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

地　道　正　行

要　旨

探索的データ解析の視点にたち，全世界の上場企業の財務データを可視化することによって得られた知見にもとづいて，売上高を従業員数と資産合計で説明するための統計モデリングを行った．正規線形モデルや対数正規線形モデルを含む考察したモデルのうち，結果として対数非対称正規線形モデルが赤池情報量規準の意味で最も良いものであるという結論を得ることができた．

キーワード：探索的データ解析（Exploratory Data Analysis），統計モデリング（Statistical Modeling），データ可視化（Data Visualization），非対称正規分布（Skew-normal Distribution），対数非対称正規線形モデル（Log-skew-normal Linear Model）

I　はじめに

地道（2014）と Jimichi and Maeda（2014）では，地道（2010-a, b）によって構築された日経 NEEDS 財務データにもとづくデータベースから抽出された東京証券取引所１部上場企業の財務データを Tukey（1977）による探索的データ解析（Exploratory Data Analysis: EDA）の視点に立ち，様々なデータ可視化（data visualization）の結果にもとづいて，売上高を従業員数と資産合計を用いた対数正規線形モデル（log-normal linear model）による統計モデリング（statistical modeling）について議論した．なお，東京証券取引所１部上場企業を対象としており，単年度当たりで扱った企業数は1500社程度

である．

本稿では，Saka and Jimichi（2015）で扱った全世界の上場企業のデータを利用し探索的データ解析の視点に立って統計モデリングを行う．実質的なデータ規模は20000社以上のものとなっていることに注意しよう．なお，この規模の拡大に起因して地道（2014）では有効であった対数正規線形モデルではデータの変動を説明することが難しいという問題が発生したため，本稿では新たな統計モデリングを試みている．

本稿の構成は以下のようなものである．まず，Ⅱ節では，本稿で扱うデータについて述べる．今回利用するデータは，対象企業を世界規模に拡大し，Bureau van Dijk（BvD）社[1] から提供される全世界上場企業データベースosiris[2] から抽出されたものを利用する．次に，Ⅲ節ではⅡ節で与えられたデータを幾つかの観点から可視化することによって分布特性に関する知見を得る．この結果の一つとして売上高が対数正規分布（log-normal distribution）よりも若干歪んだ分布に従っていることがわかる．なお，この現象を説明する一つのモデルとして，Ⅳ節では対数非対称正規分布（log-skew-normal distribution）への当てはめが議論される．また，この節での知見を生かして，Ⅴ節では売上高を従業員数と資産合計で説明する回帰モデルの構築を試みる．その際，基本的な正規線形モデル（normal linear model）の当てはめから始め，対数正規線形モデル，対数非対称正規線形モデル（log-skew-normal linear model）を利用した線形モデルを扱うことに注意されたい．なお，これらのモデルを当てはめた結果は赤池情報量規準（Akaike's Information Criterion: AIC）によって比較・検討される．最後に，Ⅵ節で本稿のまとめと今後の課題を与える．

1） ビューロー・ヴァン・ダイク社 http://www.bvdinfo.com/ja-jp/home

2） osiris（オシリス）は全世界の上場企業約80000社の情報が国際比較可能な統一のフォームで収録されたデータベースである．収録情報としては，世界の上場企業及び上場廃止・非上場企業・一般事業会社・銀行・保険会社の財務三表（BS/PL/CF）等の財務情報が平均15年（最長30年）にわたって収録されている．なお，その他にも株主／関連会社情報や株価情報，企業概要等の情報も含んでいる．

本稿ではデータ解析環境 R[3] を用いており，データの可視化には ggplot2, rgl パッケージ，データ操作には dplyr パッケージ，さらに非対称正規分布に関連する一連の計算には sn パッケージを利用している．なお，付録Aには sn パッケージの説明を与えている．また，付録Bには本稿で使用した R スクリプトを与えている．さらに，本稿は全編を通じて再成可能な研究を行うために，R Noweb（Rnw）ファイルを R による動的文書生成関数 Sweave[4] で処理することによって執筆されていることに注意しよう．

II　データ

本稿では，BvD 社から提供されるデータベース osiris から抽出された2013年決算の全世界上場企業のデータを利用する[5]．実際のデータは以下のようなものである：

表 1　データベース **osiris** から抽出した全世界の上場企業の財務データ（全データ21801件から先頭の10件を抜粋）

	firmID	country	SIC.code	sales	employees	assets.total
1	ELECTROCOMPONENTS PLC GB00647788	UNITED KINGDOM	5065	2118820	6212	1373547
2	AGA RANGEMASTER GROUP PLC GB00354715	UNITED KINGDOM	3631	412359	2516	403137
3	COBHAM PLC GB00030470	UNITED KINGDOM	3728	2947278	10090	3983939
4	REDHALL GROUP PLC GB00263995	UNITED KINGDOM	1799	182661	1225	109687
5	BRISTOL WATER PLC GB02662226	UNITED KINGDOM	4941	206207	489	766077
6	BT GROUP PLC GB04190816	UNITED KINGDOM	4899	30435053	87800	41437741
7	BP PLC GB00102498	UNITED KINGDOM	2911	379136000	83900	305690000
8	BRITISH LAND COMPANY PUBLIC LIMITED COMPANY(THE) GB00621920	UNITED KINGDOM	6531	639091	556	17939489
9	BAE SYSTEMS PLC GB01470151	UNITED KINGDOM	3721	27771636	78000	32410672
10	BRAMMER PLC GB00162925	UNITED KINGDOM	7389	1073549	3241	627595

3）R version 3.3.2（2016-10-31）

4）Sweave（http://www.statistik.lmu.de/~leisch/Sweave/）は，Norman Ramsey による Noweb（https://www.cs.tufts.edu/~nr/noweb/）をベースとして Friedrich Leisch によって開発された動的文書を作成するための環境である．Sweave のファイルは Rnw ファイルと呼ばれることに注意しよう．なお，再成可能研究と動的文書生成については，例えば，Knuth（1984），Gandrud（2015），Xie（2015）を参照されたい．

5）本稿で利用する osiris は 2015年版である．osiris は年毎に収録企業数が増加するなどの更新があることに注意しよう．

ここで，変数名は以下のようなものである：

`firmID`: 企業名と BvD 社の企業コードを結合したもの

`country`: 企業が属する国名

`SIC.code`: SIC（Standard Industrial Classification）コード[6)]

`sales`: 売上高（単位：1000米ドル）

`employees`: 従業員数（単位：人）

`assets.total`: 資産合計（単位：1000米ドル）

本稿では，このデータを用いて売上高（`sales`）を従業員数（`epmployees`）と資産合計（`assets.total`）で説明するための統計モデリングを行う．その際，次節ではこれらの3変量データを可視化することによってモデリングを行う上で重要な知見を探索する．

Ⅲ データ可視化

この節ではデータを幾つかの観点から可視化しよう．まず，対散布図（図1）を描くことによって，2変量間の関係とそれぞれの変量の分布に関する情報を得よう．

この結果から，データは1変量と2変量の両方とも原点付近に集中して分布しており，その意味での「歪み」があることから，正規分布などの左右対称の分布を仮定することが難しいことがわかる．このような歪みを解消しデータを対称に近づけるために対数をとることが一つの方法である．（例えば，Tukey (1977), Mosteller and Tukey (1977), Fox and Weisbrerg (2011)，地道（2014）などを参照されたい．）

図2は各変量の対数をとったものの対散布図である．図2から得られる重要な情報は，対数をとったものもある種の「歪み」があり，いわゆる「正規

6) アメリカ政府により定められた4桁の数字で表されるコードであり，様々な業界をその機能と製品とによって定義するために利用される．上2桁が業界の大分類を表し，下2桁が業界の小分類を表すことに注意しよう．なお，R でこのコードを利用する際には数値として扱った方が便利であることから，上1桁が0のものは省略していることに注意しよう．例えば，`0111`（小麦業界：Wheat）は `111`と表している．

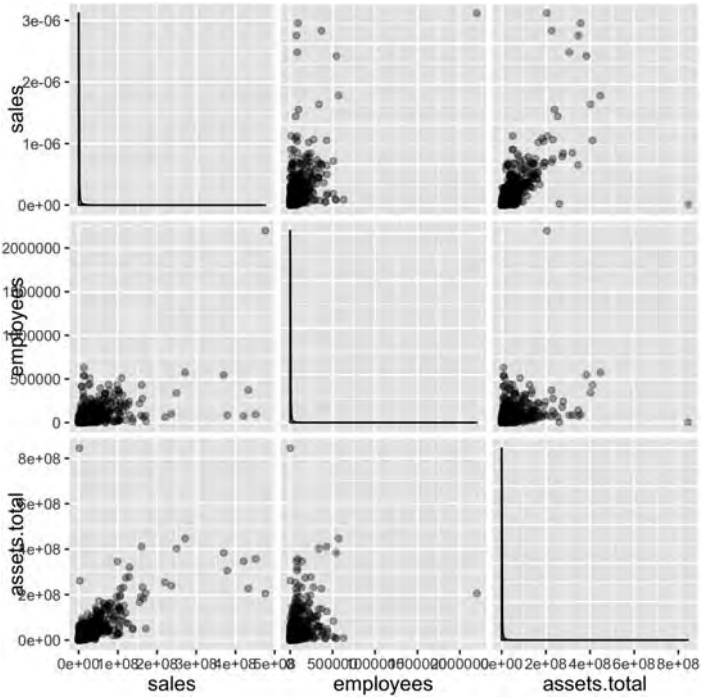

図 1　財務データの対散布図

分布」には従うとはいいがたいということがわかる．

とくに本稿では売上高（sales）を応答変数とする回帰モデルを構築することを考えているため，売上高の可視化を行うことによって，もう少し詳しくその分布構造を検討する．まず，オリジナルのスケールと対数スケールをとったヒストグラムを描こう．

図 3 から，売上高は対数をとることによって正規分布に近づけることはできるけれども，対数スケールのヒストグラムを注意深くみると，若干左側の裾が右の裾に比べて「重い」ことが見て取れる．このことを売上高の対数をとったものの正規 Q-Q プロット（normal Q-Q plot）を描くことによって確かめよう[7]．

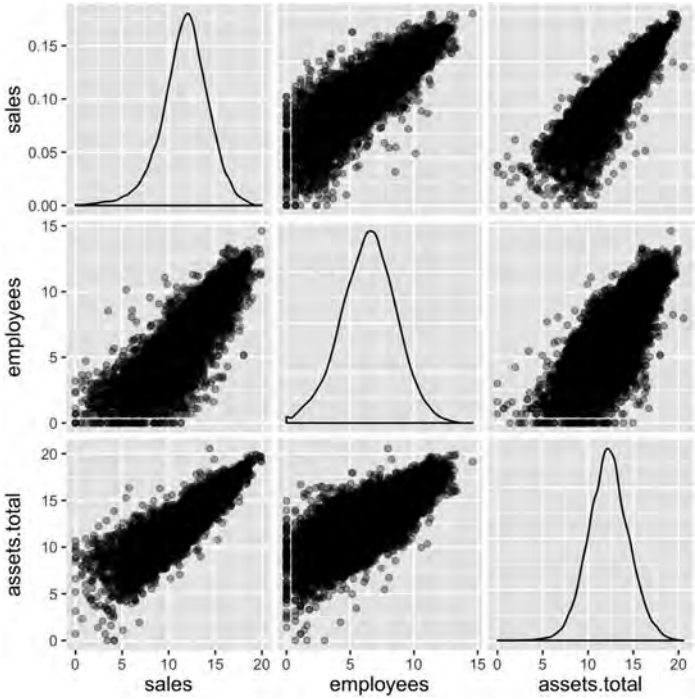

図 2　財務データの対散布図：対数スケール

正規 Q-Q プロットにおいて，正規分布にデータが従う際に理想的には直線上に点が並ぶことであるが，図 4 より明らかに左裾の部分で正規分布との開きがあることがわかる．よって，売上高の対数は正規分布よりも左に歪ん

7）　一般に，Q-Q プロットはデータが特定の分布に従うことを調べるための可視化の方法であり，分布の理論的な分位点（theoritical quantile）とデータ（標本）から求められた経験的な分位点（empirical quantile）を対にして x-y 平面上にプロットしたものである．ここではデータが正規分布に従っているかどうかを，正規分布の分位点とデータの分位点を対にしてプロットしたものである正規 Q-Q プロットを描くことによって確かめている．なお，このプロットと類似したものに理論分布の累積分布関数から求められた確率点（probability point）とデータの経験分布関数から求められた確率点の値を対にしてプロットを行う P-P プロット（P-P plot）がある．Q-Q プロットと P-P プロットの詳細については，例えば，柴田（2015）を参照されたい．

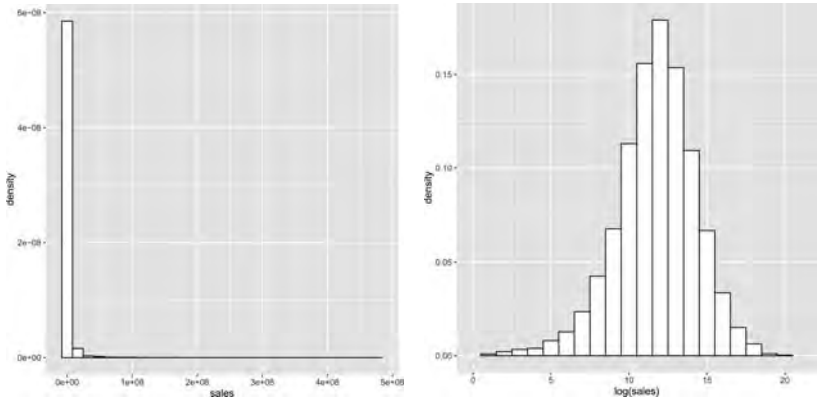

図 3　売上高のヒストグラム：オリジナルのスケール（左）と対数スケール（右）

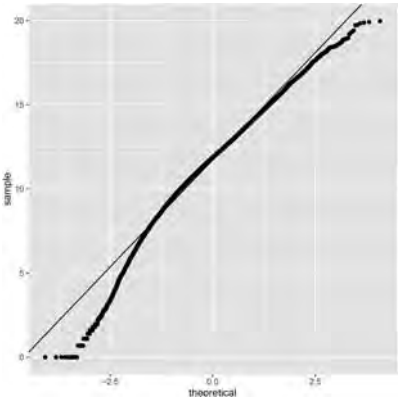

図 4　売上高の対数の正規 **Q-Q** プロット

だ分布であることがわかる．このことを数値的に確かめるためにはデータによる歪度（skewness）：

$$g_1 := \frac{m_3^2}{m_2^{3/2}}$$

を計算し，その符号をみることによって確かめることができる．ここで，$m_j := \sum_{i=1}^{n}(x_i - \bar{x})^j/n$ はデータ $\{x_1, \ldots, x_n\}$ の平均 $\bar{x} = \sum_{i=1}^{n} x_i/n$ まわりの j

次モーメントである．

売上高の対数の歪度は，

$$g_1 = -0.51 (< 0)$$

で与えられ，負の値をとることから左に歪んでいることがわかる．

以上の考察から売上高は対数正規分布には従わず，何らかの別の分布に従っていると考える方が自然である．ただし，データの対数をとったもののヒストグラム（図３の右図）と正規 Q-Q プロット（図４）をみると裾の部分（特に左裾）以外は正規分布で近似できるように推察できる．これらの考察から，正規分布の片方の裾が少し重い分布をモデルとして当てはめることが考えられ，次節では，その候補となる分布について考察する．

Ⅳ　対数非対称正規分布の売上高データへの当てはめ

前節で売上高は対数正規分布に従うとは考えにくいことをみたが，ここではまず正規分布の片方の裾を若干重くするように調整された非対称正規分布（skew-normal distribution）について説明し，次に対数をとったものが非対称正規分布になる場合，すなわち対数非対称正規分布（log-skew-normal distribution）を述べ，このモデルを売上高のデータに当てはめる．

1．非対称正規分布

非対称正規分布は Azzalini（1985）によって提案され，その後理論・応用の両面から研究がなされている[8]．その確率密度関数は以下のように与えられる：

$$f(x \mid \xi, \omega, \alpha) := \frac{2}{\omega} \phi\left(\frac{x-\xi}{\omega}\right) \Phi\left(\alpha \frac{x-\xi}{\omega}\right) \tag{1}$$

ここで，$x \in \mathbb{R} := (-\infty, \infty)$ であり，

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

8）　非対称正規分布に関する最近の研究として，総合的な観点からのものは Azzalini and Capitanio（2014）があり，応用としては大野ら（2011）がある．

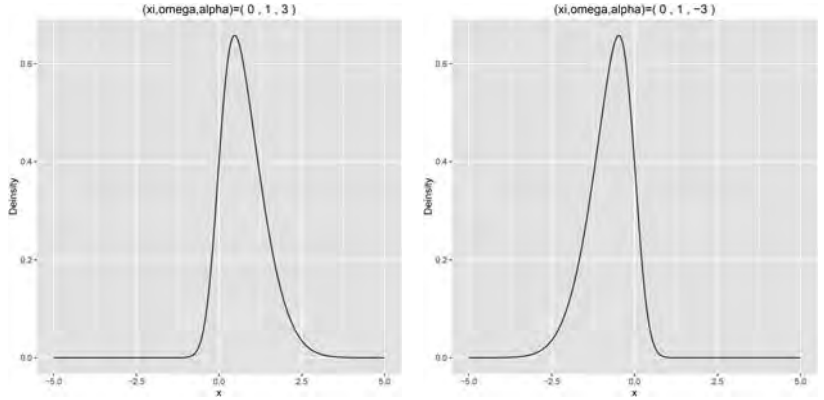

図 5　$(\xi, \omega, \alpha) = (0, 1, 3), (0, 1, -3)$ のときの確率密度関数のプロット

は非対称正規分布の未知母数である．なお，$\mathbb{R}^{+} := (0, \infty)$ である．また，

$$\phi(x) := \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)；\text{標準正規分布の確率密度関数},$$

$$\Phi(x) := \int_{-\infty}^{x} \phi(z)dz；\text{標準正規分布の累積分布関数}$$

である．確率変数 X が上の確率密度関数を持つとき，非対称正規分布に従うといわれ，

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

と記号として書かれる．特に $\alpha = 0$ のとき，

$$f(x|\xi, \omega, 0) = \frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi(0) = \frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\frac{1}{2} = \frac{1}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)$$

となり，通常の正規分布 $\mathsf{N}(\xi, \omega^2)$ の確率密度関数となる．この事実は記号的に以下のように書かれる：

$$\mathsf{SN}(\xi, \omega^2, 0) \overset{\mathrm{d}}{=} \mathsf{N}(\xi, \omega^2)$$

よって，$\alpha = 0$ のときは対称な分布となる．次に，図 5 に $(\xi, \omega, \alpha) = (0, 1, 3), (0, 1, -3)$ のときの確率密度関数のプロットを与える．

このプロットから母数 α が正のとき右に歪み，負のとき左に歪むことが

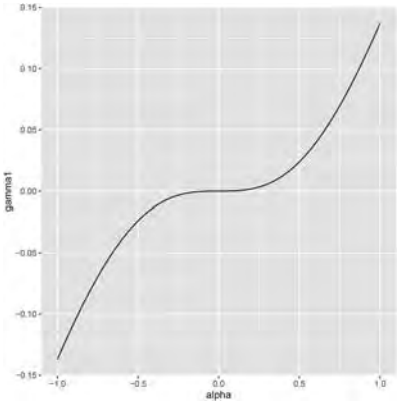

図 6　非対称正規分布の歪度のプロット：$-1 \leq \alpha \leq 1$ の場合

わかる．このことは，数値的には母集団の歪度 γ_1 によって確かめることができる．非対称正規分布の歪度は，$\delta := \alpha/\sqrt{1+\alpha^2}$ とおくことによって，

$$\gamma_1 = \frac{\sqrt{2}(4-\pi)\delta^3}{(\pi - 2\delta^2)^{3/2}} \tag{2}$$

で与えられ，この符号を調べると，

$$\gamma_1 \begin{cases} < 0, & \alpha < 0, \text{（左に歪んでいる）} \\ = 0, & \alpha = 0, \text{（対称）} \\ > 0, & \alpha > 0 \text{（右に歪んでいる）} \end{cases}$$

となり，α の値によって分布の対称-非対称が決定される．このことから，α は非対称母数（skew parameter）と呼ばれることに注意しよう．なお，歪度 (2) の範囲 $-1 \leq \alpha \leq 1$ に関するプロットは図 6 で与えられる．

注意 1（カイ自乗分布性） 確率変数 X が非対称正規分布 $\mathrm{SN}(\xi, \omega^2, \alpha)$ に従うとき，

$$Z := \frac{X - \xi}{\omega}$$

で定義される確率変数の確率密度関数は，(1) より，

$$g(z|\alpha) := 2\phi(z)\Phi(\alpha z) \tag{3}$$

となり，これは $\mathsf{SN}(0, 1, \alpha)$ の確率密度関数である[9]．確率変数 Z に関する興味深い性質としては以下が成り立つことである：

$$Z^2 \sim \chi^2_1 \text{（；自由度 1 のカイ自乗分布）}$$

（Azzalini and Capitanio（2014）の Proposition 2.1（e）を参照.）一般に標準正規分布に従う確率変数の 2 乗が自由度 1 のカイ自乗分布に従うけれども，非対称正規分布に従う確率変数の標準化されたものの 2 乗が同様にカイ自乗分布に従うことは自明ではないことに注意しよう．なお，この結果はデータが非対称正規分布に従うことを確かめる方法として Q-Q プロットや P-P プロットを描く際の理論的根拠となっていることにも注意しよう．

（注意終）

2. 対数非対称正規分布

確率変数 $Y(>0)$ に対して，その対数 $\log(Y)$ が非対称正規分布 $\mathsf{SN}(\xi, \omega^2, \alpha)$ に従うとき，Y は対数非対称正規分布 $\mathsf{LSN}(\xi, \omega^2, \alpha)$ に従うといわれる．

$$Y \sim \mathsf{LSN}(\xi, \omega^2, \alpha) \overset{\text{def}}{\iff} \log(Y) \sim \mathsf{SN}(\xi, \omega^2, \alpha)$$

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

（Azzalini and Capitanio（2014）の p. 53 を参照のこと.）

ここで，売上高（sales）が対数非対称正規分布 $\mathsf{LSN}(\xi, \omega^2, \alpha)$ に従うと仮定し，売上高の対数 log（sales）に非対称正規分布を当てはめることを考える．母数 (ξ, ω, α)[10] を最尤法[11] によって推定した結果は

$$(\hat{\xi}, \hat{\omega}, \hat{\alpha}) = (14.09, 3.48, -1.66)$$

9）正規分布に従う確率変数を標準化したものが標準正規分布に従うことのアナロジーである．

10）非対称正規分布に関する母数付け（parametrization）には，「直接母数」（Direct Parameter: DP）と「中心化母数」（Centered Parameter: CP）がある．本稿では直接母数のみを扱っていることに注意しよう．なお，これらの母数の役割については Azzalini and Capitanio（2014）を参照されたい．

11）非対称正規分布に関する母数の最尤推定についての詳細は，Azzalini and Capitanio

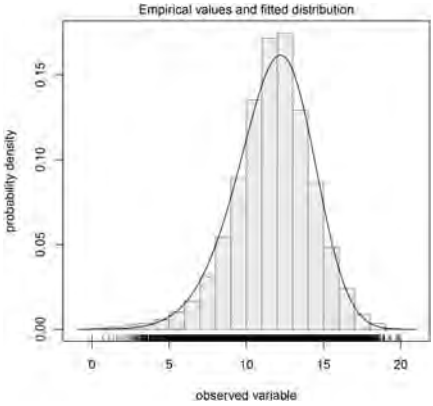

図7　売上高の対数のヒストグラムと統計モデル：ラグ付き

であり，これより推定された確率密度関数（統計モデル）

$$f(\log(\texttt{sales})\,|\,\hat{\xi},\hat{\omega},\hat{\alpha}) := \frac{2}{\hat{\omega}}\phi\left(\frac{\log(\texttt{sales})-\hat{\xi}}{\hat{\omega}}\right)\Phi\left(\hat{\alpha}\,\frac{\log(\texttt{sales})-\hat{\xi}}{\hat{\omega}}\right)$$

を売上高の対数 log（sales）のヒストグラムに重ね書きしたものは図7で与えられる．

以上の可視化の結果から売上高の対数が非対称正規分布にある程度当てはまると考え，この知見を利用して次節では売上高（sales）を従業員数（employees）と資産合計（assets.total）でモデリングする．

V　統計モデリング

この節では売上高を従業員数と資産合計で説明する統計モデリングを行う[12]．その際，Ⅲ節で得られた知見から，売上高の分布の構造から正規線形モデルが当てはまらないことが予測されるけれども，実際にそのことを確認した後，モデルを改良することを繰り返しながら探索的データ解析を実行する．

（2014）の3章を参照されたい．

12）いわゆる生産関数の一種を考えていることに注意しよう．

1．正規線形モデル

統計モデリングのベンチマークとして，通常の正規線形モデル

$$\texttt{sales}_i=\beta_0+\beta_1\texttt{employees}_i+\beta_2\texttt{assets.total}_i+\epsilon_i,\ \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0,\sigma^2),\quad i=1,\ldots,n$$

を最小自乗法によってデータに当てはめる．ここで，$\mathsf{N}(0,\sigma^2)$ は平均 0，分散 σ^2 の正規分布を表す記号であり，“$\overset{\text{i.i.d.}}{\sim}$”は独立に同一の分布に従う（independent and identically distributed: i.i.d.）ことを表す．なお，$n=21801$ であることに注意しよう．

ティー検定表（表２）の結果から，全ての回帰係数は有意となっていることに注意しよう．

表２　ティー検定表：正規線形モデル

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-89742.2976	41630.8981	-2.16	0.0311
employees	138.7339	1.7267	80.35	0.0000
assets.total	0.4478	0.0033	137.09	0.0000

また，標本回帰平面（図８）は

$$\begin{aligned}\hat{\eta}_{\mathsf{NL}}&=\hat{\beta}_0+\hat{\beta}_1\texttt{employees}+\hat{\beta}_2\texttt{assets.total}\\&=-89742.298+138.734\texttt{employees}+0.488\texttt{assets.total}\end{aligned}$$

で与えられる．

さらに，誤差分散の推定値と標準偏差の推定値，決定係数，自由度調整済み決定係数はそれぞれ以下のように与えられる：

誤差分散の推定値：$\hat{\sigma}^2=6006102.374^2$

誤差の標準偏差の推定値：$\hat{\sigma}=6006102.374$

決定係数：$R^2=0.687$

自由度調節済み決定係数：$\bar{R}^2=0.687$

特に決定係数，自由度調整済み決定係数とも約68.7％となっていることに注

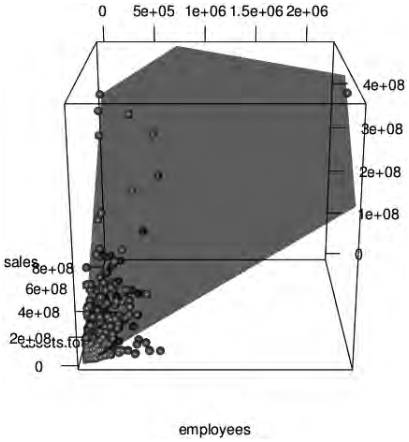

図８　標本回帰平面：正規線形モデル

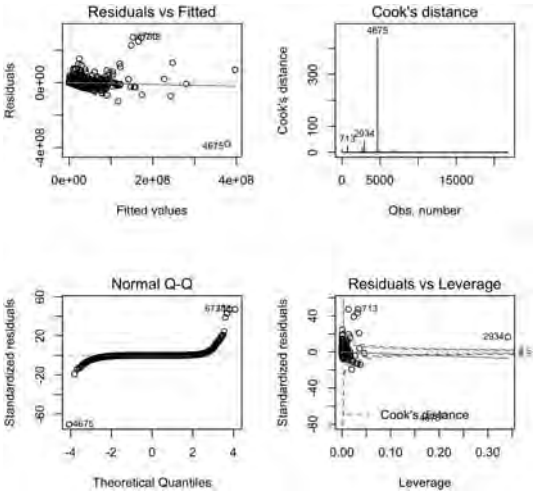

図９　正規線形モデルの当てはめに伴う回帰診断に関する各種のプロット

意しよう．この結果から正規線形モデルのデータへの当てはまりの度合いを単純に判断することには議論が分かれると思われるが，回帰診断に関する各種のプロット（図９）を見ると誤差の正規性に関する根本的な問題が存在す

ることがわかる．すなわち，残差の正規 Q-Q プロット（図 9 の（2, 1）ブロック）を見ると明らかに正規性が成り立たないことがわかる．このことから，正規線形モデルはある程度の説明力はあるものの良いものとはいいがたい．

2. 対数正規線形モデル

正規線形モデルは誤差分布の前提条件である正規性に関して問題があることがわかったので，補正の一つの選択肢として地道（2014）でも扱った対数正規線形モデル：

$$\texttt{sales}_i=\gamma\times\texttt{employees}_i^{\alpha_1}\times\texttt{assets.total}_i^{\alpha_2}\times\epsilon_i,\ \epsilon_i\overset{\text{i.i.d.}}{\sim}\mathsf{LN}(0,\sigma^2),\quad i=1,\ldots,n$$

の当てはめを行う．このモデルは両辺の対数をとることによって，

$$\log\texttt{sales}_i=\alpha_0+\alpha_1\log\texttt{employees}_i+\alpha_2\log\texttt{assets.total}_i+\log\epsilon_i,\quad \log\epsilon_i\overset{\text{i.i.d.}}{\sim}\mathsf{N}(0,\sigma^2),\ i=1,\ldots,n$$

となり，正規線形モデルに帰着することに注意しよう．ここで，$\alpha_0:=\log\gamma$ とおいた．

対数正規線形モデルを当てはめることによって得られるティー検定表（表 3）の結果から，全ての回帰係数は有意となっていることに注意しよう．

表 3　ティー検定表：対数正規線形モデル

	Estimate	Std. Error	t value	Pr(<|t|)
(Intercept)	0.8058	0.0395	20.42	0.0000
log(employees)	0.4673	0.0049	94.42	0.0000
log(assets.total)	0.6465	0.0047	136.88	0.0000

このモデルを当てはめた結果として得られる標本回帰平面（図10）は

$$\begin{aligned}\hat{\eta}_{\mathsf{LNL}}&=\hat{\alpha}_0+\hat{\alpha}_1\log\texttt{employees}+\hat{\alpha}_2\log\texttt{assets.total}\\&=0.806+0.467\log\texttt{employees}+0.646\log\texttt{assets.total}\end{aligned}$$

である．

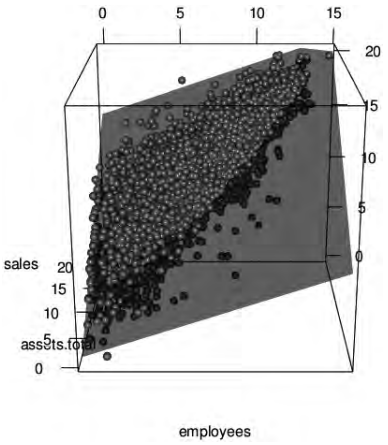

図10 標本回帰平面（対数スケール）：対数正規線形モデルの正規線形表現

誤差分散の推定値と標準偏差の推定値，決定係数，自由度調整済み決定係数はそれぞれ以下のように与えられる：

誤差分散の推定値：$\hat{\sigma}^2 = 1.003^2$

誤差の標準偏差の推定値：$\hat{\sigma} = 1.003$

決定係数：$R^2 = 0.845$

自由度調節済み決定係数：$\bar{R}^2 = 0.845$

この結果から，特に，決定係数と自由度調節済み決定係数が共に84.5％であり，当てはめの程度が飛躍的に伸びていることに注意しよう．ただし，回帰診断に関するプロット（図11）における残差の正規 Q-Q プロットを見ると，裾の部分が正規分布に当てはまっていないことがわかり，特に左裾の部分が顕著である．なお，残差のヒストグラム（図12）からも左裾の部分が通常の正規分布よりも重いことを伺うことができることに注意しよう．

3. 対数非対称正規線形モデル

これまでの考察で対数正規線形モデルも誤差の構造に対して十分な説明ができないことがわかったが，既にⅣ節でも検討したように，売上高

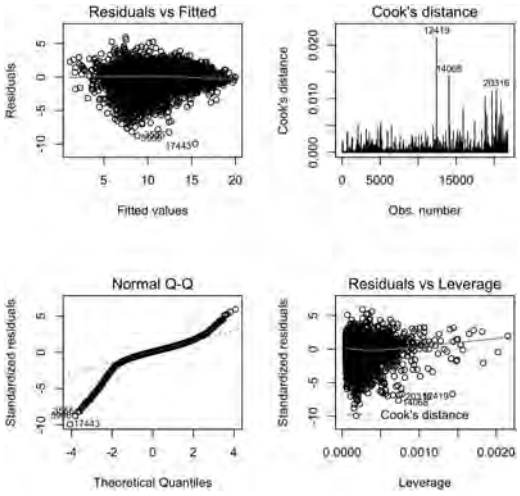

図11　対数正規線形モデルの当てはめに伴う回帰診断に関する各種のプロット

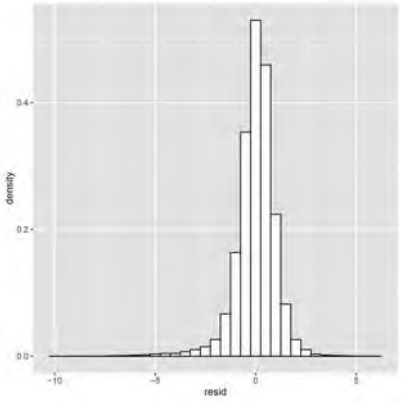

図12　対数正規線形モデルの当てはめに伴う残差のヒストグラム

（sales）は対数非対称正規分布が正規分布や対数正規分布と比較して妥当であるという結果をモデルに反映させることによって以下のようなものが提案される：

$$\texttt{sales}_i = \gamma \times \texttt{employees}_i^{\alpha_1} \times \texttt{assets.total}_i^{\alpha_2} \times \epsilon_i,\ \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{LSN}(0, \omega^2, \alpha),$$

$$i=1,\ldots,n$$

このモデルを対数非対称正規線形モデル（log-skew-normal linear model）と呼ぼう．両辺の対数をとることによって，

$$\log \mathtt{sales}_i=\alpha_0+\alpha_1 \log \mathtt{employees}_i+\alpha_2 \log \mathtt{assets.total}_i+\log \epsilon_i,$$

$$\log \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathsf{SN}(0,\omega^2,\alpha),\ i=1,\ldots,n$$

となり，これを対数非対称正規線形モデルの非対称正規線形表現と呼ぶことにする．

最尤法で推定された推定値によるゼット比（z-ratio）検定表[13] を表 4 に与える．

表 4　ゼット比検定表：対数非対称正規線形モデル

	estimate	std.err	z-ratio	Pr{>|z|}
(Intercept.DP)	1.94	0.04	52.12	0.00
log(employees)	0.36	0.01	69.71	0.00
log(assets.total)	0.69	0.00	150.32	0.00
omega	1.43	0.01	146.20	0.00
alpha	−2.26	0.04	−55.07	0.00

この表に与えられている結果から，全ての回帰係数は有意となっていることに注意しよう．また，このモデルの当てはめによる標本回帰平面（図13）は以下のように与えられる：

$$\begin{aligned}\hat{\eta}_{\mathsf{LSNL}}&=\hat{\alpha}_0+\hat{\alpha}_1 \log \mathtt{employees}+\hat{\alpha}_2 \log \mathtt{assets.total}\\&=1.94+0.36 \log \mathtt{employees}+0.69 \log \mathtt{assets.total}\end{aligned}$$

注意 2　非対称正規線形モデルの当てはまりの程度を決定係数ではかることは理にかなっていないことに注意しよう．つまり，正規線形モデルに対して最小自乗法[14] によって推定された母数によってモデルを当てはめており，正

13）ゼット比（z-ratio）にもとづく検定は，最尤法にもとづいて求められた検定統計量が帰無仮説（母数が 0）のもとで漸近的に標準正規分布に従うことを利用している．なお，記号（文字）z が利用される理由は，統計学の習慣として標準正規分布に従う統計量をこの記号で表すことによる．

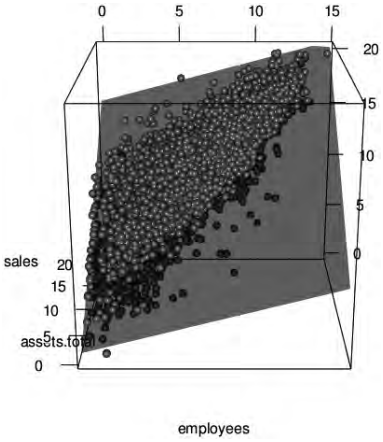

図13　標本回帰平面（対数スケール）：対数非対称正規線形モデルの非対称正規線形表現

射影にもとづく規準である決定係数は，その適合度をはかるために適当であるけれども，非対称正規線形モデルのもとでは尤度（または対数尤度）を最大化する方法（最尤法）で母数の推定を行っており，正射影にはもとづいていないため，決定係数は適合度をはかる規準として適切でない[15]．

（注意終）

以上の注意から，このモデルの当てはまりの程度は決定係数ではなく他の規準で評価する必要があり，次の項で他の場合も含めて比較・検討する．

次に，回帰診断に関するプロット（図14）を見る．このプロットは，通常の正規線形モデルをデータに当てはめた結果を診断するものと若干異なっており，特に残差の分位点と確率点に関するプロット（Q-Q プロットと P-P プロット）がカイ自乗分布に関する分位点と確率点を利用したものが描かれ

14）応答変数を説明変数の張る空間へ正射影することによって誤差平方和を最小にする方法．
15）正規線形モデルによる決定係数は非対称正規線形モデルのものよりも必ず大きくなる．

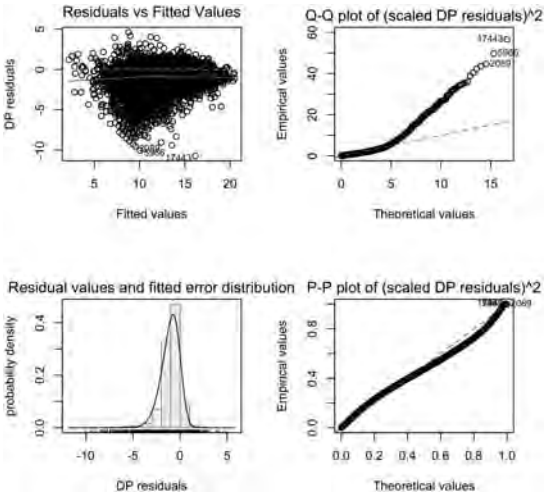

図14 対数非対称正規線形モデルの当てはめに伴う回帰診断に関する各種のプロット

ていることに注意しよう．これは，注意１で述べたように，非対称正規分布 SN$(0, 1, \alpha)$ に従う確率変数の２乗は自由度１のカイ自乗分布に従うという結果を援用して，もし観測が非対称正規線形モデルに従っている場合は，標準化された残差の２乗は近似的に自由度１のカイ自乗分布に従うことがその理論的な根拠となっていることに注意しよう．

図14を見ると，Q-Q プロット以外は妥当な結果となっていることに注意しよう．ただし，Q-Q プロットが直線（理想的な状態）からは離れていることには注意が必要である．

4. モデル比較

注意２でも述べたけれども，対数非対称正規線形モデルの非対称正規線形表現におけるモデルの当てはまりの程度を決定係数ではかることは難しいため，本稿で扱った３種類のモデルによる当てはまりをみるための共通の規準として赤池情報量規準（AIC）を利用する．正規線形モデル（firmfin.lm.

表 5　AIC 表

	df	AIC
firmfin.lm.2013	4.00	742426.07
firmfin.lm.log.2013	4.00	62018.83
firmfin.selm.log.2013	5.00	59534.75

2013），対数正規線形モデル（firmfin.lm.log.2013），対数非対称正規線形モデル（firmfin.selm.log.2013）をデータに対して当てはめたときの，それぞれのモデルに対する自由度（degree of freedom：df）（または母数の個数）と AIC の値を表 5 に与える．

この結果から，対数非対称正規線形モデルを当てはめたときの AIC の値が最小であり，このモデルが他のモデルに比べて推奨される結果が得られた．

VI　おわりに

本稿では，探索的データ解析の視点にたち，20000社を超える全世界の上場企業の財務データを可視化することによって得られた知見にもとづいて，売上高を従業員数と資産合計で説明するための統計モデリングを扱った．考察したモデルのうち，結果として対数非対称正規線形モデルが赤池情報量規準の意味で最も良いものであるという結論を得ることができた．ただし，以下のような問題が存在することには注意が必要である：

- 今回考察したモデル中では対数非対称正規線形モデルが最も良い結果を与えたけれども，Q-Q プロットの観点からは満足のいく結果となっていない．この点を改良するためには非対称ティー分布などを含む非対称分布族（Azzalini and Capitanio（2014）参照.）の当てはめも検討する必要があろう．
- 本稿においてモデルの評価規準は赤池情報量規準を利用したけれども，この規準はモデルが真の分布（データ発生メカニズム）を含まない場合には注意を必要とする．この観点から，竹内情報量規準（Takeuchi's Information Criterion：TIC）などの利用を検討する必要があろう．（小

西，北川（2004）参照.）

- 地道（2014）や Jimichi and Maeda（2014）では業種コード（類別変数）をダミー変数としてモデルに導入することによって回帰モデルを改良することが行われていたけれども，今後，業種コードのみならず国別のコードを考慮した統計モデリングを検討することも必要であろう.
- 今回の考察では単年（2013年）のものであったけれども，時間的な変化を考慮したモデリングを検討する必要があろう.

以上の事項に関しては今後の研究課題としたい.

（筆者は関西学院大学商学部教授）

謝辞

本研究の一部は，文部科学省科学研究費基盤研究 C（一般）（課題番号：16K04022，研究代表者：阪智香），関西学院大学産業研究所プロジェクト「関西経済の構造分析」（研究代表者：豊原法彦）に関する研究費ならびに関西学院大学図書館図書費 B の援助を受けている．ここに感謝の意を述べる.

参考文献

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

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

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

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

［5］ Knuth, D. E.（1984）Literate Programming, *The Computer Journal, British Computer Society,* Vol. 27, No. 2, pp. 97-111.

［6］ 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.

［7］ 地道正行（2010-b）『財務データベースサーバの構築』，関西学院大学リポジトリ，http://kgur.kwansei.ac.jp/dspace/handle/10236/6013,ISBN: 9784990553005.

［8］ 地道正行（2014）『R を利用した財務データの可視化と統計モデリング：探索的データ解析の視点から』，商学論究，第61巻，第 3 号，pp. 241-295，関西学院大学商学研究会.

［9］ Jimichi, M. and S. Maeda（2014）*Visualization and Statistical Modeling of Financial Data*

with R, Poster at The R User Conference 2014. http://user2014.stat.ucla.edu/abstracts/posters/48_Jimichi.pdf
［10］ 小西貞則，北川源四郎（2004）『情報量規準』，朝倉書店.
［11］ Mosteller, F. and J. W. Tukey（1977）*Data Analysis and Regression: A Second Course in Statistics,* Addison-Wesley, Reading Mass.
［12］ 大野忠士，山下智志，椿広計（2011）『与信判断が確率変動するときの倒産企業の信用リスク値分布のモデル化：Skew-normal 分布の応用』，統計数理，第59巻，第 1 号，pp. 3-23.
［13］ Saka, C. and M. Jimichi（2015）*Inequality evidence from accounting data visualisation,* SSRN http://ssrn.com/abstract＝2549400, pp. 1-34.
［14］ 柴田里程（2015）『データ分析とデータサイエンス』，近代科学社.
［15］ 丹後俊郎（2000）『統計モデル入門』，朝倉書店.
［16］ Tukey, J. W.（1977）*Exploratory Data Analysis,* Addison-Wesley Publishing Co.
［17］ Xie, Y.（2015）*Dynamic Documents with R and knitr,* Second edition, CRC Press.

付録A　R パッケージ sn

sn[16] は，非対称正規分布の提唱者であるパドバ大学統計科学科の Adelchi Azzalini 氏によって開発されている R パッケージであり，1 変量と多変量の非対称正規分布を含む非対称分布族[17] を扱うための関数が提供されている．本稿で利用した非対称正規分布に関する関数は，表 6 に与えられているようなものである．

付録B　R スクリプト

```
#####################################################
### パッケージとデータの読み込みとデータ処理
#####################################################
set.seed(12345)
options(width=70)
library(ggplot2)
library(GGally)
library(plyr)
```

16）パッケージは CRAN サイト（https://cran.r-project.org/）からダウンロード可能である．また，開発に関する最新の話題については http://azzalini.stat.unipd.it/SN/index.html#lib-sn から情報が得られる．
17）非対称分布族には，非対称正規分布の他に非対称ティー分布，非対称コーシー分布，非対称楕円分布などがあり，これらの計算に関する関数が sn パッケージに付属されている．

表 6　sn パッケージに付属の非対称正規分布に関する関数

関数	機能
dsn	確率密度関数の計算
psn	累積分布関数の計算
qsn	分位点の計算
rsn	乱数の生成
selm	誤差が非対称楕円分布に従う線形モデルの最尤法による当てはめ
summary	selm の結果のオブジェクトを要約（総称関数）
plot.selm	selm の結果のオブジェクトを可視化

```
library(dplyr)
library(e1071)
library(car)
library(xtable)
library(sn)
library(rgl, pos=21)
library(nlme, pos=22)
library(mgcv, pos=22)
firmfin.frame<-readRDS("firmfin.frame.rds")
firmfin2013<-select(filter(firmfin.frame,year==2013,sales>0,employees>0,assets.total>0),
                    firmID,country,SIC.code,sales,employees,assets.total)

##################################################
### 本稿で利用するデータの表示
##################################################
options(width=200)
firmfin2013[seq(10),]
options(width=70)

##################################################
### 対散布図の描画
##################################################
ggpairs(firmfin2013[c("sales","employees","assets.total")],
        upper = list(continuous = wrap("points", alpha = 0.3),
                     combo = wrap("dot", alpha = 0.4)),
        lower = list(continuous = wrap("points", alpha = 0.3),
                     combo = wrap("dot", alpha = 0.4)))

##################################################
### 対散布図の描画：対数スケール
##################################################
  ggpairs(log(firmfin2013[c("sales","employees","assets.total")]),
        upper = list(continuous = wrap("points", alpha = 0.3),
                     combo = wrap("dot", alpha = 0.4)),
        lower = list(continuous = wrap("points", alpha = 0.3),
                     combo = wrap("dot", alpha = 0.4)))

##################################################
### 売上高のヒストグラムの描画
##################################################
ggplot(firmfin2013,aes(x=sales))+
  geom_histogram(aes(y=..density..),fill="white",color="black")

##################################################
### 売上高のヒストグラムの描画：対数スケール
```

```
#####################################################
ggplot(firmfin2013,aes(x=log(sales)))+ xlim(0,21) +
  geom_histogram(aes(y=..density..),binwidth=1,fill="white",color="black")

#####################################################
### 対数変換後の売上高の正規 Q-Q プロット
#####################################################
ggplot(data=as.data.frame(qqnorm(log(firmfin2013$sales), plot=F)), mapping=aes(x=x, y=y)) +
  geom_point() +
  geom_abline(intercept=mean(log(firmfin2013$sales)),slope=sd(log(firmfin2013$sales))) +
  labs(x="theoretical",y="sample")

#####################################################
### 非対称正規分布の確率密度関数をプロットする R 関数の定義
#####################################################
ggplot.sn<-function(x=seq(-10,10,0.01),xi=0,omega=1,alpha=0)
{
  require(sn,ggplot2)
  t<-ggplot(data.frame(x,y=dsn(x,xi=xi,omega=omega,alpha=alpha)),aes(x,y))+geom_line()
  t+labs(title=paste("(xi,omega,alpha)=(",xi,",",omega,",",alpha,")"),y="Deinsity")
}

#####################################################
### 非対称正規分布の確率密度関数のプロット: alpha=3
#####################################################
ggplot.sn(x=seq(-5,5,0.01),alpha=3)

#####################################################
### 非対称正規分布の確率密度関数のプロット: alpha=-3
#####################################################
ggplot.sn(x=seq(-5,5,0.01),alpha=-3)

#####################################################
### 非対称正規分布の（母）歪度をプロットする R 関数の定義
#####################################################
ggplot.sn.skewness<-function(alpha=seq(-1,1,0.01))
{
  delta<-alpha/sqrt(1+alpha^2)
  gamma1<-sqrt(2)*(4-pi)*delta^3/(pi-2*delta^2)^(3/2)
  ggplot(data.frame(alpha=alpha,gamma1=gamma1),aes(alpha,gamma1))+geom_line()
}

#####################################################
### 非対称正規分布の（母）歪度のプロット
#####################################################
ggplot.sn.skewness()

#####################################################
### 対数変換後の売上高への非対称正規分布の最尤法による当てはめ
#####################################################
selm.sales2013<-selm(log(sales)~1,data=firmfin2013)

#####################################################
### 対数変換後の売上高のヒストグラムに推定された非対称正規分布
### の確率密度関数の重ねて描いたもの
#####################################################
plot(selm.sales2013,which=2)

#####################################################
### 正規線形モデルの当てはめ
#####################################################
firmfin.lm.2013<-lm(sales~employees+assets.total,firmfin2013)

#####################################################
```

```
### 正規線形モデルを当てはめたときのティー検定表の出力
#####################################################
print(xtable(firmfin.lm.2013,
floating=FALSE,caption=c("ティー検定表：正規線形モデル"),label="table.t.normal.linear.model"),
caption.placement = "top",table.placement="H")

#####################################################
### 3次元散布図と標本回帰平面の描画：正規線形モデル
#####################################################
library(rgl)
coef.lm.2013<-coef(firmfin.lm.2013)
plot3d(firmfin2013[c("employees","assets.total","sales")],type = "s", col = "red", size = 1)
planes3d(coef.lm.2013[2],coef.lm.2013[3],-1,coef.lm.2013[1],alpha=0.5)

#####################################################
### 回帰診断に関するプロット：正規線形モデル
#####################################################
par(mfcol=c(2,2))
plot(firmfin.lm.2013,which=c(1,2,4,5))
par(mfcol=c(1,1))

#####################################################
### 対数正規線形モデルの当てはめ
#####################################################
firmfin.lm.log.2013<-lm(log(sales)~log(employees)+log(assets.total),firmfin2013)

#####################################################
### 対数正規線形モデルを当てはめたときのティー検定表の出力
#####################################################
print(xtable(firmfin.lm.log.2013,
floating=FALSE,
caption=c("ティー検定表：対数正規線形モデル"),label="table.t.log.normal.linear.model"),
caption.placement = "top",table.placement="H")

#####################################################
### 3次元散布図と標本回帰平面の描画：対数正規線形モデル
#####################################################
coef.lm.log.2013<-coef(firmfin.lm.log.2013)
plot3d(log(firmfin2013[c("employees","assets.total","sales")]),type = "s", col = "red", size = 1)
planes3d(coef.lm.log.2013[2],coef.lm.log.2013[3],-1,coef.lm.log.2013[1],alpha=0.5)

#####################################################
### 回帰診断に関するプロット：対数正規線形モデル
#####################################################
par(mfcol=c(2,2))
plot(firmfin.lm.log.2013,which=c(1,2,4,5))
par(mfcol=c(1,1))

#####################################################
### 残差のヒストグラムの描画：対数正規線形モデル
#####################################################
ggplot(data.frame(resid=resid(firmfin.lm.log.2013)),aes(x=resid))+
  geom_histogram(aes(y=..density..),binwidth=0.5,fill="white",color="black")

#####################################################
### 対数非対称正規線形モデルの当てはめ
#####################################################
firmfin.selm.log.2013<-selm(log(sales)~log(employees)+log(assets.total),
   family="SN",data=firmfin2013)

#####################################################
### 対数非対称正規線形モデルを当てはめたときのゼット検定表の出力
#####################################################
print(xtable(summary(firmfin.selm.log.2013,"DP") @param.table,
```

```
floating=FALSE,
caption=c("ゼット比検定表:対数非対称正規線形モデル"),
label="table.z.logskewnormal.linear.model"),
caption.placement = "top",table.placement="H")

#####################################################
### 3次元散布図と標本回帰平面の描画: 対数非対称正規線形モデル
#####################################################
coef.selm.log.2013<-coef(firmfin.selm.log.2013)
plot3d(log(firmfin2013[c("employees","assets.total","sales")]),type = "s", col = "red", size = 1)
planes3d(coef.selm.log.2013[2],coef.selm.log.2013[3],-1,coef.selm.log.2013[1],alpha=0.5)

#####################################################
### 回帰診断に関するプロット: 対数非対称正規線形モデル
#####################################################
par(mfcol=c(2,2))
plot(firmfin.selm.log.2013,param.type="DP")
par(mfcol=c(1,1))

#####################################################
### AIC 表の出力:
### 正規線形モデル, 対数正規線形モデル, 対数非対称正規線形モデル
#####################################################
print(xtable(AIC(firmfin.lm.2013,firmfin.lm.log.2013,firmfin.selm.log.2013),
floating=FALSE,caption=c("AIC 表"),label="table.AIC"),
caption.placement = "top",table.placement="H")
```