ISS-P-013

健診データに対する欠損値補完手法の検討

本村　壮志†　永原　正章†　山崎　恭†
† 北九州市立大学

1. はじめに

健康寿命と平均寿命との差は不健康な期間を意味し，その期間の拡大は公的医療費の増大などに繋がる社会的な問題である．健康寿命を延ばし，この差を短縮するために，健診データを用いた健康状態の予測が注目されている．しかし，健診データには受診漏れなどに起因する欠損値が多く含まれ，それらを無視することはできない．そこで，本研究では，健診データに含まれる欠損値に対して，確率的回帰代入法と多重代入法を用いた推定を試み，その精度を分析する．

2. 健康診断で得られた健診データ

本研究では，実際の検査機関から提供を受け完全匿名加工した健診データを用いる．健診期間は 2009 年からの 10 年間で約23万人分のデータである．健診項目には検査・問診の 88 項目が含まれる．また，個人情報の秘匿性を保つために乱数を用いた加工が施され，データの統計的な特徴を失わないような秘匿化が行われている．

3. 擬似欠損データと実験

健診データをリストワイズ除去法によって完全データのみ抽出し，年齢，性別，身長，体重，BMI，腹囲，収縮期最高血圧，拡張期最低血圧の数値データ 8 項目を解析に用いる．腹囲の検査では，BMI が 20 未満の人は医師が必要でないと認めるとき省略でき，腹囲の欠損は BMI に依存して欠損すると考えられる[4]．これは，欠損が条件付きで無作為に起こる，Missing At Random (MAR) に分類できる [2]．

図1に，MAR の概念図を示す [1]．欠損を含む変数をY，その他の変数をX，Yの欠損の有無を表す確率変数をR（Yが欠損ならば1，欠損でなければ0）とする．ここでは，腹囲をY，BMI をX，腹囲の欠損の有無をRとおく．

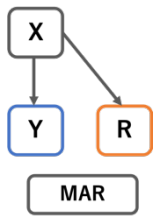

図1．MAR の概念図

MAR の欠損メカニズムに基づいて，データ数に対しての欠損が 10%，20%，30%となるように擬似的に欠損を発生させた擬似欠損データを作成した．MAR の欠損は，欠損値を除去して解析を行うと推定値に偏りが生じるため，何らかの値を代入する必要がある [2]．そこで，(1) 確率的回帰代入法，(2) 多重代入法を用いて補完値を代入する．

(1)　確率的回帰代入法は，目的変数Yを腹囲，説明変数を年齢，性別，身長，体重，BMI，収縮期最高血圧，拡張期最低血圧の 7 つとし，この順序で $X = (x_1, x_2, x_3, x_4, x_5, x_6, x_7)$ とする．回帰係数を $\beta = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7)$，定数項を α とし，推定値 $\hat{Y}$ を求める以下の回帰式を仮定する．

$$\hat{Y} = \alpha + \beta X^T$$

この回帰式に正規分布からの乱数を加えたものを補完値とする．

(2)　多重代入法は，確率的回帰代入法によって得られた擬似完全データを 20 組作成する．20 組のデータをそれぞれ独立に解析し，得られた結果を統合した値を最終的な補完値とする[3]．

4. 解析結果

欠損値を補完したデータの平均，標準偏差，標準誤差と真値の平均，標準偏差，標準誤差とを比較した結果を以下に示す．

(1)　確率的回帰代入法 ： 平均値は，欠損割合による推定の偏りはない．標準偏差，標準誤差は，欠損割合が高いほど大きくなり，データのばらつきが大きくなる．

(2)　多重代入法 ： 平均値は，欠損割合が 30%では，10%，20%のときと比べて誤差が大きくなる．標準偏差，標準誤差は，欠損割合が高くなるほど過小評価される．しかし，いずれの場合も誤差が 3%以下に収まる．

5. まとめ

多重代入法では，いずれの場合も真値に近づき，推定値のばらつきを捉えることができた．　よって，腹囲のデータ欠損に対しては，多重代入法による補完によって偏りのない推定と欠損による推定誤差を捉えた結果が得られた．しかし，本研究では回帰パラメータ推定に説明変数の選択を行なっていない．そこで，精度向上のために，主成分分析や正則化などの加工を施すことが重要である．

謝辞

本研究は平成 31 年度北九州市学術・研究振興事業調査研究助成金の助成を受けたものです．

参考文献

[1] C. K. Enders, *Applied Missing Data Analysis*, The Guilford Press, 2010.

[2] D. B. Rubin, Inference and missing data, *Biometrika*, vol. 63, pp. 581–592, 1976.

[3] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*, Wiley, 1987.

[4] 労働大臣告示第 88 号「労働安全衛生規則第 44 条第 2 項の規定に基づき厚生労働大臣が定める基準」, 労働安全衛生法に基づく健康診断の概要, 参考資料 2, 1998. 6. 24.