Mechanism of Missing Data Analysis

大阪大・基礎工　狩野　裕

欠測データ解析の仕組みの理解については今だに混乱があるように見受けられる。統計学理論に習熟しないと正確な理解に到達できないのではないか。たとえば，(Y_{obs}, Y_{mis}) という記法に戸惑う。この記法は統計ユーザの直観的理解を助けたが，背景理論をおさえたい学者や実務家を混乱させている。MAR(misssing at random) とは，観測変数を与えた下で欠測変数と欠測 (指標) とが独立である，という誤解がある。強い意味で識別可能 (strongly ignorable) と MAR との関係に悩む。筆者は，欠測という状況が（数学でいう）可測とどのような関係があるのか，よくわからなかった。本講演の前半では，欠測データ解析の基礎的な事項や概念をレビューし，正確な理解を確認する。

後半では，欠測による推定量の (漸近) バイアス評価の重要性を指摘する。欠測値データ解析に関する重要な成果である二重頑健推定法 (Bang & Robins, 2005) を考える。観測ベクトル (Y, X) において Y が欠測し得るとし，真の傾向スコアと回帰を $E(R_y|X)$，$E(Y|X)$ とする (R_y は欠測指標変数)。このとき，$\mu_y = E(Y)$ の二重頑健推定量 $\hat{\mu}_y^{(DR)}$ は，$E(R_y|X)$ または $E(Y|X)$ のどちらかを正しく特定することができれば (漸近) 不偏推定量になる，という魅力的な性質を有する。二重頑健性の証明は，推定量のバイアスを

	真	モデル
傾向スコア	$E(R_y\|X)$	$e(X)$
回帰	$E(Y\|X)$	$m(X)$

$$\hat{\mu}_y^{(DR)} - \mu_y = E_X\left[\frac{\Big\{E(R_y|X) - e(X)\Big\}\Big\{E(Y|X) - m(X)\Big\}}{e(X)}\right] + o_p(1)$$

と表現すると，よく理解できる。この公式から二重頑健推定量の性質に接近できる。

尤度に基づく推定量 $\hat{\theta} := \underset{\theta\in\Theta}{\text{argmax}}\ L(\theta|Y_{obs})$ は MAR の下で漸近不偏であることが証明されている。では MAR が満たされないときどのようなバイアスが生じるのか。そのバイアスは MAR の満たされない程度と関係するのではないか。バイアスを減少させることが期待される補助変数 (auxiliary variable) の導入は有効なのか。Kano(2015) の結果を中心に推定量のバイアス評価の方法を展開する。

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, Vol.61(4), 962-973.

Hirose, K., Kim, S., Kano, Y. et. al. (2016). Full information maximum likelihood estimation in factor analysis with a large number of missing values. *Jour. Statist. Comp. Sim.*, Vol.86(1), 91-104.

Kano, Y. (2015/July). Developments in multivariate missing data analysis. Keynote Lecture at the IMPS2015. Beijing, China.

Kano, Y. and Takai, K. (2011). Analysis of NMAR missing data without specifying missing-data mechanisms in a linear latent variate model. *Jour. Multi. Anal.*, Vol.102(Oct), 1241-1255.

Morikawa, K., Kim, J.-K. and Kano, Y. (2017). Semiparametric maximum likelihood estimation under not missing at random. *Canadian Journal of Statistics*. Vol.45(4), 393-409.

Takai, K. and Kano, Y. (2013). Asymptotic inference with missing data. *Commun. Statist. Theo. Meth.*, Vol.42(17), 3174-3190.

Takagi, Y. and Kano, Y. (in press). Bias reduction using surrogate endpoints as auxiliary variables. *Ann. Inst. Statist. Math.*

Yuan, K.-H., Jamshidian, M. and Kano, Y. (2018). Missing data mechanisms and homogeneity of means and variances-covariances. *Psychometrika*, Vol.83(2), 425-442.

狩野 裕 (2014)．NMAR の下での尤度法．日本統計学会誌，43 巻 2 号，359-377.

高井啓二・星野崇宏・野間久史 (2016)．欠測データの統計科学．岩波書店.