R を利用した非対称分布族にもとづく財務データの統計モデリング*

Statistical Modeling of Financial Data Based on Family of Skew Distributions with R

地 道 正 行

We treat visualization and statistical modeling for financial data (*e.g.*, sales, assets) of many firms which are world-wide and listed-delisted. They are based on exploratory data analysis, and are carried out with the data analysis environment R. The result that the log-skew-t linear model is very useful for explaining sales by employees and assets is obtained from comparing the Akaike's information criterions of some log-linear models which the error terms are independent and distributed to a family of log-skew distributions.

Masayuki Jimichi

JEL：C13, C21, C44, C46

キーワード：探索的データ解析、統計モデリング、非対称分布族、対数線形モデル、情報量規準

Keywords：Exploratory Data Analysis, Statistical Modeling, Family of Skew Distributions, Log-linear Model, Information Criterion

1 はじめに

本稿では, Saka and Jimichi (2017) で扱った全世界の上場企業のデータを

* 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）による研究費，平成 29 年度学際大規模情報基盤共同利用・共同研究拠点公募型共同研究（課題番号: jh171002-NWJ, 研究代表者: 地道 正行）による研究環境，ならびに関西学院大学個人研究費及び図書館図書費 B の援助を受けている．ここに感謝の意を述べる．

利用し探索的データ解析[1] (Exploratory Data Analysis: EDA) の視点に立って統計モデリングを行う．その際, 地道 (2017) で考察された対数非対称正規線形モデルを当てはめることに伴う以下のような問題と対処法を検討する[2]:

- 対数非対称正規線形モデルは通常の対数正規線形モデルよりも当てはまりは良いと思われたけれども, 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かった．この点を改良するために (対数) 非対称ティー分布 ((log-)skew-t distribution) などを含む (対数) 非対称分布族 (family of (log-)skew distributions) の当てはめを行う[3]．(Azzalini and Capitanio (2014) も参照のこと.)

- モデルが真の分布 (データ発生メカニズム) を含まない場合を考慮して, 一部の場合については, 竹内情報量規準 (Takeuchi's Information Criterion: TIC) (竹内 (1976)) でもモデルの評価を行う．

本稿の構成は以下のようなものである．2 節では, 地道 (2017) の結果にもとづいて, データの可視化と非対称正規分布をはじめとする非対称分布族を売上高の対数に当てはめる．その際, 非対称正規分布を当てはめた場合の問題を再確認し, 非対称ティー分布の当てはめも検討する．3 節では, 2 節で得られた知見を利用して, 売上高を従業員数と資産合計で説明するためにモデルを考える．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ため, 売上高の対数をとったものの統計モデリング, すなわち, 対数線形モデルを考え, 誤差分布に関する仮定を検証することを繰り返しながら探索的デー

1) 探索的データ解析については Tukey (1977) を参照のこと．

2) 地道 (2017) で扱ったデータは, 決算月数が 12 ヶ月でないものも含まれていたけれども, 本稿では 12 ヶ月のもののみ扱っていることに注意しよう．このことから, 地道 (2017) と同じ推定法でも若干の数値的な違いが生じているが, 結果への本質的な問題はないことに注意しよう．

3) 本稿では, 非対称正規分布とそれに関連する分布族 (たとえば, 非対称ティー分布や非対称コーシー分布を含むもの) を非対称分布族 (family of skew distributions) と呼ぶことにする．また, 売上高の対数が非対称分布族に属するどのような分布に従っているかを考察しており, 対数をとる前の分布を考えるときは対数非対称分布族 (family of log-skew distributions) を考えることになることに注意しよう．

タ解析を実行する．その結果にもとづき, 4 節では, 売上高の対数に 3 種類の分布 (正規分布, 数非対称正規分布, 非対称ティー分布) を当てはめた結果を赤池情報量規準と, 一部のものについては竹内情報量規準を使って比較・検討し, さらに, 売上高の対数に 3 種類の対数線形モデル (対数正規線形モデル, 対数非対称正規線形モデル, 対数非対称ティー線形モデル) に関する当てはまりをみるために赤池情報量規準を利用した考察を行う．最後に, 5 節で本稿のまとめと今後の課題を与える．

なお, 付録 A には, 本稿で扱うデータの説明を与えており, 付録 B と付録 C には, それぞれ, 非対称正規分布・対数非対称正規分布と非対称ティー分布・対数非対称ティー分布の簡単な説明を与えている．また, 付録 D には竹内情報量規準について述べている．

本稿ではデータ解析環境 R[4] を用いており, データの可視化には ggplot2, GGally, rgl パッケージ, データ操作には dplyr パッケージ, さらに非対称分布族を R で扱うために開発されたパッケージ sn を利用している．付録 E には, sn パッケージに収録されている関数のうち, 本稿で利用したものの簡単な説明を与えている．最後に, 付録 F には本稿で使用した R スクリプトを与えている[5]．

2 データ可視化と売上高の対数への非対称分布族の当てはめ

本節では, 地道 (2017) の結果にもとづいて, データの可視化と売上高の対数に非対称正規分布をはじめとする非対称分布族を当てはめる．その際, 非対称正規分布を当てはめた場合の問題を再確認し, 非対称ティー分布の当てはめも検討する．

4) R version 3.3.3 (2017-03-06)

5) 本稿は全編を通じて地道 (2017) と同様に再現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` 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．

2.1 対数非対称正規分布

地道 (2017) で示されているように本稿で扱っている粗データ (raw data) の対散布図は, 原点付近に集中して分布しており, 「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. よって, ここでは, 図 1 に各変量の対数をとったものの対散布図を与える.

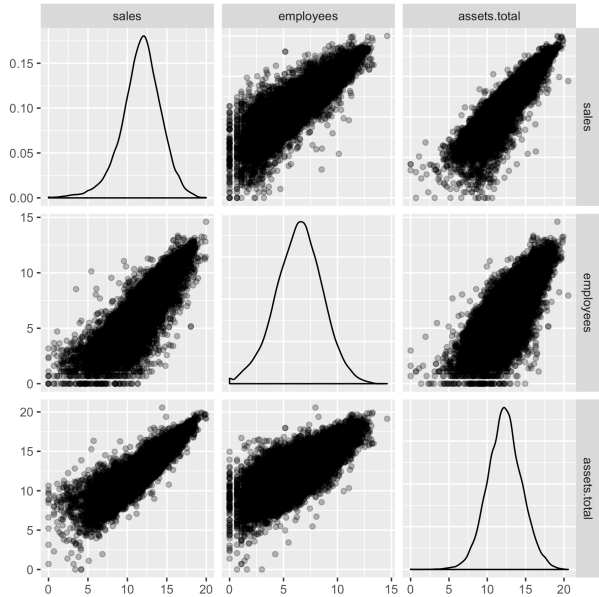

図 1: 財務データの対散布図: 対数スケール

図 1 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「正規分布」には従うとはいいがたいということである. 地道 (2017) では売上高 `sales` を応答変数とする回帰モデルを構築するために, 売上高の対数をとったもののヒストグラムや正規 Q-Q プロットを描くことによる可視化した結果から分布構造が検討されている. その結果として, 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, 若干左側の裾が右の裾に比べて重く, 売上高の対数は正規分布よりも左に歪んだ分布で

あるとの結論を得ている. ただし, 裾の部分 (特に左裾) 以外は正規分布で近似できることも指摘されており, 正規分布の片方の裾が少し「重い」分布の候補となる分布として対数非対称正規分布 (log skew-normal distribution) を採用している.

地道 (2017) でも検討されているけれども, 対数非対称正規分布 $\mathsf{LSN}(\xi, \omega^2, \alpha)$ に売上高 `sales` が従うものとし, その対数 $\log(\texttt{sales})$ に非対称正規分布 $\mathsf{SN}(\xi, \omega^2, \alpha)$ を当てはめることを再度確認する. まず, 母数 (ξ, ω, α) を最尤法によって推定した結果は,

$$(\widehat{\xi}, \widehat{\omega}, \widehat{\alpha}) = (14.09, 3.47, -1.66)$$

であり, これより推定された確率密度関数 (統計モデル)

$$f_{\mathsf{SN}}(\log(\texttt{sales}) \mid \widehat{\xi}, \widehat{\omega}, \widehat{\alpha}) := \frac{2}{\widehat{\omega}} \phi\left(\frac{\log(\texttt{sales}) - \widehat{\xi}}{\widehat{\omega}}\right) \Phi\left(\widehat{\alpha} \frac{\log(\texttt{sales}) - \widehat{\xi}}{\widehat{\omega}}\right)$$

を売上高の対数 $\log(\texttt{sales})$ のヒストグラムに重ね書きしたものと標準化された残差の 2 乗に関する Q-Q プロットを, それぞれ, 図 2 と図 3 に与える:

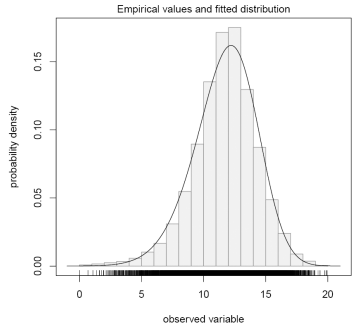

図 2: 売上高の対数のヒストグラムと非対称正規分布にもとづく統計モデル

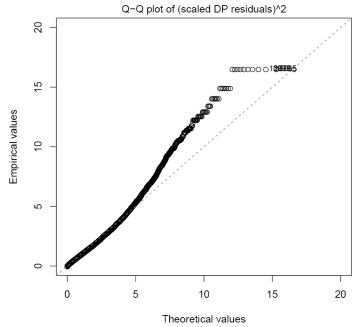

図 3: 売上高の対数に非対称正規分布を当てはめたときの標準化残差の 2 乗の Q-Q プロット

地道 (2017) では, 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統計モデリングを行っているけれども, これらの可

視化の結果を詳細に見ると, 分布の中心部あたりでモデルの当てはまりに問題のあることが見て取れる. 特に, 図 3 より, 分布の中心部から裾にかけて当てはまりの悪さが見られる.

この問題に対して, 売上高の対数に対して非対称正規分布と異なった非対称分布族に属するものを当てはめることを考える. なお, 本稿では当てはめるモデルが異なっていても母数 θ の最尤推定値は $\widehat{\theta}$ のようにハット "$\widehat{\ }$" をつけて統一的に表すことにする.

2.2 対数非対称ティー分布

非対称正規分布以外にも対称な分布の確率密度関数にある種の累積分布関数を乗じることによって得られる非対称分布族が Azzalini and Capitanio (2014) によって提案されており, 非対称ティー分布 (skew-t distribution) が代表的なものの一つである. 非対称ティー分布について簡単な説明を付録 C に与えるが, その導出や確率密度関数・モーメント等についての詳細は Azzalini and Capitanio (2014) を参照されたい.

前節で述べた売上高の対数に対して非対称正規分布を当てはめた結果として分布の中心部に対して当てはまりの悪さが見られる問題に対して, 非対称ティー分布 $\mathsf{ST}(\xi, \omega^2, \alpha, \nu)$ を当てはめることを試みる.

まず, 母数 $(\xi, \omega, \alpha, \nu)$ を最尤法によって推定した結果は,

$$(\widehat{\xi}, \widehat{\omega}, \widehat{\alpha}, \widehat{\nu}) = (13.49, 2.76, -1.08, 9.93)$$

であり, これより推定された確率密度関数 (統計モデル) は,

$$f_{\mathsf{ST}}(\log(\texttt{sales}) \mid \widehat{\xi}, \widehat{\omega}, \widehat{\alpha}, \widehat{\nu})$$
$$= \frac{2}{\widehat{\omega}} f_{\mathrm{t}}\left(\frac{\log(\texttt{sales}) - \widehat{\xi}}{\widehat{\omega}} \mid \widehat{\nu}\right) F_{\mathrm{t}}\left(\widehat{\alpha}\frac{\log(\texttt{sales}) - \widehat{\xi}}{\widehat{\omega}}\sqrt{\frac{\widehat{\nu}+1}{\left(\frac{\log(\texttt{sales})-\widehat{\xi}}{\widehat{\omega}}\right)^2 + \widehat{\nu}}} \mid \widehat{\nu}+1\right)$$

で与えられる. ここで,

$$f_{\mathrm{t}}(z \mid \nu) := \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}}\left(1 + \frac{z^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad F_{\mathrm{t}}(z \mid \nu) = \int_{-\infty}^{z} f_{\mathrm{t}}(x \mid \nu)\mathrm{d}x$$

は, それぞれ, 自由度 ν のティー分布の確率密度関数と累積分布関数である.

統計モデルを売上高の対数 log(sales) のヒストグラムに重ね書きしたものと標準化された残差の 2 乗に関する Q-Q プロットを, それぞれ, 図 4 と図 5 に与える[6]:

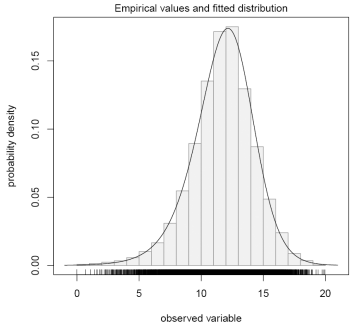

図 4: 売上高の対数のヒストグラムと非対称ティー分布にもとづく統計モデル

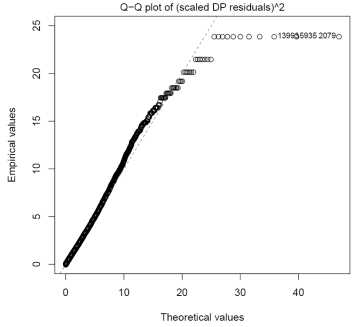

図 5: 売上高の対数に非対称ティー分布を当てはめたときの標準化残差の 2 乗の Q-Q プロット

図 5 において裾の部分には幾つかの当てはまりが悪いデータが存在するものの, データの大部分はモデルに当てはまっていると見ることができることに注意しよう. 特に, 対数非対称正規分布の場合の Q-Q プロット (図 3) と比較して対数非対称ティー分布の方がより当てはまりは良いように思われる.

次節では本節で得られた可視化の結果を踏まえて売上高を従業員数と資産合計でモデリングすることを考える.

3 売上高の対数線形モデリング

この節では売上高 sales を従業員数 employees と資産合計 assets.total で説明するためにモデル[7]:

$$\texttt{sales} = \gamma \times \texttt{employees}^{\alpha_1} \times \texttt{assets.total}^{\alpha_2} \times \epsilon$$

6) ここで描かれている Q-Q プロットはエフ分布にもとづくものであることに注意しよう. (付録 C も参照のこと.)

7) このモデルは, いわゆる, コブ・ダグラス型生産関数を考えていることに注意しよう.

を考える. その際, 2 節で得られた知見から, 売上高 `sales` 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ため, 売上高の対数 $\log(\texttt{sales})$ の統計モデリング, すなわち,

$$\log(\texttt{sales}) = \alpha_0 + \alpha_1 \log(\texttt{employees}) + \alpha_2 \log(\texttt{assets.total}) + \log(\epsilon)$$

を考え, 誤差分布に関する仮定を検証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. なお, 本稿ではこのモデルを一般的に対数線形モデル (log-linear model) と呼ぶことにする. $\alpha_0 := \log\gamma$ であることに注意しよう.

3.1 対数正規線形モデル

本稿で扱う対数線形モデルの比較のためのベンチマークとして対数正規線形モデル (log-normal-linear model):

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

の当てはめを行う[8]. ここで, "$\overset{\text{i.i.d.}}{\sim}$" は「独立に同一の分布に従う」(**i**ndependent and **i**dentically **d**istributed) ことを表すことに注意しよう. このモデルは両辺の対数をとることによって,

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

となり, 正規線形モデルに帰着することに注意しよう.

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 1) 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 1: ティー検定表: 対数正規線形モデル

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.8102	0.0395	20.50	0.0000
log(employees)	0.4670	0.0050	94.32	0.0000
log(assets.total)	0.6465	0.0047	136.79	0.0000

8) 財務データへの対数正規線形モデルの当てはめについては, 地道 (2014), 地道 (2017) も参照されたい.

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 6) は

$$\begin{aligned}\widehat{\eta}_{\mathsf{LNL}} &= \widehat{\alpha}_0 + \widehat{\alpha}_1 \log(\texttt{employees}) + \widehat{\alpha}_2 \log(\texttt{assets.total}) \\ &= 0.81 + 0.467 \log(\texttt{employees}) + 0.646 \log(\texttt{assets.total})\end{aligned}$$

である.

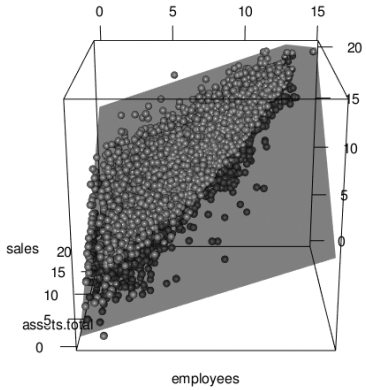

図 6: 標本回帰平面 (対数スケール): 対数正規線形モデルの正規線形表現

誤差分散の推定値, 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: $\widehat{\sigma}^2 = 1.001^2$

決定係数: $R^2 = 0.845$

自由度調節済み決定係数: $\overline{R}^2 = 0.845$

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, モデルはデータにある程度当てはまっていることがわかるけれども, 回帰診断に関するプロット (図 7) 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. (地道 (2017) も参照のこと.)

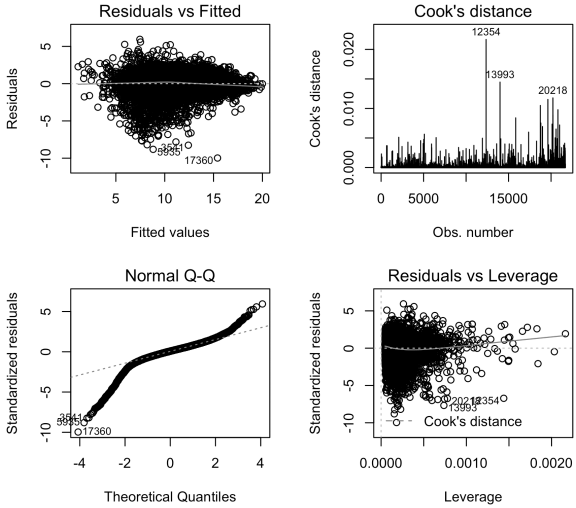

図 7: 対数正規線形モデルの当てはめに伴う回帰診断に関する各種のプロット

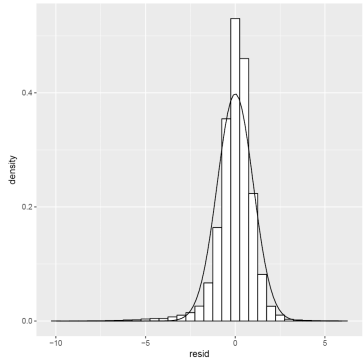

図 8: 対数正規線形モデルの当てはめに伴う残差のヒストグラムと正規分布にもとづく統計モデル

なお, 残差のヒストグラムに統計モデル (正規分布 $\mathsf{N}(0, \widehat{\sigma}^2)$ の確率密度関数) を重ねて描いたもの (図 8) 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.

3.2 対数非対称正規線形モデル

前小節でベンチマークとして与えた対数正規線形モデルは, 誤差の構造に対して十分な説明ができないことから, 2 節で指摘したように売上高 `sales` の対数は非対称分布族が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を考える. そこで, 誤差に対数非対称正規分布を仮定した以下のものを考える:

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

ここでは, このモデルを対数非対称正規線形モデル (log-skew-normal linear model) と呼ぶこととし, 両辺の対数をとることによって得られる

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

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) 検定表を表 2 に与える.

表 2: 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)	1.95	0.04	52.12	0.00
log(employees)	0.36	0.01	69.75	0.00
log(assets.total)	0.69	0.00	150.09	0.00
omega	1.42	0.01	145.84	0.00
alpha	-2.26	0.04	-54.95	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 9) は以下のように与えられる:

$$\begin{aligned} \widehat{\eta}_{\mathsf{LSNL}} &= \widehat{\alpha}_0 + \widehat{\alpha}_1 \log(\texttt{employees}) + \widehat{\alpha}_2 \log(\texttt{assets.total}) \\ &= 1.948 + 0.364 \log(\texttt{employees}) + 0.689 \log(\texttt{assets.total}) \end{aligned}$$

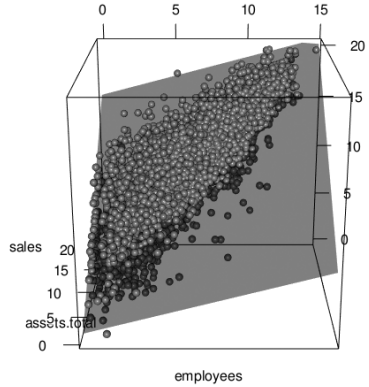

図 9: 標本回帰平面 (対数スケール): 対数非対称正規線形モデルの非対称正規線形表現

回帰診断に関するプロット (図 10) から Q-Q プロットが直線 (理想的な状態) から乖離していることが問題である. すなわち, モデルがデータ発生機構である分布の構造をとらえ切れていないと考えられる.

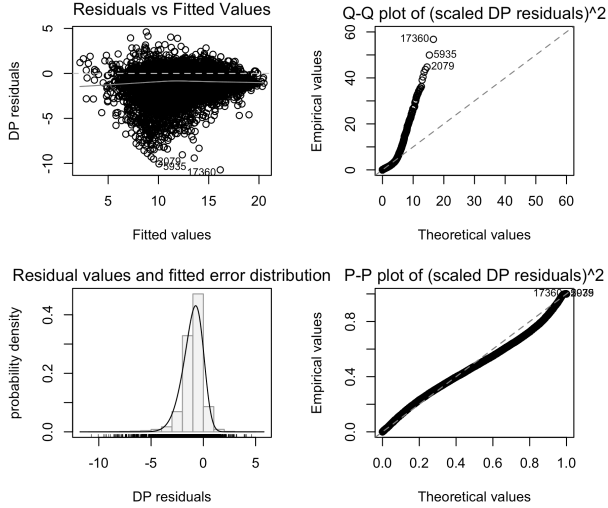

図 10: 対数非対称正規線形モデルの当てはめに伴う回帰診断に関する各種のプロット

3.3 対数非対称ティー線形モデル

地道 (2017) でも指摘されているが, 前小節で与えた対数非対称正規線形モデルは, 誤差の構造に対して説明ができないことから, 2.2 小節の結果を考慮して, 売上高 `sales` が対数非対称ティー分布に従うと仮定して, 誤差分布に対数非対称ティー分布を仮定した以下のものを考える:

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

ここではこのモデルを対数非対称ティー線形モデル (log-skew-t linear model) と呼ぶこととし, 両辺の対数をとることによって得られる

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

を対数非対称ティー線形モデルの非対称ティー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) 検定表を表 3 に与える.

表 3: ゼット比検定表: 対数非対称ティー線形モデル

	estimate	std.err	z-ratio	Pr{>|z|}
(Intercept.DP)	1.67	0.03	49.54	0.00
log(employees)	0.35	0.00	77.23	0.00
log(assets.total)	0.68	0.00	168.62	0.00
omega	0.73	0.01	67.69	0.00
alpha	-0.99	0.04	-23.47	0.00
nu	3.17	0.08	41.60	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 11) は以下のように与えられる:

$$\begin{aligned}\widehat{\eta}_{\mathsf{LSTL}} &= \widehat{\alpha}_0 + \widehat{\alpha}_1 \log(\texttt{employees}) + \widehat{\alpha}_2 \log(\texttt{assets.total}) \\ &= 1.674 + 0.352 \log(\texttt{employees}) + 0.682 \log(\texttt{assets.total})\end{aligned}$$

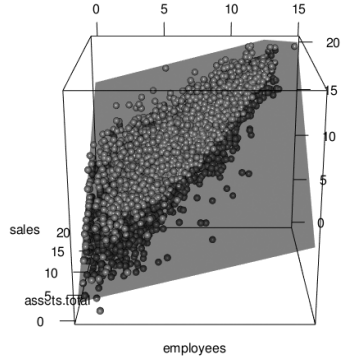

図 11: 標本回帰平面 (対数スケール): 対数非対称ティー線形モデルの非対称ティー線形表現

回帰診断に関するプロット (図 12) から Q-Q プロットが直線 (理想的な状態) から裾の部分で乖離している問題を持つけれども, 対数非対称正規線形モデルを当てはめた場合 (図 10) よりも乖離の程度は改善されているように思われる[9]. 次節では, これらの結果を情報量規準を用いて数値的に比較することによってモデルのデータへの当てはまりに関する評価を行う.

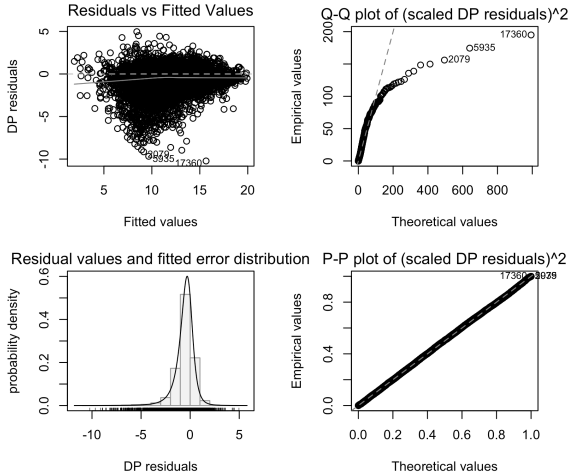

図 12: 対数非対称ティー線形モデルの当てはめに伴う回帰診断に関する各種のプロット

9) 特に, P-P プロットは理想的な当てはまりを表す結果となっていることに注意しよう.

4 情報量規準にもとづくモデル評価

ここでは, まず売上高の対数に本稿で扱った 3 種類の分布 (正規分布, 非対称正規分布, 非対称ティー分布) を当てはめた結果を赤池情報量規準と一部のものについては竹内情報量規準を使って比較・検討する. 次に, 売上高に対して 3 種類の対数線形モデル (対数正規線形モデル, 対数非対称正規線形モデル, 対数非対称ティー線形モデル) を当てはめた結果を評価するために赤池情報量規準を利用する.

4.1 売上高の分布に関するモデル評価

本稿で扱った 3 種類の分布に関する当てはまりを見るために, まず赤池情報量規準を利用する. 売上高の対数に, 正規分布 (lm.log.sales2013), 非対称正規分布 (selm.log.sales2013), 非対称ティー分布 (selm.ST.log.sales2013) を当てはめたときの, それぞれのモデルに対する母数の個数[10] と AIC の値を表 4 に与える.

表 4: AIC 表: 売上高の分布に関する比較

	df	AIC
lm.log.sales2013	2	102074.91
selm.log.sales2013	3	101407.59
selm.ST.log.sales2013	4	101164.85

表 4 から, 最小の AIC 値を与えるのは非対称ティー分布 (selm.ST.log.sales2013) を当てはめた場合であることがわかり, この結果は 2 節で考察された可視化の結果とも符合することに注意しよう.

注意 1 (竹内情報量規準によるモデル評価) 竹内 (1976) は, 想定されたモデルが必ずしもデータを発生させる真の分布を含んでいない場合を想定し, 赤池情報量規準の改良を試みた. 現在では竹内情報量規準と呼ばれ, TIC と表され

10) R の結果のオブジェクトでは自由度 (degree of freedom: df) を表す `df` と出力されることに注意しよう.

る. (付録 D も参照せよ.) ここでは, 計算が比較的簡単な正規分布と非対称正規分布の場合に対する TIC 値を表 5 に与える.

表 5: TIC 表: 売上高の分布に関する比較

	TIC
lm.log.sales2013	102076.001
selm.log.sales2013	101408.141

表 5 の結果と表 4 に共通するモデルの結果を比較すると, 若干の差異は認められるものの, ほぼそれらの値は等しいことから, それぞれのデータがモデルにある程度当てはまっている場合と考えることができよう[11].

4.2 売上高の対数線形モデリングに関するモデル評価

本稿で扱った 3 種類の対数線形モデル, すなわち, 対数正規線形モデル (lm.log.firmfin2013), 対数非対称正規線形モデル (selm.log.firmfin2013), 対数非対称ティー線形モデル (selm.ST.log.firmfin2013) に関する当てはまりをみるために赤池情報量規準を利用する.

表 6: AIC 表: 売上高の対数線形モデルに関する比較

	df	AIC
lm.log.firmfin2013	4	61607.95
selm.log.firmfin2013	5	59133.98
selm.ST.log.firmfin2013	6	55026.78

この結果から, 対数非対称ティー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.

11) ただし, この結果は竹内情報量規準を実際に求めることによってはじめてわかることである. 竹内 (1976) の 6 節も参照せよ.

5 おわりに

本稿では, 探索的データ解析の視点にたち全世界の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, 売上高の対数をとったものの分布とそれ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. 売上高の対数に関しては考察した非対称分布族の中で非対称ティー分布が当てはまるという結果を赤池情報量規準の値を比較することによって得ることができた. なお, 一部のものについては竹内情報量規準の値も求め, 大きな差異がないことを確認できた. このことは, データに対してこれらのモデルが当てはまっていることを肯定する結果であることに注意しよう. また, 対数線形モデルの当てはめに関しては, 今回考察したものの中では対数非対称ティー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果も得ることができた. ただし, これらの結果に関しても, 再び Q-Q プロットにおいて分布の裾の部分でその当てはまりに問題が存在することを確認したけれども, このことについては今後の研究課題としたい.

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] Healy, M. J. R. (1968) Multivariate normal plotting, *Applied Statistics.*, Vol. 17, pp. 157–161.

[5] 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.

[6] 地道 正行 (2017) 『R による対数非対称正規線形モデルによる財務データの統計モデリング』, 商学論究, 第 64 巻, 第 5 号, pp. 159–185, 関西学院大学商学研究会.

[7] 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

[8] 小西 貞則, 北川 源四郎 (2004) 『情報量規準』, 朝倉書店.

[9] Mosteller, F. and J. W. Tukey (1977) *Data Analysis and Regression: A Second Course in Statistics*, Addison-Wesley, Reading Mass.

[10] Saka, C. and M. Jimichi (2017) Evidence of inequality from accounting data visualisation, *Taiwan Accounting Review*, in printing.

[11] 竹内 啓 (1976) 『情報統計量の分布とモデルの適切さの規準』, 数理科学, No. 153, pp. 12–18.

[12] Tukey, J. W. (1977) *Exploratory Data Analysis*, Addison-Wesley Publishing Co.

付録

A データ

本稿で利用するデータは, Bureau van Dijk (BvD) 社[12] から提供される全世界の上場・上場廃止企業の財務データが収録されたデータベース osirisi から抽出されたものを利用する. (地道 (2017), Saka and Jimichi (2017) も参照のこと.)

実際のデータは以下のようなものである:

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

表 7: 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 21689 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

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

`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)

B 非対称正規分布と対数非対称正規分布

B.1 非対称正規分布

非対称正規分布の確率密度関数は,

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

で与えられる. ここで, $x \in \mathbb{R} := (-\infty, \infty)$ であり,

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

は未知母数である. 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. 図 13 に $(\xi, \omega, \alpha) = (0, 1, 3), (0, 1, -3)$ のときの非対称正規分布の確率密度関数のプロットを与える.

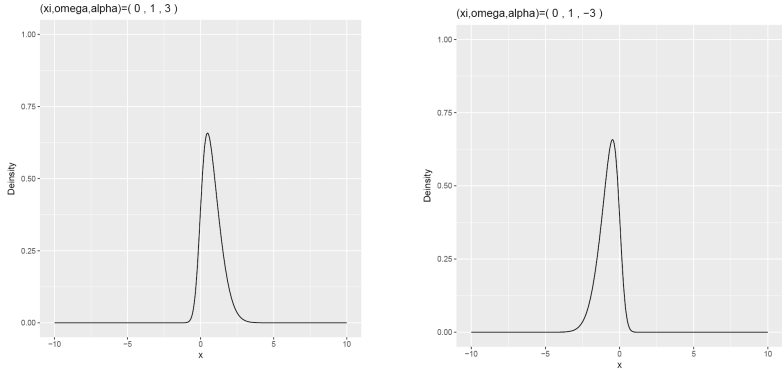

図 13: $(\xi, \omega, \alpha) = (0, 1, 3), (0, 1, -3)$ **のときの非対称正規分布の確率密度関数のプロット**

非対称正規分布に関する α は非対称母数 (skew parameter) と呼ばれ, 特に $\alpha = 0$ のとき,

$$f_{\mathsf{SN}}(x \mid \xi, \omega, 0) = \frac{1}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)$$

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. このこと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.

確率変数 X が非対称正規分布 $\mathsf{SN}(\xi, \omega^2, \alpha)$ 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,

$$g(z \mid \alpha) := 2\phi(z)\,\Phi(\alpha z) \tag{2}$$

となり, これは $\mathsf{SN}(0, 1, \alpha)$ の確率密度関数である.

Azzalini and Capitanio (2014) で与えられているように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 Z^2 は自由度 1 のカイ自乗分布に従う:

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

(Proposition 2.1 (e) 参照.) なお, データが非対称正規分布に従っているかどうかを Q-Q, P-P プロットによって判断するときは, この結果を利用することに注意しよう[13].

B.2 対数非対称正規分布

確率変数 Y 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.

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

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

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

C 非対称ティー分布と対数非対称ティー分布

C.1 非対称ティー分布

通常, 独立な確率変数 $Z \sim \mathsf{N}(0, 1)$ と $V \sim \chi^2_\nu/\nu$ を使った以下の比に関して以下のことが成り立つことを思いだそう:

$$\frac{Z}{\sqrt{V}} \sim \mathsf{t}_\nu \text{(: 自由度 } \nu \text{ のティー分布)}$$

13) いわゆる, Hearly のプロット (Hearly (1968)) と呼ばれる可視化の手法である.

ここで, 確率変数 Z_0 が非対象正規分布 $\mathsf{SN}(0,1,\alpha)$ に従うとき,

$$T := \frac{Z_0}{\sqrt{V}}$$

の分布は非対称ティー分布 $\mathsf{ST}(0,1,\alpha,\nu)$ と呼ばれる. (Azzalini and Capitanio (2014) 参照.) ここで, ν はティー分布の場合と同様に自由度と呼ばれる.

非対称ティー分布 $\mathsf{ST}(0,1,\alpha,\nu)$ の確率密度関数は,

$$\psi(z \mid \alpha,\nu) := 2f_{\mathrm{t}}(z \mid \nu)F_{\mathrm{t}}\left(\alpha z\sqrt{\frac{\nu+1}{z^2+\nu}} \mid \nu+1\right)$$

で与えられる. ここで,

$$f_{\mathrm{t}}(z \mid \nu) := \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}}\left(1+\frac{z^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

は自由度 ν のティー分布 t_ν の確率密度関数であり,

$$F_{\mathrm{t}}(z \mid \nu) = \int_{-\infty}^{z} f_{\mathrm{t}}(x \mid \nu)\mathrm{d}x$$

は自由度 ν のティー分布の累積分布関数である.

非対称正規分布 $\mathsf{SN}(0,1,\alpha)$ に従う確率変数 Z_0 に関して,

$$Z_0^2 \sim \chi_1^2$$

が成り立つことと, 非対称ティー分布の導出の過程から, $T \sim \mathsf{ST}(0,1,\alpha,\nu)$ のとき,

$$T^2 = \frac{Z_0^2}{V} \sim \frac{\chi_1^2/1}{\chi_\nu^2/\nu} \overset{\mathrm{d}}{=} \mathsf{F}_\nu^1(\text{: 自由度 } (1\nu) \text{ のエフ分布})$$

となり, これは通常のティー統計量の 2 乗がエフ分布に従う結果と一致することに注意しよう. なお, 非対称正規分布の場合と同様に, データが非対称ティー分布に従っているかどうかを Q-Q プロット等によって判断するときは, この結果を利用することに注意しよう.

確率変数 Z が非対称ティー分布 $\mathsf{ST}(0,1,\alpha,\nu)$ に従うとき,

$$X = \xi + \omega Z$$

によって得られる分布の確率密度関数は,

$$x = \xi + \omega z \Longleftrightarrow z = \frac{x-\xi}{\omega} \Longrightarrow \mathrm{d}z = \frac{1}{\omega}\mathrm{d}x$$

より,

$$\psi(z \mid \alpha, \nu)\mathrm{d}z = \psi\left(\frac{x-\xi}{\omega} \mid \alpha, \nu\right)\frac{1}{\omega}\mathrm{d}x =: f_{\mathsf{ST}}(x \mid \xi, \omega, \alpha, \nu)\mathrm{d}x$$

で与えられる. この場合の非対称ティー分布は 4 個の母数を持つ分布であり, $\mathsf{ST}(\xi, \omega^2, \alpha, \nu)$ と記号として表され, 確率変数 X がこの分布に従うとき,

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

と書かれる.

図 14 に $(\xi, \omega, \alpha, \nu) = (0, 1, 3, 1), (0, 1, -3, 1)$ のときの非対称ティー分布の確率密度関数のプロットを与える.

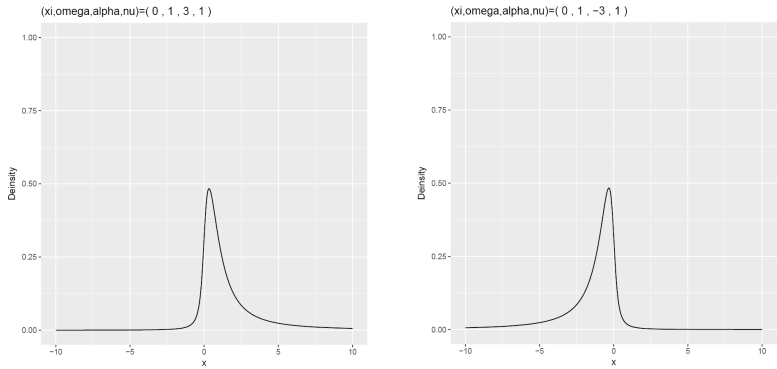

図 14: $(\xi, \omega, \alpha, \nu) = (0, 1, 3, 1), (0, 1, -3, 1)$ **のときの非対称ティー分布の確率密度関数のプロット**

注意 2 (非対称コーシー分布) 一般に, 自由度 $\nu = 1$ の場合のティー分布 t_1 はコーシー分布と呼ばれるけれども, これと類似して, 自由度 $\nu = 1$ の非対称ティー分布 $\mathsf{ST}(0, 1, \alpha, 1)$ は非対称コーシー分布 (skew Cauchy distribution) と呼ばれる. なお, 図 14 で与えられた非対象ティー分布の確率密度関数は, $\nu = 1$ の場合であるので, 厳密には非対称コーシー分布のものであることに注意しよう.

C.2 対数非対称ティー分布

確率変数 Y に対して, その対数 $\log(Y)$ が非対称ティー分布 $\mathsf{ST}(\xi, \omega^2, \alpha, \nu)$ に従うとき, Y は対数非対称ティー分布 $\mathsf{LST}(\xi, \omega^2, \alpha, \nu)$ に従うといわれる.

$$Y \sim \mathsf{LST}(\xi, \omega^2, \alpha, \nu) \overset{\text{def.}}{\Longleftrightarrow} \log(Y) \sim \mathsf{ST}(\xi, \omega^2, \alpha, \nu)$$

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

D 竹内情報量規準

データを発生させる真の分布の確率密度関数を $g(x)$ とする. 一方, 想定されるモデル族を $\mathcal{F}_{\boldsymbol{\theta}} := \{f(x \mid \boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p\}$ とする. 一般に, 真の分布は想定されるモデル族に属すとは限らないので,

$$g(x) \notin \mathcal{F}_{\boldsymbol{\theta}}$$

と仮定したとき, 真の分布に従う無作為標本 $\{X_1, \ldots, X_n\}$ の具体的な観測値 (データ) $\{x_1, \ldots, x_n\}$ が得られたときに, モデル $f(x \mid \boldsymbol{\theta})$ が適切かどうかを評価する適当な規準として以下の竹内情報量規準 (竹内 (1976), 小西, 北川 (2004) 参照) がある:

$$\mathrm{TIC} := -2\ell_n(\widehat{\boldsymbol{\theta}}) + 2\mathrm{trace}\, \widehat{\mathbf{I}}(\widehat{\boldsymbol{\theta}})\widehat{\mathbf{J}}^{-1}(\widehat{\boldsymbol{\theta}}) \tag{3}$$

ここで,

$$\ell_n(\boldsymbol{\theta}) := \sum_{i=1}^{n} \log f(x_i \mid \boldsymbol{\theta})$$

は対数尤度関数であり, $\widehat{\boldsymbol{\theta}}$ は以下で定義される最尤推定値 (ベクトル) である:

$$\widehat{\boldsymbol{\theta}} := \arg\max_{\boldsymbol{\theta} \in \Theta} \ell_n(\boldsymbol{\theta})$$

なお, $\arg\max_{x \in D} f(x)$ は $f(x)$ を最大にする $x(\in D)$ の値を表す. また,

$$\widehat{\mathbf{I}}(\widehat{\boldsymbol{\theta}}) := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}(\widehat{\boldsymbol{\theta}})\boldsymbol{g}(\widehat{\boldsymbol{\theta}})', \quad \widehat{\mathbf{J}}(\widehat{\boldsymbol{\theta}}) := -\frac{1}{n}\sum_{i=1}^{n} \mathbf{H}_i(\widehat{\boldsymbol{\theta}})$$

であり,

$$\boldsymbol{g}_i(\boldsymbol{\theta}) := \frac{\partial \log f(x_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}; \text{勾配ベクトル}, \quad \mathbf{H}_i(\boldsymbol{\theta}) := \frac{\partial^2 \log f(x_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}; \text{ヘッセ行列}$$

である.

注意 3 (竹内情報量規準と赤池情報量規準の関係) 真の分布 $g(x)$ が想定されるモデル族に属すとき, すなわち,

$$g(x) \in \mathcal{F}_{\boldsymbol{\theta}} \Longleftrightarrow \exists \boldsymbol{\theta}_0 \text{ s.t. } g(x) = f(x \mid \boldsymbol{\theta}_0)$$

が成り立つ場合は, 漸近的に $\widehat{\mathbf{I}}(\widehat{\boldsymbol{\theta}}) \overset{a}{=} \mathbf{I}(\boldsymbol{\theta}_0) = \mathbf{J}(\boldsymbol{\theta}_0) \overset{a}{=} \widehat{\mathbf{J}}(\widehat{\boldsymbol{\theta}})$ が成り立つことから,

$$\text{trace } \widehat{\mathbf{I}}(\widehat{\boldsymbol{\theta}})\widehat{\mathbf{J}}^{-1}(\widehat{\boldsymbol{\theta}}) \overset{a}{=} \text{trace } \mathbf{I}(\boldsymbol{\theta}_0)\mathbf{I}^{-1}(\boldsymbol{\theta}_0) = \text{trace } \mathbf{I}_p = p$$

となる. ここで, $X \sim G$(; 真の分布) としたとき,

$$\mathbf{I}(\boldsymbol{\theta}) := \mathrm{E}_G\left(\frac{\partial f(X \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial f(X \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right); \text{ フィッシャー情報行列,}$$

$$\mathbf{J}(\boldsymbol{\theta}) := -\mathrm{E}_G\left(\frac{\partial^2 f(X \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right)$$

であり, E_G は G のもとでの期待値を表す. また, $\mathbf{I}_p := \text{diag}(1, \ldots, 1)$ は p 次の単位行列である. この結果から,

$$\text{TIC} \overset{a}{=} -2\ell_n(\widehat{\boldsymbol{\theta}}) + 2p = \text{AIC}$$

となり, 竹内情報量規準は赤池情報量規準と (漸近的に) 一致することに注意しよう.

竹内情報量規準を求めるためには, 定義より, 対数尤度関数 $\ell(\boldsymbol{\theta})$ と最尤推定値 $\widehat{\boldsymbol{\theta}}$, 対数密度の勾配ベクトル $\boldsymbol{g}_i(\boldsymbol{\theta})$, ヘッセ行列 $\mathbf{H}_i(\boldsymbol{\theta})$ を求める必要があることがわかる. 以下に正規分布と非対称正規分布に対するこれらの量を与える.

D.1 正規分布の場合

無作為標本 $\{X_1, \ldots, X_n\}$ が正規分布 $\mathsf{N}(\mu, \sigma^2)$ に従うとき, 母数ベクトルを $\boldsymbol{\theta} = [\mu, \sigma]'$ とすると, 対数尤度関数は

$$\ell_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(x_i \mid \boldsymbol{\theta}) = \sum_{i=1}^{n} \left\{-\frac{1}{2}\log(2\pi) - \log(\sigma) - \frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right\}$$

であり, 最尤推定値は, $\widehat{\boldsymbol{\theta}} = [\widehat{\mu}, \widehat{\sigma}]' = [\overline{x}, s]'$ と陽に書くことができる. ここで, $\overline{x} := \sum_{i=1}^{n} x_i/n$(: データの平均), $s := \sqrt{s^2} = \sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2/n}$ (: データの標準偏差) である.

対数密度関数の勾配ベクトルとヘッセ行列は以下で与えられる:

$$\boldsymbol{g}_i(\boldsymbol{\theta}) = \frac{\partial \log f(x_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \dfrac{\partial \log f(x_i \mid \boldsymbol{\theta})}{\partial \mu} \\ \dfrac{\partial \log f(x_i \mid \boldsymbol{\theta})}{\partial \sigma} \end{bmatrix} = \begin{bmatrix} \dfrac{x_i - \mu}{\sigma^2} \\ -\dfrac{1}{\sigma} + \dfrac{(x_i - \mu)^2}{\sigma^3} \end{bmatrix}$$

$$\mathbf{H}_i(\boldsymbol{\theta}) = \frac{\partial^2 \log f(x_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \begin{bmatrix} \dfrac{\partial^2 \log f(x_i \mid \boldsymbol{\theta})}{\partial \mu^2} & \dfrac{\partial^2 \log f(x_i \mid \boldsymbol{\theta})}{\partial \mu \partial \sigma} \\ \dfrac{\partial^2 \log f(x_i \mid \boldsymbol{\theta})}{\partial \sigma \partial \mu} & \dfrac{\partial^2 \log f(x_i \mid \boldsymbol{\theta})}{\partial \sigma^2} \end{bmatrix}$$

$$= \begin{bmatrix} -\dfrac{1}{\sigma^2} & -2\dfrac{x_i - \mu}{\sigma^3} \\ -2\dfrac{x_i - \mu}{\sigma^3} & \dfrac{1}{\sigma^2} - \dfrac{3(x_i - \mu)^2}{\sigma^4} \end{bmatrix}$$

この結果から,

$$\widehat{\mathbf{I}}(\widehat{\boldsymbol{\theta}}) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{g}_i(\widehat{\boldsymbol{\theta}}) \boldsymbol{g}_i(\widehat{\boldsymbol{\theta}})' = \begin{bmatrix} \dfrac{1}{s^2} & \dfrac{m_3}{s^5} \\ \dfrac{m_3}{s^5} & -\dfrac{1}{s^2} + \dfrac{m_4}{s^6} \end{bmatrix},$$

$$\widehat{\mathbf{J}}(\widehat{\boldsymbol{\theta}}) = -\frac{1}{n} \sum_{i=1}^{n} \mathbf{H}_i(\widehat{\boldsymbol{\theta}}) = \begin{bmatrix} \dfrac{1}{s^2} & 0 \\ 0 & \dfrac{2}{s^2} \end{bmatrix}$$

となり, 結局,

$$\begin{aligned} \operatorname{trace} \widehat{\mathbf{I}}(\widehat{\boldsymbol{\theta}}) \widehat{\mathbf{J}}^{-1}(\widehat{\boldsymbol{\theta}}) &= \operatorname{trace} \begin{bmatrix} \dfrac{1}{s^2} & \dfrac{m_3}{s^5} \\ \dfrac{m_3}{s^5} & -\dfrac{1}{s^2} + \dfrac{m_4}{s^6} \end{bmatrix} \begin{bmatrix} \dfrac{1}{s^2} & 0 \\ 0 & \dfrac{2}{s^2} \end{bmatrix}^{-1} \\ &= \operatorname{trace} \begin{bmatrix} 1 & \dfrac{m_3}{2s^3} \\ \dfrac{m_3}{2s^3} & -\dfrac{1}{2} + \dfrac{m_4}{2s^4} \end{bmatrix} \\ &= \frac{1}{2} + \frac{m_4}{2s^4} \end{aligned}$$

となる. ここで, $m_k := \sum_{i=1}^{n} (x_i - \overline{x})^k / n$ はデータの平均まわりの k 次モーメントである.

この結果から, 正規分布の場合の竹内情報量規準は,

$$\begin{aligned} \mathrm{TIC}_{\mathsf{N}} &= -2\ell_n(\widehat{\boldsymbol{\theta}}) + 2\operatorname{trace} \widehat{\mathbf{I}}(\widehat{\boldsymbol{\theta}}) \widehat{\mathbf{J}}^{-1}(\widehat{\boldsymbol{\theta}}) \\ &= -2 \left\{ -\frac{n}{2} \log(2\pi) - n \log(s) - \frac{n}{2} \right\} + 2 \left(\frac{1}{2} + \frac{m_4}{2s^4} \right) \\ &= n \log(2\pi) + n \log(s^2) + n + 1 + \frac{m_4}{s^4} \end{aligned}$$

と陽に表現できる．なお, 赤池報量規準は,

$$\begin{aligned}\mathrm{AIC}_{\mathsf{N}} &= -2\ell_n(\widehat{\boldsymbol{\theta}}) + 2p = -2\left\{-\frac{n}{2}\log(2\pi) - n\log(s) - \frac{n}{2}\right\} + 2\times 2 \\ &= n\log(2\pi) + n\log(s^2) + n + 4\end{aligned}$$

である．

D.2 非対称正規分布の場合

無作為標本 $\{X_1, \ldots, X_n\}$ が非対称正規分布 $\mathsf{SN}(\xi, \omega^2, \alpha)$ に従うとき, 母数ベクトルを $\boldsymbol{\theta} = [\xi, \omega, \alpha]'$ とすると, 対数尤度関数は

$$\ell_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(x_i \mid \boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - n\log(\omega) - \frac{1}{2}\sum_{i=1}^{n} z_i^2 + \sum_{i=1}^{n} \zeta_0\left(\alpha z_i\right)$$

である．ここで, $z_i := (x_i - \xi)/\omega$ であり, $\zeta_0(x) := \log\{2\Phi(x)\}$ である．なお, $\Phi(x)$ は標準正規分布の累積分布関数であることを思いだそう．対数非対称正規分布の最尤推定値 $\widehat{\boldsymbol{\theta}} = [\widehat{\xi}, \widehat{\omega}, \widehat{\alpha}]'$ は残念ながら陽には表現できないので, 尤度方程式 $\partial\ell_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = \mathbf{0}$ を満たす $\boldsymbol{\theta}$ を Newton-Raphson 法などの繰り返し法で数値的に求める必要があることに注意しよう．

対数密度関数の勾配ベクトルとヘッセ行列は以下で与えられる:

$$\boldsymbol{g}_i(\boldsymbol{\theta}) = \frac{\partial \log f(x_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \dfrac{\partial \log f(x_i \mid \boldsymbol{\theta})}{\partial \xi} \\ \dfrac{\partial \log f(x_i \mid \boldsymbol{\theta})}{\partial \omega} \\ \dfrac{\partial \log f(x_i \mid \boldsymbol{\theta})}{\partial \alpha} \end{bmatrix} = \begin{bmatrix} \dfrac{1}{\omega}\left(z_i - \alpha\zeta_1\left(\alpha z_i\right)\right) \\ -\dfrac{1}{\omega}\left(1 - z_i^2 + \alpha z_i \zeta_1\left(\alpha z_i\right)\right) \\ z_i \zeta_1\left(\alpha z_i\right) \end{bmatrix}$$

$$\mathbf{H}_i(\boldsymbol{\theta}) = \frac{\partial^2 \log f(x_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} = \begin{bmatrix} \dfrac{\partial^2 \log f(x_i \mid \boldsymbol{\theta})}{\partial \xi^2} & \dfrac{\partial^2 \log f(x_i \mid \boldsymbol{\theta})}{\partial \xi \partial \omega} & \dfrac{\partial^2 \log f(x_i \mid \boldsymbol{\theta})}{\partial \xi \partial \alpha} \\ \dfrac{\partial^2 \log f(x_i \mid \boldsymbol{\theta})}{\partial \omega \partial \xi} & \dfrac{\partial^2 \log f(x_i \mid \boldsymbol{\theta})}{\partial \omega^2} & \dfrac{\partial^2 \log f(x_i \mid \boldsymbol{\theta})}{\partial \omega \partial \alpha} \\ \dfrac{\partial^2 \log f(x_i \mid \boldsymbol{\theta})}{\partial \alpha \partial \xi} & \dfrac{\partial^2 \log f(x_i \mid \boldsymbol{\theta})}{\partial \alpha \partial \omega} & \dfrac{\partial^2 \log f(x_i \mid \boldsymbol{\theta})}{\partial \alpha^2} \end{bmatrix}$$

ここで,

$$\frac{\partial^2 \log f(x_i \mid \boldsymbol{\theta})}{\partial \xi^2} = -\frac{1}{\omega^2}\left(1 - \alpha^2 \zeta_2(\alpha z_i)\right),$$
$$\frac{\partial^2 \log f(x_i \mid \boldsymbol{\theta})}{\partial \xi \partial \omega} = \frac{\partial^2 \log f(x_i \mid \boldsymbol{\theta})}{\partial \omega \partial \xi} = -\frac{1}{\omega^2}\left(2z_i - \alpha\zeta_1(\alpha z_i) - \alpha^2 z_i \zeta_2(\alpha z_i)\right),$$
$$\frac{\partial^2 \log f(x_i \mid \boldsymbol{\theta})}{\partial \xi \partial \alpha} = \frac{\partial^2 \log f(x_i \mid \boldsymbol{\theta})}{\partial \alpha \partial \xi} = -\frac{1}{\omega}\left(\zeta_1(\alpha z_i) + \alpha z_i \zeta_2(\alpha z_i)\right),$$
$$\frac{\partial^2 \log f(x_i \mid \boldsymbol{\theta})}{\partial \omega^2} = -\frac{1}{\omega^2}\left(-1 + 3z_i^2 - 2\alpha z_i \zeta_1(\alpha z_i) - \alpha^2 z_i^2 \zeta_2(\alpha z_i)\right),$$
$$\frac{\partial^2 \log f(x_i \mid \boldsymbol{\theta})}{\partial \omega \partial \alpha} = \frac{\partial^2 \log f(x_i \mid \boldsymbol{\theta})}{\partial \alpha \partial \omega} = -\frac{1}{\omega}\left(z_i \zeta_1(\alpha z_i) + \alpha z_i^2 \zeta_2(\alpha z_i)\right),$$
$$\frac{\partial^2 \log f(x_i \mid \boldsymbol{\theta})}{\partial \alpha^2} = z_i^2 \zeta_2(\alpha z_i)$$

であり,

$$\zeta_r(x) := \frac{\mathrm{d}^r}{\mathrm{d}x^r}\zeta_0(x) = \frac{\mathrm{d}^r}{\mathrm{d}x^r}\log\{2\Phi(x)\}$$

より,

$$\zeta_1(x) = \frac{\phi(x)}{\Phi(x)},$$
$$\zeta_2(x) = \frac{-x\phi(x)\Phi(x) - \phi(x)^2}{\Phi(x)^2} = -x\frac{\phi(x)}{\Phi(x)} - \left(\frac{\phi(x)}{\Phi(x)}\right)^2 = -x\zeta_1(x) - \zeta_1(x)^2$$

である.

非対称正規分布の場合は, 正規分布のときのように竹内情報量規準を簡単に書き下すことは難しいけれども, データからこれらの量を数値的に計算することによって求めることができることに注意しよう. なお, 赤池情報量規準は,

$$\mathrm{AIC}_{\mathsf{SN}} = -2\ell_n(\widehat{\boldsymbol{\theta}}) + 2p = n\log(2\pi) + n\log(\widehat{\omega}^2) + \sum_{i=1}^{n}\widehat{z}_i^2 - 2\sum_{i=1}^{n}\zeta_0(\widehat{\alpha}\widehat{z}_i) + 6$$

である. ここで, $\widehat{z}_i := (x_i - \widehat{\xi})/\widehat{\omega}$ とおいた.

E R パッケージ sn

sn は, Adelchi Azzalini 氏によって開発されている非対称分布族を扱うための R パッケージである. 本稿で利用した非対分布族に関する関数は, 表 8 に与えられているようなものである.

表 8: 本稿で利用した sn パッケージに付属する関数

関数	機能
dsn	非対称正規分布の確率密度関数の計算
psn	非対称正規分布の累積分布関数の計算
qsn	非対称正規分布の分位点の計算
rsn	非対称正規分布の乱数の生成
dst	非対称分布の確率密度関数の計算
pst	非対称ティー分布の累積分布関数の計算
qst	非対称ティー分布の分位点の計算
rst	非対称ティー分布の乱数の生成
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

F R スクリプト

```
#####################################################
### 環境設定，データ読み込み，データ整形
#####################################################
set.seed(12345)
options(width=70)
library(ggplot2)
library(GGally)
library(plyr)
library(dplyr)
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,month==12),
                    firmID,country,SIC.code,sales,employees,assets.total)


#####################################################
### 対散布図のプロット
#####################################################
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)))
```

```
#####################################################
### 売上高の対数への正規分布と非対称正規分布の当てはめ
#####################################################
lm.log.sales2013<-lm(log(sales)~1,data=firmfin2013)
selm.log.sales2013<-selm(log(sales)~1,data=firmfin2013)


#####################################################
### 売上高の対数のヒストグラムと Q-Q プロット (非対称正規分布の統計モデル付き)
#####################################################
plot(selm.log.sales2013,which=2)
plot(selm.log.sales2013,which=3,xlim=c(0,20))


#####################################################
### 売上高の対数への正規分布と非対称ティー分布の当てはめ
#####################################################
selm.ST.log.sales2013<-selm(log(sales)~1,family="ST",data=firmfin2013)


#####################################################
### 売上高の対数のヒストグラムと Q-Q プロット (非対称ティー分布の統計モデル付き)
#####################################################
plot(selm.ST.log.sales2013,which=2)
plot(selm.ST.log.sales2013,which=3)


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


#####################################################
### ティー検定表出力
#####################################################
print(xtable(lm.log.firmfin2013,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 次元散布図と回帰平面 (対数正規線形モデル) の描画
#####################################################
coef.lm.log.2013<-coef(lm.log.firmfin2013)
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)


#####################################################
### 回帰診断プロット (対数正規線形モデル)
#####################################################
par(mfcol=c(2,2))
plot(lm.log.firmfin2013,which=c(1,2,4,5))
```

```
par(mfcol=c(1,1))


#####################################################
### 残差のヒストグラムと統計モデル（対数正規線形モデル）
#####################################################
ggplot(data.frame(resid=resid(lm.log.firmfin2013)),aes(x=resid))+
  geom_histogram(aes(y=..density..),binwidth=0.5,fill="white",color="black")+
  stat_function(fun=dnorm, args=list(mean=0,sd= summary(lm.log.firmfin2013)$sigma))


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


#####################################################
### ゼット検定表出力（対数非対称正規線形モデル）
#####################################################
print(xtable(summary(selm.log.firmfin2013,"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 次元散布図と回帰平面（対数非対称正規線形モデル）の描画
#####################################################
coef.selm.log.2013<-coef(selm.log.firmfin2013)
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)


#####################################################
### 回帰診断プロット（対数非対称正規線形モデル）
#####################################################
par(mfcol=c(2,2))
plot(selm.log.firmfin2013,param.type="DP",which=1)
plot(selm.log.firmfin2013,param.type="DP",which=2)
plot(selm.log.firmfin2013,param.type="DP",which=3,xlim=c(0,60))
plot(selm.log.firmfin2013,param.type="DP",which=4)
par(mfcol=c(1,1))


#####################################################
### 対数非対称ティー線形モデルの当てはめ
#####################################################
selm.ST.log.firmfin2013<-selm(log(sales)~log(employees)+log(assets.total),
                              family="ST", data=firmfin2013)


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

```
###  ゼット検定表出力（対数非対称ティー線形モデル）
#####################################################
print(xtable(summary(selm.ST.log.firmfin2013,"DP")@param.table,
floating=FALSE,
caption=c("ゼット比検定表：対数非対称ティー線形モデル"),
label="table.z.logskew-t.linear.model"),
caption.placement = "top",table.placement="H")


#####################################################
### 3 次元散布図と回帰平面（対数非対称ティー線形モデル）の描画
#####################################################
coef.selm.ST.log.2013<-coef(selm.ST.log.firmfin2013,param.type="DP")
plot3d(log(firmfin2013[c("employees","assets.total","sales")]),type = "s",
       col = "red", size = 1)
planes3d(coef.selm.ST.log.2013[2],coef.selm.ST.log.2013[3],-1,
         coef.selm.ST.log.2013[1],alpha=0.5)


#####################################################
### 回帰診断プロット（対数非対称ティー線形モデル）
#####################################################
par(mfcol=c(2,2))
plot(selm.ST.log.firmfin2013,param.type="DP")
par(mfcol=c(1,1))


#####################################################
### AIC 表出力（売上高の分布比較）
#####################################################
print(xtable(AIC(lm.log.sales2013,selm.log.sales2013,selm.ST.log.sales2013),
floating=FALSE,caption=c("AIC 表：売上高の分布に関する比較"),
label="table.distribution.AIC"),
caption.placement = "top",table.placement="H")


#####################################################
### 竹内情報量規準の計算（正規分布，非対称性正規分布の場合）
#####################################################
TIC.norm<-function(x)
{
  n<-length(x)
  m<-mean(x)
  s<-sqrt(mean((x-m)^2))
  m4<-mean((x-m)^4)
  TIC.norm<-n*log(2*pi*s^2)+n+1+m4/s^4
  AIC.norm<-n*log(2*pi*s^2)+n+4
  list(AIC.norm=AIC.norm,TIC.norm=TIC.norm,n=n,m=m,s=s,m4=m4,p=2)
}
TIC.sn<-function(x,xi,omega,alpha)
{
  n<-length(x)
  zeta0<-function(x) log(2*pnorm(x))
  zeta1<-function(x) dnorm(x)/pnorm(x)
  zeta2<-function(x) -x*zeta1(x)-zeta1(x)^2
  z<-(x-xi)/omega
  logL <- -n*(log(2*pi))/2-n*log(omega)-sum(z^2)/2+sum(zeta0(alpha*z))
```

```
  G<-cbind((z-alpha*zeta1(alpha*z))/omega,-(1-z^2+alpha*z*zeta1(alpha*z))/omega,
           z*zeta1(alpha*z))
  I<-t(G)%*%G/n
  H<-array(0,c(3,3,n))
  H[1,1,] <- -(1-alpha^2*zeta2(alpha*z))/omega^2
  H[1,2,] <- -(2*z-alpha*zeta1(alpha*z)-alpha^2*z*zeta2(alpha*z))/omega^2
  H[2,1,] <- -(2*z-alpha*zeta1(alpha*z)-alpha^2*z*zeta2(alpha*z))/omega^2
  H[1,3,] <- -(zeta1(alpha*z)+alpha*z*zeta2(alpha*z))/omega
  H[3,1,] <- -(zeta1(alpha*z)+alpha*z*zeta2(alpha*z))/omega
  H[2,2,] <- -(-1+3*z^2-2*alpha*z*zeta1(alpha*z)-alpha^2*z^2*zeta2(alpha*z))/omega^2
  H[2,3,] <- -(z*zeta1(alpha*z)+alpha*z^2*zeta2(alpha*z))/omega
  H[3,2,] <- -(z*zeta1(alpha*z)+alpha*z^2*zeta2(alpha*z))/omega
  H[3,3,] <- z^2*zeta2(alpha*z)
  J <- -apply(H,c(1,2),mean)
  TIC.sn <- -2*logL+2*sum(diag(I%*%solve(J)))
  AIC.sn<--2*logL+6
  list(AIC.sn=AIC.sn,TIC.sn=TIC.sn,logL=logL,H=H,I=I,J=J,n=n,p=3)
}
logsales2013.TIC.n<-TIC.norm(log(firmfin2013$sales))
logsales2013.TIC.sn<-TIC.sn(log(firmfin2013$sales),
                        xi=coef(selm.log.sales2013,"DP")[1],
                        omega=coef(selm.log.sales2013,"DP")[2],
                        alpha=coef(selm.log.sales2013,"DP")[3])


#####################################################
### AIC 表出力 (売上高の対数線形モデル比較)
#####################################################
print(xtable(AIC(lm.log.firmfin2013,selm.log.firmfin2013,selm.ST.log.firmfin2013),
floating=FALSE,caption=c("AIC 表: 売上高の対数線形モデルに関する比較"),
             label="table.AIC"),
caption.placement = "top",table.placement="H")


#####################################################
### データ一部出力
#####################################################
options(width=200)
firmfin2013[seq(10),]
options(width=70)


#####################################################
### 非対称正規分布の確率密度関数を描画する関数
#####################################################
ggplot.sn<-function(x=seq(-10,10,0.01),xi=0,omega=1,alpha=0,yup=1)
{
  require(sn,ggplot2)
  t<-ggplot(data.frame(x,y=dsn(x,xi=xi,omega=omega,alpha=alpha)),aes(x,y))+
     geom_line() + ylim(0,yup)
  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描画
#####################################################
```

```
ggplot.sn(x=seq(-10,10,0.01),alpha=3)
ggplot.sn(x=seq(-10,10,0.01),alpha=-3)


#####################################################
### 非対称ティー分布の確率密度関数を描画する関数
#####################################################
ggplot.st<-function(x=seq(-10,10,0.01),xi=0,omega=1,alpha=0,nu=0,yup=1)
{
  require(sn,ggplot2)
  t<-ggplot(data.frame(x,y=dst(x,xi=xi,omega=omega,alpha=alpha,nu=nu)),aes(x,y))
     +geom_line()+ylim(0,yup)
  t+labs(title=paste("(xi,omega,alpha,nu)=(",xi,",",omega,",",alpha,",",nu,")"),
         y="Deinsity")
}


#####################################################
### 非対称ティー分布の確率密度関数の描画
#####################################################
ggplot.st(x=seq(-10,10,0.01),alpha=3,nu=1)
ggplot.st(x=seq(-10,10,0.01),alpha=-3,nu=1)
```