelated. We obtained the distribution of *t* statistics for α_1 in the regression (1.1) under the null of $\alpha_1 = -1$ by Monte Carlo simulation. Figure 3 and 4 show the histogram from 100,000 replications when *n*=100, 200 respectively. The mean, variance, skewness





and kurtosis are respectively -2.512, 171.0, 0.423, 1.012 when *n*=200. Therefore they are obviously far from *t* distribution. Table 2 shows empirical critical regions of two-sided *t* test when *n*=100 for different sizes calculated from the simulation, which should be used in testing $\alpha_1 = -1$ instead of quantiles of *t* distribution. We immediately know we face severe size distortion if we blindly apply *t* test for $\alpha_1 = -1$, because its critical region is set to be around (-∞, -2],[2,∞) in the case of test with 5% size.

Table 2. Empirical critical regions of two-sided *t* test by simulation

Size	Critical region (*n*=100)
10%	(-∞,-17.03], [16.14, ∞)
5%	(-∞,-20.17], [20.68, ∞)
1%	(-∞,-27.09], [31.08, ∞)

Moreover, we found in a simulation not reported here that *t* tends to become larger in magnitude as the sample size increases. This phenomenon is caused by the fact that standard error of the regression tends to zero as $n \to \infty$, which is proved in the following proposition.

PROPOSITION 3

Letting $s^2 = \frac{1}{n-2}\sum \{\log S_{(i)} - \hat{\alpha}_0 - \hat{\alpha}_1 \log i)^2$,

we have

(a) $E(s^2) = O(\frac{\log n}{n})$.

(b) $s^2 \to^p 0$ as $n \to \infty$.

Rewriting the estimator by Reyni representation (Reyni [1953]), straightforward application of Lindeberg-Feller central limit theorem and

Cramer device yield the following limiting distribution of $(\hat{\alpha}_0, \hat{\alpha}_1)$.

PROPOSITION 4

$$\begin{bmatrix} \frac{\sqrt{n}}{\log n} & 0 \\ 0 & \sqrt{n} \end{bmatrix} \begin{bmatrix} \hat{\alpha}_0 - \frac{\log n}{\beta} \\ \hat{\alpha}_1 - (-\frac{1}{\beta}) \end{bmatrix} \to^d N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \frac{2}{\beta^2} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \right)$$

Proposition 3 and 4 give the following result on *t* statistics for $\hat{\alpha}_1$.

PROPOSITION 5

For $t = \dfrac{\hat{\alpha}_1 - (-\beta^{-1})}{\sqrt{s^2 (X'X)^{-1}_{22}}}$, we have

$t \to^p \infty$ as $n \to \infty$, where

$$(X'X)^{-1}_{22} = \frac{n}{n\sum \log^2 i - (\sum \log i)^2} = \frac{1}{n} + o(\frac{1}{n})$$

is the (2,2)-element of $(X'X)^{-1}$.

This proposition indicates *t* value for this regression explodes asymptotically under the null of true parameter value. Therefore, when we would like to test a null hypothesis such as $\alpha_1 = -1$, we know we should never use standard *t* test especially when the sample size is large, but we should apply an asymptotic normality based test using

$$\frac{\hat{\alpha}_1 - (-\beta^{-1})}{\sqrt{2\hat{\alpha}_1^2 / n}} \to^d N(0,1)$$

as recommended in e.g. Gabaix and Ioannides (2003). $2/\beta^2$ involved in the asymptotic variance is replaced by a consistent estimator $2\hat{\alpha}_1^2$ under the null. In many application work, such as Rosen and Resnick (1980), Alperovich (1984) and Soo (2002), mechanical application of *t* test provides very large *t* values, leading to wrong conclusions.

3. MORE EFFICIENT ESTIMATION

We propose two methods of efficiency improvement in the estimation of (1.1). One is generalized least squares (GLS) method adjusting nonspherical disturbances, while the other is a trimmed least squares regression. The idea is that observing $Var[\log S_{(i)}]$ is larger for smaller *i*, and also approximation (1.1) is worse for smaller *i* in view of Proposition 1, we can expect to improve the statistical properties of the estimator by dropping some observations with smaller *i*, or larger observations.

3.1 GLS estimation

Putting

$$y' = [\log S_{(1)}, \log S_{(2)}, \cdots, \log S_{(n)}],$$

$$X' = \begin{bmatrix} 1, & 1, & \cdots, & 1 \\ \log 1, & \log 2, & \cdots, & \log n \end{bmatrix},$$

and

$$\Omega = V(y),$$

GLS estimator for α_0, α_1 is simply

$$\begin{bmatrix} \tilde{\alpha}_0 \\ \tilde{\alpha}_1 \end{bmatrix} = (X'\Omega^{-1}X)^{-1} X'\Omega^{-1} y \qquad (3.1)$$

and its variance is

$$V\left(\begin{bmatrix} \tilde{\alpha}_0 \\ \tilde{\alpha}_1 \end{bmatrix} \right) = (X'\Omega^{-1}X)^{-1}.$$

Expressions for the elements of Ω are in Lemma 1 (b), (c).

An interesting feature in (3.1) is that it is free from nuisance parameters unlike usual GLS estimation. Normally Ω involves some nuisance parameters and thus GLS estimation is infeasible, so that we need to estimate Ω in the first step in practice. In

Figure 5. Bias of GLS estimator

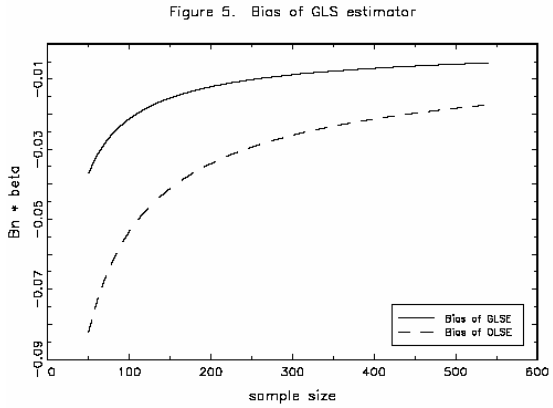
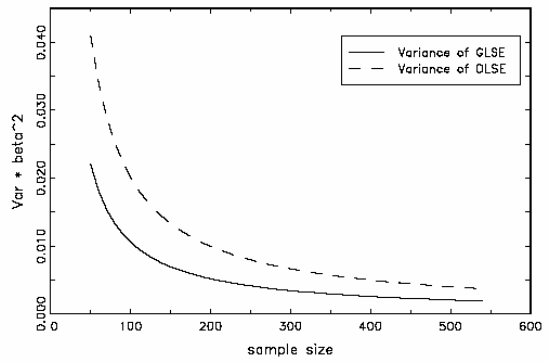


Figure 6. Variance of GLS estimator



view of Lemma 1 (b), (c), Ω itself involves an unknown parameter β, but it appears only as a multiplicative constant. Then, due to the form of (3.1), it cancels, so that (3.1) turns to be feasible. Similarly to the OLS estimator, we can obtain the exact bias and variance of GLS estimator analogous to Proposition 2, which are in Figure 5 and 6. We do not present them explicitly because of their long and tedious expressions. Table 3 provides them for the same sample size with Table 1 to compare with those for the OLS.

Table 3. Bias and variance of GLS estimator for α_1.

n	bias$\times\ \beta$	variance$\times\beta^2$
50	-0.03681	0.0220
100	-0.02128	0.0105
200	-0.01217	0.0051
300	-0.0058	0.0020

We give the following two remarks regarding this result comparing two tables. Firstly, GLS procedure reduces not only the variance but also the bias, which we did not expect because GLS is primarily developed in order to improve the efficiency, not bias reduction. Secondly, we see the variance of GLS is about a half of that of OLS, and further it approximately equals to $1/(\beta^2 n)$, which coincides with Cramer-Rao lower bound for $1/\beta$ in fact. Therefore, we anticipate GLS gives an efficient estimate, comparable with the maximum likelihood estimator (MLE).

3.2 Trimmed OLS and GLS

Proposition 1 and Lemma 1(a) imply that source of the bias of least squares estimators is the approximation error of $n^{-1}+\cdots+i^{-1}$ by $\log n-\log i$ and it is larger for smaller i. Then we conjecture the bias can be reduced by excluding observations with smaller i. Also Lemma 2(b) imply that variance of least squares estimators could become smaller if we trim observations with smaller i, though there should no doubt be trade-off between efficiency gain by exclusion of larger variance data points and efficiency loss due to the reduced sample size. Letting $(\hat{\alpha}_{0,k},\hat{\alpha}_{1,k})$ and $(\tilde{\alpha}_{0,k},\tilde{\alpha}_{1,k})$ be respectively OLS and GLS estimators from the subsample of $[\log S_{(k+1)},\cdots,\log S_{(n)}]$, where the larger k observations are excluded, we have similarly to Proposition 2,

$$
\begin{aligned}
&E(\hat{\alpha}_{1,k})-(-\frac{1}{\beta}) \\
&=\frac{1}{\beta}\left\{\frac{(n-k)\sum_{i=k+1}^{n}\log i(n^{-1}+\cdots+i^{-1}-1)}{(n-k)\sum_{i=k+1}^{n}\log^2 i-(\sum_{i=k+1}^{n}\log i)^2}+1\right\} \\
&=\frac{C_{n,k}}{\beta},
\end{aligned}
$$

and

$$
\begin{aligned}
V\left(\begin{bmatrix}\hat{\alpha}_{0,k}\\ \hat{\alpha}_{1,k}\end{bmatrix}\right)&=\begin{bmatrix} n-k & \sum_{i=k+1}^{n}\log i \\ \sum_{i=k+1}^{n}\log i & \sum_{i=k+1}^{n}\log^2 i\end{bmatrix}^{-1} \\
&\times V\left(\begin{bmatrix}\sum_{i=k+1}^{n}\log S_{(i)}\\ \sum_{i=k+1}^{n}\log i\log S_{(i)}\end{bmatrix}\right)\begin{bmatrix} n-k & \sum_{i=k+1}^{n}\log i \\ \sum_{i=k+1}^{n}\log i & \sum_{i=k+1}^{n}\log^2 i\end{bmatrix}^{-1}.
\end{aligned}
$$

Let its (2,2)-element be $D_{n,k}/\beta^2$. They are constants determined only by n and k independent of unknown quantities. We can similarly obtain the corresponding formulae for the GLS estimator, but suppress them. We tabulate the bias, variance and mean squared error (MSE) for both trimmed OLS and GLS estimators in Table 4 for n=100 and k=0,…,15. We find larger k yields smaller bias in magnitude for both OLS and GLS, while variance of OLS estimator attains the minimum when k=8 as a result of the trade-off mentioned above. GLS variance, on the other hand, increases with k, thus there is no efficiency gain but only efficiency loss by decreased sample size. Based on the above findings, we can propose an optimal trimming rule by the minimum MSE principle.

When n=100, k=9 gives the optimal trimming in OLS estimation, while in GLS estimation, k=1 is the best choice. In OLS estimation, we attain about 33% MSE improvement. In GLS estimation, variance of y is stabilized by Ω^{-1} (see (3.1)) so that we need to exclude much less observations than the OLS. We note the best trimming points depend only on n because MSE= $\beta^2(C_{n,k}^2 + D_{n,k})$ where $C_{n,k}$ and $D_{n,k}$ depend only on n and k. Table 5 gives the best trimming points for some n. As easily expected, we should exclude more observations for larger sample size.

4. CONCLUSIONS

We examined statistical properties of least squares estimators for rank size rule regression of city size under Pareto distribution. Standard method in empirical study of regional science has been OLS estimation and t test based on it. We obtained exact bias and variance of OLS estimator for the coefficient. By Monte Carlo simulation, we obtained distribution of t statistics, where we found t statistic does not have t distribution, and we will face a severe size distortion if we implement t test. Moreover we proved t value asymptotically explodes in fact.

Table 4. Bias and variance of trimmed OLS and GLS estimators (n=100)

k	bias(OLS)	var(OLS)	MSE(OLS)	bias(GLS)	var(GLS)	MSE(GLS)
0	-0.05342	0.02006	0.02292	-0.02128	0.01055	0.01101
1	-0.03525	0.01723	0.01847	-0.01913	0.01061	0.01097
2	-0.02838	0.01628	0.01708	-0.01763	0.01068	0.01099
3	-0.02449	0.01578	0.01638	-0.01650	0.01077	0.01104
4	-0.02190	0.01549	0.01597	-0.01560	0.01086	0.01110
5	-0.02001	0.01531	0.01571	-0.01486	0.01096	0.01118
6	-0.01855	0.01520	0.01554	-0.01423	0.01106	0.01126
7	-0.01738	0.01514	0.01544	-0.01369	0.01117	0.01136
8	-0.01641	0.01511	0.01538	-0.01321	0.01128	0.01146
9	-0.01559	0.01511	0.01535	-0.01279	0.01140	0.01156
10	-0.01488	0.01514	0.01536	-0.01240	0.01152	0.01167
11	-0.01426	0.01518	0.01539	-0.01206	0.01164	0.01179
12	-0.01372	0.01524	0.01543	-0.01174	0.01177	0.01190
13	-0.01323	0.01532	0.01549	-0.01145	0.01190	0.01203
14	-0.01280	0.01540	0.01557	-0.01119	0.01203	0.01216
15	-0.01240	0.01550	0.01566	-0.01094	0.01217	0.01229

Table 5. Optimal trimming points

n	OLS	GLS
50	6	1
100	9	1
200	17	1
500	39	2

We propose to apply GLS procedure because the explained variable is heteroskedastic and autocorrelated. Both of the bias and variance are significantly reduced and we believe the variance attains Cramer-Rao lower bound. As another tool of efficiency improvement, we propose a trimmed least squares method, which works well for OLS, but not so clearly effective for GLS. Obviously when we are sure of the Pareto assumption, GLS or MLE is the best, but when we are not so sure, OLS may have an advantage from robustness point of view, and we believe, trimmed OLS may have a good performance because $\log S_{(i)}$ should still have larger variance for smaller i even if the underlying distribution is not Pareto. Research toward this direction is currently under way.

5. REFERENCES

Alperovich, G.A., The Size Distribution of Cities: On the Empirical Validity of the Rank-Size Rule, *Journal of Urban Economics*, 16, 232-239, 1984.

Auerbach, F. Das Gesetz der Bevolkerungskonzentration, *Petermanns Geoghische Mitteilungen*, 59, 74-76, 1913.

Gabaix, H. and Y.M. Ioannides, The Evolution of City Size Distributions, forthcoming in Handbook of Urban and Regional Economics, vol.4, 2003.

Renyi, A., On the theory of order statistics, *Acta Math. Acad. Sci. Hung.*, 4, 191-231, 1953.

Rosen, K.T. and M. Resnick, The Size distribution of Cities: An Explanation of the Pareto Low and Primacy, *Journal of Urban Economics*, 8, 165-186, 1980.

Soo, K.T., Zipf's Law for Cities: A Cross Country Investigation, mimeo, London School of Economics, 2002.

Zipf, G.K. Human Behaviour and the Principle of Least Effort, An Introduction to Human Ecology, Cambridge, MA: Addison-Wesley, 1949.