

Rによる探索的財務データ解析と再現可能研究

—NEEDS 企業財務データの利用—

地 道 正 行

要 旨

本稿では、東京証券取引所第一部上場企業の財務データに対して、非対称分布族を考慮した誤差分布をもつ両対数モデルを用いて売上高の統計モデリングを行う。その際、探索的データ解析の視点から、データ可視化によって得られた知見を統計モデリングに利用し、さらに赤池情報量規準を利用することによってモデル選択を行う。なお、本研究は動的文書生成によって再現可能研究の立場から実施される。

キーワード：財務データ (Financial Data), 両対数モデル (Double-Log Model), 非対称分布族 (Family of Skew Distributions), 探索的データ解析 (Exploratory Data Analysis), 再現可能研究 (Reproducible Research)

I はじめに

「ビッグデータ」という用語は、2010年前後から日本においてもマスメディアなどを通じて聞かれるようになったが、現在では「ブーム」のような状態は去り、以前ほど騒がれなくなったように思われる¹⁾。このことは、Google Trends²⁾ による 'bigdata' の世界の検索動向 (図1 参照) を参照することによってもわかる。このような現状のもとでも、社会全体で収集されるデータは情報通信技術やセンサー技術の向上によって巨大化の一途をたどってい

1) 逆に、社会的にビッグデータという用語が定着したことのあらわれとも考えられる。

2) <https://trends.google.com/trends/?geo=JP>

ることは明白であり、この中から有益な情報を効率的に抽出し、新たな知見の発見や意志決定などに活用する方法を模索することは現代社会における重要な課題であることにはかわりはないであろう。

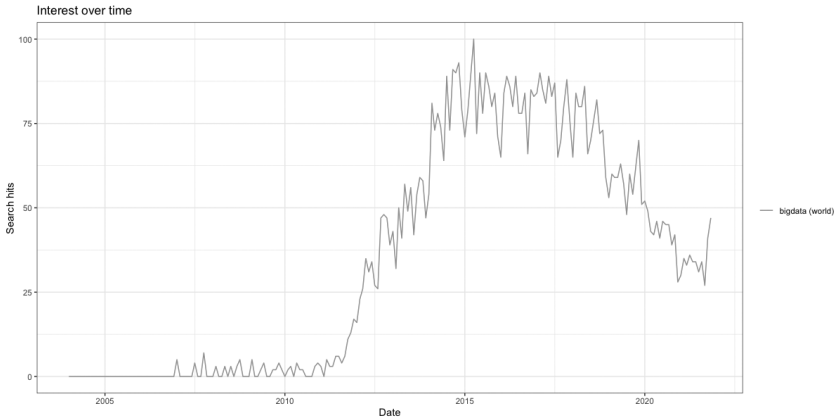


図 1 : Google Trends による 'bigdata' の検索動向 : 2004年 1月 1日 ~ 2021年 11月 11日

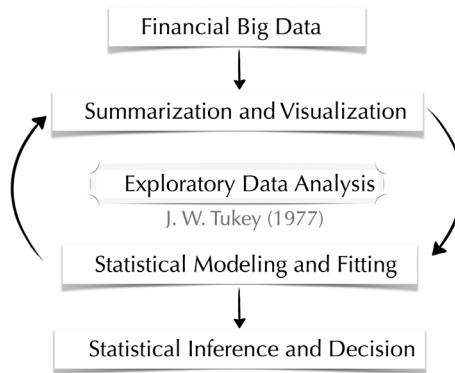


図 2 : 探索的データ解析

地道 (2014) では、当時のビッグデータ時代の到来を受けて、財務データ抽出システム (地道, 2010-a, b 参照) から、東京証券取引所第一部上場の

「全企業」（一般事業会社）に対する財務データを使用し、要約（summarization）、可視化³⁾（visualization）、統計モデリング⁴⁾（statistical modeling）、当てはめ（fitting）という循環を核とする探索的データ解析⁵⁾（Exploratory Data Analysis: EDA）を実行することによって（図2参照）、従業員数と資産合計による売上高の統計モデリングを行い、統計的推測・決定（statistical inference and decision）を行うことについて詳細に議論されている⁶⁾。

本稿では、地道（2021-a, b, c, d）によってリニューアルされた学内向け財務データ抽出システム SKWAD（スクワッド）から抽出された NEEDS 企業財務データ（一般事業会社）を、データ解析環境 R⁷⁾を利用して EDA を実行し、地道（2014）による結果を再検証する。特に、データラングリング⁸⁾や可視化には、**tidyverse** パッケージ群、**plotly** パッケージ等を利用し、モデリングには、Rにおいて非対称分布族（family of skew-symmetric distributions）を扱う **sn** パッケージを利用することによって、地道（2014）では扱うことが難しかった問題についても解決策を検討する。

本稿の構成は以下のようなものである。まず、本稿で扱う財務データの説明を行った後（Ⅱ節）、財務データを時間・空間の両面から可視化する（Ⅲ節）。この可視化によって得られた知見にもとづいてクロスセクションデータに対する回帰モデルによる統計モデリングを行い、実際にデータへ当ては

3) データ可視化（data visualization）に関する文献としては、Wilkinson (2005), Chen *et al.* (2008), Unwin (2015), Healy (2018), Kirk (2019) 等を参照されたい。また、より一般に情報可視化（information visualization）に関する文献としては、Tafte (2001), Mazza (2009), Ware (2012) 等を参照されたい。

4) 統計モデリングについては、例えば、Chambers and Hastie (1991) を参照されたい。

5) 探索的データ解析については、Tukey (1977), Mosteller and Tukey (1977) を参照されたい。また、探索的データ解析についての最近の文献としては、柴田 (2016), Wickham and Grolemond (2016), Bruce *et al.* (2020) 等を参照されたい。

6) いわゆる、コブ・ダグラス型生産関数（Cobb-Douglas type production function）の推定問題といえる（cf. Cobb and Douglas, 1928）。

7) Rについては、例えば、Kabacoff (2015), 地道 (2018) を参照されたい。なお、本稿では R version 4.1.2 (2021-11-01) を利用している。

8) データをRに読み込み、さらに分析・解析できるオブジェクトに変換する工程はデータラングリング（data wrangling）または単にラングリングと呼ばれる（cf. Wickham and Grolemond, 2016）。

めることによって、その妥当性の検証を探索的に行う（IV節）。さらに、業種情報をダミー変数として利用した両対数モデルをクロスセクションデータに当てはめることによって改良を試みた後（V節）、このモデルの経年変化にともなう安定性を決定係数や情報量規準によって検証する（VI節）。なお、両対数モデルの誤差分布として、正規分布に加えて、非対称分布族に属するものへ考察の対象を拡張する。最終節として、本稿を通じての総括を行うとともに今後の課題などについて述べる（VII節）。

付録には、本稿を作成したコンピュータ環境（付録A）とディレクトリ・ファイル構成（付録B）を与えるとともに、データラングリング（付録C）やEDAの実行に利用されたRのスクリプト（付録D）を与えている。また、本稿は、再現可能研究（reproducible research）の観点から執筆されており、Sweaveとmakeによる動的文書生成（dynamic documents）によって再現性を確保している（付録E）。特に、図3に与えるように、本稿を作成する全工程、すなわち、前処理⁹⁾からデータラングリング、探索的データ解析、動的文書生成による統計的推測・決定の公表までの全体を再現可能研究として実行している。さらに、回帰分析における感度分析に利用される指標（付録F）や非対称分布（付録G）の簡単な説明を与えている。最後に日経業種分類に関する情報を与えている（付録H）。

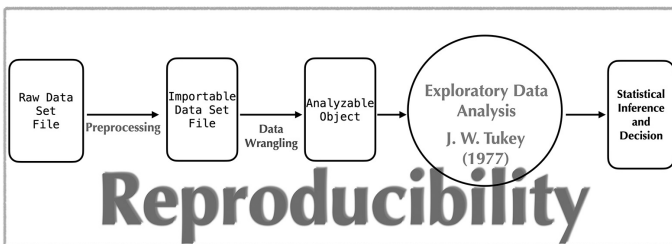


図3：前処理，データラングリング，探索的データ解析，統計的推測・決定の流れ

9) 本稿で扱ったデータはデータベースから抽出した時点で、Rに読み込めるファイル形式（CSVファイル）となっているため、特別な前処理（preprocessing）を必要としない。なお、CSV（Comma Sepelated Values）ファイルとは、項目（カラム）間がコンマ区切りのテキスト形式のファイルである。

II 財務データ

本稿で扱うデータ（表1）は、東京証券取引所第一部（以下「東証一部」と略）上場企業（一般事業会社）を母集団とする連結本決算（3月期決算分）にもとづく財務データである。

表1：日経NEEDS財務データベースから抽出した東証一部上場企業の財務データ（全データ42124件から先頭の10件を抜粋）

	name	ymd	sector1	sector2	sector3	ac	sales	employees	assets
1	KYOKUYO0000001	1989-03-01	2	35	341	1	213409	873	86649
2	KYOKUYO0000001	1990-03-01	2	35	341	1	207862	855	76786
3	KYOKUYO0000001	1991-03-01	2	35	341	1	202573	846	74061
4	KYOKUYO0000001	1992-03-01	2	35	341	1	199227	843	68312
5	KYOKUYO0000001	1993-03-01	2	35	341	1	184988	851	67760
6	KYOKUYO0000001	1994-03-01	2	35	341	1	164324	879	63693
7	KYOKUYO0000001	1995-03-01	2	35	341	1	173803	1029	61692
8	KYOKUYO0000001	1996-03-01	2	35	341	1	175202	1023	63287
9	KYOKUYO0000001	1997-03-01	2	35	341	1	183640	1000	65883
10	KYOKUYO0000001	1998-03-01	2	35	341	1	176022	976	62766

ここで、各列は以下のようなものである：

name	: 企業名+日経コード（1872社）
ymd	: 決算年月日（1984年3月期～2020年3月期）間の37年分）
month	: 決算月数
sector1	: 日経業種コード（大分類）（1：製造業，2：非製造業）
sector2	: 日経業種コード（中分類）（付録H参照）
sector3	: 日経業種コード（小分類）（付録H参照）
ac	: 会計基準（1：日本会計基準，2：米国基準，3：国際会計基準）
sales	: 売上高（単位：百万円）
employees	: 従業員数（単位：人）
assets	: 資産合計（単位：百万円）

このデータは、財務データ抽出システムSKWAD（地道, 2021-a参照）のNEEDS企業財務データ抽出機能を利用して得られたものであり、実際のデータ取得やそのラングリングについては、付録Cを参照されたい。利用

するデータの要約は以下のようなものである：

データの要約									
name		ymd	sector1		sector2		sector3		
Length:	42124	Min. :	1984-03-01	1:24467	23	:	4864	704	: 3974
Class :	character	1st Qu.:	1998-03-01	2:17657	71	:	4282	262	: 1281
Mode :	character	Median :	2006-03-01		43	:	3876	210	: 1210
		Mean :	2005-01-27		21	:	3564	225	: 1143
		3rd Qu.:	2013-03-01		07	:	3502	071	: 1091
		Max. :	2020-03-01		41	:	2603	224	: 1033
							(Other):19433	(Other):32392	
ac	sales	employees		assets					
Min. :	1.000	Min. :	239	Min. :	2	Min. :		190	
1st Qu.:	1.000	1st Qu.:	37351	1st Qu.:	855	1st Qu.:		40829	
Median :	1.000	Median :	90934	Median :	2077	Median :		93436	
Mean :	1.055	Mean :	398566	Mean :	7590	Mean :		523575	
3rd Qu.:	1.000	3rd Qu.:	257905	3rd Qu.:	5625	3rd Qu.:		279574	
Max. :	3.000	Max. :	30225681	Max. :	384586	Max. :		295849794	
		NA's :	36	NA's :	1058				

表1で与えられるデータは、一般に経時観測データ (longitudinal data) またはパネルデータ (panel data) と呼ばれるものである。この種のデータは、複数の個体 (ここでは東証一部上場企業) に対する属性 (売上高, 従業員数, 資産合計など) を (決算期において) 経時的に観測したものであり, 母集団 (ここでは, 東証一部上場企業全体) を時間・空間の両面から調査した結果として得られたものである。

III データ可視化

本稿で扱う財務データは、時間的・空間的な変動の両方を併せ持つ経時観測データであるので、その可視化には時空間のそれぞれの側面もしくは両面の観点からの以下のようなプロットが有益な情報を与える：

- すべての観測の時系列プロット
- 時点を固定した各種の散布図のプロット
- 時空間の両面からのプロット

以下にこれらのプロットを実際に描くことによってデータの可視化を行う。

1 時間的データ可視化

データの時間的な変化をみるためには各個体に対する時系列プロット (time-series plot) を描くことが最も基本的なものである。図4の左の列は、延べ1872社の個々の企業の売上高、従業員数、資産合計 (変量) に対するそれぞれの観測値を決算日において折れ線をつないだものであり、本稿で扱う売上高、従業員数、資産合計の全データがこのプロットにおいて表現されていることは注目に値する。

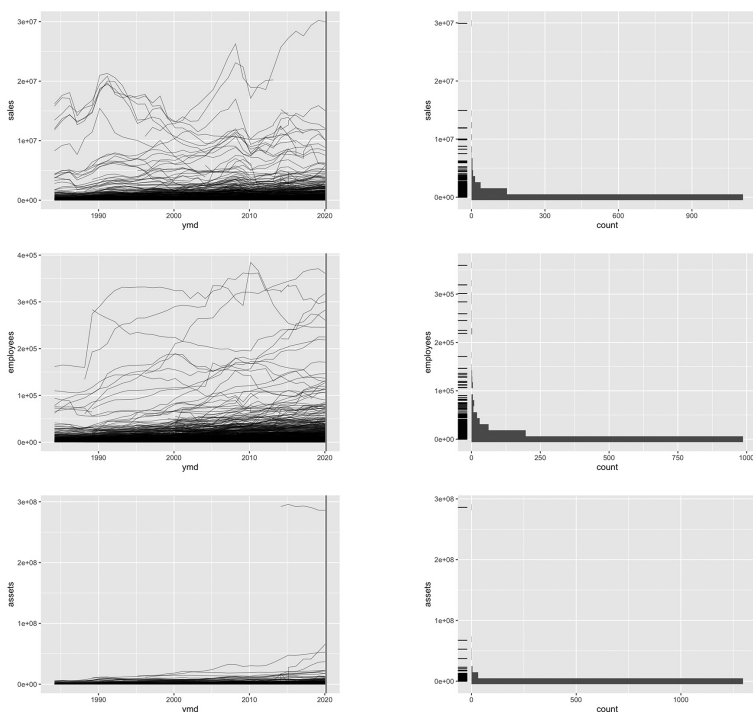


図4：東証一部上場企業の財務データの時系列プロットとヒストグラム：行列の形式で、(1, 1), (2, 1), (3, 1) ブロックに対応するプロットは、それぞれ、個々の企業の売上高 (sales), 従業員数 (employees), 資産合計 (assets) の時系列プロットであり、(1, 2), (2, 2), (3, 2) ブロックに対応するプロットは、2020年3月期で時点を固定し、横断面 (垂直線) をとったときヒストグラム (ラグ付) である。

この図から、各変量とも幾つかの「規模の大きな」企業が存在することがわかり、スケールの関係上、それ以外の企業についての変動がわかりづらい。特に、(3, 1) ブロックの資産合計の時系列プロットをみると、300兆円に迫る規模の資産合計を持つ企業（日本郵政）が2014年から存在することがわかる。このことから、全期間にわたって、売上高、従業員数、資産合計のそれぞれに対して、歪んだ分布構造を持つことがわかる。また、全期間にわたって財務データが与えられている企業があるのに対して、何らかの理由によって短期間しかデータが与えられていない企業が存在することもわかる。

2 空間的データ可視化

企業の財務データが空間的にどのように分布しているかを可視化することを考える。すなわち、時点を固定したときの母集団の分布状態を可視化するための様々なプロットを与える。一般に、ある時点で固定したもとの母集団に対する調査を行った結果として得られるデータはクロスセクションデータ（cross sectional data）または横断（面）データと呼ばれ、本稿で扱っている財務データでは、時点をたとえば2020年3月期で固定した場合が典型的なクロスセクションデータである。図4の(1, 2), (2, 2), (3, 2) ブロックは、東証一部上場企業の財務データの時系列プロットにおいて2020年3月期で時点を固定したもとのデータの分布状況をヒストグラムで可視化したものである。これらのプロットから、本稿で扱っている財務データは時点を2020年3月期で固定すると、変量毎に右に歪んだ分布（right-skewed distribution）に従うことがわかる。

次に、2変量間での同時分布を調べるためには、2組毎の変量に対する散布図を行列の形式に配置したプロット（図5）、すなわち、対散布図（pairwise scatter plot）または散布図行列（scatter plot matrix）が有益な情報を与える。図5におけるすべての散布図から原点付近でデータが「密集」しており原点から離れたところでは「疎」になっていることがわかる。この結果は1変量のヒストグラムのときにも見られたデータの歪みの2次元版と捉え

ることができる。

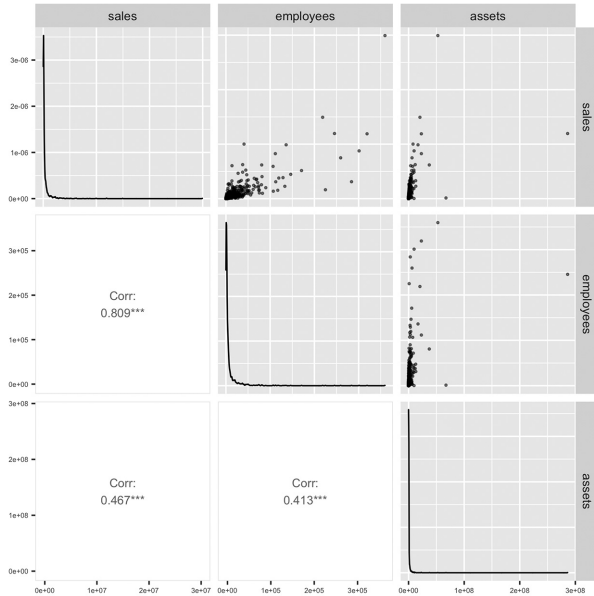


図5：東証一部における2020年3月期決算の企業の売上高，従業員数，資産合計の対散布図

さらに、3変量間での同時分布を調べるためには3次元散布図（three-dimensional scatter plot）を描くことによって実行できる。図6は2020年3月期決算の企業の売上高，従業員数，資産合計に関する3次元散布図であり、対散布図と同様に、このプロットからも原点付近でデータが「密集」しており原点から離れたところでは「疎」になっていることがわかる。この結果はデータの歪みを3次元で捉えたものと見なすことができる。

これらの結果から、2020年3月期で固定したクロスセクションデータは、原点付近で高密度をもち、原点から離れるにつれて低密度になる「歪んだ分布」に従ったものであることがわかった。地道（2014）でも、2012年3月期決算の東証一部上場企業の財務データに対する可視化の結果として同様の結

果が与えられており、8年が経過してもデータの（歪みに関する）分布構造が本質的に変化してないことがわかる。

3 時空間的データ可視化

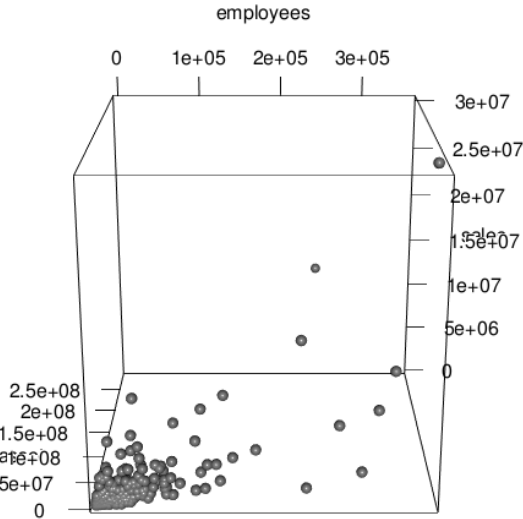


図6：東証一部における2020年3月期決算の企業の売上高，従業員数，資産合計の3次元散布図

これまでの考察は、データの時間的な推移とある時点での母集団（空間）の分布状況を別個に可視化するものであったが、これらの観点を融合し、時間的な推移に伴う母集団（空間）の分布状況を視覚で捉えることができれば、時間・空間の両面からデータの分布状況を把握することが可能となる。このことを実現するための一つの方法として、データのバブルチャート¹⁰⁾ (bub-

10) 一般に、散布図は、2変数データを $x-y$ 平面上の座標にマッピングすることによって可視化する方法であるが、バブルチャートは、さらにもう一つの変数に対するデータの値を点の大きさ（円の大きさ）にマッピングすることによって、3変数データを可視化するための統計グラフィックスである。名称は、バブル（泡）のような形状の点が平面に描かれるためと思われる。

ble chart) の時間的な推移を動的に可視化するモーションチャート (motion chart) がある。この可視化の手法は、経年変化にともなう空間的な分布状況をみることができる。モーションチャートを描くために、本稿では、Rの **ggplot2** パッケージと **plotly** パッケージを併用する方法を利用した¹¹⁾。この可視化 (図7) によって、東証一部上場企業の財務データの推移・変動を時間・空間両面から把握することができ、企業数や個々の企業の財務データに関して多少の変動はあるものの母集団における歪みを持つ分布構造に大きな変化が無いことが分かった。

4 可視化から与えられた示唆

これまでのデータ可視化の結果、データは歪みを持つことが示唆されたが、この情報を無視して正規分布にもとづく統計的推測や統計モデリングを行っても、適正な結果を得ることは難しい。この問題を解消するためには、地道 (2014) でも指摘されているように、データに対して対数 (logarithm) をとることである。このことによって、原点付近の小さな値が拡大され、かつ大きな値が圧縮されることによって、対称に近づけること (symmetrization) ができる場合がある (cf. Tukey, 1977, Moster and Tukey, 1977, Fox, 2015, Fox and Weisberg, 2019)。この観点にたち、これまでに与えられた時系列プロットとヒストグラム (図4)、対散布図 (図5)、3次元散布図 (図6)、モーションチャート (図7) を対数スケールで描きなおしたものが、それぞれ、図8, 9, 10, 11である。

11) 地道 (2014) では、モーションチャートを描くために、**googleVis** パッケージの `gvis-MotionChart` 関数を利用していたが、Adobe Flash のサポート終了にともない、利用できなくなった。

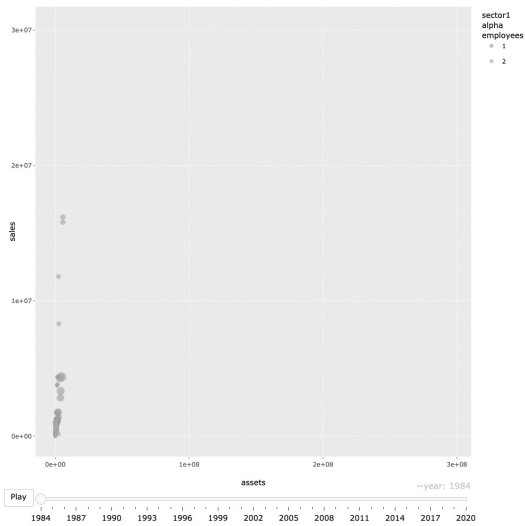
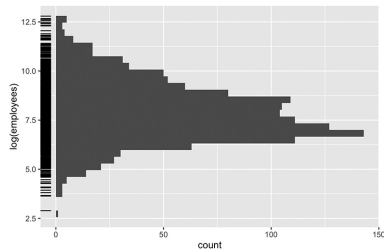
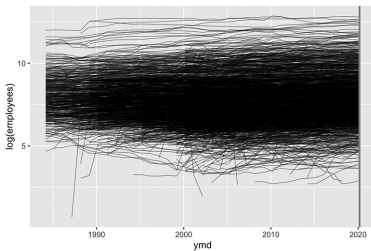
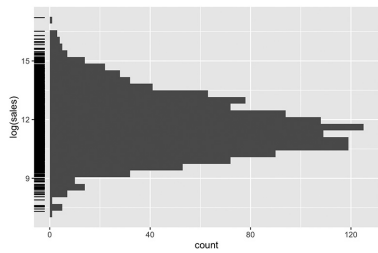
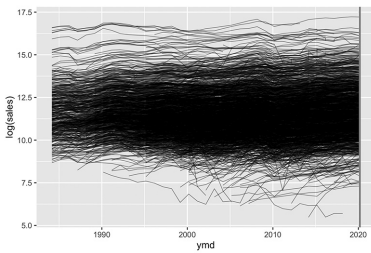


図7：ggplotlyによる東証一部における企業の売上高，従業員数，資産合計のバブルチャート：**Play**ボタンをクリックすることによってバブルチャートの時間的変遷を動的に可視化できる



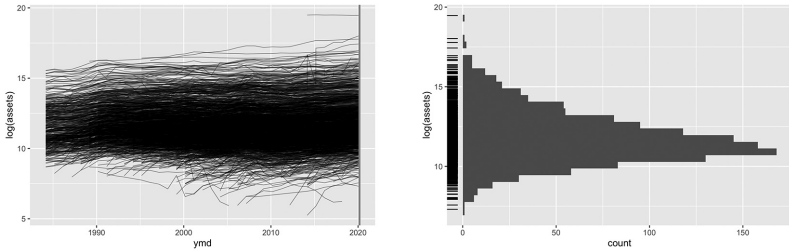


図8：東証一部上場企業の財務データ（対数スケール）の時系列プロットとヒストグラム：行列の形式で、(1, 1), (1, 2), (1, 3)ブロックに対応するプロットは、それぞれ、個々の企業の売上高、従業員数、資産合計の対数スケールの時系列プロットであり、(2, 1), (2, 2), (2, 3)ブロックに対応するプロットは、2020年3月期の時点を固定し、横断面（垂直線）をとったときのヒストグラム（ラグ付）である。

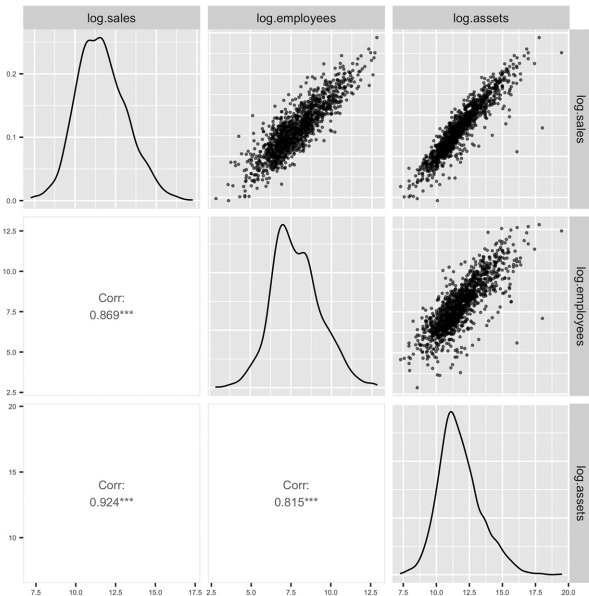


図9：東証一部における2020年3月期決算の企業の売上高、従業員数、資産合計の対散佈図（対数スケール）

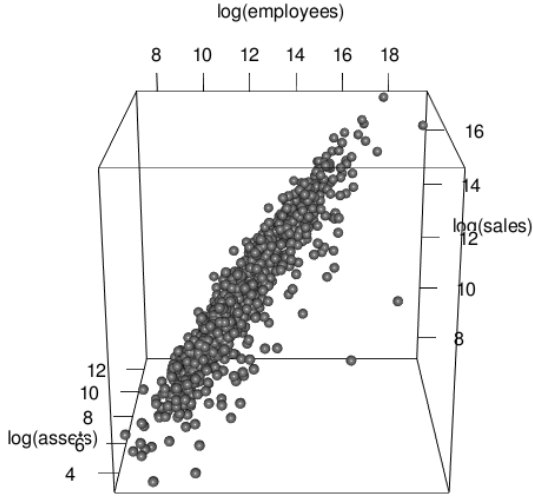


図10：東証一部における2020年3月期決算の企業の売上高，従業員数，資産合計の3次元散布図（対数スケール）

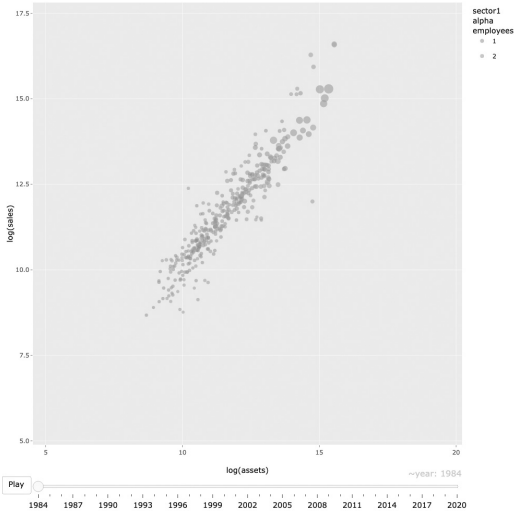


図11：gplotlyによる東証一部における企業の売上高，従業員数，資産合計（いずれも対数スケール）のバブルチャート

これらの可視化の結果から、対数スケールのデータの分布構造は経年変化を考慮しても、対称に近づくことがわかり、(多変量)正規分布をベースとして統計モデリングを行うことがある程度妥当であると考えられる(地道, 2014も参照)。しかしながら、対散布図(図9)を注意深くみると、分布は若干右に歪んでおり、対数資産合計(log.assets)と対数売上高(log.sales)の散布図((1,3)ブロック)は、楕円形をしているというよりも、右下から左上にかけて「歪曲」(slant)しているようにみることができる。このような構造をもつ分布をモデリングするためには、Azzalini (1985) や Azzalin and Capitanio (2014) によって提唱された非対称分布族に属する分布を利用することである。

次節以降で、Tukey (1977) によるEDAの視点に立ち、これらの可視化の知見をふまえて統計モデリングを行う。

IV クロスセクションデータに対する回帰モデルの当てはめ

この節では、財務データをクロスセクションの観点からとらえ、各種の回帰モデルを当てはめる。その際、時点は2020年3月期で固定する。

1 正規線形モデルの当てはめ

時点を2020年3月期で固定し、正規線形 (Normal Linear: NL) モデル

$$\text{sales}_i = \beta_0 + \beta_1 \text{employees}_i + \beta_2 \text{assets}_i + \epsilon_i, \quad \epsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathbf{N}(0, \sigma^2) \quad (1)$$

をクロスセクションデータに当てはめる。ただし、 $i=1, \dots, n$ ($n=1329$) である。正規線形モデルにおける回帰係数 (regression coefficients) $\beta_0, \beta_1, \beta_2$ を最小自乗法 (least square method) によって推定し、その最小自乗推定値 (Least Square Estimate: LSE) を $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ と書くことにする。

正規線形モデルにおける回帰係数の推定結果は表2のように与えられる。表2における 'Estimate' の列が最小自乗推定値を表しており、'Std. Error' の列が標準誤差 (standard error), 't value' の列がティ一値 (t -value), 'Pr (>|t|)' の列がピー値 (p -value) を表す。この結果から、回帰係数はすべて

5%有意である。

表2：ティール検定表：正規線形モデルの場合

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	55239.8301	22131.1163	2.50	0.0127
employees	37.2486	0.8618	43.22	0.0000
assets	0.0255	0.0027	9.35	0.0000

従業員数 (employees) と資産合計 (assets) を説明変数とする線形予測子

$$\eta := \beta_0 + \beta_1 \text{employees} + \beta_2 \text{assets}$$

は幾何学的には母回帰平面 (population regression plane) であるが、その回帰係数を最小自乗推定値でおきかえた

$$\hat{\eta}_{\text{NL}} = \hat{\beta}_0 + \hat{\beta}_1 \text{employees} + \hat{\beta}_2 \text{assets}$$

は標本回帰平面 (sample regression plane) と呼ばれ、実際に以下のように与えられる：

$$\hat{\eta}_{\text{NL}} = 55239.83 + 37.249 \text{employees} + 0.026 \text{assets} \quad (2)$$

図12に3次元散布図に標本回帰平面を描いたプロットを与える。

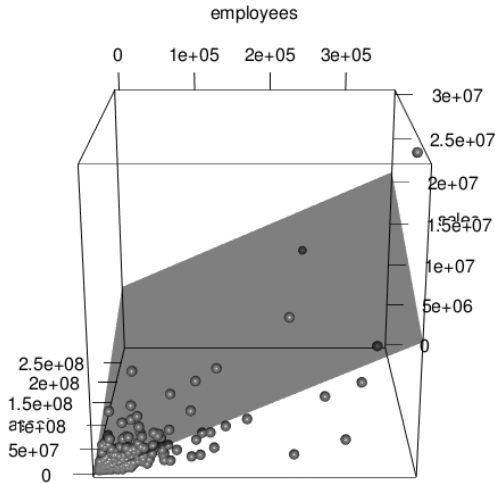


図12：2020年3月期決算の企業の財務データの3次元散布図と標本回帰平面（正規線形モデル）

また、このモデルを当てはめたときの誤差分散の推定値は、 $\hat{\sigma}^2=761146.424^2$ で与えられ、決定係数と自由度調整済み決定係数は以下のように与えられる：

$$R^2=0.6754, \quad \bar{R}^2=0.6749$$

決定率が約68%という結果をどのように見るかは判断の分かれるところであろうが、図12の標本回帰平面を勘案すると、当てはまりの悪いデータの存在が指摘される。このような状況において、回帰診断 (regression diagnostics) を行うことが推奨される (cf. Chatterjee and Hadi, 1988, Fox and Weisberg, 2019)。図13は回帰診断のための残差の各種のプロットである。これらのプロットは、誤差に関する仮定： $\epsilon_i \sim \overset{\text{i.i.d.}}{N}(0, \sigma^2)$ を検証するために利用される¹²⁾。行列形式で与えられた (1, 1) ブロックに対応する残差のインデッ

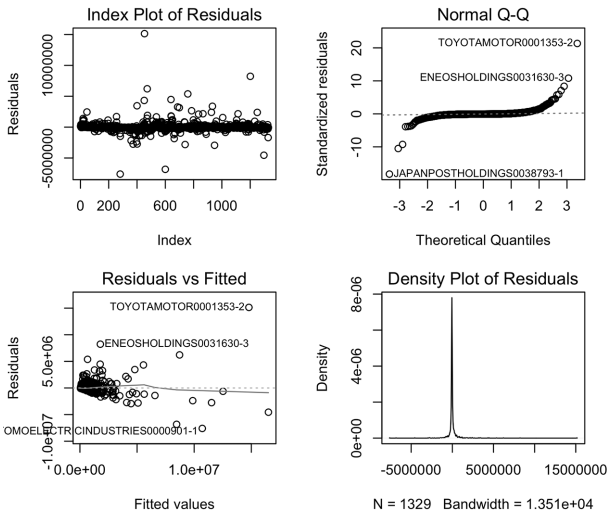


図13：2020年3月期決算の東証一部上場企業に関する財務データにもとづく正規線形モデルの当てはめ結果にもとづく残差に関する各種のプロット：行列形式で順に、(1, 1) ブロック：残差のインデックスプロット，(2, 1) ブロック：当てはめ値に対する残差のプロット，(1, 2) ブロック：残差の正規 Q-Q プロット，(2, 2) ブロック：残差の平滑化された密度関数のプロット。

12) 誤差は直接観測できないため、対応する残差を利用して誤差の仮定の検証が行われる。

クスプロットからは、相対的に大きな残差の存在が指摘され、(2, 1) ブロックの当てはめ値に対する残差のプロットは、「ファン形状」(fan-shape)を示す結果となっており、誤差の不均一分散性が指摘される (cf. Cook, 1998). さらに、(1, 2) ブロックの残差の正規 Q-Q プロットと (2, 2) ブロックの残差の平滑化された密度関数のプロットからは誤差の正規性が完全に疑われる結果となっている。

2 正規誤差をもつ両対数モデルの当てはめ

正規線形モデル (1) を2020年3月期決算の財務データに当てはめた結果から、このモデルは適切とは言いがたいことがわかった。そこで、前節で与えられた可視化による結果に基づいて統計モデリングを行う。以下のモデルを当てはめることが提案される：

$$\text{sales}_i = \gamma \times \text{employees}_i^{\alpha_1} \times \text{assets}_i^{\alpha_2} \times \epsilon_i, \quad \epsilon_i \stackrel{\text{i.i.d.}}{\sim} \text{LN}(0, \sigma^2) \quad (3)$$

このモデルは、一般には乗法モデル (productive model) と呼ばれる。ここで、誤差分布は対数正規分布 $\text{LN}(0, \sigma^2)$ である¹³⁾。

乗法モデル (3) の両辺の対数をとることによって正規線形モデルとして表現できる：

$$\begin{aligned} \log(\text{sales}_i) &= \alpha_0 + \alpha_1 \log(\text{employees}_i) + \alpha_2 \log(\text{assets}_i) + \log(\epsilon_i), \\ \log(\epsilon_i) &\stackrel{\text{i.i.d.}}{\sim} \text{N}(0, \sigma^2) \end{aligned} \quad (4)$$

ここでは、モデル (4) を正規誤差をもつ両対数モデル (double-log model) と呼ぶ¹⁴⁾。両対数モデルにおける回帰係数 $\alpha_0, \alpha_1, \alpha_2$ を最小自乗法によって推定したものを $\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2$ とおくと、このモデルにおける推定結果は表3のように与えられる。この結果から、回帰係数はすべて5%有意である。

13) モデル (3) は、コブ・ダグラス型生産関数 (cf. Cobb and Douglas, 1928) である。また、対数正規分布については、例えば、Crow and Shimizu (1988) を参照されたい。
14) 両対数モデルは、経済学や生物学などの様々な分野へ古くから応用されてきたものである。例えば、計量経済学への応用については Klein (1953, 1962)、生物学への応用については Rao (1973) を参照されたい。

表3：ティール検定表：正規誤差をもつ両対数モデルの場合

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.6819	0.1033	16.28	0.0000
log(employees)	0.3522	0.0157	22.37	0.0000
log(assets)	0.6106	0.0148	41.35	0.0000

標本回帰平面は、

$$\begin{aligned}\hat{\eta}_{\text{DLN}} &= \hat{\alpha}_0 + \hat{\alpha}_1 \log(\text{employees}) + \hat{\alpha}_2 \log(\text{assets}) \\ &= 1.682 + 0.352 \log(\text{employees}) + 0.611 \log(\text{assets})\end{aligned}\quad (5)$$

で与えられる。図14に対数スケールで描いた3次元散布図に正規誤差をもつ両対数モデルを当てはめたときの標本回帰平面を描いたプロットを与える。

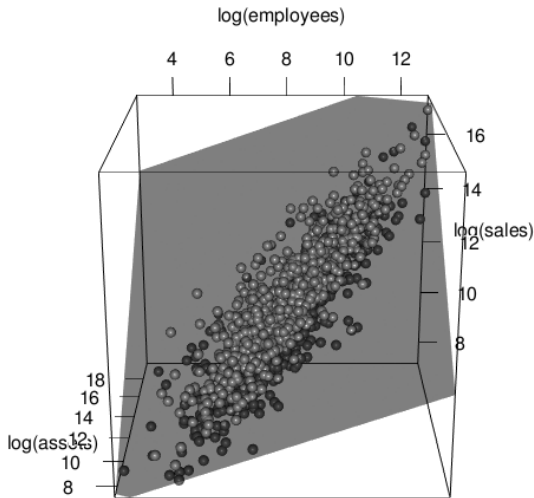


図14：2020年3月期決算の企業の財務データの3次元散布図（対数スケール）と標本回帰平面（両対数モデル：正規誤差）

このモデルを当てはめたときの誤差分散の推定値は、 $\sigma^2 = 0.501^2$ で与えられ、決定係数と自由度調整済み決定係数は以下のように与えられる：

$$R^2 = 0.8932, \quad \bar{R}^2 = 0.893$$

これらの結果において、決定率が約89%へ伸びており、図14の標本回帰平面

からも、当てはまりは向上していることがわかる。

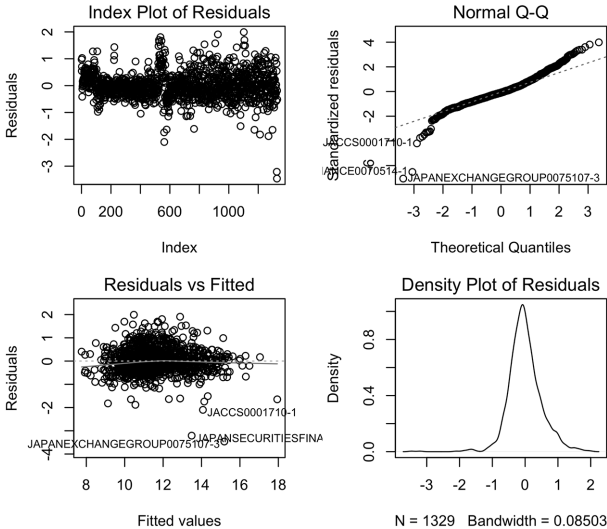


図15：2020年3月期決算の東証一部上場企業に関する財務データにもとづく両対数モデル（正規誤差）の当てはめ結果にもとづく残差に関する各種のプロット：行列形式で順に、(1, 1) ブロック；残差のインデックスプロット，(2, 1) ブロック；当てはめ値に対する残差のプロット，(1, 2) ブロック；残差の正規 Q-Q プロット，(2, 2) ブロック；残差の平滑化された密度関数のプロット。

ただし、回帰診断に関するプロット（図15）から、幾つかの影響力の強いデータの存在が指摘される。一般に、影響力あるデータを検出するための分析は感度分析（sensitivity analysis）とよばれ、専用の指標やプロットが提案されている（付録 F 参照）。ここでは、最も基本的なものであるハット値（hat value）、スチューデント化残差（Studentized residual）、クックの距離（Cook's distance）のインデックスプロットを与える（図16参照）。これらの指標を数値的に要約したものが、表 4 である。これらの結果から、JAPANEXCHANGEGROUP0075107-3（日本取引所グループ）が最も影響力の強いデータであり、続いて、JAPANSECURITIESFINANCE0070514-1（日本証券金融）

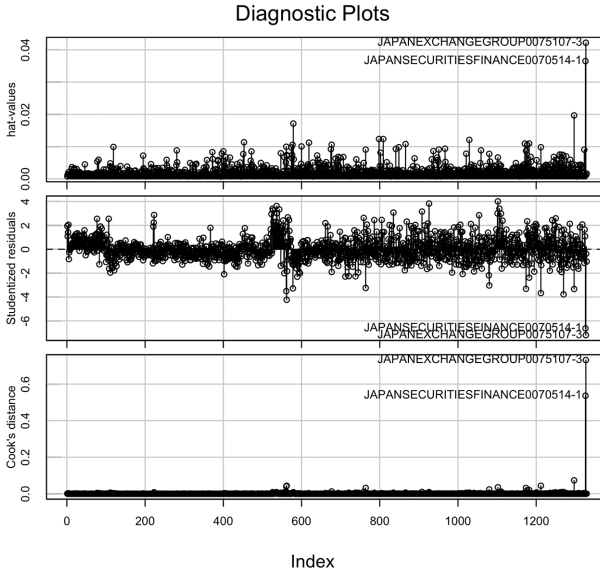


図16：2020年3月期決算の東証一部上場企業に関する財務データにもとづく両対数モデル（正規誤差）の当てはめた場合の回帰診断（感度分析）のためのプロット

JAPANPOSTHOLDINGS0038793-1（日本郵政）、JACCS0001710-1（ジャックス）が影響力が強いことが確認できる¹⁵⁾。

表4：回帰診断（感度分析）のための指標

	StudRes	Hat	CookD
JACCS0001710-1	-4.23	0.01	0.04
JAPANPOSTHOLDINGS0038793-1	-3.33	0.02	0.07
JAPANSECURITIESFINANCE0070514-1	-6.63	0.04	0.54
JAPANEXCHANGEGRUPO075107-3	-7.19	0.04	0.73

これらの指標についての簡単な説明を付録Fに与えるが、詳細は Chatterjee and Hadi (1988), Fox and Weisberg (2019) 等を参照されたい。

15) いずれも金融関連の企業であり、売上高と従業員数に対して資産合計が著しく大きい企業であることが別途わかる。

以上の感度分析の結果から、これらのデータを異質なものとして取り除き、再度正規誤差をもつ両対数モデルを当てはめる。再当てはめの結果は表5のように与えられる。

表5：ティー検定表：影響力のあるデータ除去後の正規誤差をもつ両対数モデルの場合

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.4258	0.1009	14.13	0.0000
log(employees)	0.3055	0.0155	19.70	0.0000
log(assets)	0.6643	0.0148	45.00	0.0000

この結果も、回帰係数はすべて5%有意である。標本回帰平面は、

$$\hat{\eta}_{DLN,adj} = 1.426 + 0.305 \log(\text{employees}) + 0.664 \log(\text{assets}) \quad (6)$$

で与えられ、図17に対数スケールで描いた3次元散布図に両対数モデル（正規誤差）を当てはめたときの標本回帰平面を描いたプロットを与える。

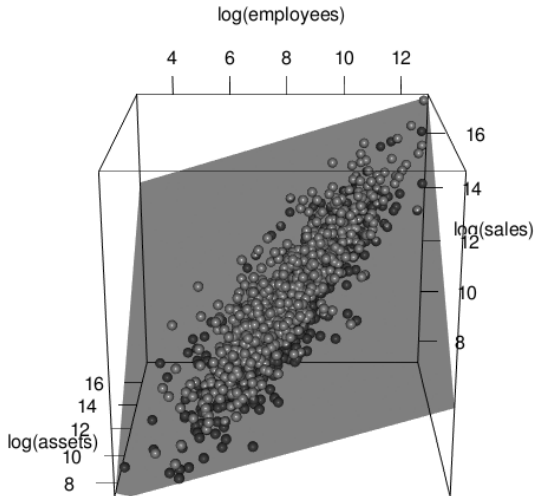


図17：2020年3月期決算の企業の財務データの3次元散布図（対数スケール）と標本回帰平面（両対数モデル：正規誤差，影響力のあるデータ除去後）

このモデルを当てはめたときの誤差分散の推定値は、 $\hat{\sigma}^2 = 0.477^2$ で与えら

れ、決定係数と自由度調整済み決定係数は以下のように与えられる：

$$R^2=0.9029, \bar{R}^2=0.9027$$

この結果において、決定率が約90%へ若干伸びていることが分かる。また、回帰診断に関するプロット（図18）からもとくに注意すべき影響力の強いデータは存在しない。

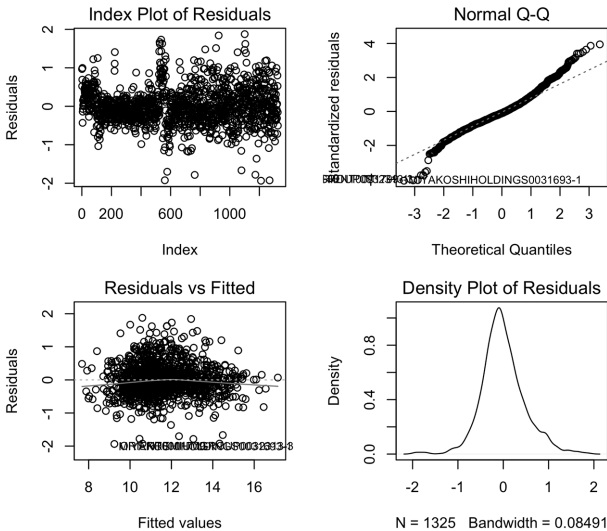


図18：2020年3月期決算の東証一部上場企業に関する財務データ（影響力があるデータ除去後）にもとづく両対数モデル（正規誤差）の当てはめ結果にもとづく残差に関する各種のプロット：行列形式で順に、(1, 1)ブロック；残差のインデックスプロット，(2, 1)ブロック；当てはめ値に対する残差のプロット，(1, 2)ブロック；残差の正規 Q-Q プロット，(2, 2)ブロック；残差の平滑化された密度関数のプロット。

しかしながら、回帰診断のプロット（図18）における残差の正規 Q-Q プロットを見ると、残差は裾の部分で正規分布に従っているか疑問が残り、このことから誤差の正規性は疑われる。なお、この現象は、地道（2014）でもみられたが、議論が不十分であった。本稿では、この問題に対して、地道（2017-a, b）、Jimichi *et al.* (2018)、地道、阪（2021）で扱った非対称正規（Skew-Normal: SN）分布や非対称ティー（Skew-t: ST）分布等の非対称分

布族に従う誤差を仮定したモデルによって説明することを考える。なお、付録 G にこれらの分布の簡単な説明を与えるが、詳細は Azzalini (1985) や Azzalini and Capitanio (2014) を参照されたい。

3 非対称正規誤差をもつ両対数モデルの当てはめ

正規誤差をもつ両対数モデル (4) の誤差分布を非対称正規分布に変更した以下のモデルを考える：

$$\begin{aligned} \log(\text{sales}_i) &= \alpha_0 + \alpha_1 \log(\text{employees}_i) + \alpha_2 \log(\text{assets}_i) + \log(\epsilon_i), \\ \log(\epsilon_i) &\stackrel{\text{i.i.d.}}{\sim} \text{SN}(0, \omega^2, \alpha) \end{aligned} \quad (7)$$

2020年3月期決算のデータから影響力のあるもの¹⁶⁾を削除したものにモデル (7) を当てはめた結果を表 6 に与える：

表 6：ゼット比検定表：非対称正規誤差をもつ両対数モデルの場合

	estimate	std.err	z-ratio	Pr{ z }
(Intercept.DP)	1.0715	0.0954	11.23	0.0000
log(employees)	0.3530	0.0161	21.97	0.0000
log(assets)	0.6259	0.0147	42.60	0.0000
omega	0.6447	0.0231	27.94	0.0000
alpha	1.5972	0.1712	9.33	0.0000

表 6 における ‘estimate’ の列には、最尤推定値 (Maximum Likelihood Estimate: MLE) が与えられており、全ての回帰係数と母数は有意になっていることがわかる¹⁷⁾。この結果から、修正標本回帰平面 (例えば、地道, 2017-b, Jimichi *et al.*, 2018, 地道, 阪, 2021参照) を求めると、

$$\begin{aligned} \hat{\eta}_{\text{DLN,adj}} &= (\hat{\alpha}_0 + \hat{\omega}b\hat{\delta}) + \hat{\alpha}_1 \log(\text{employees}) + \hat{\alpha}_2 \log(\text{assets}) \\ &= (1.072 + 0.645 \times 0.798 \times 0.848) + 0.353 \log(\text{employees}) \\ &\quad + 0.626 \log(\text{assets}) \end{aligned}$$

16) ここで、影響力のあるデータは、JAPANEXCHANGEGROUP0075107-3 (日本取引所グループ)、JAPANSECURITIESFINANCE0070514-1 (日本証券金融)、JAPANPOST-HOLDINGS0038793-1 (日本郵政)、JACCS0001710-1 (ジャックス) であり、以下の議論ではこれらのデータは除去されている。

17) 最尤法と最尤推定量の漸近的性質については、例えば、稲垣 (2003) を参照されたい。

$$=1.508+0.353\log(\text{employees})+0.626\log(\text{assets}) \quad (8)$$

となる。ここで、 $b:=\sqrt{2/\pi}$ 、 $\delta:=\hat{a}/\sqrt{1+\hat{a}^2}$ であり、図19に対数スケールで描いた3次元散布図に両対数モデル（非対称正規誤差）を当てはめたときの標本回帰平面を描いたプロットを与える。

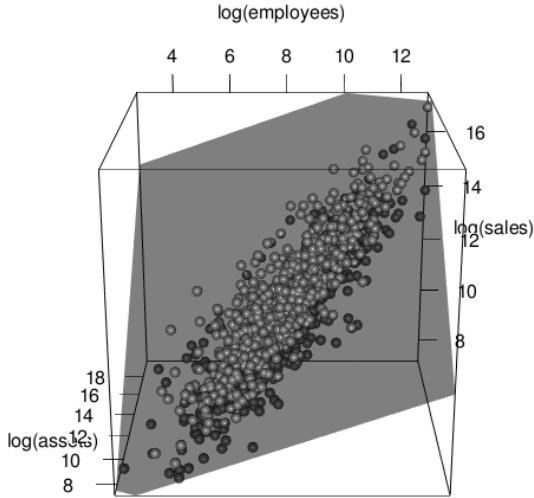


図19：2020年3月期決算の企業の財務データの3次元散布図（対数スケール）と標本回帰平面（両対数モデル：非対称正規誤差，影響力のあるデータ除去後）

回帰診断に関するプロット（図20）から、P-Pプロットが若干直線（理想的な状態）から乖離していることがわかり、モデルが誤差分布の構造を捉えきれていないと考えられる。なお、これらのプロットに利用される中心化母数残差（centered parameter (CP) residual）や尺度調整された直接母数残差（scaled direct parameter (DP) residual）については、Azzalin and Capitanio (2014), Jimichi *et al.* (2018) を参照されたい。

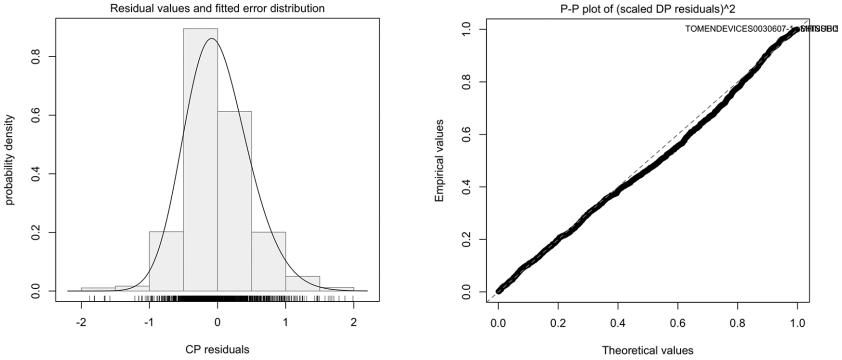


図20：非対称正規誤差をもつ両対数モデルの当てはめに伴う回帰診断に関する各種のプロット：中心化母数残差のヒストグラムと統計モデル（左），尺度調整された直接母数残差の2乗のP-Pプロット（右）

4 非対称ティール誤差をもつ両対数モデルの当てはめ

非対称ティール誤差をもつ以下の両対数モデルを考える：

$$\begin{aligned} \log(\text{sales}_i) = & \alpha_0 + \alpha_1 \log(\text{employees}_i) + \alpha_2 \log(\text{assets}_i) \\ & + \log(\epsilon_i), \quad \log(\epsilon_i) \stackrel{\text{i.i.d.}}{\sim} \text{ST}(0, \omega^2, \alpha, \nu) \end{aligned} \quad (9)$$

モデル (9) を2020年3月期決算の財務データ¹⁸⁾に当てはめた結果を表7に与える：

表7：ゼット比検定表：非対称ティール誤差をもつ両対数モデルの場合

	estimate	std.err	z-ratio	Pr{> z }
(Intercept.DP)	1.0861	0.0926	11.73	0.0000
log(employees)	0.3211	0.0156	20.60	0.0000
log (assets)	0.6585	0.0144	45.74	0.0000
omega	0.4311	0.0279	15.44	0.0000
alpha	0.9671	0.1885	5.13	0.0000
nu	4.5650	0.6412	7.12	0.0000

18) JAPANEXCHANGEGROUP0075107-3（日本取引所グループ），JAPANSECURITIESFINANCE0070514-1（日本証券金融），JAPANPOSTHOLDINGS0038793-1（日本郵政），JACCS0001710-1（ジャックス）はデータから取り除かれている。

表7に与えられている結果から、全ての回帰係数と母数は有意になっていることがわかる。この結果から、修正標本回帰平面（例えば、地道, 2017-b, Jimichi *et al.*, 2018, 地道, 阪, 2021参照）を求めると、

$$\begin{aligned}\hat{\eta}_{\text{DLS.T.adj}} &= (\hat{\alpha}_0 + \hat{\omega} b_s \hat{\delta}) + \hat{\alpha}_1 \log(\text{employees}) + \hat{\alpha}_2 \log(\text{assets}) \\ &= (1.086 + 0.431 \times 0.968 \times 0.695) + 0.321 \log(\text{employees}) \\ &\quad + 0.659 \log(\text{assets}) \\ &= 1.376 + 0.321 \log(\text{employees}) + 0.659 \log(\text{assets}) \quad (10)\end{aligned}$$

ここで、 $b_s := \sqrt{\hat{\nu}} / \pi \Gamma((\hat{\nu}-1)/2) / \Gamma(\hat{\nu}/2)$, $\hat{\delta} = \hat{\alpha} / \sqrt{1 + \hat{\alpha}^2}$ であり、図21に対数スケールで描いた3次元散布図に両対数モデル（非対称ティー誤差）を当てはめたときの標本回帰平面を描いたプロットを与える。

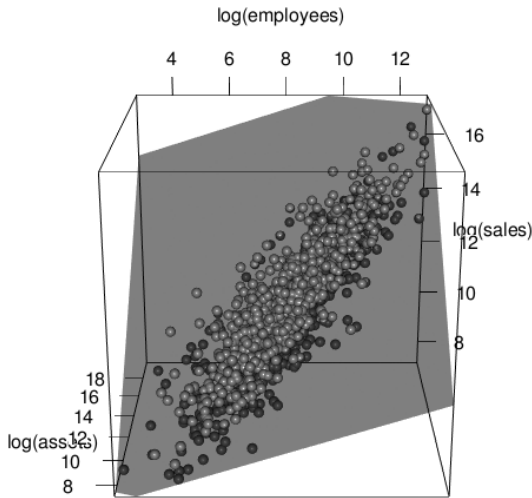


図21：2020年3月期決算の企業の財務データの3次元散布図（対数スケール）と標本回帰平面（両対数モデル：非対称ティー誤差、影響力のあるデータ除去後）

回帰診断に関するプロット（図22）からも問題がないことがわかる。

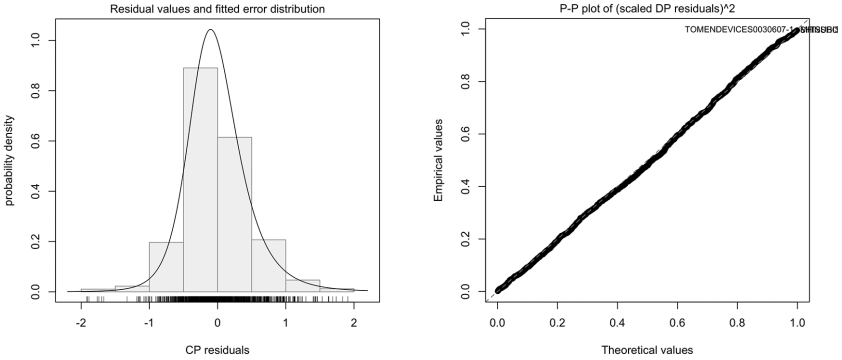


図22：非対称ティー誤差をもつ両対数モデルの当てはめに伴う回帰診断に関する各種のプロット：中心化母数残差のヒストグラムと統計モデル（左），尺度調整された直接母数残差の2乗のP-Pプロット（右）

以上の結果から，両対数モデルとしては，非対称ティー誤差をもつものが当てはまりがよいということが予想される．なお，このことに関しては次の小節で赤池情報量規準¹⁹⁾（Akaike Information Criterion: AIC）で評価する．

5 両対数モデルに関する選択

影響力のあるデータを除いたデータに，正規誤差（log.lm.x20200301.otl），非対称正規誤差（log.selm.x20200301.otl），非対称ティー誤差（log.selm.ST.x20200301.otl）を持つ両対数モデルを当てはめたときの，それぞれのモデルに対する母数（ベクトル）の次元（dim）とAICの値を表8に与える．

表8：AIC表：両対数モデルに関する比較

	dim	AIC
log.lm.x20200301.otl	4	1803.59
log.selm.x20200301.otl	5	1773.74
log.selm.ST.x20200301.otl	6	1693.37

この結果から，非対称ティー誤差をもつ両対数モデル（log.selm.ST.x

19) 赤池情報量規準については，Akaike (1973), Konishi and Kitagawa (2007) 等を参照のこと．

20200301.otl) が最も良いことがわかった。

V クロスセクションデータに対するダミー変数をもつ両対数モデルの当てはめ

これまでの統計モデリングとデータへの当てはめによって、2020年3月期決算の企業に対する財務データ（クロスセクションデータ）に対して90%近くの決定率をもつ売上高を説明するためのモデル（正規誤差をもつ両対数モデル）が構築できた。また、このモデルの当てはめに関する回帰診断から、誤差構造に非対称分布族を仮定したもの、特に非対称ティー誤差をもつものがより適切であることもわかった。

一方、バブルチャート（図23）を見ると、業種（中分類を採用）毎にデータにある種の傾向があることがわかる。具体的には、モデルの「切片」が業種毎に異なっていることが予想できる。

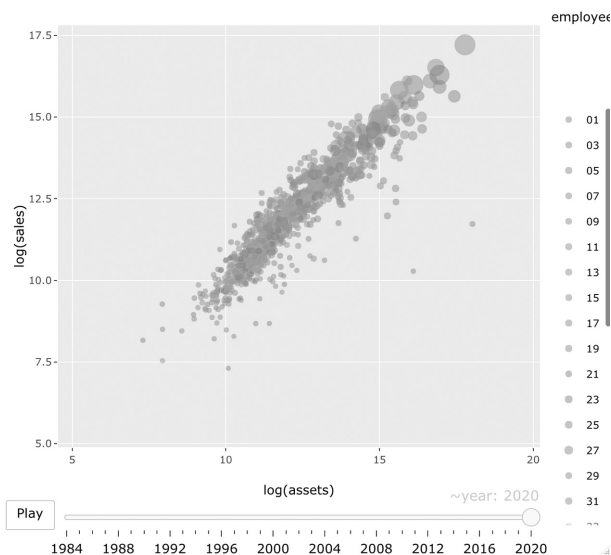


図23：モーションチャートによる2019年度（2020年3月）決算の企業に関する財務データのバブルチャート：日経業種中分類にもとづいて色分けしたもの

この可視化による情報を利用した統計モデリングの最も単純なものは、ダミー変数を利用したモデルの拡張である (cf. 地道, 2014). その際、日経業種分類 (付録 H) の中分類の情報 (33業種) を利用する²⁰⁾.

本節では、ダミー変数をモデルに入れることによって、モデルの改良を行うことを考える.

1 正規誤差とダミー変数をもつ両対数モデルの当てはめ

$$\log(\text{sales}_i) = \alpha_0 + \alpha_1 \log(\text{employees}_i) + \alpha_2 \log(\text{assets}_i) + \sum_{j=1}^m \delta_j D_{ij} + \log(\epsilon_i), \quad \log(\epsilon_i) \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2) \quad (11)$$

ここで、 $j=1, \dots, m (=33)$ であり²¹⁾,

$$D_{ij} := \begin{cases} 1, & \text{企業 } i \text{ が } j \text{ 番目の業種に属するとき,} \\ 0, & \text{企業 } i \text{ が } j \text{ 番目の業種に属さないとき} \end{cases}$$

とする. なお、推定の一意性のために $\delta_1=0$ とする. このモデルは両対数モデルに業種情報をダミー変数として追加したものである.

このモデルにける回帰係数の推定結果が表 9 に与えられている. すべての回帰係数に対する検定結果が 5% 有意という結果ではないが、ほとんどの回帰係数は有意であることがわかる²²⁾.

標本回帰平面 (群) は,

$$\begin{aligned} \hat{\eta}_{\text{DLN}_j} &= (\hat{\alpha}_0 + \hat{\delta}_j) + \hat{\alpha}_1 \log(\text{employees}) + \hat{\alpha}_2 \log(\text{assets}) \\ &= (1.446 + \hat{\delta}_j) + 0.328 \log(\text{employees}) + 0.675 \log(\text{assets}), \\ & \quad j=1, \dots, 33 \end{aligned} \quad (12)$$

で表される (図 24 参照). ここで、 $\hat{\delta}_j$ は δ_j に対する最小自乗推定値であり、標本回帰平面における業種 j 毎の切片項の調整と見なすことができる.

20) 大分類は 2 種類と少なく、小分類は 129 種類のうち各業種に属する企業数が 5 社以下となるものが 66 業種あり、逆に細分化されすぎるきらいがある.

21) たとえば、 j が 1 の場合は、業種コードが 01 の「食品業」に対応し、33 の場合は、業種コード 71 の「サービス業」に対応する.

22) とくに、日経業種コード 35 (水産業) と 37 (鉱業) に対する回帰係数は有意とはいえない. この点に関しては、水産業と鉱業に属する企業数が、それぞれ、5 社と 4 社という小数であり、これらの業種の構造上の制約が影響しているものと思われる.

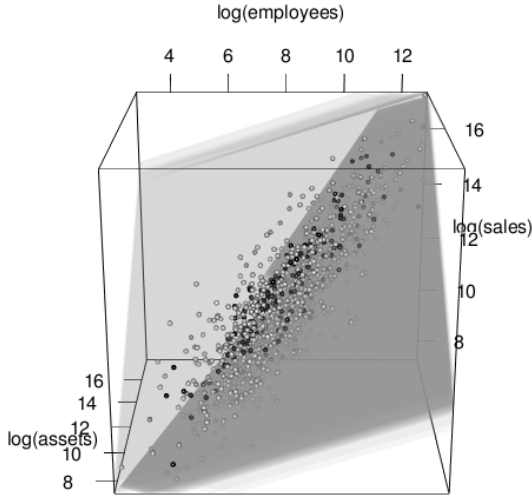


図24：2020年3月期決算の企業の財務データの3次元散佈図（対数スケール）と標本回帰平面群（両対数モデル：正規誤差，影響力のあるデータ除去，ダミー変数含む）

たとえば， $j=2$ のとき，日経業種コード（中分類）03は「繊維業」を表し，この業種に対する標本回帰平面は以下のように与えられる：

$$\begin{aligned}\hat{\eta}_{\text{DLN.textile}} &= 1.446 + (-0.611) + 0.328 \log(\text{employees}) \\ &\quad + 0.675 \log(\text{assets}) \\ &= 0.835 + 0.328 \log(\text{employees}) + 0.675 \log(\text{assets})\end{aligned}$$

なお， $j=1$ のとき，日経業種コード（中分類）01は「食品業」を表し，この業種に対する係数は $\delta_1=0$ と定義していたので，

$$\hat{\eta}_{\text{DLN.food}} = 1.446 + 0.328 \log(\text{employees}) + 0.675 \log(\text{assets})$$

となる。

ダミー変数をもつ両対数モデル（11）を当てはめたときの誤差分散の推定値は， $\hat{\sigma}^2=0.351^2$ で与えられ，決定係数と自由度調整済み決定係数は

$$R^2=0.9486, \quad \bar{R}^2=0.9472$$

となり，約95%という高い決定率をもっていることがわかる。なお，企業の

業種情報は容易に入手可能であることから、このモデルの拡張が有用であることがわかる。

表9：ティール検定表：影響力のあるデータを除去後、業種コードに対応したダミー変数をもつ両対数モデルの場合

	Estimate	Std. Error	t value	Pr (> t)
(Intercept)	1.4457	0.1012	14.29	0.0000
log(employees)	0.3279	0.0151	21.79	0.0000
log(assets)	0.6750	0.0146	46.14	0.0000
sector203	-0.6109	0.0852	-7.17	0.0000
sector205	-0.3744	0.1170	-3.20	0.0014
sector207	-0.4007	0.0602	-6.66	0.0000
sector209	-0.6599	0.0821	-8.04	0.0000
sector211	0.3779	0.1423	2.66	0.0080
sector213	-0.5954	0.1425	-4.18	0.0000
sector215	-0.4813	0.0849	-5.67	0.0000
sector217	-0.3707	0.0818	-4.53	0.0000
sector219	-0.3371	0.0713	-4.73	0.0000
sector221	-0.5810	0.0604	-9.62	0.0000
sector223	-0.5528	0.0595	-9.29	0.0000
sector225	-0.3677	0.1827	-2.01	0.0444
sector227	-0.2996	0.0720	-4.16	0.0000
sector229	-0.3690	0.1339	-2.76	0.0059
sector231	-0.6591	0.0853	-7.73	0.0000
sector233	-0.4608	0.0766	-6.02	0.0000
sector235	0.0300	0.1648	0.18	0.8558
sector237	-0.1147	0.1829	-0.63	0.5306
sector241	-0.0341	0.0629	-0.54	0.5874
sector243	0.3982	0.0587	6.78	0.0000
sector245	0.1112	0.0705	1.58	0.1149
sector252	-1.1824	0.0884	-13.38	0.0000
sector253	-0.4478	0.0813	-5.51	0.0000
sector255	-0.9716	0.0881	-11.03	0.0000
sector257	-0.3995	0.0949	-4.21	0.0000
sector259	-0.4390	0.1353	-3.25	0.0012
sector261	-0.5010	0.2094	-2.39	0.0169
sector263	-0.2975	0.0946	-3.14	0.0017
sector265	-0.3130	0.0934	-3.35	0.0008
sector267	-0.5217	0.1118	-4.66	0.0000
sector269	-0.2930	0.1521	-1.93	0.0542
sector271	-0.3440	0.0564	-6.10	0.0000

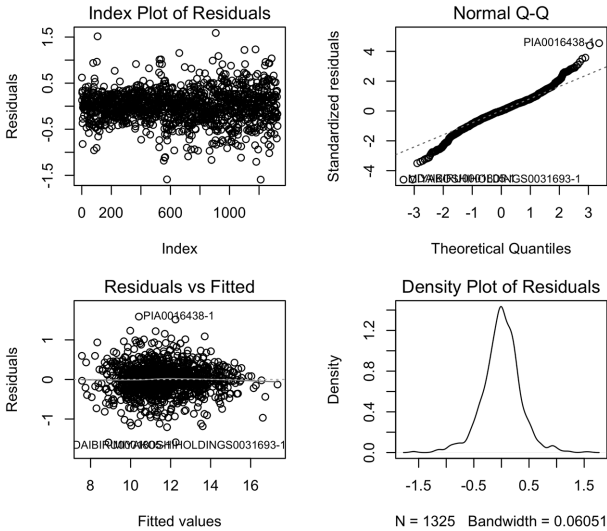


図25：2020年3月期決算の東証一部上場企業に関する財務データ（影響力があるデータ除去後）にもとづくダミー変数をもつ両対数モデル（正規誤差）の当てはめ結果にもとづく残差に関する各種のプロット：行列形式で順に、(1, 1) ブロック；残差のインデックスプロット，(2, 1) ブロック；当てはめ値に対する残差のプロット，(1, 2) ブロック；残差の正規 Q-Q プロット，(2, 2) ブロック；残差の平滑化された密度関数のプロット。

ただし、回帰診断のプロット（図25）における残差の正規 Q-Q プロットを見ると、誤差の正規性は疑われる。このことから、（ダミー変数をもたない）両対数モデルと同様に誤差分布として非対称分布族を検討する。

2 非対称正規誤差とダミー変数をもつ両対数モデルの当てはめ

非対称正規誤差とダミー変数をもつ以下の両対数モデルを考える：

$$\log(\text{sales}_i) = \alpha_0 + \alpha_1 \log(\text{employees}_i) + \alpha_2 \log(\text{assets}_i) + \sum_{j=1}^m \delta_j D_{ij} + \log(\epsilon_i), \quad \log(\epsilon_i) \stackrel{\text{i.i.d.}}{\sim} \text{SN}(0, \omega^2, \alpha) \quad (13)$$

このモデルにける回帰係数の推定結果を表11に与える。すべての回帰係数に対する検定結果が5%有意という結果ではないが、ほとんどの回帰係数は有意であることがわかる。修正標本回帰平面（群）は、

$$\begin{aligned}\hat{\eta}_{\text{DLSN}_j} &= (\hat{a}_0 + \hat{\omega}b\hat{\delta} + \hat{\delta}_j) + \hat{a}_1 \log(\text{employees}) + \hat{a}_2 \log(\text{assets}) \\ &= (1.423 + \hat{\delta}_j) + 0.313 \log(\text{employees}) + 0.687 \log(\text{assets}), \\ j &= 1, \dots, 33\end{aligned}\tag{14}$$

表10：ゼット比検定表：影響力のあるデータを除去後，非対称正規誤差と業種コードに対応したダミー変数をもつ両対数モデルの場合

	estimate	std.err	z-ratio	Pr{> z }
(Intercept.DP)	1.6804	0.1025	16.40	0.0000
log(employees)	0.3125	0.0157	19.86	0.0000
log(assets)	0.6871	0.0150	45.75	0.0000
sector203	-0.5916	0.0843	-7.02	0.0000
sector205	-0.3975	0.1139	-3.49	0.0005
sector207	-0.4194	0.0591	-7.09	0.0000
sector209	-0.6836	0.0804	-8.51	0.0000
sector211	0.4044	0.1409	2.87	0.0041
sector213	-0.6096	0.1384	-4.41	0.0000
sector215	-0.4978	0.0829	-6.01	0.0000
sector217	-0.3951	0.0801	-4.93	0.0000
sector219	-0.3460	0.0698	-4.96	0.0000
sector221	-0.5944	0.0592	-10.04	0.0000
sector223	-0.5503	0.0583	-9.44	0.0000
sector225	-0.3920	0.1772	-2.21	0.0269
sector227	-0.3013	0.0704	-4.28	0.0000
sector229	-0.3832	0.1303	-2.94	0.0033
sector231	-0.6742	0.0832	-8.10	0.0000
sector233	-0.4674	0.0749	-6.24	0.0000
sector235	0.0622	0.1619	0.38	0.7009
sector237	-0.1465	0.1789	-0.82	0.4127
sector241	-0.0569	0.0618	-0.92	0.3572
sector243	0.4185	0.0579	7.22	0.0000
sector245	0.1267	0.0691	1.83	0.0667
sector252	-1.1702	0.0866	-13.50	0.0000
sector253	-0.4101	0.0805	-5.10	0.0000
sector255	-0.9880	0.0860	-11.49	0.0000
sector257	-0.3984	0.0926	-4.30	0.0000
sector259	-0.4630	0.1323	-3.50	0.0005
sector261	-0.5220	0.2026	-2.58	0.0100
sector263	-0.2931	0.0931	-3.15	0.0016
sector265	-0.3065	0.0917	-3.34	0.0008
sector267	-0.5357	0.1096	-4.89	0.0000
sector269	-0.3215	0.1479	-2.17	0.0298
sector271	-0.3311	0.0556	-5.96	0.0000
omega	0.4315	0.0199	21.65	0.0000
alpha	-1.1235	0.1807	-6.22	0.0000

で表される (図26参照).

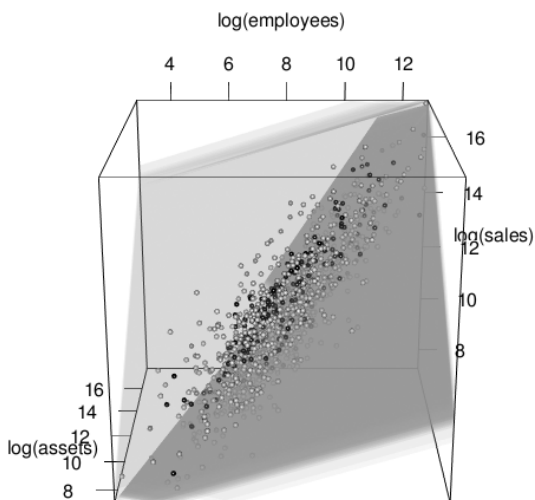


図26：2020年3月期決算の企業の財務データの3次元散佈図（対数スケール）と修正標本回帰平面群（両対数モデル：非対称正規誤差，影響力のあるデータ除去，ダミー変数含む）

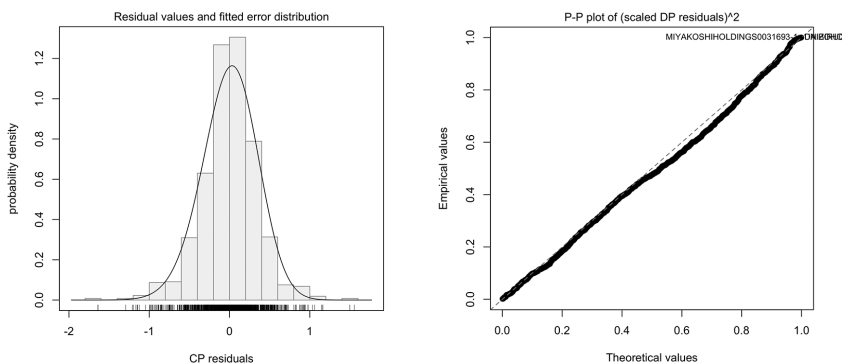


図27：2020年3月期決算の東証一部上場企業に関する財務データ（影響力があるデータ除去後）へ非対称正規誤差とダミー変数をもつ両対数モデルを当てはめた際の回帰診断に関する各種のプロット：中心化母数残差のヒストグラムと統計モデル（左），尺度調整された直接母数残差の2乗のP-Pプロット（右）

ただし、回帰診断のプロット（図27）から、当てはまりに問題があることがわかり、非対称正規誤差を仮定することは検討を要することがわかる。

3 非対称ティー誤差とダミー変数をもつ両対数モデルの当てはめ

非対称ティー誤差とダミー変数をもつ以下の両対数モデルを考える：

$$\log(\text{sales}_i) = \alpha_0 + \alpha_1 \log(\text{employees})_i + \alpha_2 \log(\text{assets})_i + \sum_{j=1}^m \delta_j D_{ij} + \log(\epsilon_i), \quad \log(\epsilon_i) \stackrel{\text{i.i.d.}}{\sim} \text{ST}(0, \omega^2, \alpha, \nu) \quad (15)$$

このモデルにける回帰係数の推定結果を表11に与える。すべての回帰係数に対する検定結果が5%有意という結果ではないが、ほとんどの回帰係数は有意であることがわかる。修正標本回帰平面（群）は、

$$\begin{aligned} \hat{\eta}_{\text{DLST},j} &= (\hat{\alpha}_0 + \hat{\omega} b_{j+1} \hat{\delta} + \hat{\delta}_j) + \hat{\alpha}_1 \log(\text{employees}) + \hat{\alpha}_2 \log(\text{assets}) \\ &= (1.294 + \hat{\delta}_j) + 0.293 \log(\text{employees}) + 0.71 \log(\text{assets}), \\ j &= 1, \dots, 33 \end{aligned} \quad (16)$$

で表される（図28参照）。

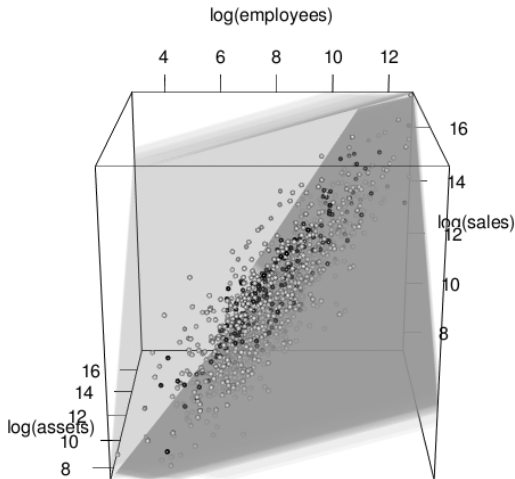


図28：2020年3月期決算の企業の財務データの3次元散佈図（対数スケール）と修正標本回帰平面群（両対数モデル：非対称ティー誤差，影響力のあるデータ除去，ダミー変数含む）

表11：ゼット比検定表：影響力のあるデータを除去後、非対称正規誤差と業種コードに対応したダミー変数をもつ両対数モデルの場合

	estimate	std.err	z-ratio	Pr {> z }
(Intercept.DP)	1.3994	0.0992	14.11	0.0000
log(employees)	0.2929	0.0157	18.65	0.0000
log(assets)	0.7103	0.0151	47.07	0.0000
sector203	-0.6386	0.0730	-8.75	0.0000
sector205	-0.4065	0.0874	-4.65	0.0000
sector207	-0.4129	0.0508	-8.13	0.0000
sector209	-0.6632	0.0680	-9.76	0.0000
sector211	0.4766	0.1799	2.65	0.0081
sector213	-0.5904	0.1034	-5.71	0.0000
sector215	-0.4689	0.0695	-6.74	0.0000
sector217	-0.3939	0.0648	-6.08	0.0000
sector219	-0.3465	0.0607	-5.71	0.0000
sector221	-0.5556	0.0509	-10.92	0.0000
sector223	-0.5311	0.0518	-10.25	0.0000
sector225	-0.3804	0.1277	-2.98	0.0029
sector227	-0.2960	0.0603	-4.91	0.0000
sector229	-0.3521	0.1031	-3.41	0.0006
sector231	-0.6508	0.0679	-9.58	0.0000
sector233	-0.4527	0.0652	-6.94	0.0000
sector235	0.1465	0.1691	0.87	0.3865
sector237	-0.1853	0.1557	-1.19	0.2338
sector241	-0.0521	0.0527	-0.99	0.3226
sector243	0.4256	0.0538	7.91	0.0000
sector245	0.2249	0.0644	3.49	0.0005
sector252	-1.1080	0.0923	-12.00	0.0000
sector253	-0.3686	0.0929	-3.97	0.0001
sector255	-0.9689	0.0715	-13.55	0.0000
sector257	-0.3749	0.0768	-4.88	0.0000
sector259	-0.4068	0.1213	-3.35	0.0008
sector261	-0.5036	0.1396	-3.61	0.0003
sector263	-0.3494	0.0826	-4.23	0.0000
sector265	-0.4010	0.0887	-4.52	0.0000
sector267	-0.6023	0.0911	-6.61	0.0000
sector269	-0.2936	0.1146	-2.56	0.0104
sector271	-0.3051	0.0509	-6.00	0.0000
omega	0.2635	0.0138	19.08	0.0000
alpha	-0.4575	0.1758	-2.60	0.0093
nu	3.7073	0.4264	8.69	0.0000

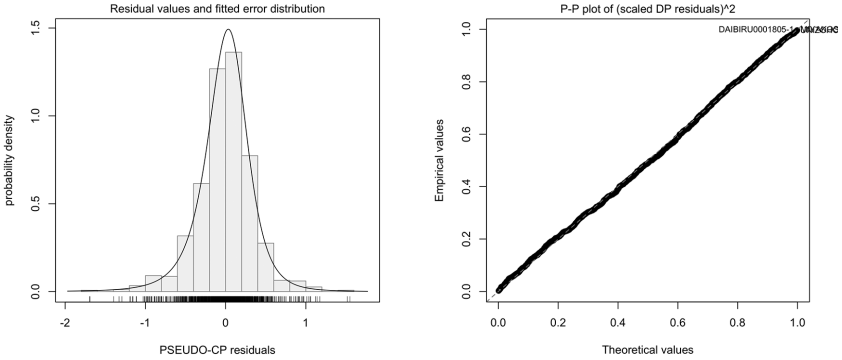


図29：2020年3月期決算の東証一部上場企業に関する財務データ（影響力があるデータ除去後）へ非対称ティー誤差とダミー変数をもつ両対数モデルを当てはめた際の回帰診断に関する各種のプロット：疑似中心化母数残差のヒストグラムと統計モデル（左）、尺度調整された直接母数残差の2乗のP-Pプロット（右）

回帰診断のプロット（図29）から、当てはまりに問題は見当たらない。なお、ヒストグラムの描画に利用される疑似中心化母数残差（pseudo-CP residual）については、Arellano-Valle and Azzalini (2013), Jimichi *et al.* (2018) を参照されたい。

4 ダミー変数を持つ両対数モデルに関する選択

影響力のあるデータを除去後に、正規誤差（log.lm.x20200301.otl.dum）、非対称正規誤差（log.selm.x20200301.otl.dum）、非対称ティー誤差（log.selm.ST.x20200301.otl.dum）をそれぞれ仮定した場合のダミー変数を持つ両対数モデルを当てはめたときの、それぞれのモデルに対する母数の次元とAICの値を表12に与える。

表12: AIC表: ダミー変数を持つ両対数モデルに関する比較

	dim	AIC
log.lm.x20200301.otl.dum	36	1025.47
log.selm.x20200301.otl.dum	37	1018.52
log.selm.ST.x20200301.otl.dum	38	908.02

この結果から、非対称ティー誤差を仮定したダミー変数を持つ両対数モデル (log.selm.ST.x20200301.otl.dum) が最も良いことがわかった。

VI 経年変化にともなうダミー変数を持つ両対数モデルの安定性

これまでの考察によって、ダミー変数をもつ両対数モデルが従業員数と資産合計で売上高を説明するために有用なものであることが分かった。ここでは、この結果が経年変化に対しても安定しているかどうかを検証する。まずは、ベンチマークとして、正規誤差を仮定したダミー変数を持つ両対数モデル (11) を経時観測データに拡張した以下のモデルを考える：

$$\log(\text{sales}_{it}) = \alpha_{0t} + \alpha_{1t} \log(\text{employees}_{it}) + \alpha_{2t} \log(\text{assets}_{it}) + \sum_{j=1}^m \delta_{jt} D_{ijt} + \log(\epsilon_{it}) \quad (17)$$

ここで、 $i=1, \dots, n_t, j=1, \dots, m_t, t=1, \dots, T$ であり、

$$D_{ijt} := \begin{cases} 1, & \text{決算年月日 } t \text{ において企業 } i \text{ が } j \text{ 番目の業種に属するとき,} \\ 0, & \text{決算年月日 } t \text{ において企業 } i \text{ が } j \text{ 番目の業種に属さないとき} \end{cases}$$

とする²³⁾。なお、推定の一意性のため $\delta_{it}=0$ とする。このモデルにおいて、決算年月日を固定 ($t=\tau$ と書く) する毎に決まるクロスセクションデータに対して、正規誤差 $\log(\epsilon_{it}) \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma_\tau^2)$ を仮定したダミー変数をもつ両対数モデル (11) を当てはめた結果として得られる決定係数等の経年変化をプロットしたものを図30に与える²⁴⁾。

23) 決算年月日 t に依存して企業と業種も変化するため、 i, j_t と書く必要があるかもしれないが、ここでは記号の簡略化をおこなった。

24) 本稿では、時点を固定するたびに決まるクロスセクションデータに対するモデリングを考えており、時間的な推移に伴う相関構造等をもつ経時観測データ (パネルデータ) に関するモデリングは別の機会に議論する予定である。

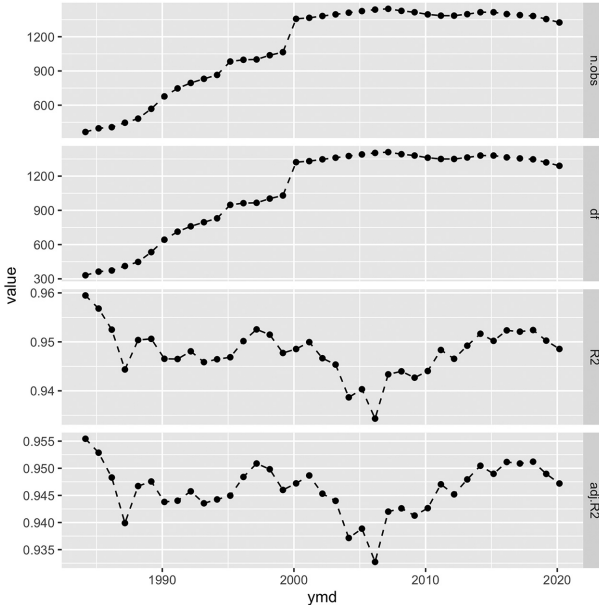


図30：1984年から2020年までの3月期決算の企業に関するクロスセクションデータヘダミー変数を持つ両対数モデル（正規誤差，影響力のあるデータ除去後）を当てはめた結果：各時点の企業数（ $n.obs$ ），自由度（ df ），決定係数（ R^2 ），自由度調整済み決定係数（ $adj.R^2$ ）の経年変化のプロット

このプロット（図30）から，最低でも約93%の決定率は確保されており，モデルの安定的な当てはまりを補償する結果となっていることに注意しよう。よって，両対数モデル（11）は東証一部上場企業の売上高を従業員数と資産合計で説明するために時間的な推移を考慮しても，ある程度妥当なものであることがわかった。

しかしながら，前節までの考察によって，両対数モデルに正規誤差を仮定することには問題があったことから，両対数モデル（17）の誤差分布として非対称正規誤差と非対称ティー誤差をもつものも検討する。その際，これらの誤差分布を仮定した場合は，最尤法によって推定が行われることから，モデル評価のためにAICを利用する。

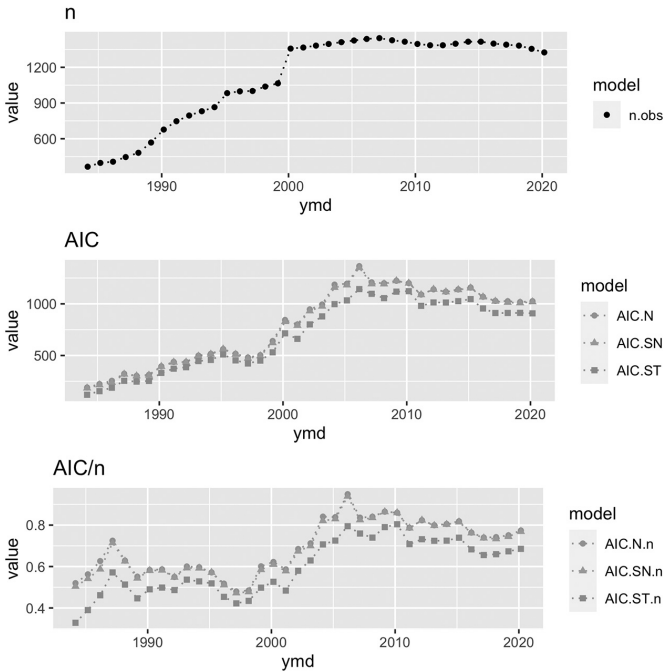


図31：1984年から2020年までの3月期決算の企業に関するクロスセクションデータヘダミー変数を持つ両対数モデル（影響力のあるデータ除去後）を当てはめた結果：各時点の企業数（ n ）、AICの経年変化（AIC）、1社あたりのAICの経年変化（AIC/ n ）のプロット

図31における企業数の経年変化（上段）より、企業数は2000年3月期に1300社を超えてからほぼ横ばいであることがわかる。また、AICの経年変化（中段）から企業数の増加とともにAICの値は増加傾向にあるが、ダミー変数と非対称ティー誤差を持つ両対数モデルが全期間を通じて良い（AICが小さい）ことがわかる。さらに、1社あたりのAICの値の経年変化（下段）からも、ダミー変数と非対称ティー誤差を持つ両対数モデルが全期間を通じて良いことがわかる。以上の結果から、ダミー変数を持つ両対数モデルの中で誤差分布としては、非対称ティー分布が経年変化を考慮しても適切であることがわかった。

VII おわりに

本稿では、地道（2021-a, b, c, d）による財務データ抽出システム SKWAD から抽出された NEEDS 企業財務データを利用して EDA を実行し、地道（2014）による結果を再検証した。時間的・空間的データ可視化の結果として得られた知見にもとづいて統計モデリングを行うことによって、企業が属する業種に対応するダミー変数をもつ両対数モデルが東証一部上場企業の売上高を従業員数と資産合計で説明するために妥当なものであることが経年的な推移を考慮しても正当化されることがわかった。その際、赤池情報量規準による検証することによって、誤差分布としては非対称ティータ分布を考慮したものが最も良い結果を与えることもわかった。また、地道（2014）でも指摘されているが経済学で扱われるコブ・ダグラス型生産関数 $Y=AL^{\alpha}K^{\beta}$ （ただし、 Y : 生産量, L : 労働, K : 資本）では、 $\alpha+\beta=1$ が成り立つ場合、生産技術が規模に関して収穫一定であることを表すが、本稿において両対数モデルを当てはめた結果として得られた各種の推定値に関しても、この関係が近似的に成り立っていることが再確認された。

結びとして、本稿では扱うことができなかった事項について以下に与える：

- (P1) 本稿において扱ったモデリングにおいて、乗法モデル (3) の両辺の対数をとる両対数モデルの枠組みで分析を行ったが、応答変数の粗データを直接予測するようなモデリングを行うことも重要なテーマであろう。この観点から、乗法モデル (3) を直接使った推測を行う方法 (cf. Bradu and Mundlak, 1970) や、両対数モデルに対する推定値を使ってある種の変換によって補正する方法 (cf. 和田, 2012) 等を考察する必要がある。
- (P2) 本稿の統計モデリングのアプローチは、時点を固定することによって母集団を固定した立場であるクロスセクションの視点からのものである。当然、時間・空間の両面からのモデリング（時系列モデル、ラン

ダム係数モデル, 混合効果モデルなど含む) も検討する必要がある。以上の問題については今後の課題としたい。

参考文献

- [1] Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle, *Proceedings of the 2nd International Symposium on Information Theory*, Petrov, B. N., and Csaki, F. (eds.), Akadimiai Kiado, Budapest: pp. 267-281.
- [2] Arellano-Valle, R. B. and A. Azzalini (2013) The centred parameterization and related quantities of the skew-t distribution. *Journal of Multivariate Analysis*, Vol. 113, pp. 73-90.
- [3] Azzalini, A. (1985) A class of distributions which includes the normal ones, *Scandinavian Journal of Statistics*, Vol. 12, No. 2, pp. 171-178.
- [4] Azzalini, A. with the collaboration of A. Capitanio (2014) *The Skew-Normal and Related Families*, Cambridge University Press, Institute of Mathematical Statistics Monographs.
- [5] Bradu, D., and Y. Mundlak (1970) Estimation in Lognormal Linear Models, *Journal of the American Statistical Association*, Vol. 65, No. 329, pp. 198-211.
- [6] Bruce, P. A. Bruce, and P. Gedeck (2020) *Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python*, O'Reilly Media, Inc. (黒川利明訳 (2020) 『データサイエンスのための統計学入門第2版: 予測, 分類, 統計モデリング, 統計的機械学習とR/Pythonプログラミング』, オライリー・ジャパン.)
- [7] Chambers, J. M., and T. J. Hastie (Editor) (1991) *Statistical Models in S*, Chapman and Hall/CRC. (柴田里程訳 (1994) 『Sと統計モデル: データ科学の新しい波』, 共立出版.)
- [8] Chatterjee, S. and Hadi (1988) *Sensitivity Analysis in Linear Regression*, John Wiley & Sons, Inc.
- [9] Chen, C., W. Härdle, and A. Unwin (editors) (2008) *Handbook of Data Visualization*, Springer.
- [10] Cobb, C. W. and P. H. Douglas (1928) A Theory of Production, *American Economic Review*, Vol. 18, pp. 139-165.
- [11] Cook, R. D. (1998) *Regression Graphics: Ideas for Studying Regressions through Graphics*, John Wiley & Sons, Inc.
- [12] Crow, E. L. and K. Shimizu (editors) (1988) *Lognormal Distributions: Theory and Applications*, Marcel Dekker.
- [13] Fox, J. (2015) *Applied Regression Analysis and Generalized Linear Models, Third Edition*, SAGE Publishing.
- [14] Fox, J. and S. Weisbrerg (2019) *An R Companion to Applied Regression, Third Edition*, SAGE Publishing.
- [15] Gandrud, C. (2020) *Reproducible Research with R and RStudio, Third Edition*, CRC

Press.

- [16] Healy, K. (2018) *Data Visualization: A Practical Introduction*, Princeton University Press. (瓜生真也, 江口哲史, 三村喬生共訳 (2021) 『実践 Data Science シリーズ: データ分析のためのデータ可視化入門』, 講談社.)
- [17] 稲垣宣生 (2003) 『数理統計学 (改訂版)』, 裳華房.
- [18] 石塚博司, 河 榮徳 (1987) 『連結財務諸表の情報効果』, 早稲田商学, 第323号, pp. 1-19.
- [19] 地道正行 (2010-a) 『日経 NEEDS 財務データにもとづくデータベースサーバの構築』, 商学論究, 第57巻, 第4号, pp. 23-80, 関西学院大学商学研究会.
- [20] 地道正行 (2010-b) 『財務データベースサーバの構築』, 関西学院大学レポジトリ, <http://hdl.handle.net/10236/6013>, ISBN: 9784990553005.
- [21] 地道正行 (2014) 『Rを利用した財務データの可視化と統計モデリング—探索的データ解析の視点から—』, 商学論究, 第61巻, 第3号, pp. 241-295, 関西学院大学商学研究会.
- [22] Jimichi, M. (2016) *Shrinkage Regression Estimators and Their Feasibilities*, Kwansei Gakuin University Press.
- [23] 地道正行 (2017-a) 『Rによる対数非対称正規線形モデルによる財務データの統計モデリング』, 商学論究, 第64巻, 第5号, pp. 159-185, 関西学院大学商学研究会.
- [24] 地道正行 (2017-b) 『Rを利用した非対称分布族にもとづく財務データの統計モデリング』, 経済学論究, 第71巻, 第2号, pp. 141-174, 関西学院大学経済学部研究会.
- [25] 地道正行 (2018-a) 『探索的財務ビッグデータ解析—前処理, データラングリング, 再現可能性—』, 商学論究, 第66巻, 第1号, pp. 1-32, 関西学院大学商学研究会.
- [26] 地道正行 (2018-b) 『探索的財務ビッグデータ解析—データ可視化, 統計モデリング, モデル選択, モデル評価, 動的文書生成, 再現可能研究—』, 商学論究, 第66巻, 第2号, pp. 1-41, 関西学院大学商学研究会.
- [27] 地道正行 (2018-c) 『データサイエンスの基礎: Rによる統計学独習』, 裳華房.
- [28] 地道正行 (2021-a) 『財務データ抽出システムの再構築—NEEDS企業財務データを中心に—』, 商学論究, 第68巻, 第3号, pp. 1-78, 関西学院大学商学研究会.
- [29] 地道正行 (2021-b) 『財務データ抽出システムの再構築—Osiris データの利用—』, 商学論究, 第69巻, 第1号, pp. 71-109, 関西学院大学商学研究会.
- [30] 地道正行 (2021-c) 『SKWAD ユーザマニュアル—NEEDS企業財務データの抽出—』, Ver. 1.0, pp. 1-88, 関西学院大学リポジトリ, <http://hdl.handle.net/10236/00029654>
- [31] 地道正行 (2021-d) 『財務データ抽出システムの再構築—Orbis データの利用—』, 商学論究, 第69巻, 第2号, pp. 65-109, 関西学院大学商学研究会.
- [32] Jimichi, M., D. Miyamoto, C. Saka, and S. Nagata (2018) Visualization and statistical modeling of financial big data: Double-log modeling with skew-symmetric error distributions, *Japanese Journal of Statistics and Data Science*, Vol. 1, No. 2, pp. 347-371, <https://>

doi.org/10.1007/s42081-018-0019-1

- [33] 地道正行, 豊原法彦 (2018) 『景気先行指数の動的文書生成にもとづく再現可能研究』, 関西学院大学産研叢書, 関西経済の構造分析, 第5章, pp. 77-111, 中央経済社.
- [34] 地道正行, 阪智香 (2021) 『財務データとESGレーティングデータによる株式時価総額の統計モデリング』, 商学論究, 第69巻, 第2号, pp. 1-64, 関西学院大学商学研究会.
- [35] Kabacoff, R. I. (2015) *R in Action: Data Analysis and Graphics with R, Second Edition*, Manning Publications Company.
- [36] Kirk, A. (2019) *Data Visualisation: A Handbook for Data Driven Design*, Second Edition, SAGE Publishing.
- [37] Klein, L. R. (1953) *A Textbook of Econometrics*, Row Peterson and Company.
- [38] Klein, L. R. (1962) *An Introduction to Econometrics*, Prentice Hall.
- [39] Konishi, S. and G. Kitagawa (2007) *Information Criteria and Statistical Modeling*, Springer.
- [40] Leisch, F. (2002) *Sweave: Dynamic generation of statistical reports using literate data analysis*, In Wolfgang Härdle and Bernd Rönz, editors, *Compstat 2002 - Proceedings in Computational Statistics*, pp. 575-580. Physica Verlag, Heidelberg. ISBN 3-7908-1517-9.
- [41] Mazza, R. (2009) *Introduction to Information Visualization*, Springer Verlag. (中本浩訳, (2011) 『情報を見える形にする技術』, ボーンデジタル.)
- [42] Mecklenburg, R. (2005) *Managing Projects with GNU Make, Third Edition*, O'Reilly Media, Inc.
- [43] Mosteller, F. and Tukey, J. W. (1977) *Data Analysis and Regression: A Second Course in Statistics*, Addison-Wesley, Reading Mass.
- [44] Peng, R. D. (2011) Reproducible research in computational science, *Science*, Vol. 334, pp. 1226-1227.
- [45] Rao, C. R. (1973) *Linear Statistical Inference and Its Applications, Second Edition*, John Wiley & Sons, Inc.
- [46] 柴田里程 (2016) 『データ分析とデータサイエンス』, 近代科学社.
- [47] Stuart, A. and J. K. Ord (1991) *Kendall's Advanced Theory of Statistics, Fifth Edition, Volume 2, Classical Inference and Relationship*, Edward Arnold.
- [48] Taddy, M. (2019) *Business Data Science: Combining Machine Learning and Economics to Optimize, Automate, and Accelerate Business Decisions*, McGraw-Hill. (上杉隼人, 井上毅郎共訳 (2020) 『ビジネスデータサイエンスの教科書』, すばる舎.)
- [49] 高橋康介 (2014) 『シリーズ Useful R 9: ドキュメント・プレゼンテーション生成』, 共立出版.
- [50] 高橋康介 (2018) 『Wonderful R 3: 再現可能性のすゝめ: RStudioによるデータ解析とレポート作成』, 共立出版.
- [51] Tufte, E. R. (2001) *The Visual Display of Quantitative Information*, Graphics Press,

Cheshire, Connecticut.

- [52] Tukey, J. W. (1977) *Exploratory Data Analysis*, Addison-Wesley Publishing Co.
- [53] Unwin, A. (2015). *Graphical Data Analysis with R*, Chapman and Hall/CRC.
- [54] 和田かず美 (2012) 『多変量外れ値の検出—繰返し加重最小二乗 (IRLS) 法による欠測値の補定方法—』, 統計研究彙報, 第69号, pp. 23-52.
- [55] Ware, C. (2012) *Information Visualization: Preception for Design*, Morgan Kaufmann.
- [56] Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis, Second Edition*, Springer. (石田基広, 石田和枝共訳 (2011) 『グラフィックスのための R プログラミング : ggplot2 入門』, シュプリンガー・ジャパン株式会社.)
- [57] Wickham, H. and G. Grolemund (2016) *R for Data Science*, O'Reilly. (黒川利明訳, (2017) 『R ではじめるデータサイエンス』, オライリー・ジャパン.)
- [58] Wilkinson, L. (2005) *The Grammar of Graphics, Second Edition*, Springer.
- [59] Xie, Y. (2015) *Dynamic Documents with R and knitr, Second Edition*, CRC Press.


謝辞

関西学院大学商学部の阪智香教授からは、本稿のドラフトに対して会計学の観点から重要なコメントをいただいた。また、総務省統計研究研修所の和田かず美氏からは、両対数モデルの推定値を使って乗法モデルの回帰曲面を推定する際の補正に関して貴重な示唆をいただいた。さらに、筆者が、2003年から2004年に在外研究で訪れたオークランド大学で、`Sweave` と `make` コマンドを利用した文書管理の手法を Ross Ihaka 氏からご教示いただいた。ここに感謝の意を表する。

なお、本研究の一部は以下の研究費より助成を得ている：

科研費 科学研究費基盤研究 C：「グラフィカル・データ・アナリシスによる格差研究と社会環境会計による解決方法の提案」(2016年～2018年), 課題番号：16K04022

科研費 科学研究費基盤研究 C：「共有価値創造 (CSV) のための社会環境会計の構築」(2019年～2021年), 課題番号：19K02006

 学際大規模情報基盤共同利用・共同研究点 (JHPCN) 課題：「財務ビッグデータの可視化と統計モデリング」(2017年度～2021年度), 課題番号：jh171002-NWJ, jh181001-NWJ, jh191002-NWJ, jh201003-NWJ, jh211001-NWJ



関西学院大学：図書館図書費 B, 研究設備費 (III), 個人研究費

付録 A コンピュータ環境

本稿の執筆に際して主に利用したコンピュータ環境の情報を与える。

ハードウェア環境

- iMac 2017:

Processor: Intel Core i7 4.2 GHz

Cores: 8

Main Memory: 64 GB

OS: macOS Big Sur (11.6)

- MacBook Pro 2018:

Processor: Intel Core i9 2.9 GHz

Cores: 6

Main Memory: 32 GB

OS: macOS Big Sur (11.6)

ソフトウェア環境

- R (R. Ihaka, R. Gentleman, R Core Team, <https://www.r-project.org/>)

- R Packages

- **car** (J. Fox, <https://socialsciences.mcmaster.ca/jfox/Books/Companion/index.html>)

- **dplyr** (H. Wickham, <http://dplyr.tidyverse.org/>)

- **GGally::ggpairs** (B. Schloerke, <http://ggobi.github.io/ggally/>)

- **ggplots2** (H. Wickham, <https://ggplot2.tidyverse.org/>)

- **magrittr** (S. M. Bache, H. Wickham, and L. Henry, <https://magrittr.tidyverse.org/>)

- **plotly** (C. Sievert, <https://plotly.com/r/>)
- **purrr** (L. Henry and H. Wickham, <https://purrr.tidyverse.org>)
- **rgl** (D. Murdoch, <https://dmurdoch.github.io/rgl/>)
- **sn** (A. Azzalini, <http://azzalini.stat.unipd.it/SN/>)
- **vroom** (J. Hester, <https://vroom.r-lib.org>)
- **xtable** (D. B. Dahl, <http://xtable.r-forge.r-project.org/>)
- **RStudio** (RStudio, <https://www.rstudio.com/>)
- **Sweave** (F. Leisch, <https://leisch.userweb.mwn.de/Sweave/>)

R 関数 `sessionInfo` を実行することによって、本稿を執筆することに利用した R に関する環境情報を以下に与える：

sessionInfo による情報

- R version 4.1.2 (2021-11-01), x86_64-apple-darwin17.0
- Locale: ja_JP.UTF-8/ja_JP.UTF-8/ja_JP.UTF-8/C/ja_JP.UTF-8/ja_JP.UTF-8
- Running under: macOS Big Sur 10.16
- Matrix products: default
- BLAS: /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRblas.0.dylib
- LAPACK: /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRlapack.dylib
- Base packages: base, datasets, graphics, grDevices, methods, stats, stats4, utils
- Other packages: car 3.0-11, carData 3.0-4, dplyr 1.0.7, forcats 0.5.1, GGally 2.1.2, ggplot2 3.3.5, gridExtra 2.3, plotly 4.10.0, purrr 0.3.4, readr 2.0.2, reshape 0.8.8, rgl 0.107.14, sn 2.0.0, stringr 1.4.0, tibble 3.1.5, tidyr 1.1.4, tidyverse 1.3.1, xtable 1.8-4
- Loaded via a namespace (and not attached): abind 1.4-5, assertthat 0.2.1, backports 1.3.0, broom 0.7.10, cellranger 1.1.0, cli 3.1.0, colorspace 2.0-2, compiler 4.1.2, crayon 1.4.2, crosstalk 1.1.1, curl 4.3.2, data.table 1.14.2, DBI 1.1.1, dbplyr 2.1.1, digest 0.6.28, ellipsis 0.3.2, fansi 0.5.0, farver 2.1.0, fastmap 1.1.0, foreign 0.8-81, fs 1.5.0, generics 0.1.1, glue 1.4.2, grid 4.1.2, gtable 0.3.0, haven 2.4.3, hms 1.1.1, htmltools 0.5.2, htmlwidgets 1.5.4, httr 1.4.2, jsonlite 1.7.2, knitr 1.36, labeling 0.4.2, lazyeval 0.2.2, lifecycle 1.0.1, lubridate 1.8.0, magrittr 2.0.1, mnormt 2.0.2, modelr 0.1.8, munsell 0.5.0, numDeriv 2016.8-1.1, openxlsx 4.2.4, pillar 1.6.4, pkgconfig 2.0.3, plyr 1.8.6, processx 3.5.2, ps 1.6.0, R6 2.5.1, RColorBrewer 1.1-2, Rcpp 1.0.7, readxl 1.3.1, reprex 2.0.1, rio 0.5.27, rlang 0.4.12, rstudioapi 0.13, rvest 1.0.2, scales 1.1.1, stringi 1.7.5, tidyselect 1.1.1, tmvnsim 1.0-2, tools 4.1.2, tzdb 0.2.0, utf8 1.2.2, vctrs 0.3.8, viridisLite 0.4.0, withr 2.4.2, xfun 0.27, xml2 1.3.2, zip 2.2.0

付録 B ディレクトリ・ファイル構成

本稿を作成するために利用したディレクトリ・ファイルの構成を図32に与える。

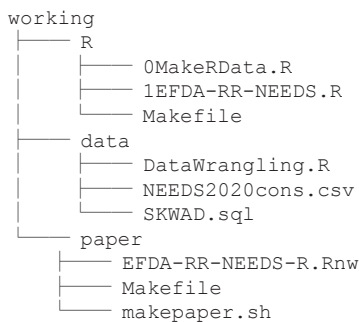


図32：本稿を作成するためのディレクトリ構成

ここで、トップディレクトリ `working` 以下のサブディレクトリには以下のファイルが納められている：

R:

`0MakeRData.R`: 探索的データ解析の結果をRの作業空間ファイル `EFDA-RR-NEEDS.RData` として保存するためのRスクリプトファイル (スクリプト4参照)

`1EFDA-RR-NEEDS.R`: Rによって探索的データ解析を実行するためのRスクリプトファイル (スクリプト5参照)

`Makefile`: Rの作業空間ファイル `EFDA-RR-NEEDS.RData` を自動生成するためのターゲット `RData` が記述されたファイル (スクリプト6参照)

data:

`SKWAD.sql`: データベースからデータを抽出するためのSQLスクリプトファイル (スクリプト1参照)

`NEEDS2020cons.csv`: 抽出されたCSVファイル (スクリプト2参照)

DataWrangling.R: Rでデータラングリングを実行するためのRスクリプトファイル (スクリプト3参照)

paper:

EFDA-RR-NEEDS-R.Rnw: 本稿のソースファイル (Sweaveファイル)

Makefile: 本稿を生成する全工程を自動処理するためのターゲットallが記述されたファイル (スクリプト6参照)

makepaper.sh: 本稿を自動生成するためのシェルスクリプトファイル (スクリプト8参照)

以降の付録において、これらのディレクトリやファイルを適宜参照・説明する。

付録C データの抽出とラングリング

ここでは、本稿で利用するデータの抽出とラングリングについて解説する。

C.1 データ抽出

本稿で扱うデータは、データ抽出システムSKWAD (地道, 2021-a, b, c, d参照)におけるNEEDS企業財務データを抽出するページからSQLスクリプトファイルSKWAD.sql (スクリプト1)を利用して抽出した。SKWADの利用とSQLスクリプト1の詳細については、地道 (2021-c)を参照されたい。

スクリプト1: SQLスクリプトファイル: SKWAD.sql

```
1 SELECT
2 f.firmname AS name,
3 b.a05 AS nikkei_firm_code,
4 b.a02 AS ym,
5 b.a35 AS sector,
6 b.a08 AS term,
7 b.a27 AS ac,
8 d.b001 AS sales,
9 h.b021 AS employees,
10 b.b001 AS assets_total
11 FROM (SELECT a02, a05, a08, a27, a29, a35, b001 FROM yb01) AS b
```

```

12 JOIN (SELECT a02, a05, a08, a27, a29, a35, b001 FROM yd01) AS d
13 ON d.a02 = b.a02 AND d.a05 = b.a05 AND d.a27 = b.a27
14 JOIN (SELECT a02, a05, a08, a27, a29, a35, b021 FROM yh01) AS h
15 ON h.a02 = b.a02 AND h.a05 = b.a05 AND h.a27 = b.a27
16 JOIN (SELECT nikkei_corp_code, firmname FROM firmlist) AS f
17 ON f.nikkei_corp_code = b.a05
18 WHERE d.a29 = '11'
19 ORDER BY b.a05, d.a02;

```

実際に、ダウンロードされたCSVファイルNEEDS2020cons.csvは、以下の様なものである：

スクリプト2：CSVファイル：NEEDS2020cons.csv（先頭10行）

```

1 name,nikkei_firm_code,ym,sector,term,ac,sales,employees,assets_total
2 KYOKUYO
3     ,0000001,198410,235341,12,1,+0000000206485,-9999999999999,+0000000093094
4 KYOKUYO
5     ,0000001,198510,235341,12,1,+0000000206512,+000000001223,+0000000082267
6 KYOKUYO
7     ,0000001,198610,235341,12,1,+0000000194353,+000000001133,+0000000082394
8 KYOKUYO
9     ,0000001,198710,235341,12,1,+0000000200304,+000000001089,+0000000085497
10 KYOKUYO
11     ,0000001,198803,235341,05,1,+0000000081843,+000000001054,+0000000082382
12 KYOKUYO
13     ,0000001,198903,235341,12,1,+0000000213409,+0000000000873,+0000000086649
14 KYOKUYO
15     ,0000001,199003,235341,12,1,+0000000207862,+0000000000855,+0000000076786
16 KYOKUYO
17     ,0000001,199103,235341,12,1,+0000000202573,+0000000000846,+0000000074061
18 KYOKUYO
19     ,0000001,199203,235341,12,1,+0000000199227,+0000000000843,+0000000068312

```

なお、これらのファイルは図32におけるdataディレクトリに格納されている。

C.2 ラングリング

上で抽出されたデータをRへ読み込み、データ解析できるオブジェクト形式へ変換する工程（データラングリング）は以下のRスクリプトファイルDataWrangling.R（スクリプト3）を実行することによって行われる。

スクリプト3：ラングリングのためのRスクリプトファイル：DataWrangling.R

```

1 # -----
2 # データラングリング
3 # -----
4 # データ読み込み
5 rawdata <- vroom::vroom("./NEEDS2020cons.csv")
6 # データ操作
7 library(dplyr)
8 library(zoo)
9 x <- rawdata %>%
10 filter(substr(ym, 5, 6) == "03", term == 12) %>%
11 mutate(employees = na_if(employees, "-999999999999999"),
12        assets_total = na_if(assets_total, "-999999999999999"),
13        sales = na_if(sales, "-999999999999999")) %>%
14 mutate(
15   name = paste0(gsub("_", "", name), nikkei_firm_code),
16   sales = as.numeric(sales),
17   employees = as.numeric(employees),
18   assets = as.numeric(assets_total),
19   sector1 = as.factor(substr(sector, 1, 1)),
20   sector2 = as.factor(substr(sector, 2, 3)),
21   sector3 = as.factor(substr(sector, 4, 6)),
22   ymd = as.Date(as.yearmon(as.character(ym), "%Y%m"))) %>%
23 filter(ymd >= "1984-03-01") %>%
24 select(name, ymd, sector1, sector2, sector3, ac, sales, employees, assets
   )

```

スクリプト3によって行われる処理は以下のようなものである：

- 5行目：CSVファイルNEEDS2020cons.csvからデータを**vroom**パッケージのvroom関数を利用して読み込み、(粗データ) rawdataオブジェクトへ付値²⁵⁾
- 6, 7行目：データマニピュレーションを行うためのパッケージ**dplyr**と時間情報を扱うための**zoo**パッケージの読み込み
- 8行目：パイプ演算子 %>% を利用して rawdata オブジェクトをパイプライン化
- 10行目：filter 関数を利用して、3月期決算のデータと決算月数が12ヶ月のものに限定
- 11~13行目：NEEDS 企業財務データでは、-999999999999999 が欠測値であることから、sales (売上高), employees (従業員数), assets_total (資産合計) におけるこの値を、mutate 関数と na_if 関数

を利用して、欠測値 NA に変換

14~22行目: `mutate` 関数を利用して、以下の列を追加・置換:

15行目: 企業名を一意化するために、`name` (企業名) を `nikkei_firm_code` (日経会社コード) と結合 (`paste0`) したもので置換

16~18行目: `sales` (売上高), `employees` (従業員数), `assets_total` (資産合計) を数値型へ変換

19~21行目: `sector` (日経業種コード) から、`substr` 関数を利用して、先頭から 1 桁 (大分類), 2~3 桁 (中分類), 4~6 桁 (小分類) を抽出し、それぞれ、`sector1`, `sector2`, `sector3` という列に追加

22行目: `zoo` パッケージの `as.yearmon` 関数を利用して、`ym` (決算年月情報) を `ymd` (決算年月日情報) に変換

23行目: 連結決算の情報開示が本格化した (持分法の全面適用が強制されるようになった²⁵⁾) 1984年3月期決算以降のデータを選択

24行目: `select` 関数でデータ解析に利用する列を選択

ここで、22行目で行った「年月」から「年月日」への変換は、「年月」という情報が R の処理で扱いづらいためである。また、ここでは3月期の決算を扱っているので31日でもよいかもしれないが、今後他の決算月を持つものを利用することを見越して、月末の日情報は月によって異なることから、全期間を通じて全て月初 (1日) とした。これはあくまでも技術的な処理である。

25) `vroom` (<https://github.com/r-lib/vroom>) は、CSV ファイルから R へデータを高速に読み書きするためのパッケージである。RStudio (<https://www.rstudio.com>) によって開発されている。

26) 石塚, 河 (1987) 参照。

付録 D R スクリプト

本稿で利用した R スクリプトを以下に与える (図32も参照)。なお、R スクリプトファイル 0MakeRData.R (スクリプト 4) は、文書の再現性を確保するために利用した R の作業空間ファイル EFDA-RR-NEEDS.RData を生成するためのものであり、R スクリプトファイル 1EFDA-RR-EDA.R (スクリプト 5) は、本稿の結果を可視化も含めて再現するためのものである。

スクリプト 4 : R の作業空間ファイル EFDA-RR-NEEDS.RData を生成するための R スクリプト: 0MakeRData.R

```

1 # -----
2 # パッケージの読み込み
3 # -----
4 library(dplyr)
5 library(zoo)
6 library(sn)
7 library(plotly)
8 # -----
9 # データラングリング
10 # -----
11 # データ読み込み
12 rawdata <- vroom::vroom("../data/NEEDS2020cons.csv")
13 # データ操作
14 x <- rawdata %>%
15   filter(substr(ym, 5, 6) == "03", term == 12) %>%
16   mutate(employees = na_if(employees, "-9999999999999999"),
17          assets_total = na_if(assets_total, "-9999999999999999"),
18          sales = na_if(sales, "-9999999999999999")) %>%
19   mutate(
20     name = paste0(gsub("_", "", name), nikkei_firm_code),
21     sales = as.numeric(sales),
22     employees = as.numeric(employees),
23     assets = as.numeric(assets_total),
24     sector1 = as.factor(substr(sector, 1, 1)),
25     sector2 = as.factor(substr(sector, 2, 3)),
26     sector3 = as.factor(substr(sector, 4, 6)),
27     ymd = as.Date(as.yearmon(as.character(ym), "%Ym"))) %>%
28     filter(ymd >= "1984-03-01") %>%
29     select(name, ymd, sector1, sector2, sector3, ac, sales, employees, assets
30            )
31 # -----
32 # 2020 年 3 月期決算の企業の財務データを抽出、欠測値の除去、行名付値
33 # -----
34 x20200301 <- x %>% filter(ymd == "2020-03-01") %>% mutate(name.ac = paste(
   name, ac, sep = "-"))
35 x20200301 <- x20200301 %>% na.omit()
```

```
35 rownames(x20200301) <- x20200301$name.ac
36 #-----
37 # plotly による可視化
38 #-----
39 gg.needs <- x %>% mutate(year = substr(ymd, 1, 4)) %>%
40   ggplot(aes(assets, sales, color = sector1)) +
41   geom_point(aes(size = employees, frame = year, ids = name, alpha = 0.5))
42 #
43 gg.needs.col <- x %>% mutate(year = substr(ymd, 1, 4)) %>%
44   ggplot(aes(assets, sales, color = sector2)) +
45   geom_point(aes(size = employees, frame = year, ids = name, alpha = 0.5))
46 #-----
47 # 2020年3月期決算の1部上場企業のクロスセクションデータ
48 # に対する正規線形モデルと両対数モデルの当てはめ
49 #-----
50 # 正規線形モデルの当てはめ
51 #-----
52 lm.x20200301 <- lm(sales ~ employees + assets, data = x20200301)
53 #-----
54 # 2020年3月期決算の企業の売上高,従業員数,資産合計の三次元散布図
55 # 標本回帰平面の当てはめ
56 #-----
57 # plotly による可視化: 対数スケール
58 #-----
59 gg.needs.log <- x %>% mutate(year = substr(ymd, 1, 4)) %>%
60   ggplot(aes(log(assets), log(sales), color = sector1)) +
61   geom_point(aes(size = employees, frame = year, ids = name, alpha = 0.5))
62 #
63 gg.needs.log.col <- x %>% mutate(year = substr(ymd, 1, 4)) %>%
64   ggplot(aes(log(assets), log(sales), color = sector2)) +
65   geom_point(aes(size = employees, frame = year, ids = name, alpha = 0.5))
66 #-----
67 # 両対数モデルの当てはめ
68 #-----
69 log.lm.x20200301 <- lm(log(sales) ~ log(employees) + log(assets), data =
70   x20200301)
71 #-----
72 # 異常値の除去とデータフレームの再生成
73 #-----
74 otl <- c("JAPANSECURITIESFINANCE0070514-1", "JACCS0001710-1", "
75   JAPANPOSTHOLDINGS0038793-1", "JAPANEXCHANGEGROUP0075107-3")
76 firms20200301 <- x20200301$name.ac
77 x20200301.otl <- x20200301 %>% filter(name.ac %in% setdiff(firms20200301,
78   otl))
79 #-----
80 # 両対数モデルの当てはめ (非対称誤差)
81 #-----
82 log.lm.x20200301.otl <- lm(log(sales) ~ log(employees) + log(assets), data
83   = x20200301.otl)
84 #
85 log.selm.x20200301.otl <- selm(log(sales) ~ log(employees) + log(assets),
86   data = x20200301.otl)
87 coef.log.selm.x20200301.otl <- coef(log.selm.x20200301.otl, param.type="DP"
88   )
```

```

83 #
84 log.selm.ST.x20200301.otl <- selm(log(sales) ~ log(employees) + log(assets
    ), family = "ST", data = x20200301.otl)
85 coef.log.selm.ST.x20200301.otl <- coef(log.selm.ST.x20200301.otl, param.
    type="DP")
86 #-----
87 # グミー変数付き両対数モデルの当てはめ (正規誤差)
88 #-----
89 log.lm.x20200301.otl.dum <- lm(log(sales) ~ log(employees) + log(assets) +
    sector2, data = x20200301.otl)
90 coef.log.lm.x20200301.otl.dum <- coef(log.lm.x20200301.otl.dum)
91 #-----
92 # グミー変数付き両対数モデルの当てはめ (非対称正規誤差)
93 #-----
94 log.selm.x20200301.otl.dum <- selm(log(sales) ~ log(employees) + log(assets
    ) + sector2, data = x20200301.otl)
95 coef.log.selm.x20200301.otl.dum <- coef(log.selm.x20200301.otl.dum, param.
    type="DP")
96 #-----
97 # グミー変数付き両対数モデルの当てはめ (非対称ティール誤差)
98 #-----
99 log.selm.ST.x20200301.otl.dum <- selm(log(sales) ~ log(employees) + log(
    assets) + sector2, family = "ST", data=x20200301.otl)
100 coef.log.selm.ST.x20200301.otl.dum <- coef(log.selm.ST.x20200301.otl.dum,
    param.type="DP")
101 #-----
102 # 決定係数の経年変化のプロット
103 #-----
104 OLS.ts <- function(obj)
105 {
106   require(dplyr)
107   require(purrr)
108   require(tidyr)
109   otl <- c("JAPANSECURITIESFINANCE0070514-1", "JACCS0001710-1", "
    JAPANPOSTHOLDINGS0038793-1", "JAPANEXCHANGEGROUP0075107-3")
110   obj <- obj %>% mutate(name.ac = paste(name, ac, sep = "-"))
111   firms <- unique(obj$name.ac)
112   x.otl <- obj %>% filter(name.ac %in% setdiff(firms, otl))
113   tp <- seq(as.Date("1984-03-01"), as.Date("2020-03-01"), by = "year")
114   x.otl <- x.otl %>% filter(ymd %in% tp)
115   n_x <- x.otl %>% group_by(ymd) %>% arrange(ymd) %>% nest()
116   mod_fun <- function(df) lm(log(sales) ~ log(employees) + log(assets) +
    sector2, data = df)
117   m_x <- n_x %>% mutate(model = map(data, mod_fun))
118   n_fun <- function(mod) length(mod$fitted.values)
119   df_fun <- function(mod) mod$df.residual
120   r_fun <- function(mod) summary(mod)$r.squared
121   adj_r_fun <- function(mod) summary(mod)$adj.r.squared
122   res <- m_x %>% transmute(ymd,
123     n.obs = map_int(model, n_fun),
124     df = map_int(model, df_fun),
125     R2 = map_dbl(model, r_fun),
126     adj.R2 = map_dbl(model, adj_r_fun))

```



```

127   data.frame(res)
128 }
129 x.OLS.ts <- OLS.ts(x)
130 plot.OLS.ts <- function(obj)
131 {
132   require(ggplot2)
133   require(reshape)
134   #require(tidyr)
135   require(dplyr)
136   #obj %>% pivot_longer(cols = ~"ymd") %>% arrange(name) %>%
137   obj %>% melt(id.vars = "ymd") %>%
138   ggplot(aes(ymd, value, group = variable)) +
139   geom_point() + geom_line(linetype = "dashed") + facet_grid(variable ~
140   ., scale="free_y")
141 }
142 #-----
143 # AICの経年変化のプロット
144 #-----
145 AIC.ts <- function(obj)
146 {
147   require(sn)
148   require(tidyverse)
149   require(lubridate)
150   #
151   otl <- c("JAPANSECURITIESFINANCE0070514-1", "JACCS0001710-1", "
152   JAPANPOSTHOLDINGS0038793-1", "JAPANEXCHANGEGROUP0075107-3")
153   obj <- obj %>% mutate(name.ac = paste(name, ac, sep = "-"))
154   firms <- unique(obj$name.ac)
155   x.otl <- obj %>% filter(name.ac %in% setdiff(firms, otl))
156   tp <- seq(as.Date("1984-03-01"), as.Date("2020-03-01"), by = "year")
157   x.otl <- x.otl %>% filter(ymd %in% tp)
158   n_x <- x.otl %>% group_by(ymd) %>% arrange(ymd) %>% nest()
159   mod_fun <- function(df) lm(log(sales) ~ log(employees) + log(assets) +
160   sector2, data = df)
161   mod_fun_SN <- function(df) selm(log(sales) ~ log(employees) + log(assets)
162   + sector2, family = "SN", data = df)
163   mod_fun_ST <- function(df) selm(log(sales) ~ log(employees) + log(assets)
164   + sector2, family = "ST", data = df)
165   lm_selm_x <- n_x %>% mutate(lmres = map(data, mod_fun),
166   selmSNres = map(data, mod_fun_SN),
167   selmSTres = map(data, mod_fun_ST))
168   n_fun <- function(lmres) length(lmres$fitted.values)
169   aic_fun <- function(lmres) AIC(lmres)
170   res <- lm_selm_x %>% transmute(ymd,
171   n.obs = map_int(lmres, n_fun),
172   AIC.N = map_dbl(lmres, aic_fun),
173   AIC.SN = map_dbl(selmSNres, aic_fun),
174   AIC.ST = map_dbl(selmSTres, aic_fun))
175   data.frame(res)
176 }
177 suppressWarnings(invisible(capture.output(x.AIC.ts <- AIC.ts(x))))
178 plot.AIC <- function(obj)
179 {
180   require(tidyverse)

```

```

176 require(gridExtra)
177 x <- obj %>% pivot_longer(-ymd, names_to = "model", values_to = "value")
178 y <- obj %>% mutate(AIC.N.n = AIC.N/n.obs, AIC.SN.n = AIC.SN/n.obs, AIC.
  ST.n = AIC.ST/n.obs) %>%
179   select(ymd, n.obs, AIC.N.n, AIC.SN.n, AIC.ST.n) %>%
180   pivot_longer(-ymd, names_to = "model", values_to = "value")
181 p1 <- x %>% filter(model == "n.obs") %>%
182   ggplot(aes(ymd, value, group = model, shape = model)) + geom_line(
  linetype="dotted") + geom_point() + ggtitle("n")
183 p2 <- x %>% filter(model != "n.obs") %>%
184   ggplot(aes(ymd, value, group = model, color = model, shape = model)) +
  geom_line(linetype="dotted") +
185   geom_point() + ggtitle("AIC")
186 p3 <- y %>% filter(model != "n.obs") %>%
187   ggplot(aes(ymd, value, group = model, color = model, shape = model)) +
  geom_line(linetype="dotted") +
188   geom_point() + ggtitle("AIC/n")
189 grid.arrange(p1, p2, p3)
190 }
191 #-----
192 # 作業空間のファイル出力
193 #-----
194 save.image("EFDA-RR-NEEDS.RData")

```

スクリプト5：本稿の結果を可視化も含めて再現するためのRスクリプト： 1EFDA-RR-NEEDS.R

```

1 #-----
2 # データラングリング
3 #-----
4 # データ読み込み
5 rawdata <- vroom::vroom("./NEEDS2020cons.csv")
6 # データ操作
7 library(dplyr)
8 library(zoo)
9 x <- rawdata %>%
10   filter(substr(ym, 5, 6) == "03", term == 12) %>%
11   mutate(employees = na_if(employees, "-999999999999999"),
  assets_total = na_if(assets_total, "-999999999999999"),
12   sales = na_if(sales, "-999999999999999")) %>%
13   mutate(
14     name = paste0(gsub("_", "", name), nikkei_firm_code),
15     sales = as.numeric(sales),
16     employees = as.numeric(employees),
17     assets = as.numeric(assets_total),
18     sector1 = as.factor(substr(sector, 1, 1)),
19     sector2 = as.factor(substr(sector, 2, 3)),
20     sector3 = as.factor(substr(sector, 4, 6)),
21     ymd = as.Date(as.yearmon(as.character(ym), "%Y%m"))) %>%
22     filter(ymd >= "1984-03-01") %>%
23     select(name, ymd, sector1, sector2, sector3, ac, sales, employees, assets
24   )

```

```
25 #-----
26 # 売上高, 従業員数, 資産合計の時系列プロットと2020年3月期決算の企業のヒスト
   グラム
27 #-----
28 library(ggplot2)
29 x %>% ggplot(aes(ymd, sales, group = name)) + geom_line(size = 0.3, alpha
   =0.5) +
30   geom_vline(xintercept = as.numeric(as.Date("2020-03-01")), lwd = 1, color
   = "red")
31 x %>% filter(ymd == "2020-03-01") %>% ggplot(aes(sales)) + geom_histogram()
   + geom_rug() + coord_flip()
32 #-----
33 x %>% ggplot(aes(ymd, employees, group = name)) + geom_line(size = 0.3,
   alpha = 0.5) +
34   geom_vline(xintercept = as.numeric(as.Date("2020-03-01")), lwd = 1, color
   = "red")
35 x %>% filter(ymd == "2020-03-01") %>% ggplot(aes(employees)) + geom_
   histogram() + geom_rug() + coord_flip()
36 #-----
37 x %>% ggplot(aes(ymd, assets, group = name)) + geom_line(size = 0.3, alpha
   =0.5) +
38   geom_vline(xintercept = as.numeric(as.Date("2020-03-01")), lwd = 1, color
   = "red")
39 x %>% filter(ymd == "2020-03-01") %>% ggplot(aes(assets)) + geom_histogram
   () + geom_rug() + coord_flip()
40 #-----
41 # 2020年3月期決算の企業の財務データを抽出, 欠測値の除去, 行名付値
42 #-----
43 x20200301 <- x %>% filter(ymd == "2020-03-01") %>% mutate(name.ac = paste(
   name, ac, sep = "-"))
44 x20200301 <- x20200301 %>% na.omit()
45 rownames(x20200301) <- x20200301$name.ac
46 #-----
47 # 2020年3月期決算の企業の売上高の分布の可視化
48 #-----
49 x20200301 %>%
50   ggplot(aes(x = sales)) + geom_histogram(aes(y = ..density..), fill = "
   white", color = "black")
51 x20200301 %>% ggplot(aes(x = log(sales))) +
52   geom_histogram(aes(y = ..density..), binwidth = 1, fill = "white", color
   = "black")
53 #-----
54 # 2020年3月期決算の企業の売上高, 従業員数, 資産合計の対散布図 (通常スケ
   ル)
55 #-----
56 library(GGally)
57 x20200301 %>% select(sales, employees, assets) %>%
58   ggpairs(
59     upper = list(continuous = wrap("points", size = 0.5, alpha = 0.5)),
60     lower = list(continuous = wrap("cor", size = 3))
61   ) +
62   theme(
63     axis.text = element_text(size = 5),
64     axis.title = element_text(size = 3)
```

```

65 | )
66 | #-----
67 | # 2020年3月期決算の企業の売上高, 従業員数, 資産合計の3次元散布図 (通常スケール)
68 | #-----
69 | library(rgl, pos=4)
70 | library(mgcv, pos=4)
71 | #install.packages("magick")
72 | #library(magick)
73 | #-----
74 | plot3d(x20200301[, c("employees", "assets", "sales")], type = "s", col = "red", size = 1)
75 | #movie3d(spin3d(axis = c(0, 0, 1), rpm = 10), movie = "sp3d", dir = "./sp3d", duration=6, type = "gif")
76 | #-----
77 | # plotlyによる可視化: 通常スケール
78 | #-----
79 | library(plotly)
80 | gg.needs <- x %>% mutate(year = substr(ymd, 1, 4)) %>%
81 |   ggplot(aes(assets, sales, color = sector1)) +
82 |   geom_point(aes(size = employees, frame = year, ids = name, alpha = 0.5))
83 | ggplotly(gg.needs)
84 | #
85 | gg.needs.col <- x %>% mutate(year = substr(ymd, 1, 4)) %>%
86 |   ggplot(aes(assets, sales, color = sector2)) +
87 |   geom_point(aes(size = employees, frame = year, ids = name, alpha = 0.5))
88 | ggplotly(gg.needs.col)
89 | #-----
90 | # 2020年3月期決算の1部上場企業のクロスセクションデータに対する正規線形モデルと両対数モデルの当てはめ
91 | #-----
92 | # 正規線形モデルの当てはめ
93 | #-----
94 | lm.x20200301 <- lm(sales ~ employees + assets, data = x20200301)
95 | lm.x20200301 %>% summary()
96 | #-----
97 | # 2020年3月期決算の企業の売上高, 従業員数, 資産合計の3次元散布図と標本回帰平面(正規線形モデル)
98 | #-----
99 | library(rgl, pos=4)
100 | library(mgcv, pos=4)
101 | #library(magick)
102 | #-----
103 | plot3d(x20200301[, c("employees", "assets", "sales")], type = "s", col = "red", size = 1)
104 | planes3d(coef(lm.x20200301)[2],
105 |         coef(lm.x20200301)[3],
106 |         -1,
107 |         coef(lm.x20200301)[1] ,
108 |         alpha=0.5)
109 | #movie3d(spin3d(axis = c(0, 0, 1), rpm = 10), movie = "sp3d-plane", dir = "./sp3d", duration=6, type = "gif")
110 | #-----
111 | # 回帰診断

```

```

112 # -----
113 par(mfcol=c(2,2))
114 plot(resid(lm.x20200301),ylab="Residuals")
115 mtext("Index_Plot_of_Residuals", 3, 0.25, cex = 1)
116 plot(lm.x20200301,which=c(1,2))
117 plot(density(resid(lm.x20200301)), main="")
118 mtext("Density_Plot_of_Residuals", 3, 0.25, cex = 1)
119 par(mfcol=c(1,1))
120 # -----
121 # 売上高, 従業員数, 資産合計の時系列プロットと2020年3月期決算の企業のヒストグラム (対数スケール)
122 # -----
123 x %>% ggplot(aes(ymd, log(sales), group = name)) + geom_line(size = 0.3,
124   alpha = 0.5) +
125   geom_vline(xintercept = as.numeric(as.Date("2020-03-01")), lwd = 1 ,color
126     = "red")
127 x %>% filter(ymd == "2020-03-01") %>% ggplot(aes(log(sales))) + geom_
128   histogram() + geom_rug() + coord_flip()
129 # -----
130 x %>% ggplot(aes(ymd, log(employees), group = name)) + geom_line(size =
131   0.3, alpha =0.5) +
132   geom_vline(xintercept = as.numeric(as.Date("2020-03-01")), lwd = 1, color
133     = "red")
134 x %>% filter(ymd == "2020-03-01") %>% ggplot(aes(log(employees))) + geom_
135   histogram() + geom_rug() + coord_flip()
136 # -----
137 x %>% ggplot(aes(ymd, log(assets), group = name)) + geom_line(size = 0.3,
138   alpha =0.5) +
139   geom_vline(xintercept = as.numeric(as.Date("2020-03-01")), lwd = 1, color
140     = "red")
141 x %>% filter(ymd == "2020-03-01") %>% ggplot(aes(log(assets))) + geom_
142   histogram() + geom_rug() + coord_flip()
143 # -----
144 # 2020年3月期決算の企業の売上高, 従業員数, 資産合計の対散佈図 (対数スケール)
145 # -----
146 library(GGally)
147 x20200301 %>% mutate(log.sales = log(sales), log.employees = log(employees
148   ), log.assets = log(assets)) %>%
149   select(log.sales, log.employees, log.assets) %>%
150   ggpairs(
151     diag = list(continuous = wrap("densityDiag", alpha=0.5)),
152     upper = list(continuous = wrap("points", size = 0.5, alpha = 0.5)),
153     lower = list(continuous = wrap("cor", size = 3),
154       combo = wrap("facethist", binwidth = 1))
155   ) +
156   theme(
157     axis.text= element_text(size = 5),
158     axis.title = element_text(size = 3)
159   )
160 # -----
161 # 2020年3月期決算の企業の売上高, 従業員数, 資産合計の3次元散佈図 (対数スケール)
162 # -----

```

```

153 library(rgl, pos=4)
154 library(mgcv, pos=4)
155 #library(magick)
156 #-----
157 plot3d(log(x20200301[, c("assets","employees","sales")])), type = "s", col =
    "red", size = 1,
158       xlab = "log(employees)", ylab = "log(assets)", zlab = "log(sales)")
159 #movie3d(spin3d(axis = c(0,0,1), rpm = 10), movie = "sp3d-log", dir = "./
    sp3d", duration=6, type = "gif")
160 #-----
161 # plotly による可視化: 対数スケール
162 #-----
163 library(plotly)
164 gg.needs.log <- x %>% mutate(year = substr(ymd, 1, 4)) %>%
165   ggplot(aes(log(assets), log(sales), color = sector1)) +
166   geom_point(aes(size = employees, frame = year, ids = name, alpha = 0.5))
167 ggplotly(gg.needs.log)
168 #
169 gg.needs.log.col <- x %>% mutate(year = substr(ymd, 1, 4)) %>%
170   ggplot(aes(log(assets), log(sales), color = sector2)) +
171   geom_point(aes(size = employees, frame = year, ids = name, alpha = 0.5))
172 ggplotly(gg.needs.log.col)
173 #-----
174 #両対数モデルの当てはめ
175 #-----
176 log.lm.x20200301 <- lm(log(sales) ~ log(employees) + log(assets), data =
    x20200301)
177 log.lm.x20200301 %>% summary()
178 #-----
179 # 2020年3月期決算の企業の売上高、従業員数、資産合計の3次元散布図(対数スケール)と標本回帰平面(両対数モデル: 正規誤差)
180 #-----
181 library(rgl, pos=4)
182 library(mgcv, pos=4)
183 #library(magick)
184 #-----
185 plot3d(log(x20200301[, c("employees","assets","sales")])), type = "s", col
    = "red", size = 1,
186       xlab = "log(employees)", ylab = "log(assets)", zlab = "log(sales)")
187 planes3d(coef(log.lm.x20200301)[2],
188         coef(log.lm.x20200301)[3],
189         -1,
190         coef(log.lm.x20200301)[1],
191         alpha=0.5)
192 #movie3d(spin3d(axis = c(0,0,1), rpm = 10), movie = "sp3d-log-plane", dir =
    "./tmp", duration=6, type = "gif")
193 #-----
194 # 回帰診断
195 #-----
196 par(mfcol = c(2, 2))
197 plot(resid(log.lm.x20200301), ylab="Residuals")
198 mtext("Index_Plot_of_Residuals", 3, 0.25, cex = 1)
199 plot(log.lm.x20200301, which=c(1,2))
200 plot(density(resid(log.lm.x20200301)), main="")

```

```

201 mtext("Density_Plot_of_Residuals", 3, 0.25, cex = 1)
202 par(mfcol = c(1, 1))
203 #
204 library(car)
205 influenceIndexPlot(log.lm.x20200301, vars=c("hat","Studentized","Cook")) #
      id.n=4, id.cex = 0.4)
206 influencePlot(log.lm.x20200301, id = list(n = 3))
207 #-----
208 # 異常値の除去とデータフレームの再生成
209 #-----
210 otl <- c("JAPANSECURITIESFINANCE0070514-1", "JACCS0001710-1", "
      JAPANPOSTHOLDINGS0038793-1", "JAPANEXCHANGEGROUP0075107-3")
211 x20200301 %>% filter(name.ac %in% otl) %>% select(name.ac, sales, employees
      ,assets)
212 firms20200301 <- x20200301$name.ac
213 x20200301.otl <- x20200301 %>% filter(name.ac %in% setdiff(firms20200301,
      otl))
214 #-----
215 # 両対数モデルの再当てはめ
216 #-----
217 log.lm.x20200301.otl <- lm(log(sales) ~ log(employees) + log(assets), data
      = x20200301.otl)
218 log.lm.x20200301.otl %>% summary()
219 #-----
220 # 回帰診断
221 #-----
222 par(mfcol = c(2, 2))
223 plot(resid(log.lm.x20200301.otl), ylab = "Residuals")
224 mtext("Index_Plot_of_Residuals", 3, 0.25, cex = 1)
225 plot(log.lm.x20200301.otl, which = c(1, 2))
226 plot(density(resid(log.lm.x20200301.otl)), main = "")
227 mtext("Density_Plot_of_Residuals", 3, 0.25, cex = 1)
228 par(mfcol=c(1,1))
229 influenceIndexPlot(log.lm.x20200301.otl, vars=c("hat","Studentized","Cook"
      )) #id.n=4, id.cex = 0.4)
230 influencePlot(log.lm.x20200301.otl, id = list(n = 3))
231 #-----
232 # 2020年3月期決算の企業の売上高、従業員数、資産合計の3次元散布図(対数スケール)と標本回帰平面(正規誤差、異常値の除去後)
233 #-----
234 library(rgl, pos=4)
235 library(mgcv, pos=4)
236 #library(magick)
237 #-----
238 plot3d(log(x20200301.otl[, c("employees","assets", "sales")])), type = "s",
      col = "red", size = 1,
239       xlab = "log(employees)", ylab = "log(assets)", zlab = "log(sales)")
240 planes3d(coef(log.lm.x20200301.otl)[2],
241          coef(log.lm.x20200301.otl)[3],
242          -1,
243          coef(log.lm.x20200301.otl)[1],
244          alpha=0.5)
245 #movie3d(spin3d(axis = c(0,0,1), rpm = 10), movie = "sp3d-log-plane-adj",
      dir = "./tmp", duration=6, type = "gif")

```

```

246 #-----
247 # 両対数モデルの当てはめ (非対称誤差)
248 #-----
249 library(sn)
250 log.selm.x20200301.otl <- selm(log(sales) ~ log(employees) + log(assets),
    data = x20200301.otl)
251 coef.log.selm.x20200301.otl <- coef(log.selm.x20200301.otl, param.type="DP"
    )
252 log.selm.x20200301.otl %>% summary(param.type = "DP")
253 par(mfcol = c(1, 2))
254 for(i in c(2, 4)) plot(log.selm.x20200301.otl, param.type = "CP", which = i
    )
255 par(mfcol = c(1, 1))
256 #
257 log.selm.ST.x20200301.otl <- selm(log(sales) ~ log(employees) + log(assets
    ), family = "ST", data = x20200301.otl)
258 coef.log.selm.ST.x20200301.otl <- coef(log.selm.ST.x20200301.otl, param.
    type="DP")
259 log.selm.ST.x20200301.otl %>% summary(param.type = "DP")
260 par(mfcol = c(1, 2))
261 for(i in c(2, 4)) plot(log.selm.ST.x20200301.otl, param.type = "CP", which
    = i)
262 par(mfcol = c(1, 1))
263 #-----
264 # 2020年3月期決算企業の売上高, 従業員数, 資産合計の3次元散布図(対数スケ-
    ル)と修正回帰平面 (非対称正規誤差, 異常値の除去後)
265 #-----
266 library(rgl, pos=4)
267 library(mgcv, pos=4)
268 #library(magick)
269 #-----
270 b <- sqrt(2/pi)
271 delta <- function(alpha) alpha/sqrt(1+alpha^2)
272 omega.b.delta <- function(omega,alpha) omega*b*delta(alpha)
273 #-----
274 plot3d(log(x20200301[, c("employees", "assets", "sales")] ), type = "s", col
    = "red", size = 1,
    xlab = "log(employees)", ylab = "log(assets)", zlab = "log(sales)")
275 planes3d(coef.log.selm.x20200301.otl[2],
    coef.log.selm.x20200301.otl[3],
    -1,
    coef.log.selm.x20200301.otl[1]
    + omega.b.delta(omega = coef.log.selm.x20200301.otl[4],
    alpha = coef.log.selm.x20200301.otl[5]
    ),
    alpha=0.5)
284 #movie3d(spin3d(axis = c(0,0,1), rpm = 10), movie = "sp3d-log-plane-SN",
    dir = "./sp3d", duration=6, type = "gif")
285 #-----
286 # 2020年3月期決算企業の売上高, 従業員数, 資産合計の3次元散布図(対数スケ-
    ル)と修正標準回帰平面(非対称テ-誤差, 異常値の除去後)
287 #-----
288 library(rgl, pos=4)
289 library(mgcv, pos=4)

```



```

290 #library(magick)
291 #-----
292 bnu <- function(nu) sqrt(nu/pi)*gamma((nu-1)/2)/gamma(nu/2)
293 delta <- function(alpha) alpha/sqrt(1+alpha^2)
294 omega.bnu.delta <- function(omega,alpha,nu) omega*bnu(nu)*delta(alpha)
295 #-----
296 plot3d(log(x20200301[, c("employees", "assets", "sales")])), type = "s", col
      = "red", size = 1,
297       xlab = "log(employees)", ylab = "log(assets)", zlab = "log(sales)")
298 planes3d(coef.log.selm.ST.x20200301.otl[2],
299          coef.log.selm.ST.x20200301.otl[3],
300          -1,
301          coef.log.selm.ST.x20200301.otl[1]
302          + omega.bnu.delta(omega = coef.log.selm.ST.x20200301.otl[4],
303                          alpha = coef.log.selm.ST.x20200301.otl[5],
304                          nu = coef.log.selm.ST.x20200301.otl[6]),
305          alpha=0.5)
306 #movie3d(spin3d(axis = c(0,0,1), rpm = 10), movie = "sp3d-log-plane-ST",
307          dir = "./sp3d", duration=6, type = "gif")
308 #-----
309 # 両対数モデルのAICによる選択 (正規誤差, 非対称正規誤差, 非対称テール誤差)
310 #-----
311 AIC(log.lm.x20200301.otl, log.selm.x20200301.otl, log.selm.ST.x20200301.otl
312     )
313 #-----
314 # グミー変数付き両対数モデルの当てはめ (正規誤差)
315 #-----
316 log.lm.x20200301.otl.dum <- lm(log(sales) ~ log(employees) + log(assets) +
317     sector2, data = x20200301.otl)
318 coef.log.lm.x20200301.otl.dum <- coef(log.lm.x20200301.otl.dum)
319 summary(log.lm.x20200301.otl.dum)
320 #-----
321 # 回帰診断
322 #-----
323 par(mfcol = c(2, 2))
324 plot(resid(log.lm.x20200301.otl.dum), ylab="Residuals")
325 mtext("Index_Plot_of_Residuals", 3, 0.25, cex = 1)
326 plot(log.lm.x20200301.otl.dum, which=c(1,2))
327 plot(density(resid(log.lm.x20200301.otl.dum)), main="")
328 mtext("Density_Plot_of_Residuals", 3, 0.25, cex = 1)
329 par(mfcol = c(1, 1))
330 #
331 influenceIndexPlot(log.lm.x20200301.otl.dum, vars=c("hat", "Studentized", "
332     Cook")) #id.n=4, id.cex = 0.4)
333 influencePlot(log.lm.x20200301.otl.dum)
334 #-----
335 # 2020年3月期決算の企業の売上高, 従業員数, 資産合計の3次元散布図(対数スケール)と標本回帰平面群(正規誤差, 異常値の除去後, グミー変数付き)
336 #-----
337 library(rgl, pos=4)
338 library(mgcv, pos=4)
339 #library(magick)
340 #-----
341 plot3d(log(x20200301.otl[, c("employees", "assets", "sales")])), type = "s",

```

```

338     col = factor(x20200301.otl$sector2),
339     size = 0.5,
340     xlab = "log(employees)", ylab = "log(assets)", zlab = "log(sales)")
341 planes3d(coef.log.lm.x20200301.otl.dum[2],
342         coef.log.lm.x20200301.otl.dum[3],
343         -1,
344         coef.log.lm.x20200301.otl.dum[1],
345         col = factor(x20200301.otl$sector2),
346         alpha=0.08)
347 for(j in 1:32)
348 {
349     planes3d(coef.log.lm.x20200301.otl.dum[2],
350         coef.log.lm.x20200301.otl.dum[3],
351         -1,
352         coef.log.lm.x20200301.otl.dum[1]
353         + coef.log.lm.x20200301.otl.dum[j+3],
354         col = factor(x20200301.otl$sector2),
355         alpha=0.08)
356 }
357 #movie3d(spin3d(axis = c(0,0,1), rpm = 10), movie = "sp3d-log-planes", dir
358 = "./sp3d", duration=6, type = "gif")
359 #-----
360 # ダミー変数付き両対数モデルの当てはめ (非対称正規誤差)
361 #-----
362 library(sn)
363 log.selm.x20200301.otl.dum <- selm(log(sales) ~ log(employees) + log(assets
364     ) + sector2, data = x20200301.otl)
365 coef.log.selm.x20200301.otl.dum <- coef(log.selm.x20200301.otl.dum, param.
366     type="DP")
367 summary(log.selm.x20200301.otl.dum, param.type = "DP")
368 par(mfcol = c(1, 2))
369 plot(log.selm.x20200301.otl.dum, which = 2)
370 plot(log.selm.x20200301.otl.dum, which = 4)
371 par(mfcol = c(1, 1))
372 #-----
373 # 2020年3月期決算の企業の売上高、従業員数、資産合計の3次元散布図(対数スケ
374     ル)と標本回帰平面群(非対称正規誤差、異常値の除去後、ダミー変数付き)
375 #-----
376 library(rgl, pos=4)
377 library(mgcv, pos=4)
378 #library(magick)
379 #-----
380 plot3d(log(x20200301.otl[, c("employees", "assets", "sales")]), type = "s",
381     col = factor(x20200301.otl$sector2),
382     size = 0.5,
383     xlab = "log(employees)", ylab = "log(assets)", zlab = "log(sales)")
384 planes3d(coef.log.selm.x20200301.otl.dum[2],
385     coef.log.selm.x20200301.otl.dum[3],
386     -1,
387     coef.log.selm.x20200301.otl.dum[1]
388     + omega.b.delta(omega = coef.log.selm.x20200301.otl.dum[36],
389         alpha = coef.log.selm.x20200301.otl.dum[37]),
390     col = factor(x20200301.otl$sector2),
391     alpha=0.08)

```

```

388 for(j in 1:32)
389 {
390   planes3d(coef.log.selm.x20200301.otl.dum[2],
391           coef.log.selm.x20200301.otl.dum[3],
392           -1,
393           coef.log.selm.x20200301.otl.dum[1]
394           + coef.log.selm.x20200301.otl.dum[j+3]
395           + omega.bnu.delta(omega = coef.log.selm.x20200301.otl.dum[36],
396                           alpha = coef.log.selm.x20200301.otl.dum[37]),
397           col = factor(x20200301.otl$sector2),
398           alpha=0.08)
399 }
400 #movie3d(spin3d(axis = c(0,0,1), rpm = 10), movie = "sp3d-log-planes-ST",
401         dir = "./sp3d", duration=6, type = "gif")
402 #-----
403 # ダミー変数付き両対数モデルの当てはめ (非対称ティー誤差)
404 #-----
405 log.selm.ST.x20200301.otl.dum <- selm(log(sales) ~ log(employees) + log(
406   assets) + sector2, family = "ST", data=x20200301.otl)
407 coef.log.selm.ST.x20200301.otl.dum <- coef(log.selm.ST.x20200301.otl.dum,
408   param.type="DP")
409 summary(log.selm.ST.x20200301.otl.dum, param.type = "DP")
410 par(mfcol = c(1, 2))
411 plot(log.selm.ST.x20200301.otl.dum, param.type = "pseudo -CP", which = 2)
412 plot(log.selm.ST.x20200301.otl.dum, param.type = "pseudo -CP", which = 4)
413 par(mfcol = c(1, 1))
414 #-----
415 # 2020年3月期決算の企業の売上高、従業員数、資産合計の3次元散布図(対数スケール)
416 # と標本回帰平面群(非対称ティー誤差、異常値の除去後、ダミー変数付き)
417 #-----
418 library(rgl, pos=4)
419 library(mgcv, pos=4)
420 #library(magick)
421 #-----
422 bnu <- function(nu) sqrt(nu/pi)*gamma((nu-1)/2)/gamma(nu/2)
423 delta <- function(alpha) alpha/sqrt(1+alpha^2)
424 omega.bnu.delta <- function(omega,alpha,nu) omega*bnu(nu)*delta(alpha)
425 #-----
426 plot3d(log(x20200301.otl[, c("employees", "assets", "sales")] ), type = "s",
427       col = factor(x20200301.otl$sector2),
428       size = 0.5,
429       xlab = "log(employees)", ylab = "log(assets)", zlab = "log(sales)")
430 planes3d(coef.log.selm.ST.x20200301.otl.dum[2],
431         coef.log.selm.ST.x20200301.otl.dum[3],
432         -1,
433         coef.log.selm.ST.x20200301.otl.dum[1]
434         + omega.bnu.delta(omega = coef.log.selm.ST.x20200301.otl.dum[36],
435                           alpha = coef.log.selm.ST.x20200301.otl.dum[37],
436                           nu = coef.log.selm.ST.x20200301.otl.dum[38]+1),
437         col = factor(x20200301.otl$sector2),
438         alpha=0.08)
439 for(j in 1:32)
440 {
441   planes3d(coef.log.selm.ST.x20200301.otl.dum[2],

```

```

438     coef.log.selm.ST.x20200301.otl.dum[3],
439     -1,
440     coef.log.selm.ST.x20200301.otl.dum[1]
441     + coef.log.selm.ST.x20200301.otl.dum[j+3]
442     + omega.bnu.delta(omega = coef.log.selm.ST.x20200301.otl.dum
443         [36],
444         alpha = coef.log.selm.ST.x20200301.otl.dum
445             [37],
446         nu = coef.log.selm.ST.x20200301.otl.dum
447             [38]+1),
448     col = factor(x20200301.otl$sector2),
449     alpha=0.08)
450 }
451 #movie3d(spin3d(axis = c(0,0,1), rpm = 10), movie = "sp3d-log-planes-ST",
452     dir = "./sp3d", duration=6, type = "gif")
453 #-----
454 # ダミー変数付き両対数モデルのAICによる選択 (正規誤差, 非対称正規誤差, 非対
455     称ティー誤差)
456 #-----
457 AIC(log.lm.x20200301.otl.dum, log.selm.x20200301.otl.dum, log.selm.ST.
458     x20200301.otl.dum)
459 #-----
460 # 決定係数の経年変化のプロット
461 #-----
462 OLS.ts <- function(obj)
463 {
464     require(dplyr)
465     require(purrr)
466     require(tidyrr)
467     otl <- c("JAPANSECURITIESFINANCE0070514-1", "JACCS0001710-1", "
468         JAPANPOSTHOLDINGS0038793-1", "JAPANEXCHANGEGROUP0075107-3")
469     obj <- obj %>% mutate(name.ac = paste(name, ac, sep = "-"))
470     firms <- unique(obj$name.ac)
471     x.otl <- obj %>% filter(name.ac %in% setdiff(firms, otl))
472     tp <- seq(as.Date("1984-03-01"), as.Date("2020-03-01"), by = "year")
473     x.otl <- x.otl %>% filter(ymd %in% tp)
474     n_x <- x.otl %>% group_by(ymd) %>% arrange(ymd) %>% nest()
475     mod_fun <- function(df) lm(log(sales) ~ log(employees) + log(assets) +
476         sector2, data = df)
477     m_x <- n_x %>% mutate(model = map(data, mod_fun))
478     n_fun <- function(mod) length(mod$fitted.values)
479     df_fun <- function(mod) mod$df.residual
480     r_fun <- function(mod) summary(mod)$r.squared
481     adj_r_fun <- function(mod) summary(mod)$adj.r.squared
482     res <- m_x %>% transmute(ymd,
483         n.obs = map_int(model, n_fun),
484         df = map_int(model, df_fun),
485         R2 = map_dbl(model, r_fun),
486         adj.R2 = map_dbl(model, adj_r_fun))
487     data.frame(res)
488 }
489 x.OLS.ts <- OLS.ts(x)
490 x.OLS.ts %>% data.frame()
491 plot.OLS.ts <- function(obj)

```

```

484 {
485   require(ggplot2)
486   require(reshape)
487   #require(tidyr)
488   require(dplyr)
489   #obj %>% pivot_longer(cols = ~"ymd") %>% arrange(name) %>%
490   obj %>% melt(id.vars = "ymd") %>%
491     ggplot(aes(ymd, value, group = variable)) +
492     geom_point() + geom_line(linetype = "dashed") + facet_grid(variable ~
493       ., scale="free_y")
494 }
494 plot.OLS.ts(x.OLS.ts)
495 #-----
496 # AICの経年変化のプロット
497 #-----
498 AIC.ts <- function(obj)
499 {
500   require(sn)
501   require(tidyverse)
502   require(lubridate)
503   #
504   otl <- c("JAPANSECURITIESFINANCE0070514-1", "JACCS0001710-1", "
505     JAPANPOSTHOLDINGS0038793-1", "JAPANEXCHANGEGROUP0075107-3")
506   obj <- obj %>% mutate(name.ac = paste(name, ac, sep = "-"))
507   firms <- unique(obj$name.ac)
508   x.otl <- obj %>% filter(name.ac %in% setdiff(firms, otl))
509   tp <- seq(as.Date("1984-03-01"), as.Date("2020-03-01"), by = "year")
510   x.otl <- x.otl %>% filter(ymd %in% tp)
511   n_x <- x.otl %>% group_by(ymd) %>% arrange(ymd) %>% nest()
512   mod_fun <- function(df) lm(log(sales) ~ log(employees) + log(assets) +
513     sector2, data = df)
514   mod_fun_SN <- function(df) selm(log(sales) ~ log(employees) + log(assets)
515     + sector2, family = "SN", data = df)
516   mod_fun_ST <- function(df) selm(log(sales) ~ log(employees) + log(assets)
517     + sector2, family = "ST", data = df)
518   lm_selm_x <- n_x %>% mutate(lmres = map(data, mod_fun),
519     selmSNres = map(data, mod_fun_SN),
520     selmSTres = map(data, mod_fun_ST))
521   n_fun <- function(lmres) length(lmres$fitted.values)
522   aic_fun <- function(lmres) AIC(lmres)
523   res <- lm_selm_x %>% transmute(ymd,
524     n.obs = map_int(lmres, n_fun),
525     AIC.N = map_dbl(lmres, aic_fun),
526     AIC.SN = map_dbl(selmSNres, aic_fun),
527     AIC.ST = map_dbl(selmSTres, aic_fun))
528   data.frame(res)
529 }
530 suppressWarnings(invisible(capture.output(x.AIC.ts <- AIC.ts(x))))
531 plot.AIC <- function(obj)
532 {
533   require(tidyverse)
534   require(gridExtra)
535   x <- obj %>% pivot_longer(~ymd, names_to = "model", values_to = "value")
536   y <- obj %>% mutate(AIC.N.n = AIC.N/n.obs, AIC.SN.n = AIC.SN/n.obs, AIC.

```

```

      ST.n = AIC.ST/n.obs) %>%
533   select(ymd, n.obs, AIC.N.n, AIC.SN.n, AIC.ST.n) %>%
534   pivot_longer(~ymd, names_to = "model", values_to = "value")
535   p1 <- x %>% filter(model == "n.obs") %>%
536   ggplot(aes(ymd, value, group = model, shape = model)) +
537     geom_line(linetype="dotted") + geom_point() + ggtitle("n")
538   p2 <- x %>% filter(model != "n.obs") %>%
539   ggplot(aes(ymd, value, group = model, color = model, shape = model)) +
540     geom_line(linetype="dotted") + geom_point() + ggtitle("AIC")
541   p3 <- y %>% filter(model != "n.obs") %>%
542   ggplot(aes(ymd, value, group = model, color = model, shape = model)) +
543     geom_line(linetype="dotted") + geom_point() + ggtitle("AIC/n")
544   grid.arrange(p1, p2, p3)
545 }
546 plot.AIC(x.AIC.ts)

```

付録 E 動的文書生成による再現可能研究

一般に、科学・技術関連の論文を作成する際、文書作成とデータ解析を別々に行い、解析結果（図、表、テキストなど）を文書に手作業（マニュアル）で挿入する方法、いわゆる、通常の文書作成手順に従って作成されたものは、全く同一のものを再度作成することが困難であるということが指摘されている。近年、研究の再現性を確保するという意味で、「再現可能研究」と呼ばれ、この問題に対する解決法が議論されている。再現可能研究を実現するための一つの方法として、データ解析言語のコードを文書に埋め込み、それらを何らかの方法で自動実行することによって、解析結果を動的に生成した後、さらにそれらを自動的に読み込んで文書を作成するという、いわゆる、「動的文書生成」が提案されている²⁷⁾。

動的文書生成を実現するためのツールは、近年も活発に開発されているものもあるが²⁸⁾、本研究では、Norman Ramsey による `noweb`²⁹⁾ をベースとし

27) 再現可能研究と動的文書生成に関しては、例えば、Xie, 2015, Gandrud, 2020, 高橋, 2014, 2018等を参照されたい。また、地道 (2018-a, b), 地道, 豊原 (2018) では景気循環や財務データの分析に関する動的文書生成による再現可能性の確保に関して議論されているので、併せて参照されたい。

28) 最近の動向としては、RStudio 上で `knitr` パッケージを利用する方法が主流となりつつある。詳細は、Xie (2015), 高橋 (2014, 2018) などを参照されたい。

29) <https://www.cs.tufts.edu/~nr/noweb/>

て Leisch (2002) によって開発された **Sweave** を文書作成のために利用し、データ解析を含む全体の制御のために **GNU make** (cf. Mecklenburg, 2005) を利用した³⁰⁾。

ここでは、本稿を作成するための動的文書生成のための環境や仕様を紹介する。まず、本稿のソースファイルは、ディレクトリ構成 (図32) の paper ディレクトリの **Sweave** ファイル³¹⁾ EFDA-RR-NEEDS-R.Rnw であり、本稿を動的に生成するためには、図32のディレクトリ paper の Makefile (スクリプト6) を利用する。

スクリプト6：動的文書生成を実行するための Makefile ファイル

```

1 all:
2     date > start-all.txt
3     (cd ../R; make RData)
4     (/bin/bash makepaper.sh)
5     date > end-all.txt
6 paper:
7     /bin/bash makepaper.sh
8 clean:
9     rm EFDA-RR-NEEDS-R-*.png
10    rm -r EFDA-RR-NEEDS-R.tex *.log *.dvi *.aux *.out *.toc

```

Makefile ファイル (スクリプト6) で与えられているターゲット all の役割とディレクトリ構成 (図32) の対応を図33に与える。

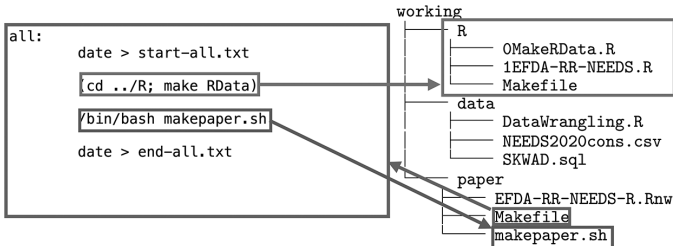


図33：Makefile ファイル (スクリプト6) のターゲット all とディレクトリ構成 (図32) の対応

30) 本稿で利用している動的文書生成のための環境は、開発から既に数十年がたっているが、安定性を重視する意味で利用している。

31) Rnw は、R noweb ファイルの拡張子である。

Makefile ファイル (スクリプト6) の1行目, 6行目, 8行目でターゲットが定義されており, 1番目のターゲット `all` を `make` コマンドで実行することによって本稿に必要なデータ解析や `Sweave` の処理, コンパイルなどの全工程が動的に実行される³²⁾. 具体的には, 図32における `paper` ディレクトリをカレントとして, 以下のように `make` コマンドを実行することによって, 動的文書生成が実行される.

make コマンドによるターゲット `all` の実行

```
% make all
```

ここで, `%` はシェルプロンプトであり, スクリプト6のターゲット `all` の実行に伴う具体的な処理の流れは以下のようなものである:

(M1) Makefile ファイル (スクリプト6) の3行目でディレクトリ `R` の Makefile (スクリプト7) で定義されたターゲット `RData` を `make` コマンドで実行している. この操作によって, ディレクトリ `R` (図32参照) の `R` スクリプトファイル `0MakeRData.R` (スクリプト4) が `Rscript` コマンドで処理され, `CSV` ファイル `NEEDS2020cons.csv` が自動的に読み込まれ, 本稿で利用されるデータ解析の結果が納められた作業空間ファイル `EFDA-RR-NEEDS-R.RData` が自動生成される (図34も参照).

スクリプト7: データ解析結果の作業空間ファイル `EFDA-RR-NEEDS-R.RData` を自動生成するためのターゲット `RData` が定義された Makefile ファイル

```
1 RData:
2     Rscript 0MakeRData.R
3 clean:
4     rm *.RData
```

32) Makefile (スクリプト6) のターゲット `paper` は, 文書 (論文) 生成に関する工程を実行するためのものであり, `clean` は中間ファイル等を削除 (クリーン) するためのものである.

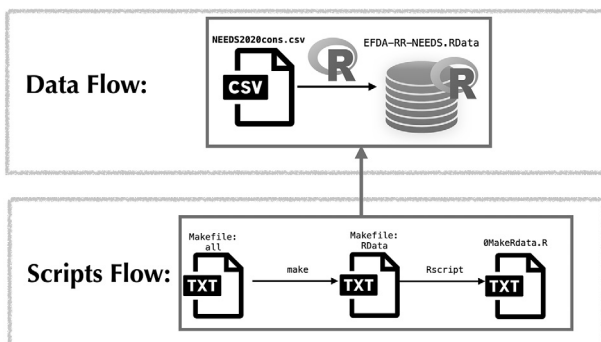


図34：データ解析結果の作業空間ファイル `EFDA-RR-NEEDS-R.RData` の自動生成に関するデータファイルとスクリプトファイルの流れと対応

(M2) Makefile ファイル (スクリプト 6) の 4 行目で動的に文書 (`EFDA-RR-NEEDS-R.pdf`) を生成するためのシェルスクリプトファイル `makepaper.sh` (スクリプト 8) が実行され、本稿が自動生成される (図35も参照)。

スクリプト 8：動的文書生成をするためのシェルスクリプトファイル `makepaper.sh`

```
1 #!/bin/sh
2 ~/Library/TeXShop/Engines/Sweave-utf8.engine EFDA-RR-NEEDS-R.Rnw
```

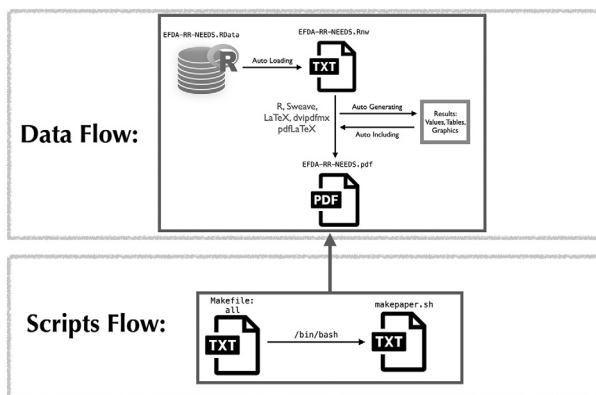


図35：本稿の自動生成に関するデータファイルとスクリプトファイルの流れと対応

シェルスクリプトファイル `makepaper.sh` (スクリプト 8) から呼び出されるシェルスクリプトファイル `Sweave-utf8.engine` のソースコードは、スクリプト 9 を参照されたい。

スクリプト 9 : Sweave による処理と L^AT_EX によるコンパイルを実行するシェルスクリプトファイル `Sweaveutf8.engine`

```

1 #!/bin/sh
2 export LANG=ja_JP.UTF-8
3 export PATH=$PATH:/Library/TeX/texbin:/Library/Frameworks/R.framework/
  Resources
4 R CMD Sweave --encoding="utf8" "$1"
5 filename=${1%.*}
6 ptex2pdf -l -ot "-synctex=1_-file-line-error" "$filename"
7 ptex2pdf -l -ot "-synctex=1_-file-line-error" "$filename"

```

(M3) 2 行目と 5 行目は処理時間の計測を行うための指定である。

Makefile ファイル (スクリプト 6) のターゲット `all` の `make` コマンドによる実行によるデータファイルと文書ファイル, スクリプトファイルの流れ, 及びスクリプトファイルから実行されるデータファイルと文書ファイルへの全処理の対応関係を図36に与える。

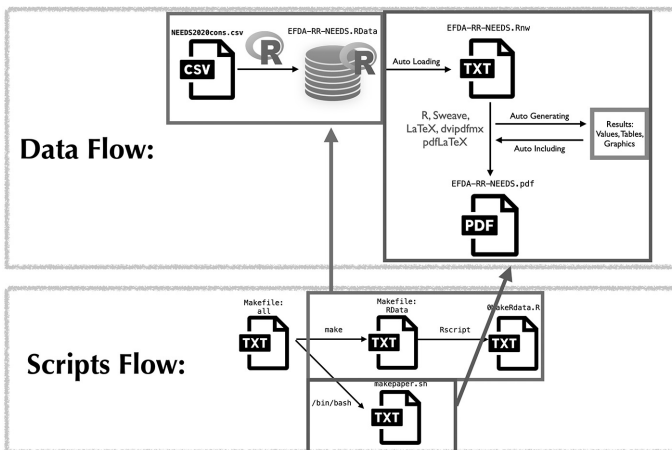


図36 : ターゲット `all` の実行に伴うデータファイル, 文書ファイル, スクリプトファイルの流れと対応

make コマンドの実行による処理時間は、スクリプト 6 における、2 行目と 5 行目の実行結果を比較することによってわかる。iMac 2017 (macOS Big Sur) 上で実行した結果を以下に与える：

iMac 2017 (macOS Big Sur) 上でターゲット a11 の処理時間の計測

```
% cat start-preprocess.txt
2021年11月7日 日曜日 13時35分33秒 JST
% cat end-preprocess.txt
2021年11月7日 日曜日 13時37分20秒 JST
```

この結果から、1分47秒である。また、MacBook Pro 2018 (macOS Big Sur) 上で実行した結果を以下に与える：

MacBook Pro 2018 (macOS Big Sur) 上でターゲット a11 の処理時間の計測

```
% cat start-preprocess.txt
2021年11月7日 日曜日 17時46分43秒 JST
% cat end-preprocess.txt
2021年11月7日 日曜日 17時48分58秒 JST
```

この結果から、2分15秒である。

本節の冒頭で述べた再現可能研究について、その達成度を表す基準として、Peng (2011) による「再現可能性スペクトル」(reproducibility spectrum) が興味深い (Peng, 2011 の Fig. 1 を参照)。それによると、再現可能性の達成度を表す以下のグレードが提示されている：

- (G1) 論文などを「公表しただけのもの」(publication only)：再現可能ではない (not reproducible)
- (G2) 「コードが管理されているもの」(code)
- (G3) 「コードとデータが管理されているもの」(code and data)
- (G4) 「コードとデータがリンクしており、実行できるもの」(linked and executable code and data)
- (G5) 「完全に再現するもの」(full replication)：ゴールドスタンダード (gold standard)

Peng (2011) は、これらの段階をスペクトルとして捉えており、これに照らすと、本研究は make によって全工程を自動実行して再現性を確保しているため、「ゴールドスタンダード」に属するものと思われる。よって、図 3 における全工程における再現可能研究を高いレベルで保証しているものといえよう。

付録 F 回帰分析における感度分析のための指標

回帰分析における感度分析に利用される主な指標の定義を与える。まず、線形回帰モデルの成分表現を

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, \dots, n$$

とし、ベクトル・行列表現を

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

とする。ここで、

$$\mathbf{y} := \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} := \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}'_1 \\ \vdots \\ \mathbf{x}'_n \end{bmatrix}, \quad \boldsymbol{\beta} := \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol{\epsilon} := \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

とおいた。ただし、 $\mathbf{x}'_i = [1, x_{i1}, \dots, x_{ip}]$ であり、プライム (') はベクトルや行列の転置を表す記号である。

回帰係数ベクトル $\boldsymbol{\beta}$ の最小自乗推定値ベクトルを

$$\hat{\boldsymbol{\beta}} := (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

とすると、当てはめ値のベクトルは、

$$\hat{\mathbf{y}} := \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{P}\mathbf{y}$$

で定義される。ここで、

$$\mathbf{P} := \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

は射影行列である。また、残差ベクトルは、

$$\mathbf{e} := \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{P}\mathbf{y} = (\mathbf{I}_n - \mathbf{P})\mathbf{y}$$

で与えられる。

i 番目の観測に対する当てはめ値 \hat{y}_i と残差 $e_i := y_i - \hat{y}_i$ は、それぞれ、当てはめ値のベクトル $\hat{\mathbf{y}}$ と残差ベクトル \mathbf{e} の第 i 成分で与えられる。

また、誤差分散 σ^2 の推定値は、

$$\hat{\sigma}^2 := \frac{1}{n-p-1} \sum_{i=1}^n e_i^2 = \frac{1}{n-p-1} \mathbf{e}'\mathbf{e}$$

で与えられる。

以上の設定のもとで、感度分析で利用される指標を構成する際の基本的かつ重要なアイデアは、ある指標についてデータ点 (\mathbf{x}'_i, y_i) を取り除いて計算したものと、データ点を取り除かないで計算されたものとを比較することによって、それらがどの程度異なっているかを見ることであり、この「差」がそのデータ点の影響力と見なされる。

まず、射影行列 \mathbf{P} の対角成分はハット値 (hat-values) と呼ばれ以下のように定義される：

$$h_i := p_{ii} := [\mathbf{P}]_{ii}$$

ハット値 h_i は、観測値 y_i に対する当てはめ値 \hat{y}_i を求める際の y_i に対する重みそのものである。すなわち、

$$\hat{y}_i = \sum_{j=1}^n p_{ij} y_j = h_i y_i + \sum_{j \neq i} p_{ij} y_j.$$

つぎに、残差 e_i を以下のように修正したものをスチューデント化残差 (Studentized residual) という：

$$e_{Ti} := \frac{e_i}{\hat{\sigma}_{(-i)} \sqrt{1-h_i}}$$

ここで、 $\hat{\sigma}_{(-i)} := \sqrt{\hat{\sigma}_{(-i)}^2}$ であり、 $\hat{\sigma}_{(-i)}^2$ は、 i 番目のデータ点 (\mathbf{x}'_i, y_i) を取り除いて計算した誤差分散 σ^2 の推定値である。

さらに、以下の指標をクックの距離 (Cook's Distance) という：

$$D_i := \frac{(\hat{\boldsymbol{\beta}}_{(-i)} - \hat{\boldsymbol{\beta}})' \mathbf{X}' \mathbf{X} (\hat{\boldsymbol{\beta}}_{(-i)} - \hat{\boldsymbol{\beta}})}{(p+1) \hat{\sigma}^2}$$

ここで、 $\hat{\boldsymbol{\beta}}_{(-i)}$ は、 i 番目のデータ点 (\mathbf{x}'_i, y_i) を取り除いて求めた $\boldsymbol{\beta}$ に対する最小自乗推定値ベクトルである。

これらの指標の詳細については、たとえば、Chatterjee and Hadi (1988) を参照されたい。

付録 G 非対称分布

G.1 非対称正規分布

定義 1 (非対称正規分布) 確率変数 X が確率密度関数 (probability density function: p.d.f.)

$$f_{\text{SN}}(x|\boldsymbol{\theta}) = \frac{2}{\omega} \phi\left(\frac{x-\xi}{\omega}\right) \Phi\left(\alpha \frac{x-\xi}{\omega}\right), x \in \mathbb{R} = (-\infty, \infty) \quad (18)$$

をもつとき、確率変数 X は非対称正規分布 $\text{SN}(\xi, \omega^2, \alpha)$ に従うと呼ばれ、

$$X \sim \text{SN}(\xi, \omega^2, \alpha)$$

と書かれる。ここで、

$$\xi \in \mathbb{R}, \quad \omega \in \mathbb{R}^+ := (0, \infty), \quad \alpha \in \mathbb{R}$$

は未知母数であり、 $\boldsymbol{\theta} = [\xi, \omega, \alpha]'$ は母数ベクトルである。また、

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right), \quad \Phi(z) = \int_{-\infty}^z \phi(x) dx \quad (z \in \mathbb{R})$$

は、それぞれ、標準正規分布 $N(0, 1)$ の p.d.f. と c.d.f. (累積分布関数) である。なお、 (ξ, ω^2, α) は直接母数 (direct parameters) と呼ばれる。

G.2 非対称テイー分布

定義 2 (非対称テイー分布) 確率変数 X が p.d.f.:

$$f_{\text{ST}}(x|\boldsymbol{\theta}) = \frac{2}{\omega} f_t\left(\frac{x-\xi}{\omega} \middle| \nu\right) F_t\left(\alpha \frac{x-\xi}{\omega} \sqrt{\frac{\nu+1}{\left(\frac{x-\xi}{\omega}\right)^2 + \nu}} \middle| \nu+1\right), x \in \mathbb{R} \quad (19)$$

をもつとき、確率変数 X は非対称テイー分布 $\text{ST}(\xi, \omega^2, \alpha, \nu)$ に従うと呼ばれ、

$$X \sim \text{ST}(\xi, \omega^2, \alpha, \nu)$$

と書かれる。ここで、

$$\xi \in \mathbb{R}, \quad \omega \in \mathbb{R}^+, \quad \alpha \in \mathbb{R}, \quad \nu \in \mathbb{R}^+$$

は未知母数であり、 $\theta = [\xi, \omega, \alpha, \nu]'$ は母数ベクトルである。

$$f_i(z|\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}} \left(1 + \frac{z^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad F_i(z|\nu) = \int_{-\infty}^z f_i(x|\nu) dx$$

は、それぞれ、自由度 ν のティー分布の p.d.f. と c.d.f. である。なお、 $(\xi, \omega, \alpha, \nu)$ は直接母数と呼ばれる。

付録 H 日経業種分類

大分類	中分類	小分類	大分類	中分類	小分類
製造業	食品	飼料	1	01	001
製造業	食品	砂糖	1	01	002
製造業	食品	製粉	1	01	003
製造業	食品	食油	1	01	004
製造業	食品	酒類	1	01	005
製造業	食品	製菓・パン	1	01	006
製造業	食品	ハム	1	01	007
製造業	食品	調味料	1	01	008
製造業	食品	乳製品	1	01	009
製造業	食品	その他食品	1	01	010
製造業	繊維	化合織	1	03	021
製造業	繊維	綿紡績	1	03	022
製造業	繊維	絹紡績	1	03	023
製造業	繊維	毛紡績	1	03	024
製造業	繊維	繊維二次加工	1	03	025
製造業	繊維	その他繊維	1	03	026
製造業	パルプ・紙	大手製紙	1	05	041
製造業	パルプ・紙	その他パルプ・紙	1	05	042
製造業	化学	大手化学	1	07	061
製造業	化学	肥料	1	07	062
製造業	化学	塩素・ソーダ	1	07	063
製造業	化学	石油化学	1	07	064
製造業	化学	合成樹脂	1	07	065
製造業	化学	酸素	1	07	066
製造業	化学	油脂・洗剤	1	07	067
製造業	化学	化粧品・歯磨	1	07	068
製造業	化学	塗料・インキ	1	07	069
製造業	化学	農薬・殺虫剤	1	07	070
製造業	化学	その他化学	1	07	071
製造業	医薬品	大手医薬品	1	09	081

大分類	中分類	小分類	大分類	中分類	小分類
製造業	医薬品	医家向医薬品	1	09	082
製造業	医薬品	大衆向医薬品	1	09	083
製造業	石油	石油精製及び販売	1	11	101
製造業	石油	石炭石油製品	1	11	102
製造業	ゴム	タイヤ	1	13	121
製造業	ゴム	その他ゴム製品	1	13	122
製造業	窯業	ガラス	1	15	141
製造業	窯業	セメント一次	1	15	142
製造業	窯業	セメント二次	1	15	143
製造業	窯業	陶器	1	15	144
製造業	窯業	耐火煉瓦	1	15	145
製造業	窯業	カーボン・その他	1	15	146
製造業	鉄鋼	鉄鋼一貫	1	17	161
製造業	鉄鋼	平電炉・単圧	1	17	162
製造業	鉄鋼	特殊鋼	1	17	163
製造業	鉄鋼	合金鉄	1	17	164
製造業	鉄鋼	鋳鍛鋼	1	17	165
製造業	鉄鋼	ステンレス	1	17	166
製造業	鉄鋼	その他鉄鋼	1	17	167
製造業	非鉄金属製品	大手精錬	1	19	181
製造業	非鉄金属製品	その他精錬	1	19	182
製造業	非鉄金属製品	アルミ加工（含ダイカスト）	1	19	183
製造業	非鉄金属製品	電線・ケーブル	1	19	184
製造業	非鉄金属製品	鉄骨・鉄塔・橋梁	1	19	185
製造業	非鉄金属製品	その他金属製品	1	19	186
製造業	機械	工作機械	1	21	201
製造業	機械	プレス機械	1	21	202
製造業	機械	繊維機械	1	21	203
製造業	機械	運搬機・建設機械・内燃機	1	21	204
製造業	機械	農業機械	1	21	205
製造業	機械	化工機械	1	21	206
製造業	機械	ミシン・編機	1	21	207
製造業	機械	軸受	1	21	208
製造業	機械	事務機	1	21	209
製造業	機械	その他機械	1	21	210
製造業	電気機器	総合電機	1	23	221
製造業	電気機器	重電	1	23	222
製造業	電気機器	家庭電器（含音響機器）	1	23	223
製造業	電気機器	通信機（含通信機部品）	1	23	224
製造業	電気機器	電子部品	1	23	225
製造業	電気機器	制御機器	1	23	226
製造業	電気機器	電池	1	23	227
製造業	電気機器	自動車関連	1	23	228
製造業	電気機器	その他電気機器	1	23	229
製造業	造船	造船	1	25	241

大分類	中分類	小分類	大分類	中分類	小分類
製造業	自動車	自動車	1	27	261
製造業	自動車	自動車部品	1	27	262
製造業	自動車	車体・その他	1	27	263
製造業	輸送用機器	車両	1	29	281
製造業	輸送用機器	自転車	1	29	282
製造業	輸送用機器	その他輸送用機器	1	29	283
製造業	精密機器	時計	1	31	301
製造業	精密機器	カメラ	1	31	302
製造業	精密機器	計器・その他	1	31	303
製造業	その他製造	印刷	1	33	321
製造業	その他製造	楽器	1	33	322
製造業	その他製造	建材	1	33	323
製造業	その他製造	事務用品	1	33	324
製造業	その他製造	その他製造業	1	33	325
非製造業	水産	水産	2	35	341
非製造業	鉱業	石炭鉱業	2	37	361
非製造業	鉱業	その他鉱業	2	37	362
非製造業	建設	大手建設	2	41	401
非製造業	建設	中堅建設	2	41	402
非製造業	建設	土木・道路・浚渫	2	41	403
非製造業	建設	電設工事	2	41	404
非製造業	建設	住宅	2	41	405
非製造業	建設	その他建設	2	41	406
非製造業	商社	総合商社	2	43	421
非製造業	商社	自動車販売	2	43	422
非製造業	商社	食品商社	2	43	423
非製造業	商社	繊維商社	2	43	424
非製造業	商社	機械金属商社	2	43	425
非製造業	商社	化学商社	2	43	426
非製造業	商社	建材商社	2	43	427
非製造業	商社	電機関連商社	2	43	428
非製造業	商社	その他商社	2	43	429
非製造業	小売業	百貨店	2	45	441
非製造業	小売業	スーパー	2	45	442
非製造業	小売業	月販店	2	45	443
非製造業	小売業	その他小売業	2	45	444
非製造業	銀行	長期信用銀行	2	47	461
非製造業	銀行	都市銀行	2	47	462
非製造業	銀行	地方銀行	2	47	463
非製造業	銀行	信託銀行	2	47	464
非製造業	銀行	相互銀行	2	47	465
非製造業	銀行	証券金融	2	47	466
非製造業	証券	証券	2	49	481
非製造業	保険	保険	2	51	501
非製造業	その他金融	その他金融業	2	52	511

大分類	中分類	小分類	大分類	中分類	小分類
非製造業	不動産	賃貸	2	53	521
非製造業	不動産	分譲	2	53	522
非製造業	鉄道・バス	大手私鉄	2	55	541
非製造業	鉄道・バス	中小私鉄	2	55	542
非製造業	鉄道・バス	バス・その他	2	55	543
非製造業	陸運	陸運	2	57	561
非製造業	海運	大手海運	2	59	581
非製造業	海運	内航	2	59	582
非製造業	海運	外航・その他	2	59	583
非製造業	空運	空運	2	61	601
非製造業	倉庫	倉庫	2	63	621
非製造業	倉庫	運輸関連	2	63	622
非製造業	通信	通信	2	65	641
非製造業	電力	電力	2	67	661
非製造業	ガス	ガス	2	69	681
非製造業	サービス	映画	2	71	701
非製造業	サービス	娯楽施設	2	71	702
非製造業	サービス	ホテル	2	71	703
非製造業	サービス	その他サービス業	2	71	704