Semiparametric Inference under Nonignorable Nonresponse

大阪大・院・基礎工　　森川 耕輔
Iowa State U., Dept. Statist.　　Jae Kwang Kim
大阪大・基礎工　　狩野 裕

1. はじめに

データの "欠測" は疫学，経済学，標本調査といったあらゆる分野で重大な問題となっている．例えば，結果変数 (Y) の平均 $E(Y)$ が興味の対象であるとする．このような統計量は，一般にある関数 $U(\cdot)$ を用いて，推定方程式 $E\{U(\boldsymbol{\theta};\boldsymbol{X},Y)\}=\boldsymbol{0}$ を満たすパラメータ $\boldsymbol{\theta}$ のクラスに属する．ここで，$\boldsymbol{X}$ は共変量である．Y が欠測している際，$E(Y)$ といった単純なパラメータでさえ，通常用いられる標本平均ではバイアスが生じてしまう場合がある．このような問題を扱う際，データが欠測する原因である，欠測メカニズムの特定が重要となる．欠測メカニズムは，$\pi(\boldsymbol{x},\ y):=P(\delta=1\mid\boldsymbol{x},y)$ で定義される．ここで，δ は Y が観測される (欠測する) 場合，1(0) をとる確率変数である．特に，欠測メカニズムが Y に依存するとき，無視できない欠測と呼ばれる．

2. 従来法

欠測メカニズムが正しく特定されれば，IPW(Inverse Probability Weightied) 推定量は $\boldsymbol{\theta}$ に対する一致性，漸近正規性を持つ推定量となる (Tsiatis, 2006)．しかし，一般には欠測メカニズムは未知であるため，$\pi(\boldsymbol{x},y;\boldsymbol{\phi})$ を正しく特定する必要がある．その際，欠測メカニズムを規定するパラメータ $\boldsymbol{\phi}$ の推定量は通常，尤度に基づき次の推定方程式の解として得られる:

$$\sum_{i=1}^{n}\left[\delta_i\boldsymbol{s}(\boldsymbol{\phi};\delta_i,\boldsymbol{x}_i,y_i)+(1-\delta_i)E_0\{\boldsymbol{s}(\boldsymbol{\phi};\delta_i,\boldsymbol{x}_i,Y)\mid\boldsymbol{x}_i\}\right]=\boldsymbol{0}.$$

ここで，$\boldsymbol{s}(\boldsymbol{\phi};\delta,\boldsymbol{x},y)=\frac{\delta-\pi(\boldsymbol{x},y;\boldsymbol{\phi})}{\pi(\boldsymbol{x},y;\boldsymbol{\phi})\{1-\pi(\boldsymbol{x},y;\boldsymbol{\phi})\}}\frac{\partial\pi(\boldsymbol{x},y;\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}$ であり，$E_0(\cdot\mid\boldsymbol{x})=E(\cdot\mid\boldsymbol{x},\delta=0)$ は $\boldsymbol{X},\delta=0$ の下での Y の条件付き期待値である．従って，この期待値を計算するためには，本来興味が無い $Y\mid\boldsymbol{X}$ の分布 $f(y\mid\boldsymbol{x})$ を正しく特定する必要がある．また，$f(y\mid\boldsymbol{x})$ は観測されるデータ ($\boldsymbol{X}$ と $\delta=1$ のときの Y) のみから欠測部分を含めた全体の分布型を特定しなければならないため，得られた推定量は客観性に欠ける．

そこで，Riddles et al. (2015) は条件付き期待値 $E_0(\cdot\mid\boldsymbol{x})$ を上手く変形することで，特定が必要なモデル $f(y\mid\boldsymbol{x})$ を $f_1(y\mid\boldsymbol{x}):=f(y\mid\boldsymbol{x},\delta=1)$ に緩めても $\boldsymbol{\theta}$ に関する一致性，漸近正規性を持つ推定量が得られることを示した．$f_1(y\mid\boldsymbol{x})$ は現在観測されているデータ ($\boldsymbol{X}$ と $\delta=1$ のときの Y) のみを用いてモデルが構築可能なため，上で述べた問題は避けられる．しかしながら，$f_1(y\mid\boldsymbol{x})$ は観測された対象者のみに対する条件付き分布であり，複雑な分布となることが想定される．

3. 提案モデル

本研究では，条件付き期待値 $E_0(\cdot\mid\boldsymbol{x})$ をカーネル平滑化によりノンパラメトリックに推定することで，$Y\mid\boldsymbol{X}$ のモデルに関する仮定を全く必要としない $\boldsymbol{\phi}$ 及び $\boldsymbol{\theta}$ に対する推定量を提案し，それらの漸近分布を導出した．即ち，我々の提案する方法により，欠測メカニズムが結果変数 Y に依存する無視できない欠測であったとしても，$Y\mid\boldsymbol{X}$ の分布に対する仮定を必要とせず，欠測メカニズムに関する仮定のみで $\boldsymbol{\theta}$ を推定することが可能となる．

4. 参考文献

[1] Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data.* Springer.

[2] Riddles, M. K., Kim, J. K. and Im, J. (2015). Propensity-score-adjustment method for nonignorable nonresponse (submitted).