—— 編集委員会依頼論文 ——

健康保険データに基づく医療費予測モデリング —正則化 two-part モデルによるアプローチ—

小暮 厚之 * 小林凌雅 ‡

2018 年 8 月 31 日投稿

概要

本論文では，健康保険データに基づく医療費の予測モデリングについて考察する．特に，医療費の代表的な統計モデルである two-part モデルを取り上げ，ラッソのような正則化手法の適用により予測精度が高まるか否かを検討する，わが国の健康保険データへ応用した結果，説明変数の個数が大きい高次元の場合には，正則化により予測精度が向上する可能性が示された．

キーワード：医療費の予測，健康保険データ，正則化，two-part モデル，スパース性，ラッソ

1 はじめに

本論文では，健康保険データに基づく医療費の予測モデリングについて考察する．特に，医療費の代表的な統計モデルである two-part モデルを取り上げ，ラッソのような正則化手法の適用により予測精度が高まるか否かを検討する，ここで用いるデータは，わが国の複数の健康保険の組合員の集団から無作為に抽出された 10,000 人の 2010 年から 2012 年までの 3 年間のレセプト及び健康診断のデータである．まず，2010 年の変数を説明変数，2011 年の総医療費を目的変数として予測モデルを構築した．次に，この予測モデルの評価を行うために，構築した予測モデルの説明変数に 2011 年の値を代入して，2012 年の医療費を予測した．この結果，説明変数の個数が大きい高次元の場合には，two-part モデルを正則化することにより，予測精度が向上する可能性が示された．

正則化による医療費の予測は，Loginov et. al (2013) によって既に報告されている．ただし，Loginov et. al (2013) では，通常の回帰モデルに正則化を適用している．本稿では，two-part モデルに対する正則化の適用について議論し，ラッソ法 (Tibshirani, R., 1996) と弾性ネット法 (Zhou and Hastie, 2005) の 2 種類のスパース正則化の適用を考察した．また，スパース性をもたない伝統的な正則化手法であるリッジ法も試みた．

本論文の構成は以下の通りである．第 2 節では，医療費の代表的な統計モデルであるトービット法，標本選択法，two-part モデルについて概説する．第 3 節では two-part モデルの推定と予測について詳述する．第 4 節では，予測モデリングにおけるバイアス-分散トレードオフについて注意を与える．第 5 節ではラッソ法を中心として two-part モデルへの正則化の適用を議論し，第 6 節でわが国健康保険データへの応用を考察する．第 7 節で結語を述べる．

* 東京経済大学 経営学部 e-mail: kogure@tku.ac.jp
† 慶應義塾大学大学院 政策・メディア研究科博士課程
‡ 本論文は，株式会社 JMDC との共同研究を通じて作成されたものであり，データの提供を含め同社から多大な支援を受けた．ここに深く感謝申し上げます．

2　医療費の統計モデル

2.1　選択バイアス

n 人の加入者からなる集団を考える．ある年における個人 i の医療費 Y_i は，その年に受けたすべての診察における医療費の合計額

$$Y_i = \sum_{j=1}^{N_i} y_{ij}$$

と考えられる．ここで，y_{ij} は個人 i の j 回目の（健康保険による）受診の医療費であり，N_i は受診件数である．受診件数 N_i がゼロの場合には，総医療費もゼロとなる．そのため，医療費がゼロとなる観察値がデータに多く含まれることが一般的である．例えば，2011 年度の医療費は 10,000 人のうち 1,646 人がゼロである．このような場合，医療費がゼロとなるデータを除いて分析すると，いわゆる選択バイアスを生じる可能性がある．

例えば，「真の医療費」（医療費に対する需要）Y^* と健康診断の検査項目の数値 x との関係が

$$Y^* = \alpha^* + \beta^* x + \varepsilon$$

とする．ここで，β^* は x が Y^* に与える影響を表すパラメータであり，ε は誤差項を表す．このとき，医療費ゼロのデータを無視して，医療費が正のデータのみを使って

$$Y = \alpha + \beta x + \varepsilon$$

を推定すると，β^* と β は一般には一致しない．もしも医療費ゼロのデータ値が健診値 x と無関係にランダムに出現するのであれば，医療費ゼロのデータを除外しても問題はないであろう．しかし，健診値 x と相関がある変数（例えば所得）によってゼロが出現するのであれば，バイアスが生じる可能性がある．

2.2　トービット・モデル

被説明変数がゼロを含む正値である問題を扱う端緒となった回帰モデルはトービット・モデル（Tobin, 1958）である．個人 i の医療費 Y_i を p 個の説明変数

$$\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$$

によってモデル化する問題を考える．トービット・モデルでは，潜在変数 Y_i^* を想定する．Y_i^* は線形回帰モデル

$$Y_i^* = \boldsymbol{\beta}' \boldsymbol{x}_i + \varepsilon_i \tag{1}$$

に従っていると仮定する．ここで，$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ はパラメータ・ベクトルであり，ε_i は誤差項を表す．$Y_i^* > 0$ のときは，その値が医療費 y_i として記録され，$Y_i^* \leq 0$ のときは，医療費はゼロと記録されると考える．したがって，観察される医療費は

$$Y_i = \begin{cases} Y_i^* & Y_i^* > 0 \text{ の場合} \\ 0 & Y_i^* \leq 0 \text{ の場合} \end{cases}$$

と表せる．このとき，医療費の期待値は

$$\mathrm{E}[Y_i] = \mathrm{E}[Y_i^* | Y_i^* > 0] \Pr(Y_i^* > 0)$$

となる．

もしもゼロである医療費を無視して，正となる医療費のみによって期待値を計算すると

$$\mathrm{E}[Y_i|Y_i>0]=\mathrm{E}[Y_i^*|Y_i^*>0]>\mathrm{E}[Y_i^*|Y_i^*>0]\Pr(Y_i^*>0)=\mathrm{E}[Y_i]$$

となり，$\Pr(Y_i^*\leq 0)=\Pr(Y_i=0)$ の割合だけ真の医療費を過大に評価することになる．

より具体的に示すために，誤差項 ε_i が平均ゼロ，分散 σ^2 の正規分布に従うと仮定すると，

$$\mathrm{E}[Y_i|Y_i>0]=\boldsymbol{\beta}'\boldsymbol{x}_i+\sigma\frac{\phi(\boldsymbol{\beta}'\boldsymbol{x}_i/\sigma)}{1-\Phi(\boldsymbol{\beta}'\boldsymbol{x}_i/\sigma)}$$

となる．ここで，ϕ，Φ は標準正規分布の密度関数及び分布関数である．

一方，ゼロを無視しない場合の医療費の期待値は

$$\mathrm{E}[Y_i]=\Phi\left(\frac{\boldsymbol{\beta}'\boldsymbol{x}_i}{\sigma}\right)\boldsymbol{\beta}'\boldsymbol{x}_i+\sigma\phi(\boldsymbol{\beta}'\boldsymbol{x}_i/\sigma)$$

となる．$\boldsymbol{\beta}'\boldsymbol{x}_i/\sigma$ が大きくなる場合には

$$\phi(\boldsymbol{\beta}'\boldsymbol{x}_i/\sigma)\approx 0,\ \Phi(\boldsymbol{\beta}'\boldsymbol{x}_i/\sigma)\approx 1,\ \frac{\phi(\boldsymbol{\beta}'\boldsymbol{x}_i/\sigma)}{1-\Phi(\boldsymbol{\beta}'\boldsymbol{x}_i/\sigma)}\approx 0\quad(\boldsymbol{\beta}'\boldsymbol{x}_i/\sigma\to\infty)$$

となるため，$\mathrm{E}[Y_i|Y_i>0]$ と $\mathrm{E}[Y_i]$ のどちらも $\boldsymbol{\beta}'\boldsymbol{x}_i$ に近似的に等しくなる．しかし，一般には，医療費がゼロとなるデータを無視して推定を行う場合には，真の関係を見誤る恐れがある．

2.3 標本選択モデル

標本選択モデル (Heckman, 1979) では，トービット・モデルで用いた医療費の潜在変数 Y_i^* の回帰モデル (1) に加え，各個人が受診するか否かを選択するプロセスを導入する．そのため，医療費の潜在変数 Y_i^* に加え，受診するか否かの基準となる潜在変数 S_i^* を考え，真のモデルを

$$\begin{cases}Y_i^*=\boldsymbol{x}_i'\boldsymbol{\beta}+\varepsilon_i\\ S_i^*=\boldsymbol{x}_i'\boldsymbol{\gamma}+\eta_i\end{cases}\tag{2}$$

とする．ただし，誤差項は 2 変量正規分布

$$\begin{pmatrix}\varepsilon_i\\ \eta_i\end{pmatrix}\sim N\left(\begin{pmatrix}0\\ 0\end{pmatrix},\begin{pmatrix}\sigma^2 & \sigma\rho\\ \sigma\rho & 1\end{pmatrix}\right)$$

に従うものとする．ここで ρ は，ε_i と η_i の相関係数である．

$S_i^*>0$ のときは，Y_i^* の値が医療費 Y_i として記録され，$S_i^*\leq 0$ のときは，医療費はゼロと記録されると考える．すなわち

$$Y_i=\begin{cases}Y_i^* & S_i^*>0\text{ の場合}\\ 0 & S_i^*\leq 0\text{ の場合}\end{cases}=Y_i^*S_i$$

である．ここで，S_i は医療費がゼロか否かを表す 2 値変数であり，

$$S_i\equiv\begin{cases}1 & S_i^*>0\text{ の場合}\\ 0 & S_i^*\leq 0\text{ の場合}\end{cases}$$

とする．

2.4 two-part モデルによる予測

(2) において，$\beta = \gamma$，$\varepsilon_i = \eta_i$ であれば，$Y_i^* = S_i^*$ となり，標本選択モデルはトービット・モデル (1) となる．その逆に，もしも，β と γ に関数関係がなく，ε_i と η_i の相関係数 ρ の値がゼロならば，Y_i の分布関数は

$$\Pr(Y_i \le y) = \begin{cases} \Pr(Y_i \le y | S_i = 1)\Pr(S_i = 1) + \Pr(S_i = 0) & y > 0 \text{ の場合} \\ \Pr(S_i = 0) & y = 0 \text{ の場合} \\ 0 & y < 0 \text{ の場合} \end{cases} \tag{3}$$

となり，標本選択モデルは，S_i のモデルと $Y_i|S_i = 1$ のモデルの 2 つの部分に分解される．これを two-part モデルという．

3 two-part モデル

two-part モデルでは，受診するか否か（S_i）は被保険者の判断によるところが大きく，医療費の大きさ $(Y_i|S_i = 1)$ は医師の判断によるところが大きいと考え，S_i と $Y_i|S_i = 1$ を別々に扱う．このため．標本選択モデルに比べ，より自由度の高いモデリングが可能となる．本稿では two-part モデルを採用する．

3.1 尤度

two-part モデルは 2 段階で考えることができる．第 1 段階は，医療費が正値かゼロかを表す S_i の統計モデルである．第 2 段階は，医療費が正値であるという条件の下における医療費 $Y_i|S_i = 1$ の回帰モデルである．(3) より，$\boldsymbol{x}_i$ を所与とする $Y_i = y_i$ の条件付き分布は

$$f_Y(y_i; \boldsymbol{\theta}|\boldsymbol{x}_i) = \begin{cases} \Pr(Y_i = 0; \boldsymbol{\theta}_1|\boldsymbol{x}_i) & y_i = 0 \text{ の場合} \\ f_Y(y_i; \boldsymbol{\theta}_2|\boldsymbol{x}_i, y_i > 0)\Pr(Y_i > 0; \boldsymbol{\theta}_1|\boldsymbol{x}_i) & y_i > 0 \text{ の場合} \end{cases} \tag{4}$$

と表現できる．ここで，$\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ は，それぞれ第 1 段階と第 2 段階のパラメータ・ベクトルであり，$\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ とする．

本論文では (4) の第 1 段階の統計モデルとして，ロジスティック回帰

$$\Pr(Y_i > 0; \boldsymbol{\theta}_1|\boldsymbol{x}_i) = \frac{1}{1 + \exp\{-\boldsymbol{\theta}_1'\boldsymbol{x_i}\}}, \quad \Pr(Y_i = 0; \boldsymbol{\theta}_1|\boldsymbol{x}_i) = \frac{\exp\{-\boldsymbol{\theta}_1'\boldsymbol{x_i}\}}{1 + \exp\{-\boldsymbol{\theta}_1'\boldsymbol{x_i}\}}$$

を採用する．

(4) の第 2 段階の回帰モデル $f_Y(y_i; \boldsymbol{\theta}_2|\boldsymbol{x}_i, y_i > 0)$ に対して，いくつかの特定化が適用されてきた．最もよく採用される特定化は，通常の線形回帰モデル

$$Y_i = \boldsymbol{\theta}_2'\boldsymbol{x}_i + \varepsilon_i$$

を当てはめることである．ここで，ε_i は誤差項であり，平均ゼロ，分散が一定値 σ^2 である正規分布に互いに独立に従うと仮定する．

医療費の非正規性を考慮して，Y_i の対数値を被説明変数とする対数線形回帰モデル

$$\log Y_i = \boldsymbol{\theta}_2'\boldsymbol{x_i} + \varepsilon_i$$

を採用する場合も多い．

また，Y_i の変換を行う代わりに，一般化線形モデル (Blough, et al. 1999) を用いることもできる．$\mu_i = \mathrm{E}[Y_i]$ とするとき，自然な特定化は，リンク関数

$$\log(\mu_i) = \boldsymbol{\theta}_2'\boldsymbol{x}_i$$

を用いたガンマ回帰

$$f_Y(y_i;\boldsymbol{\theta}_2|\boldsymbol{x}_i,y>0)=\frac{(s/\mu_i)^s}{\Gamma(s)}y^{s-1}\exp\left(-\frac{sy}{\mu_i}\right)$$

である．ここで， s はガンマ分布の形状パラメータである．

3.2 予測

共変量 $\boldsymbol{x}$ に基づく Y の予測を $\hat{Y}$ とする．$\hat{Y}$ の予測誤差の基準として平均 2 乗誤差

$$\mathrm{E}\left[\left(Y-\hat{Y}\right)^2\right]$$

を用いると，最適な予測は $\boldsymbol{x}$ を所与とする Y の条件付き期待値

$$\begin{aligned}\mathrm{E}[Y;\boldsymbol{\theta}|\boldsymbol{x}]&=\mathrm{E}[Y;\boldsymbol{\theta}_2|Y>0,\boldsymbol{x}]\Pr(Y>0;\boldsymbol{\theta}_1|\boldsymbol{x})+\mathrm{E}[Y;\boldsymbol{\theta}_2|Y=0,\boldsymbol{x}]\Pr(Y=0;\boldsymbol{\theta}_1|\boldsymbol{x})\\&=\mathrm{E}[Y;\boldsymbol{\theta}_2|Y>0,\boldsymbol{x}]\Pr(Y>0;\boldsymbol{\theta}_1|\boldsymbol{x})\end{aligned}\tag{5}$$

で与えられる．

正規回帰モデル及びガンマ回帰モデルの場合, (5) は

$$\begin{aligned}\mathrm{E}[Y;\boldsymbol{\theta}|\boldsymbol{x}]&=\mathrm{E}[\boldsymbol{x}'\boldsymbol{\theta}_2+\varepsilon]\times\Pr(Y>0;\boldsymbol{\theta}_1|\boldsymbol{x})\\&=\frac{\boldsymbol{x}'\boldsymbol{\theta}_2}{1+\exp\{-\boldsymbol{x}'\boldsymbol{\theta}_1\}}\end{aligned}\tag{6}$$

となる．対数正規のケースは, (5) は

$$\begin{aligned}\mathrm{E}[Y;\boldsymbol{\theta}|\boldsymbol{x}]&=\mathrm{E}[\exp\{\boldsymbol{x}'\boldsymbol{\theta}_2\}\exp\{\varepsilon_i\}]\times\Pr(Y>0;\boldsymbol{\theta}_1|\boldsymbol{x})\\&=\frac{\exp\{\boldsymbol{x}'\boldsymbol{\theta}_2+\sigma^2/2\}}{1+\exp\{-\boldsymbol{x}'\boldsymbol{\theta}_1\}}\end{aligned}$$

となる．

3.3 推定

実際にはパラメータは未知であり，データから推定する必要がある．標準的な方法は尤度

$$\prod_{i=1}^n f_Y(y_i;\boldsymbol{\theta}|\boldsymbol{x_i})=\prod_{i=1}^n\left[\Pr(Y_i>0;\boldsymbol{\theta}_1|\boldsymbol{x}_i)^{b_i}\Pr(Y_i=0;\boldsymbol{\theta}_1\boldsymbol{x}_i)^{1-b_i}\right]\prod_{i\in N_1}f_Y(y_i;\boldsymbol{\theta}_2|\boldsymbol{x}_i,y_i>0)\tag{7}$$

の最大化である．ここで，b_i は以下のように定義される：

$$b_i\equiv\begin{cases}1 & y_i=0\text{ の場合}\\0 & y_i>0\text{ の場合}\end{cases}$$

また，N_1 は $y_i>0$ となる観測値の集まりを表す． このように，モデル全体の尤度は，第 1 段階の尤度 $\prod_{i=1}^n\left[\Pr(Y_i>0;\boldsymbol{\theta}_1|\boldsymbol{x}_i)^{b_i}\Pr(Y_i=0;\boldsymbol{\theta}_1|\boldsymbol{x}_i)^{1-b_i}\right]$ と第 2 段階の尤度 $\prod_{i\in N_1}f_Y(y_i;\boldsymbol{\theta}_2|\boldsymbol{x}_i,y_i>0)$ の 2 つに分かれるため，各部分ごとに尤度の最大化を行えばよい．

4 バイアス-分散トレードオフ

医療費の予測を行う場合，どの統計モデルを用いるべきかというモデル選択の問題が生じる．本稿では，「すべてのモデルは間違っている．しかし，中には役に立つものもある」という Box(1976) の有名な箴言の精神にしたがって，真のモデルの選択ではなく予測に最も寄与するモデルを選択することを考える．

今ある新しい $\boldsymbol{x}$ の値に対する目標変数 y の予測を考える．2 乗誤差を最小にするという意味で理論的に最適な予測は

$$y^* = \int yp(y|\boldsymbol{x})dy$$

である．ここで，$p(y|\boldsymbol{x})$ は $\boldsymbol{x}$ の条件つきの y の密度関数である．実際には，$p(y|\boldsymbol{x})$ は未知であり，y^* は計算できない．そこで，（間違っている可能性は承知の上で）操作しやすいモデル，例えば線形モデル

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p = \boldsymbol{\beta}'\boldsymbol{x}$$

を用いる．データ $\mathcal{D}$ に基づくパラメータベクトル $\boldsymbol{\beta}$ の推定値を $\hat{\boldsymbol{\beta}}$ とし，線形モデルによる予測を $\hat{y}_{\mathcal{D}} = \hat{\boldsymbol{\beta}}'\boldsymbol{x}$ とすると，その予測誤差は

$$(\hat{y} - y)^2$$

となる．$\hat{y}$ はデータ $\mathcal{D}$ に依存しているので，まず $\mathcal{D}$ に関して平均を取ると

$$\mathrm{E}_{\mathcal{D}}[(\hat{y}_{\mathcal{D}} - y)^2]$$

となる．さらに，y に関しても平均を取ると，最終的な予測誤差は

$$\mathrm{FPE} \equiv \int \mathrm{E}_{\mathcal{D}}[(\hat{y}_{\mathcal{D}} - y)^2]p(y|\boldsymbol{x})dy$$

となる．FPE は以下のように分解される

$$\mathrm{FPE} = \int (y - y^*)^2 p(y|\boldsymbol{x})dy + (y^* - \mathrm{E}_{\mathcal{D}}[\hat{y}_{\mathcal{D}}])^2 + \mathrm{Var}_{\mathcal{D}}(\hat{y})$$

ここで，$\int (y - y^*)^2 p(y|\boldsymbol{x})dy$ は，理論的に最適な予測を行ってもなお残るノイズである，ノイズは，データ分析者にはコントロールできない．$(y^* - \mathrm{E}_{\mathcal{D}}[\hat{y}_{\mathcal{D}}])$ は，理論的に最適な予測 y^* からの $\hat{y}_{\mathcal{D}}$ の平均的なバイアスを表す．分散 $\mathrm{Var}_{\mathcal{D}}(\hat{y}_{\mathcal{D}})$ は，$\hat{y}_{\mathcal{D}}$ の平均的な変動を表す．

説明変数を多く用いれば，バイアスを低下できるが，それは同時に分散が大きくなることも意味する．このバイアスと分散のトレードオフの釣り合いを取る予測が望ましい．

5　正則化

5.1　高次元データ

バイアスと分散のトレードオフの問題に対処する標準的な手法は，予測の対象である将来医療費に関連していると思われる重要な説明変数を選択することである．このアプローチでは，p 個の説明変数のすべての部分集合の各々にモデルをあてはめ，候補となる 2^p の組み合わせの中から AIC や BIC のような基準を用いて唯一のモデルを選択する．しかし，この方法では，p が大きくなるにつれて，計算的な困難さが増していく．p が大きい高次元データに対するアプローチとして，最近では次元縮小やモデル統合などの手法が用いられるようになってきた．次元縮小では，説明変数の $M(< p)$ 個の線形結合（加重和）を作成し，これらを新しい説明変数として回帰モデルを構築する．モデル統合では，単一のモデルを選択する代わりに，複数の回帰モデルを統合する．これらについては，安道 (2014)，James, et. al. (2013) 等 を参考されたい．本稿では，次項で説明する正則化という手法を採用する．

5.2　スパース正則化

正則化は，大きなパラメータの値に罰則を与えることによってパラメータの値を縮小させる手法である．ラッソ法 (Tibshirani, 1996) は，このような正則化手法の比較的最近のアプローチである．リッジ法のよう

な伝統的な方法と異なり，ラッソ法によって縮小されたパラメータの中には厳密にゼロとなるものが生じる．これはスパース性と呼ばれる性質であり，説明変数を選択する新しい方法を与える．このため，特に高次元のデータを扱うような場合に，従来のモデル選択法に代わって，ラッソ法が広く用いられるようになってきた．

正則化を適用するためには，パラメータが大きくなることに対する（負の）罰則項を対数尤度に追加すればよい．リッジ法では $\sum_{j=1}^{p}\theta_j^2$ という L_2 の罰則項を用いてきたが，ラッソ法では $\sum_{j=1}^{p}|\theta_j|$ という L_1 の罰則項を採用する．two-part モデルに対するラッソ正則化は，two-part モデルの尤度 (7) の対数に罰則項を加えた目的関数

$$\begin{aligned} l(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n}\log f_Y(y_i;\boldsymbol{\theta}|\boldsymbol{x_i}) - \pi_1(\boldsymbol{\theta}_1) - \pi_2(\boldsymbol{\theta}_2) \\ &= \frac{1}{n}\sum_{i=1}^{n}\log\left[\Pr(Y_i>0;\boldsymbol{\theta}_1|\boldsymbol{x}_i)^{d_1}\Pr(Y_i=0;\boldsymbol{\theta}_1|\boldsymbol{x}_i)^{1-d_i}\right] - \pi_1(\boldsymbol{\theta}_1) \\ &\quad + \frac{1}{n}\sum_{i\in N_1}\log f_Y(y_i;\boldsymbol{\theta}_2|\boldsymbol{x}_i,y_i>0) - \pi_2(\boldsymbol{\theta}_2) \end{aligned} \tag{8}$$

を最大化することによって適用される．ここで，$\pi_i(\boldsymbol{\theta}_i)$ は $\boldsymbol{\theta}_i=\{\theta_j^{(i)}, j=1,\ldots,p_i\}$ に対する罰則項である．

$$\pi_i(\boldsymbol{\theta}_i) = \lambda_i\sum_{j=1}^{p_i}|\theta_j^{(i)}|, \quad i=1,2.$$

と定義される．式中の λ_i は，正則化を調節する正の値であり，チューニング・パラメータと呼ばれる．チューニング・パラメータを選択する方法はいくつか提案されているが，ここではクロスバリデーション法を用いる．

オリジナルなラッソ法が提案されて以来，スパース正則化に関して多くの進展がなされてきた．$p>n$ という状況に対処するために，弾性ネット 法 (Zou and Hastie, 2005) は，罰則項

$$\pi(\boldsymbol{\theta}) = \lambda\left[\alpha\sum_{j=1}^{p}|\theta_j| + (1-\alpha)\sum_{j=1}^{p}\frac{\theta_j^2}{2}\right]$$

を用いる．ここで，α は 0 と 1 の間の値である．もしもその値が 1 ならば，弾性ネット法 は通常のラッソ法に一致する．また，その値が 0 ならば弾性ネット法はリッジ法となる．適応型ラッソ法 (Zou, 2006) は

$$\pi(\boldsymbol{\theta}) = \lambda\sum_{j=1}^{p}w_j|\theta_j|$$

という罰則項を用いる．ここで，w_j は

$$w_j = 1/|\hat{\theta}_j|^{\gamma},\ \gamma>0$$

であり，$\hat{\theta}_j$ は θ_j の一致推定量である．適応型ラッソを用いることにより，$\hat{\theta}_j$ の値が大きいパラメータに対する罰則が緩和される．Park and Casella (2008) は，伝統的なベイズ法の枠組みの下で．ラッソの罰則項は事前分布

$$f(\boldsymbol{\theta}|\sigma,\lambda) = \prod_{j=1}^{p}\frac{\lambda}{2}\exp\left\{-\lambda\frac{|\theta_j|}{\sigma}\right\}$$

に対応していることに着目し，ベイジアン・ラッソという手法を提案している．

6　わが国健康保険データへの応用

本節では，わが国健康保険データへの応用について述べる．複数の健康保険から無作為抽出した 10,000 人の加入者について 2010 年, 2011 年, 2012 年の各年度のレセプトデータ及び健診データを用いる．次節で述べ

る 27 個の変数を共変量として用いて医療費総額の予測モデルを構築する．ケース 1 では，27 個の共変量を説明変数とする場合 ($p = 27$) を考察する．ケース 2 では，27 個の共変量の間のすべての交互作用項を説明変数に加えた高次元データを用いる場合 ($p = 378$) を考える．

それぞれのケースにおいて第 1 段階がロジスティック回帰，第 2 段階が通常の線形モデルである two-part モデルを推定し，正則化を用いる場合と用いない場合の予測誤差を比較する．

6.1 共変量

2010 年, 2011 年, 2012 年の各年度に対して，以下の変数が 10,000 人の各加入者ごとに観察されている:

- 人口統計学的属性
  性別 (SEX), 年齢 (AGE)
- 健康診断項目
  BMI(BMI), 収縮期血圧 (SBP), 拡張期血圧 (DBP), 中性脂肪 (NF), HDL コレステロール (HDLC), LDL コレステロール (LDLC), GOT(GOT), GPT(GPT), γ-GT (GGT), 空腹時血糖 (FBS), HbA1c(NGSP 値)(HbA1c), 尿糖 (US)
- 医療費
  入院医療費 (IME), 通院医療費 (OME), 薬剤医療費 (PE), 入院日数 (HID), 外来日数 (OV), 救急有無 (EMC)
- 投薬の有無
  コレステロール投薬 (CM), 糖尿病投薬 (DM), 血圧投薬 (BPM)
- 疾病の有無
  高脂血症 (HL), 糖尿病 (DB), 高血圧症 (HBP), 肝疾患 (LD)

予測の対象である．医療費総額 (TME) は，入院医療費 (IME)，通院医療費 (OME)，薬剤医療費 (PE) の合計である．各年における医療費総額の基本統計量は表 1 に掲げられている．

	最小値	第 1 四分位	中央値	平均値	第 3 四分位	最大値	標準偏差
2010 年	0	5,590	29,820	91,920	92,950	12,570,000	278,740
2011 年	0	8,868	38,350	108,600	115,000	9,389,000	291,642
2012 年	0	9,178	39,640	117,100	120,800	6,767,000	308,818

表 1　医療費総額の要約

6.2 ケース 1

ケース 1 では前節で述べた変数のうち，通院医療費 (OME) を除く 27 個の変数を説明変数として用いる．

6.2.1 データ分析

第 1 段階のロジスティック回帰の被説明変数は，2011 年の医療費総額 TME.2011 が正値かゼロかを示す 2 値変数

$$S = \begin{cases} 1 & \text{TME.2011} > 0 \text{ の場合} \\ 0 & \text{TME.2011} = 0 \text{ の場合} \end{cases}$$

とし，説明変数は 2010 年度の 27 個の変数とする．その推定結果は付録の表 9 に示されている．第 2 段階の線形回帰は，TME.2011 が正であるデータのみを対象とし，その被説明変数は $Y =$ TME.2011，説明変数は

2010 年の 27 個の変数とする．その推定結果は付録の表 10 に示されている．

これらの推定結果から求めた予測モデル (6) の説明変数に 2011 年の値を代入することによって 2012 年の総医療費の予測値 PRED.2012 を求め，それらを実際の 2012 年の医療費 TME.2012 と比較した．ただし，予測値がマイナスになるときはゼロに置き換えた．その結果は表 2 に要約されている．

	最小値	第 1 四分位	中央値	平均値	第 3 四分位	最大値
PRED.2012	0	28,920	62,950	127,300	156,600	7,005,000
TME.2012	0	9,178	39,640	117,100	120,800	6,767,000

表 2　予測値 PRED.2012 と実績値 TME.2012 の要約統計量: ケース 1(ラッソ正則化を適用しない場合）

次にラッソ正則化を適用した．第 1 段階のロジスティック回帰では，クロスバリデーションによって選択された λ の値は非常に小さく (0.0004223471)，その推定結果は正則化なしの場合と同じであった． これとは対照的に，第 2 段階の線形回帰に対する λ の値は大きく (2555.633), 表 11 に示されているように，多くのパラメータの推定値がゼロとなった．図 1 は，様々な λ の値に対する平均 2 乗誤差の値を上下の 95% 信頼区間の値とともに示している．ただし横軸は λ の対数値である．

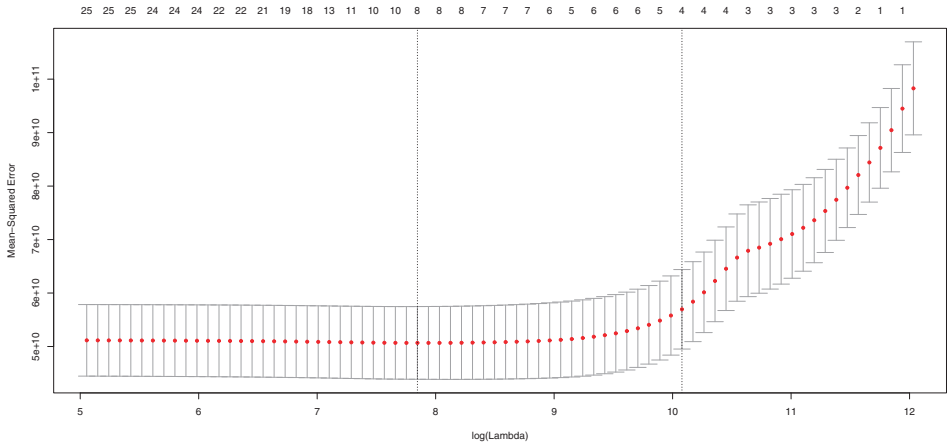

図 1　様々な λ の値に対する平均 2 乗誤差

前と同様にして予測値を求め，実績値と比較した．その要約は表 3 に掲げられている．

	最小値	第 1 四分位	中央値	平均値	第 3 四分位	最大値
PRED.2012	10140	27,760	63,360	127,100	157,400	6,678,000
TME.2012	0	9,178	39,640	117,100	120,800	6,767,000

表 3　予測値 PRED.2012 と実績値 TME.2012 の比較: ケース 1（ラッソ正則化を適用した場合）

6.2.2　予測誤差の比較

表 4 は，ラッソ正則化を適用した場合とそうでない場合の予測誤差を比較している．予測誤差の尺度として，平均 2 乗誤差

$$\mathrm{MSE} = \frac{1}{10000}\sum_{i=1}^{10000}(\mathrm{TME.2012}_i - \mathrm{PRED.2012}_i)^2,$$

を採用した．表にはリッジ法と弾性ネットを適用した場合の結果も掲げている．いずれの正則化法を用いても，正則化による予測精度の改善は小さい．特に，リッジ法を適用すると予測精度が低下してしまう．

	平均 2 乗誤差	平均 2 乗誤差 TME.2012 の分散
正則化なし	$(235623.3)^2$	0.5821
ラッソ	$(235016.2)^2$	0.5791
リッジ	$(237216.8)^2$	0.5900
弾性ネット（$\alpha = 0.5$）	$(237789.4)^2$	0.5793

表 4　予測誤差の比較：ケース 1

表の一番右の列には，TME.2012 の分散に対する平均 2 乗誤差の比が掲げられている．これは予測のターゲットである将来の医療費の変動に比べ，予測モデルで説明できなかった変動がどの程度の大きさを示している．

6.3　ケース 2

ここでは，非線形性の可能性を考慮して，ケース 1 で用いた 27 個の変数に加えて，27 変数の中のあらゆる 2 変数の組に対する交互作用を回帰式に取り込んだ．その結果，共変量の個数は 378 個となった．

6.3.1　データ分析

ラッソ正則化を用いない場合をまず考える．通常の線形回帰の決定係数の値は，ケース 1 の 0.5012 から 0.5992 に上昇した． 要約統計量は表 3 にまとめられている．

	最小値	第 1 四分位	中央値	平均値	第 3 四分位	最大値
PRED.2012	0	34,040	63,900	127,600	142,600	9,848,000
TME.2012	0	9,178	39,640	117,100	120,800	6,767,000

表 5　予測値 PRED.2012 と実績値 TME.2012 の比較: ケース 2（ラッソ正則化なし）

次にラッソ正則化を適用した．ただし，第 1 段階のロジスティック回帰にはラッソ正則化は施さずに，第 2 段階の線形回帰に対してのみラッソ正則化を施した．クロスバリデーションによる λ の値は非常に大きい (7974.432) く，その結果 378 個の説明変数のパラメータのうち 353 個の推定値がゼロとなった．図 2 は，様々な λ の値に対する平均 2 乗誤差の値を上下の 95% 信頼区間の値とともに示している．ただし横軸は λ の対数値である．

推定されたモデルを用いて各個人の医療費を予測し，予測誤差を比較した．要約統計量は表 6 に掲げられている．

	最小値	第 1 四分位	中央値	平均	第 3 四分位	最大値
PRED.2012	0	36,620	68,170	125,400	145,400	9,669,000
TME.2012	0	9,178	39,640	117,100	120,800	6,767,000

表 6　予測値 PRED.2012 と実績値 TME.2012 の比較: ケース 2 （ラッソ適用）

6.3.2　予測誤差の比較

表 7 は，ラッソ正則化を適用した場合とそうでない場合の平均平方誤差を比較している．リッジ法回と弾性ネット法を適用した場合の結果も掲げている．ケース 1 の表 4 の結果と比べると，いずれの正則化手法においても，予測精度の明確な向上が示されている．また，ラッソ正則化が最も予測誤差を向上していることも分か

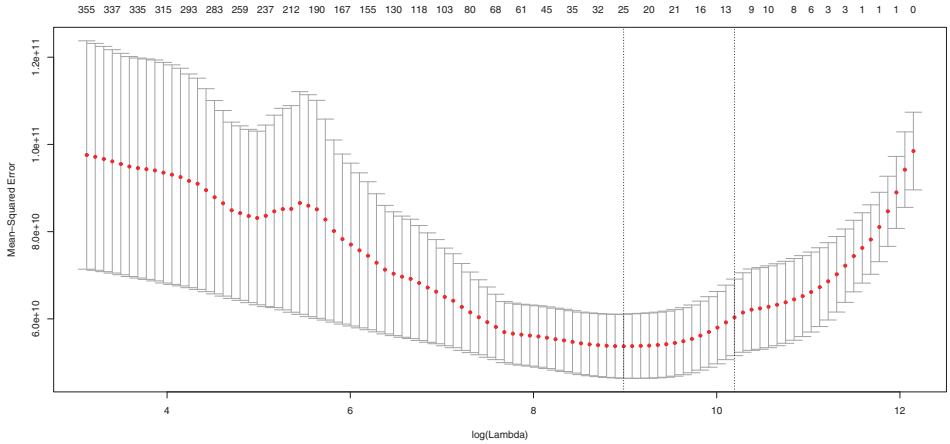

図 2　様々な λ の値に対する平均平方誤差

るであろう.

	平均 2 乗誤差	平均 2 乗誤差 / TME.2012 の分散
正則化なし	$(257820.7)^2$	0.6970
ラッソ	$(238522.5)^2$	0.5966
リッジ	$(239440.3)^2$	0.6012
弾性ネット（$\alpha = 0.5$）	$(239059.9)^2$	0.5985

表 7　予測誤差の比較：ケース 2

7　結語

本論文では，医療費の代表的な統計モデルである two-part モデルを取り上げ，正則化手法の適用により予測精度が高まるか否かを考察した．わが国健康保険データを用いて分析した結果，説明変数の個数の大きい高次元データの場合には，正則化により予測可能性が高まることが示唆された．

このような正則化による改善にも拘わらず，予測誤差は依然として大きい．これは入院医療費 (IME) の予測の困難さによるものと思われる．表 8 は，ケース 2 の場合について，医療費総額（TME）から入院医療費 (IME) を除外した医療費を予測のターゲットした場合に，正則化なしの予測誤差とラッソ正則化を適用した予測誤差を比較したものである．表 7 に比べて，将来の医療費の分散に対する予測誤差の比はかなり小さくなっている．

	平均予測誤差	平均予測誤差 / (TME.2012 − IME.2012) の分散
正則化なし	$(118107.4)^2$	0.3244078
ラッソ	$(101373.9)^2$	0.2389953

表 8　(TME-IME) に対する予測誤差の比較：ケース 2

2011 年のデータでは，IME の値が正であるデータは全体のわずか 4% である．しかし，正となる場合にはその値がかなり大きくなる傾向がある．正となる IME の値は入院日数 HID と強く関連している．HID を考慮に入れた IME の予測モデルを作ることによって，予測精度を向上できるかもしれない．

予測精度の向上を主な目的とすると，モデル統合や次元縮小のような方法の適用も可能であろう．しかし，それらの方法においては目的変数である医療費とそれを説明する共変量との関係が不明確となる．これに対して，正則化は目的変数と共変量の関係を保持しながら予測精度を向上する．また，ラッソのようなスパース性を持つ正則化は，目的変数と関連する重要な変数を選択するという追加的な利点を持つ．

参考文献

安道知寛 (2014)『高次元データ分析の方法: ― R による統計的モデリングとモデル統合―』 朝倉書店

Box, G. E. P. (1976), "Science and Statistics", *Journal of the American Statistical Association*, **71**, 791-799.

Blough, K., Madden, C.W., and Hornbrook, M.C. (1999), Modeling Risk Using Generalized Linear models. *Journal of Health Economics*, **18**, 153-171.

Duan, N., Manning, W. G. Jr., Morris, C. N., and Newhouse, J.P. (1983), A Comparison of Alternative Models for the Demand for Medical Care (Corr: V2 P413). *Journal of Business and Economic Statistics*, **1**, 115-126.

Heckman, J. (1979), Sample Selection Bias as a Specification Error. Econometrica, **47**, 153-161.

James, G. Witten, D. Hastie, T. and Tibshirani, R (2013). *An Introduction to Statistical Learning: with Applications in R* , Springer.

Jorgensen, B. (1987), Exponential dispersion models. *Journal of the Royal Statistical Society*, Series B, **49**, 127-145.

Loginov et. al. (2013), Predictive Modeling in Healthcare Costs Using Regression Models, *ARCH 2013.1 Proceedings*.

Park, T. and Casella, G. (2008), The Bayesian lasso. *Journal of the American Statistical Association*, **103**, 681-686.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society*, Series B, **58**, 267-288

Tobin, J. (1958), Estimation of relationships for limited dependent variables. *Econometrica*, **26**, 24-36.

Vidaurre, D., Bielza, C. and Pedro Larranaga, P. (2013), A Survey of L1 Regression. *International Statistical Review*, **81**, 361-387

Zou, H. (2006), The Adaptive Lasso and Its Oracle Properties, *Journal of the American Statistical Association*, **101**, 1418-1429

Zou, H. and Hastie, H. (2005) Regularization and Variable Selection via the Elastic Net, *Journal of the Royal Statistical Society*, Series B, **67**, 301-320

付録

係数	推定値	標準誤差	z 値	p 値	
(切片)	8.203e-01	4.712e-01	1.741	0.08173	.
SEX	3.192e-01	7.803e-02	4.091	4.30e-05	***
AGE.2010	-1.660e-03	3.639e-03	-0.456	0.64828	
BMI.2010	1.069e-02	1.128e-02	0.948	0.34335	
SBP.2010	-2.718e-03	3.325e-03	-0.818	0.41364	
DBP.2010	-1.160e-03	4.548e-03	-0.255	0.79871	
NF.2010	6.820e-04	4.642e-04	1.469	0.14177	
HDLC.2010	4.299e-04	2.304e-03	0.187	0.85196	
LDLC.2010	-4.030e-04	1.057e-03	-0.381	0.70290	
GOT.2010	1.048e-02	6.380e-03	1.643	0.10043	
GPT.2010	-3.434e-03	3.438e-03	-0.999	0.31788	
GGT.2010	-2.794e-04	9.099e-04	-0.307	0.75877	
FBS.2010	-3.641e-03	2.822e-03	-1.290	0.19701	
HbA1c.2010	-9.139e-03	8.351e-02	-0.109	0.91286	
US.2010	-2.382e-01	1.308e-01	-1.820	0.06869	.
TME.2010	1.669e-06	2.274e-06	0.734	0.46297	
IME.2010	-1.209e-07	2.611e-06	-0.046	0.96307	
PE.2010	2.499e-05	6.094e-06	4.101	4.12e-05	***
HID.2010	-6.387e-02	3.171e-02	-2.014	0.04400	*
OV.2010	2.676e-01	2.297e-02	11.654	< 2e-16	***
EMC.2010	-1.537e+00	6.556e-01	-2.344	0.01909	*
CM.2010	3.914e-01	3.902e-01	1.003	0.31581	
DM.2010	1.236e+00	1.067e+00	1.158	0.24674	
BPM.2010	-3.584e-01	5.366e-01	-0.668	0.50416	
HL.2010	1.234e-01	2.378e-01	0.519	0.60374	
DB.2010	4.096e-01	3.245e-01	1.262	0.20686	
HBP.2010	1.662e+00	5.565e-01	2.987	0.00282	**
LD.2010	5.895e-02	2.292e-01	0.257	0.79701	

表 9　ロジスティック回帰の係数推定値

係数	推定値	標準誤差	z 値	p 値	
(切片)	-3.117e+04	3.892e+04	-0.801	0.423204	
SEX	1.838e+02	6.109e+03	0.030	0.975992	
AGE.2010	5.241e+02	3.061e+02	1.713	0.086833	.
BMI.2010	-1.017e+03	8.889e+02	-1.144	0.252831	
SBP.2010	-1.823e+00	2.694e+02	-0.007	0.994602	
DBP.2010	5.847e+02	3.708e+02	1.577	0.114886	
NF.2010	4.158e+00	3.532e+01	0.118	0.906311	
HDLC.2010	-2.138e+01	1.853e+02	-0.115	0.908161	
LDLC.2010	6.917e+01	8.535e+01	0.810	0.417711	
GOT.2010	-8.118e+01	3.668e+02	-0.221	0.824828	
GPT.2010	-3.284e-01	2.323e+02	-0.001	0.998872	
GGT.2010	4.236e+01	6.740e+01	0.629	0.529690	
FBS.2010	-1.335e+02	2.193e+02	-0.609	0.542796	
HbA1c.2010	-3.617e+03	6.413e+03	-0.564	0.572715	
US.2010	4.335e+04	8.079e+03	5.366	8.26e-08	***
TME.2010	1.211e+00	1.968e-02	61.553	<2e-16	***
IME.2010	-1.149e+00	2.912e-02	-39.447	<2e-16	***
PE.2010	-5.253e-02	4.126e-02	-1.273	0.202995	
HID.2010	7.246e+02	1.097e+03	0.660	0.509093	
OV.2010	-9.121e+02	2.647e+02	-3.446	0.000572	***
EMC.2010	-4.385e+04	3.857e+04	-1.137	0.255718	
CM.2010	1.375e+04	1.203e+04	1.143	0.252962	
DM.2010	-1.145e+04	1.740e+04	-0.658	0.510541	
BPM.2010	5.232e+04	1.635e+04	3.200	0.001382	**
HL.2010	-1.443e+04	1.074e+04	-1.344	0.179021	
DB.2010	1.810e+04	1.208e+04	1.499	0.133893	
HBP.2010	-3.129e+04	1.637e+04	-1.911	0.056025	.
LD.2010	-2.184e+04	9.741e+03	-2.242	0.024985	*

表 10　線形回帰の係数推定値

係数	線形回帰推定値	z 値	ラッソ回帰推定値
(切片)	-3.117e+04	-0.801	
SEX	1.838e+02	0.030	·
AGE.2010	5.241e+02	1.713	241.2720
BMI.2010	-1.017e+03	-1.144	·
SBP.2010	-1.823e+00	-0.007	·
DBP.2010	5.847e+02	1.577	291.9444
NF.2010	4.158e+00	0.118	·
HDLC.2010	-2.138e+01	-0.115	·
LDLC.2010	6.917e+01	0.810	·
GOT.2010	-8.118e+01	-0.221	·
GPT.2010	-3.284e-01	-0.001	·
GGT.2010	4.236e+01	0.629	·
FBS.2010	-1.335e+02	-0.609	·
HbA1c.2010	-3.617e+03	-0.564	·
US.2010	4.335e+04	5.366	32279.08
TME.2010	1.211e+00	61.553	1.117933
IME.2010	-1.149e+00	-39.447	-1.025502
PE.2010	-5.253e-02	-1.273	·
HID.2010	7.246e+02	0.660	·
OV.2010	-9.121e+02	-3.446	·
EMC.2010	-4.385e+04	-1.137	·
CM.2010	1.375e+04	1.143	·
DM.2010	-1.145e+04	-0.658	·
BPM.2010	5.232e+04	3.200	17674.20
HL.2010	-1.443e+04	-1.344	·
DB.2010	1.810e+04	1.499	·
HBP.2010	-3.129e+04	-1.911	·
LD.2010	-2.184e+04	-2.242	-8465.799

表 11　線形回帰推定値とラッソ回帰推定値の比較：ケース 1