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―

地　道　正　行

I　はじめに

近年，情報通信技術（Information Communication Technology: ICT）のめざましい発達はインターネットを利用したビジネスの台頭を生み，その市場の拡大はめざましいものがある．このような動向の背景には，巨大ストレージ，高速通信網，コンピュータの高性能化とソフトウェアの高機能化などの総合的な環境の発展があることは容易に想像できよう．この発展の副産物として「データの爆発」とも揶揄される現象が生み出され，大量のデータが様々な分野でリアルタイムに蓄えられることによって，近年「ビッグデータ」という用語で呼ばれるようになっている．この現状のもとで，データ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の一つといえよう[1)]．

この課題に対して，Tukey（1977）によって提唱された探索的データ解析（Exploratory Data Analysis: EDA）は一つの重要なアプローチを提供する．つまり，EDA はモデルを構築する前にデータを数値的に要約（summarization）したり，図式的に可視化（visualization）することによってデータ自身のもつ情報を探索的に引き出し，それにもとづいた統計モデリング（statistical modeling）を行い，データに当てはめ（fitting），さらにその結果を可視

1)　近年，Google が開発したインフルエンザの予測モデル（http://www.google.org/flutrends/）は，この課題に対する「解答」の一つといえよう．

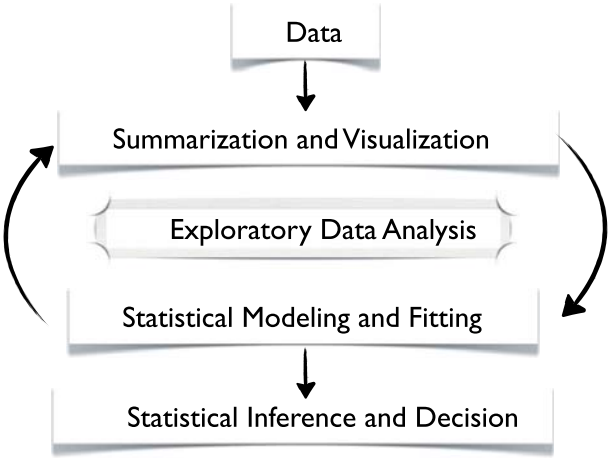

図 1　探索的データ解析とそれを取りまく環境に関する概念図

化することによってモデルを改良するというサイクルの核となる概念と具体的な方法を提供する．EDA を適切に応用することによってデータにもとづいた統計的推測・決定が実現されることに注意しよう．（図 1 参照.）この EDA における重要なパーツであるデータ可視化（data visualization），またはより広く情報可視化（information visualization）は，近年その重要性が再認識されている分野であり[2)]，統計モデリングとともに ICT 環境の進化・発展による寄与が大きい分野であることにも注意しよう[3)]．

本稿では，EDA の視点にたち，日経 NEEDS 財務データが収録されたデータベースから日本の企業に関する東京証券取引所一部上場企業全体を対象として売上高，従業員数，資産合計などの財務指標に関するデータを多期間に

2）たとえば，データ可視化に関するものとしては，Chen, Härdle, and Unwin（2008），情報可視化に関しては Tufte（1990, 1997, 2001, 2006），Mazza（2009），Ihaka（2013）等を参照のこと．

3）EDA を実現するためのソフトウェア環境の設計（design）が J. M. Chambers による S であり，この流れをくむ実装（implementation）が R. Gentleman と R. Ihaka による R であることに注意しよう．

わたって抽出したものをデータ解析環境Rを用いて可視化し，それにもとづいて統計モデリングを行うことを試みる．その際，本稿を通じて，売上高を従業員数と資産合計で説明するモデルを構築するために可視化の技法がどのように利用できるかをテーマとする[4]．

本稿の構成は以下のようなものである．まず，Ⅱ節では本稿で扱う財務データの説明を行う．次に，Ⅲ節ではⅡ節で得られた財務データを時間・空間の両面から可視化する．また，Ⅳ節では，Ⅲ節で得られた知見にもとづいて統計モデリングをおこない，Ⅴ節でそれらのモデルを実際にデータに当てはめることによって，モデルの妥当性などの検証を探索的に行う．最後に，Ⅵ節では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．なお，本稿で利用されるデータは地道（2010-a, b）で構築された学内向けデータベースサーバからR[5]を利用して抽出したものを利用しており，このデータのすべての処理もRによって行われていることに注意しよう[6]．

Ⅱ　財務データ

本稿で扱うデータは，東京証券取引所（以下「東証」と略す．）1部上場企業を母集団とする連結本決算にもとづく財務データであり，地道（2010-a, b）で構築されたデータベースを利用することによって得られたものである[7]．（表1参照．）実際のデータの取得に関して利用したスクリプトなどを付録Bに与える．なお，このデータベースの詳細については地道（2010-a, b）を参照されたい．

4）　いわゆる，生産関数（production function）の推定問題を扱う．
5）　Version 3.0.2
6）　本稿を執筆するために利用したRのスクリプトを付録Bに与えるので参照されたい．
7）　東証による公表（http://www.tse.or.jp/listing/companies/b7gje6000000pj9r-att/b7gje6000000pjqx.pdf）と比較すると，日経NEEDSに収録されているデータよりも実際には多くの企業（100社程度）が存在することがわかる．

表 1　日経 NEEDS 財務データベースから抽出した東京証券取引所一部上場企業の財務データ（全データ30928件から先頭の30件を抜粋）

	name	date	year	month	term	quarter	yearQ	sector 1	sector 2	sector 3	sales	employee	assets
1	KYOKUYO1	1984-10-31	1984	10	12	Q4	1984Q4	2	35	341	206485	NA	93094
2	KYOKUYO1	1985-10-31	1985	10	12	Q4	1985Q4	2	35	341	206512	1223	82267
3	KYOKUYO1	1986-10-31	1986	10	12	Q4	1986Q4	2	35	341	194353	1133	82394
4	KYOKUYO1	1987-10-31	1987	10	12	Q4	1987Q4	2	35	341	200304	1089	85497
5	KYOKUYO1	1988-03-31	1988	3	5	Q1	1988Q1	2	35	341	81843	1054	82382
6	KYOKUYO1	1989-03-31	1989	3	12	Q1	1989Q1	2	35	341	213409	873	86649
7	KYOKUYO1	1990-03-31	1990	3	12	Q1	1990Q1	2	35	341	207862	855	76786
8	KYOKUYO1	1991-03-31	1991	3	12	Q1	1991Q1	2	35	341	202573	846	74061
9	KYOKUYO1	1992-03-31	1992	3	12	Q1	1992Q1	2	35	341	199227	843	68312
10	KYOKUYO1	1993-03-31	1993	3	12	Q1	1993Q1	2	35	341	184988	851	67760
11	KYOKUYO1	1994-03-31	1994	3	12	Q1	1994Q1	2	35	341	164324	879	63693
12	KYOKUYO1	1995-03-31	1995	3	12	Q1	1995Q1	2	35	341	173803	1029	61692
13	KYOKUYO1	1996-03-31	1996	3	12	Q1	1996Q1	2	35	341	175202	1023	63287
14	KYOKUYO1	1997-03-31	1997	3	12	Q1	1997Q1	2	35	341	183640	1000	65883
15	KYOKUYO1	1998-03-31	1998	3	12	Q1	1998Q1	2	35	341	176022	976	62766
16	KYOKUYO1	1999-03-31	1999	3	12	Q1	1999Q1	2	35	341	171944	1000	62109
17	KYOKUYO1	2000-03-31	2000	3	12	Q1	2000Q1	2	35	341	171031	1148	60885
18	KYOKUYO1	2001-03-31	2001	3	12	Q1	2001Q1	2	35	341	166644	1145	60599
19	KYOKUYO1	2002-03-31	2002	3	12	Q1	2002Q1	2	35	341	158006	1148	57069
20	KYOKUYO1	2003-03-31	2003	3	12	Q1	2003Q1	2	35	341	162773	1162	55373
21	KYOKUYO1	2004-03-31	2004	3	12	Q1	2004Q1	2	35	341	151534	1145	58562
22	KYOKUYO1	2005-03-31	2005	3	12	Q1	2005Q1	2	35	341	152638	1123	58506
23	KYOKUYO1	2006-03-31	2006	3	12	Q1	2006Q1	2	35	341	152899	1123	65049
24	KYOKUYO1	2007-03-31	2007	3	12	Q1	2007Q1	2	35	341	157088	2791	66459
25	KYOKUYO1	2008-03-31	2008	3	12	Q1	2008Q1	2	35	341	147767	2710	57373
26	KYOKUYO1	2009-03-31	2009	3	12	Q1	2009Q1	2	35	341	147554	2682	61184
27	KYOKUYO1	2010-03-31	2010	3	12	Q1	2010Q1	2	35	341	145778	2909	64301
28	KYOKUYO1	2011-03-31	2011	3	12	Q1	2011Q1	2	35	341	162731	2753	76925
29	KYOKUYO1	2012-03-31	2012	3	12	Q1	2012Q1	2	35	341	181885	2460	84937
30	NIPPONSUISAN3	1984-03-31	1984	3	12	Q1	1984Q1	2	35	341	517134	6493	243958

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（1500社）
date：決算年月日（1976-02-29～2012-09-30間の724日分）
year：決算年（1976年～2012年の間37年分）
month：決算月
quarter：決算期（四半期）（第 1 四半期：Q1 第 4 四半期：Q4）
yearQ：決算年＋四半期
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.）
sales：売上高（単位：百万円）
employee：従業員数（単位：人）
assets：資産合計（単位：百万円）

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ことに注意しよう．

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
- Google Motion Chart による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2の左の列は，延べ1500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でつないだものであり，本稿で扱う全データがこのプロットにおいて表現されていることに注意しよう．この図からは以下のことがわかる．まず，各指標とも幾つかの「規模の大きな」企業が存在することがわかり，それ以外の企業についての変動がスケールの関係上みえ難くなっていることに注意しよう．次に，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[8)]．さらに，1980年の半ば（正確には1984年）にデータとして収録された企業数が急増してい

ることがわかる．なお，決算日が企業によって異なっていることにも注意しよう[9]．

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，時間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，時間をある時点で固定したもとで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2012年 3 月31日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 2 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2012年 3 月31日で時点を固定したもとで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．図 2 から，本稿で扱っている財務データは時点を2012年 3 月31日で固定すると，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，図 3 で与えられている 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，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 3 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かわる．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．

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 4 は2012年 3 月31日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

8） より詳しくは，決算期間が 1 年のみの企業が 3 社存在する．
9） ほとんどの企業は 3 月31日を決算日としていることが別途わかる．

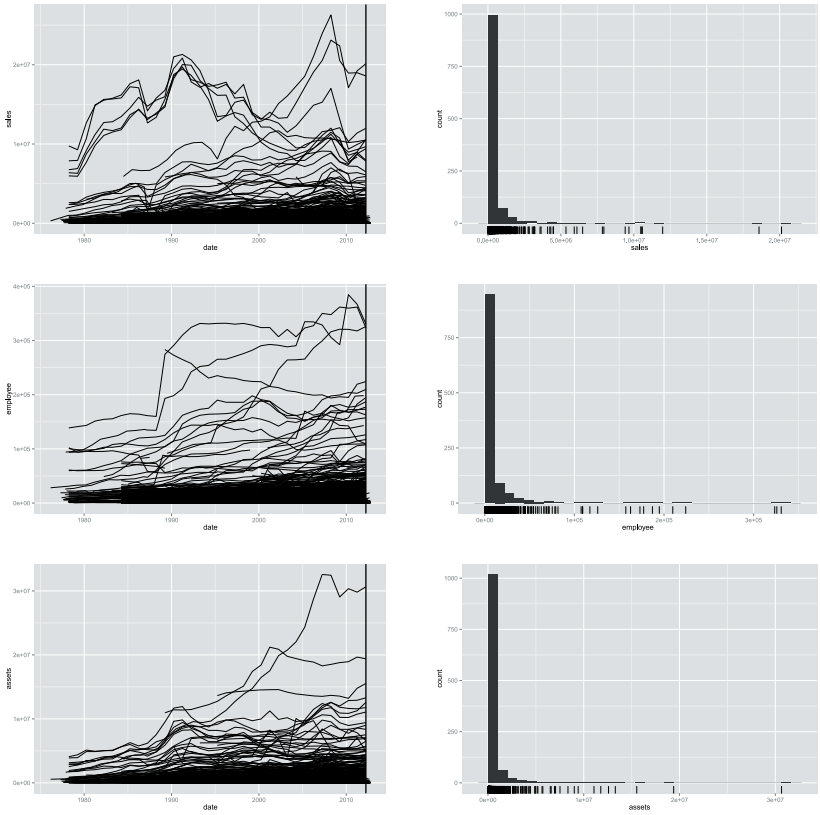

図 2　東証一部上場企業の財務データの時系列プロットとヒストグラム：行列の形式で，(1, 1)，(1, 2)，(1, 3) 成分に対応するプロットは，それぞれ，個々の企業の売上高（sales），従業員数（employee），資産合計（assets）であり，(2, 1)，(2, 2)，(2, 3) 成分に対応するプロットは，2012年 3 月31日で時点を固定し，横断面（垂直線）をとったときヒストグラム（ラグ付）である．

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かわる．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，本稿で扱っている財務データを2012年 3 月31日で固定

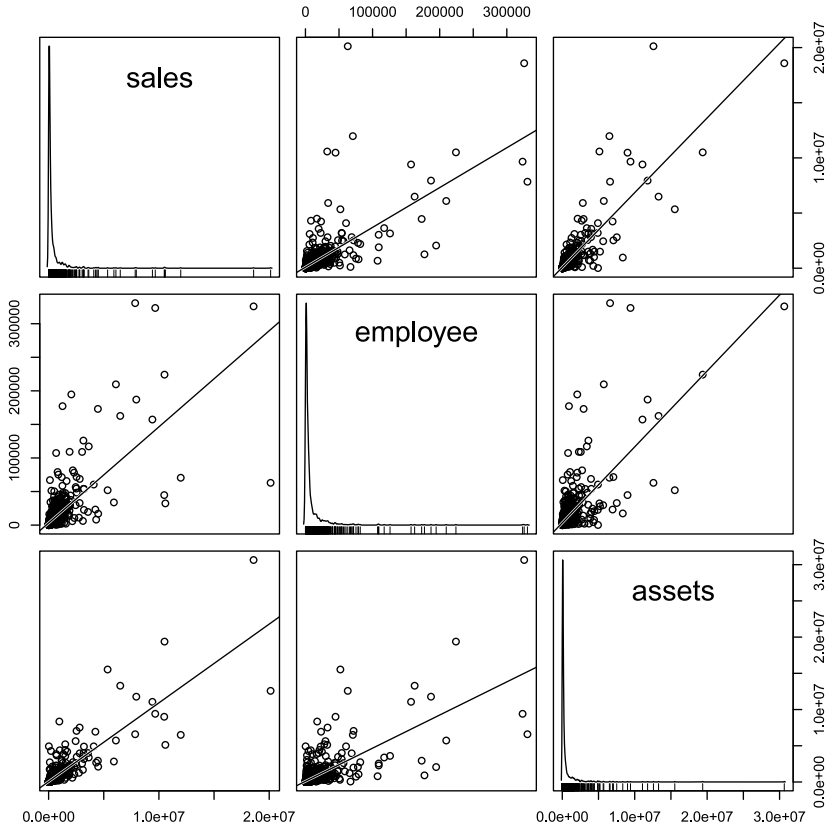

図 3　東証一部における2012年 3 月31日決算の企業の売上高，従業員数，資産合計の対散布図

した場合のクロスセクションデータは，原点付近で高密度をもち，原点から離れたところにも低いけれども密度が存在するような「歪んだ分布」に従ったものであることがわかった．このような場合には，対数（logarithm）をとることによって，原点付近の小さな値を拡大し，かつ大きな値を圧縮することによって結果的に全体を対称化（symmetrization）できる場合が多いことが知られている．（たとえば，Tukey（1977），Moster and Tukey（1977），

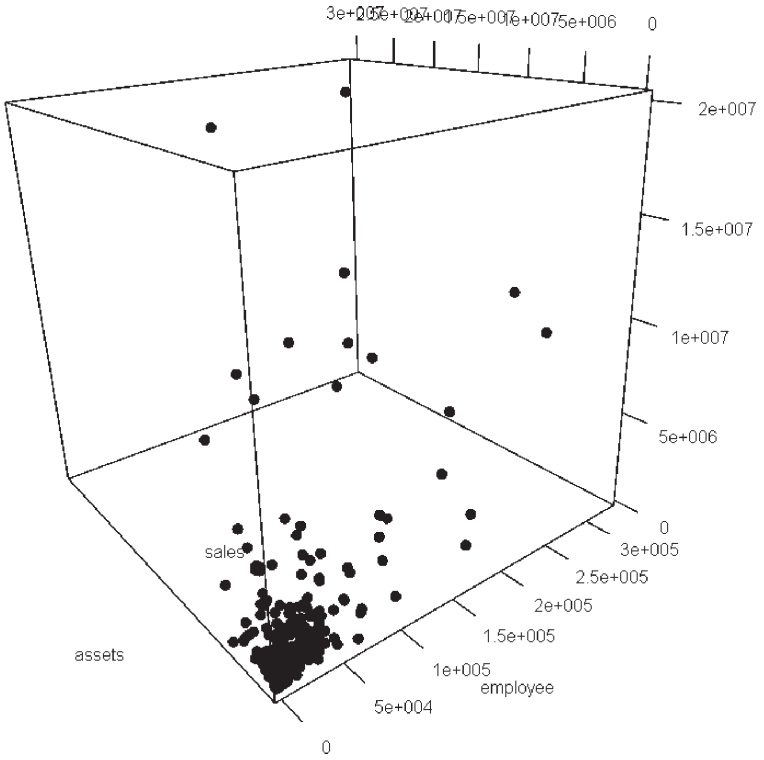

図 4　東証一部における2012年 3 月31日決算の企業の売上高，従業員数，資産合計の 3 次元散布図

Fox（2011）などを参照されたい.）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 2），対散布図（図 3），3 次元散布図（図 4）を対数スケールで描きなおしたものが，それぞれ，図 5，6，7 である.

3．時空間的データ可視化

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

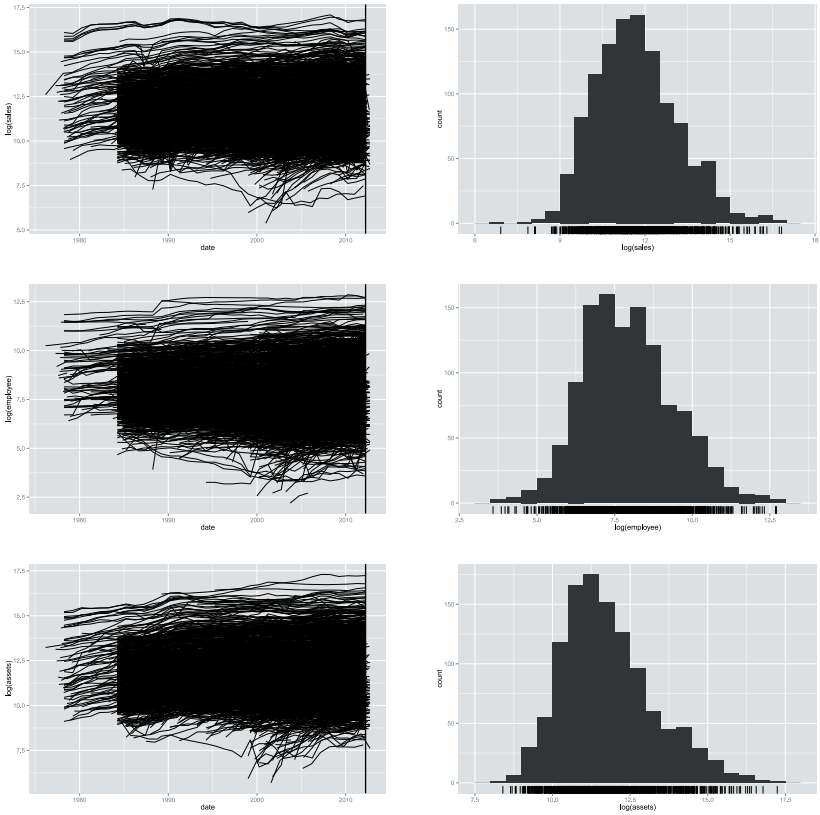

図 5　東証一部上場企業の財務データ（対数スケール）の時系列プロットとヒストグラム：行列の形式で，(1, 1)，(1, 2)，(1, 3) 成分に対応するプロットは，それぞれ，個々の企業の売上高，従業員数，資産合計の対数スケールの時系列プロットであり，(2, 1)，(2, 2)，(2, 3) 成分に対応するプロットは，2012年 3 月31日の時点を固定し，横断面（垂直線）をとったときのヒストグラム（ラグ付）である．

間的な推移に伴う母集団（空間）の状況を視覚的に見ることができれば，総合的にデータの状況を把握することが可能となる．データを時空間の両面から可視化できるフリーのアプリケーションソフトウェアとして，Google

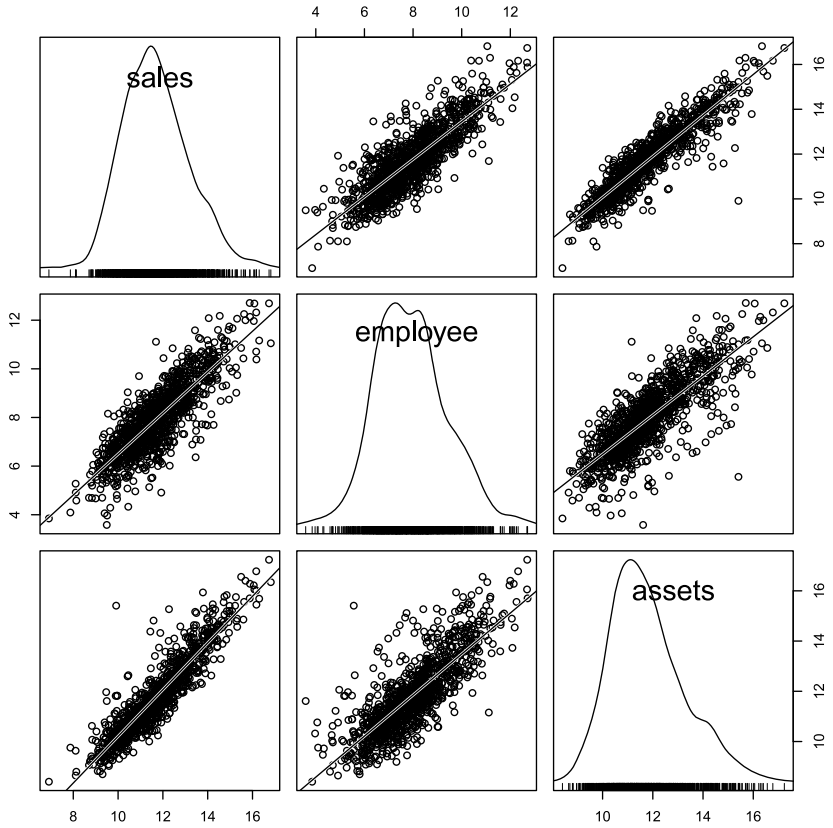

図 6　東証一部における2012年 3 月31日決算の企業の売上高，従業員数，資産合計の対散布図（対数スケール）

Motion Chart（以下，Motion Chart と呼ぶ.）が現時点で最も優れたソフトウェアのうちの一つであろう.（図 8 参照．なお，付録 C に Motion Chart に関する簡単な説明を与えている.）

動的な変化を紙面では伝えることは難しいが，ここでは Motion Chart を利用し，データを単年毎にまとめたもの[10] のバブルチャートを時間の変遷とともにプロットしたもののスナップショットをとることによって，分布状況

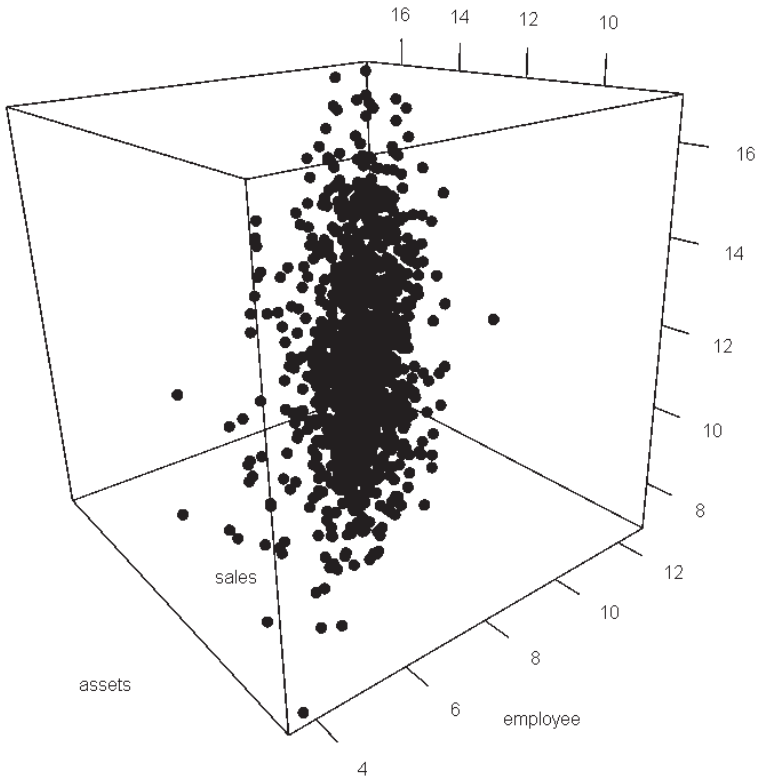

図 7　東証一部における2012年 3 月31日決算の企業の売上高，従業員数，資産合計の 3 次元散布図（対数スケール）

の時間的な推移をあらわしたものを図 9 に与える．この可視化によって，東証 1 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「分布構造」に大きな変化が無いことが分かった．

10）連続的に変化する量を何らかのカテゴリに分けることによって離散化し，可視化を行う方法は，一般に「ビニング」（binning）または「シングリング」（shingling）と呼ばれる技法であることに注意しよう．（Tidwell（2011）参照．）

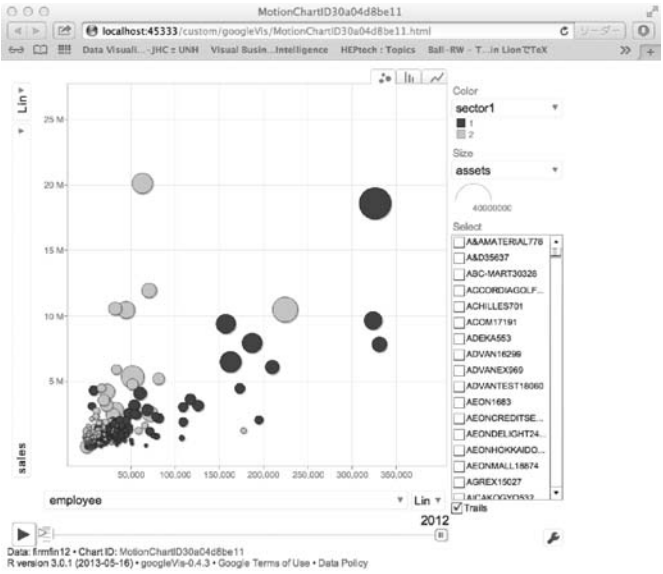

図 8　Motion Chart による2012年の東証一部における企業の売上高，従業員数，資産合計のプロット

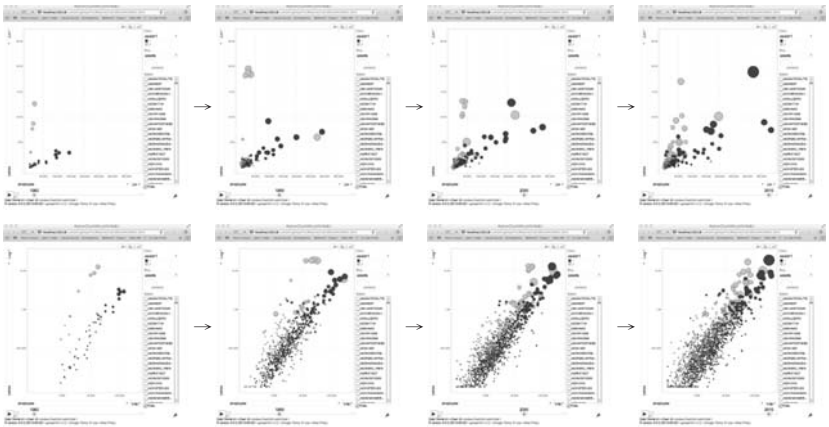

図 9　Motion Chart による1980年，1990年，2000年，2010年時点の東証一部における企業の売上高，従業員数，資産合計のプロットのスナップショット（上段：通常スケール，下段；（常用）対数スケール）

4．可視化から与えられた示唆

ここまで時空間の両面の観点からデータ可視化を行ってきたが，財務データを2012年 3 月31日で固定したクロスセクションの「分布構造」が右に歪んだものであり，その対数をとることによって対称なものに近づけることができることがわかった．このことは，各変量が対数正規分布（log-normal distribution）に従うことを示唆している．さらに，それらの変量の同時分布も，同様の理由によって，多変量対数正規分布（multivariate log-normal distribution）に従うことが示唆される．さらに，Motion Chart を利用することによって経時的に観測されたものに関してもこの構造は安定的であることが分かった．

次節では，これらの知見をふまえた統計モデリングを行う．このことは Tukey（1977）による探索的データ解析の観点にたったアプローチを実行している一つの例となっていることに注意しよう．なお，対数正規分布と多変量対数正規分布に関しては，付録Ｄに簡単な説明を与えるが，詳しくは Aitchison and Brown（1957），Crow and Shimizu（1988）などを参照されたい．

Ⅳ　財務データの統計モデリング

ここでは，前節で得られた可視化に関する情報をもとにクロスセクションの観点から財務データに対する統計モデリングを扱う．本稿で扱う統計モデリングは，加法モデル（additive model）と乗法モデル（multiplicative model）にもとづくものであることに注意しよう．なお，データの統計モデリングの詳細については，たとえば，Chambers and Hastie（1991）を参照されたい．

1．一般的モデリング

本稿で扱っている財務データにおける売上高（sales）y を従業員数（employee）x_1 と資産合計（assets）x_2 を使って一般的にモデル化することを以下のように書くことにする：

$$\texttt{sales}\sim f(\texttt{employee},\ \texttt{assets}) \Longleftrightarrow y\sim f(x_1,\ x_2)=f(\boldsymbol{x})$$

ここで，$\boldsymbol{x}:=[x_1,\ x_2]'$ とおいた．なお，$'$ は行列やベクトルの転置を表す記号である．

本稿で考察しているような経時測定データの場合は，このモデルに対して個体を表す添字 i と時間を表す添字 t を使い，

$$y_{it}\sim f(x_{i1t},\ x_{i2t})=f(\boldsymbol{x}_{it})$$

と表す．ここで，$\boldsymbol{x}_{it}:=[x_{i1t},\ x_{i2t}]'$ とおいた．また，個体（企業）が固定されている場合は，i を省略して，

$$y_t\sim f(x_{1t},\ x_{2t})=f(\boldsymbol{x}_t)$$

と書く．ここで，$\boldsymbol{x}_t:=[x_{1t},\ x_{2t}]'$ とおいた．逆に，時点（決算年月日）を固定した場合は，t を省略して，

$$y_i\sim f(x_{i1},\ x_{i2})=f(\boldsymbol{x}_i)$$

と書くこととする．ここで，$\boldsymbol{x}_i:=[x_{i1},\ x_{i2}]'$ とおいた．

2．加法モデルと正規線形モデル

統計モデリングの典型的なものとして，観測（observation）が構造的部分（structural part）[11] と確率的変動部分（stochastic part）の和によって表される場合を考え，ここでは，加法モデル（additive model）と呼ぶことにする[12]：

観測＝構造的部分＋確率的変動部分

なお，加法性の仮定については，Stuart and Ord（1991）も参照されたい．応答変数 y，説明変数ベクトル $\boldsymbol{x}(\in\mathbb{R}^p)$ に対して加法モデルを適用すると，

$$y=f(\boldsymbol{x})+\epsilon$$

と書くことができる．ここで，ϵ は観測と構造的部分の差によってあらわされる誤差（error）である．

11）系統的部分（systematic part）と呼ばれることもある．
12）加法モデルという用語は，ノンパラメトリック回帰の枠組みからの定義もあるが，ここでは構造的部分と確率的変動部分が加法的であることを表していることに注意しよう．

加法モデルにおける系統的部分に利用されている関数 f として何を採用するかが統計モデリングの一つの重要な事項であるけれども，統計的回帰問題（statistical regression problem）の観点からは応答変数の平均構造（mean structure）を利用するというのが重要な選択肢である．（付録Eも参照のこと．）具体的には，

$$f(\boldsymbol{x})=\mathrm{E}(Y|\boldsymbol{X}=\boldsymbol{x})=:\mu(\boldsymbol{x})$$

ととられ，$\mathrm{E}(Y|\boldsymbol{X}=\boldsymbol{x})$ は説明変数ベクトル $\boldsymbol{X}=\boldsymbol{x}$ が与えられたもとでの応答変数 Y（変量）の条件付き平均である．条件付き平均（回帰関数）$\mu(\boldsymbol{x})$ を構造的部分としてもつ加法モデル

$$y=\mu(\boldsymbol{x})+\epsilon$$

は回帰モデル（regression model）と呼ばれる．

平均関数が以下のように線形関数

$$\eta(\boldsymbol{x}):=\beta_0+\boldsymbol{x}'\boldsymbol{\beta}=\beta_0+\sum_{j=1}^{p}x_j\beta_j$$

で近似できる場合，すなわち，

$$\mu(\boldsymbol{x})\simeq\eta(\boldsymbol{x})$$

であるとき，加法モデルは，

$$y=\beta_0+\sum_{j=1}^{p}\beta_j x_j+\epsilon$$

となり，線形回帰モデル（linear regression model）と呼ばれる．$\eta(\boldsymbol{x})$ は線形予測子（linear predictor）と呼ばれることがあることにも注意しよう．さらに，誤差 ϵ[13] が正規分布 $\mathsf{N}(0,\ \sigma^2)$ に従うことが仮定された場合は，正規線形モデル（normal linear model）と呼ばれる．

とくに，時点 t を固定（省略）し，個体 i を意識した正規線形モデルは，

$$y_i=\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}+\epsilon_i,\ \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0,\ \sigma^2) \tag{1}$$

13）厳密には，ここで表記されている誤差はその実現値であるが，その統計的抽象化である観測（確率変数）に関してこのことが仮定される．

と書くことができる．ここで，$\overset{\text{i.i.d.}}{\sim}$ は「独立に同一の分布に従う」（independent and identically distributed: i.i.d.）ことを表す記号である．

3．乗法モデルと対数正規線形モデル

加法モデル以外のもう一つの典型的な統計モデリングが，観測が構造的部分と確率的変動部分の積によって表されるものである：

観測＝構造的部分×確率的変動部分

本稿では，このモデルを乗法モデル（multiplicative model）と呼ぶことにする．

応答変数 y，説明変数ベクトル $\boldsymbol{x}(\in\mathbb{R}^p)$ に対して乗法モデルを適用すると，

$$y=f(\boldsymbol{x})\times\epsilon$$

と書くことができる．系統的部分における関数 f として標準的なものは巾関数（power function）：

$$f(\boldsymbol{x})=\gamma\times\left(\prod_{j=1}^{p}x_j^{\alpha_j}\right)$$

であり，このときの乗法モデルは以下のように与えられる：

$$y=\gamma\times\left(\prod_{j=1}^{p}x_j^{\alpha_j}\right)\times\epsilon$$

さらに誤差 ϵ が対数正規分布 $\mathsf{LN}(0,\ \sigma^2)$ に従うことが仮定されるとき，対数正規線形モデル（log-normal linear model）と呼ばれる．時点 t を固定（省略）し，個体 i を意識した対数正規線形モデルは，

$$y_i=\gamma\times\left(\prod_{j=1}^{p}x_{ij}^{\alpha_j}\right)\times\epsilon_i,\ \epsilon_i\overset{\text{i.i.d.}}{\sim}\mathsf{LN}(0,\ \sigma^2)\qquad(2)$$

と書くことができる．ここで，注意すべきことは対数正規線形モデル(2)が両辺の（自然）対数をとることによって，

$$\log y_i=\alpha_0+\sum_{j=1}^{p}\alpha_j\log x_{ij}+\log\epsilon_i,\ \log\epsilon_i\overset{\text{i.i.d.}}{\sim}\mathsf{N}(0,\ \sigma^2)\qquad(3)$$

となり，応答変数と説明変数の対数をとったものを新たな変数と見なすことによって正規線形モデル(1)となることである．これを「対数正規線形モデルの正規線形モデル表現」と呼ぶことにする．なお，$\alpha_0 := \log \gamma$ とおいた．

V　財務データへの統計モデルの当てはめ

この節では，財務データをまずクロスセクションの観点からとらえ，統計モデルを当てはめる．その際，時点は2012年３月31日でまずは固定するが，後に時点を変化させた場合についても考察する．

１．正規線形モデルの当てはめ

時間の添字 t を2012年３月31日の時点で固定することによって省略し，正規線形モデル

$$\mathtt{sales}_i = \beta_0 + \beta_1 \mathtt{employee}_i + \beta_2 \mathtt{assets}_i + \epsilon_i,\ \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,\ \sigma^2) \quad (4)$$

を財務データに当てはめる．ただし，$i = 1, \ldots, n\ (= 1142)$ 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，その推定値（最小自乗推定値（least square estimate））を $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ と書くことにする．なお，最小自乗法による線形回帰モデルの当てはめに関して付録Ｆに簡単な説明を与えるので参照されたい．

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．この結果から，回帰係数はすべて５％有意であることに注意しよう．従業員数（employee）と資産合計（assets）の線形予測子

$$\eta := \beta_0 + \beta_1 \mathtt{employee} + \beta_2 \mathtt{assets}$$

は幾何学的には平面（回帰平面（regression plane）と呼ばれる．）であるが，その回帰係数を最小自乗推定値でおきかえた

$$\hat{\eta} := \hat{\beta}_0 + \hat{\beta}_1 \mathtt{employee} + \hat{\beta}_2 \mathtt{assets}$$

は標本回帰平面（sample regression plane）と呼ばれ，実際に以下のように

表2　ティー検定表：正規線形モデルの場合

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	40070.0893	20204.8567	1.98	0.0476
employee	10.5728	1.0542	10.03	0.0000
assets	0.5577	0.0166	33.65	0.0000

与えられる：

$$\hat{\eta} = \hat{\beta}_0 + \hat{\beta}_1 \texttt{employee} + \hat{\beta}_2 \texttt{assets}$$
$$= 40070.089 + 10.573\,\texttt{employee} + 0.558\,\texttt{assets}$$

図10に三次元散布図に標本回帰平面を描いたプロットを与える．

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，$\hat{\sigma}^2 = 640903.248^2$ 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：

$$R^2 = 0.7577,\ \overline{R}^2 = 0.7572$$

決定率が約76%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10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．（たとえば，Belsley, Kuh, and Welsch（1980）参照.）図11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：$\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,\ \sigma^2)$ を検証するために利用されることに注意しよう[14]．行列形式で与えられた（1, 1）成分に対応する残差のインデッ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）成分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．（たとえば，Cook（1998）参照.）さらに，（1, 2）成分の残差の正規 Q-Q プロットと（2, 2）成分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

14）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．

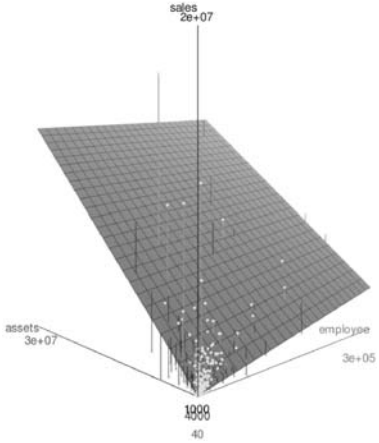

図10　2012年３月31日決算の企業の財務データの三次元散布図と標本回帰平面

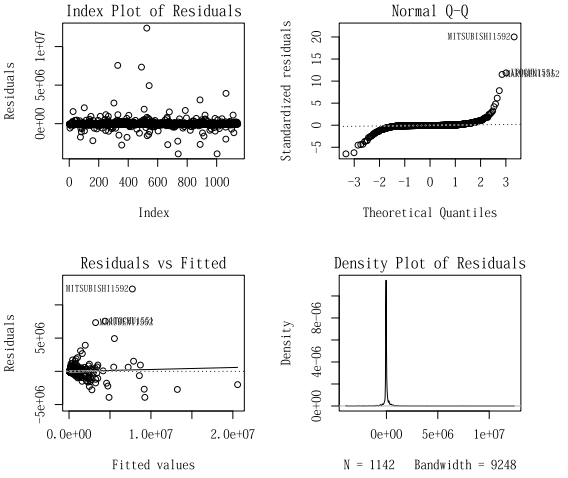

図11　2012年３月31日決算の東証一部上場企業に関する財務データにもとづく正規線形モデルの当てはめ結果にもとづく残差に関する各種のプロット：行列形式で順に，(1, 1) 成分；残差のインデックスプロット，(2, 1) 成分；当てはめ値に対する残差のプロット，(1, 2) 成分；残差の正規 Q-Q プロット，(2, 2) 成分；残差の平滑化された密度関数のプロット．

性が完全に疑われる結果となっていることに注意しよう．

2．対数正規線形モデルの当てはめ

正規線形モデル(4)を2012年3月31日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からデータが多変量対数正規分布に従うと考えられるため，この情報を積極的に利用した統計モデリングを行おう．対数正規線形モデルが多変量正規分布に親和的であることから（付録E参照.），正規線形モデルよりも適切なモデルとして以下の対数正規線形モデルを当てはめることが提案される：

$$\mathtt{sales}_i = \gamma \times \mathtt{employee}_i^{\alpha_1} \times \mathtt{assets}_i^{\alpha_2} \times \epsilon_i,\ \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathsf{LN}(0,\ \sigma^2) \qquad (5)$$

このモデルは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：

$$\log \mathtt{sales}_i = \alpha_0 + \alpha_1 \log \mathtt{employee}_i + \alpha_2 \log \mathtt{assets}_i + \log \epsilon_i,$$
$$\log \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0,\ \sigma^2) \qquad (6)$$

このモデルにおける回帰係数α_0, α_1, α_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．

表3　ティー検定表：対数正規線形モデルの場合

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.3534	0.1265	10.70	0.0000
log(employee)	0.2731	0.0179	15.28	0.0000
log(assets)	0.6942	0.0177	39.13	0.0000

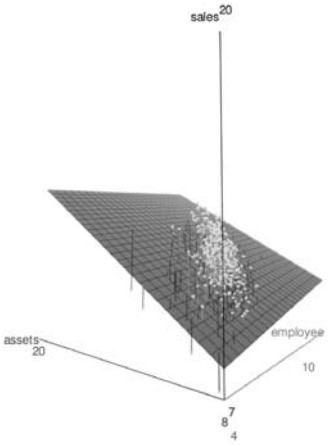

図12　2012年３月31日決算の企業の財務データの三次元散布図（対数スケール）と標本回帰平面（対数線形モデル）

標本回帰平面は，

$$\hat{\eta}_{\mathrm{LNL}} = \hat{\alpha}_0 + \hat{\alpha}_1 \log \mathtt{employee} + \hat{\alpha}_2 \log \mathtt{assets}$$
$$= 1.353 + 0.273\,\mathtt{employee} + 0.694\,\mathtt{assets}$$

で与えられ，図12に対数スケールで描いた三次元散布図に対数正規線形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．このモデルを当てはめたときの誤差分散の推定値は，$\hat{\sigma}^2 = 0.516^2$ 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：

$$R^2 = 0.8751,\ \overline{R}^2 = 0.8749$$

これらの結果において，決定率が約88%へ伸びており，図12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．

ただし，回帰診断に関するプロット（図13）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，影響力あるデータの検出などの分析は感度分析（sensitivity analysis）とよばれ，専用の指標やプロットが提案されている．（たとえば，Belsley, Kuh, and Welsch (1980), Chatterjee and Hadi (1988), Fox and Weisberg (2011) 参照.）ここでは，最も基本的なものであ

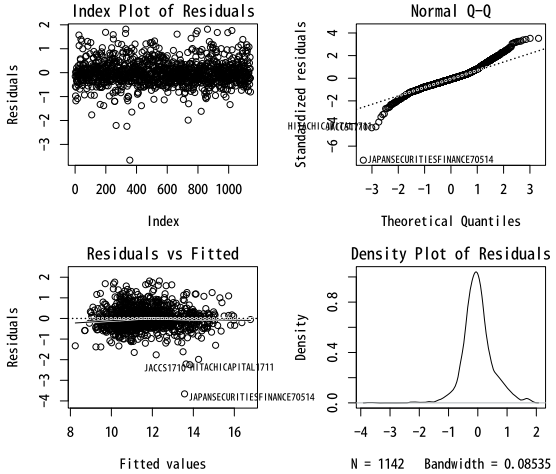

図13　2012年 3 月31日決算の東証一部上場企業に関する財務データにもとづく対数正規線形モデルの当てはめ結果にもとづく残差に関する各種のプロット：行列形式で順に，(1, 1) 成分；残差のインデックスプロット，(2, 1) 成分；当てはめ値に対する残差のプロット，(1, 2) 成分；残差の正規 Q-Q プロット，(2, 2) 成分；残差の平滑化された密度関数のプロット．

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14参照.）これらの指標についての簡単な説明を付録Gに与えているので参照されたい．なお詳細は，Chatterjee and Hadi（1988）を参照されたい．さらに，これらの指標を数値的に与えたものが表 4 である．これらの結果から，JAPAN SECURITIESFINANCE70514（日本証券金融）が最も影響力の強いデータであり，続いて，JACCS1710（ジャックス），HITACHICAPITAL1711（日立キャピタル），ORIENT1709（オリエントコーポレーション）が影響力が強いことが確認できる．この結果と回帰診断に関するプロット（図13）から，これらのデータを異質のものとして取り除き，再度対数正規線形モデルを当てはめる[15]．

再当てはめの結果は表 5 のように与えられる．この結果も，回帰係数はす

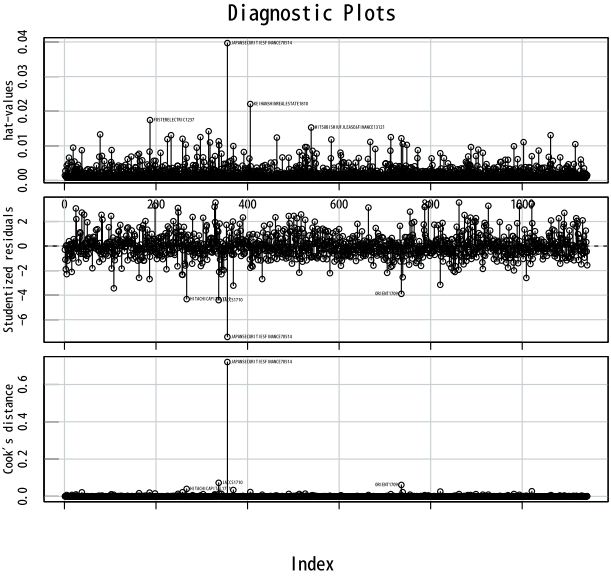

図14 2012年３月31日決算の東証一部上場企業に関する財務データにもとづく対数正規線形モデルの当てはめた場合の回帰診断（感度分析）のためのプロット

表４ 回帰診断（感度分析）のための指標

	StudRes	Hat	CookD
FOSTERELECTRIC1237	−0.87	0.02	0.07
HITACHICAPITAL1711	−4.32	0.01	0.20
JACCS1710	−4.41	0.01	0.27
JAPANSECURITIESFINANCE70514	−7.40	0.04	0.85
KEIHANSHINREALESTATE1810	−1.77	0.02	0.15
ORIENT1709	−3.89	0.01	0.25

べて５％有意であることに注意しよう．標本回帰平面は，

$$\hat{\eta}_{\mathrm{LNL.zdj}} = 1.13 + 0.231\,\mathtt{employee} + 0.742\,\mathtt{assets}$$

15）これらのデータは，資産合計がその規模に対して相対的に大きいことが別途分かる．

表5　ティー検定表：影響力のあるデータを除去後の対数正規線形モデルの場合

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.1300	0.1222	9.25	0.0000
log(employee)	0.2311	0.0175	13.23	0.0000
log(assets)	0.7423	0.0175	42.50	0.0000

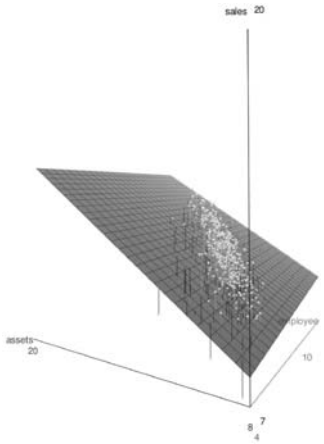

図15　2012年3月31日決算の企業の財務データの三次元散布図（対数スケール）と標本回帰平面（対数線形モデル：影響力のあるデータ除去後）

で与えられ，図15に対数スケールで描いた三次元散布図に対数正規線形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．

このモデルを当てはめたときの誤差分散の推定値は，$\hat{\sigma}^2=0.491^2$ 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：

$$R^2=0.887,\ \overline{R}^2=0.8868$$

この結果において，決定率が約89%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16）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ことに注意しよう[16]．

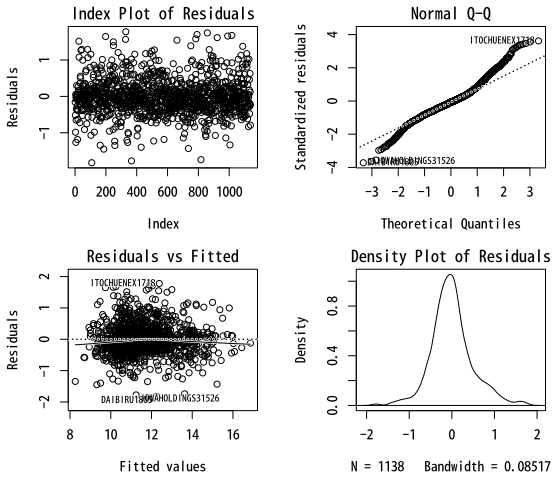

図16 2012年3月31日決算の東証一部上場企業に関する財務データ（影響力があるデータ除去後）にもとづく対数正規線形モデルの当てはめ結果にもとづく残差に関する各種のプロット：行列形式で順に，(1, 1) 成分；残差のインデックスプロット，(2, 1) 成分；当てはめ値に対する残差のプロット，(1, 2) 成分；残差の正規 Q-Q プロット，(2, 2) 成分；残差の平滑化された密度関数のプロット．

3．業種情報を踏まえた統計モデリングと当てはめ

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，2012年3月31日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が構築できたけれども，Motion Chart によるバブルチャート（図17）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．

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．その際，日経業種分類（付録H）

16）ただし，裾の部分が正規分布よりも若干重いことが残差の正規 Q-Q プロットからわかる．

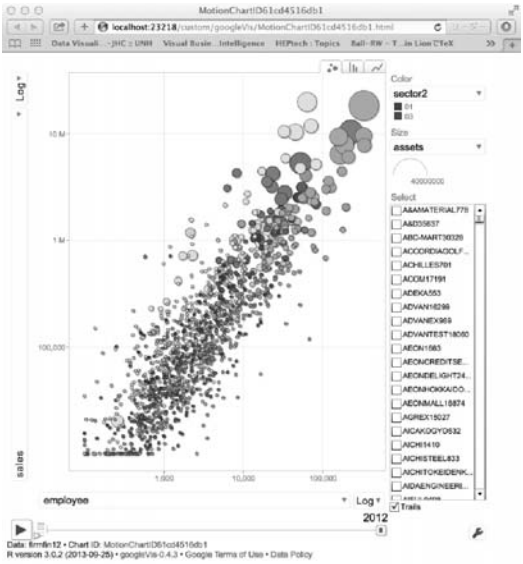

図17　Motion Chart による 2012年決算の企業に関する財務データのバブルチャート：日経業種中分類にもとづいて色分けしたもの

の中分類の情報（33業種）を利用することによって以下のようなモデリングを行う[17]．

$$\log \mathtt{sales}_i = \alpha_0 + \alpha_1 \log \mathtt{employee}_i + \alpha_2 \log \mathtt{assets}_i + \sum_{j=1}^{m} \delta_j D_{ij} + \log \epsilon_i,\ \log \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0,\ \sigma^2) \tag{7}$$

ここで，$j=1, \ldots, m\,(=33)$ であり[18]，

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

とする．なお，推定の一意性のために $\delta_1=0$ とする．このモデルは対数正

17）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71業種あり，逆に細分化されすぎるきらいがある．

18）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．

規線形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とに注意しよう．

このモデルにける回帰係数の推定結果が表 6 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[19]．

標本回帰平面（群）は，

$$\hat{\eta}_j=(1.228+\hat{\delta}_j)+0.257\,\mathtt{employee}+0.746\,\mathtt{assets},\quad j=1,\ \ldots,\ 33$$

で表される．（図18参照.）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．

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：

$$\begin{aligned}\hat{\eta}_{\mathtt{LNL.textile}}&=1.228+(-0.598)+0.257\,\mathtt{employee}+0.746\,\mathtt{assets}\\&=0.63+0.257\,\mathtt{employee}+0.746\,\mathtt{assets}\end{aligned}$$

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，

$$\hat{\eta}_{\mathtt{LNL.food}}=1.228+0.257\,\mathtt{employee}+0.746\,\mathtt{assets}$$

となる．

ダミー変数をもつ対数正規線形モデル(7)を当てはめたときの誤差分散の推定値は，$\hat{\sigma}^2=0.35^2$ で与えられ，決定係数と自由度調整済み決定係数は

$$R^2=0.9442,\ \overline{R}^2=0.9425$$

となり，約94%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も，このモデルの拡張が有用であることがわかる．

19）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，4 社と 5 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．

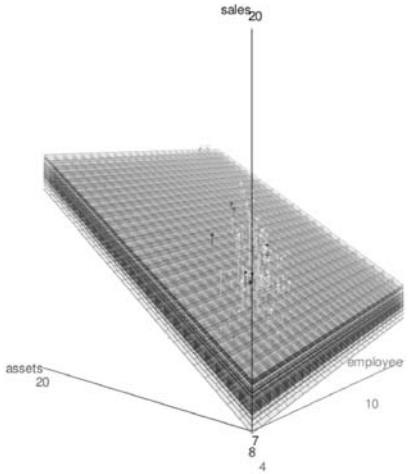

図18　2012年 3 月31日決算の企業の財務データの三次元散布図（対数スケール）と標本回帰平面群（対数線形モデル：影響力のあるデータ除去，ダミー変数含む）

5.4　経時観測の視点からの対数線形モデルの有用性

これまでの考察によって，ダミー変数をもつ対数正規線形モデル(7)が有用なものであることが分かった．では，この結果は時間を変化させても有用であるかをみよう．経時観測データにモデル(7)を拡張したものが以下のモデルである：

$$\log \mathtt{sales}_{it} = \alpha_{0t} + \alpha_{1t} \log \mathtt{employee}_{it} + \alpha_{2t} \log \mathtt{assets}_{it} + \sum_{j=1}^{m_t} \delta_{jt} D_{ijt}$$

$$+ \log \epsilon_{it}, \quad \log \epsilon_{it} \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0,\ \sigma_t^2) \tag{8}$$

ここで，$i=1, \ldots, n_t$, $j=1, \ldots, m_t$, $t=1, \ldots, T$ 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[20]．なお，推定の一意性のため $\delta_{1t}=0$ とする．

20）決算年月日 t に依存して企業と業種も変化するため，i_t, 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．

表 6　ティー検定表：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含む対数正規線形モデルの場合

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.2284	0.1155	10.63	0.0000
log(employee)	0.2569	0.0173	14.84	0.0000
log(assets)	0.7460	0.0172	43.44	0.0000
sector203	−0.5981	0.0872	−6.86	0.0000
sector205	−0.4503	0.1228	−3.67	0.0003
sector207	−0.3905	0.0638	−6.12	0.0000
sector209	−0.6580	0.0847	−7.77	0.0000
sector211	0.2814	0.1534	1.83	0.0669
sector213	−0.4740	0.1290	−3.67	0.0003
sector215	−0.5505	0.0902	−6.10	0.0000
sector217	−0.3818	0.0839	−4.55	0.0000
sector219	−0.3888	0.0745	−5.22	0.0000
sector221	−0.6145	0.0637	−9.65	0.0000
sector223	−0.5636	0.0636	−8.86	0.0000
sector225	−0.3871	0.2092	−1.85	0.0645
sector227	−0.2742	0.0761	−3.60	0.0003
sector229	−0.4468	0.1426	−3.13	0.0018
sector231	−0.7030	0.0910	−7.72	0.0000
sector233	−0.4314	0.0781	−5.53	0.0000
sector235	0.0390	0.1829	0.21	0.8314
sector237	−0.0970	0.1544	−0.63	0.5297
sector241	−0.1364	0.0659	−2.07	0.0386
sector243	0.4384	0.0633	6.93	0.0000
sector245	0.1346	0.0766	1.76	0.0791
sector252	−1.2793	0.1059	−12.08	0.0000
sector253	−0.8136	0.0875	−9.30	0.0000
sector255	−1.0425	0.0974	−10.71	0.0000
sector257	−0.4270	0.1068	−4.00	0.0001
sector259	−0.4627	0.1310	−3.53	0.0004
sector261	−0.4958	0.2094	−2.37	0.0181
sector263	−0.3532	0.1002	−3.53	0.0004
sector265	−0.1762	0.1120	−1.57	0.1161
sector267	−0.8640	0.1216	−7.10	0.0000
sector269	−0.3360	0.1657	−2.03	0.0429
sector271	−0.3481	0.0629	−5.54	0.0000

このモデルにおいて，決算年月日を，たとえば，2012年 3 月31日で固定（$t=\tau$ 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，ダミー

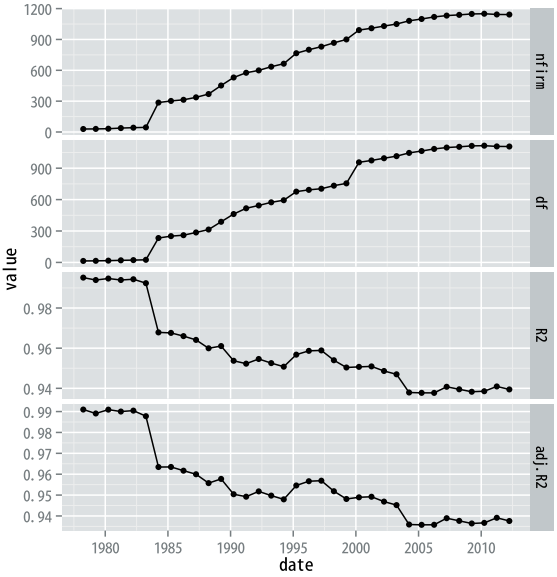

図19　1978年から2012年までの 3 月31日決算の企業に関するクロスセクションデータへ対数線形モデル（ダミー変数含む）を当てはめた結果：各時点の企業数，自由度，決定係数，自由度調整済み決定係数の時系列プロット

変数をもつ対数正規線形モデル(8)を当てはめた結果として得られる決定係数などを時系列としてプロットしたものを図19に与える．このプロットから，決定係数と自由度調整済み決定係数が時間とともに減少しており，モデルの当てはまりが悪くなっているように見受けられるかもしれないが，その範囲を見ると最低でも約94%は確保されており，年々企業数が増加することによって自由度が増加していることを勘案すると，逆に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，対数正規線形モデル(8)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妥当なものであることが時間的な推移を考慮しても正当化されることがわかった．

VI おわりに

本稿では，探索的データ解析の視点から日経 NEEDS データベースから取得された東証一部上場企業の売上高，従業員数，資産合計などの財務データに対する可視化と統計モデリングを統計解析環境 R を用いて行った．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対数正規線形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時間的な推移を考慮しても正当化されることがわかった．なお，今回得られた結果に対する補足を以下に与える：

（R1） 経済学で扱われるコブ・ダグラス型生産関数（Cobb-Douglas production function）$Y=AL^{\alpha}K^{\beta}$ 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．（ただし，Y が生産量，L が労働，K が資本を表すものとする．）今回，対数正規線形モデルを当てはめた結果においても，この関係が近似的に成り立っていることに注意しよう．

（R2） 対数正規線形モデルを実際の現象へ応用すること自体は決して新しいものではなく，経済学や生物学などの様々な分野へ応用されてきたものであることに注意しよう．（たとえば，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．）

（R3） 本稿において扱ったモデリングにおいて，応答変数（左辺）の対数をとることによって正規線形モデルの枠組みで分析を行ったが，応答変数はもとのスケールを保ったまま，説明変数（左辺）の対数をとるモデリング（たとえば，小西ら（2004）参照．）や乗法モデルを直接使った推測を行う立場（たとえば，Bradu and Mundlak（1970）参照．）もあることに注意しよう．

最後に，本稿の統計モデリングのアプローチは，時間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，時間的または時間的・空間的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）も考えることができるが，これらの観点からのモデリングとその応用については今後の課題としたい．

（筆者は関西学院大学商学部教授）

参考文献

［1］ Aitchison, J., and J. A. C. Brown（1957）*The Lognormal Distribution with Special Reference to Its Uses in Economics,* Cambridge University Press.

［2］ Belsley, D. A., E. Kuh, and R. E. Welsch（1980）*Regression Diagnostics: Identifying Infuluential Data and Sources of Collinearity,* John Wiley & Sons, Inc.

［3］ Bradu, D., and Y. Mundlak（1970）Estimation in Lognormal Linear Models, *Journal of the American Statistical Association,* Vol. 65, No. 329, pp. 198-211.

［4］ 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.）

［5］ Chatterjee, S., and Hadi（1988）*Sensitivity Analysis in Linear Regression,* John Wiley & Sons, Inc.

［6］ Chen, C., W. Härdle, and A. Unwin（editors）（2008）*Handbook of Data Visualization,* Springer.

［7］ Cook, R. D.（1998）*Regression Graphics: Ideas for Studying Regressions through Graphics,* John Wiley & Sons, Inc.

［8］ Crow, E. L., and K. Shimizu（editors）（1988）*Lognormal Distributions: Theory and Applications,* Marcel Dekker.

［9］ Faraway, J. J.（2005）*Linear Models with R,* Chapman & Hall/CRC.

［10］ Faraway, J. J.（2006）*Extending the Linear Models with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models,* Chapman & Hall/CRC.

［11］ Fox, J.（2008）*Applied Regression Analysis and Generalized Linear Models, Second Edition,* Sage.

［12］ Fox, J., and S. Weisbrerg（2011）*An R Companion to Applied Regression, Second Edition,* Sage.

［13］ Gesmann, M., and D. de Castillo（2011）Using the Google Visualisation API with R, *The R Journal,* Vol. 3, No. 2, pp. 40-44.

［14］ Ihaka, R.（2013）*Lecture Note: Information Visualisation,* `https://www.stat.auckland.ac.nz/~ihaka/courses/120/`.

［15］ 稲垣宣生（2003）『数理統計学（改訂版）』，裳華房.

［16］ 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.
［17］ 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://kgur.kwansei.ac.jp/dspace/handle/10236/6013，ISBN：784990553005.
［18］ Kabacoff, R. I.（2011）*R in Action: Data Analysis and Graphics with R,* Manning Publications Company.
［19］ Keen, K. J.（2010）*Graphics for Statistics and Data Analysis with R,* Chapman & Hall/CRC.
［20］ Klein, L. R.（1953）*A Textbook of Econometrics,* Row Peterson and Company.
［21］ Klein, L. R.（1962）*An Introduction to Econometrics,* Prentice Hall.
［22］ 小西葉子，西山慶彦，安藤知寛，川崎能典（2004）『生産関数のノンパラメトリック統計解析』，応用統計学，Vol. 33，No. 2，pp. 157-179.
［23］ 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.）
［24］ Mosteller, F., and Tukey, J. W.（1977）*Data Analysis and Regression: A Second Course in Statistics,* Addison-Wesley, Reading Mass.
［25］ Rao, C. R.（1973）*Linear Statistical Inference and Its Applications, Second Edition,* John Wiley & Sons, Inc.
［26］ Sarkar, D.（2008）*Lattice: Multivariate Data Visualization with R,* Springer.（石田基広，石田和枝共訳，（2009）『Ｒグラフィックス自由自在』，シュプリンガー・ジャパン株式会社.）
［27］ 柴田里程（2001）『データリテラシー』，共立出版.
［28］ Stuart, A., and J. K. Ord（1991）*Kendall's Advanced Theory of Statistics, Fifth Edition, Volume 2, Classical Inference and Relationship,* Edward Arnold.
［29］ Tufte, E. R.（1990）*Envisioning Information,* Graphics Press, Cheshire, Connecticut.
［30］ Tufte, E. R.（1997）*Visual Explanations,* Graphics Press, Cheshire, Connecticut.
［31］ Tufte, E. R.（2001）*The Visual Display of Quantitative Information,* Graphics Press, Cheshire, Connecticut.
［32］ Tufte, E. R.（2006）*Beautiful Evidence,* Graphics Press, LLC.
［33］ Tidwell, J.（2011）*Designing Interfaces, Second Edition,* O'Reilly.（浅野紀予訳，（2011）『デザイニング・インターフェイス第 2 版：パターンによる実践的インタラクションデザイン』，オライリー・ジャパン.）
［34］ Tukey, J. W.（1977）*Exploratory Data Analysis,* Addison-Wesley Publishing Co.
［35］ 筑波大学ビジネス科学研究科編（2003）『ビジネス数理への誘い』，朝倉書店.
［36］ Wickham, H.（2009）*ggplot2: Elegant Graphics for Data Analysis,* 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.）
［37］ Wilkinson, L.（2005）*The Grammar of Graphics, Second Edition,* Springer.

付録A　データ抽出，データブラウジング，データスクリーニング

1．データ抽出

本稿で扱うデータは，地道（2010-a, b）で構築された日経 NEEDS 財務データにもとづくデータベースサーバから R と ODBC[21] 環境とのインターフェース RODBC を利用することによって抽出したものである．その際に利用されたスクリプトは付録 B に与えられている．（図20も参照のこと.）

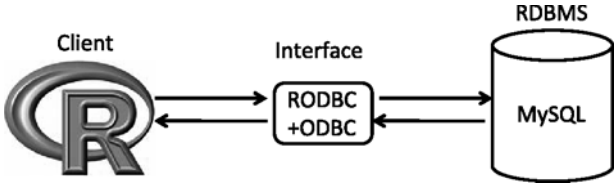

図20　R，データベース，ODBC，RODBC の関連図

2．データブラウジングとデータスクリーニング

取得されたデータを「そのまま」ながめたり，プロットすることによって俯瞰することはデータ解析を行う上で重要であり，データブラウジング（data browsing）と呼ばれることがある．（柴田（2001）参照.）財務データを時空間の観点からブラウジングすることの詳細は本文に記載されているが，ブラウジングすることによってデータを解析する前に処理するべきこと，すなわちデータスクリーニング（data screening）を行うべきこととして以下のことが検討項目としてあがった：

（DBS1）欠損値の存在
（DBS2）決算日を変更している企業が存在すること
（DBS3）決算日が企業によって異なっていること

（DBS1）については，まず，粗データ（抽出した直後のデータ）における

21）Open Database Connectivity の略．一般のアプリケーションからデータベースにネットワークを経由してアクセスするための標準的なインターフェース規格.

表 7 「極洋」に関するデータ

	name	date	year	month	term	quarter	yearQ	sector 1	sector 2	sector 3	sales	employee	assets
1	KYOKUYO1	1984-10-31	1984	10	12	Q4	1984Q4	2	35	341	206485	NA	93094
2	KYOKUYO1	1985-10-31	1985	10	12	Q4	1985Q4	2	35	341	206512	1223	82267
3	KYOKUYO1	1986-10-31	1986	10	12	Q4	1986Q4	2	35	341	194353	1133	82394
4	KYOKUYO1	1987-10-31	1987	10	12	Q4	1987Q4	2	35	341	200304	1089	85497
5	KYOKUYO1	1988-03-31	1988	3	5	Q1	1988Q1	2	35	341	81843	1054	82382
6	KYOKUYO1	1989-03-31	1989	3	12	Q1	1989Q1	2	35	341	213409	873	86649
7	KYOKUYO1	1990-03-31	1990	3	12	Q1	1990Q1	2	35	341	207862	855	76786
8	KYOKUYO1	1991-03-31	1991	3	12	Q1	1991Q1	2	35	341	202573	846	74061
9	KYOKUYO1	1992-03-31	1992	3	12	Q1	1992Q1	2	35	341	199227	843	68312
10	KYOKUYO1	1993-03-31	1993	3	12	Q1	1993Q1	2	35	341	184988	851	67760
11	KYOKUYO1	1994-03-31	1994	3	12	Q1	1994Q1	2	35	341	164324	879	63693
12	KYOKUYO1	1995-03-31	1995	3	12	Q1	1995Q1	2	35	341	173803	1029	61692
13	KYOKUYO1	1996-03-31	1996	3	12	Q1	1996Q1	2	35	341	175202	1023	63287
14	KYOKUYO1	1997-03-31	1997	3	12	Q1	1997Q1	2	35	341	183640	1000	65883
15	KYOKUYO1	1998-03-31	1998	3	12	Q1	1998Q1	2	35	341	176022	976	62766
16	KYOKUYO1	1999-03-31	1999	3	12	Q1	1999Q1	2	35	341	171944	1000	62109
17	KYOKUYO1	2000-03-31	2000	3	12	Q1	2000Q1	2	35	341	171031	1148	60885
18	KYOKUYO1	2001-03-31	2001	3	12	Q1	2001Q1	2	35	341	166644	1145	60599
19	KYOKUYO1	2002-03-31	2002	3	12	Q1	2002Q1	2	35	341	158006	1148	57069
20	KYOKUYO1	2003-03-31	2003	3	12	Q1	2003Q1	2	35	341	162773	1162	55373
21	KYOKUYO1	2004-03-31	2004	3	12	Q1	2004Q1	2	35	341	151534	1145	58562
22	KYOKUYO1	2005-03-31	2005	3	12	Q1	2005Q1	2	35	341	152638	1123	58506
23	KYOKUYO1	2006-03-31	2006	3	12	Q1	2006Q1	2	35	341	152899	1123	65049
24	KYOKUYO1	2007-03-31	2007	3	12	Q1	2007Q1	2	35	341	157088	2791	66459
25	KYOKUYO1	2008-03-31	2008	3	12	Q1	2008Q1	2	35	341	147767	2710	57373
26	KYOKUYO1	2009-03-31	2009	3	12	Q1	2009Q1	2	35	341	147554	2682	61184
27	KYOKUYO1	2010-03-31	2010	3	12	Q1	2010Q1	2	35	341	145778	2909	64301
28	KYOKUYO1	2011-03-31	2011	3	12	Q1	2011Q1	2	35	341	162731	2753	76925
29	KYOKUYO1	2012-03-31	2012	3	12	Q1	2012Q1	2	35	341	181885	2460	84937

欠損値が -999999999999 で表されているので，R での扱いを勘案して NA ですべて置換した．また，本稿は経時観測された財務データを扱うことから，時間情報である「決算日」に関して欠損値が存在するものは初期段階で取り除く処理を行った．

つぎに，（DBS2）に対しては，決算日の変更があった場合に，決算日から決算日までの期間の短縮にともなって売上高などの指標は大きく影響を受けることになる．つまり，期間が短くなったために著しく売上高が減少しているような結果[22]を与えることになり，分析に際して異常値となってしまう可能性があることを示唆している[23]．たとえば，水産業に属する「極洋」のデー

22）「見せかけの減少」とでも呼ぶべき結果となる．
23）この問題に関して，東証 1 部に上場している企業全体に対して，1987年から1988年の

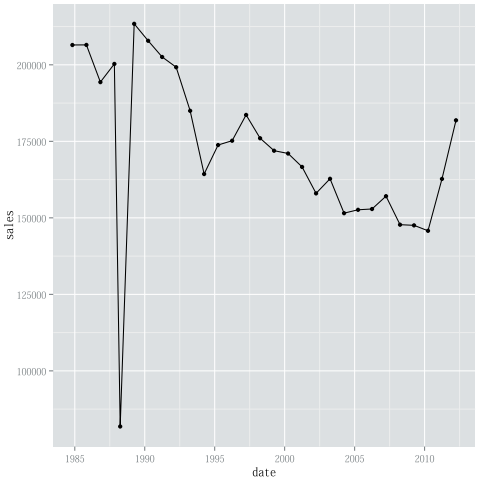

図21 「極洋」の売上高の時系列プロット

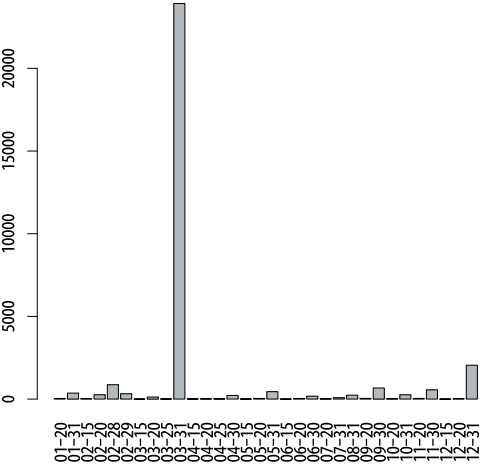

図22　決算日（全期間にわたって延べ）に関する企業数のプロット

間に決算日の変更があったことがブラウジングによってわかった．

24）本稿で行った分析は，クロスセクションデータにもとづいたものであるので，このような処理による影響は少ないが，時系列解析や経時観測データ分析などを行う場合には，改めて考える必要があろう．

タを見ると，表7のように決算日の変更があり，それによって売上高が極端に減少しているように見える．（図21も参照せよ.）この問題に対して，本稿では単純ではあるが，決算日が変更された直後のデータは削除する処理を行った[24]．

さらに（DBS3）に関しては，決算日の分布（図22）からもわかるように，ほとんどの企業が3月31日であることがわかる．このことから本稿では分析の対象となる決算日を3月31日に限定して行っていることに注意しよう．

これらの処理の他にも，粗データをRを用いて解析を行う上で扱いやすい形式[25]に整形した上で利用していることに注意しよう．詳細はRのスクリプト（付録B）を参照されたい．

付録B　Rスクリプト

本稿において利用されたRのスクリプトを以下に与える：

```
#-------------------------------------------------------------------------------------------
# Rとデータベースとのコネクションの確立
#-------------------------------------------------------------------------------------------
library(RODBC)
con <- odbcConnect('MyNEEDS')
#-------------------------------------------------------------------------------------------
# 東証一部上場企業の財務データの取得
#-------------------------------------------------------------------------------------------
rawdata<-sqlQuery(con,"SELECT
fc01.a04 as 'nikkei.code',
firmlist.shamei_en as 'name',
fc01.a02 as 'date',
fc01.a07 as 'term',
fc01.a15 as 'market',
fc01.a22 as 'sector',
fc01.b001 as 'sales',
fe01.b056 as 'employee',
fb01.b067 as 'assets'
FROM fc01
JOIN fb01     ON fc01.a04=fb01.a04 AND fc01.a02=fb01.a02
JOIN fe01     ON fc01.a04=fe01.a04 AND fc01.a02=fe01.a02
JOIN firmlist ON fc01.a04=firmlist.nikkeicode
WHERE fc01.a23='1'
ORDER BY fc01.a04,fc01.a02")
#-------------------------------------------------------------------------------------------
# -999999999999をNAで置換する
#-------------------------------------------------------------------------------------------
add.NA<-function(obj)
{
n<-dim(obj)[1]
```

25）Rにおけるデータフレーム（data frame）として再整形している．

```
for(i in 1:n) obj[i,]<-ifelse(obj[i,]==-9999999999999,NA,obj[i,])
obj
}
x<-add.NA(rawdata)
#-----------------------------------------------------------------------------------------------
# 日付における欠損値があるデータを取り除く
#-----------------------------------------------------------------------------------------------
y<-x[!is.na(x[,3]),]
#-----------------------------------------------------------------------------------------------
# 決算年月日から年と四半期時期を抽出する関数
#-----------------------------------------------------------------------------------------------
yearquarter<-function(obj)
  {
    year<-format(obj,"%Y")
    quarter<-quarters(obj)
    yearQ<-paste(year,quarter,sep="")
    list(year=year,quarter=quarter,yearQ=yearQ)
  }
#-----------------------------------------------------------------------------------------------
# 個体（会社）の一意性を確保するために会社名と日経会社コードを結合し，
# 決算年，四半期の追加と，業種コードを大中小に分け，場部の情報を削除する．
#-----------------------------------------------------------------------------------------------
firmfin.total<-data.frame(
               paste(gsub(" ","",y$name),y$nikkei.code,sep=""),
               y$date,format(y$date,"%Y"),as.integer(months(y$date,abbreviate=TRUE)),
               as.integer(y$term),quarters(y$date),yearquarter(y$date)$yearQ,
               substr(y$sector,1,1),substr(y$sector,2,3),substr(y$sector,4,6),
               y[,-seq(6)])
names(firmfin.total)<-c(
               "name","date","year","month","term","quarter","yearQ",
               "sector1","sector2","sector3",
               "sales","employee","assets")
rownames(firmfin.total)<-seq(dim(firmfin.total)[1])
firmfin.total$yearQ<-as.character(firmfin.total$yearQ)
head(firmfin.total)
#-----------------------------------------------------------------------------------------------
# 決算期間が 12 カ月に満たない観測を除去
#-----------------------------------------------------------------------------------------------
firmfin12<-subset(firmfin.total,term==12)
head(firmfin12)
#-----------------------------------------------------------------------------------------------
# 3 月 31 日決算の企業のみを抽出
#-----------------------------------------------------------------------------------------------
extract331<-function(obj)
{
  d331<-seq(as.Date("1978-03-31"),len=35,by="1 year")
  x<-obj[obj$date==d331[1],]
  for(i in 2:35) x<-rbind(x,obj[obj$date==d331[i],])
  x<-x[order(x$name),]
}
firmfin331<-extract331(firmfin12)
#-----------------------------------------------------------------------------------------------
# 2012 年 3 月 31 日決算の企業の観測を抽出
#-----------------------------------------------------------------------------------------------
firmfin20120331<-firmfin331[firmfin331$date=="2012-03-31",][,-seq(1,7)]
dimnames(firmfin20120331)[[1]]<-as.vector(firmfin331[firmfin331$date=="2012-03-31",][,1])
head(firmfin20120331)
#-----------------------------------------------------------------------------------------------
# 2012 年 3 月 31 日決算の企業の売上高，従業員数，資産合計の時系列プロットとヒストグラム
#-----------------------------------------------------------------------------------------------
library(ggplot2)
qplot(date,sales,data=firmfin.total,geom="line",group=name)
  + geom_vline(xintercept=as.numeric(as.Date("2012-03-31")),lwd=1)
qplot(sales,data=firmfin20120331,geom="histogram")
  + geom_rug()
```

```
#-----------------------------------------------------------------
qplot(date,employee,data=firmfin.total,geom="line",group=name)
  + geom_vline(xintercept=as.numeric(as.Date("2012-03-31")),lwd=1)
qplot(employee,data=firmfin20120331,geom="histogram")
  + geom_rug()
#-----------------------------------------------------------------
qplot(date,assets,data=firmfin.total,geom="line",group=name)
  + geom_vline(xintercept=as.numeric(as.Date("2012-03-31")),lwd=1)
qplot(assets,data=firmfin20120331,geom="histogram")
  + geom_rug()
#-----------------------------------------------------------------------------------------
# 2012 年 3 月 31 日決算の企業の売上高，従業員数，資産合計の時系列プロットとヒストグラム（対数スケール）
#-----------------------------------------------------------------------------------------
qplot(date,log(sales),data=firmfin.total,geom="line",group=name)
  + geom_vline(xintercept=as.numeric(as.Date("2012-03-31")),lwd=1)
qplot(log(sales),data=firmfin20120331,geom="histogram",binwidth=0.5)
  + geom_rug()
#-----------------------------------------------------------------
qplot(date,log(employee),data=firmfin.total,geom="line",group=name)
  + geom_vline(xintercept=as.numeric(as.Date("2012-03-31")),lwd=1)
qplot(log(employee),data=firmfin20120331,geom="histogram",binwidth=0.5)
  + geom_rug()
#-----------------------------------------------------------------
qplot(date,log(assets),data=firmfin.total,geom="line",group=name)
  + geom_vline(xintercept=as.numeric(as.Date("2012-03-31")),lwd=1)
qplot(log(assets),data=firmfin20120331,geom="histogram",binwidth=0.5)
  + geom_rug()
#-----------------------------------------------------------------------------------------
# 2012 年 3 月 31 日決算の企業の売上高，従業員数，資産合計の対散布図（通常スケール，対数スケール）
#-----------------------------------------------------------------------------------------
library(car)
scatterplotMatrix(firmfin20120331[,-c(1,2,3)],smooth=FALSE)
scatterplotMatrix(log(firmfin20120331[,-c(1,2,3)]),smooth=FALSE)
#-----------------------------------------------------------------------------------------
# 2012 年 3 月 31 日決算の企業の売上高，従業員数，資産合計の三次元散布図（通常スケール，対数スケール）
#-----------------------------------------------------------------------------------------
library(rgl, pos=4)
library(mgcv, pos=4)
#-----------------------------------------------------------------
plot3d(firmfin20120331[,c(4,5,6)],size=5,pch=20)
writeWebGL(width=500, height=550)
rgl.postscript("scatterplot3d.eps", fmt="eps" )
#-----------------------------------------------------------------
plot3d(log(firmfin20120331[,c(4,5,6)]),size=5,pch=20)
writeWebGL(width=500, height=550)
rgl.postscript("scatterplot3dlog.eps", fmt="eps" )
#-----------------------------------------------------------------
# googleVis(Motion Chart) による可視化
#-----------------------------------------------------------------
library(googleVis)
firmfin12$year<-as.numeric(format(firmfin12$year))
Myear<- gvisMotionChart(firmfin12,
                        idvar="name",timevar="year",
                        xvar="employee",yvar="sales",
                        colorvar="sector1",sizevar="assets",
                        options=list(width=700,height=700))
plot(Myear)
#-------------------------------------------------------------------
# 2012 年 3 月 31 日決算の 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対数線形モデルの当てはめ
#-------------------------------------------------------------------
#------------------------------
# 正規線形モデルの当てはめ
#------------------------------
lm.firmfin20120331<-lm(sales~employee+assets,data=firmfin20120331)
```

```
summary(lm.firmfin20120331)
#-------------------------------------------------------------------------------------
# 2012 年 3 月 31 日決算の企業の売上高，従業員数，資産合計の三次元散布図へ
# 標本回帰平面の当てはめ
#-------------------------------------------------------------------------------------
scatter3d(firmfin20120331$assets, firmfin20120331$sales,
          firmfin20120331$employee, fit="linear", residuals=TRUE, bg="white",
          axis.scales=TRUE, grid=TRUE, ellipsoid=FALSE, xlab="assets", ylab="sales",
          zlab="employee")
#------------------------------
# 回帰診断
#------------------------------
par(mfcol=c(2,2))
plot(resid(lm.firmfin20120331),ylab="Residuals")
mtext("Index Plot of Residuals", 3, 0.25, cex = 1)
plot(lm.firmfin20120331,which=c(1,2))
plot(density(resid(lm.firmfin20120331)),main="")
mtext("Density Plot of Residuals",3, 0.25, cex = 1)
par(mfcol=c(1,1))
#------------------------------
#対数正規線形モデルの当てはめ
#------------------------------
log.lm.firmfin20120331<-lm(log(sales)~log(employee)+log(assets),data=firmfin20120331)
summary(log.lm.firmfin20120331)
#------------------------------
# 回帰診断
#------------------------------
par(mfcol=c(2,2))
plot(resid(log.lm.firmfin20120331),ylab="Residuals")
mtext("Index Plot of Residuals", 3, 0.25, cex = 1)
plot(log.lm.firmfin20120331,which=c(1,2))
plot(density(resid(log.lm.firmfin20120331)),main="")
mtext("Density Plot of Residuals", 3, 0.25, cex = 1)
par(mfcol=c(1,1))
influenceIndexPlot(log.lm.firmfin20120331,id.n=4,
                   vars=c("hat","Studentized","Cook"),id.cex=0.4)
influencePlot(log.lm.firmfin20120331,id.n=3)
#-------------------------------------------------------------------------------------
# 2012年 3 月 31 日決算の企業の売上高，従業員数，資産合計の対数スケールの三次元散布図へ
# 標本回帰平面の当てはめ
#-------------------------------------------------------------------------------------
scatter3d(logfirmfin20120331$assets, logfirmfin20120331$sales,
          logfirmfin20120331$employee, fit="linear", residuals=TRUE, bg="white",
          axis.scales=TRUE, grid=TRUE, ellipsoid=FALSE, xlab="assets", ylab="sales",
          zlab="employee")
#--------------------------------------
# 異常値の除去とデータフレームの再生成
#--------------------------------------
tmp<-firmfin20120331[dimnames(firmfin20120331)[[1]]!="JAPANSECURITIESFINANCE70514",]
tmp<-tmp[dimnames(tmp)[[1]]!="ORIENT1709",]
tmp<-tmp[dimnames(tmp)[[1]]!="HITACHICAPITAL1711",]
firmfin20120331.new<-tmp[dimnames(tmp)[[1]]!="JACCS1710",]
logfirmfin20120331<-data.frame(firmfin20120331[,c(1,2,3)],log(firmfin20120331[,-c(1,2,3)]))
logfirmfin20120331.new<-data.frame(
                        firmfin20120331.new[,c(1,2,3)],
log(firmfin20120331.new[,-c(1,2,3)]))
#------------------------------
#対数正規線形モデルの再当てはめ
#------------------------------
log.lm.firmfin20120331.new<-lm(log(sales)~log(employee)+log(assets),data=firmfin20120331.new)
summary(log.lm.firmfin20120331.new)
#-------------------------------------------------------------------------------------
# 2012 年 3 月 31 日決算の企業の売上高，従業員数，資産合計の対数スケールの三次元散布図へ
# 標本回帰平面の当てはめ
#-------------------------------------------------------------------------------------
```

```
scatter3d(logfirmfin20120331.new$assets, logfirmfin20120331.new$sales,
          logfirmfin20120331.new$employee, fit="linear", residuals=TRUE, bg="white",
          axis.scales=TRUE, grid=TRUE, ellipsoid=FALSE, xlab="assets", ylab="sales",
          zlab="employee")
#------------------------------
# 回帰診断
#------------------------------
par(mfcol=c(2,2))
plot(resid(log.lm.firmfin20120331.new),ylab="Residuals")
mtext("Index Plot of Residuals", 3, 0.25, cex = 1)
plot(log.lm.firmfin20120331.new,which=c(1,2))
plot(density(resid
(log.lm.firmfin20120331.new)),main="")
mtext("Density Plot of Residuals", 3, 0.25, cex = 1)
par(mfcol=c(1,1))
#-------------------------------------------------------------------------------------------
# ダミー変数月対数正規線形モデルの当てはめ
#-------------------------------------------------------------------------------------------
log.lm2.firmfin20120331.new<-lm(log(sales)~log(employee)+log(assets)+sector2,
         data=firmfin20120331.new)
summary(log.lm2.firmfin20120331.new)
#-------------------------------------------------------------------------------------------
# 2012 年 3 月 31 日決算の企業の売上高，従業員数，資産合計の対数スケールの三次元散布図へ
# 標本回帰平面群の当てはめ
#-------------------------------------------------------------------------------------------
scatter3d(logfirmfin20120331.new$assets, logfirmfin20120331.new$sales,
          logfirmfin20120331.new$employee,
          fit="linear", residuals=TRUE, groups=logfirmfin20120331.new$sector2,
          parallel=TRUE, bg="white",
          surface.col=topo.colors(33),
          axis.scales=TRUE, grid=TRUE, ellipsoid=FALSE,
          xlab="assets", ylab="sales", zlab="employee")
#-------------------------------------------------------------------------------------------
# 決定係数に関する時系列プロット
#-------------------------------------------------------------------------------------------
OLS.ts331<-function(obj)
{
  date<-sort(unique(obj$date))n.ts<-NULL
  R2.ts<-NULL
  df.ts<-NULL
  for(i in date)
  {
  n.ts<-c(n.ts,dim(obj[obj$date==i,])[1])
  lm.obj<-lm(log(sales)~log(employee)+log(assets)+sector2, data=obj[obj$date==i,])
  R2.ts<-rbind(R2.ts,c(summary(lm.obj)$r.squared,summary(lm.obj)$adj.r.squared))
  df.ts<-c(df.ts,lm.obj$df.residual)
  }
  result<-data.frame(date,n.ts,df.ts,R2.ts)
  colnames(result)<-c("date","nfirm","df","R2","adj.R2")
  result
}
firmfin331.OLS.ts<-OLS.ts331(firmfin331)
plot.OLS.ts<-function(obj)
{
  require(ggplot2)
  require(reshape)
  qplot(date,value,data=melt(obj,id="date"),geom=c("point","line"),group=variable)
  + facet_grid(variable~.,scale="free_y")
}
plot.OLS.ts(firmfin331.OLS.ts)
#-------------------------------------------------------------------------------------------
```

付録C　Google Motion Chart と googleVis

ここでは，本稿で利用した可視化ツールである Google Motion Chart とそれを取りまく環境について紹介する．まず，Google Charts Tools は，Web 上でデータを可視化するために Google によって開発されたグラフィックツールである[26]．このツールの一つである Google Motion Chart は，Hans Rosling[27] を代表とする財団によって開発されたソフトウェア Gapminder Word（図23参照.）にもとづて開発されており，時間とともに変化する複数の指標

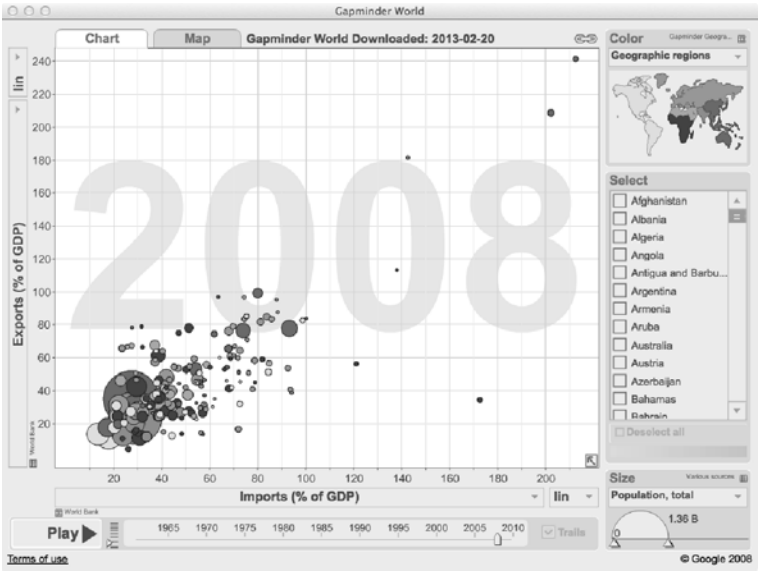

図23　Mac OS X 版 Gapminder World：世界各国の GDP に対する輸入（Imports）と輸出（Exports）の描画例（時間を固定したときはバブルチャートであることに注意）

26）Visualization API と Chart API という二つの API （Application Programming Interface）が公開されている．

27）TED talk（http://www.ted.com/）によるプレゼンテーションが契機となって広く知られるようになった．（Hans Rosling による一連の TED Talk を参照せよ.）

（変量）の関係を動的に表示するためのツールである[28]．

一方，googleVis[29] は R と Google Charts Tools のインターフェースを提供するパッケージである．（Gesmann and de Castillo（2011）参照.）本稿では，R からこのパッケージを利用することによって，Motion Chart を用いて財務データを可視化していることに注意しよう．

付録D　多変量対数正規分布

p 変量ベクトル $\boldsymbol{X}$ が同時確率密度関数：

$$f(\boldsymbol{x}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma}):=\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}\prod_{i=1}^{p}x_i}\exp\left\{-\frac{1}{2}(\log(\boldsymbol{x})-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\log(\boldsymbol{x})-\boldsymbol{\mu})\right\}$$

をもつとき，多変量対数正規分布 $\mathrm{MLN}_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ に従うといい，$\boldsymbol{X}\sim\mathrm{MLN}_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ とかく．以下のことが成り立つことに注意しよう：

$$\boldsymbol{X}\sim\mathrm{MLN}_p(\boldsymbol{\mu},\boldsymbol{\Sigma})\Longrightarrow\log\boldsymbol{X}\sim\mathrm{MN}_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$$

ここで，$\mathrm{MN}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ は平均ベクトル $\boldsymbol{\mu}$，分散共分散行列 $\boldsymbol{\Sigma}$ の多変量正規分

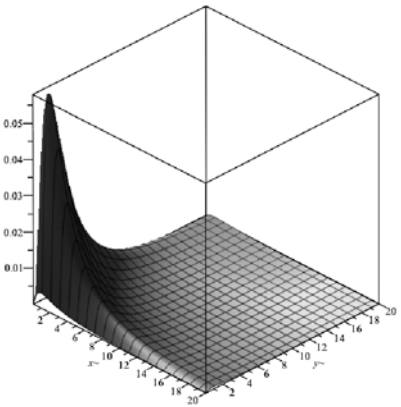

図24　二変量対数正規分布の同時確率密度関数：$\boldsymbol{\mu}=[1,\ 1]'$，$\boldsymbol{\Sigma}=\mathbf{I}_2$ のとき

28) https://developers.google.com/chart/interactive/docs/gallery/motionchart?hl=ja
29) http://code.google.com/p/google-motion-charts-with-r/

布を表す．図24に $\boldsymbol{\mu}=[1,\ 1]'$，$\boldsymbol{\Sigma}=\mathbf{I}_2$（；2次単位行列）のときの二変量対数正規分布の同時確率密度関数のプロットを与える．

付録E　データの分布法則と回帰モデルの親和性

一般に，$(p+1)$ 変量ベクトル $[\boldsymbol{X}',\ Y]'$ に関して，以下の統計的回帰問題を考える：

$$\rho(r):=\mathrm{E}|\,Y-r(\boldsymbol{X})\,|^2 \longrightarrow \min_{r\in\mathcal{L}^2}$$

ここで，ρ は予測誤差（prediction error）と呼ばれ，r は回帰関数（regression function）と呼ばれる．

この問題に対して，予測誤差を最小にする回帰関数は最良予測量（best predictor）と呼ばれ，以下のように与えられる：

$$r^*(\boldsymbol{X}):=\mathrm{E}(Y|\boldsymbol{X})$$

すなわち，最良予測量は条件付き平均で与えられる．（たとえば，稲垣（2003）を参照のこと．）

例として，$(p+1)$ 変量ベクトルが多変量正規分布に従う場合を考える．すなわち，

$$\begin{bmatrix}\boldsymbol{X}\\ Y\end{bmatrix} \sim \mathsf{MN}_{p+1}\left(\begin{bmatrix}\boldsymbol{\mu}_X\\ \mu_Y\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma}_{XX} & \boldsymbol{\sigma}_{XY}\\ \boldsymbol{\sigma}_{XY}' & \sigma_Y^2\end{bmatrix}\right) \tag{9}$$

とすると，その最良予測量は，

$$r^*(\boldsymbol{X})=\mu_Y+\boldsymbol{\sigma}_{XY}'\boldsymbol{\Sigma}_{XX}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_X)$$

で与えられる．

この予測量を実際に利用するには，未知パラメータの推定を行う必要があるが，上記の多変量正規分布に従う n 組の多変量データ

$$\begin{bmatrix}\boldsymbol{x}_i\\ y_i\end{bmatrix},\ i=1,\ \ldots,\ n$$

が与えられたとき，最尤推定値（maximum likelihood estimates）が，それぞれ，

$$\hat{\boldsymbol{\mu}}_X := \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i =: \bar{\boldsymbol{x}}, \quad \hat{\mu}_Y := \frac{1}{n}\sum_{i=1}^{n} y_i =: \bar{y}$$

$$\hat{\boldsymbol{\Sigma}}_{XX} := \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})', \quad \hat{\boldsymbol{\sigma}}_{XY} := \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(y_i-\bar{y})$$

で与えられることを利用すると，$\boldsymbol{X}=\boldsymbol{x}$ が与えられたときの Y に対する最良予測量の推定値が，

$$\hat{r}^*(\boldsymbol{x}) = \hat{\mu}_Y + \hat{\boldsymbol{\sigma}}_{XY}\hat{\boldsymbol{\Sigma}}_{XX}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_X) \tag{10}$$

で与えられる．

一方，正規線形モデル

$$Y_i = \beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta}_p + \varepsilon_i, \quad \varepsilon_i \sim \mathsf{N}(0,\ \sigma^2) \tag{11}$$

を考え，データセット $(\boldsymbol{x}_i',\ y_i) = (x_{i1},\ \ldots,\ x_{ip},\ y_i)$，$(i=1,\ \ldots,\ n)$ が与えられたときの線形回帰モデルは，

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

と書くことができる．ここで，

$$\boldsymbol{y} := \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix},\ \mathbf{X} := \begin{bmatrix} 1 & \boldsymbol{x}_1' \\ \vdots & \vdots \\ 1 & \boldsymbol{x}_n' \end{bmatrix},\ \boldsymbol{\beta} := \begin{bmatrix} \beta_0 \\ \boldsymbol{\beta}_p \end{bmatrix},\ \boldsymbol{\epsilon} := \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

とおいた．このモデルにおいて未知の回帰係数ベクトル $\boldsymbol{\beta}$ の最小自乗推定値（または同一の結果を与えるが最尤推定値）は，

$$\hat{\boldsymbol{\beta}} := (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y} = \begin{bmatrix} \bar{y} - \boldsymbol{y}'\mathbf{X}_{Cp}(\mathbf{X}_{Cp}'\mathbf{X}_{Cp})^{-1}\bar{\boldsymbol{x}} \\ (\mathbf{X}_{Cp}'\mathbf{X}_{Cp})^{-1}\mathbf{X}_{Cp}'\boldsymbol{y} \end{bmatrix} =: \begin{bmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}}_p \end{bmatrix}$$

で与えられる．線形回帰モデルをデータに当てはめる際に利用される標本回帰（超）平面は，この推定値を利用して，

$$\hat{\eta}(\boldsymbol{x}) := \hat{\beta}_0 + \hat{\boldsymbol{\beta}}_p'\boldsymbol{x} = \bar{y} + \boldsymbol{y}'\mathbf{X}_{Cp}(\mathbf{X}_{Cp}'\mathbf{X}_{Cp})^{-1}(\boldsymbol{x}-\bar{\boldsymbol{x}}) \tag{12}$$

で与えられる．ここで，$\mathbf{X}_{Cp}' := \mathbf{X}_p - \mathbf{1}\bar{\boldsymbol{x}}'$ である．ただし，$\mathbf{1} := [1,\ \ldots,\ 1]'$ とし，$\mathbf{X} := [\mathbf{1},\ \mathbf{X}_p]$ とおいた．

以上の結果において，多変量正規性(9)のもとでの最良予測量の推定値(10)と，正規線形モデル(11)のもとでの標本回帰平面(12)に関して以下の同値性が成り立

つことに注意しよう：

$$\hat{r}^*(\boldsymbol{x})=\hat{\mu}_Y+\hat{\boldsymbol{\sigma}}'_{XY}\hat{\boldsymbol{\Sigma}}_{XX}^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_X)=\overline{y}+\boldsymbol{y}'\boldsymbol{X}_{C_p}(\boldsymbol{X}'_{C_p}\boldsymbol{X}_{C_p})^{-1}(\boldsymbol{x}-\overline{\boldsymbol{x}})=\hat{\eta}(\boldsymbol{x}) \quad (13)$$

このように変量にある種の分布法則を仮定し，そのもとで一般回帰問題を解くことにをよって導かれる結果が，ある回帰モデルを適用することによって得られる推定結果と同等であるとき，本稿では，親和的（compatible）であると呼ぶことにする．すなわち，正規線形モデルは多変量正規分布に関して親和的であることに注意しよう．

親和性をもつもう一つの例を考える．（$p+1$）変量ベクトルが多変量対数正規分布に従う場合：

$$\begin{bmatrix}\boldsymbol{X}\\ Y\end{bmatrix}\sim \mathsf{MLN}_{p+1}\left(\begin{bmatrix}\boldsymbol{\mu}_X\\ \mu_Y\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma}_{XX} & \boldsymbol{\sigma}_{XY}\\ \boldsymbol{\sigma}'_{XY} & \sigma_Y^2\end{bmatrix}\right) \quad (14)$$

に，その最良予測量は，

$$r^*{}_{\mathsf{MLN}}(\widetilde{\boldsymbol{X}}):=\mu_Y+\boldsymbol{\sigma}'_{XY}\boldsymbol{\Sigma}_{XX}^{-1}(\widetilde{\boldsymbol{X}}-\boldsymbol{\mu}_X)$$

で与えられる．ここで，$\widetilde{\boldsymbol{X}}:=\log \boldsymbol{X}:=[\log X_1,\ \ldots,\ \log X_p]'$ である．

この予測量を実際に利用するには，未知パラメータの推定を行う必要があるが，上記の多変量対数正規分布に従う n 組の多変量データが与えられたとき，最尤推定値（maximum likelihood estimates）が，$\widetilde{\boldsymbol{x}}_i:=\log \boldsymbol{x}_i$, $\widetilde{y}_i:=\log y_i$ とおくことによって，それぞれ，

$$\widetilde{\boldsymbol{\mu}}_X:=\frac{1}{n}\sum_{i=1}^{n}\widetilde{\boldsymbol{x}}_i=:\overline{\widetilde{\boldsymbol{x}}},\quad \widetilde{\mu}_Y:=\frac{1}{n}\sum_{i=1}^{n}\widetilde{y}_i=:\overline{\widetilde{y}}$$

$$\widetilde{\boldsymbol{\Sigma}}_{XX}:=\frac{1}{n}\sum_{i=1}^{n}(\widetilde{\boldsymbol{x}}_i-\overline{\widetilde{\boldsymbol{x}}})(\widetilde{\boldsymbol{x}}_i-\overline{\widetilde{\boldsymbol{x}}})',\quad \widetilde{\boldsymbol{\sigma}}_{XY}:=\frac{1}{n}\sum_{i=1}^{n}(\widetilde{\boldsymbol{x}}_i-\overline{\widetilde{\boldsymbol{x}}})(\widetilde{y}_i-\overline{\widetilde{y}})$$

で与えられることを利用すると，$\boldsymbol{X}=\boldsymbol{x}$ が与えられたときの Y に対する最良予測量の推定値が，

$$\widetilde{r}^*(\widetilde{\boldsymbol{x}})=\widetilde{\mu}_Y+\overline{\boldsymbol{\sigma}}'_{XY}\widetilde{\boldsymbol{\Sigma}}_{XX}^{-1}(\widetilde{\boldsymbol{x}}-\widetilde{\boldsymbol{\mu}}_X) \quad (15)$$

で与えられる．ただし，$\widetilde{\boldsymbol{x}}:=\log \boldsymbol{x}$ とおいた．

一方，対数正規線形モデル

$$Y_i=\gamma\times\prod_{j=1}^{p}x_{ij}^{\alpha_j}\times\varepsilon_i,\quad \varepsilon_i\sim \mathsf{LN}(0,\ \sigma^2) \quad (16)$$

を考え，両辺の対数をとることによって，

$$\log Y_i = \alpha_0 + \sum_{j=1}^{p} \alpha_j \log x_{ij} + \log \varepsilon_i, \quad \log \varepsilon_i \sim \mathsf{N}(0,\ \sigma^2) \tag{17}$$

なる正規線形モデル表現が得られることに注意する．このとき，データセット $(\boldsymbol{x}_i',\ y_i) = (x_{i1},\ \ldots,\ x_{ip},\ y_i)$，$(i = 1,\ \ldots,\ n)$ に対して，$(\tilde{\boldsymbol{x}}_i',\ \tilde{y}_i) := (\log \boldsymbol{x}_i',\ \log y_i) = (\log x_{i1},\ \ldots,\ \log x_{ip},\ \log y_i)$ とおくと，対数正規線形モデルの正規線形表現に対する線形回帰モデルは，

$$\tilde{\boldsymbol{y}} = \tilde{\mathbf{X}}\boldsymbol{\alpha} + \tilde{\boldsymbol{\epsilon}}$$

と書くことができる．ここで，

$$\tilde{\boldsymbol{y}} := \begin{bmatrix} \tilde{y}_i \\ \vdots \\ \tilde{y}_n \end{bmatrix},\ \tilde{\mathbf{X}} := \begin{bmatrix} 1 & \tilde{\boldsymbol{x}}_i \\ \vdots & \vdots \\ 1 & \tilde{\boldsymbol{x}}_n' \end{bmatrix},\ \boldsymbol{\alpha} := \begin{bmatrix} \alpha_0 \\ \boldsymbol{\alpha}_p \end{bmatrix},\ \tilde{\boldsymbol{\epsilon}} := \begin{bmatrix} \tilde{\epsilon}_i \\ \vdots \\ \tilde{\epsilon}_n \end{bmatrix}$$

とおいた．このモデルにおいて未知の回帰係数ベクトル $\boldsymbol{\alpha}$ の最小自乗推定値（または同一の結果を与えるが最尤推定値）は，

$$\tilde{\boldsymbol{\alpha}} := (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\boldsymbol{y}} = \begin{bmatrix} \bar{\tilde{y}} - \tilde{\boldsymbol{y}}'\tilde{\mathbf{X}}_{Cp}(\tilde{\mathbf{X}}_{Cp}'\tilde{\mathbf{X}}_{Cp})^{-1}\bar{\boldsymbol{x}} \\ (\mathbf{X}_{Cp}'\tilde{\mathbf{X}}_{Cp})^{-1}\tilde{\mathbf{X}}_{Cp}'\tilde{\boldsymbol{y}} \end{bmatrix} =: \begin{bmatrix} \tilde{\alpha}_0 \\ \tilde{\boldsymbol{\alpha}}_p \end{bmatrix}$$

で与えられる．線形回帰モデルをデータに当てはめる際に利用される標本回帰（超）平面は，この推定値を利用して，

$$\tilde{\eta}(\tilde{\boldsymbol{x}}) := \tilde{\alpha}_0 + \tilde{\boldsymbol{\alpha}}_p'\boldsymbol{x} = \bar{\tilde{y}} + \tilde{\boldsymbol{y}}'\tilde{\mathbf{X}}_{Cp}(\tilde{\mathbf{X}}_{Cp}\tilde{\mathbf{X}}_{Cp})^{-1}(\tilde{\boldsymbol{x}} - \bar{\tilde{\boldsymbol{x}}}) \tag{18}$$

で与えられる．ここで，$\tilde{\mathbf{X}}_{Cp} ; = \tilde{\mathbf{X}}_p - \mathbf{1}\bar{\tilde{\boldsymbol{x}}}'$ である．

以上の結果から，多変量対数正規性(14)のもとでの最良予測量の推定値(15)と，対数正規線形モデル(16)のもとでの標本回帰平面(18)に関して，

$$\tilde{r}^*(\tilde{\boldsymbol{x}}) = \tilde{\mu}_Y + \tilde{\boldsymbol{\sigma}}_{XY}'\tilde{\boldsymbol{\Sigma}}_{XX}^{-1}(\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{\mu}}_X) = \bar{\tilde{y}} + \tilde{\boldsymbol{y}}'\tilde{\mathbf{X}}_{Cp}(\tilde{\mathbf{X}}_{Cp}'\tilde{\mathbf{X}}_{Cp})^{-1}(\tilde{\boldsymbol{x}} - \bar{\tilde{\boldsymbol{x}}}) = \tilde{\eta}(\tilde{\boldsymbol{x}}) \tag{19}$$

が成り立つことから，対数正規線形モデルは多変量対数正規分布に関して親和的であることがわかる．この結果から，データが多変量正規分布に従うときは，通常の正規線形モデルを当てはめるよりも，対数正規線形モデルを当てはめる方が親和性の観点から自然であると考えられる．

付録F　線形回帰モデルと最小自乗法

線形回帰モデル

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

において，

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

とおく．

最小自乗法は，誤差平方和：

$$\Delta^2(\boldsymbol{\beta}):=\sum_{i=1}^{n}\epsilon_i^2=\boldsymbol{\epsilon}'\boldsymbol{\epsilon}=(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})'(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})$$

を回帰係数ベクトル$\boldsymbol{\beta}$に関して最小にすることによって定式化される：

$$\Delta^2(\boldsymbol{\beta})\longrightarrow\min_{\boldsymbol{\beta}\in\mathbb{R}^{d+1}}$$

この最小化問題は，極値問題に置き換えることによって解かれる．

$$\frac{\partial\Delta^2(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}=\mathbf{0}\iff\mathbf{X}'\mathbf{X}\boldsymbol{\beta}=\mathbf{X}'\boldsymbol{y}$$

この方程式は正規方程式（normal equation）と呼ばれ，その解は通常最小自乗（Ordinary Least Squares: OLS）推定値と呼ばれる．

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

最小自乗推定値を係数としてもつ（超）平面：

$$\hat{\eta}(\boldsymbol{x}):=\hat{\beta}_0+\hat{\beta}_1x_1+\cdots+\hat{\beta}_px_p$$

を標本回帰（超）平面と呼ぶ．$\boldsymbol{x}=\boldsymbol{x}_i$のときの平面上の点を当てはめ値（fitted value）といい，

$$\hat{y}:=\hat{\beta}_0+\hat{\beta}_1x_{i1}+\cdots+\hat{\beta}_px_{ip}$$

で定義される．また，実際の観測点y_iと当てはめ値$\hat{y}_i$の差は残差（residual）と呼ばれ以下のように定義される：

$$e_i := y_i - \widehat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \cdots - \widehat{\beta}_p x_{ip}$$

さらに，当てはめ値のベクトルは，回帰ベクトル（regression vector）と呼ばれ，

$$\widehat{y} := \begin{bmatrix} \widehat{y}_1 \\ \vdots \\ \widehat{y}_n \end{bmatrix} = \begin{bmatrix} \widehat{\beta}_0 + \widehat{\beta}_1 x_{11} + \cdots + \widehat{\beta}_p x_{1p} \\ \vdots \\ \widehat{\beta}_0 + \widehat{\beta}_1 x_{n1} + \cdots + \widehat{\beta}_p x_{np} \end{bmatrix} = \mathbf{X}\widehat{\boldsymbol{\beta}}$$

で定義される．また，残差のベクトルは，残差ベクトル（residual vector）と呼ばれ，

$$\boldsymbol{e} := \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} y_1 - \widehat{y}_1 \\ \vdots \\ y_n - \widehat{y}_n \end{bmatrix} = \boldsymbol{y} - \widehat{\boldsymbol{y}} = \boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$$

と定義される．以下の対応がつくことに注意しよう：

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \overset{\text{relative}}{\Longleftrightarrow} y = \mathbf{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{e}$$

つまり，回帰係数ベクトル $\boldsymbol{\beta}$ と最小自乗推定値ベクトル $\widehat{\boldsymbol{\beta}}$，誤差ベクトル $\boldsymbol{\epsilon}$ と残差ベクトル $\boldsymbol{e}$ は対応していることに注意しよう．このことから，（直接観測できない）誤差に関する仮定を検証するために残差が利用されることの根拠となっている．

誤差分散 σ^2 の推定には，

$$\widehat{\sigma}^2 := \frac{1}{n-p-1}\Delta^2(\widehat{\boldsymbol{\beta}})$$

が利用され，ここで

$$\Delta^2(\widehat{\boldsymbol{\beta}}) := \sum_{i=1}^{n} e_i^2 = \boldsymbol{e}'\boldsymbol{e} = (\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$$

は，残差平方和（residuals sum of squares）である．なお，

$$\widehat{\sigma} := \sqrt{\widehat{\sigma}^2}$$

を誤差の標準誤差と呼び，モデルの精度を表すことに注意しよう．

射影行列（projection matrix）：

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

を導入することによって，回帰ベクトルと残差ベクトルは以下のように書くことができる：

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

射影行列の性質（対称性，巾等性）を利用することによって以下のような幾何学的な性質が導かれる：

加法性：　$\boldsymbol{y} = \hat{\boldsymbol{y}} + \boldsymbol{e}$

直交性：　$\hat{\boldsymbol{y}} \perp \boldsymbol{e}$

拡張されたピタゴラスの定理：　$\|\boldsymbol{y}\|^2 = \|\hat{\boldsymbol{y}}\|^2 + \|\boldsymbol{e}\|^2$

付録G　回帰分析における感度分析のための指標

ここでは，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．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（$\boldsymbol{x}_i'$, y_i）を取り除いて計算したもの同士を比較したり，データ点を取り除かないで計算されたものと比較したりすることによって，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ことに注意しよう．すなわち，

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

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，$\hat{\sigma}_{(-i)} := \sqrt{\hat{\sigma}^2_{(-i)}}$ であり，$\hat{\sigma}^2_{(-i)}$ は，i 番目のデータ点（$\boldsymbol{x}_i'$, 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（$\boldsymbol{x}_i'$, 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	化合繊	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

大分類	中分類	小分類	大分類	中分類	小分類
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
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	銑鋼一貫	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

大分類	中分類	小分類	大分類	中分類	小分類
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
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

大分類	中分類	小分類	大分類	中分類	小分類
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
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