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―

地　道　正　行

要　旨

本稿では，Bureau van Dijk 社から提供されるデータベース Osiris から抽出された世界157カ国の全上場企業（一般事業会社，上場廃止企業含む）の主要財務情報（売上高，営業利益，総資産など84項目の33年分）の財務データファイル（財務ビッグデータ）を，可視化することによって得られた知見を使って，いくつかの統計モデリングを行い，それらの選択と評価をする工程を動的文書を生成する立場から議論し，再現可能研究を実行する方法を検討する．本稿を含む一連の研究では，Tukey（1977）によって提唱された探索的データ解析をデータサイエンスを実行するカーネルとして位置付けし，財務ビッグデータから何らかの意味のある情報・知見を得ることを試みる．

キーワード：探索的データ解析（Exploratory Data Analysis），ビッグデータ（Big Data），データサイエンス（Data Science），データ可視化（Data Visualization），統計モデリング（Statistical Modeling），モデル選択（Model Selection），モデル評価（Model Evaluation），動的文書（Dynamic Documents），再現可能研究（Reproducible Research）

I　はじめに

地道（2018-a）では，データベース Osiris[1] から抽出された世界157カ国の全上場企業[2] 8万社超の主要財務情報（売上高，営業利益，総資産など84項

目，33年分）という，規模の大きなデータファイルを，データ解析環境に読み込めるファイル形式に変換する工程（前処理）と，実際にデータを解析できるオブジェクト形式に変換する工程（データラングリング）を，再現性（reproducibility）を確保することをふまえて考察した．

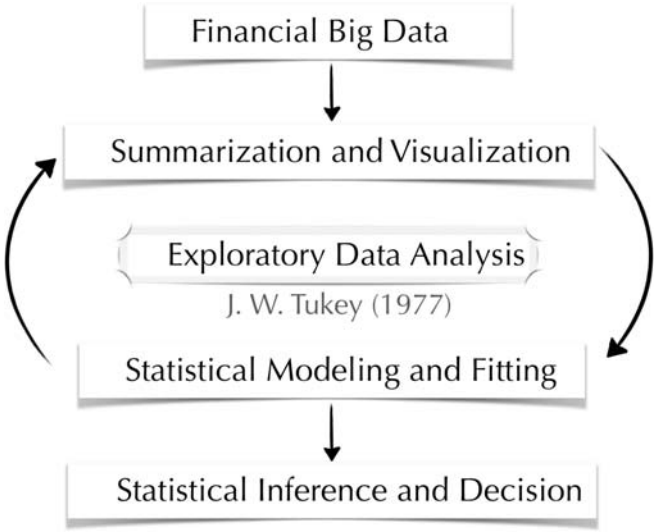

図１　探索的財務ビッグデータ解析

本稿では，この規模の大きなデータセット（ビッグデータ）に対して，Jimichi *et al.*（2018）で与えられたデータ解析（データ可視化，統計モデリング，モデル選択，モデル評価）の結果を導く工程を，動的文書（dynamic document）を生成することによって，再現可能研究（reproducible research）として実行するための方法について議論する．地道（2018-a）でも言及したが，本稿においても，Tukey（1977）によって提唱された「探索的データ解析」（Exploratory Data Analysis: EDA）を，データサイエンスを実行するた

1）　Bureau van Dijk（BvD）社（https://www.bvdinfo.com/en-gb/）から提供されるデータベースの一つ．
2）　上場廃止企業含む

めの指針とし，上記の財務ビッグデータから何らかの意味のある情報・知見を得るための核（カーネル）として位置づけする（図 1 も参照）．なお，利用したコンピュータ環境に関しては付録に与えている．

II　データ可視化

本節では，Jimichi *et al.*（2018）で与えられているデータ可視化の工程を，再現可能性を確認しながら外観する[3]．なお，可視化をおこなうための R パッケージとして，ggplot2[4] が利用されている．

一般に，データを可視化する主な目的は，データ発生メカニズムである「分布」の情報を得るためである．そのためには，シンプルであるが，ヒストグラム，対散布図，Q-Q プロットなどを描くことが，本質的に重要である．

ここでは，地道（2018-a）でおこなわれた前処理とデータラングリングの結果として得られたデータフレーム `firmfin2015` を使って，最も重要な可視化の結果である，売上高の対数（log(`sales`)）のヒストグラムを図 2 に与える（Jimichi *et al.*（2018）の Fig. 2 の右パネルも参照）．

Jimichi *et al.*（2018）では，このプロットから，売上高の対数（log(`sales`)）は，ほぼ正規分布と考えられるが，若干左に歪んでいることが指摘されている．

なお，実際に可視化をおこなった結果を文書化するために，以下のスクリプトを Rnw[5] ファイルに挿入し，`Sweave`[6] で処理する方法をとった：

3） データ可視化に関しては，Unwin（2015）などを参照されたい．

4） ggplot2 は，Wilkinson（2005）によって提唱されたグラフィックスを描くための文法（grammar of graphics）に従って R でグラフィックスを作成するためのパッケージである．Wickham（2016）によって実装された．なお，本稿では言及していないが，地道（2017-a, b）と Jimichi *et al.*（2018）では，対散布図をプロットするために，GGally パッケージも利用されている．

5） `Rnw` は，R `noweb` の略である．詳細は VI 節を参照されたい．

6） Sweave は，R のコードを埋め込んだ `Rnw` ファイルを処理して，LaTeX ファイルに変換するための R の関数である．

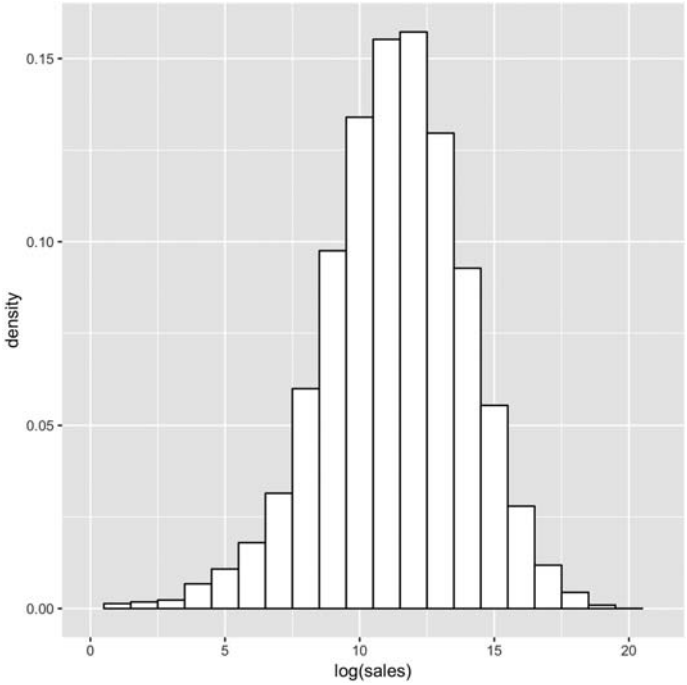

図2　売上高のヒストグラム：対数スケール

ソースコード1　売上高（対数スケール）のヒストグラムを文書に埋め込むためのスクリプト

```
<<echo=F,fig=T,png=T,pdf=F>>=
library(ggplot2)
ggplot(firmfin2015,aes(x=log(sales))) + xlim(0,21) +
  geom_histogram(aes(y=..density..),binwidth=1,fill="white",color="black")
@
```

リスト1における "<<***>>= ... @" はコードチャンク（code chunk）とよばれ，"..." にRのスクリプトを与え，"***" にオプションを与える．ここで利用されているオプションは，スクリプトの実行によるエコーバックを表示しない（echo=F）ことと，プロット結果を図として出力すること（fig=T），そして，図のファイル形式をPDFファイルではなく，PNGファ

イルとして出力することを指定（png=T,pdf=F）している[7]．

また，2行目では，ggplot2パッケージが読み込まれており，3，4行目でggplot関数を利用して，ヒストグラムが描かれている．ggplot2パッケージは，レイヤー構造を記述することができ，この機能を持つことが，このパッケージを利用する利点の一つである．このスクリプトでは，データフレームfirmfin2015に収められている売上高（sales）の対数スケール（log(sales)）をx軸に与えたレイヤー[8]を用意し，xの範囲のレイヤー（xlim(0,21)）とヒストグラムのレイヤーgeom_histogram(aes(y=..density..),binwidth=1,fill="white",color="black"）を+記号で重ね書きしている．

Rnwファイルを Sweaveで処理することによって，チャンク内のRコードが実行され，その結果がLaTeXファイルに自動的に埋め込まれる．Sweaveについての詳細は，VI節において説明する（Leisch（2002），地道，豊原（2018）も参照）．

Jimichi *et al.*（2018）では，ヒストグラムによって，売上高の対数（log（sales））が正規分布よりも若干歪んだ分布に従うという知見を，別の可視化の方法で確認するために，正規Q-Qプロット（図3）が描かれている（Jimichi *et al.*（2018）のFig. 3も参照）．

この結果として，分布の「裾」（tail）の部分で正規分布と若干の隔たりがあり，ヒストグラムから得られた結果が裏付けられている．

なお，実際の可視化は以下のスクリプトをRnwファイルに挿入することによっておこなわれた．

7） デフォルトで出力されるPDFファイルの代わりにPNGファイルを出力した理由は，本稿で扱っている全てのデータをプロットし，プリンアウトなどをおこなう際に，ベクター形式のファイル（PDFファイル）がメモリーなどに関する資源をより多く消費するのに対して，ラスター形式のファイル（PNGファイル）は抑制されるからである．

8） 情報可視化の分野では空間基盤とよばれる．

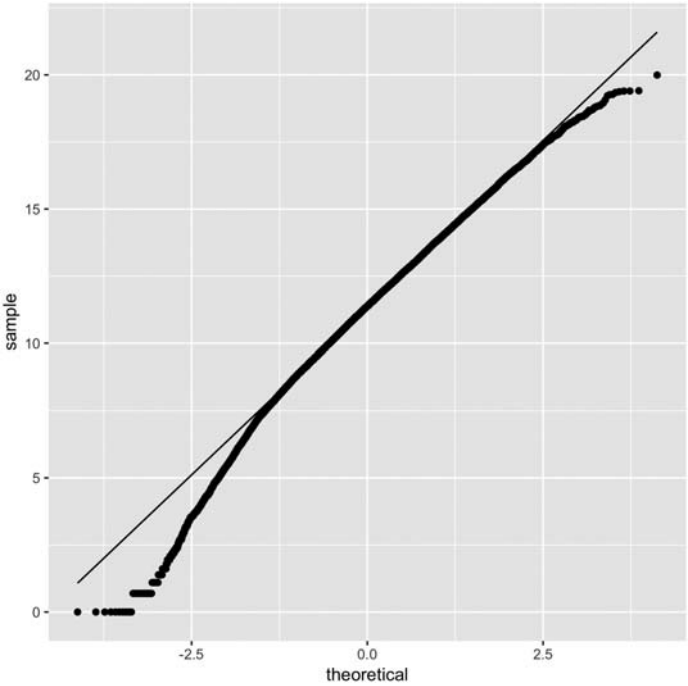

図 3　売上高（対数スケール）の正規 Q-Q プロット

ソースコード 2　売上高（対数スケール）の正規 Q-Q プロットを文書に埋め込むためのスクリプト

```
<<echo=F,fig=T,png=T,pdf=F>>=
ggplot(firmfin2015, aes(sample=log(sales))) +
  stat_qq() + stat_qq_line() +
  labs(x="theoretical",y="sample")
@
```

III　統計モデリング

本節では，前節で得られた売上高の対数をとったものが非対称な分布に従うという知見を使って，どのように統計モデリングがおこなわれたかを解説すると共に，再現性を確保する方法についても述べる．

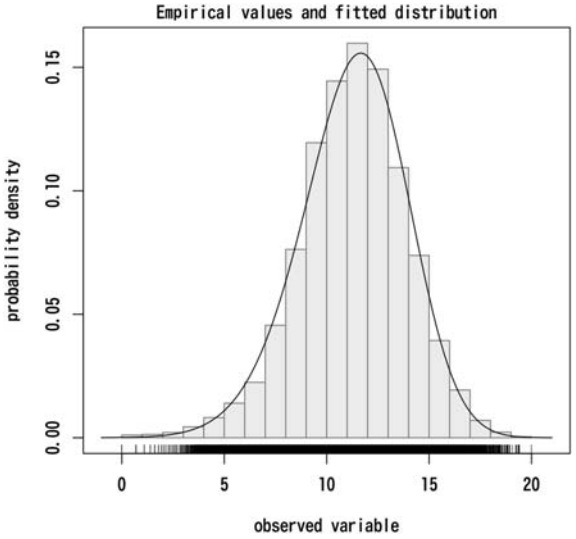

図 4　売上高の対数 **log**(**sales**) のヒストグラムと統計モデル： 非対称正規分布の場合

1．売上高（対数スケール）の分布に対するモデリング

Jimichi *et al.*（2018）では，まず，売上高の対数に，非対称正規分布（skew-normal distribution）SN（ξ, ω^2, α）と，非対称ティー分布（skew-t distribution）ST（ξ, ω^2, α, ν）が当てはめられている[9]．ヒストグラムと統計モデル（確率密度関数の母数を最尤推定値で置き換えたもの）を重ね書きしたものを，それぞれ，図 4 と図 5 に与える（Jimichi *et al.*（2018）の Fig. 4 と Fig. 5 のそれぞれの左パネルも参照）．

これらのプロットの結果を比較すると，図 4（非対称正規分布）よりも図 5（非対称ティー分布）の方が，若干当てはまりの結果が良いように思われる．この点については，IV 節で情報量規準によるモデル選択の観点からも考察を行う．

9）非対称分布に関しては，Azzalini（1985），Azzalini and Capitanio（2014）を参照のこと．

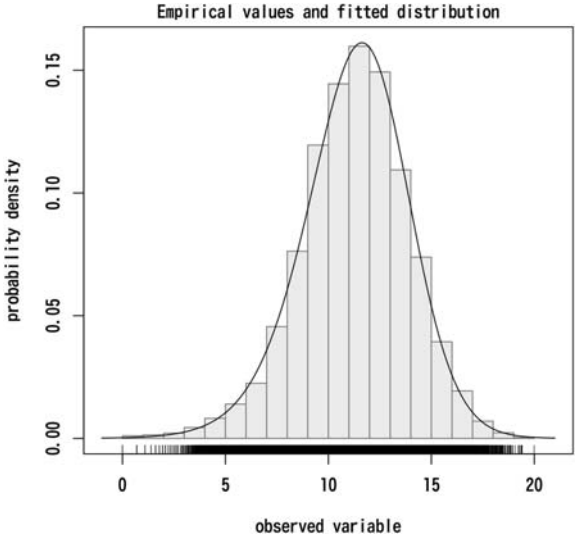

図5　売上高の対数 log(**sales**) のヒストグラムと統計モデル： 非対称ティー分布の場合

なお，実際に可視化をおこなった結果を，Sweave を利用して文書化するために，以下のスクリプトを Rnw ファイルに挿入した：

ソースコード3　ヒストグラムと統計モデルを重ね書きしたものを描くためのスクリプト

```
1 <<echo=F,fig=T>>=
2 library(sn)
3 selm.log.sales2015<-selm(log(sales)~1,data=firmfin2015)
4 plot(selm.log.sales2015,which=2)
5 @
6 <<echo=F,fig=T>>=
7 selm.ST.log.sales2015<-selm(log(sales)~1,family="ST",data=firmfin2015)
8 plot(selm.ST.log.sales2015,which=2)
9 @
```

ソースコード3の2行目において，非対称分布を扱うための R パッケージ **sn**[10] がロードされており，このパッケージに付属する関数 selm[11] を利用して，非対称正規分布（3行目）と非対称ティー分布（7行目）がそれぞれ売

上高の対数（log(sales)）に当てはめられている．なお，plot は総称関数（generic function）であり，オブジェクトのクラスに関する情報を読み取り，適切な方法が適用（メソッドディスパッチ）される．ここでは，selm.ST.log.sales2015，selm.log.sales2015 が selm クラスに属するオブジェクトであるため，plot.selm 関数が自動的に呼び出される．なお，引数 which=2 は，2番目のプロットオプションが与えられていることを表している．この場合は，1変量データへモデル（非対称分布）を当てはめることになっており，統計モデルがヒストグラムへ重ねて描かれている．

2．両対数モデルによる売上高のモデリング

Jimichi *et al.*（2018）の5節では，売上高（sales）を従業員数（employees）と総資産（asstes.total）で説明するために，経済学で古くから研究されている以下のコブ・ダグラス型生産関数（Cobb and Douglas（1928））が仮定されている：

$$\texttt{sales}_i = \gamma \times \texttt{employees}_i^{\alpha_1} \times \texttt{assets.total}_i^{\alpha_2} \times \varepsilon_i, \quad i = 1, \ldots, n \quad (1)$$

このモデルは，両辺の対数をとることによって，以下のように変形できる：

$$\log(\texttt{sales}_i) = \alpha_0 + \alpha_1 \log(\texttt{employees}_i) + \alpha_2 \log(\texttt{assets.total}_i) + \log(\varepsilon_i) \quad (2)$$

ここで，$\alpha_0 := \log \gamma$ とおいた．モデル(2)は，通常の線形回帰モデルであり，モデル(1)の両辺の対数をとることによって，線形モデルになることから両対数モデル（double-log model）と呼ばれる．

さらに，ε_i は誤差であり，誤差の対数 $\log(\varepsilon_i)$ に関して，以下の3つの仮定が検証されている：

正規分布の場合：$\log(\varepsilon_i) \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, \sigma^2)$

10）sn パッケージは，非対称分布研究の第一人者である Adelchi Azzalini 氏によって開発されている．詳細は，http://azzalini.stat.unipd.it/SN/ を参照のこと．

11）関数 selm は，非対称楕円誤差項をもつ線形回帰モデルを当てはめる（fitting linear models with skew-elliptical error term）ための関数である．

非対称正規分布の場合：$\log(\varepsilon_i) \overset{\text{i.i.d.}}{\sim} \mathsf{SN}(0,\ \omega^2,\ \alpha)$
非対称ティー分布の場合：$\log(\varepsilon_i) \overset{\text{i.i.d.}}{\sim} \mathsf{ST}(0,\ \omega^2,\ \alpha,\ \nu)$

ここで，$i=1, \ldots,\ n$, であり，記号“$\overset{\text{i.i.d.}}{\sim}$”は「独立に同一の分布に従う」(independent and identically distributed) を表すことに注意しよう．

これらの誤差分布をもつ線形モデルを当てはめるためのスクリプトは以下のようなものである：

ソースコード4　各種の誤差分布をもつ線形モデルの当てはめをおこなうためのスクリプト

```
1 <<echo=F>>=
2 lm.log.firmfin2015<-lm(log(sales)~log(employees)+log(assets.total), firmfin2015)
3 selm.log.firmfin2015<-selm(log(sales)~log(employees)+log(assets.total), family="SN",data=
      firmfin2015)
4 selm.ST.log.firmfin2015<-selm(log(sales)~log(employees)+log(assets.total), family="ST",
      data=firmfin2015)
5 @
```

ソースコード4における，2，3，4行目で，それぞれ，正規誤差，非対称正規誤差，非対称ティー誤差をもつ線形モデルを当てはめ，その結果をオブジェクトに付値していることに注意しよう．

以下，これらの当てはめた結果の回帰診断を，再現性の視点から確認する．

正規分布の場合

Jimichi *et al.* (2018) の Table 2 には，正規誤差の場合の当てはめの結果として，ティー検定表（表1）が与えられている：

表1　ティー検定表：正規誤差を仮定した場合

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.5803	0.0320	18.13	0.0000
log(employees)	0.4673	0.0045	104.36	0.0000
log(assets.total)	0.6559	0.0040	162.80	0.0000

この結果から，全ての回帰係数は有意であることがわかるけれども，残差の正規 Q-Q プロット（図7）から誤差の正規性が疑われることが指摘され

ている（Jimici *et al.*（2018）の Fig. 6 も参照）.

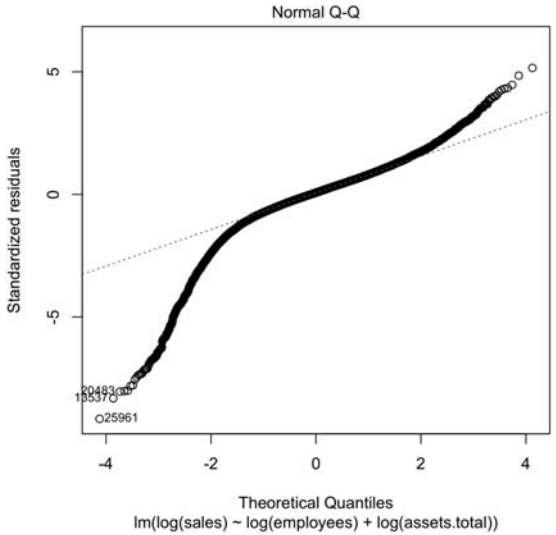

図 6　残差の正規 Q-Q プロット：正規誤差を仮定した場合

これらの結果を得るためのスクリプトは，以下のようなものである：

ソースコード 5　正規誤差をもつ線形モデルを当てはめた結果と回帰診断をおこなうためのスクリプト

```
<<echo=FALSE,results=tex>>=
library(xtable)
print(
xtable(lm.log.firmfin2015,
digits=c(0,4,4,2,4), display=c("s",rep("f",4)),
floating=FALSE,caption=c("ティー検定表:␣正規誤差を仮定した場合"),
label="table.t.log.normal.linear.model"),
caption.placement = "top", table.placement="H",
size="\\setlength{\\tabcolsep}{2pt}")
@
<<echo=F,fig=T,png=T,pdf=F>>=
plot(lm.log.firmfin2015,which=2)
@
```

ソースコード 5 における，コードチャンクのオプション `results=tex` は，結果を TeX（LaTeX）形式で出力するためのものである．このオプションを

xtable パッケージに付属する関数 xtable と併用すると，結果の表を LaTeX 形式で出力してくれる．この機能は，線形モデルの回帰係数に対するティー検定表などを動的に LaTeX ファイルに出力する場合に非常に役立つ．

非対称正規分布の場合

Jimichi *et al.*（2018）の Table 3 には，非対称正規誤差をもつ線形モデルを当てはめた結果の，ゼット比検定表（表２）が与えられている：

表２　ゼット比検定表：非対称正規誤差を仮定した場合

	estimate	std.err	z-ratio	Pr{>|z|}
(Intercept.DP)	1.6644	0.0308	54.08	0.0000
log(employees)	0.3621	0.0047	77.33	0.0000
log(assets.total)	0.7039	0.0040	178.00	0.0000
omega	1.4114	0.0088	160.55	0.0000
alpha	-2.3201	0.0393	-59.04	0.0000

この結果から，全ての母数は有意であることがわかるけれども，残差[12)]の P-P プロット（図７）をみると，誤差の非対称正規性が疑われることが指摘されている（Jimichi *et al.*（2018）の Fig. 9 の右パネルも参照）．

これらの結果を得るためのスクリプトは，以下のようなものである：

ソースコード６　非対称正規誤差をもつ線形モデルを当てはめた結果と回帰診断をおこなうためのスクリプト

```
<<echo=FALSE,results=tex>>=
print(xtable(summary(selm.log.firmfin2015,"DP")@param.table,
digits=c(0,4,4,2,4), display=c("s",rep("f",4)),
floating=FALSE,
caption=c("ゼット比検定表:␣非対称正規誤差を仮定した場合"),
label="table.z.logskewnormal.linear.model"),
caption.placement = "top",table.placement="H",
```

12）より正確には「標準化直接母数残差」（scaled direct parameter residual）である．詳細については，例えば，Azzalini and Capitanio（2014），Jimichi *et al.*（2018）を参照されたい．

```
 8 size="\\setlength{\\tabcolsep}{2pt}")
 9 @
10 <<echo=F,fig=T,png=T,pdf=F>>=
11 plot(selm.log.firmfin2015,param.type="DP",which=4)
12 @
```

ソースコード6も，ソースコード5と同様に，オプション `results=tex` と xtable パッケージに付属する関数 `xtable` を併用することによって，ゼット比検定表を得ていることに注意しよう．

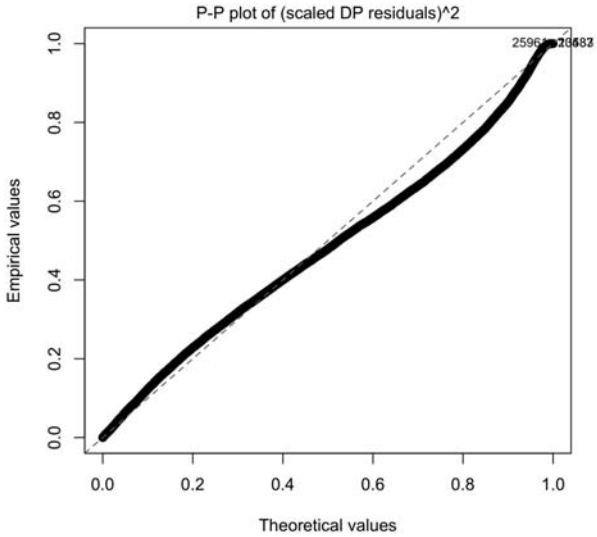

図7　残差の P-P プロット：非対称正規誤差を仮定した場合

非対称ティー分布の場合

Jimichi *et al.*（2018）の Fig. 4 では，非対称ティー誤差をもつ線形モデルを当てはめた結果の，ゼット比検定表（表3）も与えられている：

この結果から，全ての母数は有意であり，残差[13] の P-P プロット（図）をみると非対称ティー分布が誤差分布として当てはまっていることが指摘さ

13）より正確には非対称正規誤差の場合と同様に標準化直接母数残差である．

表 3　ゼット比検定表：非対称ティー誤差を仮定した場合

	estimate	std.err	z-ratio	Pr{>|z|}
(Intercept.DP)	1.3258	0.0288	45.96	0.0000
log(employees)	0.3531	0.0043	81.64	0.0000
log(assets.total)	0.7017	0.0036	195.52	0.0000
omega	0.7637	0.0105	72.40	0.0000
alpha	-1.0210	0.0405	-25.24	0.0000
nu	3.4664	0.0803	43.17	0.0000

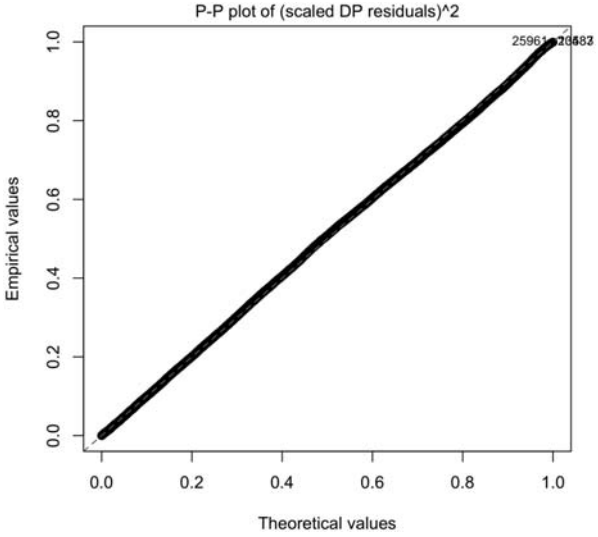

図 8　残差の P-P プロット：非対称ティー誤差を仮定した場合

れている．

これらの結果を得るためのスクリプトは，以下のようなものである：

ソースコード 7　非対称ティー誤差をもつ線形モデルを当てはめた結果と回帰診断をおこなうためのスクリプト

```
<<echo=FALSE,results=tex>>=
print(xtable(summary(selm.ST.log.firmfin2015,"DP")@param.table,
digits=c(0,4,4,2,4), display=c("s",rep("f",4)),
floating=FALSE,
```

```
caption=c("ゼット比検定表:␣非対称ティー誤差を仮定した場合"),
label="table.z.logskew-t.linear.model"),
caption.placement = "top",table.placement="H",
size="\\setlength{\\tabcolsep}{2pt}")
@
<<echo=F,fig=T,png=T,pdf=F>>=
plot(selm.ST.log.firmfin2015,param.type="DP",which=4)
@
```

ソースコード7も，ソースコード5，6と同様に，オプション `results=tex` と **xtable** パッケージに付属する関数 `xtable` を併用することによって，ゼット比検定表を得ている．

Ⅳ　モデル選択

Jimichi *et al.*（2018）では，赤池情報量規準（Akaike Information Criterion: AIC）[14] を利用して様々なモデル選択をおこなっている．

まず，売上高の対数に対しては，正規分布（lm.log.sales2015），非対称正規分布（selm.log.sales2015），非対称ティー分布（selm.ST.log.sales2015）を，それぞれ当てはめたときの結果が表4のように与えられている（Jimichi *et al.*（2018）の Table 5 も参照）．

表4　AIC 表：売上高の対数の分布

	df	AIC
lm.log.sales2015	2	127076.06
selm.log.sales2015	3	126627.07
selm.ST.log.sales2015	4	126546.63

表4における“df”はそれぞれのモデルに含まれる母数の個数を表す．この結果から，非対称ティー分布（selm.ST.log.sales2015）の場合の AIC が最も小さく，このモデルがここで考察されている中で最も良いということが結論づけられている．このことは，図5から得られた当てはまりの良さを視覚的に捉えた結果と整合する．

14）赤池情報量規準については，例えば，Akaike（1973），Konishi and Kitagawa（2008）を参照されたい．

この結果を得るためのスクリプトは，以下のようなものである：

ソースコード 8　売上高の対数に各種の分布を当てはめたときの AIC 表を与えるためのスクリプト

```
<<echo=F,results=tex>>=
lm.log.sales2015<-lm(log(sales)~1,data=firmfin2015)
print(
xtable(AIC(lm.log.sales2015,selm.log.sales2015,selm.ST.log.sales2015),
digits=c(0,0,2), display=c("s","d","f"),
floating=FALSE,caption=c("AIC␣表:␣売上高の対数の分布"),label="table.distribution.AIC",auto=
    TRUE),
caption.placement = "top",table.placement="H")
@
```

ソースコード 8 の 2 行目は，売上高の対数に正規分布を当てはめた結果をベンチマークとして利用するための入力である．

また，両対数モデルの誤差分布に関して，正規分布（lm.log.firmfin2015），非対称正規分布（selm.log.firmfin2015），非対称ティー分布（selm.ST.log.firmfin2015）をそれぞれ仮定した場合について，AIC を比較することによってモデル選択がおこなわれている（Jimichi *et al.*（2018）の Table 6 も参照）．

表 5　AIC 表：両対数モデル

	df	AIC
lm.log.firmfin2015	4	74980.13
selm.log.firmfin2015	5	71972.08
selm.ST.log.firmfin2015	6	67897.56

表 5 より，非対称ティー分布を誤差分布として仮定したモデル（selm.ST.log.firmfin2015）が最も良いと結論づけられている．この結果も，図 8 から得られた視覚的な結果と整合する．

この結果を得るためのスクリプトは，以下のようなものである：

ソースコード 9　売上高の対数に各種の誤差分布をもつ線形モデルを当てはめたときの AIC 表を与えるためのスクリプト

```
<<echo=F,results=tex>>=
print(xtable(AIC(lm.log.firmfin2015,selm.log.firmfin2015,selm.ST.log.firmfin2015),
digits=c(0,0,2), display=c("s","d","f"),
floating=FALSE,caption=c("AIC␣表:␣両対数モデル"),label="table.AIC",auto=TRUE),
caption.placement = "top",table.placement="H")
```

```
@
```

なお，結果として選択されたモデルによる標本回帰平面が，

$$\log(\texttt{sales}) = 0.795 + 0.353\log(\texttt{employees}) + 0.702\log(\texttt{assets.total}) \ (=: \tilde{\eta}_{\mathsf{LSTL}}) \tag{3}$$

となることも与えられている（Jimichi *et al.*（2018）の(8)式を参照）．

この回帰平面の方程式を得るためには，若干技巧的な表記法が必要である．この結果を得るためのスクリプトは，以下のようなものである：

ソースコード10　標本回帰平面の方程式を得るためのスクリプト

```
<<echo=F>>=
coef.selm.ST.log.firmfin2015<-coef(selm.ST.log.firmfin2015,param.type="DP")
bnu<-function(nu) sqrt(nu/pi)*gamma((nu-1)/2)/gamma(nu/2)
delta<-function(alpha) alpha/sqrt(1+alpha^2)
omega.bnu.delta<-function(omega,alpha,nu) omega*bnu(nu)*delta(alpha)
@
\begin{equation}
\Tilde{\eta}_{\mathsf{LSTL}}
=
\Sexpr{round(coef.selm.ST.log.firmfin2015[1] + omega.bnu.delta(omega=coef.selm.ST.log.
    firmfin2015[4],alpha=coef.selm.ST.log.firmfin2015[5],nu=coef.selm.ST.log.firmfin2015
    [6]+1),3)}
+ \Sexpr{round(coef.selm.ST.log.firmfin2015[2],3)} \log (\Robject{employees})
+ \Sexpr{round(coef.selm.ST.log.firmfin2015[3],3)} \log( \Robject{assets.total})
\label{regression.plane.skew-t.adj}
\end{equation}
```

ソースコード10の２行目で，非対称ティー誤差をもつ線形モデルの直接母数（direct parameter: DP）の推定値を抽出した結果を，オブジェクト `coef.selm.ST.log.firmfin2015` に付置している．このオブジェクトは，長さ６のベクトルであり，母数ベクトル（α_0, α_1, α_2, ω, α, ν）の最尤推定値ベクトル（$\hat{\alpha}_0$, $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\omega}$, $\hat{\alpha}$, $\hat{\nu}$）が，それぞれ，成分 `coef.selm.ST.log.firmfin2015[1]`, ..., `coef.selm.ST.log.firmfin2015[6]` に与えられている．

また，３，４，５行目で，標本回帰平面の補正項を計算するための関数

$$b_\nu := b(\nu) := \sqrt{\frac{\nu}{\pi}}\,\frac{\Gamma((\nu-1)/2)}{\Gamma(\nu/2)}, \quad \delta := \delta(\alpha) := \frac{\alpha}{\sqrt{1+\alpha^2}},$$
$$\omega b_\nu \delta = \omega b(\nu)\delta(\alpha)$$

を定義している．

この結果を利用して，10行目から12行目で標本回帰平面の方程式

$$\texttt{log(sales)} = (\hat{\alpha}_0 + \hat{\omega} b_{\hat{\nu}+1} \hat{\delta}) + \hat{\alpha}_1 \texttt{(employees)} + \hat{\alpha}_2 \texttt{log(assets.total)} \; (= \tilde{\eta}_{\mathrm{LSTL}})$$

のパラメータの計算結果を代入するとともに，LaTeX 形式の数式を生成している．ただし，`\Sexpr` は，行中（inline）に R コードの実行結果を埋め込むための Sweave に付属するコマンドである．また，

$$b_{\hat{\nu}+1} = b(\hat{\nu}+1) = \sqrt{\frac{\hat{\nu}+1}{\pi}} \frac{\Gamma(\hat{\nu}/2)}{\Gamma((\hat{\nu}+1)/2)}, \quad \hat{\delta} := \delta(\hat{\alpha}) = \frac{\hat{\alpha}}{\sqrt{1+\hat{\alpha}^2}}$$

である．

なお，データの３次元散布図に，標本回帰平面を当てはめた結果が，図９のように与えられている（Jimichi *et al.*（2018）の Fig. 17 の右のパネルを参照[15]）．

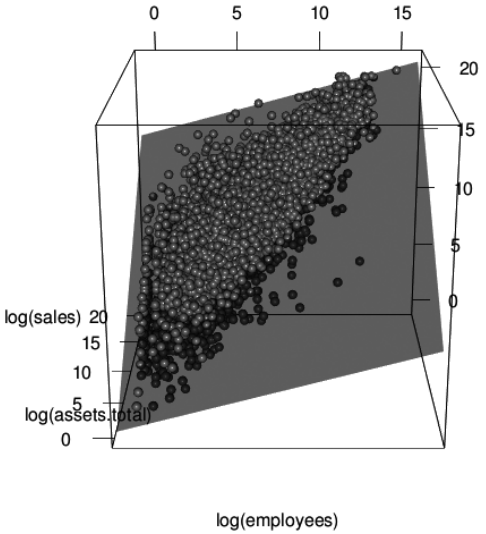

図９　標本回帰平面：非対称ティー誤差をもつ両対数モデルの場合

15）Jimichi *et al.*（2018）では軸は対数スケールとなっていることが前提となっているため，あえて log(·) という表記にはなっていない．

この３次元散布図と標本回帰平面を描き，そのプロットを文書に埋め込む方法も，３次元プロットのスナップショットを一旦ファイルに保存し，再度 LaTeX ファイルに取り組む必要があるため，技巧的な表記法が必要となる．具体的なスクリプトは，以下のようなものである：

ソースコード11　３次元散布図と標本回帰平面を描いたグラフィックを生成するためのスクリプト

```
<<echo=F>>=
library(rgl)
@
<<scatter3dlstlm-adj,echo=F, fig=T,include=FALSE,png=T,pdf=F, grdevice = rgl.Sweave,
    resolution = 100>>=
plot3d(log(firmfin2015[c("employees","assets.total","sales")]),type = "s", col = "red",
    size = 1, xlab="log(employees)",ylab="log(assets.total)",zlab="log(sales)",xlim=c
    (-1,16),ylim=c(-1,20),zlim=c(-1,21))
planes3d(coef.selm.ST.log.firmfin2015[2],
         coef.selm.ST.log.firmfin2015[3],
         -1,
         coef.selm.ST.log.firmfin2015[1]
                        +omega.bnu.delta(omega=coef.selm.ST.log.firmfin2015[4],
                          alpha=coef.selm.ST.log.firmfin2015[5],
                          nu=coef.selm.ST.log.firmfin2015[6]+1),
         alpha=0.5)
@
\setkeys{Gin}{width=0.6\textwidth}
\begin{figure}[H]
\begin{center}
\includegraphics{Ronkyu2018EDA-scatter3dlstlm-adj.png}
\caption{
標本回帰平面：非対称ティー誤差をもつ両対数モデルの場合
}
\label{figure:sample.regression.plane.log.skew.t.linear.model}
\end{center}
\end{figure}
```

ソースコード11の２行目で３次元散布図を描くためのパッケージ rgl がロードされている[16]．また，４行目でコードチャンクのラベルとオプションが与えられている．ここでは，scatter3dlstlm-adj が，コードチャンクのラベルである．このコードラベルとファイル名（Ronkyu2018EDA）をハイフン（-）で結合し，拡張子を png としたもの（Ronkyu2018EDA-scatter3dlstlm-adj.png）が画像ファイルの名称となる．そのほかのオプション

16）ここで，グラッフィクディバイスを読み込む前に，別立てのコードチャンクが用意されている．

の指定は以下のことを表す：

echo=F（R の入力に関するエコーバックを抑制）
fig=T, include=FALSE（プロットを実際に描くが，文書には読み込まない）
png=T, pdf=F（画像のファイルを PDF 形式ではなく PNG 形式で出力）
grdevice = rgl.Sweave（rgl で描かれた3次元プロットのスナップショットを Sweave の文書ファイルに挿入するためのドライバの指定）
resolution = 100（解像度（resolution）が100%であることを指定）

また，5行目で rgl パッケージに付属の関数 plot3d によって3次元散布図が描かれている．また，6行目から13行目で関数 planes3d[17] によって標本回帰平面が描かれている．

15行目から24行目において，出力された画像ファイル Ronkyu2018EDA-scatter3dlstlm-adj.png を LaTeX ファイルに読み込んでいる．

注意 Sweave を利用した通常の画像ファイルの出力と読み込みは，LaTeX の画像ファイルの読み込み環境 \begin{figure}...\end{figure} 中に直接コードチャンク <<>>=...@ を挿入すれば，Sweave 関数で Rnw ファイルを処理した際に，自動的にファイル名が付与されたファイルが出力され，コンパイルの際に自動的に PDF ファイルに読み込まれる（VI 節も参照）．一方，rgl に付属する関数で描いたプロットは，ドライバ rgl.Sweave を利用する関係上，一旦プロットのスナップショットをとったものをファイルとして出力し，再度 PDF ファイルに読み込むように設定するという2段階の

17）関数 planes3d は，(x, y, z) 空間内の平面の方程式
$$ax+by+cz+d=0$$
における係数 (a, b, c, d) を引数に与える仕様となっている．なお，13行目に与えられている alpha は平面の透明度（アルファ値）を指定するための引数であり，ここでは 0.5（半透明）が与えられている．

処理が必要となるため，このような記述となっている．

V　モデル評価

Jimichi *et al.*（2018）では，交差確認（Cross Validation: CV）法[18] によってモデルの予測精度が評価されている．具体的には，K 分割交差確認（K-fold CV）法を用い，乖離関数（discrepancy function）として予測平均 2 乗誤差（Mean Squared Error of Prediction: MSEP）とデータ 1 個あたりの AIC を利用した結果が与えられている．

交差確認法を実行するための R スクリプトを Sweave のコードチャンクに埋め込むことによって，自動実行し，動的に文書を生成することによって再現性を保つことは可能であるが，交差確認法はモンテカルロ法の一種であり，乱数を用いたシミュレーションによって実行されるため以下のことが問題となる：

（CV1）R のデフォルトの設定では乱数の生成に関して再現性が確保できない．
（CV2）文書を生成する度に，ある程度時間を要するシミュレーションを実行する必要があり，非効率的である．

問題（CV1）に対しては，関数 `set.seed` に種（seed）を与えることによって，乱数生成のための再現性を確保することができる．また，問題（CV2）に関しては，別途交差確認法を実施する環境を構築[19] 後，実行した結果を R の作業空間（`CV.RData` ファイル）として保存し，さらに，文書化する際にこのファイルを動的にロードし利用するという方法が考えられる．ただし，この方法は，再現性を確保するために，R のスクリプトとデータ，そして作

18）交差確認法についての詳細は，Efron and Tibshirani (1993), James et al. (2013), Efron and Hastie (2016) などを参照されたい．
19）本研究では，交差確認法をおこなうために，RStudio のプロジェクトを作成している．

業空間をこの環境下で厳密に管理する必要がある[20].

本稿では，交差確認法の実施に伴うシミュレーションの詳細は割愛するが，自動実行するための手順に関するイメージを図10に与える．

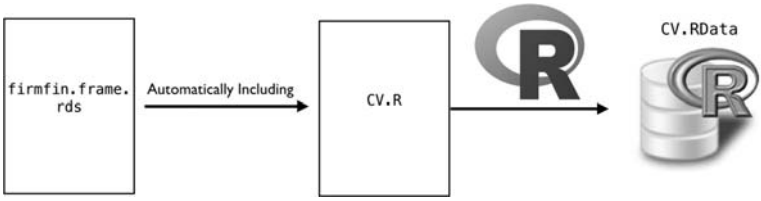

図10　交差確認法の自動実行

図10で実行されている流れを以下に説明する：

(S-CV-1) R データファイル `firmfin.frame.rds` をロード
(S-CV-2) 交差確認法を実行するため関数の定義
(S-CV-3) シュミュレーションの実行
(S-CV-4) 結果を作業空間 `CV.RData` として保存

ここで，地道（2018-a）で述べた前処理によって得られた R データセットファイル `firmfin.frame.rds` をロードしている．

以上の処理に必要な一連の R スクリプトをファイル `CV.R` に記述しておき，UNIX コマンド `Rscript` を使って自動実行することによって，作業空間のファイル `CV.RData` を出力している．ターミナルから実際に実行するコマンドは以下のようなものである：

`Rscript` コマンドの実行

```
$ Rscript CV.R
```

20）このような方法は，シミュレーションの結果を含む文書を作成するときにも有効と思われる．

ここで，$ はシェルのプロンプトである．

Jimichi *et al.*（2018）では，このような手順で生成された作業空間 CV.RData を動的に読み込み，10（$=K$）分割交差確認法を実行することによって得られた MSEP 値のボックスプロット（図11）が与えられている（Jimichi *et al.*（2018）の Fig. 13 も参照）．

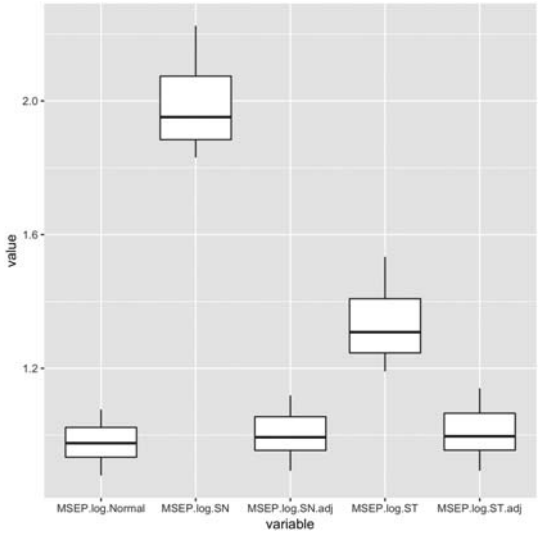

図11　MSEP 値のボックスプロット：$K=10$

ここで，各モデルのラベルの説明を表 6 に与える（Jimichi *et al.*（2018）も参照.）

表 6　モデル対応表

モデルラベル	誤差分布	予測式
MSEP.log.Normal	正規分布	通常
MSEP.log.SN	非対称正規分布	通常（直接母数）
MSEP.log.SN.adj	非対称正規分布	修正（中心母数）
MSEP.log.ST	非対称ティー分布	通常（直接母数）
MSEP.log.ST.adj	非対称ティー分布	修正（疑似中心母数）

Jimichi *et al.*（2018）では，誤差分布として正規分布を仮定した場合（MSEP.log.Normal）と，非対称正規分布を仮定し，中心母数の推定値から作られた予測式を利用する場合（MSEP.log.SN.adj），そして，非対称ティー分布を仮定し，中心母数の修正推定値から作られた予測式を利用する場合（MSEP.log.ST.adj）の予測精度はほぼ変わらないことが指摘されている．

図11を文書に埋め込むためのスクリプトは，以下のようなものである：

ソースコード12　交差確認法による MSEP 比較のプロットを文書に埋め込むためのスクリプト

```
1 <<echo=F,fig=T,png=T,pdf=F>>=
2 load("CV.RData")
3 library(reshape)
4 ggplot(melt(summary.CV.k(CV.k.firmfin2015.10)$MSEP.log),aes(variable,value))+geom_boxplot
      ()
5 @
```

ソースコード12の２行目において，交差確認法を実行した結果の作業空間（CV.RData ファイル）が読み込まれている．次に，３行目でパッケージ **reshape**[21] を読み込んだ後，４行目において，表６におけるモデルに関する MSEP の分布を確認するためにボックスプロットが描かれている．

さらに，Jimichi *et al.*（2018）では，AIC にもとづく評価として，10（$=K$）分割交差確認法を実行して得られる AIC の値のボックスプロットを描いたものが図12のように与えられている（Jimichi *et al.*（2018）の Fig. 14 も参照）．ここで，各モデルのラベルの説明は表７のようなものである．この結果から，誤差分布として非対称ティー分布を仮定し，母数を最尤法よって推定した場合（AIC.log.ST）が最も良いことが報告されており，誤差分布として非対称ティー分布を仮定したモデルが最も妥当であることが結論づけられている．

21）reshape は，データフレームの構造（正規化，非正規化）を変換するための関数が用意されたパッケージである．正規化されたデータフレームを非正規化するための関数 **melt** と，逆に非正規化されたデータフレームから正規化するための関数 **cast** が用意されている．なお，このパッケージの改良版として，tidyr パッケージが利用できる．ここでは，データフレームのサイズが小さいため，標準的な reshape パッケージを利用している．

表 7　モデル対応表

モデルラベル	誤差分布	母数推定法
AIC.log.Normal	正規分布	最小自乗法
AIC.log.SN	非対称正規分布	最尤法（直接母数）
AIC.log.ST	非対称ティー分布	最尤法（直接母数）

図12を文書に埋め込むためのスクリプトは，以下のようなものである：

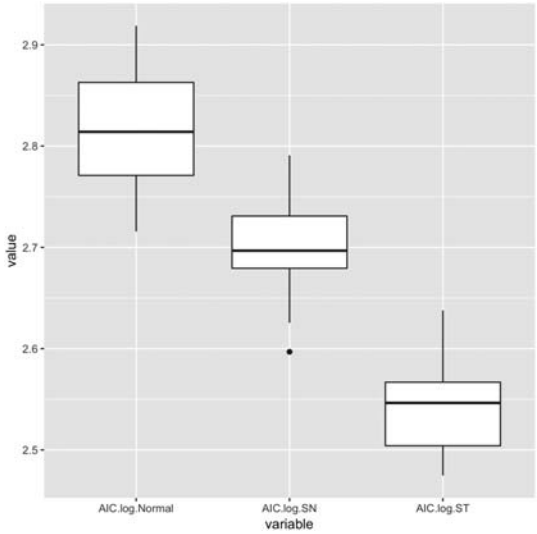

図12　AIC 値のボックスプロット：$K=10$

ソースコード13　交差確認法による AIC 比較のプロットを文書に埋め込むためのスクリプト

```
1 <<echo=F,fig=T,png=T,pdf=F>>=
2 ggplot(melt(summary.CV.AIC.k(CV.AIC.k.firmfin2015.10)$CV.AIC),aes(variable,value))+
      geom_boxplot()
3 @
```

ソースコード13の 2 行目において，表 7 におけるモデルに関する AIC の分布を確認するためにボックスプロットが描かれている．

VI 動的文書と再現可能研究

一般に，データ解析の結果を含む論文・レポートなどを作成する際に，データ解析と文書作成のそれぞれの工程は基本的には「平行」に進んでおり，前者の工程で得られた結果（図，表，テキストなど）を文書作成の段階でマニュアル操作（「手作業」）によって（静的に）コピー・アンド・ペーストすることが標準的な方法と考えられる．ここでは，このような方法を「静的文書」（static document）作成とよぶことにする（図13も参照）．

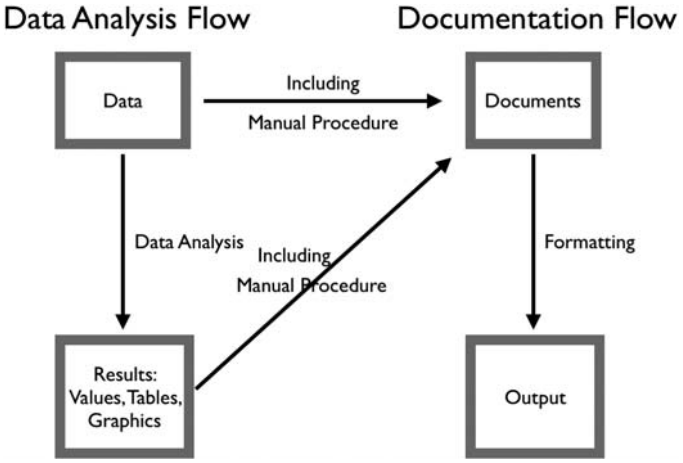

図13 静的文書生成

ただし，この方法は以下のような問題をもつことが容易にわかる：

- もしデータが刷新された場合は，もう一度「手作業」で図や結果を貼り直す必要がある．
- 再度，全く同一のものを作成することが困難である．

特に，後者の問題はデータ解析と文書作成の両方の工程の詳細な記録をとっていたとしても，細部での再現性を確保することが，非常に難しいことを意

味する．また，近年その重要性が指摘されている「再現可能研究」（reproducible research）の観点からも，この問題とどのように向き合うかはデータ解析に携わる全ての人に与えられた課題の一つといえよう（例えば，Gandrud（2015）を参照）．

そこで，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（図，表，テキストなど）を動的に生成した後，さらにそれらを自動的に読み込んだ文書を作成することができれば，このような問題を軽減することが可能となろう（図14を参照）．このような文書生成法は，近年，「動的文書」（dynamic document）生成とよばれるようになっている[22]（例えば，Xie（2015）を参照）．

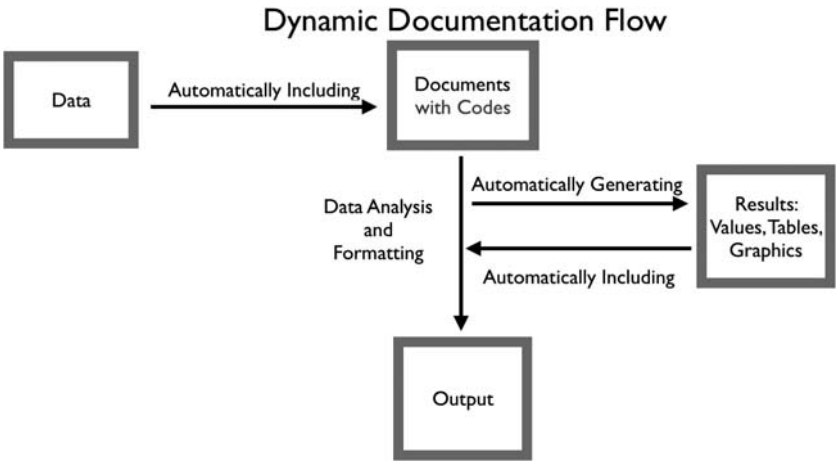

図14　動的文書生成

動的文書生成を利用することによって，再現可能研究を実現するためのツールが近年いくつか開発されている[23]．本研究では，Norman Ramsey による noweb[24] をベースとして Leisch（2002）によって開発された Sweave を利用

22）動的文書に関する原典は，Knuth（1984）であるといわれる．
23）本稿では，Sweave を動的文書を生成するために利用しているが，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．

した．Sweave の利用手順は，以下のようなものである：

（S1）R のコードを LaTeX ファイルに埋め込んだ Rnw ファイルを作成
（S2）R に標準で付属する関数 Sweave で処理

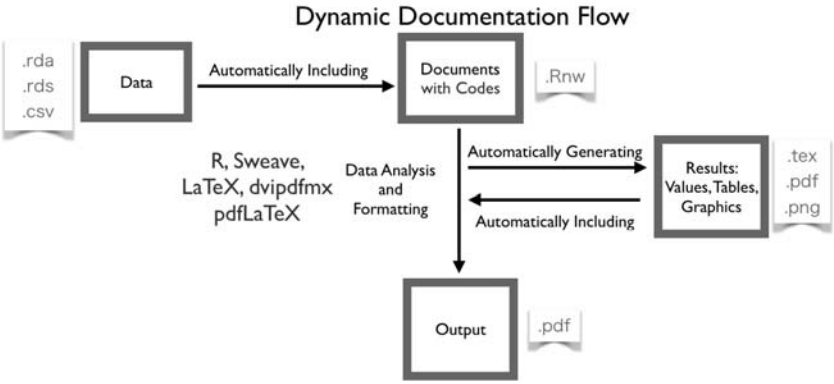

図15　Sweave による動的文書生成

このような処理によって，R によって出力された値や図表のファイルと LaTeX ファイルが出力されるので，LaTeX のコンパイラ（例えば，platex）とドライバ（例えば，dvipdfmx[25]）で処理することによって，論文などの PDF ファイルを作成することができる[26]（図15も参照）．なお，Sweave の利用上の詳細については，地道，豊原（2018）も参照されたい．

Jimichi *et al.*（2018）を作成する工程では，試行錯誤が必要となるため，macOS 上の TeX の統合環境である TeXShop[27] 用に，Sweave と LaTeX による処理を自動化する以下のようなシェル・スクリプト・ファイルを作成し

24）https://www.cs.tufts.edu/~nr/noweb/
25）dvipdfmx は，LaTeX ファイルをコンパイルすることによって得られる DVI ファイルを PDF ファイルに変換するためのソフトウェアである．
26）pdfLaTeX を利用すれば，LaTeX ファイルから直接 PDF ファイルへ変換することができる．ただし，日本語を含む文書を処理するためには，煩雑な設定が必要となるため，本稿では platex と dvipdfmx を利用した．
27）https://pages.uoregon.edu/koch/texshop/texshop.html

た：

ソースコード14　Sweave と LaTeX による処理を自動化するシェル・スクリプトファイル：**Sweave-utf8.engine**

```
#!/bin/sh
export LANG=ja_JP.UTF-8
export PATH=$PATH:/Library/TeX/texbin
R CMD Sweave --encoding="utf8" "$1"
filename=${1%.*}
ptex2pdf -l -ot "-synctex=1␣-file-line-error" "$filename"
ptex2pdf -l -ot "-synctex=1␣-file-line-error" "$filename"
```

ソースコード14の１行目から，このファイルがシェルスクリプトであることがわかる[28]．また，２行目の設定は，環境変数 LANG に ja_JP.UTF-8 を与えることによって，ロケールを設定している．３行目では TeX 関連のパスを設定し，４行目では，コマンドラインから R の関数 Sweave を実行する設定を行っている．なお，オプション -encoding="utf8" は，ファイルの文字コードを UTF-8 として出力するためのものである．さらに，５行目で処理対象となるファイルの名前から拡張子を除いた文字列を切り出したあと，６，７行目で ptex2pdf コマンドでコンパイルとドライバによる処理（タイプセット）を行っている．なお，最終的には PDF ファイルが出力される．

次に，このファイルをユーザのホームディレクトリにある TeXShop 用の個人設定ファイルが納められたディレクトリ ~/Library/TeXShop/Engines/ に保存し，TeXShop から利用できるように設定した．さらに，このファイルが，単独のスクリプトとしても実行できるように，chmod コマンドを以下のように利用することによって実行権限を与えておく必要がある：

28）正確には，shebang（シバン）とよばれ，この行が与えられたファイルに実行ビットを付与して実行すると，指定されたインタプリターによって処理が行われる．詳細は，Janssens（2014），木本他（2018）などを参照されたい．

実行権限の付与

```
$ cd ~/Library/TeXShop/Engines/
$ chmod +x Sweave-utf8.engine
```

実行権限に関する詳細については，Janssens（2014），木本他（2018）などを参照されたい．

つぎに，以下のような手順によって文書を作成した：

（J1）TeXShop のコンパイラのメニューから `Sweave-utf8` を選択
（J2）`Rnw` ファイル（`paper.Rnw`）を作成
（J3）ファイルの修正
（J4）タイプセット

以上の手順のうち，（J3）と（J4）は文書の完成まで繰り返し行うことになる．TeXShop でタイプセットするためのイメージは，図16を参照されたい．また，動的文書作成の処理を概念的に説明したものを図17に与える．

図16 TeXShop によるタイプセット：コンパイラのメニューから **`Sweave-utf8`** を選択後，タイプセット ボタンをクリックすることによって処理をおこなう

図17のタイプセットが行われる工程は，以下のように説明される：

（TS1）Sweave によって `paper.Rnw`（`Rnw` ファイル）を処理することによっ

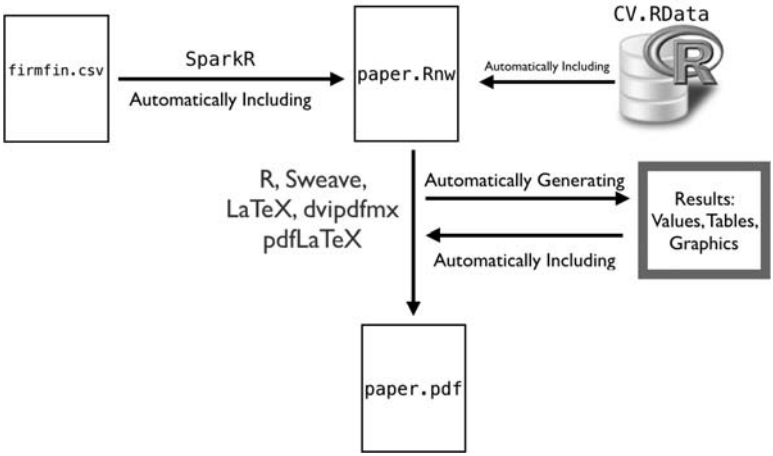

図17　Sweave による動的文書生成の概念図

て paper.tex（LaTeX ファイル）へ変換

(a) SparkR パッケージを使ってデータファイル firmfin.csv を Spark に firmfin.sdf（Spark DataFrame）として読み込む

(b) firmfin.sdf を R のデータフレーム firmfin2015 に変換

(c) 交差確認法の実行結果（作業空間 CV.RData）の読み込み

(d) paper.Rnw から paper.tex へ変換しながら，推定結果等の値や表を paper.tex に埋め込む

(e) 各種のプロットの結果を PNG ファイル（画像ファイル）に出力し，読み込むための情報を paper.tex へ埋め込む

(TS2) ptex2pdf によって paper.tex を処理することによって paper.pdf へ変換

(a) platex によって paper.tex をコンパイルし，DVI ファイル paper.dvi へ変換

(b) dvipdfmx によって，DVI ファイルから PDF ファイル paper.pdf に変換

これらの工程のうち，(TS1-a, b) は，地道 (2018-a) でデータラングリング (data wrangling) とよばれており，それ以降の部分を含む全ての工程が，ここで説明しているように自動実行されていることに注意しよう (地道 (2018-a) の IV 節を参照)．以上の処理手順 (J1)～(J4) は，TeXShop という macOS 上のアプリケーションを利用して動的に文書を生成し，再現可能性を確保する方法である．

次に，UNIX のシェル環境を利用し，さらにメタなレベルで文書生成を自動化する方法を考える．実際に利用したツールは，地道 (2018-a) でも利用した UNIX コマンド make[29] である．本研究のために以下の Makefile を用意した：

ソースコード15　Makefile

```
1  all:
2       /bin/sh ./script.sh
3       Rscript datadump.R "data.rda" "name.rda" "firmfin.csv" "firmfin.frame.rds"
4       Rscript CV.R
5       ~/Library/TeXShop/Engines/Sweave-utf8.engine paper.Rnw
6  csv:
7       /bin/sh ./script.sh
8       Rscript datadump.R "data.rda" "name.rda" "firmfin.csv" "firmfin.frame.rds"
9  RData:
10      Rscript CV.R
11 paper:
12      ~/Library/TeXShop/Engines/Sweave-utf8.engine paper.Rnw
13 paper-without-CV:
14      /bin/sh ./script.sh
15      Rscript datadump.R "data.rda" "name.rda" "firmfin.csv" "firmfin.frame.rds"
16      ~/Library/TeXShop/Engines/Sweave-utf8.engine paper.Rnw
17 preview-paper:
18      open paper.pdf
19 clean-tex:
20      rm paper-* *.out *.log *.tex *.aux
21 clean-data:
22      rm *.rda *.rda-e *.part
23 clean-pdf:
24      rm *.pdf
```

まず，図17で説明したタイプセットを実行するための，ソースコード15の11, 12行目を参照されたい．一般に，これらの 2 行は「ルール」とよばれ，11行

29) より正確には，Gnu Make (https://www.gnu.org/software/make/) である．

目（paper:）がターゲットとよばれる部分である．また，12行目がソースコード14で与えられたシェルスクリプト Sweave-utf8.engine を利用して処理が行われることを記述していることに注意しよう．

make コマンドの利用法は，macOS のターミナル上で，ファイル一式が収められているディレクトリに移動後，ターゲット名付きで make コマンドを以下のように実行する：

make コマンドの実行：ターゲット paper

```
$ make paper
```

この入力によって，図17で表されるタイプセットが実行される．

以上の説明で，本稿で述べた探索的データ解析の結果を文書化する工程を動的に行い，再現性を確保することは可能となったものと思われるが，さらに，地道（2018-a）で述べた粗データの前処理を行うことや，V 節で述べた交差確認法によるシミュレーションを実行することも make コマンドを利用することによって可能となる．つまり，ソースコード15の６，７，８行目のルールを利用すると，前処理が実行できる（地道（2018-a）の図９を参照）：

make コマンドの実行：ターゲット csv

```
$ make csv
```

また，９，10行目のルールを利用すると，V 節で議論した交差確認法を実行した結果の作業空間 CV.RData を動的に生成することができる（図10を参照）：

make コマンドの実行：ターゲット RData

```
$ make RData
```

既に，作業空間 CV.RData が得られていれば，交差確認法のシミュレーションの工程をバイパスして，粗データの前処理から始めて，探索的データ解析を実行し，その結果の PDF ファイルを動的に生成するためには，13〜

16行目のルールを利用すればよい（図18参照）：

make コマンドの実行：ターゲット paper-without-CV

```
$ make paper-without-CV
```

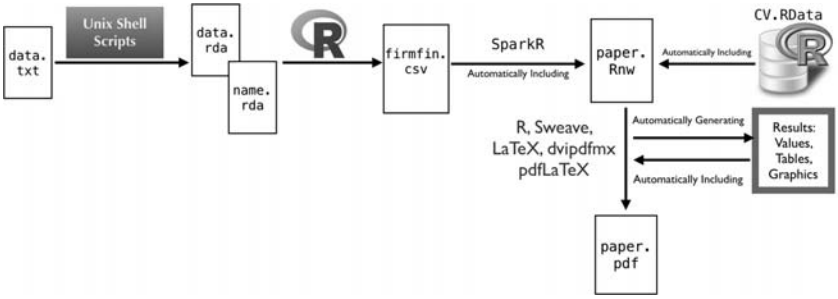

図18　粗データの前処理を含む Sweave による動的文書生成の概念図：交差確認法のシミュレーションを実行しない場合

最後に，粗データの前処理から始めて，探索的データ解析を実行し，交差確認法の実行結果の作業空間を生成した後，PDF ファイルを動的に生成する工程を全て（この意味で all）行うためのには，１～５行目のルールを利用すればよい（図19参照）：

make コマンドの実行 ：ターゲット all

```
$ make all
```

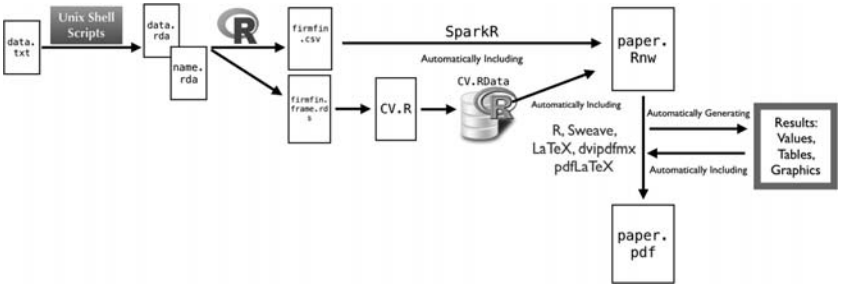

図19　全ての工程を動的生成した場合の概念図

以上の工程を処理するために必要な時間は，現時点で筆者が常時利用している環境[30]では，全ての工程（ターゲット `all`）で30分程度であり，交差確認法のシミュレーションを実行しない場合（ターゲット `paper-without-CV`）で10分程度である．

Ⅶ　おわりに

本稿では，ビッグデータという通常のソフトウェア環境では扱いにくい対象に対して，Jimichi *et al.*（2018）で与えられた探索的データ解析の結果を導く工程を例として，再現可能研究を実現する試みについて述べた[31]．図20は，地道（2018-a）で扱った前処理とデータラングリングも含めて，本研究でおこなった探索的財務ビッグデータ解析を実現するために利用した環境を簡単に説明したものである．

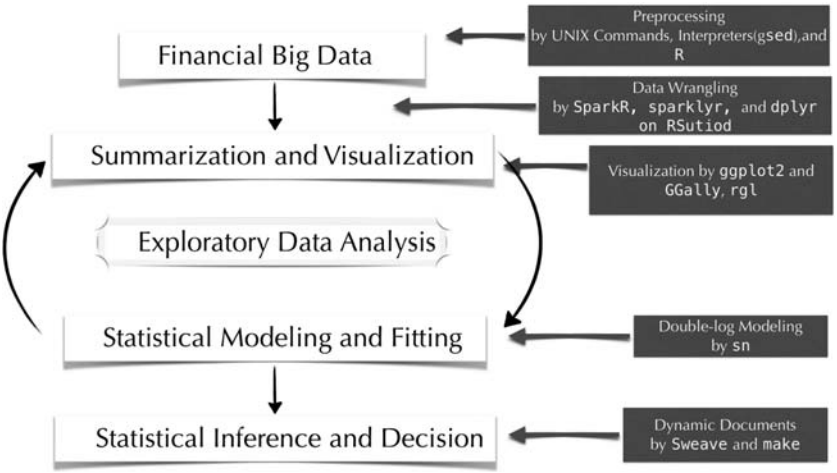

図20　再現性を確保し，探索的財務ビッグデータ解析を実行するための環境

30）MacBook Pro（15-inch, 2018, プロセッサ 2.9 GHz Intel Core i9, Storage: SSD）であり，OS は macOS High Sierra（バージョン 10.13.6）である．

31）地道（2018-b）も Sweave を利用することによって作成されている．このことから，書籍に関しても動的文書によって再現可能性を確保することが可能であることがわかる．

なお，非対称ティー誤差をもつ両対数モデルが，売上高を従業員数と総資産によって説明するために利用できるということが，探索的財務ビッグデータ解析によって得られた知見である（Jimichi *et al.*（2018）を参照）．

再現可能研究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，論文などを「公表しただけのもの」（publication only）は再現可能ではない（not reproducible）という位置付けであり，公表に加えて，「コードが管理されているもの」（code），「コードとデータが管理されているもの」（code and data），さらに「リンク情報があり，コードとデータを実行できるもの」（linked and executable code and data）の順に再現可能性のレベルが上がっていき，「完全に再現するもの」（full replication）がゴールドスタンダード（gold standard）であるとされている．このスペクトルに照らすと，本研究は make によって完全自動実行することによっ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ものと思われる．

ただし，ソフトウェア環境は常に変化していることには注意が必要である．例えば，近年の動向では macOS（オペレーティングシステム）は１，２年に１回，データ解析環境 R は，１年に１回マイナーバージョンがアップする．このアップデートにより，コマンドや関数の引数に関する仕様変更がある場合もあり，以前は問題が無かったコードが実行できなくなったり，誤動作する場合もある．このことから，時間が変化しても，安定的に利用できる環境を確保するために，常時コンピュータ及びソフトウェア環境の動向を調査し，把握しておく必要があろう．

さらに，データ解析・文書作成が「手元」でおこなわれている間は，再現性を確保できていても，原稿を出版社へ入稿し，出版に向けての校正の段階に移ると，文書ファイルは印刷所で管理されることから，厳密な意味での再現性は失われる可能性がある．この場合でも，元原稿のファイルを校正にあわせて修正することは必要となる．

最後に，今後の課題を述べる．現在扱っているさらに規模の大きなデータ

セットは，BvD 社から提供されているデータベース Orbis から世界の全上場・非上場企業の（1）連結企業（consolidated firm）主体に抽出された22,312,669社と（2）非連結企業（un-consolidated firm）主体に抽出された22,304,556社の主要財務情報（売上高，営業利益，総資産など）を最長10年分抽出したものである．サイズとしては，テキスト形式のファイルで，それぞれ，120 GB 程度であり，通常のコンピュータ環境ではデータ容量がメモリー容量を超えるため扱うことが難しいことに注意しよう．この問題に対して，東京大学情報基盤センターに設置された専有利用型リアルタイムデータ解析ノード（FENNEL）を利用し，本稿で考察した同様の研究が実行可能かどうか検証をおこなっている．なお，詳細は割愛するが，地道（2018-a）と本稿でおこなった全ての工程は，FENNEL 環境のもとでも再現可能であることは検証済みである．

（筆者は関西学院大学商学部教授）

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 Caski, F.（eds.）, Akadimiai Kiado, Budapest：pp. 267-281.

［2］ Azzalini, A.（1985）A class of distributions which includes the normal ones, *Scandinavian Journal of Statistics,* Vol. 12, No. 2, pp. 171-178.

［3］ Azzalini, A. with the collaboration of A. Capitanio（2014）*The Skew-Normal and Related Families,* Cambridge University Press, Institute of Mathematical Statistics Monographs.

［4］ Cobb, C. W. and P. H. Douglas（1928）A theory of production, *American Economic Review,* Vol. 18, pp. 139-165.

［5］ Efron, B. and T. Hastie（2016）*Computer Age Statistical Inference: Algorithms, Evidence, and Data Science,* Cambridge University Press.

［6］ Efron, B. and R. J. Tibshirani（1993）*An Introduction to Bootstrap,* Chapman and Hall/CRC.

［7］ Gandrud, C.（2015）*Reproducible Research with R and RStudio,* Second Edition, CRC Press.

［8］ James, G., D. Witten, T. Hastie, and R. Tibshirani（2013）*An Introduction to Statistical Learning with Applications in R,* Springer.

（邦訳：落海浩，首藤信通共訳（2018）『R による統計的学習入門』，朝倉書店.）
［9］ Janssens, J.（2014）*Data Science at the Command Line,* O'Reilly Media.（太田満久，下田倫大，増田泰彦監訳，長尾高弘訳（2015）『コマンドラインではじめるデータサイエンス：分析プロセスを自在に進めるテクニック』，オライリー・ジャパン.）
［10］ 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，61巻，3 号，pp. 241-295.
［11］ 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. 159-185，2017年 3 月，関西学院大学商学研究会.
［12］ 地道正行（2017-b）『R を利用した対数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. 141-174，2017年 9 月，関西学院大学経済学部研究会.
［13］ 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65巻，第 1 号，pp. 1-31，2018年 9 月，関西学院大学商学研究会.
［14］ 地道正行（2018-b）『データサイエンスの基礎：R による統計学独習』，裳華房.
［15］ 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,* https://doi.org/10.1007/s42081-018-0019-1, First Online: 10 September 2018, in printing.
［16］ 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』，豊原法彦編著『関西経済の構造分析』，第 5 章，pp. 77-111，中央経済社.
［17］ 木本雅彦，松山直道，稲島大輔共著，株式会社創夢監修（2018）『はじめて UNIX で仕事をする人が読む本』，アスキードワンゴ.
［18］ Konishi, S. and G. Kitagawa（2008）*Information Criteria and Statistical Modeling,* Springer.
［19］ Knuth, D. E.（1984）Literate programming, *The Computer Journal, British Computer Society,* Vol. 27, No. 2, pp. 97-111.
［20］ 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.
［21］ Peng, R. D.（2011）Reproducible research in computational science, *Science,* Vol. 334, pp. 1226-1227.
［22］ 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.
［23］ 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.
［24］ Tukey, J. W.（1977）*Exploratory Data Analysis,* Addison-Wesley Publishing Co.

[25] Unwin, A. (2015). *Graphical Data Analysis with R,* Chapman and Hall/CRC.
[26] Xie, Y. (2015) *Dynamic Documents with R and knitr, Second Edition,* CRC Press.
[27] Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis, Second Edition,* Springer.
(初版邦訳：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.)
[28] Wickham, H. and G. Grolemund (2016) *R for Data Science,* O'Reilly.
[29] Wilkinson, L. (2005) *The Grammar of Graphics, Second Edition,* Springer.

謝辞

筆者は，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 氏から学んだ．彼からの教示がなければ，動的文書や，再現可能性の重要性を意識した研究を行うことはなかったことと思う．ここに感謝の意を表したい．

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，研究代表者：阪智香
- 平成29年度学際大規模情報基盤共同利用・共同研究拠点（JHPCN）課題：「財務ビッグデータの可視化と統計モデリング」，課題番号：jh171002-NWJ，研究代表者：地道正行
- 平成30年度学際大規模情報基盤共同利用・共同研究拠点（JHPCN）課題：「財務ビッグデータの可視化と統計モデリング」，課題番号：jh181001-NWJ，研究代表者：地道正行
- 2018年度関西学院大学研究装置・設備等整備計画 III，研究代表者：阪智香
- 関西学院大学図書館図書費 B

また，BvD 社の増田歩氏にはデータの抽出に関して多大なるご協力いただいた．ここに感謝を申し上げる．

付録

環境

本研究を行うために主に利用した環境を以下に与える：

ハードウェア環境

- iMac Pro:

Processor: Intel Xeon W 2.3 GHz
Cores: 18
Main Memory: 128 GB
OS: macOS High Sierra

- MacBook Pro:
  Processor: Intel Core i9 2.9 GHz
  Cores: 6
  Main Memory: 32 GB
  OS: macOS High Sierra

ソフトウェア環境

- R (R. Ihaka, R. Gentleman, R Core Team, https://www.r-project.org/)
- R Packages
  - dplyr (H. Wickham, http://dplyr.tidyverse.org/)
  - GGally::ggpairs (B. Schloerke, http://ggobi.github.io/ggally/)
  - ggplots2 (H. Wickham, http://had.co.nz/ggplot2/)
  - magrittr (H. Wickham, https://github.com/tidyverse/magrittr)
  - rgl (D. Murdoch, https://cran.r-project.org/web/packages/rgl/vignettes/rgl.html)
  - sn (A. Azzalini, http://azzalini.stat.unipd.it/SN/)
  - SparkR (http://spark.apache.org/)
  - xtable (D. B. Dahl, http://xtable.r-forge.r-project.org/)
- RStudio (RStudio, https://www.rstudio.com/)
- Spark 2.3.1 (http://spark.apache.org/)
- Sweave (F. Leisch, https://leisch.userweb.mwn.de/Sweave/)

R 関連の環境の詳細は以下のようなものである：

R に関する環境

```
> sessionInfo()
R version 3.5.1 (2018-07-02)
Platform: x86_64-apple-darwin15.6.0 (64-bit)
Running under: macOS High Sierra 10.13.6

Matrix products: default
BLAS: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRlapack.dylib

locale:
[1] ja_JP.UTF-8/ja_JP.UTF-8/ja_JP.UTF-8/C/ja_JP.UTF-8/ja_JP.UTF-8

attached base packages:
```

```
[1] stats4    stats     graphics  grDevices utils     datasets
[7] methods   base

other attached packages:
[1] reshape_0.8.7   rgl_0.99.16    xtable_1.8-3    sn_1.5-2
[5] ggplot2_3.0.0   bindrcpp_0.2.2 dplyr_0.7.6

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.19            later_0.7.5
 [3] pillar_1.3.0            compiler_3.5.1
 [5] plyr_1.8.4              bindr_0.1.1
 [7] tools_3.5.1             digest_0.6.17
 [9] jsonlite_1.5            tibble_1.4.2
[11] gtable_0.2.0            pkgconfig_2.0.2
[13] rlang_0.2.2             shiny_1.1.0
[15] crosstalk_1.0.0         knitr_1.20
[17] withr_2.1.2             htmlwidgets_1.3
[19] webshot_0.5.1           manipulateWidget_0.10.0
[21] grid_3.5.1              tidyselect_0.2.4
[23] glue_1.3.0              R6_2.3.0
[25] purrr_0.2.5             magrittr_1.5
[27] promises_1.0.1          scales_1.0.0
[29] htmltools_0.3.6         assertthat_0.2.0
[31] mnormt_1.5-5            mime_0.6
[33] colorspace_1.3-2        httpuv_1.4.5
[35] numDeriv_2016.8-1       labeling_0.3
[37] miniUI_0.1.1.1          lazyeval_0.2.1
[39] munsell_0.5.0           crayon_1.3.4
```