修士論文

二段階回帰木を用いた損害保険の純保険料推定

三重大学大学院 工学研究科 情報工学専攻
博士前期課程 2006年度 修了

西 久美子

目次

図一覧

表一覧

第1章

はじめに

1.1　研究の背景

我々の日常生活の中には様々なリスク (事故や災害などの危険) がある．損害保険は，このようなリスクによる「損害賠償」という状況を想定し，被害者救済が滞らないように (加害者が金銭的リスクを回避できるように) 生まれた商品である．損害保険会社と契約を結ぶことで，事故や災害時などに補償が受けられるというものであり，火災に備える火災保険，自動車事故に備える自動車保険，海外旅行時の傷害などに備える海外旅行傷害保険などが幅広く利用されている．現在では，損害保険各社は1998年の規制緩和による保険料自由化に伴い，新しい保険商品の開発を進めており，その中でも，顧客のリスクファクター (年齢，性別など) を従来の保険商品よりも細分化して保険料を設定する「リスク細分型保険商品」が注目されている．

損害保険の価格である保険料は，一般の商品の価格と同じ考え方だが，大きく異なる点は，価格の中心となる部分が事故発生時の保険会社が支払う保険金によって構成されているため，保険を販売する時点ではあらかじめ確定していない点である．そのため保険料に含まれる保険金の支払い部分 (これを純保険料という) について，過去の保険契約をもとに科学的または工学的手法を用いて，将来の事故の支払額を計算することによって保険料が設定されている．

1.2　研究の目的と概要

本研究では1.1節の背景より，「リスク細分型保険商品」における純保険料の推定を目的とする．通常，損害保険商品の設計は過去の保険契約履歴データベースから次のような手順で行われる：

1. **リスクグループ同定**：
   リスク (支払保険金額) が均一になるような保険契約のグループを同定する.

2. **純保険料推定**：
   同じリスクグループに属する契約は同一のリスクを持つと仮定し，各リスクグループの平均支払保険金額を推定する.

現状では，リスクグループ同定は専門家のヒューリスティクスに基づいて行われ，純保険料推定は数理統計的な方法を用いて行われている．しかしながら，現状の方法について2つの問題が存在する．まず，顧客のリスクファクターを従来の保険商品よりも細かく設定することにより，リスクファクターの種類が増加し，専門家の知識だけではリスクグループ同定が困難になる．次に，支払保険金額は大部分の支払い事例が軽微な事故による低額なものであるが，稀に生じる大事故の事例により莫大な金額が支払われるため，対数正規分布のような正方向に長い裾を持つ．この特徴により，純保険料推定において最小二乗誤差回帰分析を用いると，分布の裾から生じるサンプルが推定値に大きな影響を与え，推定量の分散が大きくなってしまい，推定値が不安定になる.

このような問題を克服するために，データマイニング手法である回帰木を適用し，リスクグループの同定と純保険料推定を機械的かつ正確に行なう方法を考察する．図1.1は損害保険料設定問題を回帰木を用いて解析したイメージである．回帰木は分類や予測などを行なう際に幅広く用いられ，実際にマーケティングなどの分野にも利用されている．回帰木解析の大きな特徴として，分類にデータをグループ分けする決まり (ルール) を使用していることが挙げられる．本研究において，回帰木はリスクファクターに応じたリスクグループの自動同定を可能とするため，ニューラルネットやサポートベクトルマシンなどの他の機械学習アルゴリズムよりも損害保険料設定問題に適していると判断し，ロバスト回帰分析 [Huber 82] の枠組を利用した回帰木アルゴリズムを提案する.

本論文は以下のように構成される．第1章では，研究の概要と背景，目的を述べる．第2章では，損害保険料設定が過去の保険契約履歴をもとにリスクを推定する問題であることを述べ，問題の定式化を行う．また，損害保険データ特有の特徴をとりあげ，標準的なデータマイニング手法や統計解析手法を利用することが適切でないことを指摘する．第3章では，提案方法とその理論的な背景を説明する．第4章では，提案方法を数値シミュレーションにて検討し，自動車保険データへ適用した結果を示す．第5章では，本論文の論点をまとめる.

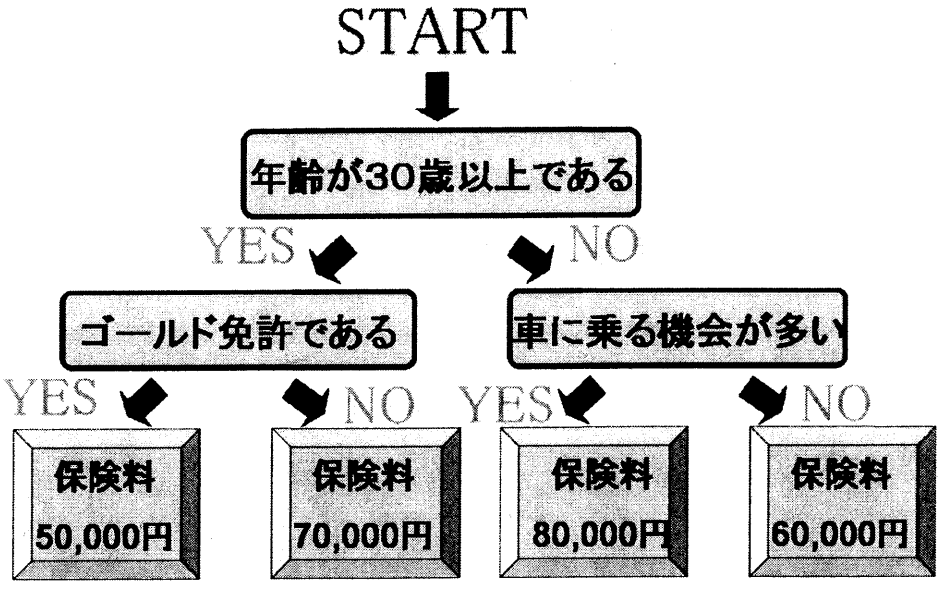


図　1.1: **損害保険料設定問題を回帰木を用いて解析したイメージ**：質問に対し YES/NO で答えるだけで，最終的にどのグループに属し保険料がいくらになるかわかるようになっている．例えば，26 歳で車に乗る機会が多い人の保険料は 80,000 円になる．

第2章

損害保険料設定と既存研究

本章では，初めに純保険料推定問題を回帰分析問題へ定式化し，次にその回帰分析問題を解く回帰木アルゴリズムについて説明する．そして，既存の方法では推定が困難である点を挙げる．

2.1　純保険料推定問題の定式化

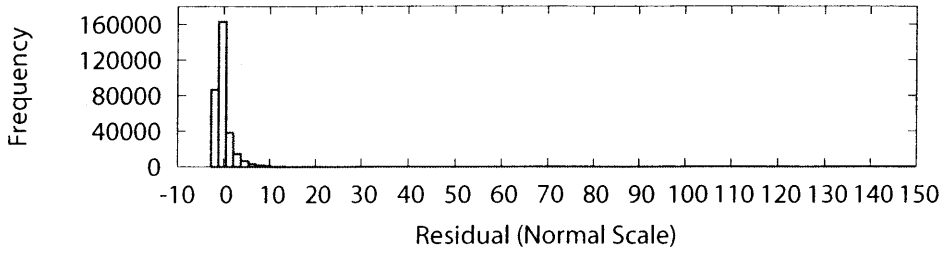
保険料は，純保険料と付加保険料から構成される．前者は将来の支払保険金に充当し，後者はその他の付加的な経費に相当する．純保険料を算出するには，将来支払われる保険金総額を予測する必要があり，通常，支払件数と平均支払金額をそれぞれ推定し，両者を乗じることによって推定される．支払件数はポアソン分布や負の二項分布で近似され，標準の統計解析手法を利用して推定することができる．本論文では，平均支払金額を推定する問題のみを扱う．以下，純保険料とは平均支払金額を意味するものとする．

p 種類のリスクファクターを $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^p$，支払金額を $Y \in \mathbb{R}^+$ と表記する．両者の関係は次のように表される：

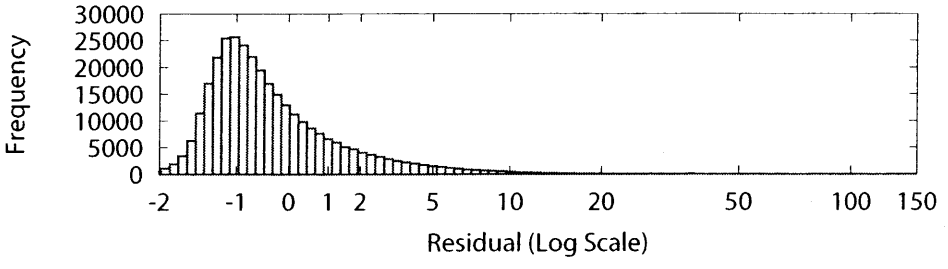
$$Y = f(\boldsymbol{x}) + \epsilon(\boldsymbol{x}),\ \ E[\epsilon(\boldsymbol{x})] = 0,\ \ \forall \boldsymbol{x} \in \mathcal{X}, \tag{2.1}$$

ここで，$f : \mathcal{X} \to \mathbb{R}^+$ はリスクファクターと支払金額との関係を表す関数，$\epsilon(\boldsymbol{x})$ は支払金額のバラツキを表す平均が0の確率変数を表している．通常の回帰分析では誤差項が $\boldsymbol{x}$ に依存しないと仮定する場合が多いが，本問題ではこの仮定があてはまらないため，誤差項が $\boldsymbol{x}$ に依存することを明示的に表し，$\epsilon(\boldsymbol{x})$ と表記する．(2.1) 式から生じた n 件の保険契約履歴を $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ と表す．純保険料推定は，リスクファクター $\boldsymbol{x} \in \mathcal{X}$ が与えられたもとでの支払金額 Y の期待値 $E[Y|\boldsymbol{x}] = f(\boldsymbol{x}) + E[\epsilon] = f(\boldsymbol{x})$ を推定する問題である．すなわち，学習データ $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ から関数 f を推定する回帰分析として定式化される．

この回帰分析で特記すべき点は，支払金額のバラツキを表す確率変数ϵが特徴的な分布に従うことである．図2.1は，第4章で用いる自動車保険データに標準的な回帰木アルゴリズム(詳細は後述)を適用した際の残差をヒストグラム形式でプロットしたものである．これはϵの分布の概形を表しており，正方向に長い裾を持つことがわかる．図2.1(b)の対数尺度の分布でさえも正方向に歪んでいることから，ϵの分布は対数正規分布よりも裾が長いものであると推察される[1]．損害保険においては，図2.1のような正方向に長い裾を持つ分布は損害保険の本質的な性質に起因するもので，この特徴を考慮した回帰分析を行う必要がある．



(a) Normal Scale



(b) Log Scale

図 2.1: **自動車保険契約318,564件に最小二乗誤差基準による回帰木を適用した際の残差のヒストグラム**：(a)は標準尺度のものを，(b)は対数尺度のものを表している．損害保険の支払金額のバラツキ((2.1)式の誤差項ϵ)の分布が正方向に長い裾を持つことが推察される．対数尺度の分布(b)も正方向に歪んでいることから，誤差項の分布は対数正規分布よりも裾が長いと考えられる．これらの残差の歪度は8.499，尖度は269.8であった．

2.2　回帰木アルゴリズム

データマイニングのアプローチを適用した既存研究としては，IBM社による対数正規尤度を最大化するような回帰木[Apte 99]，ApStat社によるMixture of Expert[Jacobs 91]のアプロー

[1]対数正規分布であれば対数尺度の分布は正規分布となる．

チに類似した階層型ニューラルネットを用いたもの [Dugas 04] がみられる.

本論文では，回帰木によるアプローチを試みる．回帰木とは，木構造に基づく回帰モデルは特徴空間を再帰的に分割し，それぞれの領域を単純なモデルで記述するものである．その特徴空間はいくつかの超直方体領域に分割され，各領域が「葉」と呼ばれる木構造の末端ノードに対応している．このようなモデルは解釈が容易であるのに加え，複数の特徴の相互作用を記述する能力を備えており，多くのデータマイニングアプリケーションで実用されている．学習データから木構造モデルを作成するさまざまなアルゴリズムが提案されている [Breiman 84, Quinlan 93].

木構造モデルの生成アルゴリズムは「分割基準」と「サイズ決定規則」によりおおまかに特徴づけられる.

分割基準とはどのように特徴空間を分割するべきかを定量化したものである．ほとんどのアプリケーションでは，最小二乗誤差基準に基づく回帰木 (L_2 木) が用いられている．学習データにより生成された L_2 木は条件付平均 $E[Y|\boldsymbol{x}]$ の予測器として用いられる．特徴量 $\boldsymbol{x}_0$ を持つデータが与えられたとすると，$\boldsymbol{x}_0$ が木構造にしたがっていずれかの葉に分類され，分類された葉に集まった学習データのサンプル平均値が $E[Y|\boldsymbol{x}_0]$ の予測値となる．ロバストな性能を持つとされる最小絶対誤差基準に基づく回帰木 (L_1 木) も提案されている．学習データにより生成された L_1 木は条件付中央値 $F^{-1}_{Y|\boldsymbol{x}}(0.5)$ の予測器として用いられる．特徴量 $\boldsymbol{x}_0$ を持つデータが与えられたとすると，$\boldsymbol{x}_0$ が木構造にしたがっていずれかの葉に分類され，分類された葉に集まった学習データのサンプル中央値が $F^{-1}_{Y|\boldsymbol{x}_0}(0.5)$ の予測値となる.

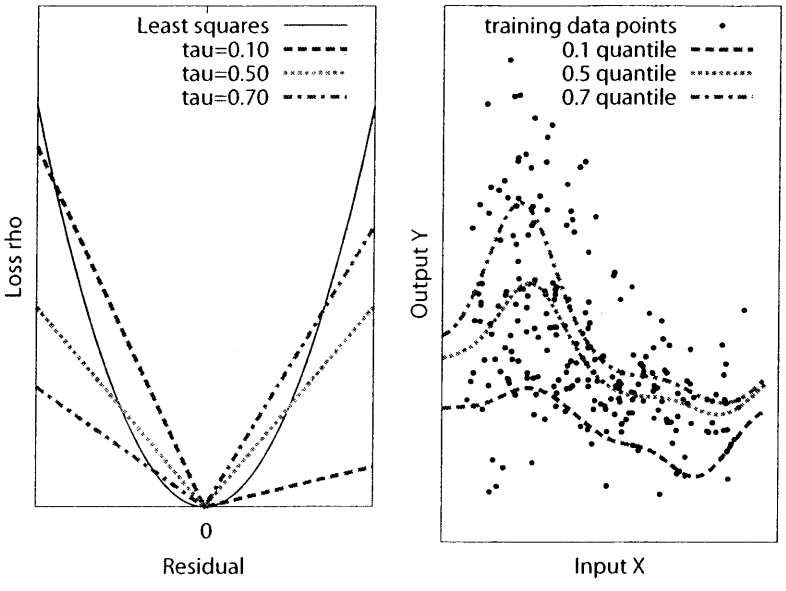
本研究では L_1 木を拡張した分位点回帰木を用いる．L_1 木は条件付中央値，すなわち，条件付 0.5 分位点，$F^{-1}_{Y|\boldsymbol{x}}(0.5)$，を推定するものであるが，これを一般の条件付 $\tau \in (0,1)$ 分位点，$F^{-1}_{Y|\boldsymbol{x}}(\tau)$，を推定するものへ拡張する．条件付分位点関数を推定するための方法論は分位点回帰分析 (Quantile Regression) と呼ばれ，Koenker を中心として研究されている [Koenker 78, Koenker 94, Koenker 05, Takeuchi 06]．分位点回帰分析は次のように定式化される：

$$\hat{f}_\tau = \arg\min_{f\in\mathcal{F}} \sum_{i=1}^{n} \psi_\tau\{y_i - f(\boldsymbol{x}_i)\}, \tag{2.2a}$$

$$\psi_\tau(r) = \begin{cases} (1-\tau)|r|, & \text{if } r \le 0, \\ \tau|r|, & \text{if } 0 < r. \end{cases} \tag{2.2b}$$

ここで，$\mathcal{F}$ はなんらかの関数のクラスを表している．損失関数 ψ_τ は，最小絶対誤差基準の損失関数を一般化したもので，$\tau = 0.5$ のとき L_1 回帰に一致する．図 2.2 に，$\tau = 0.1, 0.5, 0.7$ における損失関数と解析例を示す．分位点回帰木は分位点損失関数 (2.2b) 式を最小にするような分割基準を用いて木構造を生成する．学習データにより生成された τ-分位点回帰木 (Q_τ 木) は条件付 τ 分位点 $F^{-1}_{Y|\boldsymbol{x}}(\tau)$ の予測器として用いられる．特徴量 $\boldsymbol{x}_0$ を持つデータが与えられたとす

ると，$\boldsymbol{x}_0$ が木構造にしたがっていずれかの葉に分類され，分類された葉に集まった学習データのサンプル τ 分位点が $F^{-1}_{Y|\boldsymbol{x}_0}(\tau)$ の予測値となる．L_1 木は分位点回帰木の一例として $Q_{0.5}$ 木と表すことができるため，以降では，まとめて分位点回帰木と表現する．



(a) 分位点回帰の損失関数　　(b) 分位点回帰の例

図 2.2: **分位点回帰分析の損失関数と解析例**：(a) は損失関数を，(b) は解析例を表している．

サイズ決定規則とは，回帰木のサイズを定めるモデル選択基準で，モデルの汎化性能に大きく影響を及ぼす．サイズ決定規則には，分割を途中で停止するアプローチと，最大の木を作成した後に不必要な枝を剪定するアプローチに大別できる．前者は，計算量が少なく実装も容易であるが，汎化性能の観点からは後者が優れているとの報告が多い．本研究では CART [Breiman 84] で用いられている交差検証法による剪定アルゴリズムを採用する．以下にその概要を説明する．回帰木のあるノードを t，そのノードをルートノードとする部分木を T とし，それぞれの学習誤差を $\mathrm{TrErr}(t)$，$\mathrm{TrErr}(T)$ と表す．汎化誤差を推定するため，学習誤差に葉数に比例する項を加えた $\mathrm{GeErr}_\alpha(T) = \mathrm{TrErr}_\alpha(T) + \alpha|T|$ という指標を導入する．ここで，$\alpha \in \mathbb{R}^+$ は剪定の程度を制御するパラメータ，$|T|$ は木 T の葉数を表している．ノード t を葉数が 1 の木とみなすと，t に対する同様の指標は $\mathrm{GeErr}_\alpha(t) = \mathrm{TrErr}_\alpha(t) + \alpha$ と表される．$\mathrm{GeErr}_\alpha(T) = \mathrm{GeErr}_\alpha(t)$ となるような α を α_t とすると，$\alpha_t < \alpha$ のときノード t 以下の部分木は過学習と判断され剪定される．最大の木から α_t が最小となるノード以下の部分木を順次剪定していき，剪定パラメータの候補列 $\alpha_1 < \alpha_2 < \cdots$ を得る．これらの候補のうち最適な α_{best} が 10-fold 交差検証法により決定され，$\alpha_t < \alpha_{\mathrm{best}}$ となるノード t 以下の部分木は剪定される．

2.3　純保険料の推定に伴う問題点

誤差項ϵの分布の裾が長いとき，最小二乗誤差回帰分析 (L_2 回帰) が不安定であることはよく知られている．このような問題に対処するためにロバスト統計 [Huber 82] と呼ばれる分野が大きく発展し，様々なロバスト回帰推定量 [Rousseeuw 87] が考案されてきた．もっともよく使われているロバスト回帰推定量は残差の絶対値を最小化する最小絶対誤差回帰分析 (L_1 回帰) である．誤差項が (均一分散な) 正規分布であるならば L_2 回帰が最尤推定となることはよく知られているが，L_1 回帰は誤差項がラプラス分布の場合の最尤推定に対応している[2]．

誤差項ϵの分布にかかわらず，L_2 回帰は条件付平均関数 $E[Y|\boldsymbol{x}] = f(\boldsymbol{x})$ を，L_1 回帰は条件付中央値関数 $F^{-1}_{Y|\boldsymbol{x}}(0.5) = f(\boldsymbol{x}) + F^{-1}_{\epsilon(\boldsymbol{x})}(0.5)$ を推定する枠組を提供する[3]．ここで，F_Z は確率変数 Z の累積分布関数を，F_Z^{-1} はその逆関数を表すものとする[4]．誤差項の分布が対称であるならば平均値と中央値は一致するため，条件付平均を推定する目的で L_1 回帰を用いることができる．しかしながら，損害保険の支払金額の分布は図 2.1 のように非対称であるため，条件付平均関数と条件付中央値関数は一致しない．このため，純保険料推定問題に L_1 回帰や他のロバスト回帰推定量を用いると条件付平均を過小に見積ってしまい，将来の支払に備える準備金が不足してしまう．

誤差項ϵの分布が既知であるとき，最尤推定を用いることができる．IBM 社の [Apte 99] のように損害保険の支払金額のパラメトリックモデルとして，対数正規分布を用いる場合が多い．しかしながら，図 2.1 に例示されるように，支払金額が対数正規分布よりも裾の長い分布であることが指摘されつつあり，パレート分布などの利用が試みられ始めている．また前述したように，支払金額のバラツキ $\epsilon(\boldsymbol{x})$ はリスクファクターに依存しているため，最尤推定を行うには分散関数も既知である必要がある．一般化線形モデル [McCullaugh 89] は均一分散正規誤差モデルを指数分布族へ拡張したものである．このうち，ガンマ分布モデルを損害保険の支払金額に適用する試みがなされている [Haberman 96]．

また他に，誤差項が非対称である場合，Y に Box-Cox 変換 [Box 64] 等の変数変換を適用すると有効な場合が多い．適切な変換を行えば，変換後のデータの誤差項を対称にすることができ，標準的な統計解析方法を用いてよい推定量を構築できる．しかしながら，条件付平均 $E[Y|\boldsymbol{x}]$ を推定する問題においては，ロバスト統計量の場合と同様に，変換バイアス [Miller 84] と呼ばれるバイアスを生じてしまう．変数変換を h，その逆変換を h^{-1} とすると，変換後の条件付平均

---

[2] ラプラス分布は分布の裾が 1 次の指数関数で減衰し正規分布に比べて長い裾を持っている

[3] 一変量サンプル $y_1, y_2, \cdots, y_n$ に対し，二乗誤差最小化：$\arg\min_\mu \sum_{i=1}^n (y_i - \mu)^2$ はサンプル平均と一致し，絶対誤差最小化：$\arg\min_\mu \sum_{i=1}^n |y_i - \mu|$ はサンプル中央値 (メディアン) と一致することを確認されたい．

[4] 累積分布関数の逆関数 F_Z^{-1} は分位点関数と呼ばれる．

$E[h(Y)|\boldsymbol{x}]$ を求めるよい推定量を構築することができるがこれを逆変換して元の尺度に戻したもの $h^{-1}(E[h(Y)|\boldsymbol{x}])$ は $E[Y|\boldsymbol{x}]$ と一致しない．変換バイアスを補正する方法としてSmearing推定量 [Duan 83] が提案されている．

第3章

提案手法 - 二段階回帰木

本論文では，第2章で議論した支払金額分布の特徴を考慮した回帰分析を回帰木 [Breiman 84, Quinlan 93] の枠組で実装する．提案する回帰木アルゴリズムは，第1段階でロバスト推定量の基準を用いた回帰木を作成し，第2段階で条件付平均関数に対するバイアス補正を行う．以降，提案するアプローチを二段階回帰木 (Two-Stage Regression Tree) と呼ぶ．

まず，3.1節で提案手法の理論的な背景となる二段階回帰推定量 [Takeuchi 02, Kanamori 06] を一般的な回帰分析の枠組で考察する．次に3.2節で，[Takeuchi 02, Kanamori 06] のアプローチの問題点を指摘するとともに，二段階回帰推定量を回帰木の枠組で実装した二段階回帰木を提案する．最後に，3.3節で二段階回帰木のアルゴリズムをまとめる．

3.1　二段階回帰推定量

第2章で議論したように，損害保険の支払金額は正方向に長い裾を持つ分布に従う．最小二乗誤差基準による回帰木 (L_2 木) は分布の裾から生じるサンプル (はずれ値) の影響を強く受けて不安定になってしまう．一方，分位点回帰木 (Q_τ 木) は支払金額の条件付分位点の推定量であるため，そのまま用いると，推定値が条件付平均 $E[Y|\boldsymbol{x}]$ から大きくずれてしまう．

純保険料推定問題の推定対象である条件付平均関数は次のように表すことができる：

$$E[Y|\boldsymbol{x}] = F^{-1}_{Y|\boldsymbol{x}}(\tau) + \{E[Y|\boldsymbol{x}] - F^{-1}_{Y|\boldsymbol{x}}(\tau)\}. \tag{3.1}$$

右辺第1項は条件付分位点を，右辺第2項は条件付平均と条件分位点のずれを表している．次のような二段階推定量を考える：

第1段階 (3.1) 式第1項の条件付分位点，$F^{-1}_{Y|\boldsymbol{x}}(\tau)$，を分位点回帰により推定する．

第2段階 (3.1) 式第2項の条件付平均と条件付分位点の差，$E[Y|\boldsymbol{x}] - F^{-1}_{Y|\boldsymbol{x}}(\tau)$，を最小二乗回帰により推定する．

これまで考察してきたように，第1段階でははずれ値の影響が少なく，$F^{-1}_{Y|\boldsymbol{x}}(\tau)$ を比較的安定して求めることができる．しかしながら，第2段階では最小二乗誤差基準を使わざるを得ず，推定量が不安定になってしまう．[Takeuchi 02, Kanamori 06] などの研究により，誤差項の不均一性に関して特定の仮定が可能な場合，(3.1) 式の右辺第2項の推定が容易になることが示されている．

加法誤差モデル

加法誤差 (Additive Noise) モデルは次のように定式化される：

$$Y = f(\boldsymbol{x}) + c \cdot e, \tag{3.2}$$

ここで，$c \in \mathbb{R}^+$ は正のスカラー定数，e は $\boldsymbol{x}$ に依存しない平均0，分散1の確率変数を表す．これは，誤差項が均一分散 (homoscedastic) である場合に相当している．このとき，条件付平均関数は

$$E[Y|\boldsymbol{x}] = F^{-1}_{Y|\boldsymbol{x}}(\tau) + cF^{-1}_{e}(\tau) \tag{3.3}$$

と表され，(3.1) 式第2項は $\boldsymbol{x}$ に依存しない定数となる．すなわち，不安定な最小二乗法で推定されるのは1つのパラメータのみでよい．

加法乗法誤差モデル

加法乗法誤差モデル (Additive and Multiplicative Noise) モデルは次のように定式化される：

$$Y = f(\boldsymbol{x}) + \{c + d \cdot f(\boldsymbol{x})\} \cdot e, \tag{3.4}$$

ここで，$c, d \in [0, \infty)$ は $c \geq 0, d \geq 0, (c, d) \neq (0, 0)$ となるようなスカラー定数，e は $\boldsymbol{x}$ に依存しない平均0，分散1の確率変数を表す．このとき，条件付平均関数は

$$E[Y|\boldsymbol{x}] = F^{-1}_{Y|\boldsymbol{x}}(\tau) + (c + dE[Y|\boldsymbol{x}])F^{-1}_{e}(\tau)$$

と表される．これを $E[Y|\boldsymbol{x}]$ についてまとめると，

$$E[Y|\boldsymbol{x}] \;=\; \frac{1}{1 - dF^{-1}_{e}(\tau)} F^{-1}_{Y|\boldsymbol{x}}(\tau) + \frac{cF^{-1}_{e}(\tau)}{1 - dF^{-1}_{e}(\tau)} \tag{3.5}$$

と表され，第1段階で推定された条件付分位点 $F^{-1}_{Y|\boldsymbol{x}}(\tau)$ のアフィン関数として条件付平均が求められる．すなわち，第2段階では2つパラメータのみを最小二乗法で推定すればよい．

誤差項 e の分布が非対称で正方向に長い裾を持つとき，誤差項の不均一性に関する (3.2) 式や (3.4) 式のような仮定が正しければ，(3.3) 式や (3.5) 式を利用した二段階推定量は汎化誤差を小さくできることが理論的に検証されている．

3.2　二段階回帰木

3.1節で考察した二段階推定法では，特徴空間$\mathcal{X}$全体にて誤差構造が均一であると仮定している．損害保険の支払金額は，分布形状そのものがリスクファクターに依存する場合も多く，上述の仮定は現実を反映していない．本論文では，特徴空間全体に共通の誤差構造を仮定するのではなく，局所的に誤差構造が均一であるとみなすアプローチを考察する．特徴空間$\mathcal{X}$をいくつかの領域に分割し，各領域での誤差構造が均一であるとみなすことができれば，各領域にて(3.3)式や(3.5)式のような補正をすることが可能となる．しかしながら，分割する領域を増やすほど第2段階の最小二乗法で推定するパラメータ数が増え推定量が不安定になる．回帰木によるモデリングは適切な特徴空間の分割を行う枠組を提供するため，本論文では，回帰木による上述のアプローチの実装を行う．

第1段階では適当な分位点$\tau \in (0,1)$を選択し分位点回帰木を作成する．まず，学習データに対し(2.2)式のτ分位点誤差を最小にする回帰木を可能なかぎりの深さまで作成する．続いて，交差検証法により最適な剪定パラメータを選択し，2.2節で概説した規則により剪定を行う．第2段階で作成される回帰木は条件付分位点モデルから条件付平均モデルへのバイアス修正に用いられる．誤差構造に加法誤差モデル (3.2)式を用いる場合には(3.3)式の加法修正 (additive correction) が，加法乗法誤差モデル (3.4)式を用いる場合には(3.5)式のアフィン修正 (affine corection) が行われる．条件付分位点関数は安定して推定できるが，バイアス修正は不安定な最小二乗法で行わざるを得ないため，第1段階では深い回帰木 (複雑なモデル) を用い，第2段階では浅い回帰木 (単純なモデル) を用いることが重要となる．もし第2段階で第1段階よりも深い木を作成してしまうと (第1段階の影響が消えて) 最小二乗回帰木と同じ不安定性を有してしまう．本論文で提案する二段階剪定アプローチでは第1段階で作成された木をさらに剪定することにより第2段階でより浅い木が作成されることを保証する．さらに，二段階で剪定を行うと，第2段階で作成される回帰木が第1段階で作成される回帰木の一部を共有するため，同定されたリスクグループの解釈が容易となる．図3.1に4.3節で考察する自動車保険データ318,564件に対して作成された二段階回帰木 ($\tau = 0.7$, アフィン修正) が例示されている．

3.3　二段階回帰木のアルゴリズム

二段階回帰木アルゴリズムは以下のように整理される：

- アルゴリズムへの入力は次のようになる：

入力：学習データ $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$，分位点 $\tau \in (0,1)$，修正オプション CO $\in$ {additive, affine}；

分位点 τ の選択に関しては第4章で実例を交えて考察する．修正オプションは，局所的な誤差構造として加法誤差モデルを想定するか，加法乗法誤差モデルを想定するかの選択に相当している．

- 第1段階は次のような手順をとる：

第1段階：学習データ $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ に関し (2.2) 式の分位点回帰誤差を最小にする回帰木 Q_τ を作成する．

ここでは通常の分位点回帰木の生成を行う．分位点誤差基準に関して最適な最大の回帰木を作成した後，交差検証法を用いて剪定を行う．

- 第2段階では第1段階で作成した分位点回帰木 Q_τ をさらに剪定する．修正オプションによって操作の詳細が異なる．修正オプションが CO = additive の場合は次のような手順をとる：

第2段階 (CO=additive)：Q_τ をさらに剪定する．剪定は次の基準を交差検証法の評価データに関して最小化するように行われる：

$$\sum_{\ell \in \mathcal{L}} \sum_{i \in \ell} [y_i - \{\hat{a}_\ell + Q_\tau(\boldsymbol{x}_i)\}]^2,$$

ここで，$\mathcal{L}$ は木に含まれる葉全体の集合を，ℓ は個々の葉を，$i \in \ell$ は葉 ℓ に分類されるデータの添字を表している．$\hat{a}_\ell$ は葉 ℓ の加法修正パラメータで葉 ℓ に集まる学習データに関して基準を最小化する値に設定されている．$Q_\tau(\boldsymbol{x}_i)$ は τ 分位点回帰木 Q_τ に特徴 $\boldsymbol{x}_i$ を与えたときの予測値を表している．

各葉の加法修正パラメータ $\hat{a}_\ell$ は (3.3) 式の $cF_e^{-1}(\tau)$ の最小二乗推定値となっている．

- 修正オプションが CO = affine の場合は次のような手順をとる：

第2段階 (CO=affine)：Q_τ をさらに剪定する．剪定は次の基準を交差検証法の評価データに関して最小化するように行われる：

$$\sum_{\ell\in\mathcal{L}}\sum_{i\in\ell}[y_i-\{\hat{a}_\ell+\hat{b}_\ell Q_\tau(\boldsymbol{x}_i)\}]^2,$$

ここで，$\hat{a}_\ell$,$\hat{b}_\ell$ は葉 ℓ のアフィン修正パラメータで葉 ℓ に集まる学習データに関して基準を最小化する値に設定されている．

各葉のアフィン修正パラメータ $(\hat{a}_\ell,\hat{b}_\ell)$ はそれぞれ (3.5) 式の $\frac{cF_e^{-1}(\tau)}{1-dF_e^{-1}(\tau)}$ ならびに $\frac{1}{1-dF_e^{-1}(\tau)}$ の最小二乗推定値となっている．

- アルゴリズムの出力は次のようになる．

出力：分位点回帰木 Q_τ，バイアス修正用回帰木 (C と表記)；

Q_τ は第1段階で得られる回帰木，C は第2段階で Q_τ をさらに剪定して得られる回帰木である．

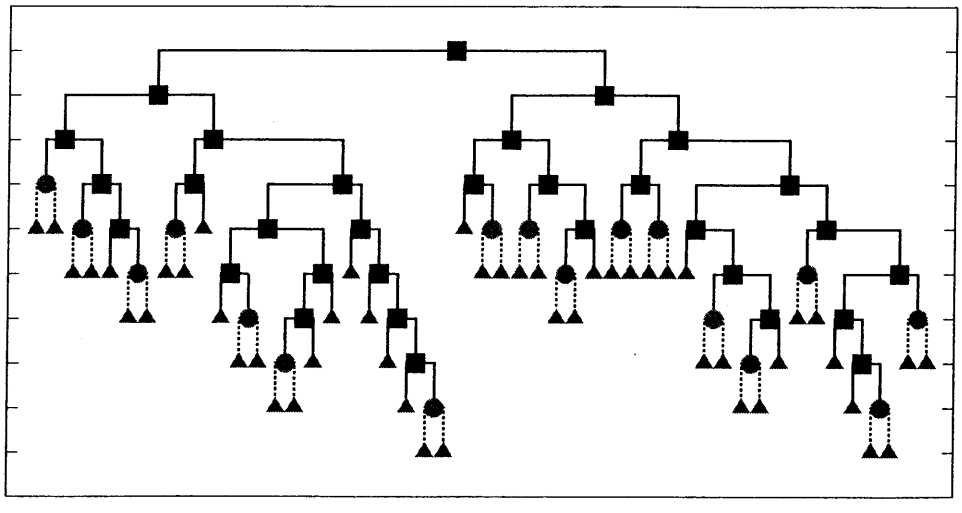


図　3.1: **自動車保険契約 318,564 件に対して作成された二段階回帰木**：実線ならびに点線で表された枝全体から構成される回帰木が第1段階で作成された 0.7 分位点回帰木を表し，実線で表された枝のみから構成される回帰木が第2段階で作成されたアフィン修正用の回帰木を表している．△は第1段階の 0.7 分位点回帰木の葉を，○は第2段階で再剪定されてアフィン修正用回帰木の葉となったノードを，□は両者共通のノードを表している．

リスクファクター $\boldsymbol{x}_{\text{new}} \in \mathcal{X}$ を持つ新しい保険契約の平均支払金額は以下の手順で推定される．まず，$\boldsymbol{x}_{\text{new}}$ を分位点回帰木 Q_τ のいずれかの葉に分類する．分類された葉に集まった学習データのサンプル τ 分位点を $\hat{F}^{-1}_{Y|\boldsymbol{x}_{\text{new}}}(\tau)$ とする．続いて，$\boldsymbol{x}_{\text{new}}$ をバイアス修正用回帰木 C のいずれかの葉に分類する．分類された葉の修正パラメータ (CO = additive の場合，$\hat{a}_\ell$，CO = affine の場合，$\hat{a}_\ell, \hat{b}_\ell$，) を利用し，加法修正の場合には

$$\hat{E}[Y|\boldsymbol{x}_0] = \hat{a}_\ell + \hat{F}^{-1}_{Y|\boldsymbol{x}_{\text{new}}}(\tau),$$

アフィン修正の場合には

$$\hat{E}[Y|\boldsymbol{x}_0] = \hat{a}_\ell + \hat{b}_\ell\, \hat{F}^{-1}_{Y|\boldsymbol{x}_{\text{new}}}(\tau),$$

により条件付平均値，すなわち，平均支払金額を推定する．

第4章

数値例

本章では数値実験により二段階回帰木の検証を行う．まず，加法誤差構造，加法乗法誤差構造を持つ人工データに対して本手法を適用したシミュレーションについて報告する．続いて，北米の某損害保険会社から提供された自動車保険データ318,564件に本手法を適用した結果を示す．

4.1　実験の概要

本実験では，最小二乗誤差基準による回帰木(LeastSquareと表記)，対数正規分布を仮定し剪定を用いる回帰木(LogNormalと表記)，ならびに[Apte 99]による回帰木(ProbEと表記)[1]を比較対象とした．第1段階にτ分位点回帰木を作成し，第2段階で加法修正ならびにアフィン修正を行った二段階回帰木をそれぞれAdd-τ，Aff-τと表記する．さらに，文献[Takeuchi 02]ならびに[Kanamori 06]のアプローチとの比較も行った．これらのアプローチは第2段階において特徴空間全体に共通な加法修正やアフィン修正を行う場合，すなわち，第2段階の修正木がルートノードのみから構成される場合に対応している．このため，加法修正，アフィン修正のものそれぞれをAdd-Root-τ，Aff-Root-τと表記する．

4.2　人工データによるシミュレーション例

4.2.1　人工データの発生方法

学習データ$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$は以下のように作成した．$p$次元入力データ$\boldsymbol{x}_i$は$\{0, 1, \cdots, m\}^p$を母集団とする$p$次元離散一様分布から発生させた．出力データ$y_i$は(3.4)式:$y_i = f(\boldsymbol{x}_i)+\{c+d\cdot f(\boldsymbol{x}_i)\}\cdot e$

[1]対数正規分布を仮定した回帰木で保険数理的信頼度基準と呼ばれるものを用いて分割を途中停止するアルゴリズムで詳細は[Apte 99]を参照のされたい．

により作成した．条件付平均関数は共通の係数 $\beta \in \mathbb{R}$ を持つ線形関数 $f(\boldsymbol{x}) = \beta \sum_{j=1}^{p} x_{ij}$ とした．誤差項 e は，図2.1の残差の経験分布関数を分散が1になるよう正規化したものから発生させた．誤差構造は加法誤差構造 $(c = 1, d = 0)$，ならびに，加法乗法誤差構造 $(c = 0.5, d = 1/\beta pm)$ の二通りの場合を考察した[2]．学習データ数は $n = 10,000$，入力変数次元は $p = 2$，入力データの最大値は $m = 9$，線形モデルの係数は $\beta = 0.1$ とした．

4.2.2　実験結果と考察

上記の学習データを用いた回帰木を作成し，ノイズ成分のないテストデータ1,000,000個に対して評価するプロセスを異なる乱数のシードで10回行った．評価値に平均二乗誤差 (MSE) を用いた．図4.1は加法誤差構造の場合の，図4.2は加法乗法誤差構造の場合の実験結果を示している[3]．

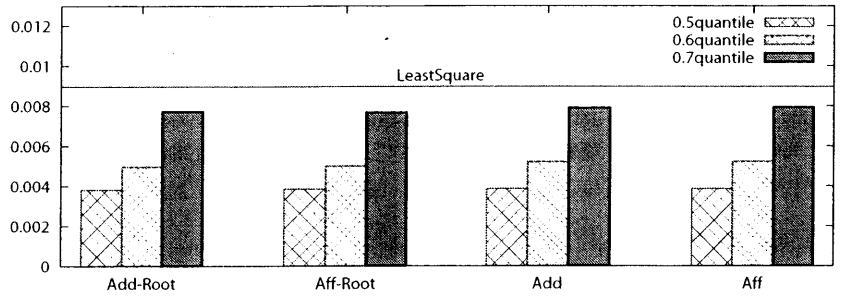


図 4.1: シミュレーション結果 (加法誤差構造の場合)

対数正規尤度最大化の分割基準を用いたLogNormal回帰木とProbE回帰木はLeastSquare回帰木よりも大幅に悪い結果となった．これは誤差項の分布が対数正規分布よりも裾の長い分布を持つ (図2.1参照) ためと推察される．Add-Root-τ 回帰木は，4.1の加法誤差構造に基づくデータの場合には，その仮定が正しいため，対応するAdd-τ 回帰木よりも誤差が小さくなっているが，図4.2の加法乗法誤差構造に基づくデータの場合には，その仮定が満たされないため，$(\tau = 0.5, 0.6$ のとき$)$ 誤差が大きく増大している．Aff-Root-τ 回帰木は，どちらの誤差構造の場合でもその仮定が正しいため，対応するAff-τ 回帰木よりも誤差が小さくなっている．図4.1

[2]加法乗法誤差構造の場合のパラメータ d は，$f(\boldsymbol{x}_i)$ に比例して誤差項の標準偏差が0.5から1.5の間を変化するように定めた．

[3]LogNormal回帰木，ProbE回帰木のMSEは，加法誤差構造ではそれぞれ0.019，0.038，加法乗法誤差構造ではそれぞれ0.014，0.038となり，ベンチマークのLeastSquare回帰木よりも大幅に誤差が大きくなった．

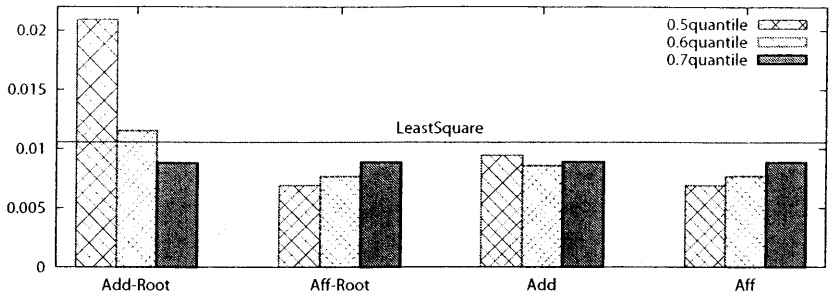


図 4.2: シミュレーション結果 (加法乗法誤差構造の場合)

の加法誤差構造の場合，$\tau = 0.5, 0.6, 0.7$ の順に，誤差が増加している．加法誤差構造では第 1 段階で推定される条件付平均関数の '形状' は任意の条件付分位点関数のものと同じであり，'位置' のみが第 2 段階で推定される．ゆえに，第 1 段階はより密度の大きい $\tau = 0.5$ 付近で推定するのが好ましく，このような結果が得られたものと思われる．一方，図 4.2 の加法乗法誤差構造では，第 1 段階で推定した条件付分位点関数のアフィン修正をするため，第 2 段階の不安定性が推定量全体に大きく影響を与える．そのため，修正量の少ない $\tau = 0.7$ の場合に，比較的よい結果が得られている[4].

4.3　自動車保険データへの適用例

4.3.1　自動車保険データの詳細

北米の某損害保険会社より提供された 318,564 件の自動車保険契約履歴を用いた[5]．利用可能なリスクファクター 30 種類のうち，使用地域，使用目的，等級，契約車の年式，運転者が契約車の所有者か否か，免許証の種類，運転者の年齢，性別，運転者の交通違反回数，事故歴，免許停止回数の 11 種類を用いた．各リスクファクターは，バイナリ変数，順序なしカテゴリ変数ならびに順序付きカテゴリ変数のいずれかとしてコーディングした．なお，保険金額の単位はカナダドル (CAD) である．

[4]本シミュレーションで誤差分布として用いた図 2.1 の残差分布では，0.5 分位点 $= -0.2919$，0.6 分位点 $= -0.1739$，0.7 分位点 $= -0.0041$ であった．残差分布の平均が 0 であることを考慮すると，条件付 0.7 分位点関数と条件付平均関数はほぼ一致しており，$\tau = 0.7$ の場合には，第 2 段階の修正量が少なくて良いことがわかる．

[5]提供された保険契約レコードは約 7,000,000 件で，そのうち，支払のあった事例 318,564 件を実験で用いた．

4.3.2　実験結果と考察

本実験ではランダム抽出したn件を学習データとして回帰木を作成し残りの$318,564-n$件に対して評価するプロセスを異なる乱数のシードで10回行った．評価値にMSEを用いた．表4.1に$n=100,000$の場合の結果を示す．各回帰木の評価値からLeastSquare回帰木の評価値を引いた値が示されている．$n=10,000$の場合もほぼ同様の結果が得られたが，$n=5,000$の場合は提案手法とLeastSquare回帰木に有意な$(p<0.05)$差はみられなかった．$n=10,000$においては，ニューラルネット，カーネルマシンとの比較も行った．両者ともに，L_2誤差基準ならびにL_1誤差基準を用いた場合の推定量を構築した．ニューラルネットは標準的な3層パーセプトロンを用い，そのキャパシティは，交差検証法で制御した．カーネルマシンのL_1回帰推定量はϵ-insensitive SV regressionのϵを0としたものと等しい．カーネルマシンのL_2回帰推定量はカーネルリッジ回帰 [Saunders 98] と呼ばれる．ニューラルネット，カーネルマシンともに，L_1回帰推定量は大きな下方バイアスを生じた．L_2回帰推定量も提案する二段階回帰木よりも有意に悪い$(p<0.05)$結果となった．

シミュレーションの場合と同様に，LogNormal回帰木とProbE回帰木はLeastSquare回帰木よりも有意に悪い結果となった．特徴空間全体に共通の誤差構造を仮定したAdd-Root-τならびにAff-Root-τは，Add-τならびにAff-τと比べて悪い性能を示した．図2.1の残差分布では，0.5分位点$=-0.2919$，0.6分位点$=-0.1739$，0.7分位点$=-0.0041$であった．残差の平均値は0であるので，第1段階で$\tau=0.5$あるいは0.6とする場合，第2段階で推定する修正量が$\tau=0.7$の場合に比べて大きくなると考えられる．$\tau=0.5,0.6$のときにAdd-Root-τやAff-Root-τの性能がAdd-τやAff-τに比べて悪いのは，特徴空間全体が共通の誤差構造を持つという仮定が不適切であることを示唆しており，二段階決定木のように局所的な誤差構造を同定するアプローチが有効であると思われる．$\tau=0.5$とする二段階決定木の性能が$\tau=0.6$や0.7の場合と比べて悪いことから，第1段階におけるτの選択が重要であることを示唆している．これには残差を用いて条件付平均に近い分位点τを推定するアプローチなどが有用であると思われる．

表 4.1: **自動車保険データにおける実験結果**：“MSE Diff.” は各回帰木の平均二乗誤差 (MSE) から LeastSquare 回帰木の MSE を引いた値を表している．“Signif.” は paired-sample t-test に基づく差の有意性を表しており，○○○ (●●●) は有意水準 0.01 で，○○ (●●) は有意水準 0.05 で，各回帰木の MSE が LeastSquare 回帰木の MSE よりも小さい (大きい) ことを表している．○ (●) は，前者が後者より小さい (大きい) ものの，有意な差がみられないことを表している．“Ave. #(leaf)” は各回帰木の葉数，すなわち，同定されたリスクグループ数の平均値を表している．二段階回帰木では，第 1 段階の葉数を左列に，第 2 段階の葉数を右列に表示している．

Tree-type	MSE Diff.	Signif.	Ave. #(leaf)	
LeastSquare	0000.00 ± 000.00	N.A.	14.2	
LogNormal	+8203.32 ± 460.72	●●●	9.5	
ProbE	+6294.35 ± 447.75	●●●	12.7	
Add-Root-0.5	+6257.58 ± 437.52	●●●	17.7	1.0
Add-Root-0.6	−77.34 ± 394.38	○	22.1	1.0
Add-Root-0.7	−951.15 ± 415.42	○○	20.7	1.0
Aff-Root-0.5	+3423.58 ± 422.42	●●●	17.7	1.0
Aff-Root-0.6	−78.22 ± 413.02	○	22.1	1.0
Aff-Root-0.7	−2175.03 ± 384.63	○○○	20.7	1.0
Add-0.5	−697.47 ± 358.89	○	17.7	11.4
Add-0.6	−2125.22 ± 374.77	○○○	22.1	10.7
Add-0.7	- 2814.26 ± 356.29	○○○	20.7	18.5
Aff-0.5	−438.75 ± 367.21	○	17.7	9.9
Aff-0.6	−1927.80 ± 376.79	○○○	22.1	12.2
Aff-0.7	−2734.16 ± 364.01	○○○	20.7	7.8

第5章

おわりに

本論文では損害保険の純保険料推定に特化した回帰木を考察した．提案した二段階回帰木は，二段階推定を行なう．第1段階にて条件付分位点関数を推定し，第2段階にて条件付平均関数とのバイアスを修正する．条件付分位点関数は安定して推定できるが，バイアス修正は不安定な最小二乗法で行わざるを得ないため，第1段階では深い回帰木 (複雑なモデル) を用い，第2段階では浅い回帰木 (単純なモデル) を用いることが重要となる．本論文で提案した二段階の剪定を行う方法は，このような枠組のひとつの実装として位置づけられる．シミュレーションと自動車保険データへの適用の結果から提案するアプローチが従来のものと比べて性能が良いことを例証した．自動車保険データへ適用した結果，一般的な回帰木である `LeastSquare` 回帰木の root MSE は 2329.29CAD，最高の性能を示した `Add-0.7` 回帰木の root MSE は 2328.68CAD であった．これは，1支払契約あたりの純保険料推定誤差を約 0.61CAD 程度減らせることを意味しており，損害保険会社のビジネス上メリットとしても有意であると思われる．

参考文献

[Apte 99] Apte, C., Grossman, E., Pednault, E. P. D., Rosen, B. K., Tipu, F. A., and White, B.: Probabilistic estimation-based data mining for discovering insurance risks, *IEEE Intelligent Systems*, Vol. 14, No. 6, pp. 49–58 (1999)

[Box 64] Box, G. E. P. and Cox, D. R.: An Analysis of Transformations, *Journal of the Royal Statistical Society, Ser. B*, Vol. 26, pp. 211–246 (1964)

[Breiman 84] Breiman, L., Friedman, J., Olshen, R., and Stone, C.: *Classification and regression trees*, Wadsworth, Monterrey, CA (1984)

[Duan 83] Duan, N.: Smearing estimate: a nonparametric retransformation method, *Journal of the American Statistical Association*, Vol. 78, No. 383, pp. 605–610 (1983)

[Dugas 04] Dugas, C., Chapados, N., Bengio, Y., Vincent, P., Denoncourt, G., and Fournier, C.: Neural network applied to automobile insurance ratemaking, in Jain, L. and Shapiro, A. F. eds., *Intelligent and Other Computational Techniques in Insurance: Theory and Applications*, chapter 4, pp. 137–192, World Scientific (2004)

[Haberman 96] Haberman, S. and Renshaw, A. E.: Generalized linear models and actuarial science, *The Statistician*, Vol. 45, No. 4, pp. 407–436 (1996)

[Huber 82] Huber, P. J.: *Robust Statistics*, John Wiley & Sons Inc. (1982)

[Jacobs 91] Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E.: Adaptive mixtures of local experts, *Neural Computation*, Vol. 3, pp. 79–87 (1991)

[Kanamori 06] Kanamori, T. and Takeuchi, I.: Conditional mean estimation under asymmetric and heteroscedastic error by linear combination of quantile regressions, *Computational Statistics and Data Analysis*, Vol. 50, No. 12, pp. 3605–3618 (2006)

[Koenker 78] Koenker, R. and Bassett, G.: Regression quantiles, *Econometrica*, Vol. 46. No. 1, pp. 33–50 (1978)

[Koenker 94] Koenker, R., Ng, P., and Portnoy, S.: Quantile smoothing splines, *Biometrika*, Vol. 81, pp. 673–680 (1994)

[Koenker 05] Koenker, R.: *Quantile regression*, Cambridge University Press (2005)

[McCullaugh 89] McCullaugh, P. and Nelder, J. A.: *Generalized linear models*, Chapman and Hall (1989)

[Miller 84] Miller, D. M.: Reducing transformation bias in curve fitting, *The American Statistician*, Vol. 38, pp. 124–126 (1984)

[Quinlan 93] Quinlan, J. R.: *C4.5: programs for machine learning*, Morgan Kaufmann (1993)

[Rousseeuw 87] Rousseeuw, P. J. and Leroy, A. M.: *Robust regression and outlier detection*, John Wiley & Sons Inc. (1987)

[Saunders 98] Saunders, C., Gammerman, A., and Vork, V.: Ridge Regression Learning Algorithm in Dual Variables, in *Proceedings of the 15-th International Conference on Machhine Learning*, pp. 515–521 (1998)

[Takeuchi 02] Takeuchi, I., Bengio, Y., and Kanamori, T.: Robust regression with asymmetric heavy-tail noise distributions, *Neural Computation*, Vol. 14, No. 10, pp. 2469–2496 (2002)

[Takeuchi 06] Takeuchi, I., Le, Q. V., Sears, T., and Smola, A. J.: Nonparametric quantile estimation, *Journal of Machine Learning Research*, Vol. 7, pp. 1231–1264 (2006)

謝辞

本研究を進めるにあたり，終始適切な御指導を賜わりました，三重大学工学部教授 成瀬 央先生に深く感謝いたします．本研究に対し貴重な御助言を賜わりました，三重大学工学部助教授 児玉 哲司先生に厚く御礼申し上げます．また，終始一貫して熱心な御指導を賜わりました，三重大学工学部助手 竹内 一郎先生に厚く御礼申し上げます．学生生活を共にした，織田君，金谷君，辻川君，吉川君そして院生１年の皆をはじめ，４年の皆，成瀬研究室の皆様に感謝します．最後に，末筆ながら大学における日頃の研究活動を暖かく見守ってくれた家族に感謝します．

発表論文リスト

学術論文

(1) 西久美子，竹内一郎，"二段階回帰木による損害保険の純保険料推定"，人工知能学会論文誌，Vol.22，No.2，pp183-190(2007)

国際会議

(1) Kumiko Nishi，Ichiro Takeuchi，"Casuality Insurance Pure Premium Estimation Using Two-Stage Regression Tree"，The International Workshop on Data-Mining and Statistical Science，September 25-26，2006，Century Royal Hotel，Sapporo，Japan

国内会議

(1) 西久美子，竹内一郎，"決定木による損害保険データ解析に関する一考察"，平成18年電気関係学会東海支部連合大会，9月28日-29日，2006，岐阜大学

(2) 西久美子，竹内一郎，"決定木による損害保険データ解析に関する一考察"，第20回東海ファジィ研究会，2月26日-27日，2006，日間賀島公民館