共変量シフト下での医療費予測モデリング

――我が国健康保険データへの応用――

小　暮　厚　之

1　はじめに

保健医療水準の改善と長寿化が進行する中で，医療費の高騰は多くの国々において喫緊の課題となっている。この課題の解決を図る上で，将来の医療費の予測は不可欠であろう。本論文では，医療費予測に用いる統計モデリングにおける共変量シフトの問題を考察する[1)]。

ある年（t 年とする）において翌年（$t+1$ 年）の医療費 Y_{t+1} を予測するために，以下の 2 つのステップの手続きを取ることが多い：

1. t 年における医療費 Y_t を目的変数，前年の共変量 $\boldsymbol{X}_{t-1}$ を説明変数とする予測モデルを構築
2. 構築した予測モデルの説明変数の値を t 年の共変量 $\boldsymbol{X}_t$ に置き換え，Y_{t+1} を予測。

第 1 ステップにおける予測モデルは，前年の共変量 $\boldsymbol{X}_{t-1}=\boldsymbol{x}$ を所与とした今年の医療費 Y_t の条件つき期待値 $g(\boldsymbol{x})$ を推定することによって得られることが多い。g の推定値を $\hat{g}$ とするとき，翌年の医療費 Y_{t+1} は，$\hat{g}(\boldsymbol{X}_t)$ によって予測される。

この予測の手続きは，「$(\boldsymbol{X}_t, Y_{t+1})$ の確率分布が $(\boldsymbol{X}_{t-1}, Y_t)$ の確率分布から変化していない」ということを暗黙裡に想定している。この想定の下では，予測値 $\hat{g}(\boldsymbol{X}_t)$ はターゲットである Y_{t+1} に近いと考えられるであろう。しかし，現実のいわゆる「リアルワールドデータ」を用いる場合には，これら二つの確率分布の間に何らかの変化が生じている可能性があろう。その場合，この手続きに基づいた予測の信頼性は失われる。機械学習の世界では，このような問題を「データシフト」という名称で扱い，その困難さを軽減するための適応（adaptation）の手法を議論している。

機械学習の用語を用いると，$(\boldsymbol{X}_{t-1}, Y_t)$ は，訓練データ，$(\boldsymbol{X}_t, Y_{t+1})$ はテストデータに対応する。データシフトとは，訓練データの分布とテストデータの分布の間の不一致を指す。そのような不一致はいくつかのタイプに分類できるが，本稿では共変量シフトと呼ばれるタイプの状況を考える。すなわち，訓練データの共変量とテストデータの共変量の分布の間には何らかの変化があるが，共変量を所与とする医療費の条件付き分布は両者間で変化はない

ものとする。そのような状況は，例えば，保険会社が新しい医療保険の市場に参入する際に，既存の市場のデータを用いて新市場の医療費の予測を行う場合などで生じるであろう。

本論文の目的は，実際の健康保険データを用いて，共変量シフトに対する適応手段を用いることによって医療費の予測精度が向上するかどうかを検討することである。予測モデルの構築にあたっては，医療費データの中にゼロの値となるものが含まれることを考慮して two-part モデルを用いる。また，利用するデータは，我が国の健康保険組合から無作為に抽出された 1 万人の加入者に関する 2010–2012 年の 3 年間のレセプトデータ及び健診データである[2)]。

2　共変量シフト

共変量 $\boldsymbol{X}$ と目的変数 Y の同時分布 $p(\boldsymbol{x}, y)$ は

$$p(\boldsymbol{x}, y) = p(y|\boldsymbol{x})p(\boldsymbol{x}) = p(\boldsymbol{x}|y)p(y) \tag{1}$$

と分解される。ここで，$p(y|\boldsymbol{x})$ は $\boldsymbol{X}=\boldsymbol{x}$ を所与とする Y の条件付き分布，$p(\boldsymbol{x}|y)$ は $Y=y$ を所与とする $\boldsymbol{X}$ の条件付き分布，$p(\boldsymbol{x})$ と $p(y)$ は，それぞれ $\boldsymbol{X}$ と Y の周辺分布とする。

(1) の分解に基づいて，Moreno-Torres and et al.（2011）は目的変数がクラス変数である場合について，データシフトをいくつかのタイプに区分している。以下では，p の添え字の train, test によって訓練データの分布かテストデータの分布かを区別する。

1. 共変量シフト（Covariate shift）：

$$p_{\text{train}}(y|\boldsymbol{x}) = p_{\text{test}}(y|\boldsymbol{x}) \text{ かつ } p_{\text{train}}(\boldsymbol{x}) \neq p_{\text{test}}(\boldsymbol{x}) \tag{2}$$

2. 事前確率シフト（Prior probability shift）：

$$p_{\text{train}}(\boldsymbol{x}|y) = p_{\text{test}}(\boldsymbol{x}|y) \text{ かつ } p_{\text{train}}(y) \neq p_{\text{test}}(y)$$

3. コンセプトシフト（Concept shift）：

$$p_{\text{train}}(y|\boldsymbol{x}) \neq p_{\text{test}}(y|\boldsymbol{x}) \text{ かつ } p_{\text{train}}(\boldsymbol{x}) = p_{\text{test}}(\boldsymbol{x})$$

または

$$p_{\text{train}}(\boldsymbol{x}|y) \neq p_{\text{test}}(\boldsymbol{x}|y) \text{ かつ } p_{\text{train}}(y) = p_{\text{test}}(y)$$

本論文では，共変量シフト（2）を考察する。1 節で述べた 2 段階の予測手続きでは，こ

のデータシフトは

1. 今年の共変量 $\boldsymbol{X}_t$ の分布は前年の共変量 $\boldsymbol{X}_{t-1}$ の分布から変化している。
2. しかし，今年の共変量 $\boldsymbol{X}_t$ を所与とする来年の医療費 Y_{t+1} の条件付き分布は，前年の共変量 $\boldsymbol{X}_{t-1}$ を所与とする今年の医療費 Y_t の分布から変化していない。

ことを意味する。Shimodaira（2000）を端緒として，共変量シフトに対する適応手段に関して様々な研究がなされている。詳細については，例えば，Sugiyama and et al.（2017）や Sugiyama and Kawanabe（2012）を参考にされたい。

3　経験リスクに基づくパラメータ推定

本節では，統計モデルを用いて予測を行う場合のパラメータ推定について述べる。最尤法を含む多くの推定法は，経験リスクを最小にするようにパラメータを推定する。しかし，共変量シフトが生じている場合には，通常の経験リスクに基づく方法が不適切であることを見る。

訓練データ $\{(\boldsymbol{X}_i, Y_i), i=1, \cdots, n\}$ の各観測値は，同一分布

$$p_{\text{train}}(\boldsymbol{x}, y) = p_{\text{train}}(y|\boldsymbol{x})p_{\text{train}}(\boldsymbol{x})$$

に互いに独立に従うものとする。また，$p_{\text{train}}(y|\boldsymbol{x})$ の統計モデルとして

$$\{f(y|\boldsymbol{x}, \theta), \theta \in \Theta \subset \mathbb{R}^m\}$$

を採用する。このとき，ある $\hat{\theta} \in \Theta$ を選択し，$f(y|\boldsymbol{x}, \hat{\theta})$ を，共変量の値が $\boldsymbol{x}$ であるときの予測分布とする。 回帰問題であれば，予測分布の平均

$$g(\boldsymbol{x}, \hat{\theta}) \equiv \int y f(y|\boldsymbol{x}, \hat{\theta}) dy$$

を $\boldsymbol{x}$ に対するアウトプット Y に対する点予測としてしばしば用いる。

予測分布 $f(y|\boldsymbol{x}, \theta)$ を用いたときの損失を $\text{loss}(\boldsymbol{x}, y, \theta)$ とする。このとき，損失の期待値であるリスク

$$\text{TRAIN RISK} = \text{E}^{\text{train}}[\text{loss}(\boldsymbol{x}, y, \theta)] \tag{3}$$

を最小にする θ が最適な値と考えられる。ここで，E^{train} は訓練分布に関する期待値を表す。このとき，経験リスク

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{loss}(Y_i, g(\boldsymbol{X}_i, \theta)) \tag{4}$$

を最小にする

$$\hat{\theta} = \arg\min_{\theta}\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{loss}(\boldsymbol{X}_i, Y_i, \theta)\right]$$

を θ の点推定値として用いることが自然であろう。もしも損失関数が

$$\mathrm{loss}(\boldsymbol{x}, y, \theta) = -\log f(y|\boldsymbol{x}, \theta)$$

ならば，$\hat{\theta}$ は最尤推定値である。

医療費の予測を考えるとき，訓練データの共変量に対する予測ではなく，テストデータの共変量に対する予測に焦点を合わせるべきであろう。この場合，ターゲットとなるのは

$$\text{TEST RISK} = \mathrm{E}^{\mathrm{test}}_{X,Y}[\mathrm{loss}(\boldsymbol{X}, Y, \theta)] \tag{5}$$

である。ここで，$\mathrm{E}^{\mathrm{test}}$ はテスト分布に関する期待値を表す。

データシフトが生じていない場合には

$$\text{TEST RISK} = \mathrm{E}^{\mathrm{test}}_{X,Y}[l(\boldsymbol{X}, Y, \boldsymbol{\theta})] = \mathrm{E}^{\mathrm{train}}_{X,Y}[l(\boldsymbol{X}, Y, \boldsymbol{\theta})]$$

となるため，経験リスク（4）の最小化は適切な推定法である。

しかし，共変量シフトが生じている場合は

$$\begin{aligned}\text{TEST RISK} &= \mathrm{E}^{\mathrm{test}}_{X}[\mathrm{E}^{\mathrm{test}}_{Y|X}[l(\boldsymbol{X}, Y, \boldsymbol{\theta})]] \\ &= \mathrm{E}^{\mathrm{test}}_{X}[\mathrm{E}^{\mathrm{train}}_{Y|X}[l(\boldsymbol{X}, Y, \boldsymbol{\theta})]]\end{aligned}$$

となり，経験リスクの最小化は不適切となる。

モデルが正しい場合，すなわち，ある $\theta^* \in \Theta$ に対して

$$p_{\mathrm{train}}(y|\boldsymbol{x}) = f(y|\boldsymbol{x}, \theta^*)$$

となる場合には，（5）式の TEST RISK を最小にする θ と（3）式の TRAIN RISK を最小にする θ の値は一致し，共変量シフトによる問題は生じない。しかし，現実の場面では，モデルが正しく特定化されていることは必ずしも期待できない。そのような場合，共変量シフトの問題から逃れることはできない。

4　加重経験リスクに基づくパラメータ推定

4.1　加重経験リスク

共変量シフトに対処するために，経験リスクの代わりに，加重経験リスク

$$\frac{1}{n}\sum_{i=1}^{n} w(\boldsymbol{X}_i) l(\boldsymbol{X}_i, Y_i, \boldsymbol{\theta}), \tag{6}$$

を最小化することを考える。ここで，$w(\cdot)$ は，各 $\boldsymbol{X}_i$ の値に応じた加重である。$n \to \infty$ になるにつれ，加重経験リスクは

$$\frac{1}{n}\sum_{i=1}^{n} w(\boldsymbol{X}_i) l(\boldsymbol{X}_i, Y_i, \boldsymbol{\theta}) \to \mathrm{E}_{X}^{\mathrm{train}}[w(\boldsymbol{X})\mathrm{E}_{Y|X}^{\mathrm{train}}[l(\boldsymbol{X}, Y, \boldsymbol{\theta})]]$$

に近づく。w を

$$w(\boldsymbol{x}) \propto \frac{p_{\mathrm{test}}(\boldsymbol{x})}{p_{\mathrm{train}}(\boldsymbol{x})},$$

と設定すれば

$$\begin{aligned}
&\mathrm{E}_{X}^{\mathrm{train}}[w(\boldsymbol{X})\mathrm{E}_{Y|X}^{\mathrm{train}}[l(\boldsymbol{X}, Y, \boldsymbol{\theta})]] \\
&\quad \propto \int p_{\mathrm{train}}(\boldsymbol{x}) \frac{p_{\mathrm{test}}(\boldsymbol{x})}{p_{\mathrm{train}}(\boldsymbol{x})} \int p_{\mathrm{train}}(y|\boldsymbol{x}) l(\boldsymbol{x}, y, \boldsymbol{\theta}) dy d\boldsymbol{x} \\
&\quad = \int p_{\mathrm{test}}(\boldsymbol{x}) \int p_{\mathrm{train}}(y|\boldsymbol{x}) l(\boldsymbol{x}, y, \boldsymbol{\theta}) dy d\boldsymbol{x} \\
&\quad = \mathrm{E}_{X}^{\mathrm{test}}[\mathrm{E}_{Y|X}^{\mathrm{train}}[l(\boldsymbol{X}, Y, \boldsymbol{\theta}]].
\end{aligned}$$

となり，共変量シフト下でも TEST RISK に一致する。したがって，共変量シフトが生じている場合でも，適切な荷重を用いた加重経験リスク（6）を最小化することにより，TEST RISK の最小化を実現できると期待できる。

4.2　加重の推定：ロジスティック回帰

実際に加重経験リスクを用いるためには，加重として用いる密度比

$$\frac{p_{\mathrm{test}}(\boldsymbol{x})}{p_{\mathrm{train}}(\boldsymbol{x})}$$

を推定する必要がある。

簡単な方法は，ロジスティック回帰の手法を用いることである。ある観測値 $(\boldsymbol{X}, y)$ に対して，ダミー変数 Z を

$$Z = \begin{cases} 1 & \text{観測値がテストデータに属する場合} \\ 0 & \text{観測値が訓練データに属する場合} \end{cases}$$

と定義する。このとき，Z は共変量が $\boldsymbol{X}=\boldsymbol{x}$ のロジスティック回帰に従うと仮定する。すなわち，

$$\Pr(Z=1|\boldsymbol{x}) = \frac{1}{1+\exp(-\boldsymbol{\beta}'\boldsymbol{x})},\quad \Pr(Z=0|\boldsymbol{x})=\frac{\exp(-\boldsymbol{\beta}'\boldsymbol{x})}{1+\exp(-\boldsymbol{\beta}'\boldsymbol{x})} \tag{7}$$

とする。この結果

$$\begin{aligned}\frac{p_{\text{test}}(\boldsymbol{x})}{p_{\text{train}}(\boldsymbol{x})} &= \frac{\Pr(\boldsymbol{X}=\boldsymbol{x}|Z=1)}{\Pr(\boldsymbol{X}=\boldsymbol{x}|Z=0)}\\ &= \frac{\Pr(Z=1|\boldsymbol{x})\Pr(Z=1)}{\Pr(Z=0|\boldsymbol{x})\Pr(Z=0)}\\ &\propto \frac{\Pr(Z=1|\boldsymbol{x})}{\Pr(Z=0|\boldsymbol{x})}=\exp(\boldsymbol{\beta}'\boldsymbol{x}).\end{aligned}$$

となる。従って，Z を被説明変数とするロジスティック回帰（7）のパラメータ $\boldsymbol{\beta}$ の推定値 $\widehat{\boldsymbol{\beta}}$ を求め，加重を

$$w(\boldsymbol{x}) \propto \exp(\widehat{\boldsymbol{\beta}}'\boldsymbol{x})$$

とすればよい。

5　医療費の回帰モデル

n 人の加入者からなる集団を考える。1 年間における加入者 i の医療費は

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

と表せる。ここで，y_{ij} は加入者 i の j 回目の受診の医療費であり，N_i は 1 年間の受診件数である。受診件数 N_i がゼロの場合には，1 年間の総医療費もゼロとなる。実際の統計分析では，医療費がゼロとなる観察値がデータに含まれることが多い。このような場合，ゼロとなるデータを除いて回帰分析を行うと，推定されたパラメータにバイアスが生じることが知られている。このようなバイアスを考慮して，ゼロを含む正値を説明する回帰モデルとして，トービット・モデル（Tobin, 1958），標本選択モデル（Heckman, 1979），Two-part モデル（Duan et al., 1983）などが提案されている。

加入者 i の医療費 Y_i を p 個の共変量

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

によって説明する問題を考える。トービット・モデルでは，観察される Y_i の背後に，潜在変数 Y_i^* を想定する。Y_i^* は「真の医療費」を表し，線形回帰モデル

$$Y_i^* = \boldsymbol{\beta}'\boldsymbol{x}_i + \varepsilon_i \tag{8}$$

に従っていると仮定する。ここで，$\boldsymbol{\beta}=(\beta_1, \cdots, \beta_p)$ 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，Y_i 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，$\Pr(Y_i^*\leq 0)=\Pr(Y_i=0)$ の割合だけ実際の医療費を過大に評価することになる。

標本選択モデルでは，トービット・モデルで用いた医療費の回帰モデルに加え，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，(Y_i^*, S_i^*) が

$$\begin{cases} Y_i^* = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i \\ S_i^* = \boldsymbol{x}_i'\boldsymbol{\gamma} + \eta_i \end{cases} \tag{9}$$

であると仮定する。ただし，誤差項は 2 変量正規分布

$$\begin{pmatrix} \varepsilon_i \\ \eta_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \sigma\rho \\ \sigma\rho & 1 \end{pmatrix} \right) \tag{10}$$

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

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

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。

(9) 式において，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

$$\Pr(Y_i \leq y) = \begin{cases} \Pr(Y_i \leq y | S_i=1)\Pr(S_i=1)+\Pr(S_i=0) & y>0 \text{ の場合} \\ \Pr(S_i=0) & y=0 \text{ の場合} \\ 0 & y<0 \text{ の場合} \end{cases} \tag{11}$$

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

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想定し，S_i の統計モデルと $Y_i|S_i=1$ の統計モデルを独立に設定する。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 モデルを予測モデルとして採用する。

6　Two-part モデルによる予測

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。(11) より，$\boldsymbol{x}_i$ を所与とする $Y_i=y_i$ の条件付き分布は

$$f_Y(y_i ; \boldsymbol{\theta}|\boldsymbol{x}_i) = \begin{cases} \Pr(Y_i=0 ; \boldsymbol{\theta}_1|\boldsymbol{x}_i) & y_i=0 \text{ の場合} \\ f_Y(y_i ; \boldsymbol{\theta}_2|\boldsymbol{x}_i, y_i>0)\Pr(Y_i>0 ; \boldsymbol{\theta}_1|\boldsymbol{x}_i) & y_i>0 \text{ の場合} \end{cases} \tag{12}$$

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・ベクトルである。それらをまとめて，$\boldsymbol{\theta}=(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ とする。

本論文では（12）式の第 1 段階の統計モデルとして，ロジスティック回帰

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

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

を採用する。

（12）式の第 2 段階の回帰モデル $f_Y(y_i ; \boldsymbol{\theta}_2|\boldsymbol{x}_i, y_i>0)$ として，対数線形回帰モデル

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

を当てはめる。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。

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

$$\mathrm{E}[(Y-\widehat{Y})^2]$$

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

$$\begin{aligned}\mathrm{E}[Y;\boldsymbol{\theta}|\boldsymbol{x}] &= \mathrm{E}[Y;\boldsymbol{\theta}_2|Y>0,\boldsymbol{x}]\Pr(Y>0;\boldsymbol{\theta}_1|\boldsymbol{x})+\mathrm{E}[Y;\boldsymbol{\theta}_2|Y=0,\boldsymbol{x}]\Pr(Y=0;\boldsymbol{\theta}_1|\boldsymbol{x}) \\ &= \mathrm{E}[Y;\boldsymbol{\theta}_2|Y>0,\boldsymbol{x}]\Pr(Y>0;\boldsymbol{\theta}_1|\boldsymbol{x}) \qquad (13)\end{aligned}$$

で与えられる。

誤差項が正規分布である場合には，(13) 式は

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

となる。詳細については小暮・小林（2018）を参照されたい。

7　我が国健康保険データへの応用

7.1　データ

分析に用いるデータは，我が国の健康保険組合から無作為に抽出された 1 万人の加入者に関する 2010–2012 年の 3 年間のレセプトデータ及び健診データである。

予測のターゲットは，2012 年の医療費 Y_{2012} である。そのため，2011 年の共変量 X_{2011}，2011 年の医療費，2010 年の共変量 X_{2010} が利用可能である。

共変量としては，

- 人口統計学的な属性
  SEX（性別），AGE（年齢）

- 健康診断の変数
  BMI（ボディマス指数），
  SBP（収縮期血圧），DBP（拡張期血圧），
  NF（中性脂肪），
  HDLC（HDL コレステロール），LDLC（LDL コレステロール），
  GOT（GOT），
  GPT（GPT），
  GGT（GGT），
  FBS（空腹時血糖）
  HbA1c（ヘモグロビン A1c）。

を用いた。

ここでは，訓練データは高年齢の集団から得られており，テストデータはそれよりも低年齢の集団から得られている想定し，訓練データとテストデータを以下のように設定した。

- 訓練データは，45 歳を超える加入者に対する 2011 年の医療費（Y_{2011}）及び 2010 年の共変量（X_{2010}）。
- テストデータは，45 歳以下の加入者に対する 2012 年の医療費（Y_{2012}）及び 2011 年の共変量（X_{2011}）

7.2　サブサンプリンング

以下の手順に従って，10 個のサブサンプリングを作成する：

1. 全体のデータを大きさが等しい 10 個のブロックに等分する。k 番目のブロックは，$1000(k-1)+1$ から $1000k$ までの観測値からなる。ここで，$k=1, 2, \cdots, 10$ 。
2. k 番目のブロックを除き，残りの 9000 人の観測値からなるサブ・サンプルを作る。
3. 各サブサンプルに対して，訓練データとテストデータを以下のように作成する。

- 訓練データは，45 歳を超える加入者に対する 2011 年の医療費（Y_{2011}）及び 2010 年の共変量（X_{2010}）。
- テストデータは，45 歳以下の加入者に対する 2012 年の医療費（Y_{2012}）及び 2011 年の共変量（X_{2011}）

4. 各サブサンプルについて，共変量シフトに対する適応手段を施した場合（加重経験リスクを用いてパラメータを推定）とそうしなかった場合（通常の経験リスクを用いてパラメータ推定）の予測精度を比較

7.3　予測誤差の尺度

予測精度の比較のために，各サブサンプルに対して，次の二つの尺度を計算した：

- 平方根平均二乗誤差

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(Y_i - \widehat{Y}_i)^2}$$

- 平均絶対値偏差

$$\mathrm{MAE} = \frac{1}{m}\sum_{i=1}^{m}|Y_i - \widehat{Y}_i|$$

ここで，m は，各サブサンプルにおけるテストデータの大きさである。また，Y_i は各テストデータにおける 2012 年の医療費の i 番目の観測値であり，$\widehat{Y}_i$ はその予測値である。

これら二つの尺度の各々について，共変量シフトに対する適応手段を施した場合とそうしなかった場合の予測誤差の比

$$\text{RPE} = \frac{\text{適応手段を施した場合の予測誤差}}{\text{適応手段を施さなかった場合の予測誤差}}$$

を計算した。REP の値が 1 より小さければ，適応手段による改善が見られたことになる。

7. 4　推定結果

7. 4. 1　結果 1

訓練データとテストデータを以下のように作成した。

- 訓練データは，45 歳を超える加入者に対する 2011 年の医療費（Y_{2011}）及び 2010 年の共変量（$\boldsymbol{X}_{2010}$）。
- テストデータは，45 歳以下の加入者に対する 2012 年の医療費（Y_{2012}）及び 2011 年の共変量（$\boldsymbol{X}_{2011}$）

予測誤差の比（RPE）

サブサンプル	1	2	3	4	5
RMSE	0.4586997	0.7558779	0.5937718	0.7296496	0.5165025
MAE	0.3130207	0.3211196	0.333876	0.3270688	0.3309153
サブサンプル	6	7	8	9	10
RMSE	0.4226051	0.804832	0.3685778	0.5503828	0.5193979
MAE	0.3222189	0.3059576	0.5193979	0.3713025	0.3175049

この表から，RMSE と MAE のどちらの尺度を用いても，共変量シフトに対する適応手段が予測精度を向上させていることが分かる。

7. 4. 2　結果 2

ここでは，訓練データとテストデータを以下のように作成した。

- 訓練データは，40 歳を超える加入者に対する 2011 年の医療費（Y_{2011}）及び 2010 年の共変量（$\boldsymbol{X}_{2010}$）。
- テストデータは，50 歳以下の加入者に対する 2012 年の医療費（Y_{2012}）及び 2011 年の

共変量（$\boldsymbol{X}_{2011}$）

従って，訓練データとテストデータの年齢層には重なる部分がある。

予測誤差の比（RPE）

サブサンプル	1	2	3	4	5
RMSE	0.9500305	0.9402613	0.962066	0.9596291	0.9536439
MAE	0.259719	0.2561888	0.2550798	0.2518912	0.2625094
サブサンプル	6	7	8	9	10
RMSE	0.8996982	0.9543823	0.9419313	1.00291	0.9720088
MAE	0.2606326	0.2553133	0.2602852	0.2692236	0.2710581

この場合には，MAE に関しては予測精度の向上が見られるが，RMSE に関する予測精度は向上していない。

8　おわりに

本稿では，医療費の予測モデリングにおける共変量シフトの問題に着目し，健康保険データへ適用を通じて，この問題に対する適応手段が現実のデータに対して有用かどうかを検討した。訓練データが高年齢集団，テストデータがそれよりも低年齢の集団という設定の下でデータ分析を行い，予測精度が向上する可能性を示した。また，訓練データとテストデータをいかに設定するかによって予測精度が大きく変化することも見出した。

しかし，今回の分析は規模も小さく，またその設定はいささか恣意的である。そのため，7 節の結果から得られる含意も限定的である。共変量シフトに対する適応手段の現実的な有用性を確かめるために，様々な設定の下でさらなる分析を進めて行きたい。

謝辞

本論の作成にあたって，株式会社 JMDC からデータの提供を受けました。同社に深く感謝申し上げます。

附記：本稿は 2018 年度個人研究助成費による研究成果の一部である。

注 ――――――――

1）本論は，2018 年 11 月 10 日に開催された日本保険・年金リスク学会 第 17 回研究発表大会における報告に基づいている。

2）ここで，加入者とは被保険者とその被扶養者とする。

参 考 文 献

小暮厚之・小林凌雅（2018）「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」日本保険・年金リスク学会誌第 10 巻第 1 号 21–35

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.

Jose G. Moreno-Torres, J. G., Raeder, T., Rodriguez, R. A, Chawla, N. V. and Herrera, F. (2012), A unifying view on dataset shift in classification. *Pattern Recognition*, **49**, 521–530.

Shimodaira, H. (2000) Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, **90**. 227–244.

Sugiyama, M., Krauledat, M. and Muller, K.-R. (2007), Covariate shift adaptation by importance weighted cross validation, *Jounal of Machine Learning Research*, **8**, 985–1005.

Sugiyama, M. and Kawanabe, M. (2012), *Machine leaning in non-stationary environments: introduction to covariate shift adaptation*, The MIT Press, Cambridge, Massachusetts.

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

Wei, D. Ramamurthy, K. N. and Varshney, K. R. (2015), Health Insurance Market Risk Assessment: Covariate Shift and k-Anonymity. *Proceedings of the 2015 SIAM International Conference on Data Mining*, 226–234.