JSPS Grants-in-Aid for Creative Scientific Research

Understanding Inflation Dynamics of the Japanese Economy

Working Paper Series No.22

ランクサイズ回帰の検定について

小西葉子
西山慶彦

May 19, 2008

Research Center for Price Dynamics
Institute of Economic Research, Hitotsubashi University
Naka 2-1, Kunitachi-city, Tokyo 186-8603, JAPAN
Tel/Fax: +81-42-580-9138
E-mail: sousei-sec@ier.hit-u.ac.jp
http://www.ier.hit-u.ac.jp/~ifd/

ランクサイズ回帰の検定について*

小西葉子†　　西山慶彦‡

2008 年 5 月

概 要

多くの実証研究では都市サイズ，企業の資産や売上高の規模などの研究対象がパレート性を持つことを，ランクサイズ回帰で観察してきた．具体的には，順位の対数値をその規模の対数値に回帰することにより，その係数が-1 になるかを調べる．また，パレート性の有無には，二次項の係数が 0 であることも条件になるので，本稿では，二次項を加えたものを回帰モデルとする．パレート性の検証には，一次項，二次項それぞれの t 検定と，一次項の係数が-1，二次項の係数が 0 という複合仮説が成立しているかを F 検定で調べる方法がある．しかし，分析対象がパレート分布に従う時，データ数が大きくなると，t 値は発散してしまうため通常の t 検定を行えないことがわかっており，F 検定でも同様の問題が観察された．そこで本稿では，F 値の棄却域をシミュレーションによって構成し，ランクサイズ回帰の複合仮説を検証可能とし，パレート性の検定の新たな手法として提案した．

JEL Classification: C12, C16, R12,

*一橋大学経済研究所，定例研究会報告 (2008 年 2 月 27 日) に対する討論者であった野口晴子氏のコメントに感謝します．また，同定例研究会の参加者からも多くの有意義なコメントを頂きました．記して感謝致します．

†独立行政法人 経済産業研究所

‡京都大学経済研究所

1　はじめに

都市・地域経済学でランクサイズ回帰のあてはまりがよく，古くから応用されている分野に都市人口分布の分析がある．まず一国の都市の人口を大きい順に並べ替え，1 位，2 位 . . . と順位（ランク）をつける．ランクサイズ回帰とは，都市の人口規模の対数値を当該都市の順位 (ランク) の対数値に回帰したものである．すると，多くの国において定数項がほぼサンプルサイズの対数に等しく，傾きはほぼ −1 に等しくなるという結果が得られる．つまり人口規模が 1 番大きい都市から順に 2 番目の都市は 1/2 の人口，3 番目は 1/3, . . . と減少していく．

S_i,$i = 1, \ldots, n$ をある国の都市 i の人口とし，$S_{(i)}$ をそれを大きい順に並べ替えた順序統計量とする．つまり $S_{(1)} \geq S_{(2)} \geq \ldots \geq S_{(n)}$ である．

$$\log S_{(i)} \approx \alpha_0 + \alpha_1 \log i, i = 1, \ldots, n \tag{1}$$

$\alpha_0 > 0$, $\alpha_1 < 0$ のとき，(1) 式はランクサイズ回帰モデルと呼ばれ，$\alpha_1 = -1$ のとき Zipf's law(ジップの法則) が成り立っているといえる．またパレート分布のパラメータが 1 の場合も同様の関係が観察される．

このような現象は様々な分野やトピックで観察されており，経済学では所得や資産が上位層に集中することや，有能な少数の従業員が全体の生産に大きく貢献することなどが事例として挙げられている．これは，パレートの 80 : 20 の法則とも呼ばれ，集中度や不平等度の指標の一つとして用いられてきたが，近年では，マーケティングや経営の分野でも広く用いられている．この種の方法は，自然科学や実験など大規模データが利用可能な分野で応用されてきたが，近年マイクロデータの利用可能性が高まることによって，都市経済学以外の分野でも分析対象がパレート性をもつか，あるいはある種のべき乗分布に従うか否かを検証する研究が行われはじめた．齊藤・渡辺 (2007) では，我が国の法人企業の $\frac{1}{3}$ をカバーする約 82 万社のデータを用いて企業間関係の分析を行っている．そこでは，企業間のネットワーク構造に着目して，当該企業と「仕入先」,「販売先」,「大株主」に関する企業間ネットワークに，パレート性やべ

き乗の関係があることが発見されている．

このようにランクサイズ回帰による分析が盛んな理由の一つは，最小二乗法中心の簡便な方法で分析ができることにある．

==図 1 挿入==

図 1 は各国の都市人口について縦軸に人口の対数値，横軸にその大きさに対応するランクの対数値をとってプロットしたグラフである．グラフは概ね右下がりの直線になっている．実際，図 1 のデータで (1) 式についての回帰を行った場合，推定値が -1 に非常に近い．都市経済学におけるランクサイズ回帰では，長い間通常の t 検定によって $\alpha_1 = -1$ という帰無仮説が調べられてきたが，通常 t 値が非常に大きく，α_1 の推定値は -1 にかなり近いにも関わらず帰無仮説が棄却されるという実証研究が非常に多かった．この点に関して，数多くはないものの，先行研究が蓄積されてきている．詳しいことは後述するが，近年の研究から，都市のサイズがパレート分布に従っている場合にはランクサイズ回帰の，t 統計量が漸近的に発散することがわかり，当然このような場合は，通常の t 検定の棄却域は使用できない．

他方，Rosen and Resnick(1980) 等では，被説明変数がパレート分布に従っているかの簡便的な検定として，(1) 式の一次の項の二乗項を説明変数として加え，その推定値が 0 であれば，パレート分布に従っているとする分析方法も提案されている．ここでも t 検定が用いられているが，この場合も上と同じ問題を含む．Nishiyama, Osada and Sato(2007) では，パレート性の検定のために，一次の項と二次の項をそれぞれの帰無仮説の下で t 検定できるようにシミュレーションによって棄却域を求めた．しかし，この定式化で，パレート性を調べるのならば，一次項の係数が -1，二次項の係数が 0 という複合仮説が成立しているかを F 検定により調べるのが自然である．ただし，t 検定の際に問題になったように，真の分布がパレート分布であっても，サンプルサイズが大きくなると共に F 値が発散し，F 検定を行うと帰無仮説を棄却しやすくなることが懸念される．

そこで，本稿では，同様に二乗の項を説明変数に考慮して推定を行い，F 値の棄却域をシミュレーションによって構成し，ランクサイズ回帰のパレート性の検定手法として新たに提案

する．

さらに企業の資産規模のデータを用いて，二次項も含めてランクサイズ回帰の実証分析を行う．また，本稿で得られた t 値，F 値の臨界値を用いてパレート性の検定も行う．その際先行研究で指摘されていない，二次項を含んだ場合の説明変数間の相関の高さに着目した．一次項と二次項の相関係数について，漸近的にどのような挙動をとるのかを調べ新たな事実がわかった．

次節では，先行研究のレビューを行う．3 節ではランクサイズ回帰のシミュレーションを行い，t 統計量と F 統計量の棄却域を構成する．またそれに基づく実証分析を行う．4 節では，結論と今後の課題，付録では実証結果で得られた知見より，ランクサイズ回帰の説明変数間の相関係数の挙動について調べている．

2　先行研究

経済学でランクサイズ回帰のあてはまりがよく，古くから応用されている分野に都市経済学がある．これは，都市の人口規模の対数値をその大きさのランクの対数値に回帰すると，定数項がほぼサンプルサイズの対数に等しく，傾きはほぼ –1 に等しくなるというものである．この関係は，都市規模が i.i.d. でパラメータの値が 1 のパレート分布に従っているときに成立することが知られている．この文脈で様々な国に関して先駆的かつ包括的な分析を行ったのは Rosen and Resnick(1980) であり，現在もなおこの分野で必ず引用される文献である．Soo (2005) はそれを更新したデータについて調べている．これらの論文では，最小二乗推定 (OLS) によりランクサイズ回帰を行って点推定値を得ている．それと同時に，1. ランクの対数の係数に関する t 検定に基づいて，傾きが –1 であるという仮説の検定，および 2. 回帰モデルにランクの対数とその 2 乗項を説明変数に含めて後者の係数がゼロかどうかを t 検定で調べることによるパレート性の検定が行われている．しかし，被説明変数は順序統計量であり，その結果，定義上被説明変数は，分散不均一と自己相関をもつ．そのため，古典的な回帰理論を適

用することはできない．それらを考慮して推定量の性質を調べたのが Gabeix and Ioannides (2003)，Gabaix and Ibragimov (2005) などであり，一致性，漸近正規性が証明されている．また，Nishiyama and Osada (2005)，Nishiyama, Osada and Sato(2007) は，OLS よりも有効性のある推定方法として trimmed OLS と GLS 法による推定方法を提案している．

それらの研究において，推定に関しては実は統計的にはあまり大きな問題は生じないことが示されているが，検定に関しては，Nishiyama and Osada (2005)，Nishiyama, Osada and Sato(2007) らが分散不均一と自己相関のために，傾き -1 であるという帰無仮説を標準的な t 検定で調べることはできないことを示している．加えて，帰無仮説の下でも $s^2 = \sum$(回帰残差)$^2/(n -$ 説明変数の数) が漸近的にゼロに収束するという問題が指摘されている．そのため，真の分布がパレート分布であっても，サンプルサイズが大きくなると共に t 値が発散し，t 検定を行うと帰無仮説を棄却しやすくなる．この問題を回避するため，それらの論文では修正した t 検定が提案されている．そこでは，通常の t 値が検定統計量として用いられているが，棄却域はシミュレーションによって構成されている．

Nishiyama, Osada and Sato(2007) では，パレート性の検定のために，1. で述べたように説明変数に二乗の項も含み，一次の項と二次の項をそれぞれの帰無仮説の下で t 検定できるようにシミュレーションによって棄却域を求めた．しかし，この定式化で，パレート性を調べるのならば，一次項の係数が -1，二次項の係数が 0 という複合仮説が成立しているかを F 検定により調べるのが自然である．ただし，t 検定の際に問題になったように，ここでも帰無仮説の下でも $s^2 = \sum$(回帰残差)$^2/(n -$ 説明変数の数) が漸近的にゼロに収束し，真の分布がパレート分布であっても，サンプルサイズが大きくなると共に F 値が発散し，F 検定を行うと帰無仮説を棄却しやすくなることが懸念される．

そこで，本稿では，同様に二乗の項を説明変数に考慮して推定を行い，F 値の棄却域をシミュレーションによって構成し，パレート性の検定の手法として提案する．

これらの提案が，非常に初歩的なことからもわかるようにランクサイズ回帰は，汎用的で長い間多くの分野で利用されてきているが，その統計的性質が明らかになっていないのが現状

である．

3　モンテカルロシミュレーションと実証分析

前節で述べたように，Gabeix and Ioannides (2003)，Gabaix and Ibragimov (2005) で，(2) 式のタイプのランクサイズ回帰の推定値の一致性と漸近正規性が証明された．Nishiyama, Osada and Sato(2007) は，パレート性の検証のために，Rosen and Resnick(1982) や Soo (2005) タイプの (4) 式についてシミュレーションを行い，帰無仮説を一次の項が -1，二次の項が 0 としてそれぞれの t 値の臨界値を求めている．さらに Soo (2005) がアップデートしたデータを用いてパラメータ推定し検定を行っているが，その推定値の挙動や問題点については指摘していない．ここで (4) 式は通常定義されるランクサイズ回帰の逆回帰となっていることに注意されたい．ただし，逆回帰であっても，一次の項に関して (2) 式が有する統計的性質はおそらく変わらないと予想されるが，理論的な検証は今後の課題である．

本節ではまず，パレート性を調べるために，(3)，(4) 式についてモンテカルロシミュレーションを行い，各推定値，t 統計量，回帰残差の二乗和を計算し，シミュレーションベースの t 統計量の臨界値を得る．次に，F 検定によって一次項の係数が -1，二次項の係数が 0 という複合仮説が成立しているかを調べる．その際，F 値の棄却域をシミュレーションによって構成し，パレート性の検定の手法とする．また，実証分析では日経 NEEDS のデータを用い，上場企業の資産についてランクサイズルールが成立しているかを調べる．

3.1　モンテカルロシミュレーション

各シミュレーションでは，パラメータが 1 のパレート分布から $n = 50, 100, 200, 500, 1000, 3000$ のデータを発生させ，(3) 式，(4) 式について OLS を行った．繰り返し計算は 10000 回行っている．

$$\log S_{(i)} = c + \alpha_1 \log i, i = 1, \ldots, n \tag{2}$$

$$\log S_{(i)} = c + \alpha_1 \log i + \alpha_2 \log^2 i, i = 1, \ldots, n \tag{3}$$

$$\log i = c + \beta_1 \log S_{(i)} + \beta_2 \log^2 S_{(i)}, i = 1, \ldots, n \tag{4}$$

表 1 はサンプルサイズごとのシミュレーション結果で，記述統計量である．$\alpha_1, \alpha_2, \beta_1, \beta_2$ は推定値，$\alpha_1 t$ は帰無仮説が $\alpha_1 = -1$，$\beta_1 t$ は帰無仮説が $\beta_1 = -1$ の両側検定，$\alpha_2 t$ は帰無仮説が $\alpha_2 = 0$，$\beta_2 t$ は帰無仮説が $\beta_2 = 0$ の両側検定を行ったときの t 統計量，SSR は回帰の残差二乗和である．

α_1 については，サンプルサイズが 50，500，3000 と大きくなる程，平均値が -1 に近くなり，α_2 は同様にサンプルサイズが大きくなると平均値が 0 に近くなり，両係数とも一致性が成り立っているようにみえる．

図 2 は (3) 式について $n = 50, 500, 3000$ のとき回帰を行い，その推定値の密度関数を描いたものである．左の列は一次項 α_1，右の列は二次項 α_2 の密度関数である．サンプル数が大きくなるほど，一次項は -1，二次項は 0 の周りに推定値が集まっているのがわかる．(4) 式の一次項 β_1, 二次項β_2 も (3) 式の結果と同じ挙動になっている．一方，表 1 より (3) 式，(4) 式通じて，サンプル数が大きくなる程 t 値の平均値は絶対値で大きくなり，範囲（レンジ）は広がっている．これは先行研究で指摘されているように，サンプルサイズが大きくなると t 値が発散傾向にあることと矛盾がない．図 3 は，左の列が一次項 α_1 の帰無仮説 $\alpha_1 = -1$ の下での t 統計量の密度関数，右の列が二次項 α_2 の帰無仮説 $\alpha_2 = 0$ の下での t 統計量の密度関数である．サンプルサイズが大きくなるほど，分散が大きくなっている．Nishiyama, Osada and Sato(2007) で指摘されているように，t 統計量の分母を構成する s^2 が，n が大きくなる程 0 に近づいていることが関係していると考えられる．

=表 1 挿入=

=図 2 挿入=

=図 3 挿入=

このようにパレート分布から発生させたデータを用いても，推定値に関して t 検定を行うと，ランクサイズルールが成立していても $\alpha = -1$ や $\alpha = 0$ を棄却しやすくなってしまう．そのため，シミュレーションによって得られた棄却域を用いることが解決策の一つになる．

表 2 は，(3) 式，(4) 式のランクサイズ回帰において，一次項，二次項の係数の有意性を t 検定で検証するためのシミュレーションで得られた棄却域である．α_1 は -1，α_2 は 0 が帰無仮説であり，各サンプルサイズに対する両側 $1\%, 5\%, 10\%$ 水準の臨界値である．

サンプル数が多くなるほど，臨界値の絶対値は大きくなり，棄却域が狭くなっている．t 検定は，t 値が絶対値で 2 より大きければ，当該係数に関して帰無仮説が棄却される．シミュレーションで得られた臨界値は，2 よりかなり大きく，サンプル数の増加に伴い大きくなっており，通常の t 統計量の臨界値を用いるより帰無仮説を棄却しにくくなっている．以上より，順回帰 ((3) 式) であっても逆回帰 ((4) 式) であっても，t 検定を行う場合は通常の棄却域は用いることが適切でないことわかった．

=表 2 挿入=

前述したが，(3) 式と (4) 式で，パラメータ 1 のパレート性の成立を調べるのならば，一次項の係数が -1，二次項の係数が 0 という複合仮説が成立しているかを調べるのが自然である．

ただし，t 検定の際に問題になったように，$s^2 = \sum(\text{回帰残差})^2/(n - \text{説明変数の数})$ が漸近的にゼロに収束し，真の分布がパレート分布であっても，サンプルサイズが大きくなると共に F 値が発散し，F 検定を行うと帰無仮説を棄却しやすくなることが懸念される．

図 4 は (3) 式について，シミュレーションで発生させたデータを用いて帰無仮説 $\alpha_1 = -1$, $\alpha_2 = 0$，(4) 式について帰無仮説 $\beta_1 = -1$,$\beta_2 = 0$ の下で計算した F 統計量の密度関数である．n が大きくなるほど，F 値が大きくなっていることがわかる．表 3 は F 統計量の棄却域である．

通常の F 統計量は，制約の数が等しい場合，サンプルサイズが大きくなるほど，臨界値が

小さくなる．(3) 式と (4) 式で,，パラメータ 1 のパレート性の成立を調べる場合，通常の F 統計量の臨界値は $n = 50$ 以上だとおおよそ 3 である．しかし，表 3 の結果では，t 統計量と同じように，サンプル数が大きくなる程，臨界値が大きくなっており，通常の F 検定を行うと，帰無仮説 $\alpha_1 = -1$,$\alpha_2 = 0$，帰無仮説 $\beta_1 = -1$,$\beta_2 = 0$ を棄却しやすくなってしまう．よって，順回帰 ((3) 式) であっても逆回帰 ((4) 式) であっても，F 検定を行う場合は通常の棄却域は用いることが適切でないことわかった．以上より，本稿では，シミュレーションで得た F 値の棄却域をパレート性の検定の手法として提案する．

=表 3 挿入=

=図 4 挿入=

3.2 実証研究

ここでは，日経 NEEDS の 2006 年の上場企業の資産のデータを用いてパラメータ 1 のパレート性の検定を行う．サンプル数は 1736 社である．図 5 は Y 軸が企業の資産総額の対数値で，X 軸は当該企業の順位の対数値である．概ね線形で傾きも -1 に近く何らかのべき乗関数に従っているように見える．しかし，順位の低いところにデータが集中しそこが非線形になっている．そのため，パラメータ 1 のパレート分布に従っているかは，二次項まで含めたランクサイズ回帰を行い，その係数の推定値が 0 になるかを調べることが望ましい．表 4 の左側は (2)〜(4) 式の推定結果である．表の中央は通常の t 検定，F 検定を行った場合の検定結果を示している．通常の，t 検定，F 検定の臨界値を用いると各式において帰無仮説を棄却してしまう．一方右側は，本稿で得られたシミュレーションベースの棄却域による検定結果である．これを用いた場合，(2) 式，(3) 式では，パレート性を棄却しなかった．

=表 4 挿入=

=図 5 挿入=

ここで，(2) 式と (3) 式の推定値に着目したい．パラメータ 1 のパレート分布に従う場合，対数をとって回帰モデルの形にすると，一次項のパラメータが -1 の (2) 式の形になる．(2) 式の推定値は -1.05 で t 検定でも棄却されなかった．そこでさらに，二次項を加えその係数が 0 になるか否かを確認したのが (3) 式である．検定結果では，F 検定でパレート性は棄却されなかった．t 検定においても，一次項，二次項それぞれの帰無仮説も棄却しなかったが，α_1 は 0.4 で，符号が逆となり，値が大きく異なった．このような症状の代表的な要因として多重共線性が考えられる．特にランクサイズ回帰では，$\log$ (ランク) と $\log^2$ (ランク) を説明変数に含むため多重共線性が起きやすい．実際サンプル 1000 では，説明変数間の相関は 0.989 と非常に 1 に近い．

本稿の実証分析でも，二次の項を加えたことで推定値が大きく変わった．また指摘はされていないが，Nishiyama, Osada and Sato(2007) らの都市規模に関するランクサイズ回帰でも，多重共線性がおきているようにみえるものとそうでないものがある．

このことについて，本稿では，$\log$ (ランク) と $\log^2$ (ランク) の相関係数の挙動を調べた．詳しい導出は付録を参照されたい．ランクサイズ回帰においては，n が大きいところでは説明変数間の相関係数が 1 になることがわかった．このような状況下では，通常推定自体が困難となることが多い．しかし，前小節のシミュレーション結果においては，二次の項を含んだ回帰でも，各推定値は漸近的に一致性があるように見受けられる．このような特殊な状況下で推定に際して何が起こっているのかに関する理論的な検討は今後の課題とする．

4 おわりに

本稿では，古くから様々な分野で応用されてきたランクサイズ回帰のパレート性の検定手法を提案した．多くの実証研究では都市サイズ，企業の資産や売上高の規模などの研究対象がパレート性を持つことを，ランクサイズ回帰を行うことで観察してきた．具体的には，規模の対数値にその順位の対数値を回帰することによって，その係数が -1 になることを調べる．ま

た，パレート性の有無には，二次項の係数が 0 であることも条件になるため，本稿では，二次項を加えて推定を行った．パレート性の検証には，一次項，二次項それぞれの t 検定と，一次項の係数が -1，二次項の係数が 0 という複合仮説が成立しているかを F 検定で調べる方法がある．しかし，分析対象がパレート分布に従う時，サンプルサイズが大きくなると，t 値は発散してしまうため通常の t 検定を行うことはできないことがわかっており，F 検定でも同様の問題が観察された．

そのため本稿では，F 値の棄却域をシミュレーションによって構成し，一次項の係数が -1，二次項の係数が 0 という複合仮説を検証可能とし，パレート性の検定の新たな手法として提案した．これが本稿の主な貢献である．

実証研究では，本稿で得られた t 値，F 値の臨界値を用いて二次項も含めてランクサイズ回帰を行ったところ，検定によりパレート性があることが検証されたが，二次項を含んだ場合に多重共線性の症状がみられた．このことより，一次項と二次項の相関係数について調べ，漸近的に相関係数が 1 となることが明らかになり，このことが，実証結果で推定値の大きさに影響を与えている可能性があることがわかった．

今後の課題は，二次項と F 統計量に関する統計的性質を明らかにすること，一次項と二次項の相関係数が漸近的に 1 になり完全な多重共線性を起こすことが推定値や検定統計量にどのような影響を与えるのかを調べることである．

付録　説明変数間の相関について

以下では n が大きくなったときの相関係数挙動を調べる．証明には以下のレンマを用いる．各レンマの証明は，Nishiyama and Osada (2005)，Nishiyama, Osada and Sato(2007) で示されている．

- (a) $\sum \log i = n \log n - n + \frac{1}{2} \log n + O(1)$
- (b) $\sum \log^2 i = n \log^2 n - 2n \log n + 2n + \frac{1}{2} \log^2 n + O(\log n)$

- (c) $\sum \log i(\frac{1}{n} + \ldots + \frac{1}{i}) = n\log n - 2n + \frac{1}{4}\log^2 n + O(\log n)$
- (d) $\sum \frac{\log i}{i} = \frac{\log^2 n}{2} + o(\log^2 n)$
- (e) $\sum \frac{\log^2 i}{i} = \frac{\log^3 n}{3} + o(\log^3 n)$
- (f) $\sum \frac{\log^2 i}{i^2} = O(1)$

レンマ 1

$n \to \infty$ のとき $\log i$ と $\log^2 i$ の相関係数は 1 になる.

証明

$\log i = [\log 1, \log 2, \ldots, \log n]'$ と $\log^2 i = [\log^2 1, \log^2 2, \ldots, \log^2 n]'$ の相関係数は以下で定義される.

$$\frac{\frac{1}{n}\sum_{i=1}^{n}\log^3 i - \left(\frac{1}{n}\sum_{i=1}^{n}\log i\right)\left(\frac{1}{n}\sum_{i=1}^{n}\log^2 i\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\log^2 i - \left(\frac{1}{n}\sum_{i=1}^{n}\log i\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\log^4 i - \left(\frac{1}{n}\sum_{i=1}^{n}\log^2 i\right)^2}}$$

まず分子の第 1 項から計算する. レンマ (a),(b),(e) 使用

$$\begin{aligned}
&\sum_{i=1}^{n}\log^3 i = n\log^3 n - \sum i\{\log^3(i+1) - \log^3 i\} \\
&= n\log^3 n - \sum i\left[\left(\log i + \log\left(1+\frac{1}{i}\right)\right)^3 - \log^3 i\right] \\
&= n\log^3 n - \sum i\left[3\log^2 i\left(\log\left(1+\frac{1}{i}\right)\right) + 3\log i\left(\log^2\left(1+\frac{1}{i}\right)\right) + \log^3\left(1+\frac{1}{i}\right)\right] \\
&= n\log^3 n - \left[3\sum\left(\log^2 i - \frac{\log^2 i}{2i} + \frac{\log^3 i}{3i^2}\right) + 3\sum\left(\frac{\log i}{i} - \frac{\log i}{i^2}\right) + \sum\left(\frac{1}{i^2} + \frac{3}{2i^3}\right)\right] \\
&= n\log^3 n - \left[3\left(n\log^2 n - 2n\log n + 2n\right) - \frac{3}{2}\left(\frac{\log^3 n}{3}\right) + 3\left(\frac{\log^2}{2}\right)\right] \\
&= n\log^3 n - 3n\log^2 n + 6n\log n - 6n
\end{aligned}$$

よって, 分子第 1 項は以下のようになる.

$$\frac{1}{n}\sum_{i=1}^{n}\log^3 i = \log^3 n - 3\log^2 n + 6\log n - 6$$

分子の第 2 項は，レンマ (a),(b) より

$$\left(\frac{1}{n}\sum\log i\right)\left(\frac{1}{n}\sum\log^2 i\right) \approx \frac{1}{n}\left(n\log n - n + \frac{1}{2}\log n\right)\frac{1}{n}\left(n\log^2 n - 2n\log n + 2n\right)$$
$$\approx (\log n - 1)\left(\log^2 n - 2\log n + 2\right) = \log^3 n - 3\log^2 n + 4\log n - 2$$

となる．よって分子は，$2\log n - 4$ である．レンマにより，分子は以下の近似が成立する．

$$\frac{1}{n}\sum_{i=1}^{n}\log^3 i - \left(\frac{1}{n}\sum_{i=1}^{n}\log i\right)\left(\frac{1}{n}\sum_{i=1}^{n}\log^2 i\right)$$

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

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

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

次に，分母 $\frac{1}{n}\sum_{i=1}^{n}\log^2 i - \left(\frac{1}{n}\sum_{i=1}^{n}\log i\right)^2$ と $\frac{1}{n}\sum_{i=1}^{n}\log^4 i - \left(\frac{1}{n}\sum_{i=1}^{n}\log^2 i\right)^2$ について考える．
レンマ (b) より，

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

$$= \log^2 n - 2\log n + 2 + \tfrac{\log^2 n}{2n} - \tfrac{\log n}{n} + O(n^{-1}) - \left[\log n - 1 + \tfrac{\log n}{2n} + O(n^{-1})^2\right]$$

$$= 1 - \frac{\log^2 n}{2n} + O(n^{-1})$$

$\frac{1}{n}\sum_{i=1}^{n}\log^3 i$, の近似で用いたのと同様の変形により，$\sum_{i=1}^{n}\log^4 i = n\log^4 n - 4n\log^3 n - 3n\log^2 n + 6n\log n - 6n + O(\log n)$ であることがわかる．従って，レンマ (b) を用いると

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

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

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

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

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.

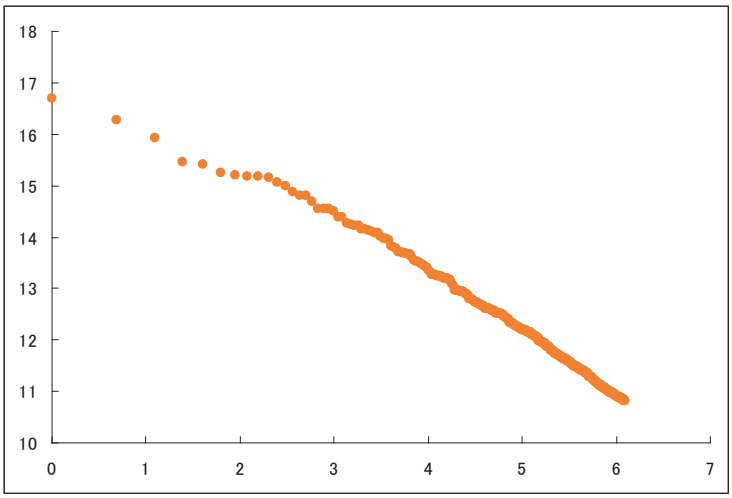
4. Y. Nishiyama and S. Osada (2004), Statistical theory of rank size rule regression under Pareto distribution, *CAEA Discussion Paper* **009**, Kyoto University.

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*.

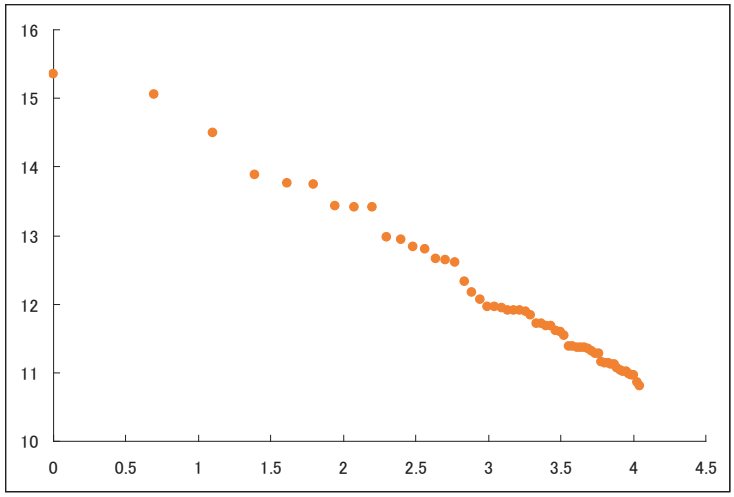
6. K.T. Rosen and M. Resnick (1980), The size distribution of cities: An explanation of the Pareto law and primacy, *Journal of Urban Economics* **8**, pp. 165-186.

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

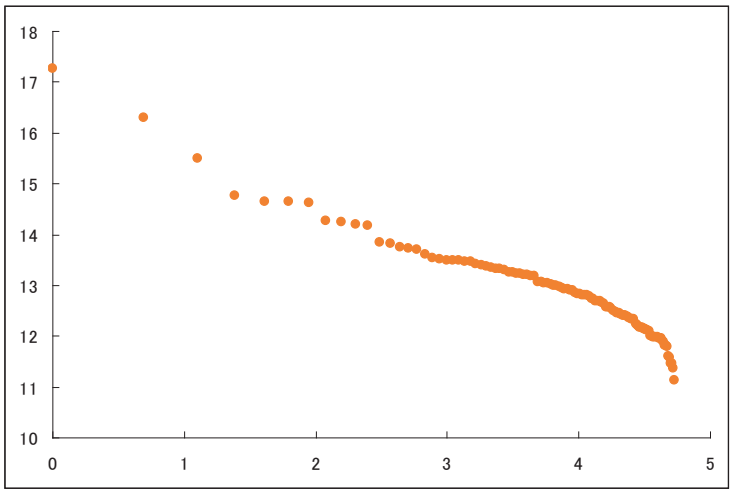
アメリカ合衆国（2005）



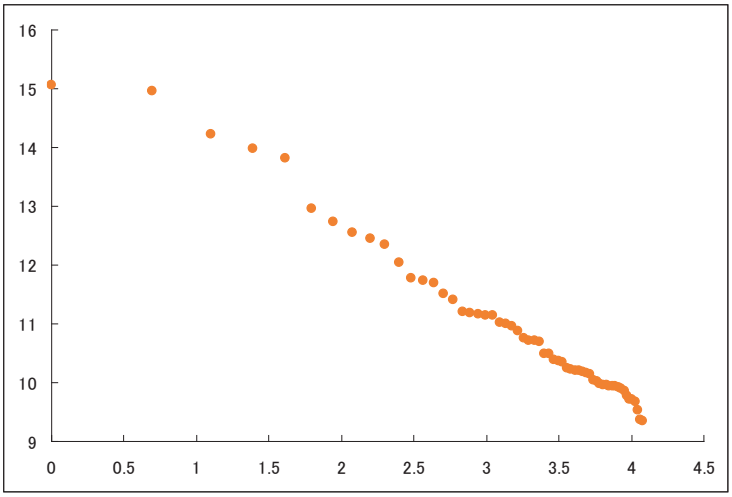
カナダ（2001）



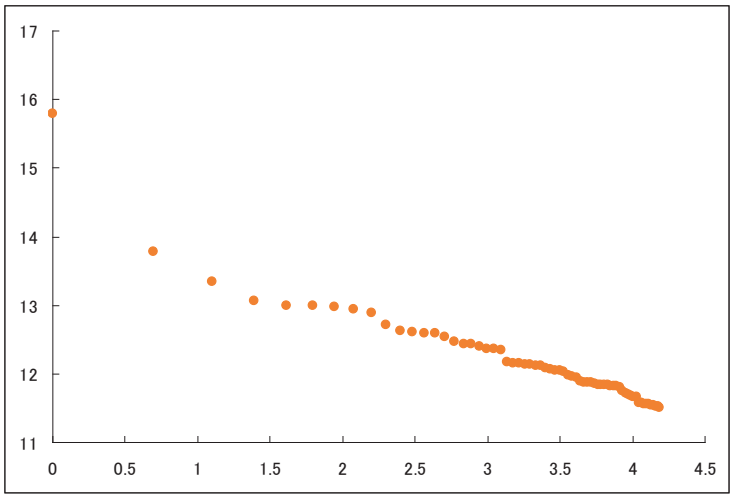
日本（2000）



オーストラリア（2001）



イギリス（2001）



フランス（2004）

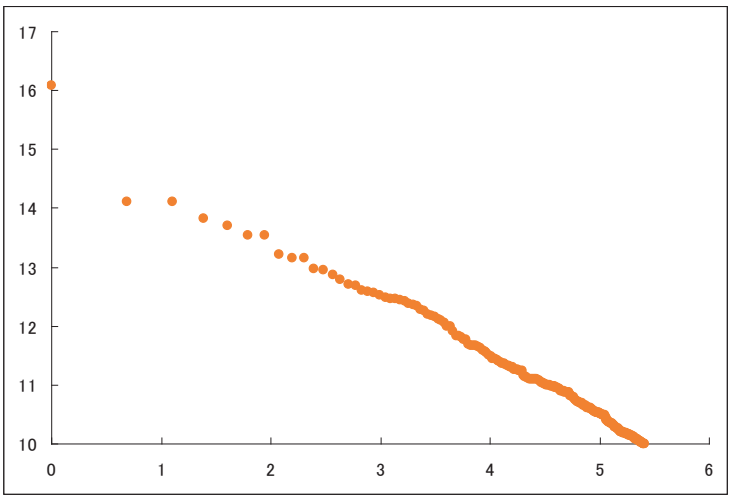
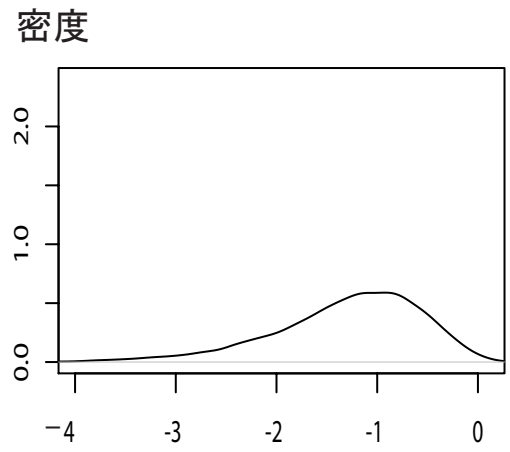


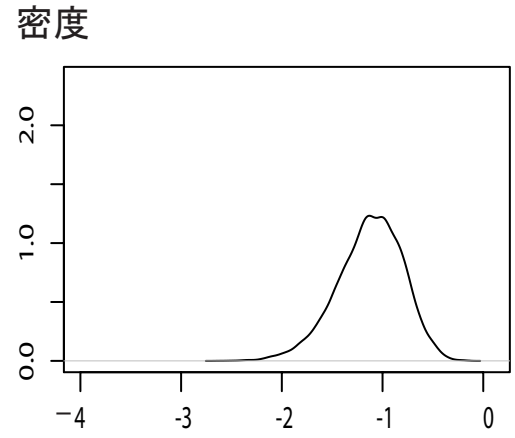
図１：各国都市の人口規模と順位の散布図
（横軸：log(rank)，縦軸：log(Size)）

http://www.citypopulation.de/より作成

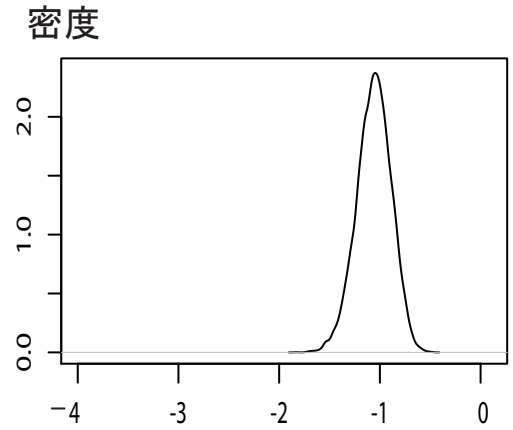
(1)一次項の密度関数



（ア）一次項α1, n=50

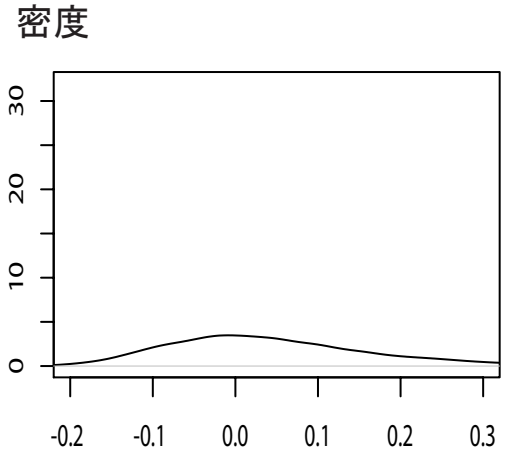


（イ）一次項α1, n=500

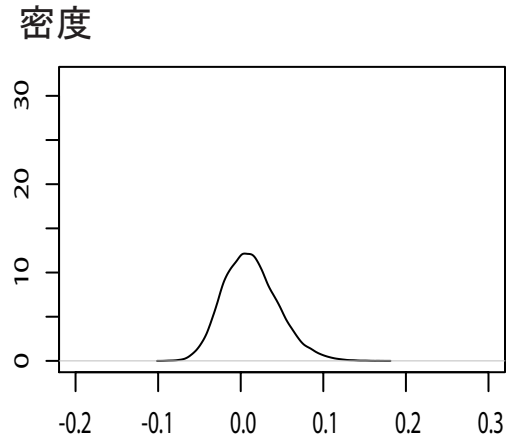


（ウ）一次項α1, n=3000

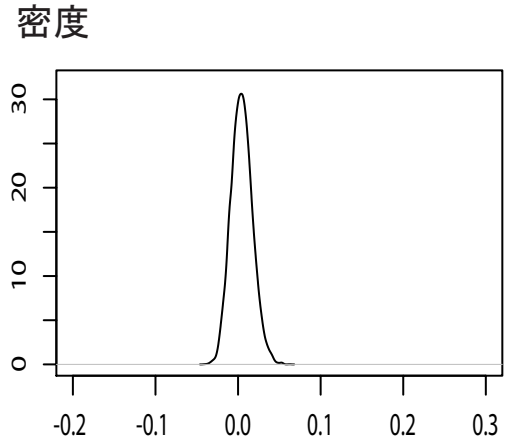
(2)二次項の密度関数



（エ）二次項α2, n=50



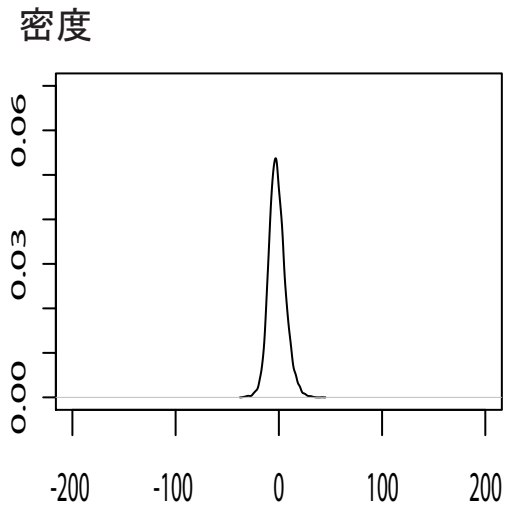
（オ）二次項α2, n=500



（カ）二次項α2, n=3000

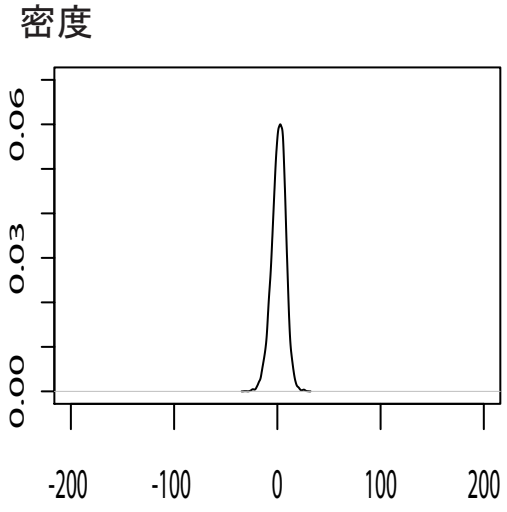
図２：(3)式　log(Size)=c+α1*log(rank)+α2*log^2(rank)の
推定値のシミュレーション結果，繰り返し回数：10000回

(1) α1のt値の密度関数:
帰無仮説：α1=-1

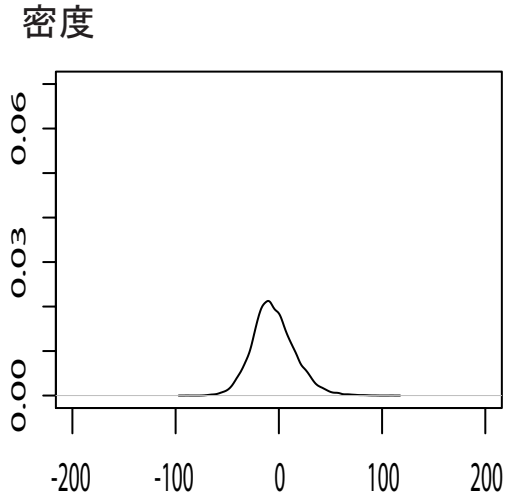


（ア） α1のt値, n=50

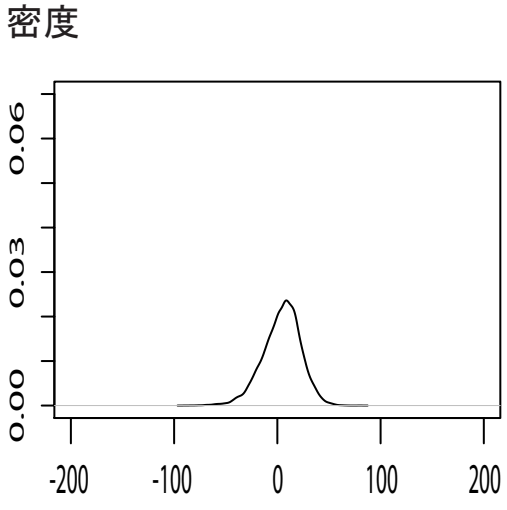
(2) α2のt値の密度関数:
帰無仮説：α2=0



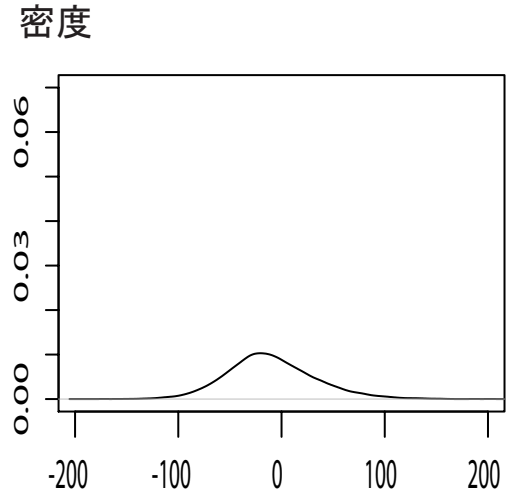
（エ） α2のt値, n=50



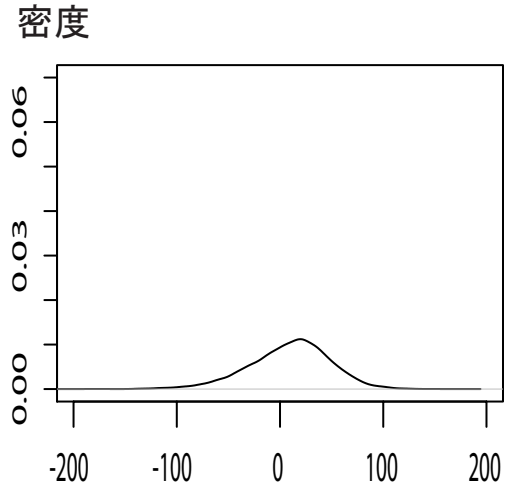
（イ） α1のt値, n=500



（オ） α2のt値, n=500



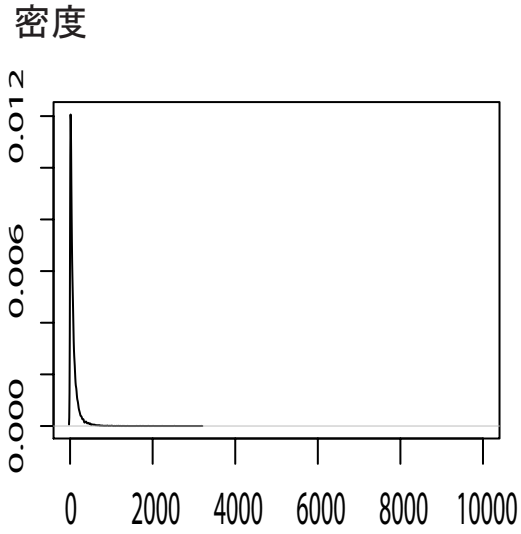
（ウ） α1のt値, n=3000



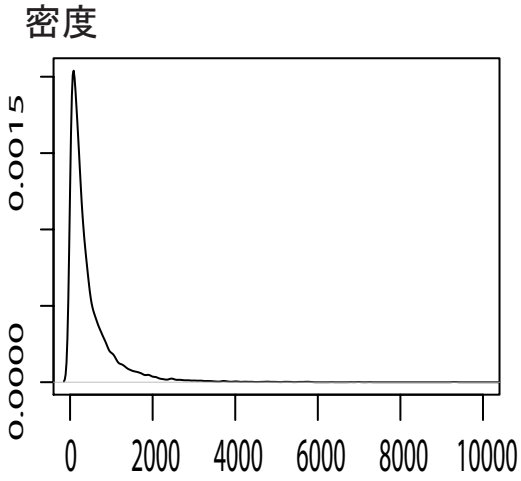
（カ） α2のt値, n=3000

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

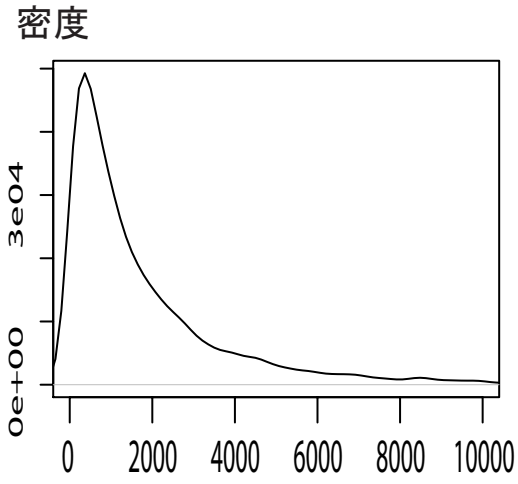
(1) 式(3)のF値の密度関数
帰無仮説： $\alpha 1=-1$, $\alpha 2=0$



（ア） 式(3), n=50

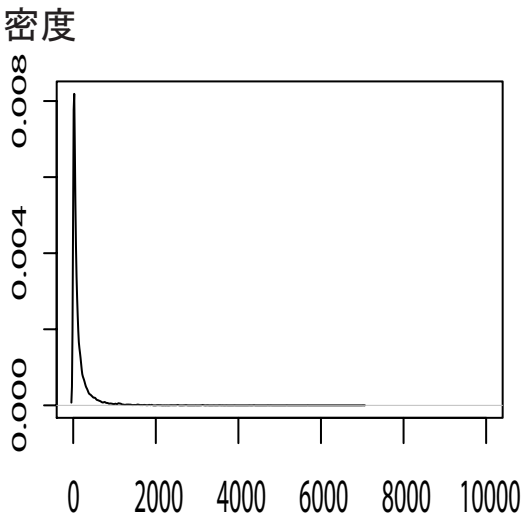


（イ） 式(3), n=500

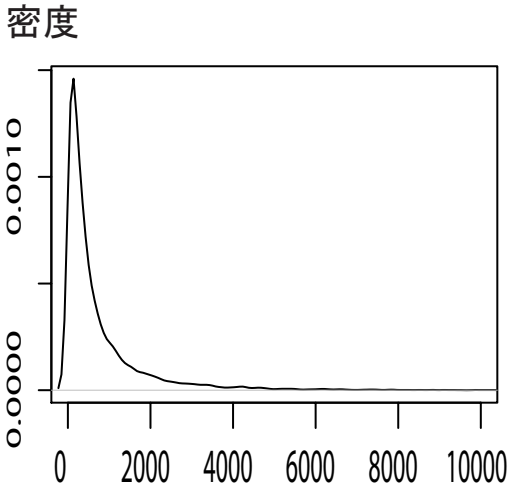


（ウ） 式(3), n=3000

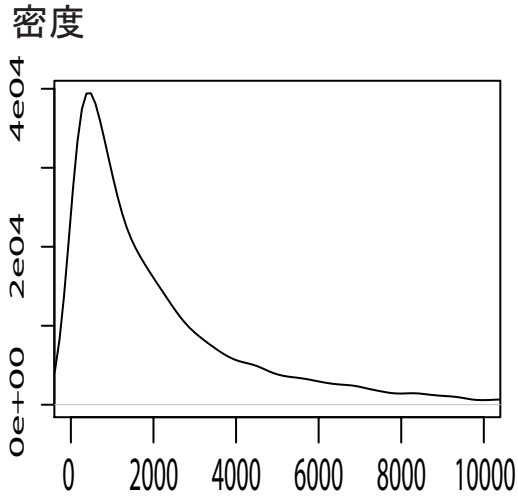
(2) 式(4)のF値の密度関数
帰無仮説： $\beta 1=-1$, $\beta 2=0$



（エ） 式(4), n=50



（オ） 式(4), n=500



（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

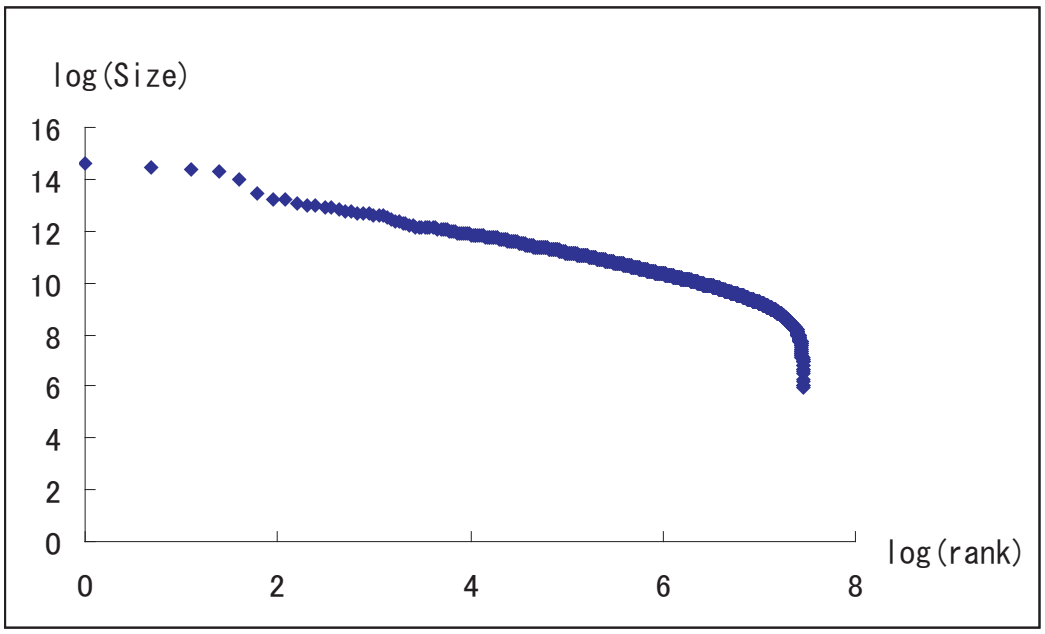


図５：上場企業の資産規模と順位の散布図

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.