欠測値補定の診断手法としての多重代入法

統計センター　高橋　将宜

1. はじめに

調査観測データには欠測値が付き物である。欠測がランダム(MAR)ならば、補定(imputation)により、データ内の偏りを是正できる (Little and Rubin, 2002)。しかし、補定値は単なる推定値に過ぎず、診断を行う必要性があるが、欠測値に対応する真値は常に欠測しており、補定値を真値と比較して評価できない。これは、補定におけるパラドクスと言える。この問題に対する解決策は、観測データのみに依拠し、欠測の前提を間接的に検証することである(Abayomi *et al.*, 2008)。本稿では、補定プロセスにおいて確率的単一代入法により回帰補定を行い、診断プロセスにおいて多重代入法(multiple imputation)を診断ツールとして使用する新たな方法の提案を行う。

2. 診断アルゴリズムのメカニズム

M 個の多重代入済データセットにおける補定値の変動は、補定における推定不確実性を反映している(Honaker, King, and Blackwell, 2011)。本稿のアルゴリズムは、この変動を利用している。補定モデルは、式(1)のとおり確率的回帰補定であり、診断モデルは式(2)のとおり多重代入法（~は事後分布からの無作為抽出）である。多重代入法のアルゴリズムとしては、計算効率の高い EMB アルゴリズムを採用した(高橋, 伊藤, 2014)。もし補定値が安定しているなら、$\hat{\beta} \approx \tilde{\beta}$と期待される。すなわち、$\mathrm{sd}(\tilde{Y}_{ij}) \approx 0$である。

$$\hat{Y}_{ij} = Y_{i,-j}\hat{\beta} + \sigma_{\hat{u}_i}\hat{\varepsilon}_i \quad (1)$$

$$\tilde{Y}_{ij} = Y_{i,-j}\tilde{\beta} + \tilde{\varepsilon}_i \quad (2)$$

3. 経済データによる例示

表 3.1 と表 3.2 は、企業 100 社の売上高(turnover)と費用(cost)の擬似データであり、ID 3 の企業の売上高が欠測している。表 3.1 では、費用と売上高の線形回帰モデルによる補定値（単一代入値）が 2758 だったとする。100 回の多重代入法(imp1 から imp100)による補定値の平均値も約 2750 だが、標準偏差は 867 と非常に大きい。すなわち、このケースでは、ID 3 の売上高の点推定値として 2758 という補定値を採用して良いか疑問である。

表 3.1: 変動の大きいケース (不確実性 = 高)

ID	turnover	cost	imp1	imp2	⋯	imp99	imp100
1	10514.630	12152.540	10514.630	10514.630	⋯	10514.630	10514.630
2	3272.958	2247.895	3272.958	3272.958	⋯	3272.958	3272.958
3	**欠測**	1038.320	2597.715	4056.021	⋯	1790.942	2834.807
⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮
99	5397.015	4493.931	5397.015	5397.015	⋯	5397.015	5397.015
100	32950.010	37870.540	32950.010	32950.010	⋯	32950.010	32950.010

注：売上高(turnover)と費用(cost)の単位は 100 万円である。

表 3.2 では、自然対数モデルによる補定値（単一代入値）が 1192 だったとする。100 回の多重代入法による 補定値の平均値は 1195 であり、標準偏差は 71 である。すなわち、このケースでは、ID 3 の売上高の点推定値の信頼度は高く、1192 という補定値を点推定値として採用しても大きな問題はないと言える。

表 3.2: 変動の小さいケース (不確実性 = 低)

ID	turnover	cost	imp1	imp2	⋯	imp99	imp100
1	10514.630	12152.540	10514.630	10514.630	⋯	10514.630	10514.630
2	3272.958	2247.895	3272.958	3272.958	⋯	3272.958	3272.958
3	**欠測**	1038.320	1226.377	1152.078	⋯	1295.926	1191.531
⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮
99	5397.015	4493.931	5397.015	5397.015	⋯	5397.015	5397.015
100	32950.010	37870.540	32950.010	32950.010	⋯	32950.010	32950.010

注：売上高(turnover)と費用(cost)の単位は 100 万円である。

4. R 関数 diagimpute

100 回の多重代入法に基づき、補定モデルと補定値の安定性を診断する R 関数 diagimpute を開発し、経済センサス - 活動調査のシミュレーションデータなどを用いて、予備的な検証を行った。

参考文献

[1] Abayomi, Kobi, Andrew Gelman, and Marc Levy. (2008). "Diagnostics for Multivariate Imputations," *Applied Statistics* vol.57, no.3, pp.273-291.

[2] Honaker, James, Gary King, and Matthew Blackwell. (2011). "Amelia II: A Program for Missing Data," *Journal of Statistical Software* vol.45, no.7.

[3] Little, Roderick J. A. and Donald B. Rubin. (2002). *Statistical Analysis with Missing Data*, Second Edition. New Jersey: John Wiley & Sons.

[4] 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.