論文内容の要旨

論文題目　確率的補完法による欠測データの解析
―大規模疫学研究 (日本動脈硬化縦断研究; JALS) データへの適用―

指導教員 松山裕助教授

東京大学大学院医学系研究科

平成 16 年 4 月進学

博士後期課程

健康科学・看護学専攻

土居 主尚

1　背景

多くの疫学研究において、予定していた全てのデータを測定できることは稀であり、データ解析の際には欠測データの問題に少なからず直面する。これまで提案されている欠測データに対する統計解析手法のほとんどは、結果変数の欠測に伴うバイアスの補正を念頭としているが、説明変数の欠測に伴うバイアスの問題も回帰分析を主要な解析手法とする疫学研究においては重要な問題である。

2　目的

本研究では、結果変数の欠測に対してだけでなく、複数の説明変数に欠測値が存在する場合にも適用可能な新しい統計解析手法である確率的補完法（stochastic imputation method）

を提案する。

3　対象と方法

提案する確率的補完法は、欠測データ解析の手法として現在広く使用されている多重補完法（multiple imputation method）を改良した手法である。主な改良点は、多重補完法では、補完モデルの選択によっては常に同じ値が補完される可能性があるのに対して、提案する方法は、欠測値がとりうる全ての値を考え、それらの値が実現する可能性（観察確率）に応じて欠測データを重み付けする点である。確率的補完法によって得られる推定値の信頼区間は、ブートストラップ法により求めることを提案する。シミュレーション実験を通して、説明変数に複数の欠測値が存在する場合の標準的な解析手法である complete case 解析と、2 通りの多重補完法（ロジスティックモデルと傾向スコアによる補完）、及び確率的補完法の性能を比較検討する。さらに、大規模疫学研究である日本動脈硬化縦断研究（Japan Arteriosclerosis Longitudinal Study; JALS）データにおける、脳梗塞発症に対するリスクファクターの検討の際に見られた複数の説明変数の欠測値に対処するために、提案する方法の当てはめを行う。本研究で対象としたコホートは 13 コホート、総対象者数は 42,546 人（男性 16,521 人、女性 26,025 人）、総観察人年は 358,369 人年（男性 131,759 人年、女性 226,610 人年）である。またいくつかの説明変数の中で欠測が観察された 5 変数について、その分布をコホートごとに集計したものを表 1 に示す。このデータに対して complete case 解析（解析モデルはポアソン回帰）を行うと、解析に寄与する対象者数は、男性 10,735 人（65.0%）、女性 10,625 人（40.8%）であった。

4　結果

シミュレーション実験の結果、欠測過程がランダムでない状況（not missing at random; NMAR）では、多重補完法は確率的補完法よりもバイアスが小さい場合もあったが、補完モデルによって結果が大きく変化し、complete case 解析よりもバイアスが大きくなる状況も存在した。一方で確率的補完法は常に complete case 解析よりもバイアスは小さかった。

欠測過程がランダムな欠測（missing at random; MAR）の状況に近づくにつれて、確率的補完法ではバイアスが 0 に近付いた一方で、多重補完法ではバイアスが生じた。JALS データに適用した結果を表 2 に示す。脳梗塞発症に対する糖尿病既往のハザード比は、complete case 解析では男性 1.92（95%CI, 1.38-2.69）、女性 1.76（95%CI, 1.15-2.71）であったが、確率的補完法では男性 1.56（95%CI, 1.24-1.89）、女性 1.68（95%CI, 1.22-2.18）、ロジスティックモデルに基づく多重補完法では男性 1.59（95%CI, 1.20-2.11）、女性 1.67（95%CI, 1.18-2.34）、傾向スコアに基づく多重補完法では男性 1.54（95%CI, 1.16-2.04）、女性 1.69（95%CI, 1.21-2.37）であり、complete case 解析の結果と比べて、確率的補完法や多重補完法を用いると点推定値は減少する傾向が見られるものの依然として統計的に有意であった。男性と比べて、女性の方が complete case 解析と、確率的補完法や多重補完法とのハザード比の点推定値の差は小さかった。また、男性の飲酒や喫煙でも糖尿病既往と同じように確率的補完法や多重補完法を用いると点推定値が減少する傾向が見られた。

5　考察

従来の脱落データに対する統計解析手法のうち、説明変数の欠測には容易に拡張可能なものは数少ないが、汎用性があり、なおかつ既存の統計解析パッケージで容易に実行可能な説明変数にも適用できる方法として、多重補完法が挙げられる。欠測値が取りうる全ての値を考慮していないという多重補完法の問題を対処しつつ、多重補完法と同様に既存の統計解析パッケージで容易に実行可能であることは、確率的補完法の利点である。ブートストラップ法を用いて信頼区間を構成したが、対象者数が非常に大きく 1 回の計算に時間がかかる JALS データへの適用を念頭に置いたために反復回数に制限が生じ、やや被覆確率が低くなった。効率のよい信頼区間の構成方法のさらなる検討が必要である。確率的補完法や多重補完法を JALS データに適用した結果、解析に寄与した対象者数は complete case 解析と比べ、男性では約 1.5 倍、女性では約 2.5 倍となり、確率的補完法や多重補完法の脳梗塞発症対する糖尿病既往のハザード比の点推定値は complete case 解析と比べ減少した。しかしながら対象者数が男性と比べて大幅に増加した女性の方が、その点推定値の減少の

程度は小さかった。この相違を検討するために、糖尿病既往の観測の有無を結果変数、脳梗塞発症の有無を説明変数としたロジスティック回帰を男女別に実行したところ、男性では脳梗塞発症による糖尿病既往の観測オッズ比は 0.56（95%CI, 0.47-0.68）である一方で、女性は 1.03（95%CI, 0.83-1.28）であった。糖尿病既往が脳梗塞発症に対するリスク因子であることを考慮すると、女性では脳梗塞発症とは関係なく糖尿病既往の欠測が起きており、脳梗塞発症例に欠測データが多かった男性においてより欠測値の補正の影響が見られたと考えられる。Complete case 解析では多くの対象者が除かれてしまう JALS データのような解析においては、今後本研究で提案した方法を含むいくつかの欠測を考慮した解析を行い complete case 解析との結果を比較し、得られた結果の不確実性を評価する感度解析を行う必要がある。シミュレーション実験では、確率的補完法は欠測過程が MAR の状況で多重補完法よりもバイアスが小さく、共変量が多く測定されている状況で有用であることが示唆された。また多重補完法は complete case 解析よりもバイアスが大きい状況がある一方で、確率的補完法は常に complete case 解析よりもバイアスが小さく、真のデータ発生の過程が分からない現実のデータに対しては、確率的補完法の適用が望まれる。本研究では、二値データの欠測変数を補完する方法として確率的補完法を提案したが、二値データだけでなく多値データや連続量にも拡張可能と考えられ、今後さらなる研究が望まれる。

6　結論

複数の説明変数に欠測値が存在する場合にも適用可能な新しい解析手法である確率的補完法を提案した。シミュレーション実験の結果、提案した方法は、欠測過程が NMAR では多重補完法ほど極端に大きなバイアスが生じず、MAR に近い状況では多重補完法よりもバイアスが小さいことが示された。JALS データに適用した結果、脳梗塞発症に対する糖尿病既往のハザード比は、complete case 解析と比べて確率的補完法や多重補完法では点推定値が減少する傾向が見られ、従来の complete case 解析の結果がハザード比を過大評価している可能性が示唆された。