Hot Deck 補完の下での
標本平均の漸近正規性について

青山学院大学経済学部准教授
元山　斉

1. はじめに

多くの社会・経済データやそれらの集計データの基礎となるデータの多くは対象となる母集団に対する全数または標本調査に基づいて作成されている。しかしながら、調査においては、調査対象者の不在ないしは調査拒否等による調査対象者からの回答を得られないこと(unit non-response)や個別の質問項目がsensitiveであることや質問文の意図が分かりにくいこと等による無回答による欠損値が存在する(item non-response)。

欠損値に対処する方法として、統計作成の実務においては長い間、補完(補定、代入、imputation)と呼ばれる手法で、欠損値に対してもっともらしいと考えられる数値を代入する方法が長い間用いられてきた。補完の方法については、回答が得られたレコードの平均値を代入する方法(平均値補完)や関心のある変数と相関の高い補助変数による回帰の予測値を代入する方法(回帰補完)などが提案されているが、それらの中でもHot Deck補完は欠損値をその欠損値が含まれるデータセット内の、欠損値に何らかの意味で似ていると考えられる完全データの値によって補完する方法であり、マッチング代入とよばれる補完方法において欠損値と同じデータセットを用いる特別な場合と考えることもできる。

欠損値補完を行ったデータセットに基づいて統計的推論を行う際には、統計量の分布の標準偏差や分布そのものについて知ることが重要となるが、Abadie and Imbens(2012)は、統計的因果推論でしばしば用いられるMatching Estimatorがマルチンゲールという性質を持った確率変数列から構成されることを示し、マルチンゲール中心極限定理に基づいてMatching Estimatorの漸近正規性を証明した。また、その導出の際に用いられた考え方がCell Hot Deckとよばれる方法で補完を行った際の標本平均にも適用できることを示しCell Hot Deck補完を行った際の標本平均の漸近正規性を証明した。本稿はAbadie and Imbens(2012)の当該箇所の概略をまとめたノートである。本稿は上記論文の数理的行間の一部を埋めた以外は、理論的にも視点的にも新しい結果の提示をしているわけではなく、上記論文の流れに沿った紹介にとどまっているが、今後、応用面でさらに有益な結果を導出するための足掛かりとなることを意図している。

2. Cell Hot Deck補完とその下での標本平均の漸近正規性

いま関心のある変数をY、その変数と関係の強い共変量をXとし、大きさNの標本$(Y_i, X_i), i = 1, \dots, N$が調査によって調べられたとする。しかしながら$Y_i, i = 1, \dots, N$の一部には何らかの理由で欠損値が生じているとする。このようなとき、Yの値と関係のある共変量Xの値の同じセル内の回答された完全なレコードの値で補完することがしばしば行われる。

いま共変量Xの値に基づいて、観測値の全体が重なることのないT個のクラス(セル)$C_1, \dots, C_t, \dots C_T$に分割されているとする。そして$Y$の欠損値は同じセル内の完全なレコードによって補完されるとする。

Hot Deck補完には上記のCell Hot Deck以外の方法も存在するが（Andridge and Little(2010)）、応用上、最も多く用いられ一般的な枠組みと考えられるCell Hot Deckについて取り扱うことにする。

Wを定義関数(指標関数)でYが観測されたとき、すなわち完全データであったときに１、それ以外に０という値をとるとする。ここで、$X \in C_t$を与えたときにYは(X, W)と独立で同一の分布であると仮定する(cell mean assumption)、この仮定から同一セル内のYの分布の平均は共通であることがわか

る。また、標本抽出は単純無作為抽出を仮定する。

いま、$\mu = \mathrm{E}[Y]$、$\mu(x) = \mathrm{E}[Y|X = x]$、$\mu_t = \mathrm{E}[Y|X \in C_t]$、$\sigma_t^2 = \mathrm{V}(Y|X \in C_t)$とする。また、$j(i)$を$i$番目の観測値に対する Y の値を補完するのに用いられた観測値の添え字とする(i番目の観測値のYの値が観測されたときは$j(i) = i$となる)。上記の補完を行った上でのYの標本平均を$\bar{Y}$とすると

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_{j(i)} = \frac{1}{N}\sum_{i=1}^{N} W_i\left(1 + K_{N,i}\right)Y_i \tag{1}$$

と表すことができる。ここで$K_{N,i}$は不完全なレコードを補完するためにi番目の観測値が何回用いられたかを表す数である。(Note:$K_{N,i}$の値は補完がどのように行われるか、例えば同じセルの完全なレコードから無作為に選ぶ方法(random cell hot deck)や、同じセル内で直前に観測された完全なレコードで補完する方法(sequential cell hot deck)によって変わる。)

ここで(1)式および$\sum_{i=1}^{N} W_i\left(1 + K_{N,i}\right)\mu(X_i) = \sum_{i=1}^{N} \mu(X_{j(i)})$より、

$$\bar{Y} - \mu = \frac{1}{N}\sum_{i=1}^{N}(Y(X_i) - \mu) + \frac{1}{N}\sum_{i=1}^{N} W_i\left(1 + K_{N,i}\right)(Y_i - \mu(X_i)) + \frac{1}{N}\sum_{i=1}^{N}(\mu(X_{j(i)}) - \mu(X_i)) \tag{2}$$

が成立する。ここで cell mean assumption より、すべてのiに対して$\mu\left(X_{j(i)}\right) - \mu(X_i) = 0$ となるので(2)式の右辺第 3 項は 0 となる。いま、セルC_t に含まれる要素の数をN_tとして、$K_{N,i}$ の 2 次のモーメントの存在($\mathrm{E}\left[K_{N,i}^2\right] < \infty$)を仮定し、

$$\sigma^2 = \mathrm{E}\left[\sum_{t=1}^{T}\left(\frac{N_t}{N}\right)(\mu_t - \mu)^2\right] + \mathrm{E}\left[\sum_{t=1}^{T}\left(\frac{N_t}{N}\right)\sigma_t^2 \sum_{i=1}^{N} 1_{\{X_i \in C_t\}} W_i\left(1 + K_{N,i}\right)^2\right] \tag{3}$$

とおき

$$\xi_{N,k} = \begin{cases} \frac{1}{\sigma\sqrt{N}}(Y(X_k) - \mu) & 1 \le k \le N \\ \frac{1}{\sigma\sqrt{N}} W_{k-N}\left(1 + K_{N,k-N}\right)\left(Y_{k-N} - \mu(X_{k-N})\right) & N+1 \le k \le 2N \end{cases} \tag{4}$$

と定義すると、(2)より

$$\frac{\bar{Y} - \mu}{\sigma / \sqrt{N}} = \sum_{k=1}^{2N} \xi_{N,k} \tag{5}$$

と表すことができる。

ここで、$\chi_N = \{X_1, \ldots, X_N\}$ と $\omega_N = \{W_1, \ldots, W_N\}$ とおき、σ-加法族の列 $F_{N,k}$ を $F_{N,k} = \sigma\{W_1, \ldots, W_k, X_1, \ldots, X_k\}(1 \le k \le N)$, $F_{N,k} = \sigma\{\chi_N, \omega_N, Y_1, \ldots, Y_k\}(N+1 \le k \le 2N)$と定義すると、$F_{N,k}$は単調非減少なσ-加法族の列(フィルトレーション)となる。ここでσ{ }は、{ }内の確率変数を可測にするような最小のσ-加法族である。

このとき、$\sum_{j=1}^{i} \xi_{N,i}\,(1 \le i \le 2N)$は$F_{N,i}(1 \le i \le 2N)$についてマルチンゲールとなる。よって、マルチンゲール中心極限定理と漸近分散が 1 となることを示すことができることから

$$\frac{\bar{Y} - \mu}{\sigma/\sqrt{N}} \xrightarrow{d} N(0,1) \tag{6}$$

を得る。ここで$\xrightarrow{d}$は分布収束を表す。Abadie and Imbens(2012)の得た結果を定理の形でまとめると以下となる(Abadie and Imbens(2012), Theorem2)。

定理（1）$\{X_1, \ldots, X_N\}$は母集団から無作為に抽出され、(2) $P(W = 1|X \in C_t) > 0, t = 1, \ldots, T$、(3) $X \in C_t (t = 1, \ldots, T)$を与えたときに$Y$は$(X, W)$と独立で、(4)$\mathrm{V}(\mathrm{Y}) > 0$で、(5) ある$\delta > 0$に対して$\mathrm{E}[|\mathrm{Y}|^{2+\delta}] < \infty$であるとき、(6)が成立する。

最後にσ^2の推定方法を考える。最初にYについて欠損値を補完したことを考慮に入れない通常の分散推定量

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{j(i)} - \bar{Y}\right)^2 \tag{7}$$

を考える。各セルごとの$Y_{j(i)}$の平均を$\bar{Y}_{\mathrm{t}} = \sum_{i=1}^{N} Y_{j(i)}\, 1_{\{X_i \in C_t\}}/\mathrm{N_t}$とおく。各セルを分散分析における個別母集団(級)と考えると上の分散は級間分散(セルの間の平均値$\bar{Y}_{\mathrm{t}}$の分散)と級内分散(各セル内の$Y_{j(i)}$の値の分散)に分解されることから

$$\left| \hat{\sigma}^2 - \sum_{t=1}^{T}\left(\frac{N_t}{N}\right)(\mu_t - \mu)^2 - \sum_{t=1}^{T}\left(\frac{N_t}{N}\right)\sigma_t^2 \right| \xrightarrow{p} 0 \tag{8}$$

を得る。ここで$\xrightarrow{p}$は確率収束を表す。σ^2の定義式((3)式)と(8)式左辺の絶対値内の第 2 項、第 3 項を比較することより$\hat{\sigma}^2$に$\sum_{t=1}^{T}\left(\frac{N_t}{N}\right)\sigma_t^2\left(\sum_{i=1}^{N} 1_{\{X_i \in C_t\}} W_i\left(1 + K_{N,i}\right)^2 - 1\right)$と漸近的に同等な項を加えることで$\sigma^2$の一致推定量が構成されることが予想される。

ここで、$\sum_{i=1}^{N} 1_{\{X_i \in C_t\}} W_i\left(1 + K_{N,i}\right) = N_t$より

$$\frac{1}{N_t}\sum_{i=1}^{N} 1_{\{X_i \in C_t\}} W_i\left(1 + K_{N,i}\right)^2 = 1 + \frac{1}{N_t}\sum_{i=1}^{N} 1_{\{X_i \in C_t\}} W_i\left(K_{N,i}^2 + K_{N,i}\right) \tag{9}$$

が成り立つことからσ^2の推定量として、

$$\begin{aligned}\hat{\sigma}^2_{\mathrm{adj}} &= \hat{\sigma}^2 + \frac{1}{N}\sum_{t=1}^{T}\left(\sum_{i=1}^{N} 1_{\{X_i \in C_t\}} W_i\left(K_{N,i}^2 + K_{N,i}\right)\right)\hat{\sigma_t}^2 \\ &= \hat{\sigma}^2 + \sum_{t=1}^{T}\left(\frac{N_t}{N}\right)\left(\frac{1}{N_t}\sum_{i=1}^{N} 1_{\{X_i \in C_t\}} W_i\left(K_{N,i}^2 + K_{N,i}\right)\right)\hat{\sigma_t}^2\end{aligned} \tag{10}$$

が提案される。ここで、$\hat{\sigma_t}^2$はセルC_t内の観測されたレコードによるYの分散の推定量である。

3. まとめ

本稿では Abadie and Imbens(2012)によって示された Cell Hot Deck 補完を行った際の標本平均の漸近正規性の概略を記した。彼らの結果は類似の補完方法に容易に拡張が可能な汎用性の高いものであり、今後、彼らの結果をより一般的な標本抽出の枠組みや標本平均以外の統計量に拡張することで、応用上の可能性がさらに広がっていくものと期待される。

参考文献

Abadie, A and G.W. Imbens(2012). A Martingale Representation for Matching Estimators. *Journal of the American Statistical Association*, **107(498)**, 833-843.

Andridge, R.R. and R.J.A. Little(2010). A Review of Hot Deck Imputation for Survey Non-response. *International Statistical Review*, **78(1)**, 40-64.