September 9, 2015

EMBアルゴリズムの新たな応用による多重比率補定(高橋将宜)

Masayoshi Takahashi, *National Statistics Center of Japan*

EMB アルゴリズムの新たな応用による多重比率補定

独立行政法人統計センター　高橋　将宜

1. はじめに

米国センサス局、英国国家統計局、オランダ統計局など、公的統計における欠測値は、ratio imputation (比率補定)により処理されることが多い(高橋, 阿部, 野呂, 2015)。一方、通常の比率補定は、推定不確実性を評価できず、multiple imputation (多重代入法)の使用が推奨されるが、これまで multiple ratio imputation (多重比率補定)の研究はされていない。本研究では、ブートストラップに期待値最大化法を適用する Expectation-Maximization with Bootstrapping (EMB)アルゴリズムに基づく新たな多重比率補定法を提唱する。本報告では、独自に開発した多重比率補定の *R* 関数をシミュレーションデータに適用して検証する。また、実データを用いてその有用性を示す。

2. EMB アルゴリズムによる多重比率補定

多重比率補定における第一段階では、推定不確実性を反映させるために、適切な事後分布から平均値ベクトルの無作為抽出を行う。EMB アルゴリズム(Honaker *et al.*, 2011)は、事後分布からの無作為抽出という複雑なプロセスをノンパラメトリック・ブートストラップに置き換えることで簡略化している。つまり、標本サイズ n の既存の標本データを擬似的な母集団として用い、ここから標本サイズ n の再標本(resample)の無作為な復元抽出を M 回実行し、補定済みデータを M 回生成することで、推定不確実性を反映させる。しかし、不完全データからのブートストラップ再標本もまた不完全である可能性が高い。

第二段階では、EM アルゴリズムにより推定値の改善を図る。つまり、平均値、分散、共分散等のパラメータ初期値を設定し、期待値ステップにおいて観測データとパラメータ推定値を条件として完全データの十分統計量の期待値を推定し、最大化ステップにおいて完全データの推定された十分統計量の値をもとにパラメータ推定値を更新する。これらのステップを繰り返して収束した値は最尤推定値となることが知られている。

多重比率補定モデルは、$\tilde{Y}_{i1} = \tilde{\omega} Y_{i2} + \tilde{\varepsilon}_i$であり、ここで~ (tilde)は欠測データの適切な事後分布からの無作為抽出値であることを表す。すなわち、このモデルでは、切片が 0 であり、$\tilde{\omega}$は適切な事後分布から無作為抽出された平均値の比率のベクトルとして傾きを表すことにより推定不確実性を反映させており、$\tilde{\varepsilon}_i$は誤差項として根本的な不確実性を反映させている。なお、ブートストラップ再標本に EM アルゴリズムを適用して得られた最尤推定値は、ベイズ統計における事後分布からの無作為抽出による推定値と漸近的に等価である(Little & Rubin, 2002)。

3. モンテカルロ実験

比推定のモデルが正しいモデルとなるような任意の平均値ベクトルと分散・共分散行列からなる 2 次元正規分布から生成した 2 変数を用い、1,000 回のモンテカルロ実験を行った。各々の実験において、標本サイズを 50～1,000 までの 5 パターンを用意し、平均欠測率は 10%～40%の範囲で 3 種類を用意した。欠測メカニズムは、MCAR、MAR、NI の 3 種類である。結果は、RMSE (Root Mean Square Error)により評価する。実験結果の詳細については、当日報告する。

4. ソフトウェア *MrImputation*

独自に開発した *R* 関数 `mrimpute` と `mranalyze` から構成され、*R* 関数 `mrimpute` は多重比率補定を実行し、*R* 関数 `mranalyze` は多重比率補定済みデータを用いた統計解析を実行する。

参考文献

[1] Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: a program for missing data. *Journal of Statistical Software*, 45(7), 1-47.

[2] Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, second edition. Hoboken, NJ: John Wiley & Sons.

[3] 高橋将宜, 阿部穂日, 野呂竜夫. (2015).「公的統計における欠測値補定の研究:多重代入法と単一代入法」,『製表技術参考資料』第 30 号, pp.1-95.