# 「裾の厚い分布」を用いたベイジアン頑強回帰モデルのシミュレーション分析

林　　由　子*

要旨

本稿では，ベイジアン線形回帰モデルにおける外れ値問題について検討を行っていく。ベイズ統計学における外れ値問題を「裾の厚い分布」によって自動的に検出し棄却するモデルがDawid（1973）によって示された。これを線形回帰モデルに適応した際の外れ値検出のパフォーマンスについてシミュレーション分析を行い，裾の厚い分布を用いたベイジアン頑強回帰モデルが有効な範囲の特定化を試みた。その結果，モデルは外れ値が外れ値以外のデータ集団がとる範囲の近傍では頑強であるが，大きく離れた場合には頑強モデルとして有効ではないことが明らかになった。

## １．は　じ　め　に

本稿では，ベイジアンモデルにおける外れ値への対処法について検討を行っていく。

1852年，Peirce によって問題提起されて以来，外れ値問題は，150年以上にわたり統計学における研究対象として関心を持たれてきた重要な課題の一つである。また外れ値問題は，頻度論学派のみならず，ベイジアンにとっても重要なテーマである。

ベイズ統計学における外れ値問題は，De Finetti（1961）によって提起されたものである。その後，汚染された分布との混合モデルを想定した Box and Tiao（1968），モデルを想定しない Zellner（1975），裾の厚い分布によって対処を行った Dawid（1973）等によって展開を見せた。

Dawid（1973）の「裾の厚い分布」による外れ値への対処法は，O'Hagan（1979）によって離散モデルが示されたのち，O'Hagan（1988）による応用例が示された。また，Andrade and O'Hagan（2006）においては，尺度パラメータへの拡張がなされた。

ベイズ統計学を用いたこの裾の厚い分布による外れ値への線形回帰モデルへの応用例にO'Hagan（1988）がある。O'Hagan（1988）では定数項のない線形回帰モデルにおいてどのように外れ値が自動的に検出され棄却されていくかについてのパフォーマンスが示されている。しかしこれに対して Peña et al.（2009）では，裾の厚い分布を用いたベイジアン線形回帰モデルは一般には頑強ではないことが示された。そこで本稿では，線形回帰モデルを定数項を持つモデルに拡張し，シミュレーション分析から，モデルが有効である範囲の

---

＊ 小稿を，奥田聡先生に捧げるとともに先生のご冥福を心よりお祈り申し上げます。

＊＊ 本稿は，2008/2009年度大阪経済大学共同研究費の成果の一部である。

特定化を行っていく。

本稿の構成は以下のとおりである。まず第２節で「外れ値」についての考察を行った後，頻度論学派による外れ値への対処法について渉猟をおこなう。そして第３節でベイジアンモデルにおける外れ値の対処法の一つである，Dawid（1973）の裾の厚い分布を用いたモデルの紹介をおこなう。そして第４節でシミュレーション結果を示す。最後に，第５節でまとめをおこなう。

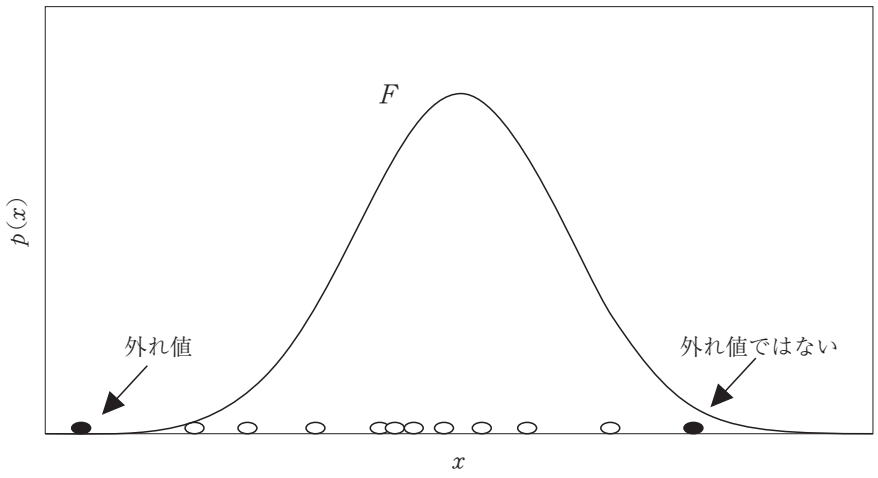
## ２．外　れ　値

本節では，「外れ値（outlier)」をどのように定義し，取り扱ってきたかについての検討を基本的に Barnett and Lewis（1994）に従って行っていく。

「「外れ値」は，そのデータセットにおける他の観測値の変動パターンに属さない観測値」とされる（Johnson and Wichern（2007））。

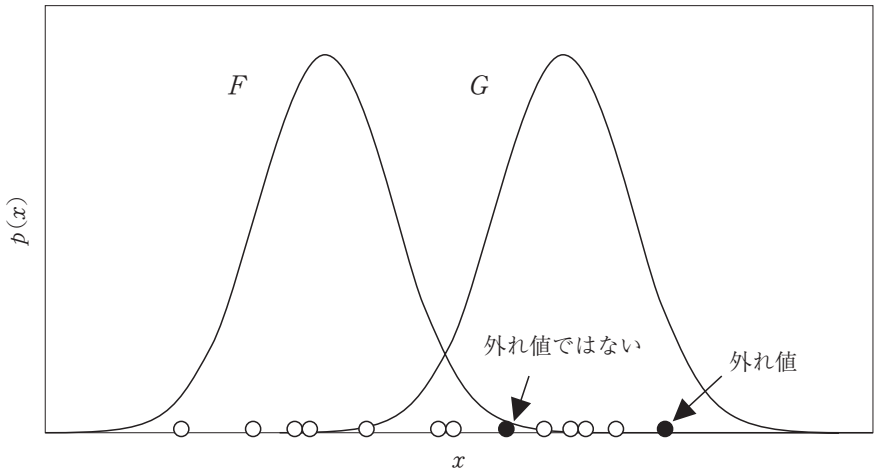
「外れ値」をあらわす表現としては，「みせかけの（spurious）データ」，「意外な（surprising）データ」，「だまし（rogue）データ」といったものに加え，「汚染値（contaminant)」あるいは，「異端値（maverick)」といった表現が見られる。「外れ値」のなかでも，「不一致値（discordant value)」と呼ばれる他のデータと異なる性質をもつものを一般に区別する。一方で，Barnett and Lewis（1994）は，「外れ値」を「意外」かつ「不一致（discordant)」な観測値であるとしている。

外れ値は，それを含むデータセットから示された情報を誤解させることから，汚染されたデータとみなされてきた。しかし外れ値は，それが想定された分布からの１データとして出現しているならば，当然重要な情報をもたらしうる。それゆえ，自動的にデータから除くことは情報を減らしかねない。こうしたケースにおいて，Barnett and Lewis（1994）




出典：Barnett and Lewis（1994）

［図１　外れ値と極値］




出典：Barnett and Lewis（1994）

［図２　外れ値と汚染値］

は，すべてのデータを用いた頑強モデルを構築することが望ましいとし，このことを「調和（accommodation）」と呼んでいる。

外れ値の分析を行うために，最初に，「外れ値」，「極値（extreme value）」，「汚染値」の相違について述べていく。「極値」とは，観測されたデータセットにおける最大値あるいは最小値であり，図１に見られるように，必ずしも「外れ値」であるとは限らない。図１において，最小値の極値は外れ値となっているが，最大値の極値は外れ値ではない。

図２は，「汚染値」と「外れ値」の違いを示している。「汚染値」とは他の観測値とは異なる分布から発生したデータである。ここでは，分布 $F$ と異なる分布 $G$ からデータが発生している。黒点はいずれも分布 $G$ から出現したものであるが，右の黒点は分布 $F$ においては外れ値ではない。

## 2.1　外れ値発生の原因

Barnett and Lewis（1994）は，外れ値発生の原因として，本質的な変動，測定誤差，実行エラーの三つを挙げている。そしてこれらの原因に対し，「調和（accommodation）」と「不一致検定（discordancy test）」の二つの対処方法を挙げている。

最初に，外れ値発生の原因について述べる。外れ値発生の原因は，決定論的な要因と確率的な要因に分かれる。Barnett and Lewis（1994）では，それらの発生原因を以下の三つのカテゴリーに分けている。

（1）本質的変動

母集団の性質から生じる外れ値の原因。たとえば裾の厚い分布から生じたデータを裾の薄い分布を仮定して取り扱うことで，外れ値とみなされる。

（2）測定誤差
　決定論的な要因の一つによるものであり，主に記述ミスにより生じる外れ値の原因。
（3）実行エラー
　偏ったサンプルを用いたデータによって生じる外れ値の原因。

本質的変動および実行エラーは確率的な要因を含む。こうした確率的な発生原因に対し，Barnett and Lewis（1994）では，検定および頑強推定を行っていくために，エラーも異質データも含まない分布 $F$ を帰無仮説とした場合の確率的な変動を持つ対立仮説として，以下の仮説を挙げている。①外れ値が基の分布 $F$ とは異なる，分布 $G$ から生じていると仮定する「本質的対立仮説（Inherent alternative）」。②外れ値がある小さな確率 $\alpha$ で異なる分布 $G$ から生じると仮定することで全体の分布を分布の加法型 $(1-\alpha)F+\alpha G$ を仮定する「混合対立仮説（Mixture alternative）」。③外れ値が，分布は同じであるが異なる位置パラメータあるいは尺度パラメータを分布から生じている仮定する「ずれ対立仮説（Slippage alternative）」。④外れ値が同じ確率で $n$ 個の中 $k$ 個，異なる分布 $G$ から生じると仮定する「可換性対立仮説（Exchangeable alternative）」等を示している。

## 2.3　一変量モデルにおける検定

既述のような外れ値に対して，検出の基準として以下のようなものがある。

（1）Chauvenet's criterion
　外れ値が $n$ 個のデータすべての観測値による平均値からの偏差の大きさをもつ確率を他のデータも持つ時，その確率が $1/2n$ よりも小さくなるならば外れ値は取り除かれる。
（2）Smirnov=Grubbs test
　Smirnov=Grubbs test では以下の統計量に基づいてデータの棄却をおこなう。

$$SG=\frac{|x_i-\bar{x}|}{\sqrt{\hat{\sigma}_x^2}}$$

（3）Cochran test statistic
　Cochran test statistic は，正規分布ではなく，むしろガンマ分布に従う変数に対する検定統計量である。

$$C=\frac{x_n}{\sum_{i=1}^{n}x_i}$$

**外れ値領域（Outlier region）**

ある変数が正規分布 $N(\mu, \sigma^2)$ に従う時，信頼係数 $\alpha(0<\alpha<1)$ を用いて，

$$\text{Outlier region}=\{x : |x-\mu|>z_{1-\alpha/2}\cdot\sigma\}$$

となる（Ben-Gal（2005））。外れ値が $m$ 個存在する場合には，信頼係数 $\alpha$ は，Bonferroni の補正を用いて以下の修正がなされる。

$$\alpha_m = 1-(1-\alpha)^{1/m}$$

### 2.3　頑強性

「頑強」の統計学的概念は，一般的には外れ値に対して統計学的に抵抗を持つことをいう[1)]。Anderson（2008）では，Mosteller and Tukey の以下のような頑強性についての定義を示している。

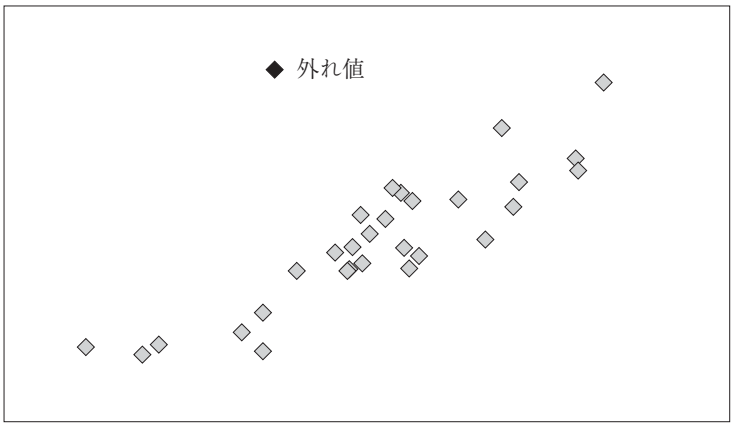
頑強推定量とは，以下の二つの条件を満たすものである。

（1）Robustness of validity
　　データの小さな変化が推定値に大きな変化をもたらさない。

（2）Robustness of efficiency
　　分布に関する仮定が大きく推定値の精度に影響しない。

### 2.4　多変量モデルにおける外れ値と検定

ここでは，Fieller（2005）に従って多変量モデルに対する外れ値の定義および検出方法について述べていく。一変量モデルにおいては，外れ値は，最大値あるいは最小値を見ることで容易に識別することができる。これに対して，多変量モデルにおける外れ値の識別は困難を伴う。図３に二変量モデルにおける外れ値を図示している。二変量モデルにおける外れ値は，図のように，二変数の相関関係を損なうデータとして位置付けることができる。しかしこの外れ値は，Ｘ軸であらわされる変数，Ｙ軸であらわされる変数のどちらか




［図３　二変量モデルにおける外れ値］

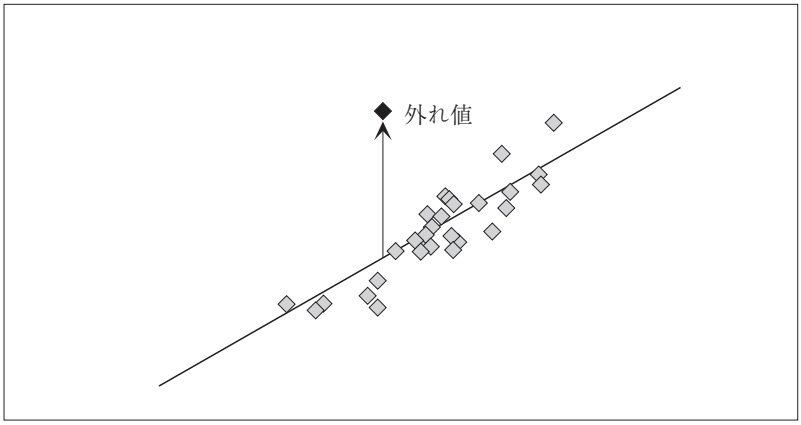
1）これに対して Barnett and Lewis（1994）では，「頑強性」は外れ値問題のみならずより一般的な概念であると述べ，特に外れ値に対する頑強推定を「調和（accommodation）」と区別している。

らも外れ値とは考えにくい。

変数の数が２を超えると外れ値を平面上であらわすことはできない。こうした場合には主成分分析を用いて，第１および第２主成分のプロットから外れ値を図示する手法がとられる。また，多変量モデルの場合の外れ値であるか否かの数量的検討にはマハラノビスの距離が用いられる。

### 2.5　線形回帰モデルにおける外れ値

線形回帰モデルにおける外れ値もまた，二変量モデルと同様に，外れ値は散布図上で図示しうる。しかし線形回帰モデルにおいては，外れ値はむしろ従属変数の条件付き期待値からどれだけ離れているかを基準として定義され，一変量における外れ値の一つのケースとして捉えることができる。回帰直線は，独立変数で条件づけられた期待値であるから，外れ値は従属変数に生じる確率的要因である。




［図４　線形回帰モデルにおける外れ値］

線形回帰モデルにおける頑強性

Anderson（2008）では，外れ値を「調和（accommodate）する」頑強回帰の定義として以下の二つを提示している[2)]。

（1）誤差項に裾の厚い分布を用いて外れ値を調和させる回帰モデル

（2）妥当性の頑強性と効率性の頑強性の両方を併せ持つ回帰モデル

## ３．ベイジアン推論における外れ値

頻度論学派が外れ値を外部的な影響とみなしていることに対して，Barnett and Lewis（1994）は，そうした対処は，ベイズ統計学おける「不当な先入観（unwarranted preoccupation）」であるとしている。

2）Anderson（2008）では，このほかに分散不均一に関する線形回帰モデルの頑強性について触れている。

De Finetti（1961）が述べているように，ベイジアンモデルにおいてすべての推論は事後分布によってあらわされるものであり，外れ値の棄却もまた事後分布に依存してなされるべきである。そしてデータを取り除くことで大きな影響が生じないことが望ましいモデルである。このことは頻度論学派とはまったく逆の見方である。

Freeman（1980）では，ベイジアンモデルにおいて外れ値の定義は，あるデータセットの大部分を生じさせるメカニズムによって生じていない観測値であるとしている。これは，外れ値は汚染されたデータであり，異なる分布から発生していることを基本に置く Box and Tiao（1968）のモデルの中心的考え方である[3)]。一方で，裾の厚い分布を用いたベイジアン頑強モデルが存在する。これは，外れ値とそれ以外のデータに異なる分布を仮定することなく，事後分布において，外れ値を自動的に調和あるいは棄却するものである。

### 3.1　ベイジアンモデルにおける外れ値（正規分布のモデル）

平均 $\mu$，分散の $\sigma^2$ 正規分布に従う変数 $x$ の観測値が 1 個のみ存在するとする。分散 $\sigma^2$ は既知とする。それゆえ尤度は，以下の（1）式となる。

$$f(x|\mu)=\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{1}$$

平均 $\mu$ についての事前分布にも平均 $\mu_0$，分散 $\sigma_0^2$ の正規分布を仮定する。

$$f(\mu)=\frac{1}{\sqrt{2\pi}\,\sigma_0}\exp\left[-\frac{(x-\mu_0)^2}{2\sigma_0^2}\right] \tag{2}$$

この時，事後分布は，容易に以下のように陽表的に整理することができる。

$$f(\mu|x)\propto\exp\left[-\left\{\frac{(x-\mu)^2}{2\sigma^2}+\frac{(x-\mu_0)^2}{2\sigma_0^2}\right\}\right]=\exp\left[-\left\{\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\}\right] \tag{3}$$

ここで $\mu_1=\dfrac{\mu_0\sigma^2+x\sigma_0^2}{\sigma^2+\sigma_0^2}$，$\sigma_1^2=\dfrac{\sigma^2\sigma_0^2}{\sigma^2+\sigma_0^2}$ である。

それゆえ，尤度および事前分布が正規分布に従う時，事後分布もまた正規分布に従う。

さらに，事後分布の分散を整理することで以下の（4）式が得られる。分散の逆数は精度（precise）と定義される。精度が大きければ，情報の集中が大きいことを示す。（4）式より，事後分布の精度が，事前分布と尤度の精度の和に等しいため，事後分布の精度が，事前分布よりも尤度よりも大きくなることがわかる。このことは，追加的な情報によって正確性が増すことを意味する。

$$\sigma_1^2=\frac{\sigma^2\sigma_0^2}{\sigma^2+\sigma_0^2}=\frac{1}{1/\sigma^2+1/\sigma_0^2} \tag{4}$$

### 3.2　矛盾する事後分布（Conflict）

上述のように，事前分布と尤度が正規分布に従う場合，事後分布もまた正規分布に従う。

---

3）このモデルを用いたものとして，例えば Kitagawa and Akaike（1982）がある。

そしてこのとき事後分布の平均は，事前分布と尤度の加重平均となる。つまり事後分布の平均$\mu_1$は事前分布の平均$\mu_0$と尤度の平均$\mu$との間に存在する。このことから事前分布と尤度が大きく離れて分布する場合に「矛盾する事後分布（conflict)」が生じるのである（O'Hagan (1988), and O'Hagan and Forster (2004))。この「矛盾する事後分布」について例を用いて見ていく。

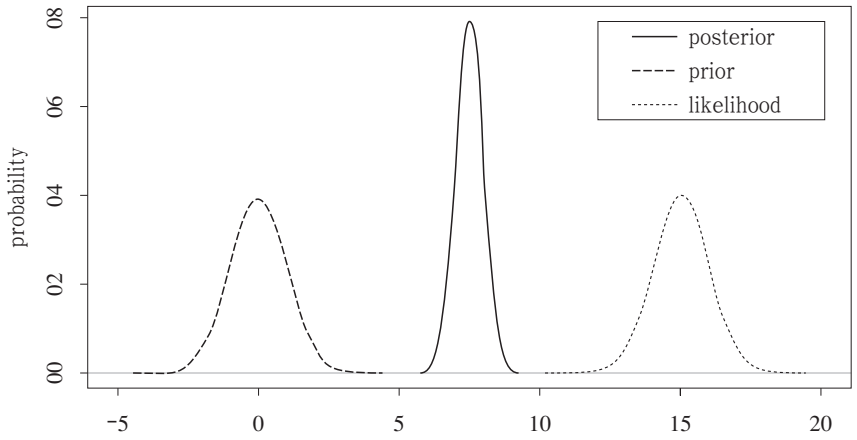
いま，平均$\mu$の事前分布も，尤度も分散が1であるとする（$\sigma_0^2=\sigma^2=1$)。そして，平均$\mu$の事前分布の平均が0であるとする。

$f(\mu)=N(0,\ 1)$：$\mu$の事前分布

$f(x|\mu)=N(\mu,\ 1)$：データ$x$の尤度

図5に，観測値が一つだけ存在し，その値が$x=15$である場合の事前分布，尤度，事後分布を示している。この場合の事後分布は平均7.5，分散0.5となり，事前分布からも尤度からも離れたところに位置することとなる。

この場合，事後分布から発生した尤もらしいと考えうる，たとえば$\mu_1\pm2\sigma_1$（あるいは$\mu_1\pm3\sigma_1$）の範囲内に存在する平均は，事前分布からも尤度からも生じにくい。




［図5　矛盾する事後分布］

### 3.3　ベイジアンモデルにおける頑強性

ここでは，ベイジアンモデルにおける頑強性について主にO'Hagan and Forster (2004)に従って紹介していく。

ベイジアンモデルにおいて，最終的な結論を導く事後分布は，我々が信じている事前分布と尤度に依存する。それゆえ，択一的な事前分布や尤度が存在する場合もある。こうした場合，前者のモデルが後者のモデルに比べてよいと判断できるか否かという問題に直面する。

この問題を評価するために，モデルがいかに頑強であるかを測る方法として包括的感度

および局所的感度がある（O'Hagan and Forster（2004））。

### 3.3.1　包括的感度（Global Robustness（sensitivity））

包括的感度は，同じ分布集合$\Omega$のモデルの中で，あるモデルが頑強であるか否かを評価することにある。そして，その分布族のモデルの全ての推論が十分にそのモデルに近ければモデルは頑強であるとみなす。

集合$\Omega$の代表的なものに以下が挙げられる。

（1）パラメータによる範囲（Parameter range）

$$\Omega=\{f(\theta):c1<\theta<c2\}$$

（2）分布による範囲（Nonparametrical range）

$$l(\theta)<f(\theta)<u(\theta)$$

（3）汚染度による範囲（Contaminated range）

$$\Omega=\{(1-\eta)f(\theta)+\eta q(\theta):q(\theta)\in\Gamma\}$$

こうした集合の中で，$\theta$の統計量$t(\theta)$について以下の事後的統計量を用いて頑強性についての評価をおこなう。

$$T=\sup_{\Omega}E\{t(\theta)\,|\,x\}=\sup_{\Omega}\frac{\int t(\theta)f(\theta)f(x\,|\,\theta)d\theta}{\int f(\theta)f(x\,|\,\theta)d\theta}$$

### 3.3.2　局所的感度（Local sensitivity）

局所的感度はある事後分布をもたらす事前分布における小さな変化による影響を評価するものである。モデル全体の事前分布として以下の$f(\theta)$を仮定する。

$$f(\theta)=(1-\eta)f_0(\theta)+\eta q(\theta)$$

ここでは$q(\theta)$汚染要因の事前分布である。そして，事後分布の推論を吟味するために，事後分布の統計量$h(\eta)$の評価をおこなう。

$$h(\eta)=(1-\eta^*)h_0+\eta\cdot h_q\quad(\text{as }\eta\to0)$$

それゆえ，測度として以下の$H(q)$の絶対値が定義される。

$$H(q)=\lim_{\eta\to0}h(\eta)-h_0=(h_q-h_0)\frac{f_q(x)}{f_0(x)}$$

ここで，$f_0(x)$, $f_q(x)$はそれぞれ以下のような汚染値が含まれない事前分布$f_0(\theta)$と汚染要因の事前分布$q(\theta)$による$x$の予測分布を表す。

$$f_0(x)=\int f(x\,|\,\theta)f_0(\theta)d\theta$$

$$f_q(x)=\int f(x\,|\,\theta)q(\theta)d\theta$$

### 3.3.3　頑強性の評価基準

Peña et al.（2009）では，頑強性の評価基準として，外れ値を含まないデータセット（$X_\alpha$）と全てのデータを含むデータセット（$X$）による事後分布に基づく divergence function, D を用いている。そしてその上限が∞とならないことを基準としている。

$$\sup D(p(\theta|X_\alpha),\ p(\theta|X))<\infty$$

## 3.4　Surprise

ベイジアンモデルにおける外れ値が生じているか否かについて，O'Hagan and Forster（2004）ではこれに対する測度と検定統計量を示している。

Barnet and Lewis（1994）で述べられているように，サプライズの度合いは，外れ値の重要な特質である。ここでは，ベイジアンモデルにおけるサプライズの測度および検定方法についてみていく。

### 3.4.1　Surprising Index

外れ値に対するサプライズインデックスは，どの程度データが外れた位置に生じているのかを測るものである。事後分布は，データ $x$ のとる予測確率 $f_0(x)$ が小さければ，あまり生じにくい状況が生じている，つまり事後分布が疑わしいということになる。これは，$f(\theta)f(x|\theta)$ の事後分布への影響が大きくなることから分かる。

このことは，事前情報が適切か否かとの指標（$\mathrm{SI}_{\mathrm{prior}}$）

$$SI_{prior}=\{f_0(x)\}^{-1}\sup_{\theta\in\Theta}f(x|\theta)$$

と，データが適切であるか否かの指標（$\mathrm{SI}_{\mathrm{data}}$）

$$SI_{data}=\{f_0(x)\}^{-1}\sup_{x}f_0(x)$$

によって測られる。つまり，これらの指標は，データ $x$ がまれにしか起こらないとされるならば，大きくなる測度である（O'Hagan and Forster（2004））。

### 3.4.2　診断

サプライズインデックスは相対的な外れ度を示すものであったのに対し，ここでは，そのデータが外れ値か否かについての診断を行う。O'Hagan and Forster（2004）では以下の２つの診断基準を紹介している[4)]。

「矛盾する事後分布」の項でみてきたように，もしデータ $x_i$ が予測分布 $f(x_i|x_{\setminus i})$ の裾に位置するならば，データ $x_i$ は事後分布に矛盾を引き起こす。このことをこの予測分布に基づく期待値および分散による統計量 $d(x)$ を用いて診断を行う。

$$d(x)=\sum_{i=1}^{n}\frac{(x_i-E(x_i|x_{\setminus i}))^2}{\mathrm{var}(x_i|x_{\setminus i})}$$

4）ただしこれらの統計量は複数の外れ値が存在する場合には適用できない。

ここで $x_{\backslash i}$ は，$x_i$ を除いたデータセットである。

統計量 $d(x)$ が自由度 $n$ の $\chi^2$ 分布に従うものとして検定を行う。ただし，この手法はあくまでも近似的手法であることに注意を要する。

択一的に，以下の定義が用いられる。

$$e_i=\frac{x_i-E(x_i|x_{\backslash i})}{\mathrm{var}(x_i|x_{\backslash i})^{\frac{1}{2}}}$$

### 3.5　ベイジアンモデルにおける外れ値（裾の厚い分布のモデル）

De Finetti（1961）において，ベイジアンモデルを用いた外れ値問題がはじめて論ぜられた。Dawid（1973）は，De Finetti（1961）に基づき，裾の厚い分布のモデルによって，事後分布が外れ値を自動的に棄却するモデルを示した。

O'Hagan and Forster（2004）で述べられているように，この裾の厚い分布のモデルは「モデル批判」に対する解決法あるいは頑強モデルを構築するためのモデルである。ここでは，この裾の厚い分布のモデルについて若干の直感的説明を行った後，Dawid（1973）の仮定および正則変動関数を用いて簡潔に示した Andrade and O'Hagan（2006）の仮定を見ていく。

#### 3.5.1　位置パラメータに対するモデルへの仮定

Dawid（1973）では，連続分布の下で，以下の（5）式の定理が成立する十分条件を示している。これは O'Hagan（1979）によって離散分布に拡張され，Andrade and O'Hagan（2006）では正則変動関数を用いた証明が示された。

まず，Dawid（1973）で提示されている以下の（5）式が成立するための仮定について見ていく[5)]。

$$E(\Theta|X=x)\to E(\Theta)\qquad \text{as}\quad x\to\infty \tag{5}$$

この定理を示すために，Dawid（1973）では以下の 3 つの条件が必要であることを示している。

（1）$\varepsilon>0$ かつ $h>0$ の下で，以下を満たす $A$ が存在する。

$$\text{if } y>A, \text{ then } |f_0(y')-f_0(y)|<\varepsilon f_0(y) \qquad \text{whenever } |y'-y|<h$$

（2）ある定数 $B$ と $M$ は，以下を満たす。

$$0<f_0(y')<Mf_0(y) \qquad \text{whenever } y'>y>B$$

（3）$k(\mu)=\sup_y\{f_0(y-\mu)/f_0(y)\}$ において，

5）このことは，逆に，尤度に正規分布を，事前分布がより裾の厚い分布を用いた際には，事前分布が棄却される。つまり事前分布が信頼性のあるデータに基づく尤度にそぐわないケースとみなしうる（West（1984））。

(a) $\int^{\infty} k(\mu)f(\mu)d\mu < \infty$

(b) $\int^{\infty} |\mu| k(\mu)f(\mu)d\mu < \infty$

仮定１と２は $y$ の密度関数が $y$ が大きくなるときにほぼ平らになることを，仮定３は $y$ の密度関数が事前分布よりも厚い裾を持つことを意味するものである。

これに対して Andrade and O'Hagan（2006）では，正則変動関数を用いて以下のような単純化された条件を提示している。

尤度 $y|\mu \overset{D}{\sim} f(y-\mu)$ と事前分布 $\mu \overset{D}{\sim} p(\mu)$ は以下の４つの条件を満たす。

(1) $f \in RV_{-\rho}$ （$\rho > 0$）
(2) $f > 0$ が実数全体で連続
(3) $f(y)$ が $y \geq B_2$ において減少し，$y \leq B_1$ において増加するような $B_1$ と $B_2 \geq B_1$ となる $B_2$ が存在する。さらに，$\dfrac{d\log f(y)}{dy}$ が存在し，かつ $y \geq B_2$ において増加する。
(4) $p \notin R$ or $p \in RV_{-c}$ （as $\mu \to \infty$）であり，$c > \rho + 1$

ここで，$f(x) \in RV_{-\rho}$ は，$f$ が $x \to \infty$ において指数 $\rho$ の正則変動関数であることを示す（Resnick（2006））。

### 3.5.2　Credence

Credence とは裾のウェイトの大きさを測り，裾の厚い分布のモデルによる外れ値を棄却するか否かを決定する条件をもたらす測度である。

事前分布の credence が尤度の credence よりも大きいとき，データが棄却され，事後平均は事前平均に漸近的に近づいていく。つまり credence は，矛盾する事後分布が生じうる場合において我々がどの程度二つの情報に信頼を持っているのかを測る速度である（Andrade and O'Hagan（2006））。

**Credence**

ある密度関数 $f$ が credence $c$ を持つ（$cred(f) = c$）とは，$f$ が以下を満たすことをいう。

全ての実数 $x$ について $g \leq (1+x^2)^{c/2} f(x) \leq G$（for $G \geq g > 0$）となる時，実数上の $f$ は credence $c$ を持つ。

自由度 $v$ の $t$ 分布において，データが∞に近づくとき，credence は $v+1$ となる。これに対して正規分布において credence は∞となる。

しかし O'Hagan（1990）で定義された credence は，定義より，非対称の分布にのみ適用が可能である。そこで Andrade and O'Hagan（2006）は，正則変動関数を用いて非対称な分布における credence として，RV－Credence を以下のように定義している。

**RV-Credence**

$f(x)\in RV_{-d}$（as $x\to\infty$）ならば，$f$ の right（finite）RV-Credence は $d$ となる（$d>0$）。
$f(x)\in RV_{-c}$（as $x\to-\infty$）ならば，$f$ の left（finite）RV-Credence は $c$ となる（$c>0$）。

## 4．シミュレーション

第3節では，事前分布と尤度が同じ厚みの裾をもつときに，「矛盾する事後分布」の問題が生じることについて見てきた。ここでは，尤度に事前分布よりも裾の厚い分布として $t$ 分布を用いて，データにおける外れ値モデルに対する頑強モデルのパフォーマンスについて検討をおこなっていく。

### 4.1　一変量モデル

まず，尤度に $t$ 分布を，平均 $\mu$ の事前分布に標準正規分布 $N(0, 1)$ を用いて，データが1つだけ存在するケースによる頑強性のパフォーマンスを見ていく。そこで，データを－10から10まで1ずつ動かし，事後平均を観察していく。

また，分布の厚みの違いを見るために，$t$ 分布には自由度3と自由度20の二種類の分布を用いて比較をおこなっていく。

$f(\mu)=N(0, 1)$：平均 $\mu$ の事前分布
$f(x|\mu)=t_d(\mu, 1)$：データ $x$ の尤度

シミュレーションには，WinBUGS ver. 1.4 を用いた。以下，いずれのシミュレーションにおいても一回のシミュレーションのサンプル数を10,000とし，最初の999個のサンプルを取り除いた。また，平均値に 0, 1, 10 の三種類の初期値を用いて，シミュレーションが収束しているか否かのチェックをおこなった。ここでは事後平均のみに興味があるため thinning は行っていない。

モデルの要約は以下のとおりである。

$$\begin{cases}\mu\overset{D}{\sim}N(0, 1)\\ x\overset{D}{\sim}t_d(\mu, 1)\end{cases}$$

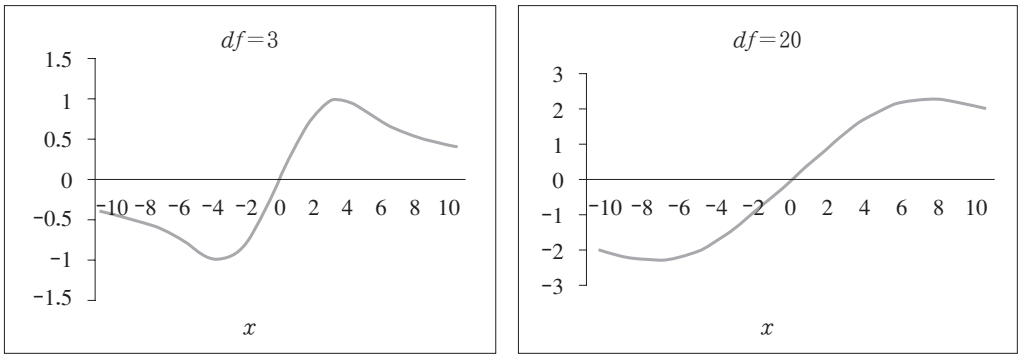
$df=3$　and　$df=20$

$x=-10, -9, -8, -7, -6, -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$

図6に上記のモデルによる事後平均を示している。3節でみたように，尤度にも分散1の正規分布を用いたとすると，事後分布の平均は $x/2$ となる。これに対し，自由度3お

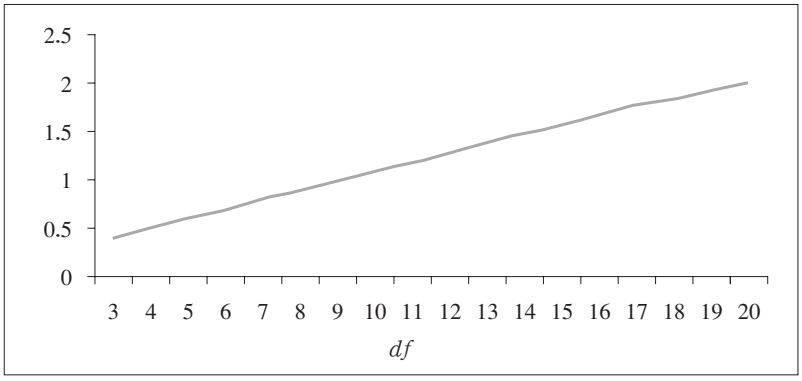
よび20のいずれの $t$ 分布を用いた場合にも，平均は，ある程度まで上昇したのち０に収束していく様子がうかがえる。

自由度３の $t$ 分布を用いたケースにおいては，事後分布の平均は，データの値が絶対値で３あたりに転換点を持つ。これに対して，自由度20の $t$ 分布を用いたケースにおいては，事後分布の平均は，データの値が絶対値で７あたりに転換点を持つ。標準正規分布を仮定したここでの事前分布のもとで，$|x|>2$ あるいは $|x|>3$ となるデータは，通常，外れ値とみなされる。それゆえ，事前分布に対して尤度の裾が厚い限り漸近的には外れ値は棄却されるが，頑強モデルとして $t$ 分布を用いる際には，十分に自由度の小さな分布を用いることが実用上望ましいといえよう。




［図６　位置パラメータの事後平均］

ここでは，自由度の大きさによってどの程度の頑強モデルが構築されているのかをみるために，上記のモデルにおいて，データ $x$ を10に固定し，自由度を３から20まで変化させたときの事後平均を観測した。




［図７　位置パラメータの事後平均］

それゆえ，ここでのモデルの要約は以下のとおりである。

$$\begin{cases}\mu \overset{D}{\sim} N(0,\ 1)\\ x \overset{D}{\sim} t_d(\mu,\ 1)\end{cases}$$

$df = 3,\ 4,\ 5,\ 6,\ 7,\ 8,\ 9,\ 10,\ 11,\ 12,\ 13,\ 14,\ 15,\ 16,\ 17,\ 18,\ 19,\ 20$

$x = 10$

図7にシミュレーション結果を示している。これを見ると，事後平均は自由度が大きくなるのに比例して大きくなり，自由度20のモデルにおいては，1.995まで上昇する。

## 4.2　定数項を持たない線形回帰モデル

ここでは，O'Hagan（1988）に従って，以下の（6）式のような定数項を持たない線形回帰モデルにおける裾の厚い分布を用いた頑強回帰モデルのパフォーマンスについて見ていく。このモデルでは，外れ値以外のデータ全体と外れ値の credence の大小によって，外れ値が自動的に検出され，棄却されるかが決まる。

$$Y_i = \beta x_i + \varepsilon_i \qquad (6)$$

ここでは，O'Hagan（1988）で提示された $x$ の最大値に外れ値を持つデータ（MODEL 1）と $x$ の中央値に外れ値を持つデータ（MODEL 2）から外れ値に対するモデルの頑強性についての比較をおこなっていく。O'Hagan（1988）とのモデルの比較のために，自由度20の $t$ 分布を用いる。それゆえ4つの標本グループの credence は84（$=4*(20+1)$）であり，外れ値の credence は21となる。それゆえ，この credence の差により外れ値は棄却される。

図8と9の右図はそれぞれのモデルにおいて，外れ値が存在しない場合の条件付き期待値よりも10だけプラス方向に標本が位置する外れ値（$y_{1,5} = 20,\ y_{2,3} = 13$）による通常最少二乗法による結果を示している。

MODEL 1 : $y_{1i} = \{1,\ 1,\ 4,\ 4,\ y_{1,5}\}$　　$x_{1i} = \{1,\ 2,\ 3,\ 4,\ 10\}$

MODEL 2 : $y_{2i} = \{1,\ 3,\ y_{2,3},\ 6,\ 9\}$　　$x_{2i} = \{1,\ 2,\ 3,\ 4,\ 10\}$

モデルの要約は以下のとおりである。傾き $\beta$ には，一様分布を仮定する。この時誤差項に正規分布を用いたモデルは，通常最少二乗法による推定値と同じ事後平均を持つ。裾の厚い分布モデルにおいては，自由度20の $t$ 分布を用いる。

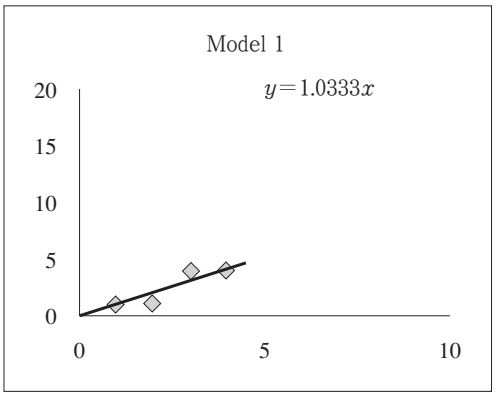
$$\begin{cases}\beta \overset{D}{\sim} U\\ \varepsilon_i \overset{D}{\sim} N(0,\ 1)\end{cases}$$

$$\begin{cases}\beta \overset{D}{\sim} U\\ \varepsilon_i \overset{D}{\sim} t_d(0,\ 1)\end{cases}$$

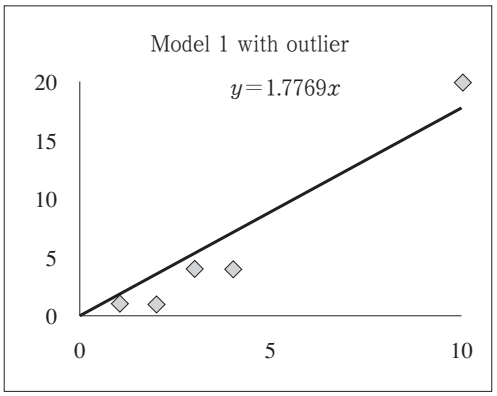
$df = 20$

$y_{1,5} = 10,\ 15,\ 20,\ 25,\ 30,\ 35,\ 40,\ 45,\ 50$

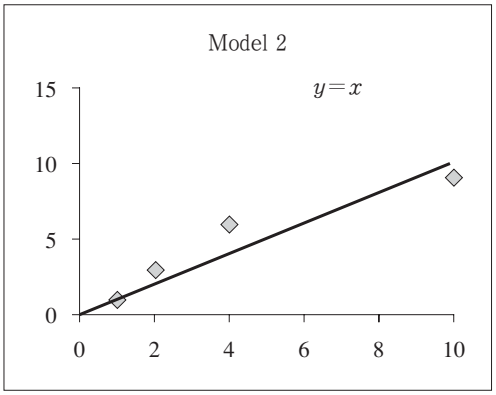
$y_{2,3} = 3,\ 10,\ 15,\ 20,\ 25,\ 30,\ 35,\ 40,\ 45,\ 50$



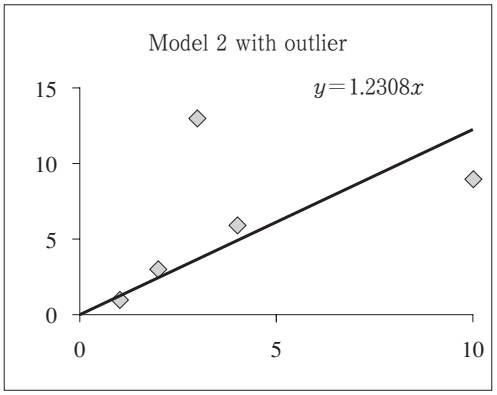

［図 8a］



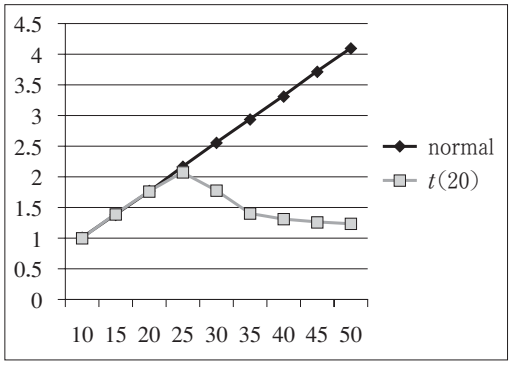

［図 8b］



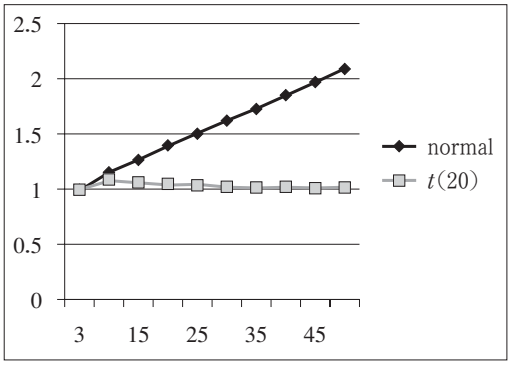

［図 9a］




［図 9b］




［図10　Model 1］




［図11　Model 2］

図10と図11に結果を示している。それぞれ，外れ値が存在しない場合の $x$ による条件付き期待値である３と10から50までのシミュレーション結果を示している。これをみると外れ値がデータの中心付近に存在する場合には，データの端に存在する場合に比べ，モデルは外れ値を棄却しやすいことが分かる。

### 4.3　定数項を持つ線形回帰モデル

ここでは，より一般的に定数項を持つ線形回帰モデルにおける裾の厚い分布を用いてRousseeuw and Leroy（2003）で考察がなされている型の外れ値による頑強性についての考察をおこなっていく。

4.2節において，定数項のない線形回帰モデルを用いた場合，外れ値が棄却されることが確認された。しかし，Peña et al.（2009）では，裾の厚い分布を用いたベイジアン線形回帰モデルは一般には頑強ではないことが示された。そこでここでは，モデルを定数項を持つモデルに拡張し，シミュレーション分析から，モデルが有効である範囲の特定化を行っていく。

ここでは，独立変数が一つの以下の（７）式の線形回帰モデルを用いる。

$$Y_i=\alpha+\beta x_i+\varepsilon_i \qquad (7)$$

ここでのシミュレーションには，傾きおよび相関がほぼ等しく，外れ値以外の標本数が２倍（credenceが２倍）となるデータセット（MODEL 3, MODEL 4）を用いている。

MODEL 3 : $y_{3i}=\{4,\ 6,\ 4,\ 8,\ y_{3,\ 5}\}$　　$x_{3i}=\{1,\ 2,\ 3,\ 4,\ x_{3,\ 5}\}$

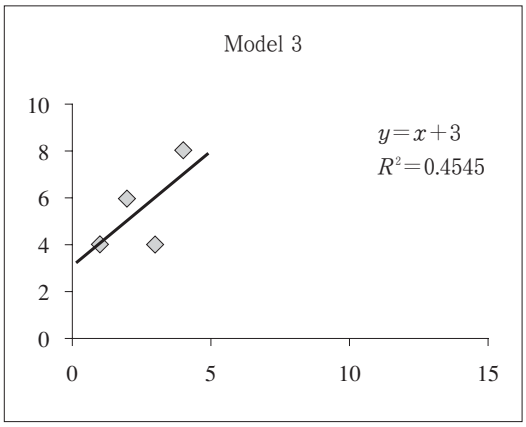
MODEL 4 : $y_{4i}=\{4,\ 6,\ 4,\ 7,\ 2,\ 3,\ 7,\ 5,\ y_{4,\ 9}\}$　　$x_{4i}=\{1,\ 2,\ 3,\ 4,\ 1,\ 2,\ 3,\ 4,\ x_{4,\ 9}\}$

モデルの要約は以下のとおりである。定数項 $\alpha$ および傾き $\beta$ の事前分布に一様分布を用いた。裾の厚い分布モデルには自由度４の $t$ 分布を用いた。この時誤差項に正規分布を用いたモデルは，定数項のないモデルと同様に，通常最少二乗法による推定値と同じ事後平均を持つ。

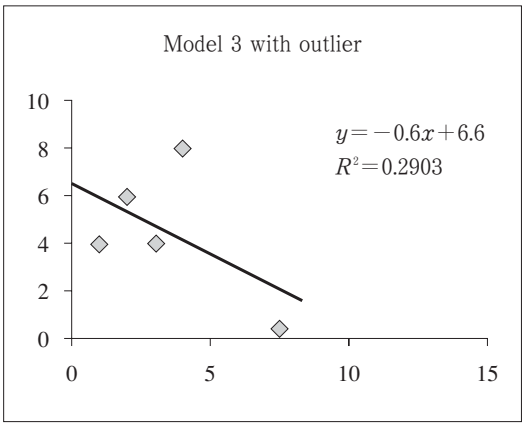
$$\begin{cases}\alpha \overset{D}{\sim} U\\ \beta \overset{D}{\sim} U\\ \varepsilon_i \overset{D}{\sim} N(0,\ 1)\end{cases}$$

$$\begin{cases}\alpha \overset{D}{\sim} U\\ \beta \overset{D}{\sim} U\\ \varepsilon_i \overset{D}{\sim} t_d(0,\ 1)\end{cases}$$

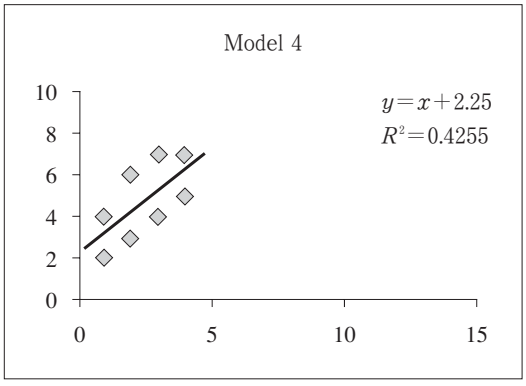
$$df=4$$



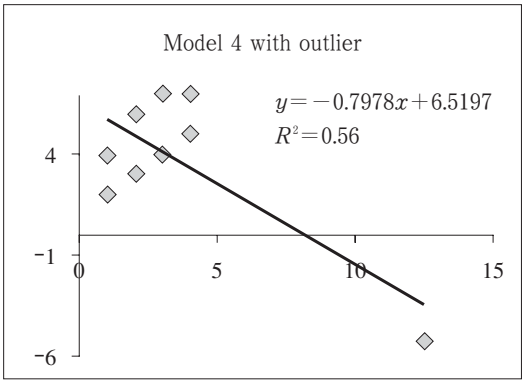

［図12a］




［図12b］




［図13a］




［図13b］

まず，定数項のないモデルと同様に独立変数の最大値（$x=4$）が，回帰直線に対して鈍角および鋭角の残差を持つモデルについてのシミュレーションを行う。表１に，鈍角をとる外れ値のデータを，表２に鋭角をとる外れ値のデータを示している。

図14に鈍角のデータを，図15に鋭角のデータを用いた事後平均を示している。図14より，O'Hagan（1988）で得られた結果と同様に，外れ値が棄却されていることが分かる。また，すぐに外れ値が棄却し始めていることが分かる。これに対して，図15の鋭角のケースでは，同じ回帰直線からの乖離幅においても鈍角に比べて，傾きの事後平均は外れ値から相対的に大きく影響を受けているがやはり外れ値は棄却されていくことが分かる。

外れ値とその他のデータのグループ化がはっきりとあらわれるケースとして，データ $x$ の平均値における外れ値がある。図16に見られるように，定数項が棄却されている様子が分かる。

以下では，データが棄却される範囲について考察を行っていくために，外れ値が存在しない場合の回帰直線に対してそれぞれの平均値（$\bar{x}$, $\bar{y}$）から直交する直線上の標本点を外

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $y_{3,5}$ | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 | 47 | 52 | 57 |
| $x_{3,5}$ | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| $y_{4,9}$ | 6.25 | 11.25 | 16.25 | 21.25 | 26.25 | 31.25 | 36.25 | 41.25 | 46.25 | 51.25 | 56.25 |
| $x_{4,9}$ | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |

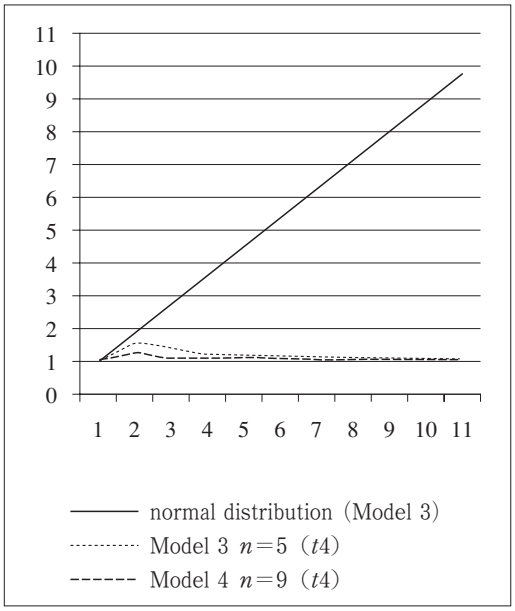
［表１　鈍角の外れ値］

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $y_{3,5}$ | 7 | 2 | −3 | −8 | −13 | −18 | −23 | −28 | −33 | −38 | −43 |
| $x_{3,5}$ | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| $y_{4,9}$ | 6.25 | 1.25 | −3.75 | −8.75 | −13.75 | −18.75 | −23.75 | −28.75 | −33.75 | −38.75 | −43.75 |
| $x_{4,9}$ | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |

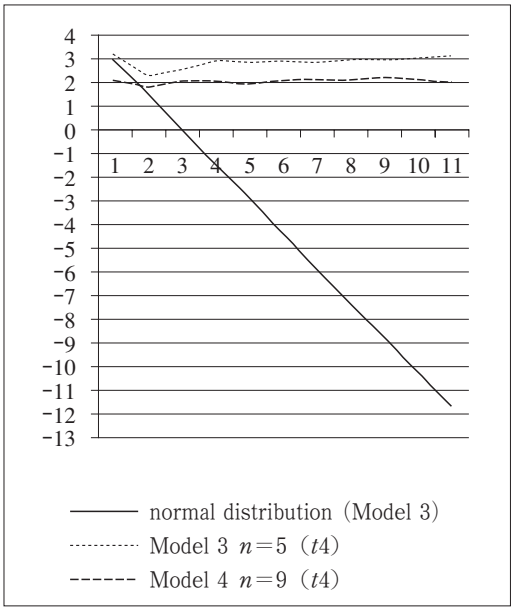
［表２　鋭角の外れ値］

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $y_{3,5}$ | 5.5 | 0.5 | −4.5 | −9.5 | −14.5 | −19.5 | −24.5 | −29.5 | −34.5 | −39.5 | −44.5 |
| $x_{3,5}$ | 2.5 | 7.5 | 12.5 | 17.5 | 22.5 | 27.5 | 32.5 | 37.5 | 42.5 | 47.5 | 52.5 |
| $y_{4,9}$ | 4.75 | −0.25 | −5.25 | −10.25 | −15.25 | −20.25 | −25.25 | −30.25 | −35.25 | −40.25 | −45.25 |
| $x_{4,9}$ | 2.5 | 7.5 | 12.5 | 17.5 | 22.5 | 27.5 | 32.5 | 37.5 | 42.5 | 47.5 | 52.5 |

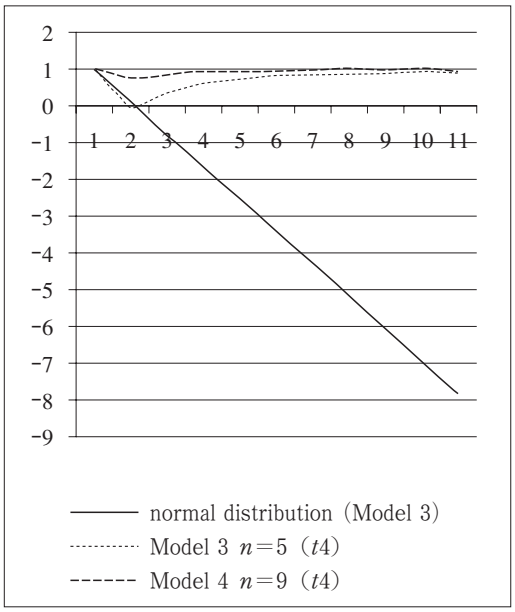
［表３　直交する直線上の外れ値］



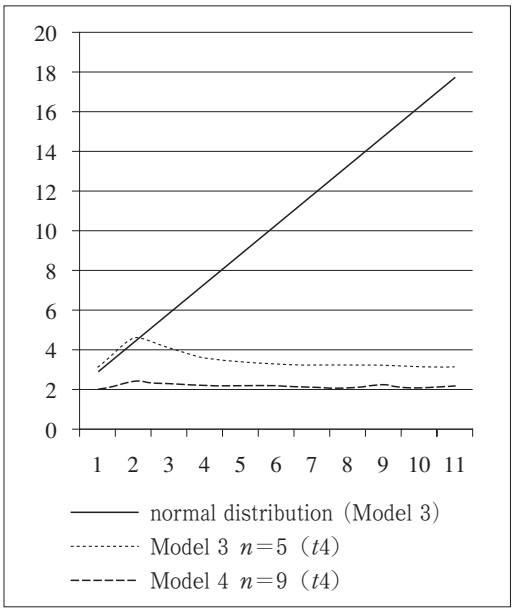

［図14a　傾きの事後平均］（鈍角の外れ値）



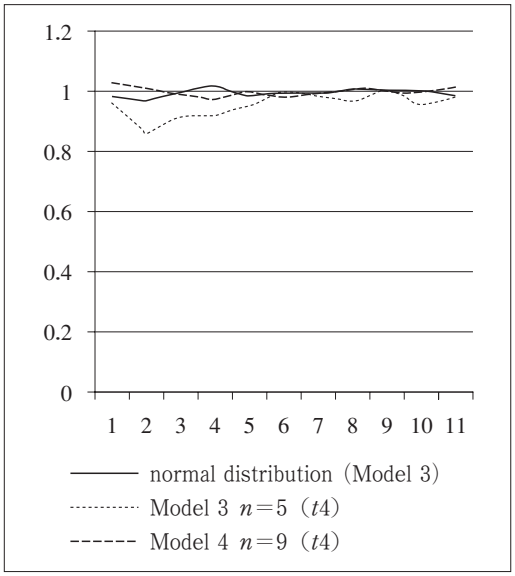

［図14b　定数項の事後平均］（鈍角の外れ値）



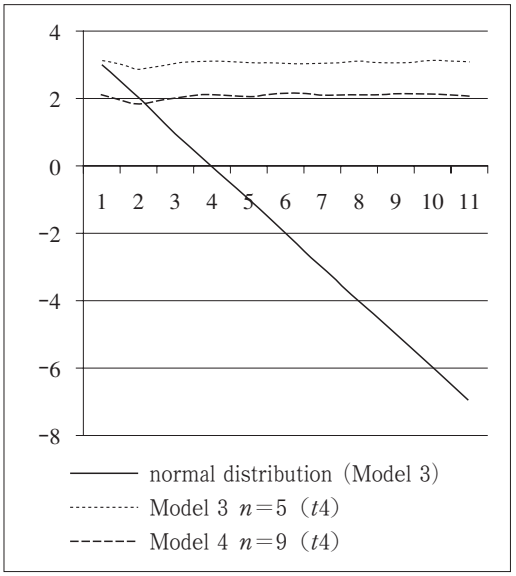

［図15a　傾きの事後平均］（鋭角の外れ値）




［図15b　定数項の事後平均］（鋭角の外れ値）

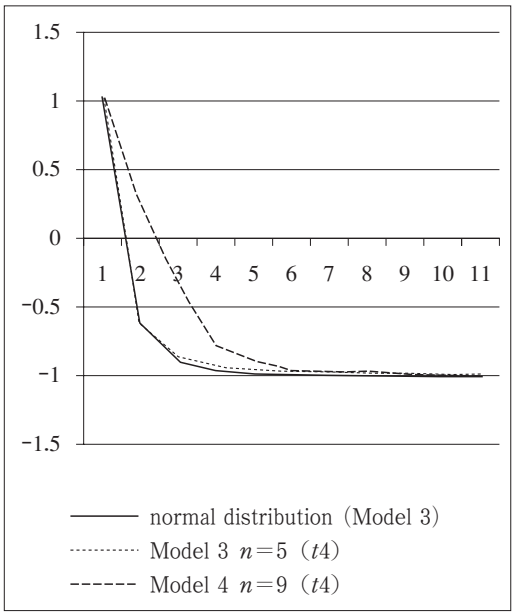



［図16a　傾きの事後平均］
（平均値における外れ値）

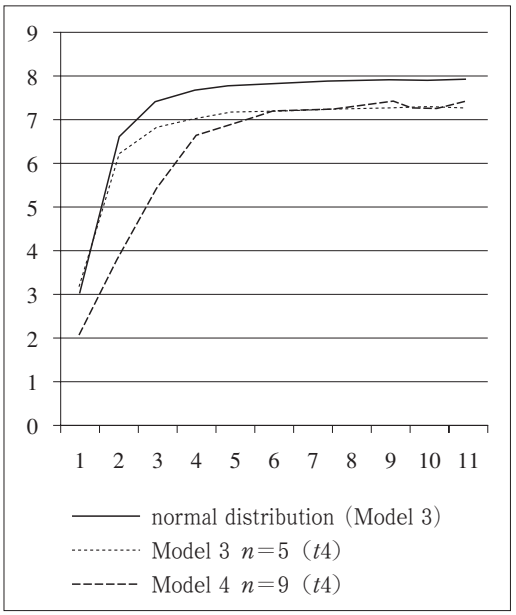



［図16b　定数項の事後平均］
（平均値における外れ値）

れ値に取っていく。用いた標本点を表３に示している。図17にシミュレーション結果を示している。これをみると外れ値が全く棄却されず，最少二乗法による推定値，−1 にいずれのモデルも収束していくことが分かる。つまり，回帰直線に直交する直線上付近の外れ



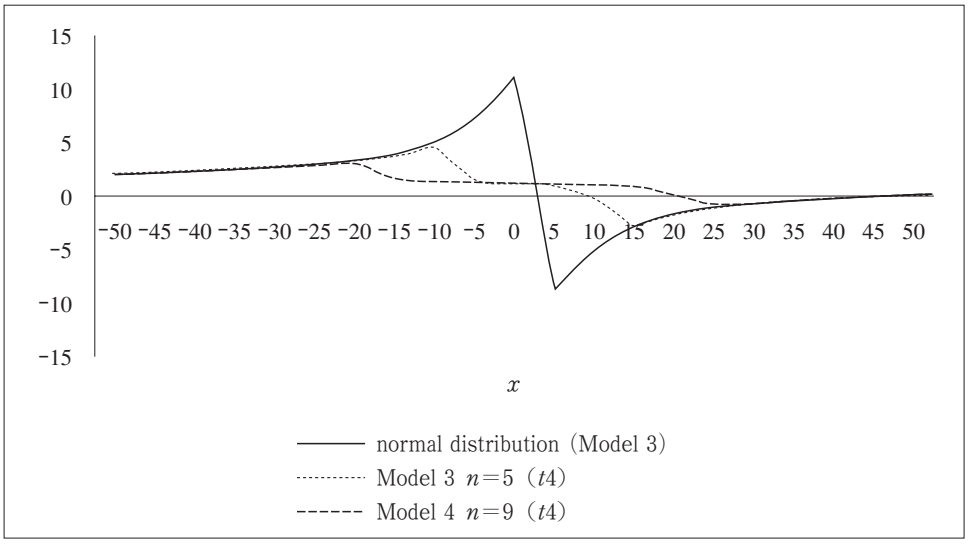

［図17a　傾きの事後平均］（直交する外れ値）




［図17b　定数項の事後平均］（直交する外れ値）

値は，モデルが外れ値として検出できない。

以上でみてきたように，裾の厚い分布を用いた頑強回帰モデルは，頑強である部分とない部分とを持つ。そこで，これらの範囲を示すために，MODEL 3，MODEL 4 のいずれについても，$x$＝－50 から50まで５刻みで外れ値の乖離幅が－50となるように外れ値を動かし（$y$＝－97 から３まで），外れ値が棄却されているか否かについてのシミュレーション




［図18　傾きの事後平均］

を行った。その結果，図18に見られるように，MODEL 3 では－5 から 5 まで，MODEL 4 においては－15から15においては，外れ値が棄却されるが，外れ値が外れ値以外のデータ集団がとる $x$ の範囲から大きく離れた場合には，外れ値が他のデータと比べて十分に回帰直線から離れないため棄却されない。この結果は，Peña et al.（2009）で得られたデータの $x$ の値のみを∞方向に近づけた際に，divergence の値が増加するとの結果と矛盾しない。

## ５．ま　と　め

本稿では，ベイズ統計学における外れ値問題を裾の厚い分布によって自動的に検出し棄却するモデルを線形回帰モデルに適応し，外れ値検出のパフォーマンスについてのシミュレーション分析を行った。その結果，裾の厚い分布を用いたベイジアン頑強回帰モデルは，外れ値が外れ値以外のデータの独立変数がとる範囲の近傍では頑強であるが，大きく離れた場合には頑強モデルとして有効ではないことが明らかになった。

しかしここで得られた結果は，一例であり，今後様々なケースにおける「裾の厚い分布」を用いた頑強モデルのパフォーマンスを更に検証していくことが求められる。

参考文献

Anderson, R.（2008）, *Modern Methods for Robust Regression,* SAGE publication

Andrade, J. A. A. and O'Hagan, A.（2006）, "Bayesian Robustness Modeling Using Regularly Varying Distribution," Bayesian Analysis, vol. 1, no. 1, pp. 169-188.

Barnett, V. and Lewis, T.（1994）, *Outlier in Statistical data,* Wiley

Bayarri, M. J. and Moraes, J.（2003）, "Bayesian measures of surprise for outlier detection," Journal of Statistical Planning and Inference, Vol. 111, no.（1-2）, pp. 3-22.

Ben-Gal, I.（2005）, "Outlier Detection", In: Maimon O. and Rockach L.（Eds.）Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers, luwer Academic Publishers

Box, G. E. P. and Tiao, and G. C.（1968）"A Bayesian approach to some outlier problems" Biometrika, vol. 55, pp. 119-129.

Dawid, A. P.（1973）, "Posterior expectations for large observations," Biometrika, vol. 60, pp. 664-667.

De Finetti, B.（1961）, "The Bayesian Approach to the Rejection of Outliers," In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, 199-210.

Fieller, N.（2005）, "Multivariate Outliers", *Encyclopedia of Statistics,* in Everitt, B. and Howell, D. C.（Eds.）, vol. 3, pp. 1379-1384.

Freeman, P. R.（1981）, and "On the number of outliers in data from a linear model," Bayesian Statistics, pp. 349-365.

Johnson, R. A. and Wichern, D. W.（2007）, *Applied Multivariate Statistical Analysis* Prentice-Hall

Kitagawa, G. and Akaike, H.（1982）, "A Quasi Bayesian Approach to Outlier Detection," Annual Institute Statistic and Mathematics, vol. 34, Part B, pp. 389-398.

O'Hagan, A. (1979), "On Outliers Rejection Phenomena in Bayes Inference," Journal of the Royal Statistical Society, Series B, no. 41, pp. 358-367.

O'Hagan, A. (1988), "Modelling with Heavy Tails." In Bayesian Statistics," vol. 3, pp. 345-359. Oxford: Claredon Press.

O'Hagan, A. (1990), "On Outliers and Credence for Location Parameter Inference," Journal of the American Statistical Association, vol. 85, pp. 172-176.

O'Hagan, A. and Forster, J. J. (2004), *Kendall's Advanced Theory of Statistics,* vol. 2B, *Bayesian Inference,* Arnold.

Peña, D., Zamar, R. and Yan, G. (2009), "Bayesian likelihood robustness in linear models," Journal of Statistical Planning Inference, vol. 139, pp. 2196-2207.

Resnick, S. I. (2006), *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling,* Springer

Rousseeuw, P. J. and Leroy, A. M. (2003), *Robust Regression and Outlier Detection,* Wiley.

West, M. (1984), "Outlier models and prior distribution in Bayesian linear regression," Journal of the Royal Statistical Society, Series B, vol. 46, pp. 431-439.

Zellner, A. (1975), "Bayesian Analysis of regression error terms," Journal of the American Statistical Association, vol. 70, pp. 138-144.