$$\frac{1}{n}\sum_{i=1}^{n}\log^4 i - \left(\frac{1}{n}\sum_{i=1}^{n}\log^2 i\right)^2$$

$$= \log^4 n - 4\log^3 n - 3\log^2 n + 6\log n - 6 + o(1)$$

$$-\{\log^2 n - 2\log n + 2 + o(1)\}^2$$

$$= 4\log^2 n - 16\log n + 20 + o(1)$$

となる．よって，相関係数の分母は

$$\sqrt{1 - \frac{\log^2 n}{2n} + O(n^{-1})4\log^2 n - 16\log n + 20 + o(1)}$$

$$= \sqrt{4\log^2 n - 16\log n + 20 + o(1)},$$

となり，相関係数の２乗は以下のように近似できる．

$$\rho^2 = \frac{4\log^2 n - 16\log n + 16 + o(1)}{4\log^2 n - 16\log n + 20 + o(1)}$$

$$\approx 1 - \frac{1}{\log^2 n - 4\log n + 5}$$

テイラー展開を用いて，相関係数は

$$\rho \approx 1 - \frac{1}{2\log^2 n} + O(\log^{-3} n)$$

と近似される．以上より，ランクサイズ回帰においては，説明変数間の相関係数が漸近的に 1 になることがわかった．

独立行政法人 経済産業研究所

小西葉子

参考文献

1. 齊藤 (梅野) 有希子，渡辺努 (2007),「企業関係と企業規模」,『経済研究』, **第 58 巻第 4 号** , pp. 302-313.

2. X. Gabaix and R. Ibragimov (2006), Log(Rank-1/2): A SimpleWay to Improve the OLS Estimation of Tail Exponents, *Harvard Institute of Economic Research Discussion Paper No. 2106.*

3. X. Gabaix and Y.M. Ioannides (2004), *The evolution of city size distributions*, Handbook of Urban and Regional Economics, Vol.**4**, Chap.53.

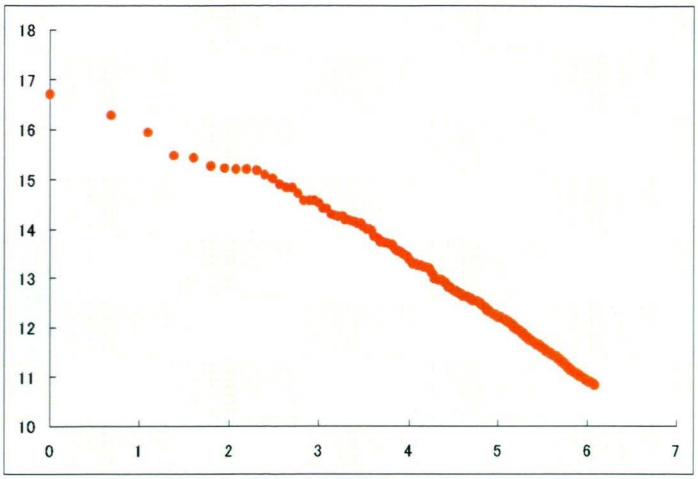
4. Y. Nishiyama and S. Osada (2004), Statistical theory of rank size rule regression under Pareto distribution, *CAEA Discussion Paper* **009**, Kyoto University.

5. Y. Nishiyama, S. Osada and Y. Sato (2007), OLS estimation and the t test revisited in rank-size rule regression, *forthcoming in Journal of Regional Science.*

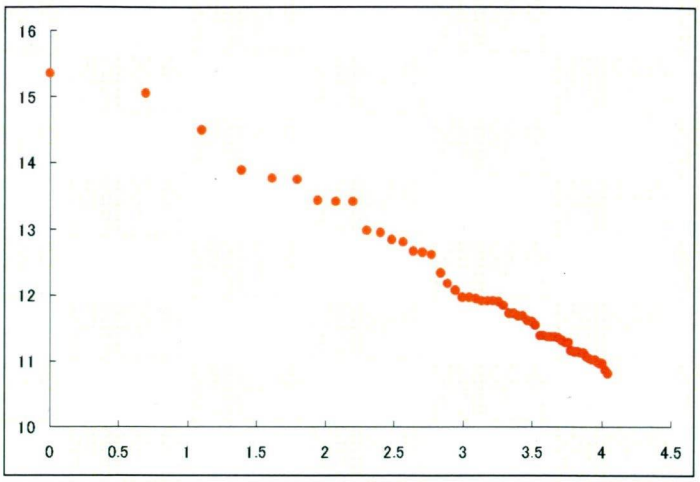
6. K.T. Rosen and M. Resnick (1980), The size distribution of cities: An explanation of the Pareto law and primacy, *Journal of Urban Economics* **8**, pp. 165-186.

7. K.T. Soo (2005), Zipf's law for cities: A cross country investigation, *Regional Science and Urban Economics* **35**, pp. 239-263.

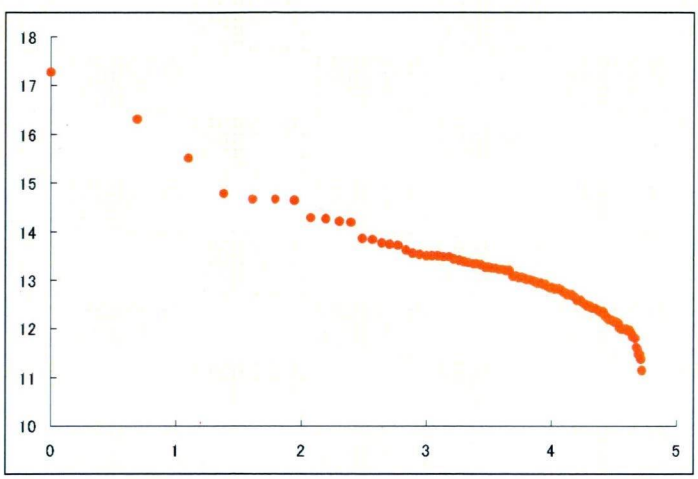
アメリカ合衆国（2005）



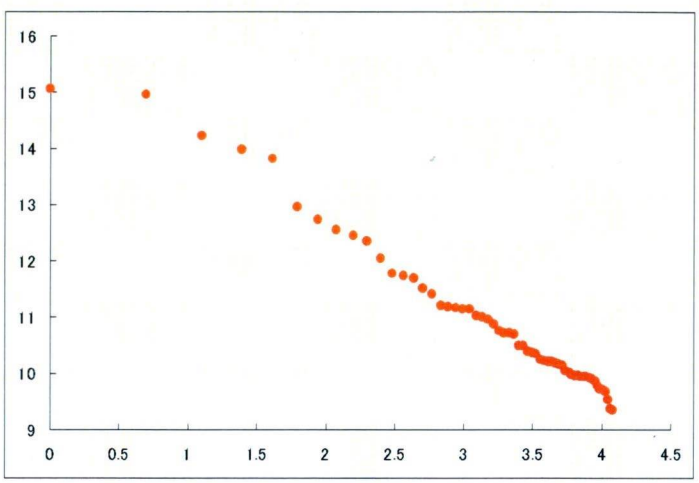
カナダ（2001）



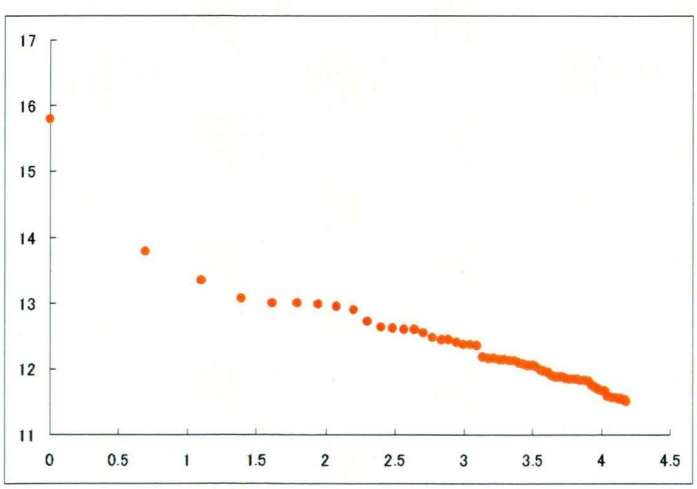
日本（2000）



オーストラリア（2001）



イギリス（2001）



フランス（2004）

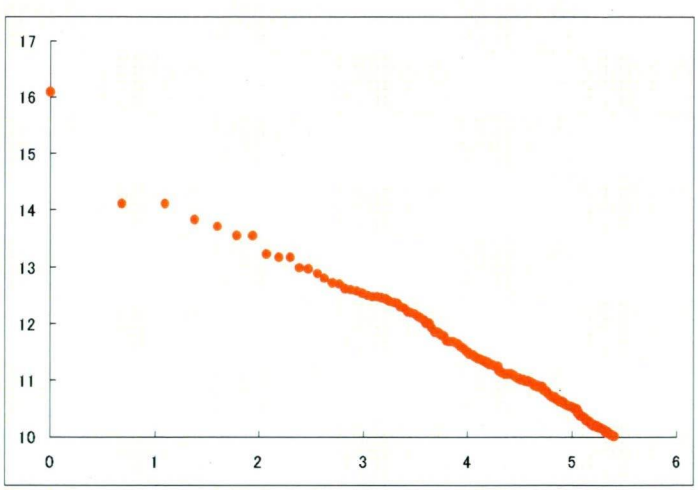
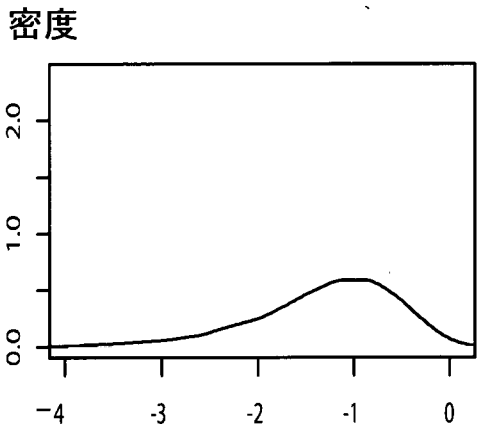
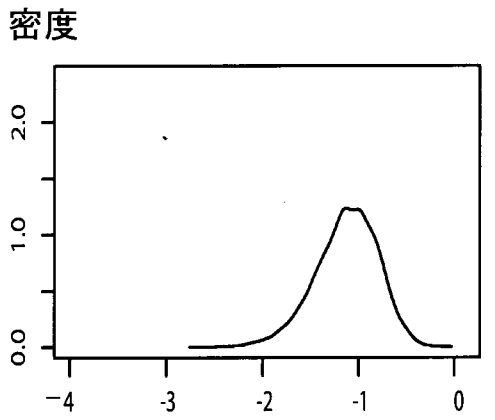


図１：各国都市の人口規模と順位の散布図
（横軸：log(rank)，縦軸：log(Size)）
http://www.citypopulation.de/より作成

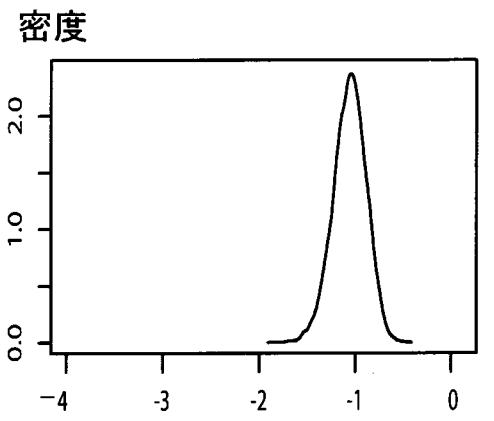
(1)一次項の密度関数



（ア）一次項α1, n=50

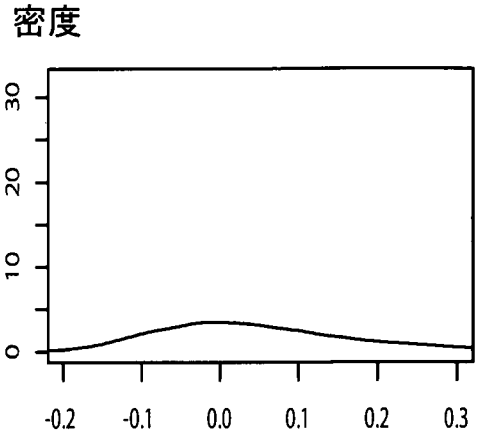


（イ）一次項α1, n=500

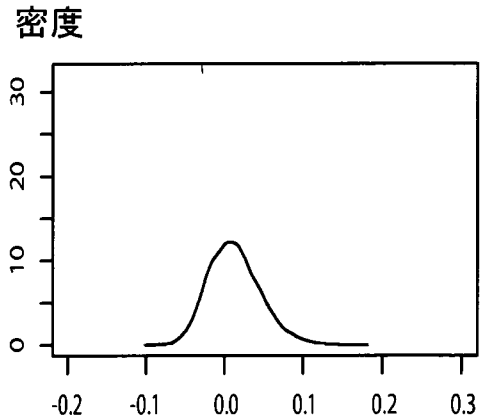


（ウ）一次項α1, n=3000

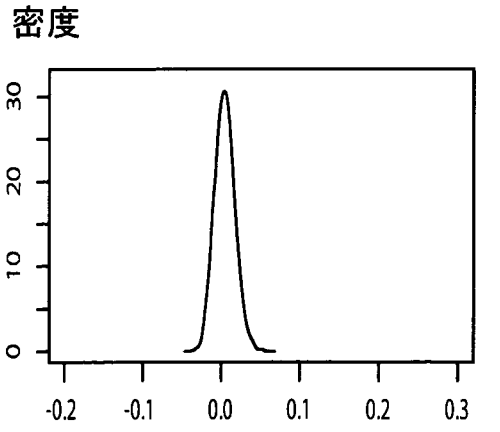
(2)二次項の密度関数



（エ）二次項α2, n=50



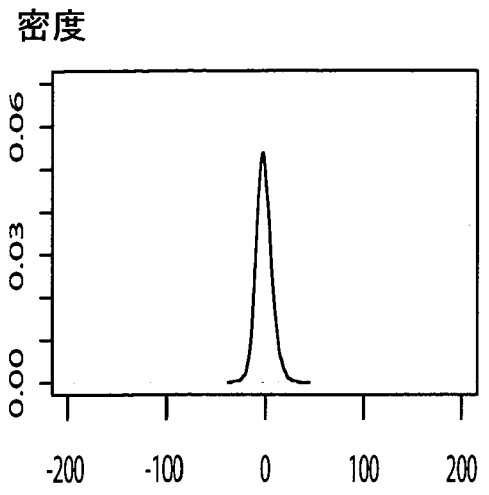
（オ）二次項α2, n=500



（カ）二次項α2, n=3000

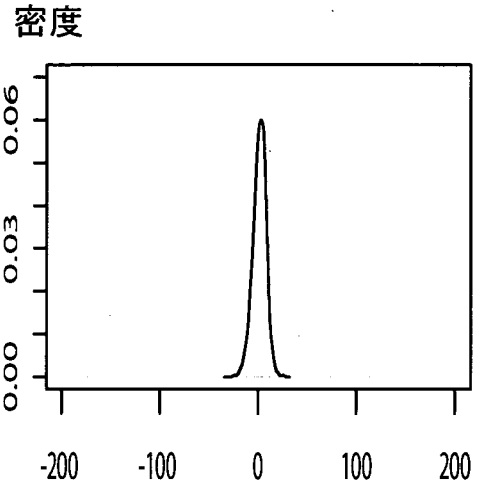
図２：(3)式　log(Size)=c+α1*log(rank)+α2*log^2(rank)の
推定値のシミュレーション結果，繰り返し回数：10000回

(1) α1のt値の密度関数:
帰無仮説：α1=-1

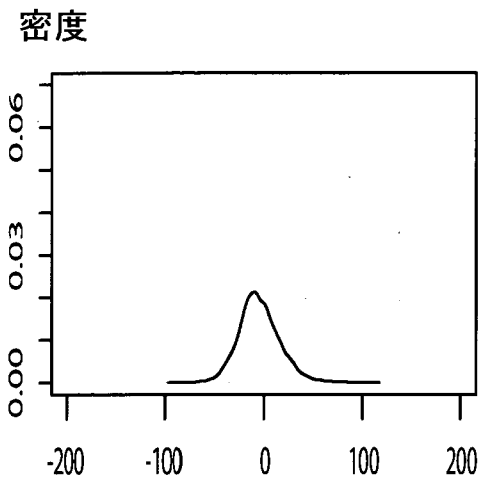


（ア） α1のt値, n=50

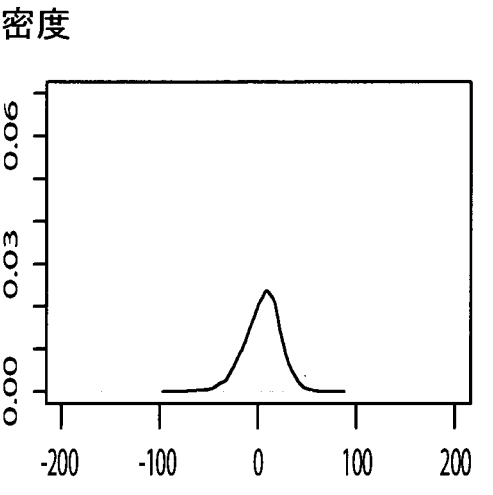
(2) α2のt値の密度関数:
帰無仮説：α2=0



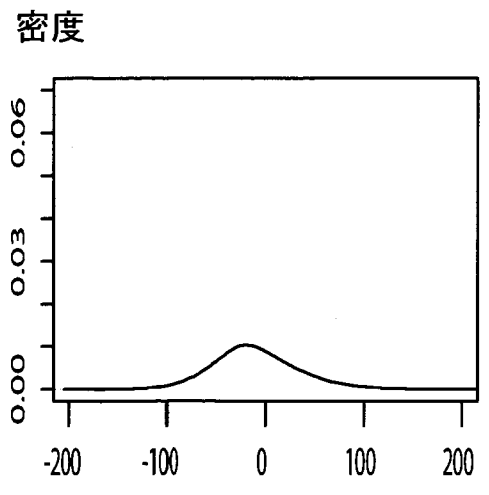
（エ） α2のt値, n=50



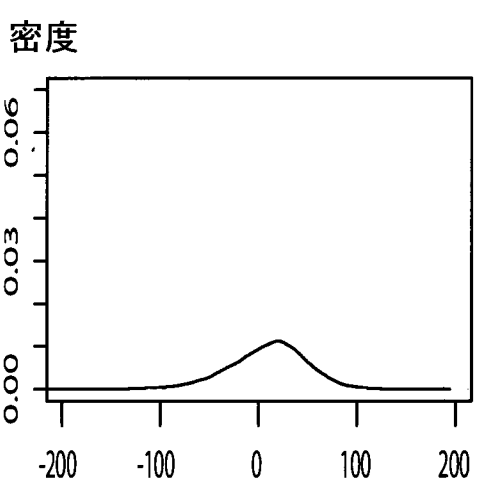
（イ） α1のt値, n=500



（オ） α2のt値, n=500



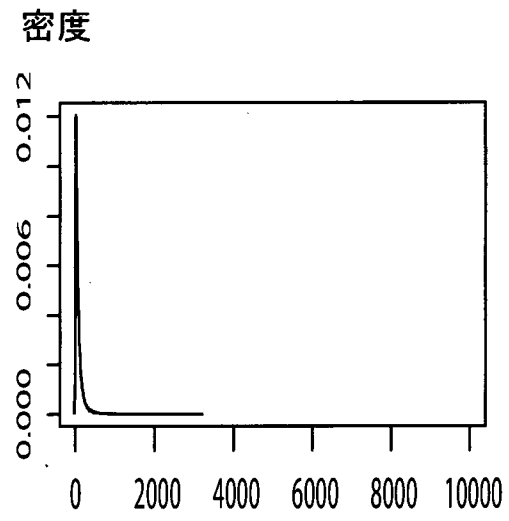
（ウ） α1のt値, n=3000



（カ） α2のt値, n=3000

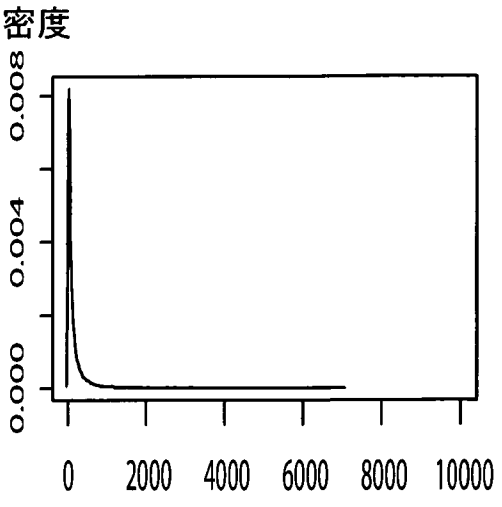
図３：(3)式　log(Size)=c+α1*log(rank)+α2*log^2(rank)の
t値のシミュレーション結果，繰り返し回数：10000回

(1) 式(3)のF値の密度関数
帰無仮説：$\alpha 1=-1$, $\alpha 2=0$

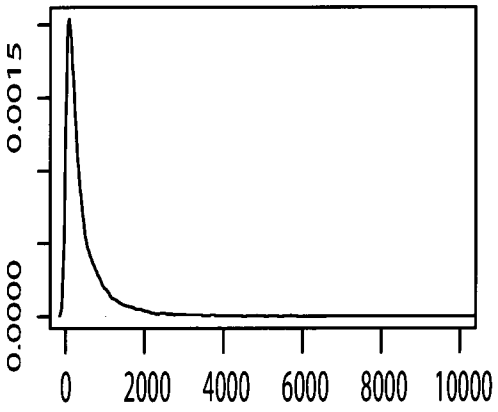


（ア）式(3), n=50

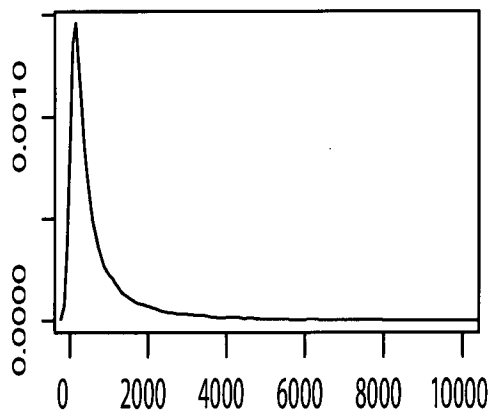
(2) 式(4)のF値の密度関数
帰無仮説：$\beta 1=-1$, $\beta 2=0$



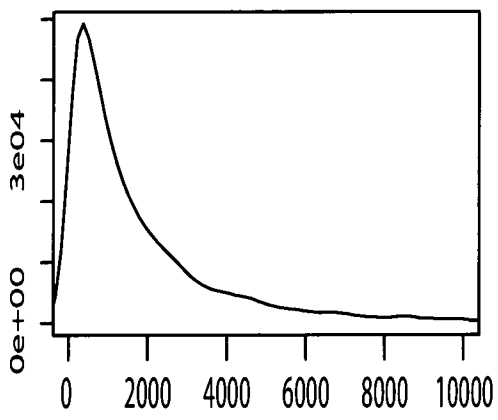
（エ）式(4), n=50



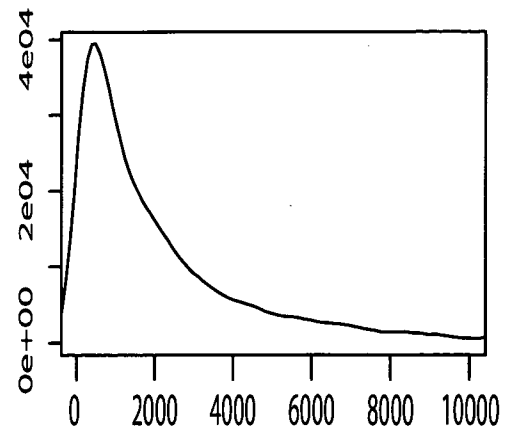
（イ）式(3), n=500



（オ）式(4), n=500



（ウ）式(3), n=3000



（カ）式(4), n=3000

図４：式(3), 式(4)のF値のシミュレーション結果：
繰り返し回数10000回

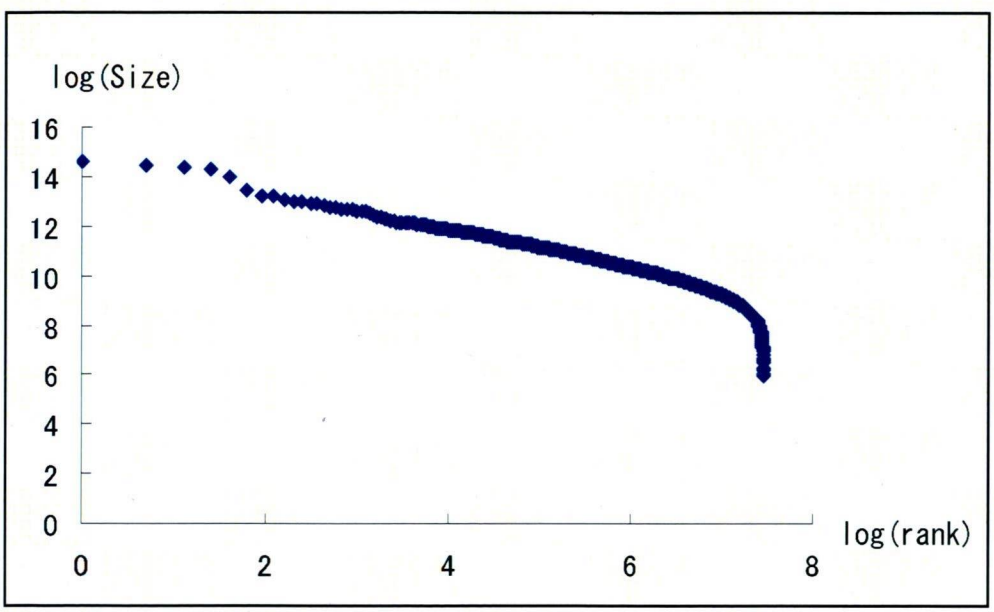


図５：上場企業の資産規模と順位の散布図

n=50	(3)式： log(Size)=c+α_1*log(rank)+α_2*log^2(rank)					(4)式： log(rank)=c+β_1*log(Size)+β_*log^2(Size)				
	α_1	α_2	α_1_t	α_2_t	SSR	β_1	β_2	β_1_t	β_2_t	SSR
平均	-1.306	0.047	-1.688	1.246	0.022	-1.007	0.008	-0.313	3.306	0.014
中央値	-1.191	0.029	-2.240	1.647	0.015	-1.007	0.031	-0.157	2.563	0.012
標準偏差	0.748	0.130	8.127	6.769	0.027	0.270	0.100	5.651	7.711	0.008
分散	0.560	0.017	66.048	45.814	0.001	0.073	0.010	31.937	59.467	0.000
尖度	1.492	1.414	0.833	0.553	66.892	0.474	7.235	0.452	1.033	5.439
歪度	-0.952	0.905	0.337	-0.276	6.055	0.005	-1.868	-0.153	0.611	1.743
範囲	6.760	1.213	76.305	61.268	0.585	2.593	1.358	52.532	71.839	0.088

n=500	(3)式： log(Size)=c+α_1*log(rank)+α_2*log^2(rank)					(4)式： log(rank)=c+β_1*log(Size)+β_*log^2(Size)				
	α_1	α_2	α_1_t	α_2_t	SSR	β_1	β_2	β_1_t	β_2_t	SSR
平均	-1.127	0.012	-4.931	4.176	0.005	-1.027	0.010	-4.949	11.394	0.004
中央値	-1.105	0.010	-6.639	5.834	0.004	-1.028	0.014	-4.700	9.698	0.003
標準偏差	0.325	0.033	20.752	18.418	0.006	0.094	0.028	15.658	22.521	0.002
分散	0.106	0.001	430.626	339.232	0.000	0.009	0.001	245.161	507.188	0.000
尖度	0.272	0.291	0.842	0.712	76.411	0.019	0.793	0.235	1.414	7.550
歪度	-0.440	0.441	0.453	-0.446	6.347	0.124	-0.800	-0.101	0.618	2.124
範囲	2.445	0.254	197.942	169.172	0.144	0.712	0.208	146.398	254.013	0.022

n=3000	(3)式： log(Size)=c+α_1*log(rank)+α_2*log^2(rank)					(4)式： log(rank)=c+β_1*log(Size)+β_*log^2(Size)				
	α_1	α_2	α_1_t	α_2_t	SSR	β_1	β_2	β_1_t	β_2_t	SSR
平均	-1.059	0.004	-9.881	9.065	0.001	-1.016	0.005	-12.336	20.961	0.001
中央値	-1.054	0.004	-13.205	12.422	0.001	-1.016	0.006	-12.719	20.380	0.001
標準偏差	0.171	0.013	43.271	39.578	0.001	0.043	0.012	33.020	44.795	0.001
分散	0.029	0.000	1872.362	1566.387	0.000	0.002	0.000	1090.311	2006.587	0.000
尖度	0.112	0.131	1.010	0.840	41.001	-0.005	0.113	0.108	0.453	8.439
歪度	-0.216	0.231	0.453	-0.446	4.843	-0.012	-0.337	0.022	0.165	2.330
範囲	1.350	0.103	461.735	418.064	0.028	0.348	0.088	271.673	436.014	0.008

表1：シミュレーション結果：係数値(α_1、α_2、β_1、β_2)、各t統計量(α_1_t、α_2_t、β_1_t、β_2_t)、回帰残差の二乗和(SSR)の記述統計量

一次項に対するt統計量： (3)式の帰無仮説はα_1=-1, (4)式の帰無仮説β_1=-1.
二次項に対するt統計量： (3)式の帰無仮説はα_2=0, (4)式の帰無仮説はβ_2=0.
nはサンプル数, 繰り返し回数は10000回

	α_1		α_2		β_1		β_2	
n=50	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
10%	[−∞, −13.74)	(12.42, ∞]	[−∞,−10.8)	(11.38, ∞]	[−∞,−9.81)	(8.68, ∞]	[−∞,−7.9)	[17.05, ∞]
5%	[−∞,−16.37)	(16.03, ∞]	[−∞,−13.12)	(13.7, ∞]	[−∞,−11.82)	(10.56, ∞]	[−∞,−9.8)	[20.6, ∞]
1%	[−∞,−23.07)	(23.23, ∞]	[−∞,−18.48)	(18.42, ∞]	[−∞,−16.49)	(14.16, ∞]	[−∞,−13.98)	[28.8, ∞]

	α_1		α_2		β_1		β_2	
n=100	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
10%	[−∞,−18.17)	(16.46, ∞]	[−∞,−13.98)	(15.29, ∞]	[−∞,−14.08)	(11.03, ∞]	[−∞,−10.81)	[24.79, ∞]
5%	[−∞,−21.36)	(21.38, ∞]	[−∞,−17.81)	(17.86, ∞]	[−∞,−16.87)	(13.37, ∞]	[−∞,−13.44)	[30.36, ∞]
1%	[−∞,−28.55)	(31.54, ∞]	[−∞,−25.94)	(23.49, ∞]	[−∞,−22.96)	(18.50, ∞]	[−∞,−19.19)	[43.44, ∞]

	α_1		α_2		β_1		β_2	
n=200	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
10%	[−∞,−24.4)	(21.25, ∞]	[−∞,−18.6)	(20.8, ∞]	[−∞,−19.93)	(13.48, ∞]	[−∞,−13.97)	(34.44, ∞]
5%	[−∞,−39.62)	(40.5, ∞]	[−∞,−34.36)	(32.67, ∞]	[−∞,−32.22)	(24.26, ∞]	[−∞,−25.42)	(59.18, ∞]
1%	[−∞,−28.65)	(27.76, ∞]	[−∞,−23.68)	(24.37, ∞]	[−∞,−23.6)	(16.48, ∞]	[−∞,−17.72)	(42.04, ∞]

	α_1		α_2		β_1		β_2	
n=500	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
10%	[−∞,−36.7)	(30.98, ∞]	[−∞,−27.73)	(32.05, ∞]	[−∞,−30.64)	(19.88, ∞]	[−∞,−21.95)	(50.94, ∞]
5%	[−∞,−42.17)	(40.32, ∞]	[−∞,−36.05)	(37.18, ∞]	[−∞,−36.30)	(25.17, ∞]	[−∞,−27.63)	(61.06, ∞]
1%	[−∞,−55.09)	(59.80, ∞]	[−∞,−54.68)	(48.15, ∞]	[−∞,−46.92)	(35.13, ∞]	[−∞,−39.91)	(83.22, ∞]

	α_1		α_2		β_1		β_2	
n=1000	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
10%	[−∞,−48.64)	(40.10, ∞]	[−∞,−36.28)	(43.24, ∞]	[−∞,−41.98)	(26.17, ∞]	[−∞,−28.98)	(66.11, ∞]
5%	[−∞,−57.03)	(50.79, ∞]	[−∞,−67.01)	(67.54, ∞]	[−∞,−49.54)	(32.94, ∞]	[−∞,−36.89)	(80.55, ∞]
1%	[−∞,−75.85)	(75.61, ∞]	[−∞,−46.23)	(50.43, ∞]	[−∞,−66.78)	(44.90, ∞]	[−∞,−53.73)	(114.29, ∞]

	α_1		α_2		β_1		β_2	
n=3000	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
10%	[−∞,−75.09)	(66.23, ∞]	[−∞,−59.92)	(68.60, ∞]	[−∞,−65.91)	(42.09, ∞]	[−∞,−50.27)	(95.34, ∞]
5%	[−∞,−87.96)	(83.91, ∞]	[−∞,−76.78)	(79.53, ∞]	[−∞,−75.67)	(52.73, ∞]	[−∞,−65.03)	(112.89, ∞]
1%	[−∞,−115.14)	(126.25, ∞]	[−∞,−113.91)	(104.92, ∞]	[−∞,−97.50)	(72.97, ∞]	[−∞,−91.58)	(149.50, ∞]

表2:シミュレーションによるt値の棄却域(繰り返し回数は10000回)

(3)式: log(Size)=c+α_1*log(rank)+α_2*log^2(rank)、(4)式: log(rank)=c+β_1*log(Size)+β_*log^2(Size)
一次項に対するt検定: (3)式の帰無仮説はα_1=-1, (4)式の帰無仮説β_1=-1の両側検定.
二次項に対するt検定: (3)式の帰無仮説はα_2=0, (4)式の帰無仮説はβ_2=0の両側検定.

	(3)式			(4)式		
n	10%	5%	1%	10%	5%	1%
50	88.24	279.13	577.85	374.16	603.09	1407.78
100	387.18	574.49	1209.84	776.63	1246.99	2638.46
200	554.05	800.71	1575.5	1011.48	1584.86	3464.72
500	1113.95	1621.68	3187.27	1853.66	2888.15	5898.67
1000	2006.01	2922.18	5405.54	2945.61	4500.72	9021.17
3000	4597.47	6715.73	12816.42	6197.66	8922.91	16318.68

表3:シミュレーションによるF値の棄却域(繰り返し回数は10000回)

(3)式: log(Size)=c+α_1*log(rank)+α_2*log^2(rank)
(4)式: log(rank)=c+β_1*log(Size)+β_*log^2(Size)
各F検定は, (3)式については、帰無仮説はα_1=-1、α_2=0の両側検定.
各F検定は, (4)式については、帰無仮説はβ_1=-1、β_2=0の両側検定.

式番号	推定結果			通常の棄却域			シミュレーションによる棄却域		
	一次項 (t値)	二次項 (t値)	F値	一次項 t検定	二次項 t検定	F検定	一次項	二次項	F検定
(2)	−1.056 (−6.18)			棄却			棄却されない		
(3)	0.401 (31.40)	−0.132 (−33.09)	578.96	棄却	棄却	棄却	棄却されない	棄却されない	棄却されない
(4)	2.163 (139.30)	−0.152 (−132.78)	11610.97	棄却	棄却	棄却	棄却	棄却	棄却

表4: 上場企業の資産データ(2006年)を用いたランクサイズ回帰結果とパレート性の検定
サンプル数: 1736, 括弧の中はt値、検定は5%有意水準で行っている.

Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Mathematics and Computers in Simulation 79 (2009) 2869–2878
MATHEMATICS AND COMPUTERS IN SIMULATION
www.elsevier.com/locate/matcom

Hypothesis testing in rank-size rule regression

Yoko Konishi [a,*], Yoshihiko Nishiyama [b]

[a] *Research Institute Economy, Trade and Industry, 1-3-1 Kasumigaseki, Chiyoda-ku,Tokyo, Japan*
[b] *Institute of Economic Research, Kyoto University, Kyoto, Japan*

Received 17 August 2008; received in revised form 17 October 2008; accepted 21 October 2008
Available online 20 November 2008

Abstract

This note examines testing methods for Paretoness in the framework of rank-size rule regression. Rank-size rule regression describes a relationship found in the analysis of various topics such as city population, words in texts, scale of companies and so on. In terms of city population, it is basically an empirical rule that $\log(S_{(i)})$ is approximately a linear function of $\log(i)$ where $S_{(i)}$ is the number of population of ith largest city in a country. This is closely related to the so-called Zipf's law. It is known that this kind of empirical observation is found when the city population is a random variable following a Pareto distribution. Thus one may be willing to test if city size has a Pareto distribution or not. Rosen and Resnick [K.T. Rosen, M. Resnick, The size distribution of cities: an explanation of the Pareto law and primacy, Journal of Urban Economics 8 (1980), 165–186] and Soo [K.T. Soo, Zipf's law for cities: a cross country investigation, Regional Science and Urban Economics (35) 2005, 239–263] regress $\log(S_{(i)})$ on $\log(i)$ and $\log^2(i)$ and test the null of Paretoness by standard t-test for the latter regressor. It is found that t-statistics take large values and the Paretoness is rejected in many countries. We study the statistical properties of the t-statistic and show that it explodes asymptotically, in fact, by simulation and thus the t-test does not provide a reasonable testing procedure. We propose an alternative test statistic which seems to be asymptotically normally distributed. We also propose a test with the null hypothesis that the city size distribution is Pareto with exponent unity, which is a modification of the F-test.

Keywords: Rank-size rule regression; Paretoness

1. Introduction

This paper examines testing methods for Paretoness in the framework of rank-size rule regression which is done by Rosen and Resnick [5] and Soo [6]. Rank-size rule regression describes a relationship often found in the analysis of various topics such as city population, words in texts, scale of companies and so on. In terms of city population, it is basically an empirical fact that $\log(S_{(i)})$ is approximately a linear function of $\log(i)$, namely,

$$\log(S_{(i)}) \approx \alpha_0 + \alpha_1 \log(i),$$

where $S_{(i)}$ is the population of the ith largest city in a country. This is closely related to the so-called Zipf's law. This relationship motivates to run the regression

$$\log(S_{(i)}) = \alpha_0 + \alpha_1 \log(i) + u_i \tag{1}$$

* Corresponding author.
E-mail address: konishi-yoko@rieti.go.jp (Y. Konishi).

0378-4754/$36.00 © 2008 IMACS. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.matcom.2008.10.012

to estimate the parameters α_0 and α_1. It is observed that α_1 and α_0 are approximately -1 and $\log(S_{(1)})$, respectively. It is known that this relationship holds when the city population is a random variable following a Pareto distribution. Thus one may be willing to test if city size has a Pareto distribution or not. Rosen and Resnick [5] examine the Paretoness for many countries based on the t-test in a reverse regression

$$\log(i) \approx \beta_0 + \beta_1 \log(S_{(i)}) + \beta_2 \log^2(S_{(i)}). \tag{2}$$

They test if $\beta_1 = -1$ by t-test in the first step and test if $\beta_2 = 0$ by t-test again[1]. The latter t-test corresponds to the test for Paretoness, which comes from the following motivation. If Pareto distribution is the correct specification, $\log(S_{(i)})$ and $\log(i)$ have a linear relationship asymptotically (see Nishiyama et al. [4], Proposition 1), and thus if the relationship is nonlinear, the Paretoness hypothesis must be incorrect. Therefore, $\beta_2 = 0$ is an expression describing the null of Paretoness. In many countries, both the null hypotheses are rejected in their studies.

The statistical properties of the OLS estimation for (1) have been studied by some papers. Gabaix and Ioannides [2] obtain the asymptotic distribution of the estimators, while Gabaix and Ibragimov [1] point out that estimators have asymptotic bias, and propose a simple way of removing it based on the asymptotic expansions of the estimators.

Nishiyama et al. [4] examine the properties of t-statistic for α_1 in regression (1) and show that it asymptotically explodes, in fact, since the estimate of the error variance

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

decays to zero as $n \to \infty$ in probability. Therefore, it does not make sense to compare t-value with quantiles of the t-distribution. Such a result comes from the fact that $\log(S_{(i)})$ is autocorrelated and heteroskedastic in a manner described in Nishiyama et al. [4]. They apply a modified test for $\beta_1 = -1$ with correct size, which does not reject the null in many countries, unlike results of Rosen and Resnick [5] and Soo [6].

In this paper, we first examine the validity of the t-test for nonlinear (square) term in rank-size rule regression. More precisely, we study the properties of t-statistics of α_2 in the following regression:

$$\log(S_{(i)}) \approx \alpha_0 + \alpha_1 \log(i) + \alpha_2 \log^2(i). \tag{3}$$

We take this formulation rather than (2) because $\log(S_{(i)})$, not $\log(i)$, is the random variable in this problem and thus it is more natural to regard it as the regressand. We investigate the statistical behavior of the t-statistic for α_2 by simulation. We show similar phenomenon occurs in the case of (2) as found in Nishiyama et al. [4], namely the variance estimate, or the square of the denominator of t-value, converges to zero and thus t-value tends to take very large values. We secondly examine the distribution of suitably normalized statistics of the OLS esitmator for α_2. We find it is asymptotically normally distributed. We propose to test the null of $\alpha_2 = 0$ using this property and apply the methods to empirical data of the Japanese Urban Employment Areas (UEAs), and UK and US population of Urban Aggromelation. The new test seems to work reasonably well in our results.

The following section describes some tests for Paretoness including rank-size rule regression-based tests. Section 3 shows simulation results investigating their small sample performances. Some empirical application is presented in Section 4. Section 5 concludes with some remarks.

2. Tests for paretoness

We consider the problem oF-testing for Paretoness in the context of the rank-size rule regression (3). If Pareto specification is correct, we must have $\alpha_2 = 0$, as explained above, and thus one way of describing the null of Paretoness is

$$H_0 : \alpha_2 = 0. \tag{4}$$

[1] We think it would be more natural to regard $S_{(i)}$, the population, as a random variable than i, the rank, so that we consider the formulation (1) here, making $\log(S_{(i)})$ be the dependent variable. Rosen and Resnick [5] and some other previous researcher consider the reverse regression version where $\log(i)$ is regressed on $\log(S_{(i)})$. Since $\log(S_{(i)})$ and $\log(i)$ must have a linear relation under the null of Paretoness, they extend the reverse regression as in (2) and propose to examine the Paretoness by a siginificance test for additional regressor of $\log^2(S_{(i)})$.

Rosen and Resnick [5] take this approach to examine the Paretoness. They also consider models including $\log^3(i)$ in addition to $\log(i)$ and $\log^2(i)$. Soo [6] follows their approach to examine an updated dataset. In the standard regression analysis, such kind of null hypothesis is tested using t-value. Under the i.i.d. normal disturbances, the t-value has a t-distribution with $n-3$ degree of freedom, and without the normality, it has an asymptotic standard normal distribution. But in the present problem, the t-statistic does not have a t-distribution, and more seriously, we show later that it seems to explode asymptotically. We will propose an alternative test statistic which is asymptotically normally distributed.

In the current problem, one possible alternative approach is to set the null hypotheses is as follows:

$$H_0 : (\alpha_1, \alpha_2) = (-1, 0). \tag{5}$$

This setting tries to test the null that the underlying distribution is a Pareto distribution with unit exponent. (4) only considers Paretoness with any exponent value. If we would like to test the joint hypothesis as above, an F-test is the natural procedure. It is a Wald test checking if $\hat{\alpha}_1 \approx -1$ and $\hat{\alpha}_2 \approx 0$. Konishi and Nishiyama [3] proposed to use the F-statistic, but one difficulty is that it does not have the standard distribution. They construct the critical region by simulation and apply it to empirical data.

A classical standard method of testing goodness-of-fit is the Kolmogorov–Smirnov test. We briefly mention this test later.

2.1. *t-Test and its modification*

We may like to apply the standard testing procedure for (4), namely a t-test based on an OLS estimation of (3). The t-value for α_2 is defined by

$$t_2 = \frac{\sqrt{n}\hat{\alpha}_2}{sQ^{22}}$$

where $\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2$ are the OLS estimates of $\alpha_1, \alpha_2, \alpha_3$, respectively,

$$s^2 = \frac{1}{n-3}\sum_{i=1}^{n}\{\log(S_{(i)}) - \hat{\alpha}_0 - \hat{\alpha}_1 \log(i) - \hat{\alpha}_2 \log^2(i)\}^2,$$

and Q^{22} is the 3,3 element of $(X'X)^{-1}$ with

$$X = \begin{pmatrix} 1 & \log(1) & \log^2(1) \\ 1 & \log(2) & \log^2(2) \\ \vdots & \vdots & \vdots \\ 1 & \log(n) & \log^2(n) \end{pmatrix}.$$

We will show by simulation that t_2 appears to explode asymptotically similarly to the t-value examined by Nishiyama et al. [4]. To avoid this problem, we propose to use $S \equiv \sqrt{2n}\hat{\alpha}_2$ as the test statistic. The value 2 in the square root comes from the asymptotic variance of $\hat{\alpha}_2$.

2.2. *F-test*

In the standard regression analysis, (5) is tested by $F-$test. The F-value in the present context is defined as

$$F = \frac{(R\hat{\alpha} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\alpha} - r)/2}{s^2}$$

where

$$\hat{\alpha} = (\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2)',$$
$$r = (-1, 0)',$$
$$R = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$
$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\{\log(S_{(i)}) - \hat{\alpha}_0 - \hat{\alpha}_1 \log(i) - \hat{\alpha}_0 \log^2(i)\}^2.$$

The F-statistic has an F-distribution under the n.i.d. disturbances, while it has an χ^2 distribution under i.i.d. disturbances as $n \to \infty$. In the present problem, however, the F-value neither has an F-distribution, nor follows an χ^2distribution, even approximately. It is obviously because of the non-i.i.d. structure of disturbances

$$\epsilon_i = \log(S_{(i)}) - \alpha_0 - \alpha_1 \log(i) - \alpha_2 \log^2(i).$$

We show by simulation that the F-value seems to explode asymptotically, in fact, similarly to the t-value examined by Nishiyama et al. [4]. To avoid this problem, we propose the following modified test statistic that is easy to calculate, and is the numerator of the F-statistic,

$$T = \frac{(R\hat{\alpha} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\alpha} - r)}{2}. \tag{6}$$

This statistic is motivated by the fact that F explodes asymptotically because it appears s^2 tends to zero asymptotically. The critical value is calculated by simulation. The asymptotic distribution looks quite like the $\chi^2(2)$ distribution, but it has a heavier tail. The theoretical investigation is currently underway.

2.3. *Kolmogorov–Smirnov test for goodness-of-fit*

From a statistical point of view, we can also test whether a sample has any particular distribution by various methods such as the Kolmogorov–Smirnov (KS) test and its variants. We briefly explain them. In the next section, we compare a small sample performance of these methods by simulation when data is a randomly sampled from a Pareto distribution with parameter unity. It is known that this test does not have much power.

Given a random sample, let n, $F_n(x)$, $F^0(x)$ be the sample size, empirical cumulative distribution, and the null distribution, respectively. Then the test statistic is

$$KS = \sqrt{n} \sup_{-\infty<x<\infty} |F_n(x) - F^0(x)|.$$

This has the asymptotic null distribution

$$H(x) = \left\{1 - 2\sum_{j=1}^{\infty}(-1)^{j-1} \exp(-2j^2x^2)\right\} I_{(0,\infty)}(x).$$

This is a general nonparametric test of goodness-of-fit.

3. Monte carlo simulation

This section provides some simulation results to examine the performance of the tests described in the previous section. Our primary interest is in the testing procedure, but we first examine the properties of the OLS estimator of (3) in Section 3.1 because the test statistic involves them. Section 3.2 gives simulation results for the t-test and its modification S. Section 3.3 provides simulation results of the F-statistic and its modification, T.

Table 1
Quantiles of t-statistic by simulation, from Table 2 in Konishi and Nishiyama [3] (null: $\alpha_2 = 0$ in $\log(S_{(i)}) = \alpha_0 + \alpha_1 \log(i) + \alpha_2 \log^2(i)$).

Size	$n = 50$	$n = 100$
10%	$(-\infty, -10.8], [11.38, \infty)$	$(-\infty, -13.98], [15.29, \infty)$
5%	$(-\infty, -13.12], [13.7, \infty)$	$(-\infty, -17.81], [17.86, \infty)$
1%	$(-\infty, -18.48]$, $[18.42, \infty)$	$(-\infty, -25.94]$, $[23.49, \infty)$
Size	$n = 200$	$n = 500$
10%	$(-\infty, -18.6], [20.8, \infty)$	$(-\infty, -27.73], [32.05, \infty)$
5%	$(-\infty, -34.36], [32.67, \infty)$	$(-\infty, -36.05], [37.18, \infty)$
1%	$(-\infty, -23.68], [24.37, \infty)$	$(-\infty, -54.68], [48.15, \infty)$
Size	$n = 1000$	$n = 3000$
10%	$(-\infty, -36.28], [43.24, \infty)$	$(-\infty, -59.92], [68.60, \infty)$
5%	$(-\infty, -67.01], [67.54, \infty)$	$(-\infty, -76.78], [79.53, \infty)$
1%	$(-\infty, -46.23], [50.43, \infty)$	$(-\infty, -113.91], [104.92, \infty)$

3.1. Properties of the estimator

Tests introduced in Sections 2.1 and 2.2 use an OLS estimator of (3) as shown in the next section, and thus we first investigate the statistical properties of the estimator. The consistency property under the Paretoness is studied in Konishi and Nishiyama [3]. Namely, if the original sample is a random sample from a distribution function

$$F(x) = 1 - \frac{1}{x},$$

the estimators have the following probability limits:

$$\hat{\alpha}_1 \to^p -1, \quad \hat{\alpha}_2 \to^p 0.$$

In this paper, we examine the order of convergence as well as the limiting distribution of the OLS estimators.

Figs. 1 and 2 show the histogram of $\sqrt{n}(\hat{\alpha}_1 + 1)/\log n$ and $\sqrt{n}\hat{\alpha}_2$ when the data is generated from the Pareto distribution with exponent unity. We find that both distributions look like normal distributions and the normalization constants $\sqrt{n}/\log n$ and $\sqrt{n}$, respectively, work well as the stabilizers. Obviously, the current problem is non-standard regression with autocorrelated and heteroskedastic disturbances, and thus any convergence rates could appear. In view

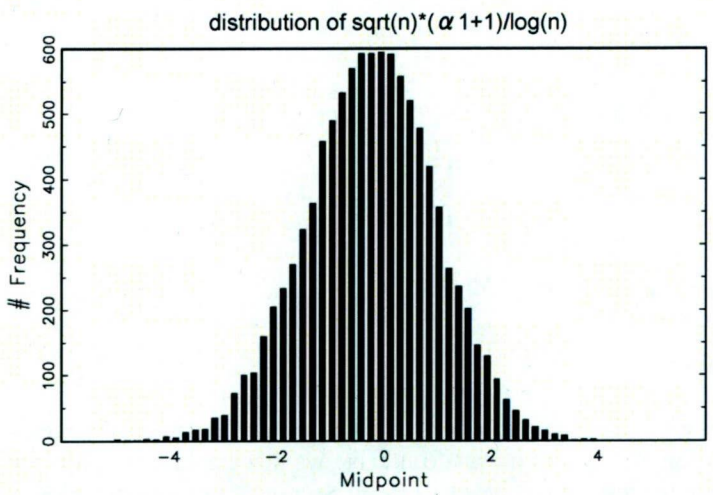


Fig. 1. Histgram of $\sqrt{n}(\alpha_1 + 1)/\log(n)$.

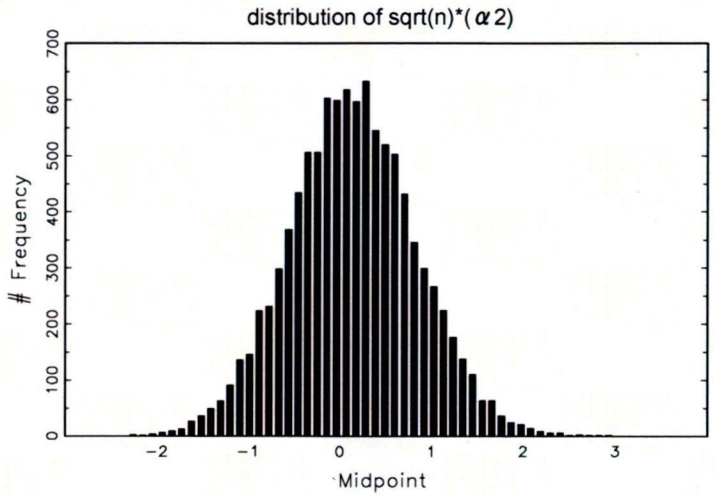


Fig. 2. Histgram of $\sqrt{n}(\alpha_2)$.

Table 2
Descriptive statistics of $\sqrt{n}(\hat{\alpha}_1 + 1)/\log n$.

n	Mean	Var.	3rd moment	4th moment
100	−0.502	1.533	−1.354	8.919
200	−0.510	1.435	−1.024	7.409
300	−0.497	1.414	−0.879	6.912
500	−0.479	1.396	−0.764	6.766
1000	−0.438	1.351	−0.623	6.346
3000	−0.398	1.350	−0.351	5.655
5000	−0.388	1.377	−0.324	6.070
7000	−0.366	1.398	−0.304	6.047
10000	−0.360	1.457	−0.188	6.605

of the moments of the normalized statistics provided in Tables 2 and 3, we anticipate the following asymptotic normality results:

$$\frac{\sqrt{n}}{\log n}(\hat{\alpha}_1 + 1) \to^d N(0, 1.5),$$

$$\sqrt{n}\hat{\alpha}_2 \to^d N(0, 0.5). \tag{7}$$

Their theoretical verification is currently under way.

Table 3
Descriptive statistics of $\sqrt{n}\hat{\alpha}_2$.

n	Mean	Var.	3rd moment	4th moment
100	0.306	0.675	0.390	1.746
200	0.313	0.600	0.277	1.323
300	0.308	0.582	0.232	1.176
500	0.293	0.553	0.191	1.064
1000	0.266	0.518	0.151	0.942
3000	0.241	0.492	0.078	0.750
5000	0.234	0.491	0.072	0.776
7000	0.220	0.493	0.067	0.757
10000	0.214	0.507	0.040	0.800

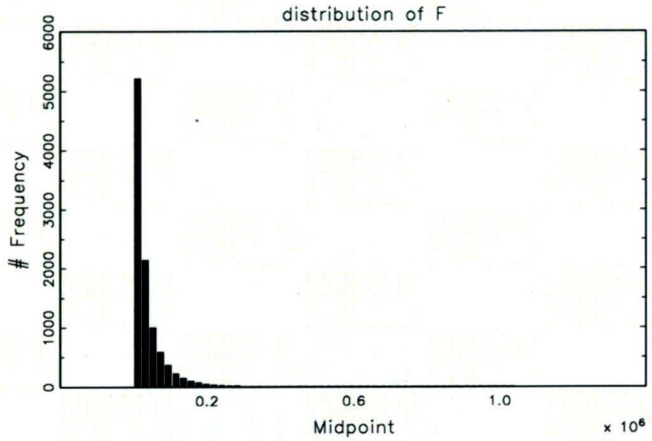


Fig. 3. Distribution of F-statistic.

3.2. *t-Test and a new test (S) with asymptotic normality*

t_2 is the t-value for α_2. Under the i.i.d. setting, it is asymptotically normally distributed. However, it appears its variance is getting larger asymptotically as shown in Table 1, which reports the quantile of t_2 for different sample sizes. For example, the 97.5% quantile of t_2 is 13.7 when $n = 50$, while it becomes 37.18 for $n = 500$, and it further increases to 79.53 for $n = 3000$. We note that the 97.5% quantile of the t-distribution is approximately 2, which converges to 1.96, the 97.5% quantile of N(0,1). As the table shows, any of the 10%, 5%and1% quantiles becomes larger as the sample size increases. The reason is found to be the same as in Nishiyama et al. [4], namely s^2 decays to zero. Therefore, we cannot apply a standard t-test for the null (4). If we did so, we would tend to reject the null too often, namely the size is much larger than the prefixed nominal size. Furthermore, the size approaches 100% asymptotically. Thus a t-test using the t_2 value does not work.

Noting the fact that $s^2 \to^p 0$ is the reason why t_2 explodes, we can think of using only the numerator of t_2 as the test statistic. We propose to use the result of (7). As $\sqrt{2n}\hat{\alpha}_2$ is approximately normally distributed under the null, we simply compare the value with quantiles of N(0,1).

3.3. *F and F-type tests*

Taking into account that we are interested in not only Paretoness ($\alpha_2 = 0$) but also the unit Pareto exponent ($\alpha_1 = -1$), Konishi and Nishiyama [3] point out that it might be another natural approach to test the joint hypothesis that the coefficient of $\log(i)$ is -1 and that of squared $\log(i)$ is 0, rather than Rosen and Resnick [5] and Soo [6]. An F-test is a standard approach to such a kind oF-testing problem of joint null hypothesis. However, there is a difficulty in that it does not have the standard distribution in the present problem because of the heteroskedasticity and autocorrelation. Therefore, they construct the critical region of the F-test by simulation. The critical values, however, depend on sample size, and thus they need to obtain them for each sample size, which is inconvenient for empirical researchers. We would like to use a statistic with some limiting distribution. T, we believe, has such a property.

Figs. 3 and 4 provide the null distribution of F and T defined by (6) for $n = 100,000$. Tables 4 and 5 provide simulation estimates of the first moments of F and T under the null of the Pareto distribution with unit exponent. The former table presents $E(F)$, $Var(F)$, $E\{F - E(F)\}^3$, $E\{F - E(F)\}^4$ and the same values for T in Table 5. We find that all the moments of the F-statistic rapidly increase as n becomes larger, while those of T do not. Obviously, F may not be a convenient statistic for testing the null of Paretoness. T, on the other hand, it looks to converge to a distribution. Therefore, we think that T must be a valid test statistic. We implement a Q–Q plot of the empirical distribution of T when $n = 100,000$ against an χ^2 distribution with two degree of freedom, which is Fig. 5. The plot suggests the χ^2 distribution with two degree of freedom may be a reasonably good approximation to the distribution of T multiplied by a constant.

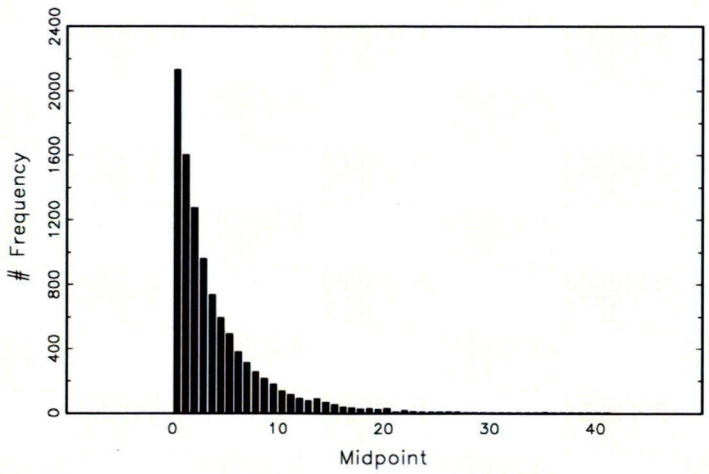


Fig. 4. Distribution of T-statistic.

Table 4
Descriptive statistics of F.

n	Mean	Var.	3rd moment	4th moment
100	138.50	51095.80	81326607	2.55e+011
200	231.88	100887.78	1.311e+008	3.18e+011
300	311.43	195611.82	4.83e+008	2.79e+012
500	478.84	439771.75	1.20e+009	6.10e+012
1000	816.76	1294226.5	6.42e+009	6.27e+013
3000	1957.59	7876407.2	1.12e+011	3.55e+015
5000	2938.63	16421381	2.90e+011	1.01e+016
7000	3918.47	29185099	6.38e+011	2.66e+016
10000	5328.51	55249185	1.78e+012	1.16e+017

4. Empirical applications

We apply the methods proposed above to empirical data of the Japanese Urban Employment Areas (UEAs) and UK and US population of Urban Aggromelation. We estimate the rank-size regression (4) for these data. In the Japanese case, we re-examine the OLS results in Nishiyama et al. [4]. About the UK and US, we use the updated Urban Aggromeration data which Soo [6] and Nishiyama et al. [4] analyzed. These datasets are available at the website of the Center for Spatial Information Science at the University of Tokyo and in Tohmas Brinkhoff (2004): City Population. Fig. 6 shows the scatter plot of $\log(S_{(i)})$ and $\log(i)$ with respect to the cities of Japan, the UK and the US for years 2000, 2005 and 2007, respectively. The correlation coefficients between $\log(S_{(i)})$ and $\log(i)$ are close to -1, and rank-size rule

Table 5
Descriptive statistics of T.

n	Mean	Var.	3rd moment	4th moment
100	3.057	19.248	309.91	8519.34
200	3.385	21.393	356.21	11198.63
300	3.475	21.626	344.78	10100.31
500	3.731	24.220	447.86	16738.66
1000	3.769	23.539	447.29	17991.47
3000	3.865	21.879	305.25	8218.55
5000	3.953	22.265	319.14	8998.00
7000	3.963	21.868	279.59	6715.27
10000	4.081	23.059	316.44	8926.79

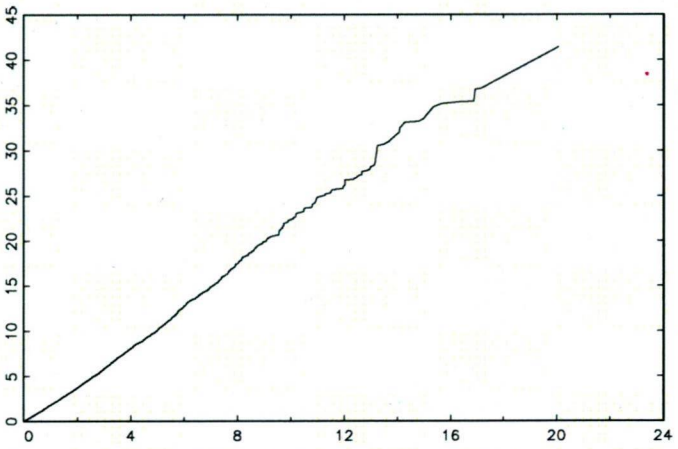


Fig. 5. Q–Q plot of T vs. χ^2.

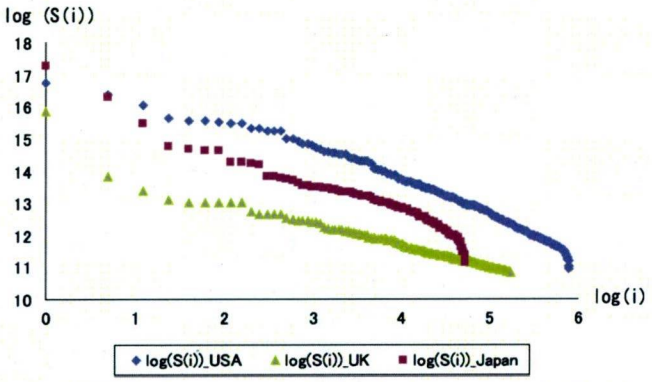


Fig. 6. Rank-size rule for Japanese UEA and UK and US's population of Urban Aggromelation.

Table 6
Testing for paretoness in rank-size rule regression (null: $\alpha_2 = 0$ in $\log(S_{(i)}) = \alpha_0 + \alpha_1 \log(i) + \alpha_2 \log^2(i)$).

	α_2	t_2	S	n	Year
Japan	−0.005	−0.354	−0.082	113	2000
UK	0.030	5.257**	0.584	189	2005
USA	−0.098	−46.76**	−2.657**	363	2007

** significant at 5%.

is likely to hold well for the three countries in view of Fig. 6. In order to test the Paretoness, we estimate the coefficients in regression (4) and focus on the results of α_2. Table 6 reports the estimates of α_2, t_2, and S, which were proposed in the previous section. The null hypothesis is $\alpha_2 = 0$ in $\log(S_{(i)}) = \alpha_0 + \alpha_1 \log(i) + \alpha_2 \log^2(i)$. As mentioned before, we cannot apply the standard t-test in rank-size regression because the t-statistic explodes as n increases. If the standard t-test is applied mechanically, Paretoness is rejected for the UK and the US at the 5% significance level, with a huge t_2 value for the US. On the other hand, in the results of new test S, it is not rejected for Japan and the UK. It means the Pareto specification is acceptable in these two countries. In the US, the null hypothesis is rejected. This finding is consistent with our results that a standard t-test (t_2) tends to reject the null too often, or the true size is greater than the nominal size.

5. Concluding remarks

We investigate the statistical properties of estimators of rank-size rule regression when we include both log-rank (or $\log(i)$) and squared log-rank ($\log^2(i)$) as explanatory variables. It is, we believe, interesting that the order of convergence