CIRJE-J-203

Lasso 分位点回帰の理論と損害保険への応用

東京大学大学院経済学研究科大学院生
加藤賢悟

東京大学大学院経済学研究科
国友直人

中央三井アセット信託銀行
増田智巳

2008 年 8 月

Lasso-Quantile Regression and its Application to a Non-life Insurance Problem

We summarize the recent developments on the statistical method of Lasso-Quantile Regression and we apply it to a Non-life Insurance problem. We discuss the asymptotic properties of the Quantile Regression estimator, the computational aspects related to the Linear Programming problem and the selection of Quantile regressors. We illustrate the practical aspects of measuring risk factors by using a Non-life insurance data.

Lasso分位点回帰の理論と損害保険への応用*

加藤賢悟† ・国友直人‡ ・増田智巳§

平成20年8月29日

概 要

分位点回帰 (Quantile Regression) の方法を損害保険におけるリスク要因のデータ解析へ応用した．分位点回帰における推定問題を大規模な線形計画問題として問題を定式化した上で数値的にLasso法を実現し，自動車保険の保険請求 (クレーム) 額の分析を行った．分位点に依存する説明変数の分析より中位クレーム・高額クレーム別のリスク要因を特定化した．

1 はじめに

統計的分析を利用する多くの応用分野と同様，近年では金融リスク管理や保険リスク管理で遭遇する様々な場面においても回帰分析 (regression analysis) が利用されている．例えば，ある変数（被説明変数と呼ぶ）のリスク要因分析として幾つかの説明変数に回帰し、求められた回帰係数をリスクに対する説明変数の感応度と解釈して統計的な解析が行われることが少なくない．統計学においてはよく知られているように，こうした統計的分析で用いられる線形回帰モデルは被説明変数の期待値（ある意味での確率分布の平均値）を定数を含む幾つかの説明変数の線形結合と誤差により「平均的に」説明する統計的モデルである．実際的な応用の場では、被説明変数のある種の平均的傾向を説明することに重要な意味がある場合もむろん少なくないが，例えば金融や保険におけるリスク分析，特に損害保険データによる保険リスク分析などではこうした回帰モデルの利用や分析結果の解釈に際しでは注意すべき基本的な問題もなお存在すると考えられる．例えば，本稿で実例として用いる損害保険の損失データは大きく右に歪んでいるのが一般的である．このような場合には算術平均は必ずしも「分布の平均」を代表するとは限らない, ことを統計学は教えている．また，近年になり目覚ましく発展をとげている金融リスク管理の問題では，損失額リスクを損失額分布 (loss distribution) として確率分布で表現するとき，少額の損失と多

*KKM08-8-20. 本稿は増田・国友 [16] の改訂稿である．本稿で利用したデータを提供してくれた損保ジャパンの足立尚人氏に感謝する．なお，本稿の内容に関する責任は著者のみにあり，損保ジャパン及び中央三井アセット信託銀行の見解を示すものではない.

†東京大学大学院経済学研究科 (大学院)

‡東京大学大学院経済学研究科

§中央三井アセット信託銀行

額の損失とではリスクの意味合いがかなり異なっていることが，議論されてきている．したがって，こうした金融リスク管理における応用問題では，通常の回帰分析のように被説明変数の期待値を説明するだけではなく，何らかの統計的方法により被説明変数の確率分布の特性を他の説明変数により説明することが必要であり，こうした既存の分析方法の問題点を克服する新たな統計的方法の開発も求められている．

本稿では被説明変数の分布の分位点を説明する回帰モデルを考察するが，そうした統計モデルは分位点回帰 (Quantile Regression) モデルと呼ばれている．この分位点回帰モデルは，統計学では古くから知られている最小絶対偏差法 (Least Absolute Deviation Method, 略して LAD 法) の発展型として，Koenker and Basssett [10] によって提唱され，発展している統計的モデル，統計的方法である．例えば，線形回帰モデルの下では誤差の絶対偏差最小化により得られる推定値は中央値回帰 (Median Regression) 問題の解に対応していることがよく知られている．このことは特に説明変数が定数のみであれば，被説明変数の中央値 (median) が定数の推定値として妥当であることから類推すると直観的にも分かり易いだろう．分位点回帰問題では中央値を含む任意の分位点において説明変数の効果が同一とは限らないことが一つの特徴である．本稿では分位点回帰の統計的理論についての考察と損害保険分野への応用を議論する．これまでに分位点回帰モデルの統計的な推定方法としては Koenker and Bassett [10] が提案した方法がよく知られているが, 本稿で説明するこの一般化絶対偏差最小化による推定方法は漸近的にはよい統計的性質を持つことが知られている．他方, 実データを用いるモデルの推定では最小二乗法のように解析的で明示的な表現を持つ解を導くことはできないが, このことがそれほど多くの応用例が報告されていない原因とも考えられる．しかしながら，本稿で説明するように, 分位点回帰問題は線形計画問題として定式化できるので，近年における計算技術の進歩と相まって，大規模なデータに対しても推定値を数値的に計算することは比較的容易となっている．このように，分位点回帰の統計的理論と解法は標準的な統計学の教科書で議論されている通常の線形回帰分析の議論とはかなり異なる側面があるので，ここで多くの応用家に理解しやすいように説明することにも意義があると考えられる．さらに，線形回帰問題における実際的に重要な問題として、説明変数の選択問題が古くから議論されているが, 分位点回帰問題においては説明変数のリストが分位点にも依存しうる，というさらなる問題も存在する．この問題に対処する為に，本稿では Tibshirani [23] が線形回帰モデル分析を念頭に提唱している, Lasso (least absolute shrinkage and selection operator) と呼ばれる統計的方法を分位点回帰問題に応用した Lasso 分位点回帰法をも考察する．

本稿の主たる目的は分位点回帰の統計的方法を比較的わかりやすく説明すると共に，損害保険分野においてリスク解析が必要とされる具体的なデータ例に対し “Lasso 分位点回帰” の理論を応用することである．近年では生命保険・年金・損害保険など保険の分野においてはリスクの多様化に対応し，様々な保険商品が登場してきている．そうした中で民間の保険会社は開発し販売している保険商品について，発生しうるクレームについて適切に支払保険金を見積もり，適切な保険料を定める必要がある．特に損害保険分野では損失分布は右に大きく歪んでいることが一般的であることより，分位点回帰を用いることで支払い保険金，つまりクレーム額のデータ分析を行うことが重要であろう．そして特にリス

ク要因として被保険者の特性に依存した支払い保険金の分布をより正確に推定することが必要となる，と考えられる．すなわち，保険リスクのデータ分析においては分位点回帰は広い応用が期待できるのである．本稿では具体的な実例として実際に観察された自動車保険のクレーム額についてのデータ分析を行い，その分析結果も報告する．我々のデータ分析により特に分位点回帰に Lasso 罰則を付与することにより，中位分位点や高分位点における分位点回帰の変数選択を行うことで，保険支払リスクの要因を詳細に分析することが可能であることが明らかとなった．

ここであらかじめ本稿の構成を説明しておこう．第 2 節では分位点回帰モデルと推定の漸近理論など必要な統計理論を説明する．次に第 3 節では Lasso 分位点回帰法について説明し，第 4 節ではその方法を用いて実際に自動車保険データについての実証分析を行った結果を報告し，第 5 節では結論を述べる．なお補論において、本稿で利用した数学的定理の証明、証明に必要となる補題、および実証分析で得られた幾つかの図を与えておく．

2 分位点回帰の理論

分位点回帰は Koenker and Bassett [10] により導入された統計モデルであり，被説明変数の条件付分位点を推定する方法と見なすことができる．通常の回帰分析においては二乗損失関数を用いて統計モデルの母数を推定し，条件付平均関数を求める最小二乗法を利用するのが一般的である．これに対して分位点回帰では一般化絶対偏差 (“check loss” と呼ばれる) 損失関数を用いて母数を推定すると，$(0,1)$ 上の任意の下側 τ 点に対応する条件付分位点を推定することができる．このことから条件付分布の中央値についての統計的推論が可能であるだけでなく，さらに分布の裾の挙動についても、そのリスク要因の統計的解析を行うことができる．分位点回帰法に関する様々な問題については Koenker [9] がかなり包括的に議論しているが，本稿では応用に興味を持つ読者を想定して，まず分位点回帰法における重要な論点に絞って説明しておこう．

2.1 分位点回帰モデル

統計的分析の対象であるリスクを含む変数 Y を確率変数と考える．そして, Y の変動を説明するリスク要因として幾つかの説明変数を考え、説明変数ベクトル $\mathbf{X} = (X_1, \ldots, X_p)'$ が与えられた時の確率変数 Y の条件付分布関数を $\mathrm{P}(Y \leq y|\mathbf{X}) = F_Y(y|\mathbf{X})$，条件付 τ 分位点を $Q_\tau(Y|\mathbf{X}) = \inf\{y|F_Y(y|\mathbf{X}) \geq \tau\}$ とする．このとき分位点回帰モデルは

$$
\begin{aligned}
Q_\tau(Y|\mathbf{X}) &= \alpha(\tau) + \beta_1(\tau)X_1 + \cdots + \beta_p(\tau)X_p \\
&= \alpha(\tau) + \mathbf{X}'\boldsymbol{\beta}(\tau)
\end{aligned} \tag{1}
$$

と表現される．ただし $\boldsymbol{\beta}(\tau) = (\beta_1(\tau), \ldots, \beta_p(\tau))'$ は定数項を表す母数 $\alpha(\tau)$ を除く未知母数ベクトルである．ここで確率変数 U を

$$
U = Y - \{\alpha(\tau) + \mathbf{X}'\boldsymbol{\beta}(\tau)\}
$$

により定義すれば，分位点回帰モデルは

$$Y = \alpha(\tau) + \mathbf{X}'\boldsymbol{\beta}(\tau) + U \tag{2}$$

と表現できる．この形式は統計的線形回帰モデルに類似しているが，右辺の母数の係数ベクトルが τ に依存し，誤差項 U の分布関数を $F_U(u)$ とすると

$$F_U(0) = \mathrm{P}\left(Y \leq \alpha(\tau) + \mathbf{X}'\boldsymbol{\beta}(\tau)\right) = \tau \tag{3}$$

となるので，若干の注意が必要である．

ここで被説明変数 Y と説明変数ベクトル $\mathbf{X}$ について互いに独立な n 個のデータの組が得られる状況を想定する．さらに $\mathbf{y} = (y_1, \ldots, y_n)'$ を被説明変数ベクトル $(n \times 1)$，$\mathbf{x}_i = (x_{1i}, \ldots, x_{pi})'$ $(i = 1, \ldots, n)$ を説明変数ベクトル $(p \times 1)$ としよう．本稿では議論の簡単化の為に分位点回帰モデルにおいて条件 $n \geq p+1$ が成立し，被説明変数の分布関数は Lebesgue 測度に関して絶対連続の場合のみを考察する[1]．この仮定の下では密度関数が存在するので分位点関数は分布関数の逆関数として一意的に定義でき，統計的分析がかなり簡単化されることになる．こうした仮定は損害保険などで実際に生じる多くの応用上ではそれほど大きな問題は生じない，と考えられる．

次に被説明変数 Y と説明変数ベクトル $\mathbf{X}$ の n 組のデータより母数ベクトル $((p+1) \times 1)$ $(\alpha(\tau), \boldsymbol{\beta}(\tau)')'$ を推定する問題を考えよう．この問題について，Koenker and Bassett [10] は損失関数として

$$L(u) = \rho_\tau(u) = (\tau - \mathbf{1}(u < 0))u \tag{4}$$

を用いることを提案している．(ここで記号 $\mathbf{1}(\omega)$ は ω が成立すれば 1, その他は 0 という指示関数とする.) n 組のデータより評価基準

$$\min_{\alpha,\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_\tau(y_i - \alpha - \mathbf{x}_i'\boldsymbol{\beta}) \tag{5}$$

を最小化する推定方法を考察する．この最小化問題の解を $(\widehat{\alpha}(\tau), \widehat{\boldsymbol{\beta}}(\tau)')'$ と表す．また説明をより容易にする為に，定数項を含む説明変数ベクトルのデータを $\mathbf{x}_i^* = (1, \mathbf{x}_i')'$ $(i = 1, \ldots, n)$ と表しておこう．

2.2 推定量の漸近的性質

分位点回帰モデルでは母係数ベクトルの推定が重要な統計的問題である．母係数ベクトルを $\boldsymbol{\delta}(\tau) = (\alpha(\tau), \boldsymbol{\beta}(\tau)')'$，推定量ベクトルを $\widehat{\boldsymbol{\delta}}(\tau) = (\widehat{\alpha}(\tau), \widehat{\boldsymbol{\beta}}(\tau)')'$ としておこう．分位点回帰推定量は標本数 n が大きいときには一致性 (consistency) と漸近正規性 (asymptotic normality) を持つことが知られている．推定量の漸近的性質の分析は評価関数が非線形の

[1]離散分布の場合にも議論を拡張することは可能であるが，例えば漸近理論はより複雑になる．

場合には一般に複雑になるが，分位点回帰問題の場合には次のようにすると推定量の漸近的性質を比較的容易に導くことができる．本稿では確率変数列 $(y_i, \mathbf{x}_i)$ $(i = 1, \cdots, n)$ について次のような標準的な正則条件[2] を用いる．

(A1) 確率変数列 $(y_i, \mathbf{x}_i)$ $(i = 1, \cdots, n)$ は互いに独立で同一分布 (i.i.d.) にしたがう．
(A2) $\mathbf{x}_i$ を所与とする u_i の条件付分布関数 $F_U(\cdot|\mathbf{x}_i)$ は ($\mathbf{x}_i$ に依存しない) 原点の近傍上で正の密度関数 $f_U(\cdot|\mathbf{x}_i)$ をもつ．さらにこの近傍上で，$s \mapsto f_U(s|\mathbf{x}_i)$ は ($\mathbf{x}_i$ について一様に) 連続となる．
(A3) 行列 $\mathbf{C} = \mathrm{E}[\mathbf{x}_i^* \mathbf{x}_i^{*'}]$ は正定値行列となる．
(A4) 行列 $\mathbf{D} = \mathrm{E}[f_U(0|\mathbf{x}_i)\mathbf{x}_i^* \mathbf{x}_i^{*'}]$ は正定値行列となる．

以上の仮定の下で次のような理論的な結果が成り立つ．証明は 6 節の補論に与えておくので参照されたい．

定理 1. 分位点回帰推定量 $\widehat{\boldsymbol{\delta}}(\tau)$ について $n \to \infty$ のとき次の性質が成り立つ．

(i) 条件 *(A1)*〜*(A3)* のもとで

$$\widehat{\boldsymbol{\delta}}(\tau) = \begin{bmatrix} \widehat{\alpha}(\tau) \\ \widehat{\boldsymbol{\beta}}(\tau) \end{bmatrix} \xrightarrow{p} \boldsymbol{\delta}(\tau) = \begin{bmatrix} \alpha(\tau) \\ \boldsymbol{\beta}(\tau) \end{bmatrix} \tag{6}$$

が成り立つ．

(ii) 条件 *(A1)*〜*(A4)* のもとで

$$\sqrt{n}\{\widehat{\boldsymbol{\delta}}(\tau) - \boldsymbol{\delta}(\tau)\} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \tau(1-\tau)\mathbf{D}^{-1}\mathbf{C}\mathbf{D}^{-1}) \tag{7}$$

が成立する．

なお，以上の説明では説明変数 $\mathbf{x}_i$ が確率的である場合を扱ったが，$\mathbf{x}_i$ が非確率的変数である場合には条件

(A1)' 確率変数列 u_i $(i = 1, \ldots, n)$ は互いに独立に同一分布 (i.i.d.) にしたがう．
(A2)' u_i の分布関数 F_U は原点の近傍で連続かつ正な密度関数 f_U を持つ．
(A3)' 正定値行列 $\mathbf{C}$ が存在し，$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \mathbf{x}_i^* \mathbf{x}_i^{*'} = \mathbf{C}$ となる．

のもとで，分位点回帰推定量が一致性と漸近正規性を持つことを示せる．ただし，この場合には漸近分布は

$$\mathcal{N}(\mathbf{0}, \tau(1-\tau)\{f_U(0)\}^{-2}\mathbf{C}^{-1}) \tag{8}$$

と表現される．

ところで, 6 節の補論では分位点回帰推定量の一致性・漸近正規性を，比較的簡明な方法で導出した．その証明方法は Knight [8] が LAD 推定量の漸近的性質を調べた時に用いた

[2] むろん通常の回帰分析での議論と同様にここで利用している正則条件 (A1)-(A4) を弱めたり，あるいは様々な方向の条件の下での議論に拡張できる．

議論を拡張したものである．補論で説明した方法はこれまでに知られている推定量の漸近正規性に関する証明方法よりもかなり一般的でありかつ簡明であると思われる[3]．

2.3 漸近共分散の推定

分位点回帰法においては，漸近分布の共分散行列に誤差の密度関数が表れるので実際に分析を行う際には漸近共分散を推定する必要が生じる．漸近共分散を直接推定する方法としてはブートストラップ法など統計的リサンプリング法の利用も考えられる．またカーネル密度推定を用いて密度関数を推定することで共分散を推定する方法（Powell [19]）も提唱されている．本稿ではより直観的にも分かり易い Hall and Sheather [6] の方法を紹介する．なお，ここでは簡単のため説明変数 $\mathbf{x}_i$ が非確率的である場合を扱う．

このとき，$f_U(0) = f_Y(F_Y^{-1}(\tau))$ であることに注意する (Y の分布関数および密度関数をそれぞれ $F_Y(\cdot)$，$f_Y(\cdot)$ で表す)．そこで

$$s(\tau) = \{f_Y(F_Y^{-1}(\tau))\}^{-1} \tag{9}$$

を推定することを考える．いま $F_Y(F_Y^{-1}(\tau)) = \tau$ であるから，

$$\frac{d}{d\tau} F_Y^{-1}(\tau) = s(\tau)$$

となる．この関係を上手く利用して Siddiqui [21] は被説明変数 Y の経験分布関数 $\widehat{F}_n$ を用いて $s(\tau)$ を

$$\widehat{s}_n(\tau) = \{\widehat{F}_n^{-1}(\tau + h_n) - \widehat{F}_n^{-1}(\tau - h_n)\}/2h_n \tag{10}$$

により推定することを提案している．こうした推定方法を漸近的に正当化するにはバンド幅 h_n を $h_n \to 0$ $(n \to \infty)$ とする必要がある．例えば Hall and Sheather [6] は幾つかの数学的仮定の下でエッジワース展開を評価して，信頼区間を構成する際の最適なバンド幅として

$$h_n = n^{-1/3} z_\alpha^{2/3} \left\{ 1.5 \frac{s(\tau)}{s''(\tau)} \right\}^{1/3} \tag{11}$$

を提案している．ここで α は有意水準，$\Phi(\cdot)$ は標準正規分布の分布関数，z_α は $\Phi(z_\alpha) = 1 - \alpha/2$ をみたす点とする．ここで $s(\tau)$ と $s''(\tau)$ については例えば経験分布関数を利用して推定することが可能である．$s(\tau)$ と $s''(\tau)$ に関しては一致推定量さえ構成できればよいことに注意する．

さらに Koenker [9] は正規分布を用いると $s(t)/s''(t)$ が位置・尺度変換について不変であることから，より簡便な方法として F として標準正規分布を用いることを提案している．以上の考察より本稿で報告する実証分析では，この漸近分散・共分散の推定方法を採用した．

[3]例えば絶対偏差最小化問題については Amemiya [1]4 節の説明が標準的であろう。特に評価関数が凸関数である場合には極小値推定量 (extremum estimator) の収束や漸近分布の導出についてより一般的な証明方法の展開も可能である．

2.4 線形計画問題としての分位点回帰法

分位点回帰問題の推定は線形計画問題として書き表すことができる．特に$\tau = 0.5$の場合，つまり絶対偏差最小化 (LAD) 回帰については古くから研究されている．例えばBarrodale and Roberts [2] はLAD回帰の性質を利用し，単体法 (simplex method) を用いて最適化問題を効率的に解く方法を提案している．彼らのアルゴリズムはその後，単体法を用いて分位点回帰問題を解く場合にも広く使われるようになっている．さらに線形計画法においては1980年代に内点法 (interior point method) が登場したことにより，大規模線形計画問題に対する計算速度も大幅に改善できるようになった[4]．分位点回帰モデルにおける推定問題は線形計画問題

$$\mathrm{P_{qr}} \left| \begin{array}{ll} \min & \tau \mathbf{1}_n' \mathbf{u} + (1-\tau)\mathbf{1}_n' \mathbf{v} \\ \text{subject to} & \mathbf{y} - \mathbf{1}_n \alpha - \mathbf{X}\boldsymbol{\beta} = \mathbf{u} - \mathbf{v}, \\ & \mathbf{u} \geq \mathbf{0}_n,\ \mathbf{v} \geq \mathbf{0}_n \end{array} \right.$$

として表現できる．ここで$\mathbf{0}_n$と$\mathbf{1}_n$はそれぞれ0と1をn個並べたベクトルであって，$\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_n]'$である．主問題では$\mathrm{P_{qr}}$の目的関数を$(\alpha, \boldsymbol{\beta}', \mathbf{u}', \mathbf{v}')'$という$2n+p+1$個の変数について最小化する．ここで，$\mathbf{X}^* = [\mathbf{1}_n, \mathbf{X}]$，　$\boldsymbol{\delta} = (\alpha, \boldsymbol{\beta}')'$とおく．問題$\mathrm{P_{qr}}$を単体法により解く場合には，初期実行可能基底解は

$$\begin{aligned} &\boldsymbol{\delta} = \mathbf{X}^*(h)^{-1}\mathbf{y}(h),\ h \in \mathcal{K}, \\ &\mathbf{u}(h) = \mathbf{v}(h) = \mathbf{0}_{p+1},\ \mathbf{u}(\bar{h}) = (\mathbf{y} - \mathbf{X}^*\boldsymbol{\delta}(h))_+,\ \mathbf{v}(\bar{h}) = (\mathbf{y} - \mathbf{X}^*\boldsymbol{\delta}(h))_-, \end{aligned}$$

ただし，

$$\begin{aligned} &\mathbf{X}^*(h)' = (\mathbf{x}_i^*)_{i\in h},\ \mathbf{y}(h) = (y_i)_{i\in h},\ h \subset \{1, \ldots, n\}, \\ &\mathcal{K} = \{h \subset \{1, \ldots, n\} \mid \mathrm{rank}\mathbf{X}^*(h) = p+1\},\ \bar{h} = \{1, \ldots, n\} \backslash h \end{aligned}$$

とすればよい．さらに主問題$\mathrm{P_{qr}}$の双対問題は

$$\mathrm{D1_{qr}} \left| \begin{array}{ll} \max & \mathbf{y}'\mathbf{d} \\ \text{subject to} & \mathbf{1}_n'\mathbf{d} = 0, \\ & \mathbf{X}'\mathbf{d} = \mathbf{0}_p, \\ & \mathbf{d} \in [\tau - 1, \tau]^n \end{array} \right.$$

で与えられる (Koenker [9] のTheorem 2.1を参照) [5]．ここで変数を定義し直し，$\mathbf{a} = \mathbf{d} + (1-\tau)\mathbf{1}_n$とすれば線形計画問題は

$$\mathrm{D2_{qr}} \left| \begin{array}{ll} \max & \mathbf{y}'\mathbf{a} \\ \text{subject to} & \mathbf{1}_n'\mathbf{a} = (1-\tau)n, \\ & \mathbf{X}'\mathbf{a} = (1-\tau)\mathbf{X}'\mathbf{1}_n, \\ & \mathbf{a} \in [0, 1]^n \end{array} \right.$$

[4] こうした問題については例えばPortnoy and Koenker [18] を参照されたい．

[5] 線形計画法における主問題と双対問題の定式化や双対定理 (duality theorem) については竹内 [22] 5章の説明が分かりやすい．

なる有界変数問題になる．初期値となる内点には $(1-\tau)\mathbf{1}_n$ を選んでやればよい[6]．主問題としての定式化では非負条件に加えて制約条件数は標本数 n と等しくなっている．したがって標本数 n が大きくなると[7]計算負荷量が増大する．他方，双対問題では制約条件数は説明変数 $p+1$ 程度であるので推定に必要な計算量は遙かに小さくなる．

分位点回帰問題を解くアルゴリズムについて単体法ベースの Barrodale-Roberts 法に対して，内点法ベースの Frisch-Newton 法が Portnoy and Koenker [18] で提案されている．分位点回帰問題では $(0,1)$ 上の任意の τ に対して線形計画問題を解く必要がある．この為に線形計画法において効率的計算方法として知られているパラメトリック線形計画法の利用も考えられる．ここで議論している線形計画計画問題ではパラメトリック線形計画法を用いて $(0,1)$ 上のすべての最適基底解を求めることが可能である．例えば Koenker and d'Orey [11] では，$(0,1)$ 上のすべての τ に対してパラメトリック線形計画法と分位点回帰問題の最適基底解についての性質 (Koenker [9] の定理 2.1) を利用し，解（最適基底解）を連続的に得るアルゴリズムを提案している．分位点回帰問題の解が連続的に得られる点ではこの方法が良いと思われるが，n が大きくなると解が変化する場合の数も多くなることが応用上の問題になってくる．

本稿で扱った問題を検討する過程では説明変数と誤差の両方について，すべて相関の無い標準正規乱数を用いてシミュレーション実験により数値計算の効率性を確かめてみた．例えば $n=3,000$，$p=8$ とした場合には解が変化する回数は約 6,000 回だった．また p は変えずに $n=10,000$ とした場合は約 20,000 回の解の変化が確認された．n がこの程度の大きさならば計算時間についてはあまり考慮する必要は無いが，例えば $n=500,000$ 程度になると解が変化する回数とともに計算時間もかなり大きくなると予想される．ここでこのような状況は損害保険会社が実際のデータで分析を行う際には考えられなくもない状況であり，こうしたときには計算は非効率となる．したがって，大きなデータの場合には実用上は τ について一定の幅をとりながら内点法により解を得る方法が良いと云えよう．

3 Lasso 分位点回帰

3.1 Lasso 法

Lasso (least absolute shrinkage and selection operator) 法は Tibshirani [23] により提案された統計的手法であり，もともとその開発の目的は，回帰分析における予測精度の改善と変数選択であった．Lasso はモデルの定数項以外のパラメータに絶対値 (L_1) ノルムの罰則をつけて最小二乗推定を適用する手法である．すなわち，Lasso 推定量 $(\widehat{\alpha}^{lasso}(\lambda), \widehat{\boldsymbol{\beta}}^{lasso}(\lambda)')'$

[6]内点法については小島・土屋・水野・矢部 [13] で分かりやすく説明されている．単体法が実行可能多面体の頂点を辿りながら最適解を探索することに対し，内点法では多面体の内部を通って最適解を探す．

[7]例えば 4 節で報告する実証分析で利用したデータは $n=6,113$ であるが，実務的な応用まで考慮すると n は遙かに大きくなる．

は最小化問題

$$\min_{\alpha,\boldsymbol{\beta}} \sum_{i=1}^{n}(y_i - \alpha - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \quad \lambda \geq 0 \tag{12}$$

の最適解として定義される．ここで λ はチューニング・パラメータと呼ばれている．(12) 式を λ 形式の Lasso と呼ぼう．λ 形式に対して t 形式の Lasso も定義しておこう．t 形式の Lasso 推定量 $(\widehat{\alpha}^{lasso}(t), \widehat{\boldsymbol{\beta}}^{lasso}(t)')'$ は次の最小化問題問題の最適解として定義される．

$$\begin{aligned} &\min_{\alpha,\boldsymbol{\beta}} \sum_{i=1}^{n}(y_i - \alpha - \mathbf{x}_i'\boldsymbol{\beta})^2 \\ &\text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t, \quad t \geq 0. \end{aligned} \tag{13}$$

λ 形式と t 形式の最小化問題は数学的には同値である．

Lasso 法と類似の統計的方法としては Ridge 回帰法が古くから知られている．Ridge 回帰法は罰則付き最小二乗法という意味では Lasso 法と類似の統計的方法と見なすことができる．ただし，Lasso 法は Ridge 法と異なり，パラメータの値を正確にゼロと推定することが可能である．すなわち, Lasso 法では推定と同時に変数選択も実行することが可能なのである．この点から Lass 法は最近になり注目されるようになってきたのである．ここで参考として Ridge 回帰問題は

$$\begin{aligned} &\min_{\alpha,\boldsymbol{\beta}} \sum_{i=1}^{n}(y_i - \alpha - \mathbf{x}_i'\beta)^2 \\ &\text{subject to } \sum_{j=1}^{p} \beta_j^2 \leq t, \quad t \geq 0 \end{aligned} \tag{14}$$

と定式化できる事に言及しておく．この Ridge 回帰は元々説明変数行列が線形独立でない場合に線形回帰分析を行う実用的な目的に為に提案されたことは興味深い．

3.2 L_1 罰則を加えた場合の分位点回帰

分位点回帰と L_1 罰則を組み合わせることにより，分位点回帰問題においても分位点に依存する変数選択が可能となる．ここで罰則が L_1 ノルムであることから，Lasso 分位点回帰もまた線形計画問題として定式化できることに注意する必要があろう．

L_1 罰則を加えた t 形式の分位点回帰問題は

$$\begin{aligned} &\min_{\alpha,\beta} \sum_{i=1}^{n} \rho_\tau(y_i - \alpha - \mathbf{x}_i'\boldsymbol{\beta}) \\ &\text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t, \quad t \geq 0 \end{aligned} \tag{15}$$

である．この (15) は線形計画問題

$$
\mathrm{P}_{\mathrm{lasso}}(t)\left|\begin{array}{ll}
\min & \tau\mathbf{1}_n'\mathbf{u}+(\mathbf{1}-\tau)\mathbf{1}_\mathbf{n}'\mathbf{v} \\
\text{subject to} & \mathbf{y}-\mathbf{1}_n(\alpha^+-\alpha^-)-\mathbf{X}(\boldsymbol{\beta}^+-\boldsymbol{\beta}^-)=\mathbf{u}-\mathbf{v}, \\
& \mathbf{1}_p'(\boldsymbol{\beta}^++\boldsymbol{\beta}^-)\le t, \\
& \alpha^+\ge 0,\ \alpha^-\ge 0, \\
& \boldsymbol{\beta}^+\ge \mathbf{0}_p,\ \boldsymbol{\beta}^-\ge \mathbf{0}_p, \\
& \mathbf{u}\ge \mathbf{0}_n,\ \mathbf{v}\ge \mathbf{0}_n
\end{array}\right.
$$

として表現できる..

Lasso 分位点回帰の場合には最終的にチューニング・パラメータ t を選ぶ必要があるので，すべての $t\ge 0$ に対して推定値を計算する必要がある．この問題については，例えば Kato [7] はパラメトリック単体法をベースに解のパスを効率的に計算するアルゴリズムを提案している．本稿 4 節では Kato [7] の方法をもとに Lasso 分位点回帰の数値計算を行った．なお前節でも述べたように．n が極端に大きい場合 (たとえば n が 100,000 を超えるような場合) にパラメトリック単体法を適用すると，ステップ数がかなり大きくなることが予想される．そうした場合には双対問題を考え，t について一定の幅を取りながら内点法を逐次適用する方法などが考えられる．ここでは n が $10,000\sim 20,000$ 程度であれば，Kato [7] の方法で数値的にも問題が生じることなく解を求めることも指摘しておく．いずれにしても，本稿で扱うデータ解析の範囲では数値的な問題は生じなかった．

3.3 チューニング・パラメータの選択規準

分位点回帰において用いられるモデル選択規準としては，SIC (Schwartz Information Criterion, Koenker et al. [12]) や GACV (Generalized Approximate Cross Validation, Yuan [24]) などが提案されている．また，SIC の公式においてペナルティ項における $\log n$ を 2 に変えることにより，形式的に AIC (Akaike Information Criterion) も定義できる．SIC, GACV, AIC はそれぞれ

$$
\mathrm{SIC}(t)=\log\left\{\frac{1}{n}\sum_{i=1}^n\rho_\tau(y_i-\mathbf{x}_i^{*'}\widehat{\boldsymbol{\delta}}(\tau,t))\right\}+\frac{\log n}{2n}df(t), \tag{16}
$$

$$
\mathrm{GACV}(t)=\frac{\sum_{i=1}^n\rho_\tau(y_i-\mathbf{x}_i^{*'}\widehat{\boldsymbol{\delta}}(\tau,t))}{n-df(t)}, \tag{17}
$$

$$
\mathrm{AIC}(t)=\log\left\{\frac{1}{n}\sum_{i=1}^n\rho_\tau(y_i-\mathbf{x}_i^{*'}\widehat{\boldsymbol{\delta}}(\tau,t))\right\}+\frac{1}{n}df(t) \tag{18}
$$

で与えられる．ここで $df(t)$ はモデルの自由度 (degrees of freedom)，または有効パラメータ数 (effective number of parameters) と呼ばれ，チューニング・パラメータ t によってコントロールされるモデルの複雑さを表す指標である．このモデルの自由度に関しては Efron [3]

が議論しているが，Yuan [24] および Li and Zhu [15] はモデル $\widehat{f}_{\tau,t}(\mathbf{x}_i) = \widehat{\alpha}(\tau,t) + \mathbf{x}_i'\widehat{\beta}(\tau,t)$ に対して

$$\widehat{df}(t) = \sum_{i=1}^{n} \frac{\partial \widehat{f}_{\tau,t}(\mathbf{x}_i)}{\partial y_i} \tag{19}$$

を自由度 $df(t)$ の推定値として用いることを提案している．また，Li and Zhu [15] はいくつかの条件の下で，

$$\sum_{i=1}^{n} \frac{\partial \widehat{f}_{\tau,t}(\mathbf{x}_i)}{\partial y_i} = \#\{j \mid \widehat{\delta}_j(\tau,t) \neq 0\} \tag{20}$$

が成り立つことを示している．ここで，右辺はモデル $\widehat{f}_{\tau,t}$ に含まれるパラメータの個数を表しているので，直観的にもモデルの複雑さを表す指標として適切なものであると見なせよう．そこで (20) の右辺の値を自由度 $df(t)$ の推定値として採用しよう．

ここで例えば，SIC に基づいてチューニング・パラメータを選択することを考えよう．すなわち

$$\hat{t} = \arg\min_{t \geq 0} \mathrm{SIC}(t) \tag{21}$$

となる $\hat{t}$ を用い，$\widehat{\boldsymbol{\delta}}(\tau,\hat{t})$ を最終的な推定値とすること，が考えられる．なおデータより $\hat{t}$ を選ぶときには，すべての $t \geq 0$ の中で $\mathrm{SIC}(t)$ を最小にするものを選ぶ必要はないが，このことは次のような議論から正当化されよう．

ここで t を増やすことは制約条件が緩くなることを意味するから $\sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^{*\prime}\widehat{\boldsymbol{\delta}}(\tau,t))$ は t に関して単調減少となる．そこで $\widehat{df}(t)$ の t の関数としての挙動を考えると，正則化経路 (パス)$\{\widehat{\boldsymbol{\delta}}(\tau,t), t \geq 0\}$ は t に関して区分的に線形，すなわち $0 = t^0 < t^1 < \cdots < t^K = \sum_{j=1}^{p} |\widehat{\beta}_j(\tau)|$ が存在し，$\{\widehat{\boldsymbol{\delta}}(\tau,t), t \in [t^{k-1}, t^k]\}$ は $\widehat{\boldsymbol{\delta}}(\tau,t^{k-1})$ と $\widehat{\boldsymbol{\delta}}(\tau,t^k)$ を線形に結んだ経路に等しい．ただし，$\widehat{\beta}_j(\tau)$ は無制約のもとでの分位点回帰推定値である．このとき，(20) で与えられる $\widehat{df}(t)$ の値は $t \in (t^{k-1}, t^k)$ のとき一定であり，$t \uparrow t^k$ のとき一定か，1 減るかのいずれかである．いずれの場合にも $\widehat{df}(t) \geq \widehat{df}(t^k),\ (t \in (t^{k-1}, t^k])$ より $\mathrm{SIC}(t) \geq \mathrm{SIC}(t^k),\ t \in (t^{k-1}, t^k]$ が成り立つことがわかる．したがって

$$\hat{k} = \arg\min_{k=0,\ldots,K} \mathrm{SIC}(t^k) \tag{22}$$

なる $\hat{k}$ を選べば，$\hat{t} = t^{\hat{k}}$ となることがわかる (GAVC，AIC に関しても手順は同一である)．

なお，ここで説明したチューニング・パラメータの選択規準はいずれも直観的な議論に基づいて導出されたものである．よく知られている統計的方法としてクロスバリデーション法の応用も考えられるが，ここではデータ数が大きくなると計算量の観点から実用的であるとは言い難い．こうした理由から本稿では上述した規準に基づいてチューニング・パラメータの選択を行った．

4 自動車保険の分析

4.1 自動車保険について

自動車保険では自動車の所有・使用・管理に関連して生ずる損害を填補する目的の保険であり，損害保険会社において中心的な保険である．また自動車保険の担保種目については対人賠償保険・自損事故保険・搭乗者傷害保険・対物賠償保険・車両保険・無保険車傷害保険の 6 種類があるが，本稿ではその中でも総額の支払い保険金が比較的大きい対物賠償保険について，分位点回帰を用いたデータ分析の結果を報告する．ここで対物賠償保険とは,“ 自動車の所有，使用，管理に起因して他人の財物を滅失，破損または汚損すること (対物事故) により，法律上の賠償責任を負担することによる損害を填補する保険 ”を意味する．損害保険料率算出機構がまとめた自動車保険統計[8]によると，2006 年度における対物賠償保険の支払保険金は約 6,800 億円であり，対人賠償保険・搭乗者傷害保険・車両保険を抑えて最大であった．強制保険である自賠責保険では対物事故については支払は行われないので，対物賠償保険に関するリスクは損害保険会社にとって重要な分析対象であろう．

4.2 データについて

本稿のデータ分析では $n = 6,113$ 個の自動車保険対物事故について，ある期間を無作為に選び，各曜日について同じ数だけ無作為抽出したデータで, 被説明変数としてクレーム額，事故や個人に関する質的データを説明変数として用いた．クレーム額が 1 万円以下のデータについては省いたが，これはリスク分析とは直接的に関係のない要因による支払いと見なした為である．利用可能な説明変数としては，運転者の性別や年齢，事故が起こった曜日と時間帯，車種などで，すべてダミー変数である．ここで被説明変数・説明変数についての情報をまとめておこう．

- Pay: クレーム額（保険請求額）.
- Male: 男性ならば 1，その他は 0.
- Car3: 用途車種が自家用普通乗用車 (3 ナンバー) ならば 1，その他は 0.
- Holiday: 事故が起こった曜日が土日なら 1，その他は 0.
- Midnight: 事故が起こった時間帯が 23 時か 5 時までの間ならば 1，その他は 0.
- Age20s: 運転者年齢が 20 歳代ならば 1，その他は 0.
- Age40s: 運転者年齢が 40 歳代ならば 1，その他は 0.
- Gold: ゴールド免許所持者ならば 1，その他は 0.
- IUnlim: 対物保険金額 (事故が発生した場合に損害保険会社が支払う保険金の限度額) が無制限ならば 1，その他は 0.

[8]損害保険料率算出機構の HP（http://www.nliro.or.jp/）上で公開されている．

表 1: クレーム額 (Pay) の記述統計量

最小値	標準偏差	1st-Qu	中央値	平均	3rd-Qu	最大値	歪度	尖度
0.101	2.544	0.696	1.361	2.088	2.527	43.58	43.98	4.795

- IUplim500: 対物保険金額が 500 万円以下ならば 1，その他は 0.

表 1 と図 1 はそれぞれクレーム額 (Pay) についての記述統計量，ヒストグラムを表し単位は 10 万円である．図より明らかなようにクレーム額の分布は非対称であり，右裾が重い分布になっている．また各説明変数が与えられた場合（値が 1 となるとき）の個体数とクレーム額の平均，中央値，最大値，標準偏差を表 2 にまとめておく．例えば 4 行目より深夜ダミー変数の影響は平均的にクレーム額が他よりかなり大きくなっていることや，8 行目より対物保険金額は無制限の人の割合が多いことがわかる．

ここで説明変数ベクトル $\mathbf{x}$ を用いた分位点回帰の計測モデルを

$$
\begin{aligned}
(23)\quad Q_\tau(\mathrm{Pay}|\mathbf{x}) &= \beta_0(\tau) + \beta_1(\tau)\mathrm{Male} + \beta_2(\tau)\mathrm{Car3} + \beta_3(\tau)\mathrm{Holiday} + \beta_4(\tau)\mathrm{Midnight} \\
&\quad + \beta_5(\tau)\mathrm{Age20s} + \beta_6(\tau)\mathrm{Age40s} + \beta_7(\tau)\mathrm{Gold} + \beta_8(\tau)\mathrm{IUnlim} + \beta_9(\tau)\mathrm{IUplim500}.
\end{aligned}
$$

と表しておこう．我々は統計モデル (23) を用いて主に分位点 $\tau \in (0,1)$ 全体の傾向，中位点，上側分位点 $\tau \in [0.9, 0.995)$ などについてデータ分析を行った．

ところで，Lasso 分位点回帰を用いる時には，制約条件と係数の整合性の観点より説明変数・被説明変数に関して基準化を行うことが適切と考えられる．そこで我々のデータ分析では

$$
(24)\qquad \sum_{i=1}^{n} y_i = 0,\ \sum_{i=1}^{n} x_{ij} = 0,\ \sum_{i=1}^{n} x_{ij}^2 = 1,\ (j = 1, \ldots, p)
$$

と云う変換により説明変数の基準化を行った．本稿では主に $\tau = 0.5$，0.95 について Lasso 分位点回帰を行ったデータ分析の結果を説明するが，それぞれの分位点 τ に対し図 12・図 13 が対応する．この図はチューニング・パラメータ t を横軸にとったときの推定値 $\widehat{\boldsymbol{\beta}}^{lasso}(\tau, t)$ のグラフである．それぞれの折れ線に対応する変数のインデックスを図の t 軸近くに印しておいた．

4.3 実証結果

分位点回帰の推定結果を見ておこう．表 3 は最小二乗法による回帰分析の結果と $\tau = 0.05, 0.1, 0.5, 0.9, 0.95$ についての分位点回帰の結果であり，数字の上側がパラメータの推定値でその下の括弧内の数字が標準誤差である．図 2 －図 11 では，横軸が τ で縦軸が推定値である．図の中の横に横断する灰色の直線が最小二乗推定量でその上下にある同色の点線が 90%信頼区間を表しており，黒い点を直線で結んだ線は分位点回帰推定値，周りの灰色

図 1: クレーム額のヒストグラム

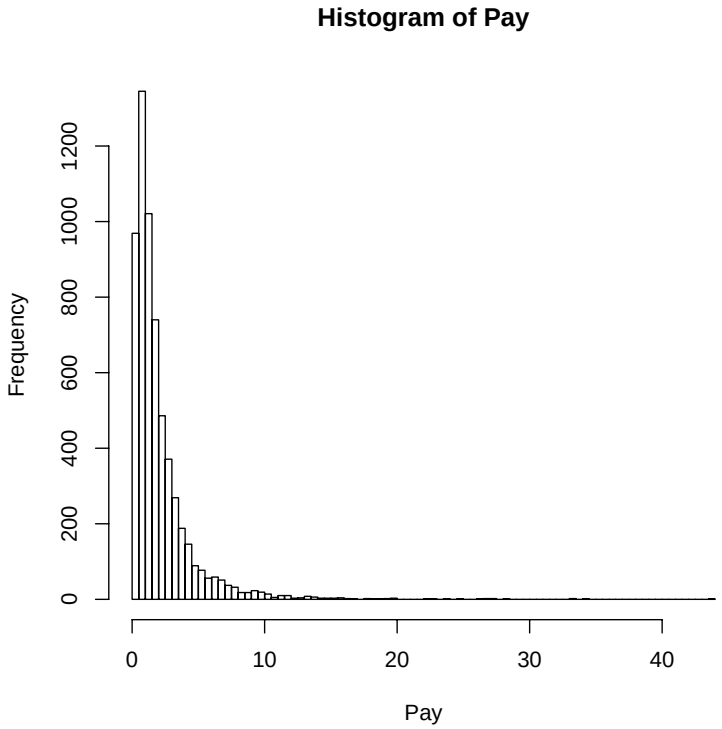

表 2: 説明変数を与えたときのクレーム額の記述統計量

変数名	件数（固体数）	平均	中央値	最大値	標準偏差
Male	3252	2.250	1.449	43.58	2.764
Car3	1714	2.343	1.479	43.58	3.059
Holiday	1758	2.258	1.468	28.35	2.580
Midnight	328	3.216	1.872	34.40	4.194
Age20s	1282	2.257	1.546	27.13	2.567
Age40s	1063	1.876	1.247	18.42	2.017
Gold	2568	1.958	1.286	33.24	2.335
IUnlim	5367	2.071	1.352	43.58	2.526
IUplim500	86	2.477	1.514	19.84	3.224

の領域がその 90%信頼区間である．これらの推定結果を踏まえて通常の線形回帰分析についての結果を見ると，運転者が男性 (Male)，車種が 3 ナンバー車 (Car3)，土日 (Holiday)，深夜 (Midnight)，40 歳代 (Age40s)，ゴールド免許所持 (Gold) については最小二乗推定量は 90%以上の有意性を示しているが，他の説明変数については有意と云えないことがわかる．以降「有意性を示す」とは 90%有意のことを意味する．

まず性別について観察してみよう．一般に女性に比べ男性の方がクレーム額が大きくなると云われている．我々の分析結果を見ても，男性については全体的に推定値が正で有意になっていることから，男性効果が確認できた．最小二乗法による推定でも有意，上側の分位点回帰モデルでも有意であり，他の説明変数に比べてリスクとしては大きくなる．

車種についてであるが，自家用普通乗用車については分位点回帰推定量が男性の場合と同じようなパスを描いている．上側分位点においては男性の場合程顕著な有意性を示さないが，それでも十分にリスク要因といえるだろう．

平日と土日の効果については，運転の目的がかなり異なると考えられる．分析結果からは土日についてはクレーム額の分布が全体的に右にずれていることが確認できる．上側でもほぼすべての分位点について推定値は正で有意となっている．したがって土日に起こる事故が平日のそれよりクレーム額は大きくなることが予想される．

時間帯効果としての深夜効果を見てみよう．深夜効果についての最小二乗推定量はすべての説明変数の中で最も有意な値になっている．分位点回帰でも深夜の事故のクレーム額の分布は，そうでない場合の事故に比べて全体が大きく右にずれている．したがってこの説明変数は最も大きなリスク要因だと考えられる．

年齢について 20 歳代と 40 歳代を見てみよう．20 歳代はクレーム額が大きくなる傾向がある．最小二乗法では有意性を示さなかったが，分位点回帰の結果を見れば 20 歳代のクレーム額分布は下側から中央，そして $\tau = 0.8$ 付近にかけて正で有意となっていることが分かった．しかし，上側分位点をみると有意性は示さないものの負の値をとっている．このことから，そのクレーム額分布は平均的に右よりだが，右裾は厚くないといえるであろう．次に 40 歳代を見てみよう．30 歳以上の運転者についてはしばしば保険金額が安くなる傾向がある．我々の分析結果を見てみると最小二乗法では 40 歳代という要因は負で有意性を示している．分位点回帰によれば $\tau = 0.8$ 付近で少し有意性が確かめられるものの，特別にリスクが小さくなるとは云い難い．

ゴールド免許所持者に関しては τ について全体的に推定値が負に有意になっている．上位分位点においても観察されているので，ゴールド免許所持者についてのクレーム額はやはり小さくなると云えるであろう．

対物支払保険については，対物支払保険金額を無制限にしている人は事故に対するリスクを大きく見積もっているのではと考えられる．最小二乗法では有意性が確かめられなかったが，分位点回帰においては上側，特に $\tau \in (0.9, 0.95)$ あたりで推定値が負で有意になっている．逆に対物保険金額を 500 万円としている人については上側分位点の $\tau \in (0.94, 0.985)$ 付近で正で有意となっている．つまり他の説明変数を固定した場合の対物保険金額を 500 万円としている人のクレーム額の分布は，そうでない人にくらべ右裾が重くなっていることが推測される．すなわち事故についてのリスクを小さく見積もる人は，高額の支払を要

求しているのではないかと推測できよう.

以上で説明したように分位点回帰によるデータ解析を用いるとこれまでの線形回帰分析のみを用いた場合に比べ，リスク要因についてのより詳しい結果が得られることが分かる.

次に Lasso 分位点回帰により説明変数の選択を行った簡単な結果を紹介しておこう[9]．まず $\tau = 0.95$（図 12）について推定結果を見ておこう．t を 0 から大きくしていくと，深夜の推定値が最初にゼロから非ゼロとなる．すなわち，リスク要因として最初に深夜効果が現れると言えよう．徐々に t を大きくしていくと，次に男性がリスク要因として検出され，さらに t を大きくしていくと各推定値の絶対値はより大きくなる．その後，対物保険金額が無制限，休日，自家用車普通自動車，ゴールド免許所持者の順で推定値がゼロでなくなっている．それぞれの推定値の符号は負，正，正，負となっており，特に対物保険金額が無制限とゴールド免許所持者の被保険者はリスクが小さくなると云えるであろう．年齢効果については図を見る限り，そのリスク要因としての重要性はそれほど大きくない．また対物保険金額が 500 万円以下の被保険者についてはゼロでなくなる最初の t が他に比べ大きいことをみれば，有意か否かについてのより慎重な議論が必要であろう．以上の議論から判断すれば，特に深夜変数と男性変数はリスク要因として重要であると云えるであろう.

続いて中央値 ($\tau = 0.5$) についてのデータ解析の結果（図 13) について簡単なコメントを加えておこう．最初にリスク要因として検出されるのは，上側分位点の場合と異なり 20 歳代であった．次に男性，深夜，休日の推定値が正で検出されている．すなわち，上側の高分位点に比べて，結果が大分違ってくるのがわかる．特に 20 歳代が上側の分位点にくらべ中央値において，変数の重要性が増していると云えるであろう．また，中央値付近では対物保険金額についての変数は重要性が小さいように思われる.

4.4 モデル選択の結果

損害保険データに対して Lasso 分位点回帰を適用し，モデル選択規準 SIC，GACV，AIC の値をそれぞれ計算し，データ分析の結果を以下の図 14〜図 17 にまとめておいた．$\tau = 0.95$ のケースでは，各選択規準の最小値近傍におけるグラフが見にくいので，大きい t に対して拡大したグラフを掲載しておいた．ここで選択された分位点回帰モデルを表 4 と表 5 にまとめておいた．$\tau = 0.5$ のケースでは GACV と AIC はともに右端の t において最小値をとっている．また，$\tau = 0.95$ のケースではすべての規準の下で同じモデルが選択された.

我々の分析により SIC 選択された分位点回帰モデルは無制約モデルにおいて有意でない係数を数値的にゼロとして再推定した結果と見なすことができる。AIC と GACV による結果は無制約分位点回帰モデルの結果に一致している。なお、これらモデル選択基準により得られた分位点回帰モデルの差はかなり小さく、モデル選択基準にはあまり依存しない分位点回帰モデルの推定結果と見なすこともできよう.

[9]図 12・図 13 では数値は t を動かしたときの各変数の係数推定値の変化を示している．$t = 0$ の時はすべての推定値は 0 となり，t の値が大きくなるにつれ制約が緩くなって推定値が 0 から離れていく．右端における推定値は制約がない場合の推定値に一致している.

表 3: 分位点回帰の結果

Covariates	0.05	0.10	0.50	0.90	0.95	LS
(Intercept)	0.231 (0.026)	0.312 (0.034)	1.195 (0.071)	4.207 (0.280)	6.076 (0.446)	1.907 (0.116)
Male	0.015 (0.015)	0.017 (0.020)	0.167 (0.040)	0.757 (0.160)	1.116 (0.256)	0.271 (0.066)
Car3	0.026 (0.016)	0.053 (0.022)	0.181 (0.044)	0.564 (0.176)	0.632 (0.280)	0.287 (0.072)
Holiday	0.035 (0.016)	0.081 (0.021)	0.150 (0.044)	0.451 (0.172)	0.797 (0.275)	0.202 (0.071)
Midnight	0.045 (0.032)	0.097 (0.043)	0.388 (0.088)	2.893 (0.350)	3.579 (0.558)	1.076 (0.145)
Age20s	0.025 (0.019)	0.039 (0.025)	0.179 (0.052)	−0.220 (0.206)	−0.365 (0.328)	0.038 (0.085)
Age40s	−0.013 (0.020)	−0.031 (0.026)	−0.078 (0.054)	−0.277 (0.213)	−0.057 (0.339)	− 0.197 (0.088)
Gold	0.004 (0.015)	0.010 (0.020)	−0.089 (0.042)	−0.267 (0.165)	−0.664 (0.262)	− 0.150 (0.068)
IUnlim	−0.022 (0.023)	−0.001 (0.031)	−0.022 (0.063)	−0.544 (0.251)	−0.722 (0.400)	− 0.095 (0.104)
IUplim500	0.000 (0.065)	0.001 (0.086)	0.127 (0.176)	0.106 (0.698)	2.804 (1.114)	0.384 (0.289)

表 4: 選択されたモデル ($\tau = 0.5$)

	SIC	GACV & AIC
Selected t	27.571	37.685
(Intercept)	-0.72438	-0.71568
Male	5.9725	6.4818
Car3	4.6336	6.3775
Holiday	3.2380	5.2478
Midnight	5.2293	6.7440
Age20s	5.1353	5.6664
Age40s	-1.1156	-2.3105
Gold	-2.2425	-3.4222
IUplim	0	-0.55356
IUplim500	0	0.83435

表 5: 選択されたモデル ($\tau = 0.95$)

	SIC & GACV & AIC
Selected t	229.25
(Intercept)	4.2630
Male	43.062
Car3	20.341
Holiday	26.318
Midnight	60.743
Age20s	-7.3406
Age40s	0
Gold	-25.716
IUplim	-19.890
IUplim500	25.842

5 おわりに

本稿では分位点回帰と Lasso 分位点回帰の統計的理論を説明すると共に，実際に自動車対物賠償保険のクレーム額に対して応用した結果を報告した．被説明変数であるクレーム額の分位点に応じたリスク要因、特に高分位点のリスク要因について興味ある結果が得られたと言う意味では，通常の線形回帰モデルよりも分位点回帰モデルの方が保険リスク分析に適しているという結論が得られた．

損害保険の分野においては例えば, 近年ではリスク細分型保険といわれる保険契約も登場している．例えばリスク細分型自動車保険では運転者年齢や地域，走行距離，目的などによりリスク要因を細かくして保険料を定めていると思われる．こうしたリスク要因のクレーム頻度との関係の分析を行うために，分位点回帰モデルは有用な統計的分析法を提供しているのではないかと考えられる．

ところで分位点回帰は本稿で議論した自動車保険をはじめとする損害保険に限らず，生命保険や第三分野保険などの保険分野，あるいはより広い金融分野で利用可能である．例えば近年では金融機関におけるリスク管理においても標準的なリスク指標として VaR(Value-at-Risk) がある．VaR は金融資産の収益率分布の下側分位点であり，分位点回帰を用いた一つの応用例が Engle and Manganelli [4] によって報告されている．こうしたデータ解析例からは，金融機関におけるリスク管理問題などでの分位点回帰法の今後の有用性が期待されよう．

最後になるが，本稿で取りあげた分位点回帰や Lasso 分位点回帰については更に検討すべき様々な理論的問題や計算上の問題があることを指摘しておく．例えば理論面では有限標本において漸近的議論がどれほど有効であるかはまだよく分かっていない．モデル選択基準についても通常の回帰分析や時系列分析などでは AIC など予測の基準を巡る議論が活

発であるが、分位点回帰における予測の意味などを検討する必要があろう。さらに、データ数や変数の数が極めて大きい場合の説明変数の選択なども重要な検討課題であろう.

参考文献

[1] Amemiya, T. (1985), *Advanced Econometrics*, Blackwell, New York.

[2] Barrodale, I. and Roberts, F. (1973), "An improved algorithm for discrete l_1 linear approximation," *SIAM Journal of Numerical Analysis*, **10**, 839-848.

[3] Efron, B. (2004), "The estimation of prediction error: covariance penalties and cross validation," *Journal of the American Statistical Association*, **99**, 619-632.

[4] Engle, R. and Manganelli, S. (2004), "CAViaR: Conditional autoregressive value at risk by regression quantiles," *Journal of Business and Economic Statistics*, **22**, 367-381.

[5] 福島雅夫 (2001), "非線形最適化の基礎," 朝倉書店，東京.

[6] Hall, P. and Sheather, S. (1988), "On the distribution of a studentized quantile," *Journal of the Royal Statistical Society, Series B*, **50**, 381-391.

[7] Kato, K. (2008), "Solving ℓ_1 regularization problems with piecewise linear losses," Preprint.

[8] Knight, K. (1998), "Limiting distributions for L_1 regression estimators under general conditions," *Annals of Statistics*, **26**, 755-770.

[9] Koenker, R. (2005), *Quantile Regression*, Cambridge University Press.

[10] Koenker, R. and Bassett, G. (1978), "Regression quantiles," *Econometrica*, **46**, 33-50.

[11] Koenker, R.and d'Orey, V. (1987), "Computing regression quantiles," *Applied Statistics*, **36**, 383-393.

[12] Koenker, R., Ng, P., and Portnoy, S. (1994), "Quantile Smoothing Splines," *Biometrika*, **81**, 673-680.

[13] 小島政和・土屋隆・水野眞治・矢部博 (2001), "内点法," 朝倉書店，東京.

[14] 今野浩 (1987), "線形計画法," 日科技連，東京.

[15] Li, Y. and Zhu, J. (2008), "L_1-norm quantile regression," *Journal of Computational and Graphical Statistics*, **17**, 1-23.

[16] 増田智巳・国友直人 (2008), “Lasso 分位点回帰の理論と損害保険への応用,” 東京大学経済学研究科 CIRJE, Research Report R-7,1-23.

[17] Pollard, D. (1991), ”Asymptotics for least absolute deviation regression estimators,” *Econometric Theory*, **7**, 186-199.

[18] Portnoy, S. and Koenker, R. (1997), ”The Gaussian hare and the Laplacian tortoise: computability of squared-error vs absolute error estimators,” *Statistical Science*, **12**, 279-300.

[19] Powell, J. L. (1991), ”Estimation of monotonic regression models under quantile restrictions,” In W. Barnett, J. Powell and G. Tauchen (Ed.), *Nonparametric and Semiparametric Models in Econometrics*, Cambridge University Press, Cambridge.

[20] Rockafellar, R.T. (1970), *Convex Analysis*, Princeton University Press, Princeton.

[21] Siddiqui, M. (1960), ”Distribution of quantiles from a bivariate population,” *Journal of Research of the National Bureau of Standards*, **64**, 145-150.

[22] 竹内啓 (1966), “線形数学,” 培風館，東京.

[23] Tibshirani, R. (1996), ”Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society, Series B*, **58**, 267-288.

[24] Yuan, M. (2006), ”GACV for Quantile Smoothing Splines,” *Computational Statistics and Data Analysis*, **5**, 813-829.

A 補論

A.1 定理 1 の証明

まず条件 (A1)〜(A3) のもとで一致性を示そう．係数ベクトルを $\boldsymbol{\phi} = \boldsymbol{\delta} - \boldsymbol{\delta}(\tau)$ と変換すると，評価関数の差は

$$\frac{1}{n}\sum_{i=1}^{n}\{\rho_\tau(y_i - \mathbf{x}_i^{*'}\boldsymbol{\delta}) - \rho_\tau(u_i)\} = \frac{1}{n}\sum_{i=1}^{n}\{\rho_\tau(u_i - \mathbf{x}_i^{*'}\boldsymbol{\phi}) - \rho_\tau(u_i)\} =: S_n(\boldsymbol{\phi}) \tag{25}$$

と表現されるので，$\widehat{\boldsymbol{\delta}}(\tau) - \boldsymbol{\delta}(\tau)$ は $S_n(\boldsymbol{\phi})$ の最小化点となる．次に $S_n(\boldsymbol{\phi})$ がある関数 $S(\boldsymbol{\phi})$ に確率収束し，収束先 $S(\boldsymbol{\phi})$ は $\boldsymbol{\phi} = \mathbf{0}$ で一意な最小化点を持つことを示す．ここで Knight の等式 (Knight [8])

$$\rho_\tau(u - v) - \rho_\tau(u) = -v\{\tau - \mathbf{1}(u < 0)\} + \int_0^v \{\mathbf{1}(u \le s) - \mathbf{1}(u \le 0)\}ds \tag{26}$$

を利用して,$S_n(\boldsymbol{\phi})$ を

$$\begin{aligned} S_n(\boldsymbol{\phi}) &= -\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^{*'}\boldsymbol{\phi}\{\tau - \mathbf{1}(u_i < 0)\} + \frac{1}{n}\sum_{i=1}^{n}\int_0^{\mathbf{x}_i^{*'}\boldsymbol{\phi}}\{\mathbf{1}(u_i \le s) - \mathbf{1}(u_i \le 0)\}ds \\ &=: S_{1n}(\boldsymbol{\phi}) + S_{2n}(\boldsymbol{\phi}) \end{aligned}$$

と展開する．ただし $\mathcal{E}[\tau - \mathbf{1}(u_i \le 0)|\mathbf{x}_i] = 0$ であり，条件 (A3) より各項の分散が有限となるので，大数の法則 (law of large numbers) より

$$S_{1n}(\boldsymbol{\phi}) \xrightarrow{p} 0 \tag{27}$$

となる．他方，$S_{2n}(\boldsymbol{\phi})$ については

$$\left[\int_0^{\mathbf{x}_i^{*'}\phi}\{\mathbf{1}(u_i \le s) - \mathbf{1}(u_i \le 0)\}ds\right]^2 \le 4(\mathbf{x}_i^{*'}\phi)^2$$

かつ $\mathcal{E}[(\mathbf{x}_i^{*'}\phi)^2] < \infty$ であるので，大数の法則より

$$\begin{aligned} S_{2n}(\boldsymbol{\phi}) &\xrightarrow{p} \mathcal{E}\left[\int_0^{\mathbf{x}_i^{*'}\boldsymbol{\phi}}\{\mathbf{1}(u_i \le s) - \mathbf{1}(u_i \le 0)\}ds\right] \\ &= \mathcal{E}\left[\int_0^{\mathbf{x}_i^{*'}\boldsymbol{\phi}}\{F_U(s|\mathbf{x}_i) - F_U(0|\mathbf{x}_i)\}ds\right] \\ &=: S(\boldsymbol{\phi}) \end{aligned} \tag{28}$$

となる．したがって，(27) と (28) より

$$S_n(\boldsymbol{\phi}) \xrightarrow{p} S(\boldsymbol{\phi}) \tag{29}$$

を得る．さらに条件 (A2) より，$\mathbf{x}_i^{*'}\boldsymbol{\phi} \ne 0$ なる $\boldsymbol{\phi}$ に対して，

$$\int_0^{\mathbf{x}_i^{*'}\boldsymbol{\phi}}\{F_U(s|\mathbf{x}_i) - F_U(0|\mathbf{x}_i)\}ds > 0 \tag{30}$$

となる．また条件 (A3) より，a.s.(almost surely) で $\mathbf{x}_i^{*'}\boldsymbol{\phi} = 0$ なる $\boldsymbol{\phi}$ は $\boldsymbol{\phi} = \mathbf{0}$ のみであるから，$\boldsymbol{\phi} = \mathbf{0}$ は $S(\boldsymbol{\phi})$ の一意な最小化点となる．

次に $S_n(\boldsymbol{\phi})$ の最小化点 $\boldsymbol{\phi} = \widehat{\boldsymbol{\delta}}(\tau) - \boldsymbol{\delta}(\tau)$ が $S(\boldsymbol{\phi})$ の一意な最小化点 $\boldsymbol{\phi} = \mathbf{0}$ に確率収束することを示す．$\epsilon > 0$ を任意に固定すると，$S_n(\boldsymbol{\phi})$ は凸関数であるから，$\|\mathbf{h}\| = 1$ なる $\mathbf{h} \in \mathbb{R}^{p+1}$ と $l > \epsilon$ に対して

$$\left(1 - \frac{\epsilon}{l}\right)S_n(\mathbf{0}) + \frac{\epsilon}{l}S_n(l\mathbf{h}) \ge S_n(\epsilon\mathbf{h})$$

となる．そこで $\Delta_n(\boldsymbol{\phi}) = S_n(\boldsymbol{\phi}) - S(\boldsymbol{\phi})$ とおけば，上の不等式から

$$\begin{aligned} \frac{\epsilon}{l}\{S_n(l\mathbf{h}) - S_n(\mathbf{0})\} &\ge S_n(\epsilon\mathbf{h}) - S_n(\mathbf{0}) \\ &= \{S(\epsilon\mathbf{h}) - S(\mathbf{0})\} - \{\Delta_n(\epsilon\mathbf{h}) - \Delta_n(\mathbf{0})\} \end{aligned}$$

が得られる．したがって $\|\mathbf{h}\| = 1$ なる $\mathbf{h} \in \mathbb{R}^{p+1}$ と $l > \epsilon$ に対し

$$\frac{\epsilon}{l}\{S_n(l\mathbf{h}) - S_n(\mathbf{0})\} \geq \eta - 2\Delta_n \tag{31}$$

が成り立つ．ただし，ここで $\eta = \inf_{\|\mathbf{h}\|=1} |S(\epsilon\mathbf{h}) - S(\mathbf{0})|$，$\Delta_n = \sup_{\|\boldsymbol{\phi}\| \leq \epsilon} |\Delta_n(\boldsymbol{\phi})|$ である．また $\boldsymbol{\phi} = \mathbf{0}$ は $S(\boldsymbol{\phi})$ の一意な最小化点であるので $\eta > 0$ となることに注意する．いま $\boldsymbol{\phi} = \widehat{\boldsymbol{\delta}}(\tau) - \boldsymbol{\delta}(\tau)$ は $S_n(\boldsymbol{\phi})$ の最小化点であるから，$\|\widehat{\boldsymbol{\delta}}(\tau) - \boldsymbol{\delta}(\tau)\| > \epsilon$ なら (31) 式の右辺は負となる．すなわち，

$$\left\{\|\widehat{\boldsymbol{\delta}}(\tau) - \boldsymbol{\delta}(\tau)\| > \epsilon\right\} \subset \{\Delta_n > \eta/2\}$$

である．この包含関係から

$$\mathrm{P}\left(\|\widehat{\boldsymbol{\delta}}(\tau) - \boldsymbol{\delta}(\tau)\| > \epsilon\right) \leq \mathrm{P}(\Delta_n > \eta/2) \tag{32}$$

が得られる．したがって (29) と補題 1(6.2 節を参照) を用いると $\Delta_n \xrightarrow{p} 0$ となるので，(32) の右辺は 0 に収束する．いま $\epsilon > 0$ は任意であったから

$$\widehat{\boldsymbol{\delta}}(\tau) \xrightarrow{p} \boldsymbol{\delta}(\tau)$$

が得られた．

次に条件 (A1)〜(A4) のもとで推定量の漸近正規性を示そう．これまでの議論より基準化された推定量 $\sqrt{n}\{\widehat{\boldsymbol{\delta}}(\tau) - \boldsymbol{\delta}(\tau)\}$ は (局所的) 評価関数

$$Z_n(\boldsymbol{\phi}) = \sum_{i=1}^{n}\{\rho_\tau(u_i - \mathbf{x}_i^{*'}\boldsymbol{\phi}/\sqrt{n}) - \rho_\tau(u_i)\} \tag{33}$$

の最小化点であることに注意し，$Z_n(\boldsymbol{\phi})$ の漸近的な挙動を調べる．再び Knight の等式 (26) を用いると $Z_n(\boldsymbol{\phi})$ は

$$\begin{aligned} Z_n(\boldsymbol{\phi}) &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i^{*'}\boldsymbol{\phi}\{\tau - \mathbf{1}(u_i < 0)\} + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathbf{x}_i^{*'}\boldsymbol{\phi})\int_0^1\{\mathbf{1}(u_i \leq \mathbf{x}_i^{*'}\boldsymbol{\phi}s/\sqrt{n}) - \mathbf{1}(u_i \leq 0)\}ds \\ &=: Z_{1n}(\boldsymbol{\phi}) + Z_{2n}(\boldsymbol{\phi}) \end{aligned}$$

と展開できる．ここで第 2 項 $Z_{2n}(\boldsymbol{\phi})$ をさらに

$$\begin{aligned} Z_{2n}(\boldsymbol{\phi}) &= \mathcal{E}[Z_{2n}(\boldsymbol{\phi})|\mathbf{x}_1, \ldots, \mathbf{x}_n] + \{Z_{2n}(\boldsymbol{\phi}) - \mathcal{E}[Z_{2n}(\boldsymbol{\phi})|\mathbf{x}_1, \ldots, \mathbf{x}_n]\} \\ &=: Z_{2n}^{(1)}(\boldsymbol{\phi}) + Z_{2n}^{(2)}(\boldsymbol{\phi}) \end{aligned}$$

と分解する．まず

$$Z_{2n}^{(1)}(\boldsymbol{\phi}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^{*'}\boldsymbol{\phi})\int_0^1\sqrt{n}\{F_U(\mathbf{x}_i^{*'}\boldsymbol{\phi}s/\sqrt{n}|\mathbf{x}_i) - F_U(0|\mathbf{x}_i)\}ds$$

の漸近的な挙動を評価する．いま

$$\frac{d}{dt}F_U(\mathbf{x}_i^{*'}\boldsymbol{\phi}st/\sqrt{n}|\mathbf{x}_i) = \frac{\mathbf{x}_i^{*'}\boldsymbol{\phi}s}{\sqrt{n}}f_U(\mathbf{x}_i^{*'}\boldsymbol{\phi}st/\sqrt{n}|\mathbf{x}_i)$$

であるから，

$$\sqrt{n}\{F_U(\mathbf{x}_i^{*'}\boldsymbol{\phi}s/\sqrt{n}|\mathbf{x}_i) - F_U(0|\mathbf{x}_i)\} = \frac{\mathbf{x}_i^{*'}\boldsymbol{\phi}s}{\sqrt{n}}\int_0^1 f_U(\mathbf{x}_i^{*'}\boldsymbol{\phi}st/\sqrt{n}|\mathbf{x}_i)dt \tag{34}$$

と表せる．したがって，

$$\begin{aligned} & Z_{21n}(\boldsymbol{\phi}) - \frac{1}{2n}\sum_{i=1}^n(\mathbf{x}_i^{*'}\boldsymbol{\phi})^2 f_U(0|\mathbf{x}_i) \\ = \ & \frac{1}{n}\sum_{i=1}^n(\mathbf{x}_i^{*'}\boldsymbol{\phi})^2\int_0^1\int_0^1 s\{f_U(\mathbf{x}_i^{*'}\boldsymbol{\phi}st/\sqrt{n}|\mathbf{x}_i) - f_U(0|\mathbf{x}_i)\}dtds \end{aligned}$$

が成り立つ．ここで条件 (A3) から $\max_{1\le i\le n}\|\mathbf{x}_i\| = o_p(\sqrt{n})$ となることに注意すると，条件 (A2) より

$$\max_{1\le i\le n}\left|\int_0^1\int_0^1 s\{f_U(\mathbf{x}_i^{*'}\boldsymbol{\phi}st/\sqrt{n}|\mathbf{x}_i) - f_U(0|\mathbf{x}_i)\}dtds\right| \xrightarrow{p} 0$$

を得る．したがって条件 (A4) より

$$Z_{2n}^{(1)}(\boldsymbol{\phi}) \xrightarrow{p} \frac{\boldsymbol{\phi}'\mathbf{D}\boldsymbol{\phi}}{2}$$

となる．

次に $Z_{2n}^{(2)}(\boldsymbol{\phi}) \xrightarrow{p} 0$ を示そう．いま

$$e_i = \int_0^1[\{\mathbf{1}(u_i \le \mathbf{x}_i^{*'}\boldsymbol{\phi}s/\sqrt{n}) - \mathbf{1}(u_i \le 0)\} - \{F_U(\mathbf{x}_i^{*'}\boldsymbol{\phi}s/\sqrt{n}|\mathbf{x}_i) - F_U(0|\mathbf{x}_i)\}]ds$$

とおくと，$n^{-1/2}\sum_{i=1}^n \mathbf{x}_i^{*'}\boldsymbol{\phi}e_i \xrightarrow{p} 0$ となることを示せばよい．まず $\mathcal{E}[e_i|\mathbf{x}_1,\ldots,\mathbf{x}_n] = 0$ $(i = 1,\ldots,n)$ および

$$\begin{aligned} \mathcal{E}[e_i^2|\mathbf{x}_1,\ldots,\mathbf{x}_n] &\le \int_0^1 \mathcal{E}[\{\mathbf{1}(u_i \le \mathbf{x}_i^{*'}\boldsymbol{\phi}s/\sqrt{n}) - \mathbf{1}(u_i \le 0)\}^2|\mathbf{x}_1,\ldots,\mathbf{x}_n]ds \\ &\le \int_0^1 |F_U(\mathbf{x}_i^{*'}\boldsymbol{\phi}s/\sqrt{n}|\mathbf{x}_i) - F_U(0|\mathbf{x}_i)|ds \end{aligned}$$

に注意する．最後の不等式は，

$$\begin{aligned} & \{\mathbf{1}(u_i \le \mathbf{x}_i^{*'}\boldsymbol{\phi}s/\sqrt{n}) - \mathbf{1}(u_i \le 0)\}^2 \\ & = \mathbf{1}(u_i \le \mathbf{x}_i^{*'}\boldsymbol{\phi}s/\sqrt{n}) + \mathbf{1}(u_i \le 0) - 2\mathbf{1}(u_i \le \mathbf{x}_i^{*'}\boldsymbol{\phi}s/\sqrt{n})\mathbf{1}(u_i \le 0) \\ & = \mathbf{1}(u_i \le \max\{\mathbf{x}_i^{*'}\boldsymbol{\phi}s/\sqrt{n}, 0\}) - \mathbf{1}(u_i \le \min\{\mathbf{x}_i^{*'}\boldsymbol{\phi}s/\sqrt{n}, 0\}) \end{aligned}$$

から導かれる．従って

$$(35)\quad \mathcal{E}\left[\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathbf{x}_i^{*'}\boldsymbol{\phi})e_i\right\}^2\right] \leq \frac{1}{n}\sum_{i=1}^{n}\mathcal{E}\left[(\mathbf{x}_i^{*'}\boldsymbol{\phi})^2\int_0^1 |F_U(\mathbf{x}_i^{*'}\boldsymbol{\phi}s\frac{1}{\sqrt{n}}|\mathbf{x}_i) - F_U(0|\mathbf{x}_i)|ds\right]$$

が成り立つ．さらに (34) 式を用いると

$$\int_0^1 |F_U(\mathbf{x}_i^{*'}\boldsymbol{\phi}s/\sqrt{n}|\mathbf{x}_i) - F_U(0|\mathbf{x}_i)|ds \leq \frac{1}{2n}\max_{1\leq i\leq n}|\mathbf{x}_i^{*'}\boldsymbol{\phi}| \xrightarrow{p} 0$$

となる．すると，Lebesgue の収束定理より (35) 式の右辺が 0 に収束するから，$Z_{2n}^{(2)}(\boldsymbol{\phi}) \xrightarrow{p} 0$ を得る．

したがって $Z_n(\boldsymbol{\phi}) = \widetilde{Z}_n(\boldsymbol{\phi}) + o_p(1)$ および

$$(36)\qquad \widetilde{Z}_n(\boldsymbol{\phi}) = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i^{*'}\boldsymbol{\phi}\{\tau - \mathbf{1}(u_i < 0)\} + \frac{\boldsymbol{\phi}'\mathbf{D}\boldsymbol{\phi}}{2}$$

が示された．ここで $\mathbf{D}$ は正定値行列だから，$\widetilde{Z}_n(\boldsymbol{\phi})$ の一意な最小化点は $\boldsymbol{\phi} = \mathbf{D}^{-1}\mathbf{W}_n$ で与えられる．ただし $\mathbf{W}_n = (1/\sqrt{n})\sum_{i=1}^{n}\mathbf{x}_i^*\{\tau - \mathbf{1}(u_i < 0)\}$ である．このとき，補題 (1(6.2 節を参照) より，任意の $K > 0$ に対して

$$(37)\qquad \sup_{\|\boldsymbol{\phi}\|\leq K}|Z_n(\boldsymbol{\phi}) - \widetilde{Z}_n(\boldsymbol{\phi})| = \sup_{\|\boldsymbol{\phi}\|\leq K}|\{Z_n(\boldsymbol{\phi}) + \boldsymbol{\phi}'\mathbf{W}_n\} - \frac{\boldsymbol{\phi}'\mathbf{D}\boldsymbol{\phi}}{2}| \xrightarrow{p} 0$$

が成り立つ．したがって一致性の証明と同様に

$$(38)\qquad \sqrt{n}\{\widehat{\boldsymbol{\delta}}(\tau) - \boldsymbol{\delta}(\tau)\} = \mathbf{D}^{-1}\mathbf{W}_n + o_p(1)$$

を得る．最後に中心極限定理より

$$(39)\qquad \mathbf{W}_n \xrightarrow{d} \mathcal{N}(\mathbf{0}, \tau(1-\tau)\mathbf{C})$$

でありから，基準化された分位点回帰推定量の漸近共分散行列は

$$(40)\qquad \tau(1-\tau)\mathbf{D}^{-1}\mathbf{C}\mathbf{D}^{-1}$$

で与えられることがわかる． □

A.2 補題の証明

この数学補論では分位点回帰推定量の一致性と漸近正規性を示すときに用いた補題 1 を証明する．補題 1 は本質的には Rockafellar[20] の Theorem 10.8 をランダムな凸関数列のケースに拡張したものである．Pollard [17] では補題 1 を凸性の補題（CONVEXITY LEMMA）と呼び，この補題を使って LAD 推定量の漸近分布の証明を与えている．補題 1 の証明としては，対角線論法を用いて Rockafellar [20] の Theorem 10.8 に帰着させる方法も考えられるが，ここでは Pollard [17] に従い自己充足的な証明を与える．

補題 1. $g_n : \mathbb{R}^d \to \mathbb{R}$ をランダムな凸関数列とする．$g : \mathbb{R}^d \to \mathbb{R}$ を非確率的な関数とし，各 $\mathbf{u} \in \mathbb{R}^{\mathbf{d}}$ に対して，$g_n(\mathbf{u}) \xrightarrow{p} g(\mathbf{u})$ が成り立っているとする．このとき，g は凸関数であって， 任意のコンパクト集合 $K \subset \mathbb{R}^d$ に対して

$$\sup_{\mathbf{u}\in K} |g_n(\mathbf{u}) - g(\mathbf{u})| \xrightarrow{p} 0$$

が成り立つ．

証明. 簡単のため $d = 1$，$K = [0, 1]$ の場合を示す (一般の場合も記号がより複雑になるがほぼ同様に証明できる)．$\epsilon > 0$ を任意に固定する．まず $g(u)$ は凸関数だから，特に連続関数である (福島 [5] 参照)．したがって，$g(u)$ は $[0, 1]$ 上で一様連続であるから，$k \in \mathbb{N}$ を十分大きくとれば，$|u - v| < 1/k$ なる $u, v \in [0, 1]$ に対して $|g(u) - g(v)| < \epsilon$ が成り立つ．そこで，$[0, 1]$ 区間を k 個の区間

$$[0, 1/k],\ [1/k, 2/k],\ \ldots\ , [(k-1)/k, 1] \tag{41}$$

に分割する．分割 (41) に含まれる区間は有限個だから，各 $j \in \{1, \ldots, k\}$ に対して

$$\lim_{n\to\infty} \mathrm{P}\left(\sup_{u\in[(j-1)/k, j/k]} |g_n(u) - g(u)| > C\epsilon \right) = 0 \tag{42}$$

を示せばよい．ここで C は ϵ によらない正の定数である．(42) を示すには，次の (43) と (44) を示せば十分である：

$$\lim_{n\to\infty} \mathrm{P}\left(\sup_{u\in[(j-1)/k, j/k]} \{g_n(u) - g(u)\} > C_1\epsilon \right) = 0, \tag{43}$$

$$\lim_{n\to\infty} \mathrm{P}\left(\inf_{u\in[(j-1)/k, j/k]} \{g_n(u) - g(u)\} < -C_2\epsilon \right) = 0. \tag{44}$$

ただし，C_1，C_2 は ϵ によらない正の定数である．

(43) の証明：$u \in [(j-1)/k, j/k]$ を $(j-1)/k$ と j/k の凸結合の形で表して

$$u = \alpha\frac{(j-1)}{k} + (1-\alpha)\frac{j}{k}$$

と書く．ただし $\alpha \in [0, 1]$ である．すると $g_n(u)$ は凸関数であるから，

$$\begin{aligned}
g_n(u) &\le \alpha g_n((j-1)/k) + (1-\alpha) g_n(j/k) \\
&= g(u) + \alpha\{g_n((j-1)/k) - g(u)\} + (1-\alpha)\{g_n(j/k) - g(u)\} \\
&\le g(u) + \alpha\{|g_n((j-1)/k) - g((j-1)/k)| + |g((j-1)/k) - g(u)|\} \\
&\quad + (1-\alpha)\{|g_n(j/k) - g(j/k)| + |g(j/k) - g(u)|\} \\
&\le g(u) + \max\{|g_n((j-1)/k) - g((j-1)/k)|, |g_n(j/k) - g(j/k)|\} + \epsilon
\end{aligned}$$

が成り立つ．最右辺の第 2 項と第 3 項は u に依存しないから，

$$\sup_{u\in[(j-1)/k,j/k]}\{g_n(u)-g(u)\}\le\max\{|g_n((j-1)/k)-g((j-1)/k)|,|g_n(j/k)-g(j/k)|\}+\epsilon$$

が成り立つことがわかる．ここで，$g_n(u)$ は各点で $g(u)$ に確率収束するから，右辺第 1 項は 0 に確率収束する．したがって，

$$\lim_{n\to\infty}\mathrm{P}\left(\sup_{u\in[(j-1)/k,j/k]}\{g_n(u)-g(u)\}>2\epsilon\right)=0$$

を得る．

(44) の証明：$u\in[(j-1)/k,j/k]$ を任意にとる．すると j/k は $(j+1)/k$ と u の凸結合で表せる：

$$\frac{j}{k}=\beta u+(1-\beta)\frac{(j+1)}{k}.$$

ただし，$\beta\in[1/2,1]$ である．$g_n(u)$ は凸関数であるから，

$$g_n(j/k)\le\beta g_n(u)+(1-\beta)g_n((j+1)/k)$$

より，

$$\begin{aligned}\beta g_n(u)&\ge g_n(j/k)-(1-\beta)g_n((j+1)/k)\\&=g(j/k)+\{g_n(j/k)-g(j/k)\}-(1-\beta)g((j+1)/k)\\&\quad-(1-\beta)\{g_n((j+1)/k)-g((j+1)/k)\}\\&=\beta g(u)+\{g(j/k)-g(u)\}+\{g_n(j/k)-g(j/k)\}\\&\quad-(1-\beta)\{g((j+1)/k)-g(u)\}-(1-\beta)\{g_n((j+1)/k)-g((j+1)/k)\}\\&\ge\beta g(u)-|g_n(j/k)-g(j/k)|-|g_n((j+1)/k)-g((j+1)/k)|-2\epsilon\end{aligned}$$

を得る．$\beta\ge1/2$ であることに注意すると，

$$\inf_{u\in[(j-1)/k,j/k]}\{g_n(u)-g(u)\}\ge-2|g_n(j/k)-g(j/k)|-2|g_n((j+1)/k)-g((j+1)/k)|-4\epsilon$$

が成り立つことがわかる．$g_n(u)$ は各点で $g(u)$ に確率収束するから，右辺第 1 項と第 2 項は 0 に確率収束する．したがって，

$$\lim_{n\to\infty}\mathrm{P}\left(\inf_{u\in[(j-1)k,j/k]}\{g_n(u)-g(u)\}<-5\epsilon\right)=0$$

を得る． □

A.3　データ解析の結果：幾つかの図

本稿 4 節で報告したデータ解析において得られた幾つかの図をここに示しておく．

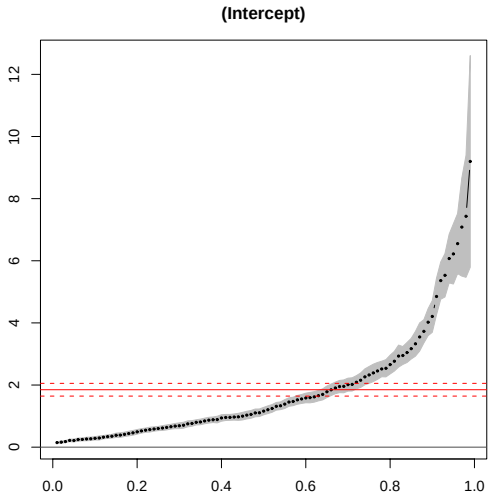

図 2: 切片

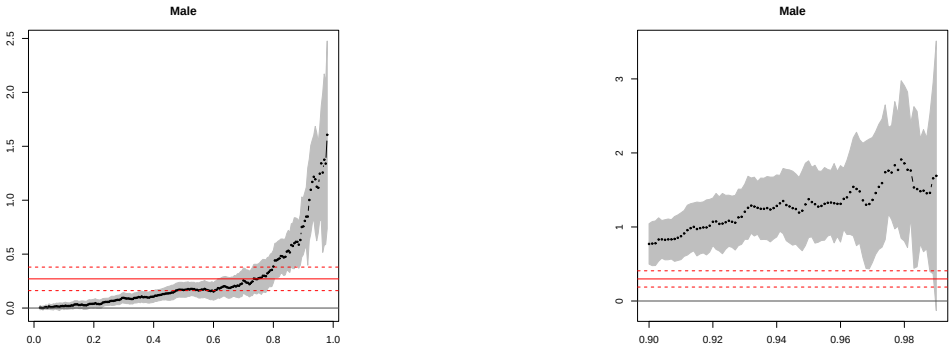

図 3: 男性

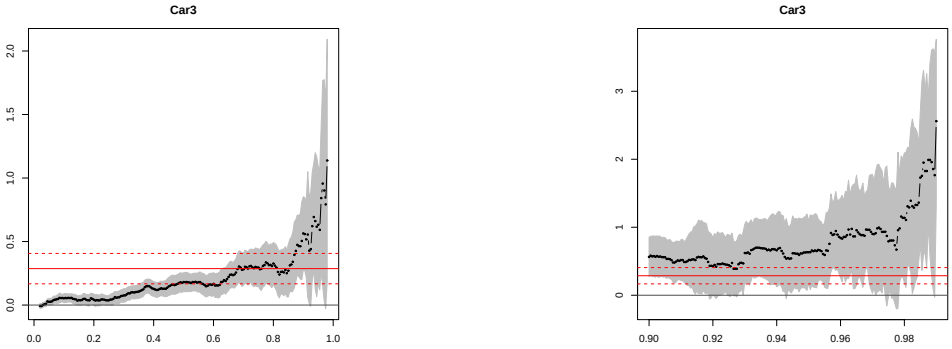

図 4: 自家用普通乗用車 (3 ナンバー)

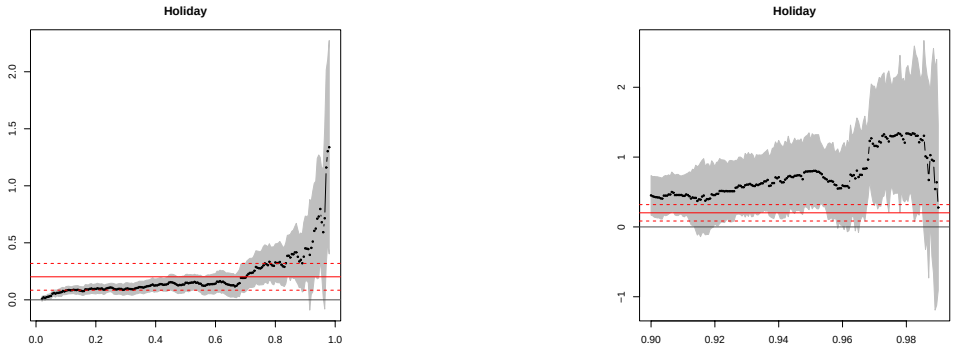

図 5: 土日

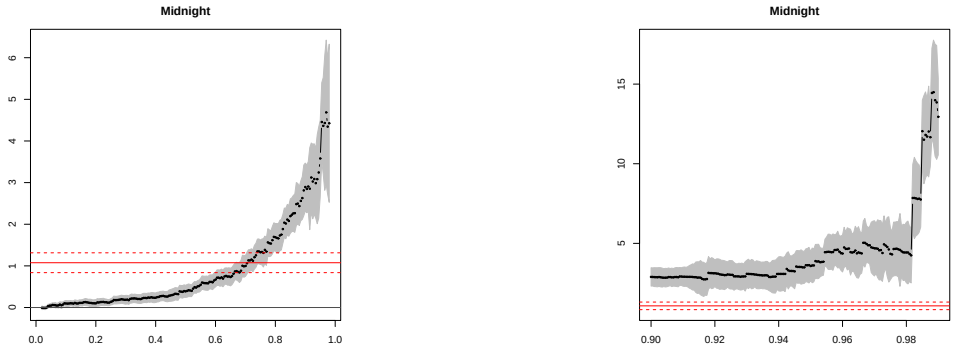

図 6: 深夜

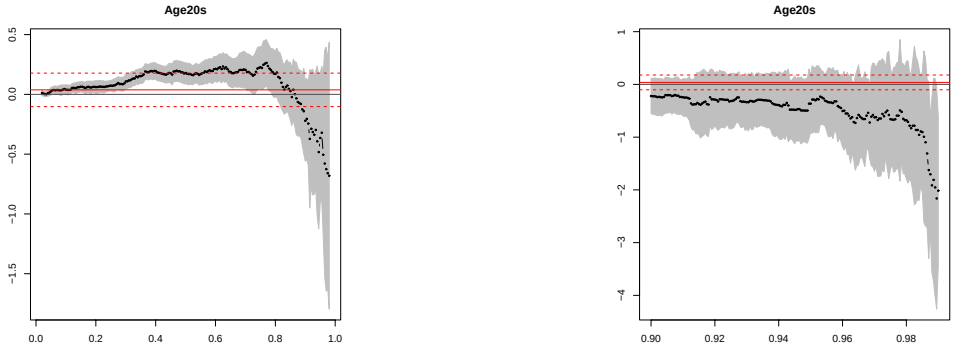

図 7: 年齢が 20 歳代

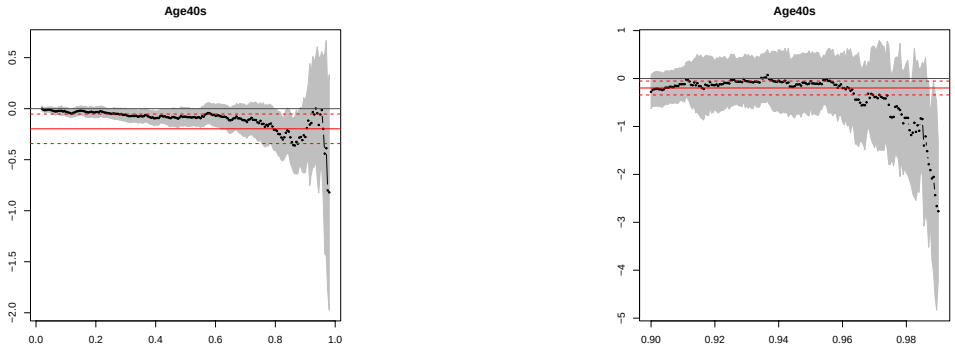

図 8: 年齢が 40 歳代

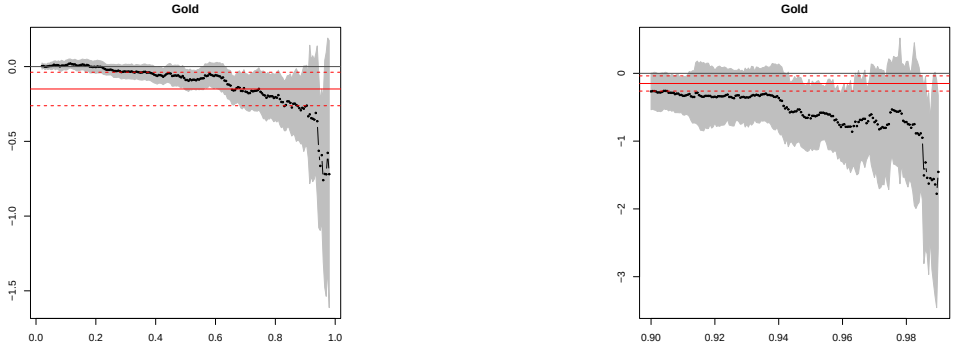

図 9: ゴールド免許所持者

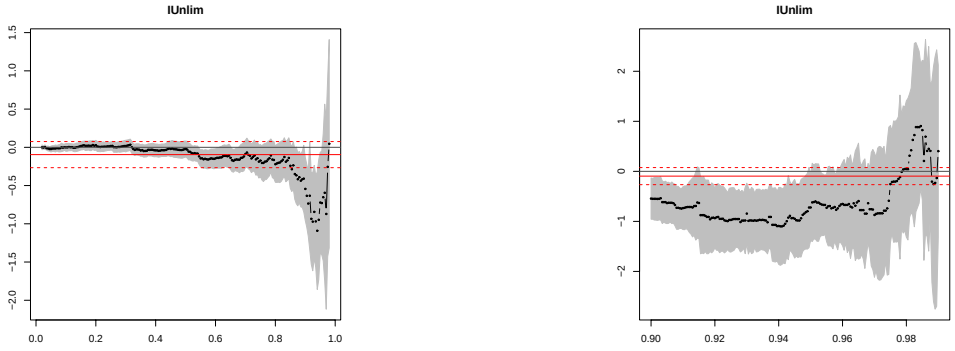

図 10: 対物保険金額が無制限

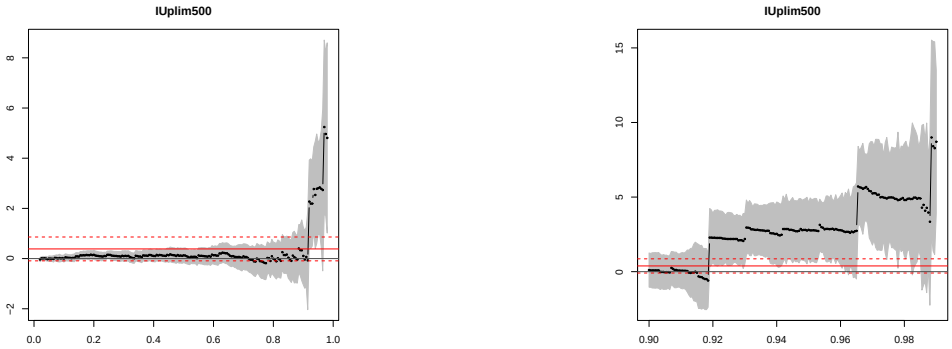

図 11: 対物保険金額が 500 万円以下

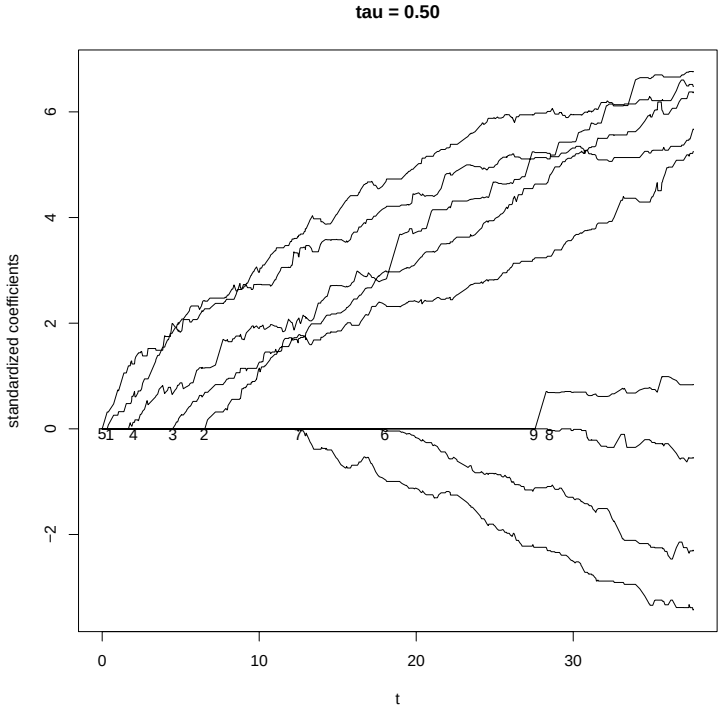

図 12: Lasso 分位点回帰の結果 ($\tau = 0.50$)

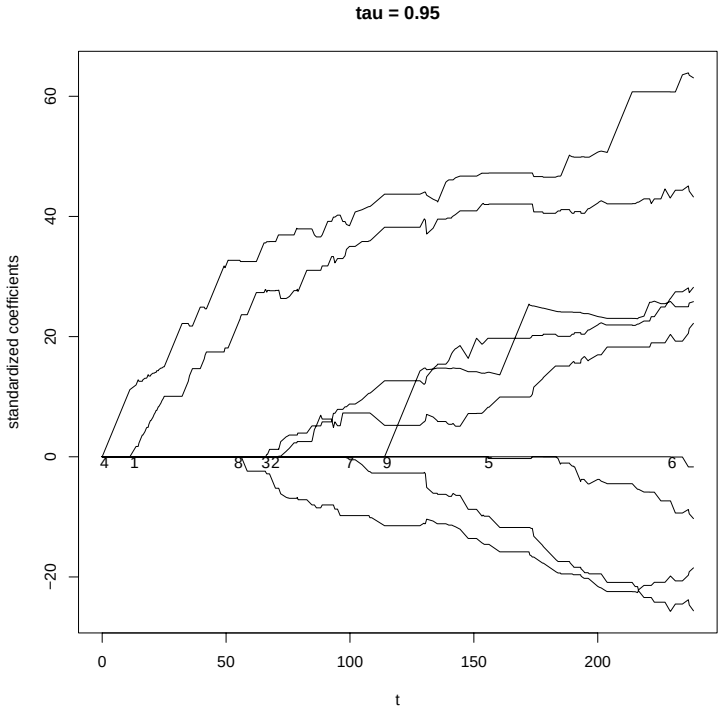

図 13: Lasso 分位点回帰の結果 ($\tau = 0.95$)

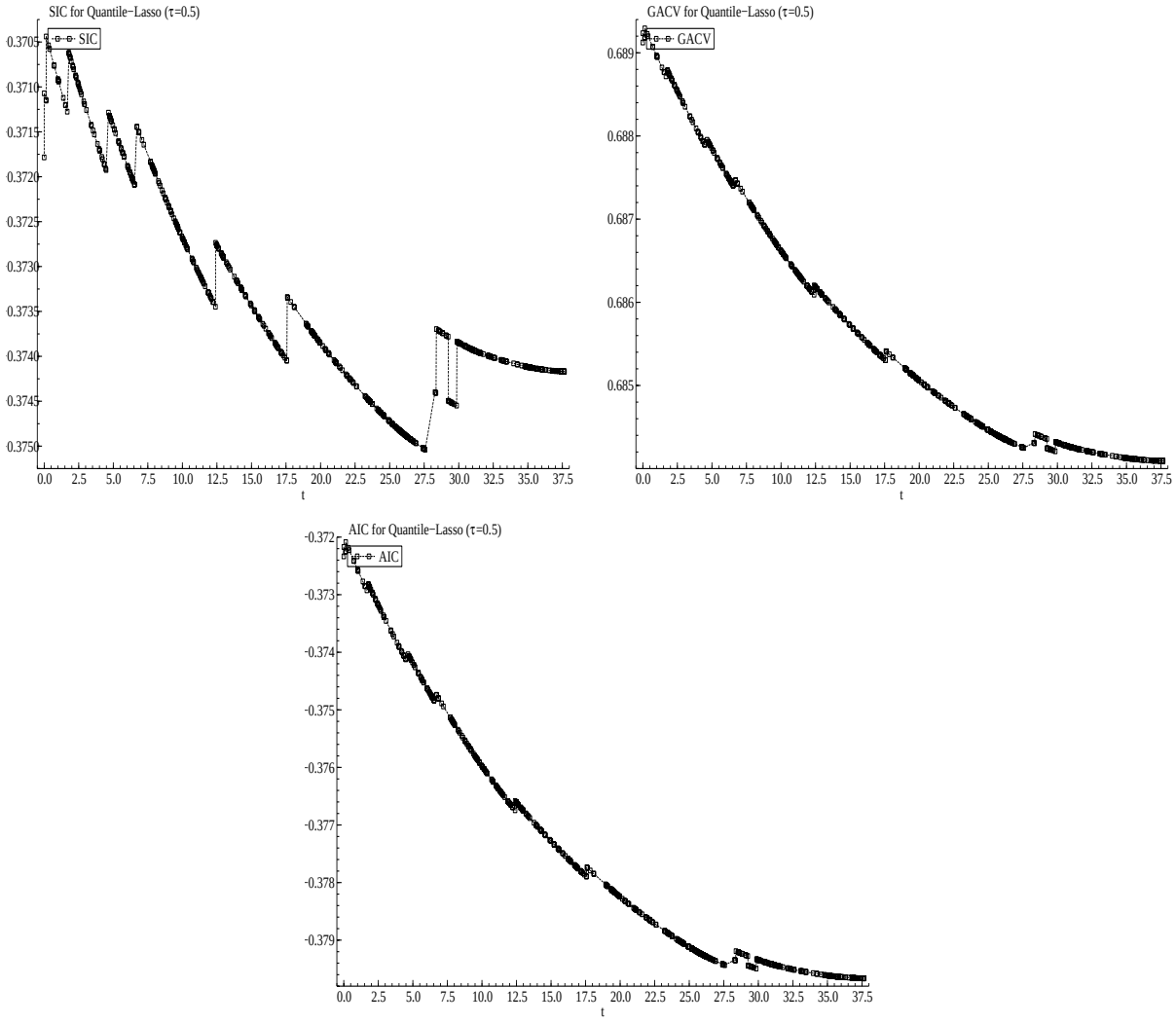

図 14: Lasso 分位点回帰 ($\tau = 0.5$) におけるモデル選択規準の値：上段左：SIC，上段右：GACV，下段：AIC

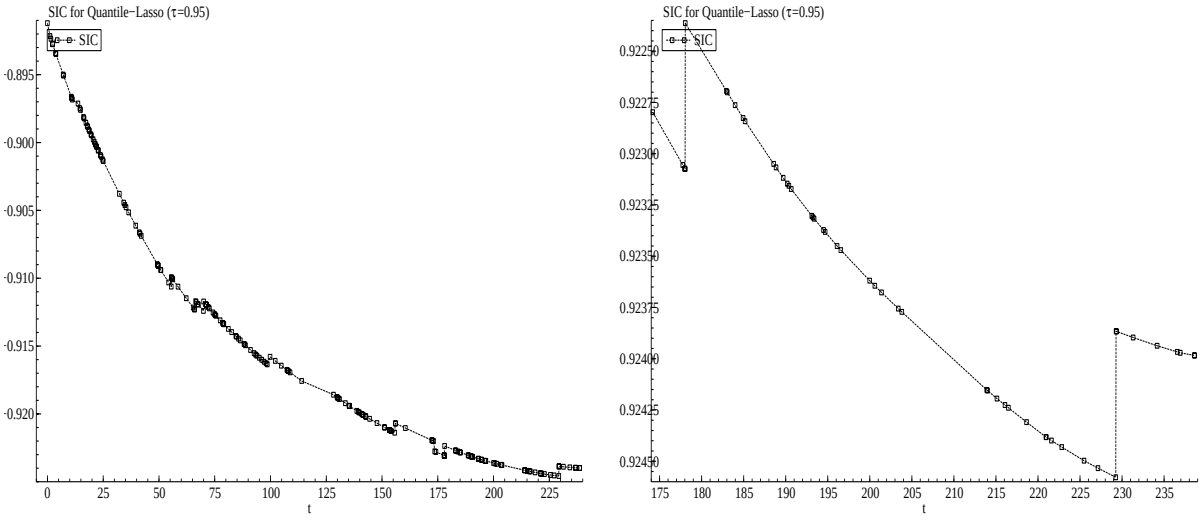

図 15: Lasso 分位点回帰 ($\tau = 0.95$) における SIC の値：右は大きい t に対する SIC の値

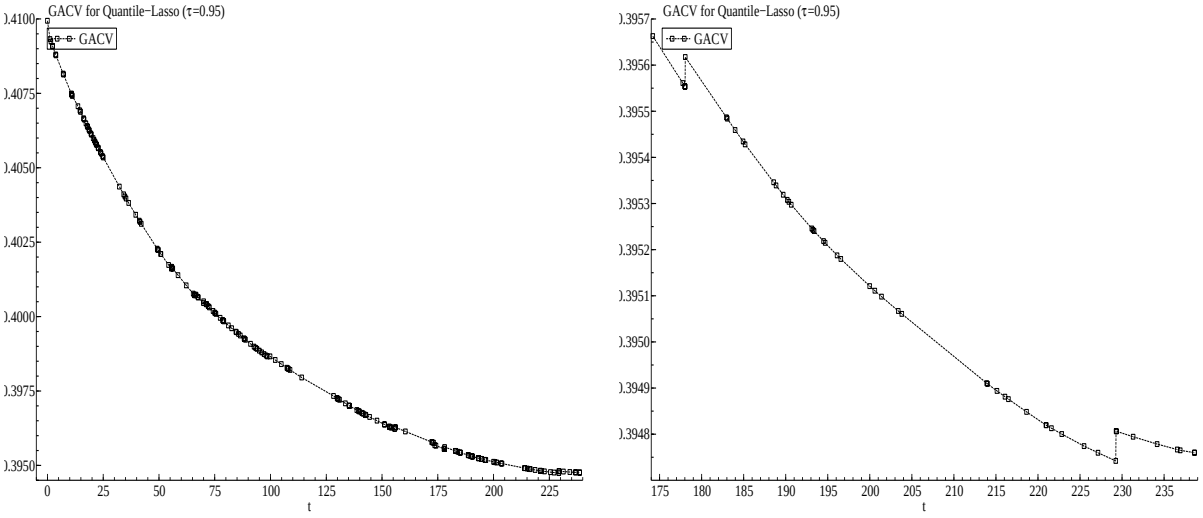

図 16: Lasso 分位点回帰 ($\tau = 0.95$) における GACV の値：右は大きい t に対する GACV の値

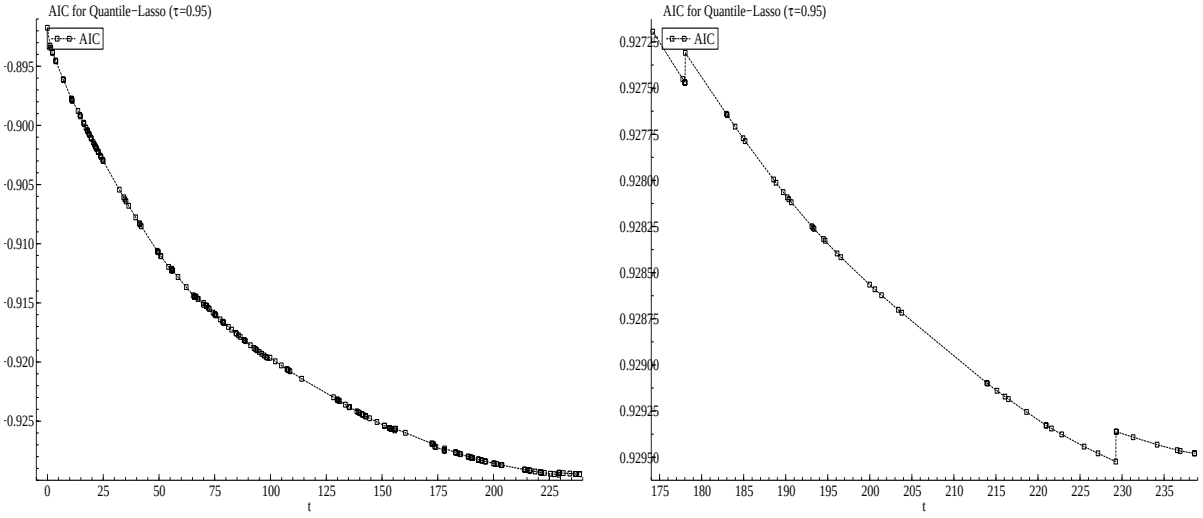

図 17: Lasso 分位点回帰 ($\tau = 0.95$) における AIC の値：右は大きい t に対する AIC の値