公的統計における多重代入法の利活用方法の可能性
～諸外国における適用を例に～[1]

高橋　将宜（統計センター）

序論

多重代入法(Multiple Imputation)は、欠測値の対処法として確立された手法だが、我が国の公的統計において活用された実績はない。米国センサス局や国連欧州経済委員会のワークセッションを視察し、米国、ドイツ、スイスにおける多重代入法の利用実態を調査した[2]。また、本稿では、補定(Imputation)と多変量外れ値検出法を組み合わせた自動エディティングを提案し、多重代入法を補定値の診断手法として活用する案を提示する。

1.　米国における欠測値補定の現状

1.1　米国センサス局における経済センサスの補定

経済センサスの補定は、Office of Statistical Methods and Research for Economic Programs にて行われている。基本的に、税務情報など計測された値を用いることが原則だが、埋められない場合、統計モデルにより補定する（上田, 2013）。具体的には、Plain Vanilla というシステムにて比率補定を多用している。米国の経済センサスは単一代入法(Single Imputation)を用いており、多重代入法を導入する予定はない。Rubin の言うところの適切な補定(Proper Imputation)の実装が課題となるだけでなく、計算負荷といった実務上の問題のためである。

1.2　米国センサス局における多重代入法の研究事例

Survey of Income and Program Participation では、現在、確定的ホットデック手法を用いているが、米国センサス局の Center for Statistical Research and Methodology では、TEA という独自ソフトウェアを用いて、ランダムホットデックと「逐次抽出型回帰による多重代入法」(SRMI: Sequential Regression Multiple Imputation)を検証している(García *et al.*, 2014)。SRMI は、本質的には、完全条件付指定(Fully Conditional Specification)と同じ連鎖方程式による多重代入法である(高橋, 伊藤, 2014)。検証の結果、SRMI は、データセット内のどのような変数でもモデルに組み込むことができるという利点があるが、計算負荷が高いという問題がある。

1.3　米国における多重代入法の導入事例

米国センサス局では、多重代入法を実務に導入していないが、National Center for Health Statistics の National Health Interview Survey において、多重代入法を適用して、世帯所得と個人収入の補定を行った(Schenker *et al.*, 2006)。$M = 5$ の補定済みデータを公開している[3]。

2.　ドイツにおける研究と導入事例

ドイツ連邦統計局の Department for Mathematical-Statistical Methods では、2010 年の農業センサスデータを用いて、多重代入法の研究を行っている(Spies *et al.*, 2014)。このデータセットでは、14,213 の農場のうち、欠測値が 2,301 含まれている（欠測率 16.2%）。ランダムホ

---

[1] 本稿の内容は、執筆者の個人的見解を示すものであり、機関の見解を示すものではない。
[2] 本調査結果は、公表論文、現地での質疑応答、メールでのフォローアップなどの情報に基づいている。
[3] http://www.cdc.gov/nchs/nhis/2010imputedincome.htm（2014 年 8 月 12 日アクセス）

ットデックと予測平均値マッチング(PMM: Predictive Mean Matching)を検証した結果、ホットデックでは極端な値が補定値として採用される可能性があるが、PMM では現実的な補定値が採用される可能性が高いことが分かった。

また、研究と並行して、2010 年農業センサスにおいて多重代入法を初めて採用した。予備的な導入であり、非常に限られた範囲での運用だが、研究部門において推定値の算出を行い、結果を農業部門に提出し、そこから欧州統計局(Eurostat)に転送した。ドイツにおいても、多重代入法の手法とソフトウェアについて研究を行っている段階であり、実務における大規模な導入については未定である。

3. スイスにおける導入事例

スイスでは、2010年のStatistics on Income and Living Conditions において、IVEwareを用いて所得の欠測値補定を行った(Swiss Federal Statistical Office, 2012)。スイス連邦統計局のDaniel Kilchmannによると、「複数の補定済みデータセットをどう扱うか、また、これらをユーザーに提供すべきかどうかが実務上の大きな問題になった。また、補定済みデータセットが複数算出されることで、データのボリュームも実務上の問題になった。」という。

4. 日本での導入を目指して：自動エディティングと補定値の診断

諸外国の公的統計においても、多重代入法の研究は行われており、一部では実務に導入されているものの、導入事例はあまり多くない。そもそも、我が国では、多重代入法に関する文献自体が少なく、十分に普及していないといった側面があるが、加えて、諸外国と同様に複数の補定済みデータセットをどう扱うかといった実務上の問題が存在する。

そこで、本稿では、多重代入法を補定手法として使用するのではなく、補定値の診断手法として活用する案を提示する。多重代入法による補定値のゆらぎの大きさに基づいて、補定モデルの安定性と信頼性を検証し、問題のある補定値を外れ値として視覚的に検出する R 関数 `diagimpute` を開発した。2012 年の経済センサス - 活動調査のデータに基づくシミュレーションデータを用いて、R 関数 `diagimpute` の予備的な検証を行っている。また、混淆正規分布モデルによりエラーを検出し、確率的単一代入法によりエラーを訂正し、多重代入法により補定値を評価するという 3 段階の自動エディティング手法を考案中である。

参考文献

[1] García, Matría, Chandra Erdman, and Ben Klemens. (2014). “Multiple Imputation Methods for Imputing Earnings in the Survey of Income and Program Participation,” *Work Session on Statistical Data Editing, United Nations Economic Commission for Europe*, Paris, France, April 28-30 2014.

[2] Schenker, Nathaniel, Trivellore E. Raghunathan, Pei-Lu Chiu, Diane M. Makuc, Guangyu Zhang, and Alan J. Cohen. (2006). “Multiple Imputation of Missing Income Data in the National Health Interview Survey,” *Journal of the American Statistical Association* vol. 101, no. 475, pp.924-933.

[3] Spies, Lydia, Sven Schmiedel, and Katrin Schmidt. (2014). “Simulating Multiple Imputation of Water Consumption in the German Agricultural Census 2010,” *Work Session on Statistical Data Editing, United Nations Economic Commission for Europe*, Paris, France, April 28-30 2014.

[4] Swiss Federal Statistical Office. (2012). “Statistics on Income and Living Conditions (SILC) 2010 Data: Codebook, Description of Microdata,” Swiss Foundation for Research in Social Sciences.

[5] 高橋将宜, 伊藤孝之. (2014).「様々な多重代入法アルゴリズムの比較～大規模経済系データを用いた分析～」, 『統計研究彙報』第 71 号, no.3, pp.39-82.

[6] 上田聖. (2013).「米国センサス局を訪問して」, *Estrela* no.226, pp.30-34.（詳細版：羽渕達志, 上田聖, 小高敦, 高橋将宜, 小泉英希. (2012).「平成 24 年度米国センサス局出張報告」, 独立行政法人統計センター.）