統計数理(2014)
第 62 巻 第 1 号 103–122

特集「疫学研究のデザインとデータ解析：最近の理論的展開と実践」
[総合報告]

欠測データに対するセミパラメトリックな解析法

―その理論的背景について―

逸見 昌之†

(受付 2013 年 12 月 10 日；改訂 2014 年 3 月 12 日；採択 3 月 12 日)

要　　旨

疫学におけるコホート研究のような観察研究では，様々な要因により統計解析の結果に偏りが生じる可能性があるが，そのうちの 1 つはデータの欠測によるものである．統計的な欠測データの解析法には，大きく分けて最尤法や多重代入法などのパラメトリックな手法と，逆確率重み付け法や二重頑健推定法などのセミパラメトリックな手法があるが，本稿では後者の方に焦点を当てる．後者の方法は，近年主に医学統計学の分野で急速に発展し，実際の適用例はまだそれほど多くはないが，有力な方法として注目されており，今後その需要はさらに高まってくるものと思われる．セミパラメトリックな統計手法は，特に医学統計学の分野では他にも様々な場面で用いられているが，本稿ではその理論的背景の理解にも資するように，セミパラメトリック推測の一般論に基づいた解説を行う．

キーワード：MAR，影響関数，推定関数，局外接空間，漸近有効性．

1. はじめに

疫学におけるコホート研究のような観察研究では，様々な要因により統計解析の結果に偏りが生じる可能性がある． 例えば，ある曝露(喫煙など)がある疾患(肺癌など)の原因であるかどうかを調べるような因果推論における交絡などはその代表的なものであるが，データの欠測もまた，結果にバイアスをもたらし得る要因として考慮すべきものである．コホート研究などの疫学研究では，その参加者から常に(最後まで)データが観測できるとは限らない． このとき，もし欠測の仕方がその研究で対象としている結果変数(ある疾患に罹るかどうかなど)に影響のあるものであれば，つまり，例えば二群比較の際，曝露群において悪い結果となる参加者ほど欠測しやすい傾向にあれば，欠測を考慮しない解析では曝露効果を過小評価してしまう．欠測データを取り扱うための統計的方法は，大きく分けて 2 種類のものがある．1 つは最尤法や多重代入法(Rubin, 1987; Carpenter and Kenward, 2013)などのパラメトリックな手法であり，もう 1 つは逆確率重み付け法(inverse probability weighting; Robins et al., 1994; Seaman and White, 2011)や二重頑健推定法(doubly robust estimation; Scharfstein et al., 1999; Bang and Robins, 2005)などと呼ばれるセミパラメトリックな手法である． 後者の方法は近年，主に医学統計学の分野で急速に発展し，実問題への適用例はまだそれほど多くはないが，有力な方法として注目されており，今後その需要はさらに高まってくるものと思われる．本稿の主な目的は，これらの方法をセミパラメトリック推測の一般論の観点から解説し，その理論的な背景を明らかにすること

†統計数理研究所：〒 190–8562 東京都立川市緑町 10–3

である．数学的な議論の詳細(厳密な正則条件等)にはあまり深く立ち入らず，手法導出の道筋をはっきりさせることに重点を置くが，特に理論的な立場から，二重頑健推定法の仕組みとその役割を理解することを 1 つの目標とする．

本稿の構成は以下の通りである． 第 2 節ではまず，統計的な欠測データ解析の出発点として欠測メカニズムの分類について述べ，単純な方法の問題点に触れた後，パラメトリックな解析法として最尤法について簡単に述べる．そして，それと対比しながらセミパラメトリックな解析法の必要性について触れ，ごく簡単な場合に，逆確率重み付け法と二重頑健推定法を紹介する． 続いて第 3 節では，理論的な取り扱いの準備として，必要な範囲でセミパラメトリック推測の一般論の概説を行う． そしてそれをもとに第 4 節では，二重頑健推定量の導出を中心に，セミパラメトリックな欠測データ解析法について理論的な立場から解説する．最後に第 5 節として，注意事項や関連事項等について述べる．

2. 欠測データ解析の基本事項

2.1 欠測メカニズムの分類

欠測データの解析を行う際に最も留意すべきことの 1 つは，データの欠測がどのようなメカニズムによって生じているかということである．ここでは，欠測データ解析の出発点としてまず，Rubin(1976)によって提唱されて以来，広く受け入れられている欠測メカニズムの 3 つの分類について述べる． ある(疫学)研究において，n 人の対象者から m 次元の多変量データ

$$\boldsymbol{Z}_1 = (Z_{11}, \ldots, Z_{1m}), \ldots, \boldsymbol{Z}_n = (Z_{n1}, \ldots, Z_{nm}) \tag{2.1}$$

を測定する状況を想定し，これらはある確率分布からのランダムサンプル(互いに独立で同一の確率分布に従う確率ベクトル)であるとする．但し，これらのデータは必ずしも全て観測されるとは限らないとし，例えば，対象者 1 に対しては Z_{11} から Z_{1m} までが観測されるが，対象者 n に対しては Z_{n1} と Z_{n2} しか観測されないということが起こり得るとする． このとき，観測の指示変数 R_{ij} を，Z_{ij} が観測されれば 1, 欠測すれば 0 の値を取る確率変数として定義し，対象者 i に対する観測データ($\boldsymbol{Z}_i$ の各成分のうち，実際に観測される部分)を $\boldsymbol{Z}_i$ と $R_{i1}, \ldots, R_{im}$ の関数として $O(\boldsymbol{Z}_i, R_{i1}, \ldots, R_{im})$ と表すことにすると，データの欠測メカニズムは以下のように分類される(Rubin, 1976; Little and Rubin, 2002)．

- 完全にランダムな欠測(Missing Completely At Random, MCAR)
  $\forall r_1 \in \{0,1\}, \ldots, \forall r_m \in \{0,1\}$,
  $P(R_{i1} = r_1, \ldots, R_{im} = r_m | \boldsymbol{Z}_i) = P(R_{i1} = r_1, \ldots, R_{im} = r_m)$
- ランダムな欠測(Missing At Random, MAR)
  $\forall r_1 \in \{0,1\}, \ldots, \forall r_m \in \{0,1\}$,
  $P(R_{i1} = r_1, \ldots, R_{im} = r_m | \boldsymbol{Z}_i) = P(R_{i1} = r_1, \ldots, R_{im} = r_m | O(\boldsymbol{Z}_i, r_1, \ldots, r_m))$
- ランダムでない欠測(Not Missing At Random, NMAR)
  $\exists r_1 \in \{0,1\}, \ldots, \exists r_m \in \{0,1\}$ $s.t.$
  $P(R_{i1} = r_1, \ldots, R_{im} = r_m | \boldsymbol{Z}_i) \neq P(R_{i1} = r_1, \ldots, R_{im} = r_m | O(\boldsymbol{Z}_i, r_1, \ldots, r_m))$

「完全にランダムな欠測(MCAR)」とは，文字通りデータの欠測が完全にランダムに生じていて，上述のようにデータの各観測(あるいは欠測)パターンの確率がデータそのものの値には全く依らないことを意味する．この場合は，観測されたデータのみ用いて解析を行っても統計的推測の結果にバイアスは生じないが(但しデータの不足による精度の低下は起こる)，データの欠測を意図的に制御しない限り実際の場面では稀である． 2 番目の「ランダムな欠測(MAR)」

は，しばしば「観測（あるいは欠測）確率が観測データのみに依存する」と言い表されるが，より正確に表すと上述のようになる．最も単純な場合として，$m = 2$ で Z_{i1} の方は常に観測されるという状況を考えてみると，この場合，観測の指示変数は R_{i2} の方だけ考えればよく，MAR は $P(R_{i2} = r|Z_{i1}, Z_{i2}) = P(R_{i2} = r|Z_{i1})$ $(r = 0, 1)$ と表される．これはつまり，Z_{i2} が観測されるかどうかの傾向は（常に観測される）Z_{i1} の値のみによって決まることを意味している．例えば，ある集団に対する健康診断において血圧の値を（日を改めて）2 回に分けて測定する場合，1 回目の血圧 Z_{i1} は全員に対して測定されるが，2 回目の血圧 Z_{i2} を測定するかどうかは 1 回目の結果を見て健診対象者自身が決められるものとすると，Z_{i2} の欠測のメカニズムは MAR であると考えられる． 最後の「ランダムでない欠測（NMAR）」は，文字通り MAR でないということであるが，これはつまりデータのある観測パターンの確率が，観測データだけでなく欠測データの値にも依存してしまうということを意味する．

統計的な欠測データ解析法の多くは MAR の仮定の下で構築されているが，データの欠測メカニズムが MAR であるかどうかは観測されたデータだけからは（原理的に）検証できない．したがって，データ取得のデザインなどから MAR であることが保証されていない限りは，まず MAR を仮定して解析を行ったとしても，その仮定からのずれの影響を調べる感度解析を行うことが望ましいとされている．本稿における手法の解説は基本的に MAR の仮定の下で行っていくが，感度解析などについては，例えば第 5 節で挙げる文献等を参照されたい．

2.2 単純な方法の問題点

近年までの疫学研究では，欠測データの取り扱い方として多くの場合，Complete-Case 解析や単一代入法といった単純な方法が用いられているが（Eekhout et al., 2012），これらは非常に限られた状況でしか正当化されない．例えば，上述の血圧測定の例において（記号を見やすくするために）$X_i = Z_{i1}$, $Y_i = Z_{i2}$, $R_i = R_{i2}$ $(i = 1, \ldots, n)$ とおき，2 回目の血圧 Y_i の平均 $\mu = E(Y_i)$ に興味があるとすると，Complete-Case 解析による μ の推定量は

$$\hat{\mu}_{CC} = \frac{1}{N}\sum_{i=1}^{n} R_i Y_i \quad \left(N = \sum_{i=1}^{n} R_i\right)$$

で与えられるが，これは欠測メカニズムが MCAR でないと（漸近的に）バイアスが生じ得る．実際，大数の法則より（$n \to \infty$ のとき）$\hat{\mu}_{CC}$ は $E(Y_i|R_i = 1)$ に確率収束し，漸近バイアス

$$E(Y_i|R_i = 1) - E(Y_i) = \{E(Y_i|R_i = 1) - E(Y_i|R_i = 0)\}(1 - p) \quad (p = P(R_i = 1))$$

は MCAR 以外の場合では 0 とは限らない．一方，単一代入法については，例えば回帰代入法による μ の推定量は以下のように与えられる．

$$\hat{\mu}_{RI} = \frac{1}{n}\sum_{i=1}^{n}\{R_i Y_i + (1 - R_i)m(X_i; \hat{\boldsymbol{\beta}})\}. \tag{2.2}$$

ここで $\hat{\boldsymbol{\beta}}$ は，ある回帰モデル $E(Y_i|X_i) = m(X_i; \boldsymbol{\beta})$（例えば線型回帰なら，$m(X_i; \boldsymbol{\beta}) = \beta_0 + \beta_1 X_i$, $\boldsymbol{\beta} = (\beta_0, \beta_1)$）に対して，$\boldsymbol{\beta}$ を「観測データだけ」を用いて推定したときの（最小 2 乗）推定量である．つまり，推定量 $\hat{\mu}_{RI}$ は，Y_i が欠測しているところについては回帰モデルに基づく予測値で補完してから，全体で標本平均を取ったものである．このとき，欠測のメカニズムが MAR で，かつ回帰モデルが正しく特定されていれば，$\hat{\mu}_{RI}$ は（μ に対して）一致性を持ち漸近的なバイアスは 0 となる．しかしながら，推定の対象を例えば Y_i の 2 次モーメント $\nu = E(Y_i^2)$ に置き換えると，回帰代入法によるその推定量

$$\hat{\nu}_{RI} = \frac{1}{n}\sum_{i=1}^{n}[R_i Y_i^2 + (1-R_i)\{m(X_i;\hat{\boldsymbol{\beta}})\}^2]$$

は，MAR の下で回帰モデルが正しく特定されていても，もはや一致性を持たない．回帰代入法以外にも平均値代入法や Hot deck などの方法があるが，一般にこのような単一代入法は正当性が非常に限られるので，注意が必要である．

2.3 パラメトリックな解析法 ―最尤法―

冒頭でも述べたように，本稿の主な目的は欠測データに対するセミパラメトリックな手法について解説することであるが，その前に，より基本的なパラメトリックな手法として，最尤法について簡単に述べておく．

多変量データ(2.1)に対して $\boldsymbol{Z}_i$ の分布(確率密度関数)が，あるパラメトリックモデル

$$\mathcal{S} = \{p(\boldsymbol{z};\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \Theta \subset \boldsymbol{R}^q\}$$

に属していると仮定し，$\boldsymbol{Z}_i$ の各成分の(条件付き)同時観測確率もまた

$$P(R_{i1} = r_1, \ldots, R_{im} = r_m | \boldsymbol{Z}_i = \boldsymbol{z}) = \omega(r_1, \ldots, r_m, \boldsymbol{z}; \boldsymbol{\psi}) \ (\boldsymbol{\psi} \in \Psi \subset \boldsymbol{R}^r)$$

と(パラメトリックに)モデル化されているとすると，全対象者の観測データ

$$\boldsymbol{Z}_i^o = O(\boldsymbol{Z}_i, R_{i1}, \ldots, R_{im}), R_{i1}, \ldots, R_{im} \quad (i = 1, \ldots, n)$$

に基づくパラメータ $\boldsymbol{\theta}, \boldsymbol{\psi}$ の尤度は

$$L(\boldsymbol{\theta}, \boldsymbol{\psi}) = \prod_{i=1}^{n} \int p(\boldsymbol{Z}_i;\boldsymbol{\theta})\omega(R_{i1}, \ldots, R_{im}, \boldsymbol{Z}_i;\boldsymbol{\psi}) d\boldsymbol{Z}_i^m$$

で与えられる．但し，$\boldsymbol{Z}_i^m$ は $\boldsymbol{Z}_i$ における欠測部分を表す(つまり，$\boldsymbol{Z}_i^m = O(\boldsymbol{Z}_i, 1-R_{i1}, \ldots, 1-R_{im})$)．ここでもし，データの欠測メカニズムが MAR であるとすると，$\omega(R_{i1}, \ldots, R_{im}, \boldsymbol{Z}_i;\boldsymbol{\psi})$ は $\boldsymbol{Z}_i^o$ にのみ依存し $\boldsymbol{Z}_i^m$ には依らないので

$$L(\boldsymbol{\theta}, \boldsymbol{\psi}) = \prod_{i=1}^{n} p_o(\boldsymbol{Z}_i^o;\boldsymbol{\theta})\omega(R_{i1}, \ldots, R_{im}, \boldsymbol{Z}_i;\boldsymbol{\psi}), \ p_o(\boldsymbol{Z}_i^o;\boldsymbol{\theta}) = \int p(\boldsymbol{Z}_i;\boldsymbol{\theta}) d\boldsymbol{Z}_i^m.$$

したがって，MAR に加えて，2 つのパラメータ $\boldsymbol{\theta}, \boldsymbol{\psi}$ が互いに無関係に変化し得るとすると，パラメータ $\boldsymbol{\theta}$ の最尤推定量は単に

$$L_o(\boldsymbol{\theta}) = \prod_{i=1}^{n} p_o(\boldsymbol{Z}_i^o;\boldsymbol{\theta}) = \prod_{i=1}^{n} \int p(\boldsymbol{Z}_i;\boldsymbol{\theta}) d\boldsymbol{Z}_i^m$$

を最大化することによって得られる．これは，各対象者に対して多変量データ $\boldsymbol{Z}_i$ のどの部分が欠測するかが予め分かっているときの(パラメータ $\boldsymbol{\theta}$ に対する)尤度であり，通常「観測データ尤度(observed-data likelihood)」と呼ばれる．このように，(a)欠測メカニズムが MAR でかつ(b)パラメータ $\boldsymbol{\theta}$ と $\boldsymbol{\psi}$ が互いに無関係(distinct)であれば，最尤法においては(同時)観測確率 $P(R_{i1} = r_1, \ldots, R_{im} = r_m | \boldsymbol{Z}_i)$ をモデル化する必要はなく，この意味で(a)と(b)は欠測が無視可能(ignorable)であるための条件と言われるが，これはパラメータの推測法に依存する概念であることに注意すべきである(実際，これから述べるセミパラメトリック推測の場合は，これらの条件が成り立っても欠測は無視可能でない)．MAR という概念は，もともと Rubin(1976)によって，最尤法やベイズ法などのパラメトリックな方法で欠測データを扱うことを想定して導入されたが，以降の節(第 4 節)で見るように，セミパラメトリックな方法を用いる際にも重要

な役割を果たす．

2.4 IPW 法と DR 法 ―単純な場合―

観測データから最尤法によって完全データ $\boldsymbol{Z}_i$ $(i=1,\ldots,n)$ の分布に関する推測を行う際には，上述の $p(\boldsymbol{z};\boldsymbol{\theta})$ のように $\boldsymbol{Z}_i$ の分布（確率密度関数）が完全にパラメトリックにモデル化されている必要があるが，実際に推測の対象となる興味あるパラメータは，$\boldsymbol{\theta}$ の一部（あるいは多対一関数）であることが多い．例えば，2.2 節で触れた血圧測定の例で見てみると，興味あるパラメータは 2 回目の血圧 Y_i の平均 $\mu=E(Y_i)$ であるが，MAR $(P(R_i=r|X_i,Y_i)=P(R_i=r|X_i)\ (r=0,1))$ の下で最尤法によって μ の推定を行おうとすると，Y_i だけでなく X_i と Y_i の同時分布をモデル化する必要がある．そこで，仮にその同時分布を平均ベクトル $\boldsymbol{m}$，分散共分散行列 Σ の 2 変量正規分布とすれば，$\boldsymbol{m}$ と Σ（の各成分をベクトルとしてまとめたもの）がパラメータ $\boldsymbol{\theta}$ に相当し，興味あるパラメータ μ はその一部となる．さらにもし，ここで MAR を成立させている X_i がもっと高次元のものであれば，同時分布をモデル化するためのパラメータ数も増大し，モデルの誤特定の可能性も増してくる．一方，セミパラメトリックな方法では，完全データの同時分布をフルにモデル化する必要はない．例えばこの例の場合，セミパラメトリック推定量の 1 つである μ の逆確率重み付け推定量（IPW 推定量）は，以下のように与えられる．

$$\tilde{\mu}_{IPW}=\frac{1}{n}\sum_{i=1}^{n}\frac{R_iY_i}{e(X_i)}.$$

ここで $e(X_i)$ は，X_i が与えられた下で Y_i が観測される条件付き確率を表し（*i.e.* $e(X_i)=P(R_i=1|X_i)$），しばしば傾向スコア（propensity score）と呼ばれる．この推定量は，Complete-Case（$R_i=1$）の各対象者に対して，Y_i の値をその観測確率（傾向スコア）の逆数倍することによって欠測値を疑似的に復元し，それらについて標本平均を取るという考え方に基づいているが，これによって，$\tilde{\mu}_{IPW}$ は MAR の下で μ の不偏推定量となる．しかしながら傾向スコア $e(X_i)$ は通常未知なので，その場合は $e(X_i)$ を観測データから，例えばロジスティック回帰モデル

$$e(X_i)=\frac{\exp(\alpha_0+\alpha_1X_i)}{1+\exp(\alpha_0+\alpha_1X_i)}(=e(X_i;\boldsymbol{\alpha}),\boldsymbol{\alpha}=(\alpha_0,\alpha_1))$$

等で推定する必要があり，実際は

$$\hat{\mu}_{IPW}=\frac{1}{n}\sum_{i=1}^{n}\frac{R_iY_i}{e(X_i;\hat{\boldsymbol{\alpha}})}$$

の形で用いることが多い．ここで，$\hat{\boldsymbol{\alpha}}$ は上のモデルに基づく $\boldsymbol{\alpha}$ の最尤推定量を表すが，このとき $\hat{\mu}_{IPW}$ は MAR の下で μ の一致推定量になる．IPW 推定量は直観的に分かりやすく，完全データに対するモデル化も不要であるが，通常は傾向スコアに対するモデル化が必要となるのでそれが誤特定されれば，（漸近的に）バイアスが生じる．また，そのモデルが正しく特定されていたとしても，Y_i が欠測している対象者の X_i の情報は（推定量の構成そのものには）用いていないので，μ に対する推定効率（漸近分散）は一般にあまり良くない．これらの点を改善するのが二重頑健推定量（DR 推定量）であるが，それは以下のように与えられる．

$$\hat{\mu}_{DR}=\frac{1}{n}\sum_{i=1}^{n}\left[\frac{R_iY_i}{e(X_i;\hat{\boldsymbol{\alpha}})}+\left\{1-\frac{R_i}{e(X_i;\hat{\boldsymbol{\alpha}})}\right\}m(X_i;\hat{\boldsymbol{\beta}})\right].$$

ここで $\hat{\boldsymbol{\beta}}$ は，ある回帰モデル $E(Y_i|X_i)=m(X_i;\boldsymbol{\beta})$ に対して，$\boldsymbol{\beta}$ を観測データから（最小二乗）推定したものである．推定量 $\hat{\mu}_{DR}$ は，形式的には回帰代入法による推定量（2.2）において，R_i を $R_i/\hat{e}(X_i)$ に置き換えたものになっているが，MAR の下で以下の性質を持つ．

(1) 傾向スコア $e(X_i)$ に対するモデル $e(X_i;\boldsymbol{\alpha})$ と条件付き期待値 $E(Y_i|X_i)$ に対するモデル(回帰モデル)$m(X_i;\boldsymbol{\beta})$ のうち，少なくともどちらか一方が正しく特定されていれば，DR 推定量 $\hat{\mu}_{DR}$ は μ に対して一致性を持つ．

(2) どちらのモデルも正しく特定されていれば，DR 推定量 $\hat{\mu}_{DR}$ の漸近分散は IPW 推定量 $\hat{\mu}_{IPW}$ よりも小さくなり，さらにセミパラメトリック漸近有効となる．

(1)の性質は $\hat{\mu}_{DR}$ の二重頑健性(double robustness)と呼ばれ，二重頑健推定量という名称はこの性質に由来する．また，(2)において「セミパラメトリック漸近有効」というのは，MAR(と傾向スコアに対するパラメトリックモデル)によって規定されるセミパラメトリックモデルの下で，μ に対する(一致性と漸近正規性を持つ)推定量のあるクラスを考えたとき，$\hat{\mu}_{DR}$ の漸近分散が，そのクラスに属する推定量の漸近分散の下限(semiparametric efficiency bound)に達するという意味であるが，これは次節で明らかになる．

疫学や医学研究で用いられる統計的手法においては，経時データ解析の一般化推定方程式(GEE)や生存時間解析の Cox 回帰モデルなど，データの欠測が全くなかったとしても，そもそもセミパラメトリックな統計モデルを用いることが多い．このような場合は，欠測に対処する際にそもそも最尤法などのパラメトリックな方法を用いることはできず，その意味でも欠測データに対するセミパラメトリックな解析法は重要な役割を果たす．

一般論を述べる前にごく簡単に歴史的なことについて触れておくと，IPW 推定量の基になる考え方自体は古くからあり，例えば標本調査の分野における Horvitz-Thompson 推定量(Horvitz and Thompson, 1952)なども同様の考え方に基づいているが，欠測データ解析に(セミパラメトリック推測の観点を伴って)導入されたのは比較的最近であり，Robins et al.(1994)が初期のものである．また二重頑健推定量は，その後，セミパラメトリック推測理論に基づく欠測データ解析法の研究の進展に伴って見い出されたものであるが，Scharfstein et al.(1999)によって最初に導入されて以来，現在でも多くの研究が行われている．

3. セミパラメトリック推測の一般論

次節において，欠測データに対するセミパラメトリックな解析法，特に二重頑健推定量についてその導出を中心に解説するが，ここではまずその準備として，必要な範囲でセミパラメトリック推測の一般事項について述べる．(なお，本節および次節の議論は主に Tsiatis, 2006 に基づく．ここでは概略のみ述べるが，より詳しくは Tsiatis, 2006 を参照されたい．)

多変量データ(2.1)に対して $\boldsymbol{Z}_i$ の分布(確率密度関数)が，あるセミパラメトリックモデル

$$\mathcal{S} = \{p(\boldsymbol{z};\boldsymbol{\theta},\eta) \mid \boldsymbol{\theta} \in \Theta \subset \boldsymbol{R}^r, \eta \in H\} \tag{3.1}$$

に属しているとする(但しここでは，$\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_n$ は全て観測されるとする)．ここで，$\boldsymbol{\theta}$ は有限次元(r 次元)の興味あるパラメータ，η は無限次元の局外パラメータで，応用上はしばしば(未知の)密度関数や回帰関数などとして現れる．例えば，Y を(連続的な)結果変数，$\boldsymbol{X}$ を(連続的な)説明変数ベクトル，$\boldsymbol{\theta}$ を有限次元の回帰パラメータとして，回帰モデル

$$E(Y|\boldsymbol{X}) = m(\boldsymbol{X};\boldsymbol{\theta}) \tag{3.2}$$

を考えたとき，$\boldsymbol{Z} = (Y,\boldsymbol{X})$ の確率密度関数は

$$p(\boldsymbol{z};\boldsymbol{\theta},\eta_1,\eta_2) = \eta_1(y - m(\boldsymbol{x};\boldsymbol{\theta}),\boldsymbol{x})\eta_2(\boldsymbol{x}) \quad (\boldsymbol{z} = (y,\boldsymbol{x}))$$

と書ける．但しここで，η_1,η_2 は

$$\int \eta_1(\epsilon, \boldsymbol{x})d\epsilon = 1,\ \int \epsilon\eta_1(\epsilon, \boldsymbol{x})d\epsilon = 0\ \ (\forall \boldsymbol{x}),\ \int \eta_2(\boldsymbol{x})d\boldsymbol{x} = 1$$

を満たす(未知の)非負値関数である(η_1 は $\boldsymbol{X} = \boldsymbol{x}$ の下での $\epsilon = Y - E(Y|\boldsymbol{X})$ の条件付き密度関数，η_2 は $\boldsymbol{X}$ の周辺密度関数に相当する)．これはつまり，(3.2)のモデルが，$\boldsymbol{\theta}$ を興味あるパラメータ，η_1, η_2 を(無限次元の)局外パラメータとするセミパラメトリックモデルを規定していることを意味する．

セミパラメトリック推測論の 1 つの主題は，与えられたセミパラメトリックモデル(3.1)の下で，興味ある(有限次元)パラメータ $\boldsymbol{\theta}$ に対する「最良の」推定量を見い出すことである．そのためにはまず，どのような推定量のクラスで考えるかということを定めなくてはならないが，ここでは次のような関係式を満たす推定量 $\hat{\boldsymbol{\theta}}$ を考える．

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \boldsymbol{\phi}(\boldsymbol{Z}_i, \boldsymbol{\theta}, \eta) + o_p(1)\ \ (\forall \boldsymbol{\theta} \in \Theta, \forall \eta \in H). \tag{3.3}$$

ここで，$\boldsymbol{\phi}(\boldsymbol{z}, \boldsymbol{\theta}, \eta)$ は

$$E\{\boldsymbol{\phi}(\boldsymbol{Z}_i, \boldsymbol{\theta}, \eta)\} = \boldsymbol{0},\ E\{\|\boldsymbol{\phi}(\boldsymbol{Z}_i, \boldsymbol{\theta}, \eta)\|^2\} < \infty\ \ (\forall \boldsymbol{\theta} \in \Theta, \forall \eta \in H)$$

を満たすある r 次元のベクトル値関数であり，推定量 $\hat{\boldsymbol{\theta}}$ の影響関数(influence function)と呼ばれる(但し，$o_p(1)$ は $n \to \infty$ のときに 0 に確率収束する項を表す)．式(3.3)を満たす推定量 $\hat{\boldsymbol{\theta}}$ は通常，漸近線形推定量(asymptotic linear estimator)と呼ばれるが，影響関数 $\boldsymbol{\phi}(\boldsymbol{z}, \boldsymbol{\theta}, \eta)$ は $\hat{\boldsymbol{\theta}}$ に対して一意に定まる(ことが示せる)ので，漸近線形推定量は影響関数によって特徴付けられる．また，大数の法則と中心極限定理により，漸近線形推定量 $\hat{\boldsymbol{\theta}}$ に対して一致性と漸近正規性が成り立つ．すなわち，$n \to \infty$ のとき

一致性：$\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}$(確率収束)，漸近正規性：$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to N(\boldsymbol{0}, \mathrm{Avar}(\hat{\boldsymbol{\theta}}))$(分布収束)．

ここで，$\mathrm{Avar}(\hat{\boldsymbol{\theta}})$ は $\hat{\boldsymbol{\theta}}$ の漸近分散共分散行列と呼ばれ，

$$\mathrm{Avar}(\hat{\boldsymbol{\theta}}) = E\{\boldsymbol{\phi}(\boldsymbol{Z}_i, \boldsymbol{\theta}, \eta)\boldsymbol{\phi}(\boldsymbol{Z}_i, \boldsymbol{\theta}, \eta)^T\}$$

で与えられる．右辺はちょうど影響関数の分散共分散行列に相当するが，漸近線形推定量についてはこの量が(対称行列の順序関係の意味で)小さいほど，漸近的に精度の良い推定量である．

漸近線形推定量の典型的な例は，推定関数(あるいは推定方程式)から得られる推定量(M-推定量)である． 推定関数とは，例えば，以下のような条件

$$E\{\boldsymbol{u}(\boldsymbol{Z}_i, \boldsymbol{\theta})\} = \boldsymbol{0},\ E\{\|\boldsymbol{u}(\boldsymbol{Z}_i, \boldsymbol{\theta})\|^2\} < \infty,\ \det E\left\{\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{\theta}}(\boldsymbol{Z}_i, \boldsymbol{\theta})\right\} \neq 0\ \ (\forall \boldsymbol{\theta} \in \Theta, \forall \eta \in H)$$

を満たす r 次元ベクトル値関数 $\boldsymbol{u}(\boldsymbol{z}, \boldsymbol{\theta})$ のことであるが，ランダムサンプル $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ に対し，推定方程式

$$\sum_{i=1}^{n} \boldsymbol{u}(\boldsymbol{Z}_i, \boldsymbol{\theta}) = \boldsymbol{0}$$

の解として得られる $\boldsymbol{\theta}$ の推定量を $\hat{\boldsymbol{\theta}}$ とすると，ある適当な正則条件の下で $\hat{\boldsymbol{\theta}}$ は一致性を持ち，また以下の関係式が成り立つ．

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[E\left\{\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{\theta}}(\boldsymbol{Z}_i, \boldsymbol{\theta})\right\}\right]^{-1} \boldsymbol{u}(\boldsymbol{Z}_i, \boldsymbol{\theta}) + o_p(1)\ \ (\forall \boldsymbol{\theta} \in \Theta, \forall \eta \in H).$$

つまり，M-推定量 $\hat{\boldsymbol{\theta}}$ は $\left[E\left\{\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{\theta}}(\boldsymbol{Z}_i,\boldsymbol{\theta})\right\}\right]^{-1}\boldsymbol{u}(\boldsymbol{z},\boldsymbol{\theta})$ を影響関数とする漸近線形推定量であり，それゆえ漸近正規性が成り立ち，漸近分散共分散行列は

$$\begin{aligned}(3.4)\qquad &\mathrm{Avar}(\hat{\boldsymbol{\theta}})\\ &=\left[E\left\{\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{\theta}}(\boldsymbol{Z}_i,\boldsymbol{\theta})\right\}\right]^{-1}E\{\boldsymbol{u}(\boldsymbol{Z}_i,\boldsymbol{\theta})\boldsymbol{u}(\boldsymbol{Z}_i,\boldsymbol{\theta})^T\}\left[E\left\{\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{\theta}}(\boldsymbol{Z}_i,\boldsymbol{\theta})\right\}\right]^{-T}\end{aligned}$$

で与えられる．パラメータ $\boldsymbol{\theta}$ の(漸近的な)信頼領域などを構成するためには，(3.4)をデータから推定する必要があるが，通常，(その形から)サンドイッチ推定量と呼ばれる以下のような推定量が用いられる．

$$\begin{aligned}(3.5)\qquad &\hat{\mathrm{Avar}}(\hat{\boldsymbol{\theta}})\\ &=\left[\hat{E}\left\{\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{\theta}}(\boldsymbol{Z}_i,\hat{\boldsymbol{\theta}})\right\}\right]^{-1}\hat{E}\{\boldsymbol{u}(\boldsymbol{Z}_i,\hat{\boldsymbol{\theta}})\boldsymbol{u}(\boldsymbol{Z}_i,\hat{\boldsymbol{\theta}})^T\}\left[\hat{E}\left\{\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{\theta}}(\boldsymbol{Z}_i,\hat{\boldsymbol{\theta}})\right\}\right]^{-T}.\end{aligned}$$

但し，$\hat{E}$ は経験分布による期待値を表す．すなわち $\boldsymbol{Z}_i$ の関数 $f(\boldsymbol{Z}_i)$ に対して $\hat{E}\{f(\boldsymbol{Z}_i)\}=\frac{1}{n}\sum_{i=1}^n f(\boldsymbol{Z}_i)$ である．推定関数(推定方程式)は，セミパラメトリックモデルの下でのパラメータ推定法としてよく用いられるものであり，次節で述べる欠測データに対するセミパラメトリックな解析法においても重要な役割を果たすが，推定関数について詳しくは例えば，van der Vaart (2000)の第 5 章，Boos and Stefanski(2013)の第 7 章等を参照されたい．

さて，漸近線形推定量は一致性と漸近正規性を持ち，また M-推定量がそのクラスに含まれるということから自然な要請であると考えられるが，これだけでは例えば，超有効性と呼ばれる性質を持つ極めて不自然な挙動を示す推定量まで含まれてしまうことになる．そこで，そのようなものを除外するためにある正則条件を付加するのが通常であり，その条件を満たす漸近線形推定量のことを RAL 推定量と呼ぶ．(RAL とは Regular Asymptotic Linear の略．ここでは，正則条件の内容については述べないが，詳しくは Tsiatis, 2006 を参照のこと．)

セミパラメトリックモデル(3.1)が与えられたとき，興味あるパラメータ $\boldsymbol{\theta}$ の RAL 推定量(あるいはそれを特徴付ける影響関数)はどのようなものになるであろうか？ 次にそれを見い出すための手がかりになる事実(性質)について述べるが，まず，そのために必要となるいくつかの基礎概念を定義しておく．

定義 1.(パラメトリックサブモデル)セミパラメトリックモデル $\mathcal{S}$ の各点 $p(\boldsymbol{z};\boldsymbol{\theta},\eta)$ に対し

$$(3.6)\qquad p(\boldsymbol{z};\boldsymbol{\theta},\eta)\in\mathcal{S}_{sub}\subset\mathcal{S}$$

を満たすパラメトリックモデル

$$\mathcal{S}_{sub}=\{p(\boldsymbol{z};\boldsymbol{\theta},\boldsymbol{\gamma})\mid\boldsymbol{\theta}\in\Theta\subset\boldsymbol{R}^r,\boldsymbol{\gamma}\in\Gamma\subset\boldsymbol{R}^s\}\ \ (s\text{ はある自然数})$$

を，$p(\boldsymbol{z};\boldsymbol{\theta},\eta)$ における $\mathcal{S}$ のパラメトリックサブモデルという．

定義 2.(局外接空間)セミパラメトリックモデル $\mathcal{S}$ の各点 $p(\boldsymbol{z};\boldsymbol{\theta},\eta)$ に対し，上記のパラメトリックサブモデル $\mathcal{S}_{sub}$ の局外接空間を

$$T^N_{\boldsymbol{\theta},\boldsymbol{\gamma}}(\mathcal{S}_{sub})=\{\boldsymbol{\lambda}^T\boldsymbol{s}_{\boldsymbol{\gamma}}(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\gamma})\mid\boldsymbol{\lambda}\in\boldsymbol{R}^s\}$$

として定義する．ここで，$\boldsymbol{\gamma}$ は $p(\boldsymbol{z};\boldsymbol{\theta},\eta)$ に対応するもの(つまり $p(\boldsymbol{z};\boldsymbol{\theta},\eta)=p(\boldsymbol{z};\boldsymbol{\theta},\boldsymbol{\gamma})$)であり，$\boldsymbol{s}_{\boldsymbol{\gamma}}(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\gamma})$ は $\boldsymbol{\gamma}$ に関するスコア関数，すなわち

$$\boldsymbol{s}_{\boldsymbol{\gamma}}(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\gamma})=\frac{\partial}{\partial\boldsymbol{\gamma}}\log p(\boldsymbol{z};\boldsymbol{\theta},\boldsymbol{\gamma})$$

である．このとき

$$T^N_{\boldsymbol{\theta},\eta}(\mathcal{S}) = \overline{\cup_{\mathcal{S}_{sub}} T^N_{\boldsymbol{\theta},\gamma}(\mathcal{S}_{sub})} \tag{3.7}$$

を，$p(\boldsymbol{z};\boldsymbol{\theta},\eta)$ における $\mathcal{S}$ の局外接空間(nuisance tangent space)という．但し，ここでパラメトリックサブモデル $\mathcal{S}_{sub}$ の局外接空間 $T^N_{\boldsymbol{\theta},\gamma}(\mathcal{S}_{sub})$ は

$$< h_1, h_2 >_{\boldsymbol{\theta},\eta} = E\{h_1(\boldsymbol{Z})h_2(\boldsymbol{Z})\} \quad (\forall h_1, h_2 \in \mathcal{H}_{\boldsymbol{\theta},\eta})$$

を内積とするヒルベルト空間

$$\mathcal{H}_{\boldsymbol{\theta},\eta} = \{h \mid E\{h(\boldsymbol{Z})\} = 0,\ E[\{h(\boldsymbol{Z})\}^2] < \infty\} \tag{3.8}$$

($\boldsymbol{Z}$ は $p(\boldsymbol{z};\boldsymbol{\theta},\eta)$ を密度関数とする確率ベクトル，h は $\boldsymbol{Z}$ の値域で定義された実数値可測関数)に含まれるとし，$\cup_{\mathcal{S}_{sub}}$ は全てのパラメトリックサブモデル $\mathcal{S}_{sub}$ について和(集合)を取ることを，上付きの横線は $\mathcal{H}_{\boldsymbol{\theta},\eta}$ において閉包を取る(つまりその集合の要素から成る無限点列の極限であるような $\mathcal{H}_{\boldsymbol{\theta},\eta}$ の点を全て含める)ことを意味するものとする． また，(3.7)の右辺は，一般には $\mathcal{H}_{\theta,\eta}$ の線形部分空間になるとは限らないが，ここではそのようになる場合についてのみ考える．(実際，応用上現れる多くの例で，そのようになっている．)

さて，セミパラメトリックモデルの局外接空間と RAL 推定量の影響関数の間には，次のような関係がある．

命題 1. (影響関数の満たすべき条件)セミパラメトリックモデル(3.1)の下で，パラメータ $\boldsymbol{\theta}$ の RAL 推定量 $\hat{\boldsymbol{\theta}}$ の影響関数 $\boldsymbol{\phi}(\boldsymbol{z},\boldsymbol{\theta},\eta)$ は，必ず以下の条件を満たす．

$$\boldsymbol{\phi}(\boldsymbol{z},\boldsymbol{\theta},\eta) \perp T^N_{\boldsymbol{\theta},\eta}(\mathcal{S}), \tag{3.9}$$

$$E\{\boldsymbol{\phi}(\boldsymbol{Z},\boldsymbol{\theta},\eta)\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{Z},\boldsymbol{\theta},\eta)^T\} = I_r \ (r \text{ 次の単位行列}). \tag{3.10}$$

ここで，(3.9)は影響関数 $\boldsymbol{\phi}(\boldsymbol{z},\boldsymbol{\theta},\eta)$ の各成分が(ヒルベルト空間 $\mathcal{H}_{\boldsymbol{\theta},\eta}$ の内積に関して)局外接空間 $T^N_{\boldsymbol{\theta},\eta}(\mathcal{S})$ に直交していることを意味し，(3.10)における $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{z},\boldsymbol{\theta},\eta)$ は $\boldsymbol{\theta}$ に関するスコア関数を表す．すなわち

$$\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{z},\boldsymbol{\theta},\eta) = \frac{\partial}{\partial\boldsymbol{\theta}} \log p(z;\boldsymbol{\theta},\eta).$$

(3.9)の条件は別の言い方をすれば，影響関数の各成分は常に局外接空間の($\mathcal{H}_{\boldsymbol{\theta},\eta}$ における)直交補空間に属するということであるが，このことから逆に，RAL 推定量の影響関数がどのようなものであるかを知るにはまず，その直交補空間を求めればよいということが示唆される．また，推定関数から得られる M-推定量については影響関数と推定関数が(多次元の)比例関係にあるので，局外接空間の直交補空間を求めることは推定関数を求めるための手がかりにもなる．(実際，推定関数の各成分もその直交補空間に属している．)

さて，RAL 推定量の漸近分散共分散行列は影響関数の分散共分散行列で与えられたので，条件(3.9)と(3.10)を満たす $\boldsymbol{\phi}(\boldsymbol{z},\boldsymbol{\theta},\eta)$ のうち，その分散共分散行列が(対称行列の順序関係の意味で)最小となるものが見い出せれば，それが漸近分散共分散行列の「下限」を与えることになる．これについて以下のことが成り立つ．

命題 2. (セミパラメトリック Cramer-Rao 不等式)セミパラメトリックモデル(3.1)の下で，パラメータ $\boldsymbol{\theta}$ の任意の RAL 推定量 $\hat{\boldsymbol{\theta}}$ の漸近分散共分散行列 $\mathrm{Avar}(\hat{\boldsymbol{\theta}})$ に対し，以下の不等式が成り立つ．

$$\mathrm{Avar}(\hat{\boldsymbol{\theta}}) \geq [E\{\boldsymbol{s}^E_{\boldsymbol{\theta}}(\boldsymbol{Z},\boldsymbol{\theta},\eta)\boldsymbol{s}^E_{\boldsymbol{\theta}}(\boldsymbol{Z},\boldsymbol{\theta},\eta)^T\}]^{-1}. \tag{3.11}$$

ここで，$\boldsymbol{s}_{\boldsymbol{\theta}}^{E}(\boldsymbol{z},\boldsymbol{\theta},\eta)$ は $\boldsymbol{\theta}$ に関するスコア関数 $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{z},\boldsymbol{\theta},\eta)$（の各成分）を局外接空間 $T_{\boldsymbol{\theta},\eta}^{N}(\mathcal{S})$ の直交補空間に直交射影したものとして定義され，$\boldsymbol{\theta}$ に関する有効スコア関数（efficient score function）と呼ばれる．また，等号は $\hat{\boldsymbol{\theta}}$ の影響関数が

$$\boldsymbol{\phi}^{E}(\boldsymbol{z},\boldsymbol{\theta},\eta)=[E\{\boldsymbol{s}_{\boldsymbol{\theta}}^{E}(\boldsymbol{Z},\boldsymbol{\theta},\eta)\boldsymbol{s}_{\boldsymbol{\theta}}^{E}(\boldsymbol{Z},\boldsymbol{\theta},\eta)^{T}\}]^{-1}\boldsymbol{s}_{\boldsymbol{\theta}}^{E}(\boldsymbol{z},\boldsymbol{\theta},\eta)$$

で与えられるときに成立し，この $\boldsymbol{\phi}^{E}(\boldsymbol{z},\boldsymbol{\theta},\eta)$ を $\boldsymbol{\theta}$ に関する有効影響関数（efficient influence function）と呼ぶ．（有効影響関数は，条件 (3.9) と (3.10) を満たす $\boldsymbol{\phi}(\boldsymbol{z},\boldsymbol{\theta},\eta)$ のうち，最小の分散共分散行列を持つものである．但しこれは，影響関数として対応する RAL 推定量が存在するか否かに依らないことに注意.）不等式 (3.11) の右辺はしばしば，セミパラメトリック効率限界（semiparametric efficiency bound）と呼ばれるが，これはパラメトリックモデルにおける（漸近分散に対する）Cramer-Rao の下限のセミパラメトリック版に相当する．

有効スコア関数は一般に未知の（無限次元）局外パラメータにも依存するが，例えば生存時間解析における比例ハザードモデル（Cox 回帰モデル）のような特別なセミパラメトリックモデルに対しては，興味ある（有限次元の）パラメータのみに依存し，そのような場合には有効スコア関数を基にした推定関数によって，常に漸近有効な RAL 推定量（大域的（漸近）有効推定量）が得られる． しかしながら，例えば経時データ解析における一般化推定方程式（GEE）のセミパラメトリックモデル（制限モーメントモデル）では，有効スコア関数は未知の局外パラメータにも依存する部分（結果変数の分散共分散行列）を含み，有効スコア関数を基に推定方程式を構成する際にはその部分を推定する必要がある．ただこの場合は，それを推定するためのモデルが誤特定されていたとしても，（漸近有効性は失われるが）少なくとも一致推定量（RAL 推定量）は得られるという性質を有しており，このような推定量は，そのモデルが正しく特定されているときのみ漸近有効性を持つという意味で，局所（漸近）有効推定量と呼ばれる．これらについて詳しくは，Tsiatis (2006) の第 4–5 章を参照されたい．

4. セミパラメトリックな解析法 ―二重頑健推定量の導出―

2.4 節において，ごく簡単な場合に，欠測データに対するセミパラメトリック推定量として逆確率重み付け推定量（IPW 推定量）と二重頑健推定量（DR 推定量）を紹介したが，本節ではセミパラメトリック推測の一般論に基づき，さらにより一般の設定の下でこれらの推定量について述べる．

4.1 観測データセミパラメトリックモデル

2.3 節では多変量データ (2.1) に対し，$\boldsymbol{Z}_i$ の確率密度関数がパラメトリックモデルに属しているとしたが，ここでは以下のようなセミパラメトリックモデル

$$\mathcal{S}^{F}=\{p(\boldsymbol{z};\boldsymbol{\beta},\eta)\mid\boldsymbol{\beta}\in B\subset\boldsymbol{R}^{p},\eta\in H\}$$

に属しているとする．ここで，$\boldsymbol{\beta}$ は有限次元（p 次元）の興味あるパラメータ，η は無限次元の局外パラメータである．R_{ij} を Z_{ij} の観測指示変数（つまり Z_{ij} が観測されれば 1, 欠測すれば 0 の値を取る確率変数）とし，データの欠測メカニズムが MAR であるとすると，2.3 節と同様の議論により，対象者 i に対する観測データ

$$\boldsymbol{Z}_i^{o}=O(\boldsymbol{Z}_i,\boldsymbol{R}_i),\ \boldsymbol{R}_i=(R_{i1},\ldots,R_{im})^{T}$$

の（同時）確率密度関数は

$$q(\boldsymbol{z}^o, \boldsymbol{r}) = p_o(\boldsymbol{z}^o; \boldsymbol{\beta}, \eta)\omega(\boldsymbol{z}^o, \boldsymbol{r})$$

と書ける．但しここで

$$p_o(\boldsymbol{z}^o; \boldsymbol{\beta}, \eta) = \int p(\boldsymbol{z}; \boldsymbol{\beta}, \eta) d\boldsymbol{z}^m \quad (\boldsymbol{z} = (\boldsymbol{z}^o, \boldsymbol{z}^m)), \tag{4.1}$$

$$\omega(\boldsymbol{z}^o, \boldsymbol{r}) = P(\boldsymbol{R}_i = \boldsymbol{r} | O(\boldsymbol{Z}_i, \boldsymbol{r}) = \boldsymbol{z}^o) \tag{4.2}$$

である． すなわち，観測データの分布もまた，これらによって規定されるセミパラメトリックモデルに属する． 興味あるパラメータ $\boldsymbol{\beta}$ の二重頑健推定量は，このセミパラメトリックモデルに対して前節の一般論を適用することにより導かれるが，以下ではそのことについて順を追って説明する．

4.2 観測データ局外接空間とその直交補空間 1 ―データ観測確率が既知の場合―

まず，データの(条件付き)観測確率(4.2)が既知の場合について考える．前節の命題 1 によれば，興味あるパラメータに対する RAL 推定量の(影響関数の)クラスを知るには，まずセミパラメトリックモデルの局外接空間とその直交補空間を求めることが重要であったが，観測データセミパラメトリックモデル

$$\mathcal{S}^O_\star = \{q(\boldsymbol{z}^o, \boldsymbol{r}; \boldsymbol{\beta}, \eta) \mid \boldsymbol{\beta} \in B \subset \boldsymbol{R}^p, \eta \in H\}$$

の局外接空間およびその直交補空間は，以下のように与えられる．

命題 3. (観測データ局外接空間とその直交補空間― データ観測確率が既知の場合 ―)完全データモデル $\mathcal{S}^F$ の(点 $p(\boldsymbol{z}; \boldsymbol{\beta}, \eta)$ における)局外接空間を T^{NF}，完全データヒルベルト空間

$$\mathcal{H}^F = \{h \mid E\{h(\boldsymbol{Z})\} = 0,\ E[\{h(\boldsymbol{Z})\}^2] < \infty\}$$

における T^{NF} の直交補空間を Λ^F とし，また，観測データモデル $\mathcal{S}^O_\star$ の(点 $q(\boldsymbol{z}^o, \boldsymbol{r}; \boldsymbol{\beta}, \eta)$ における)局外接空間を T^{NO}_η，観測データヒルベルト空間

$$\mathcal{H}^O = \{g \mid E\{g(\boldsymbol{Z}^o, \boldsymbol{R})\} = 0,\ E[\{g(\boldsymbol{Z}^o, \boldsymbol{R})\}^2] < \infty\}$$

における T^{NO}_η の直交補空間を Λ^O_η とする．このとき，$\mathcal{H}^F$ から $\mathcal{H}^O$ への線形作用素(線形写像) $\mathcal{L}$ と $\mathcal{H}^O$ から $\mathcal{H}^F$ への線形作用素 $\mathcal{L}^*$ をそれぞれ

$$\mathcal{L}(h)(\boldsymbol{Z}^o, \boldsymbol{R}) = E\{h(\boldsymbol{Z}) | \boldsymbol{Z}^o, \boldsymbol{R}\} \quad (\forall h \in \mathcal{H}^F)$$
$$\mathcal{L}^*(g)(\boldsymbol{Z}) = E\{g(\boldsymbol{Z}^o, \boldsymbol{R}) | \boldsymbol{Z}\} \quad (\forall g \in \mathcal{H}^O)$$

によって定義すると，

$$T^{NO}_\eta = \mathcal{L}(T^{NF}) = \{\mathcal{L}(h) \mid h \in T^{NF}\} \tag{4.3}$$

$$\Lambda^O_\eta = (\mathcal{L}^*)^{-1}(\Lambda^F) = \{g \in \mathcal{H}^O \mid \mathcal{L}^*(g) \in \Lambda^F\} \tag{4.4}$$

が成り立つ．

式(4.4)は，Λ^O_η が線形方程式 $\mathcal{L}^*(g) = h$(h は Λ^F の任意の要素)の解集合を($h \in \Lambda^F$ について)全て併合したものであることを示しているが，$I(\boldsymbol{R} = \boldsymbol{r})$ を $\boldsymbol{R} = \boldsymbol{r}$ であれば 1，そうでなければ 0 の値を取る 2 値の確率変数とし，$\mathbf{1} = (1, \ldots, 1)^T \in \boldsymbol{R}^m$, また $P(\boldsymbol{R} = \mathbf{1} | \boldsymbol{Z}) > 0$ と仮定すると

$$g(\boldsymbol{Z}^o, \boldsymbol{R}) = \frac{I(\boldsymbol{R} = \mathbf{1}) h(\boldsymbol{Z})}{P(\boldsymbol{R} = \mathbf{1} | \boldsymbol{Z})}$$

がその線形方程式の 1 つの解(特殊解)であることから

$$\Lambda_\eta^O = \frac{I(\boldsymbol{R}=\boldsymbol{1})\Lambda^F}{P(\boldsymbol{R}=\boldsymbol{1}|\boldsymbol{Z})} + \mathrm{Ker}\,(\mathcal{L}^*) \tag{4.5}$$

が成り立つ．ここで，$\mathrm{Ker}\,(\mathcal{L}^*)$ は線形作用素 $\mathcal{L}^*$ の核，つまり $\mathrm{Ker}\,(\mathcal{L}^*) = \{h \in \mathcal{H}^O | \mathcal{L}^*(h) = 0\}$ であるが，$\mathcal{L}^*$ の定義と観測指示変数ベクトル $\boldsymbol{R}$ が離散的である(有限個の値しか取らない)ことから，$\mathrm{Ker}\,(\mathcal{L}^*)$ の要素は一般に次のように書ける．

$$k(\boldsymbol{Z}^o, \boldsymbol{R}) = \sum_{\boldsymbol{r}\neq\boldsymbol{1}} \left\{ I(\boldsymbol{R}=\boldsymbol{r}) - \frac{I(\boldsymbol{R}=\boldsymbol{1})P(\boldsymbol{R}=\boldsymbol{r}|\boldsymbol{Z})}{P(\boldsymbol{R}=\boldsymbol{1}|\boldsymbol{Z})} \right\} h_{\boldsymbol{r}}(\boldsymbol{Z}_{\boldsymbol{r}}^o). \tag{4.6}$$

但し，$\sum_{\boldsymbol{r}\neq\boldsymbol{1}}$ は $\boldsymbol{1}$ 以外の全ての $\boldsymbol{r}$ の値に関して和を取ることを意味し，$h_{\boldsymbol{r}}(\boldsymbol{Z}_{\boldsymbol{r}}^o)$ は $\boldsymbol{Z}_{\boldsymbol{r}}^o = O(\boldsymbol{Z}, \boldsymbol{r})$ のある実数値関数を表す．式(4.6)の表現は一般にはこれ以上簡単にならないが，単純な場合として，完全データ $\boldsymbol{Z}$ が $\boldsymbol{Z} = (\boldsymbol{V}^T, \boldsymbol{W}^T)^T$ のように予め 2 つの部分に分かれ，$\boldsymbol{V}$ は常に観測されるが $\boldsymbol{W}$ は観測されるか(全ての成分が)欠測するかのどちらかであるという場合(2 水準の欠測)について考えてみると，(4.6)は以下のように表される．

$$k(\boldsymbol{Z}^o, \boldsymbol{R}) = \frac{R - e(\boldsymbol{V})}{e(\boldsymbol{V})} h(\boldsymbol{V}). \tag{4.7}$$

但し，R は $\boldsymbol{Z}$(の全ての成分)が観測されれば 1, そうでなければ(つまり $\boldsymbol{W}$ が欠測すれば)0 の値を取る 2 値の観測指示変数であり，$e(\boldsymbol{V}) = P(R=1|\boldsymbol{V})$, また $h(\boldsymbol{V})$ は $\boldsymbol{V}$ のある実数値関数である．

4.3 観測データ局外接空間とその直交補空間 2 ―データ観測確率が未知の場合―

さて，次にデータの観測確率(4.2)が未知の場合について考える．このときは，(4.2)を観測データから推定する必要があるが，そのためにそれをパラメトリックにモデル化し，一般に

$$P(\boldsymbol{R}_i = \boldsymbol{r} | O(\boldsymbol{Z}_i, \boldsymbol{r}) = \boldsymbol{z}^o) = \omega(\boldsymbol{z}^o, \boldsymbol{r}; \boldsymbol{\alpha}) \quad (\boldsymbol{\alpha} \in A \subset \boldsymbol{R}^q)$$

と表すことにすると，観測データモデル

$$\mathcal{S}^O = \{q(\boldsymbol{z}^o, \boldsymbol{r}; \boldsymbol{\beta}, \boldsymbol{\alpha}, \eta) \mid \boldsymbol{\beta} \in B \subset \boldsymbol{R}^p, \boldsymbol{\alpha} \in A \subset \boldsymbol{R}^q, \eta \in H\} \tag{4.8}$$

$$\text{但し，} q(\boldsymbol{z}^o, \boldsymbol{r}; \boldsymbol{\beta}, \boldsymbol{\alpha}, \eta) = p_o(\boldsymbol{z}^o; \boldsymbol{\beta}, \eta)\omega(\boldsymbol{z}^o, \boldsymbol{r}; \boldsymbol{\alpha}) \tag{4.9}$$

の局外接空間とその直交補空間は以下のように与えられる．

命題 4. (観測データ局外接空間とその直交補空間― データ観測確率が未知の場合 ―) 観測データモデル $\mathcal{S}^O$ の(点 $q(\boldsymbol{z}^o, \boldsymbol{r}; \boldsymbol{\beta}, \boldsymbol{\alpha}, \eta)$ における)局外接空間を T^{NO}, 観測データヒルベルト空間 $\mathcal{H}^O$ における T^{NO} の直交補空間を Λ^O とすると，

$$T^{NO} = T_\eta^{NO} \oplus T_{\boldsymbol{\alpha}}^{NO}, \quad T_\eta^{NO} \perp T_{\boldsymbol{\alpha}}^{NO} \tag{4.10}$$

$$\Lambda^O = \{h - \Pi(h|T_{\boldsymbol{\alpha}}^{NO}) \mid h \in \Lambda_\eta^O\} \tag{4.11}$$

が成り立つ．ここで，$T_{\boldsymbol{\alpha}}^{NO}$ は $\mathcal{S}^O$ の(有限次元の)局外パラメータ $\boldsymbol{\alpha}$ に関する接空間，すなわち

$$T_{\boldsymbol{\alpha}}^{NO} = \{\boldsymbol{\lambda}^T \boldsymbol{s}_{\boldsymbol{\alpha}}(\boldsymbol{z}^o, \boldsymbol{r}, \boldsymbol{\alpha}) \mid \boldsymbol{\lambda} \in \boldsymbol{R}^q\}$$

$$\text{但し，} \boldsymbol{s}_{\boldsymbol{\alpha}}(\boldsymbol{z}^o, \boldsymbol{r}, \boldsymbol{\alpha}) = \frac{\partial}{\partial \boldsymbol{\alpha}} \log \omega(\boldsymbol{z}^o, \boldsymbol{r}; \boldsymbol{\alpha})$$

であり，また $\Pi\left(h|T_{\boldsymbol{\alpha}}^{NO}\right)$ は，ヒルベルト空間 $\mathcal{H}^O$ における，h の $T_{\boldsymbol{\alpha}}^{NO}$ への直交射影を表す．

ちなみに，局外接空間 T^{NO} が(4.10)のように直交分解されることは，観測データの密度関数が(4.9)のように分解されることからの帰結であり，(4.11)もその直交分解から自然に導かれる．

4.4 観測データ影響関数とその最適性

一般に興味あるパラメータの RAL 推定量の影響関数は，局外接空間の直交補空間に属することに加えて命題 1 の(3.10)も満たす必要があるが，そのことも考慮すると(4.5)，(4.11)から，観測データに基づくパラメータ $\boldsymbol{\beta}$ の RAL 推定量の影響関数(観測データ影響関数)は，一般に次のように表されることが分かる．

$$\boldsymbol{\phi}^O(\boldsymbol{Z}^o,\boldsymbol{R},\boldsymbol{\beta},\boldsymbol{\alpha},\eta)=\boldsymbol{\phi}^O_\star(\boldsymbol{Z}^o,\boldsymbol{R},\boldsymbol{\beta},\boldsymbol{\alpha},\eta)-\Pi\{\boldsymbol{\phi}^O_\star(\boldsymbol{Z}^o,\boldsymbol{R},\boldsymbol{\beta},\boldsymbol{\alpha},\eta)|T^{NO}_{\boldsymbol{\alpha}}\} \tag{4.12}$$

ここで，

$$\boldsymbol{\phi}^O_\star(\boldsymbol{Z}^o,\boldsymbol{R},\boldsymbol{\beta},\boldsymbol{\alpha},\eta)=\frac{I(\boldsymbol{R}=\mathbf{1})\boldsymbol{\phi}^F(\boldsymbol{Z},\boldsymbol{\beta},\eta)}{\omega(\boldsymbol{Z},\mathbf{1};\boldsymbol{\alpha})}+\boldsymbol{k}(\boldsymbol{Z}^o,\boldsymbol{R},\boldsymbol{\alpha})$$

であり，$\boldsymbol{\phi}^F(\boldsymbol{Z},\boldsymbol{\beta},\eta)$ は完全データに基づく $\boldsymbol{\beta}$ の RAL 推定量の影響関数(但し，命題 1 の(3.9)と(3.10)を満たすという意味において)，$\boldsymbol{k}(\boldsymbol{Z}^o,\boldsymbol{R},\boldsymbol{\alpha})$ はその各成分が $\mathrm{Ker}\ (\mathcal{L}^*)$ に属する $\boldsymbol{R}^p$ 値関数である．また(4.12)の右辺第 2 項は，成分ごとに直交射影を取るものとする．

さて，RAL 推定量の漸近分散共分散行列はその影響関数の分散共分散行列によって与えられたので，(4.12)において任意性のある完全データ影響関数 $\boldsymbol{\phi}^F(\boldsymbol{Z},\boldsymbol{\beta},\eta)$ と関数 $\boldsymbol{k}(\boldsymbol{Z}^o,\boldsymbol{R},\boldsymbol{\alpha})$ を変化させたときに，観測データ影響関数 $\boldsymbol{\phi}^O(\boldsymbol{Z}^o,\boldsymbol{R},\boldsymbol{\beta},\boldsymbol{\alpha},\eta)$ の分散共分散行列が最も小さくなる場合が最適な場合となる(つまりそのときに，(4.12)は有効影響関数となる)．しかしながら，その意味で最適な場合は一般に複雑となるので，まず，完全データ影響関数を 1 つ固定したときの最適性について考える．これについて，以下のことが成り立つ．

命題 5. (最適な観測データ影響関数 ―完全データ影響関数を固定した場合―)完全データ影響関数 $\boldsymbol{\phi}^F(\boldsymbol{Z},\boldsymbol{\beta},\eta)$ を 1 つ固定したとき，(4.12)によって与えられる観測データ影響関数の分散共分散行列が最小となるのは

$$\boldsymbol{k}(\boldsymbol{Z}^o,\boldsymbol{R},\boldsymbol{\alpha})=-\Pi\left\{\frac{I(\boldsymbol{R}=\mathbf{1})\boldsymbol{\phi}^F(\boldsymbol{Z},\boldsymbol{\beta},\eta)}{\omega(\boldsymbol{Z},\mathbf{1};\boldsymbol{\alpha})}\middle|\,\mathrm{Ker}\ (\mathcal{L}^*)\right\}$$

の場合である．このとき，その観測データ影響関数は

$$\boldsymbol{\phi}^O(\boldsymbol{Z}^o,\boldsymbol{R},\boldsymbol{\beta},\boldsymbol{\alpha},\eta)=\frac{I(\boldsymbol{R}=\mathbf{1})\boldsymbol{\phi}^F(\boldsymbol{Z},\boldsymbol{\beta},\eta)}{\omega(\boldsymbol{Z},\mathbf{1};\boldsymbol{\alpha})}-\Pi\left\{\frac{I(\boldsymbol{R}=\mathbf{1})\boldsymbol{\phi}^F(\boldsymbol{Z},\boldsymbol{\beta},\eta)}{\omega(\boldsymbol{Z},\mathbf{1};\boldsymbol{\alpha})}\middle|\,\mathrm{Ker}\ (\mathcal{L}^*)\right\} \tag{4.13}$$

となる．

式(4.13)の右辺第 2 項は一般には複雑で明示的には表せないが，特別な場合，例えば 4.2 節で触れた「2 水準の欠測」の場合には，$\mathrm{Ker}\ (\mathcal{L}^*)$ の要素が(4.7)のように簡単な形になることから，(4.13)は以下のように表される．

$$\boldsymbol{\phi}^O(\boldsymbol{Z}^o,R,\boldsymbol{\beta},\boldsymbol{\alpha},\eta)=\frac{R\boldsymbol{\phi}^F(\boldsymbol{Z},\boldsymbol{\beta},\eta)}{e(\boldsymbol{V};\boldsymbol{\alpha})}-\frac{R-e(\boldsymbol{V};\boldsymbol{\alpha})}{e(\boldsymbol{V};\boldsymbol{\alpha})}E\{\boldsymbol{\phi}^F(\boldsymbol{Z},\boldsymbol{\beta},\eta)|\boldsymbol{V}\}. \tag{4.14}$$

ここで，$\boldsymbol{Z}=(\boldsymbol{V}^T,\boldsymbol{W}^T)^T$ であり，$\boldsymbol{V}$ は常に観測され，$\boldsymbol{W}$ は観測されるか(全ての要素が)欠測するかのどちらかである．また，R は $\boldsymbol{Z}$(の全ての成分)が観測されるかどうかを表す 2 値の指示変数で，$e(\boldsymbol{V};\boldsymbol{\alpha})$ は傾向スコア $P\ (R=1|\boldsymbol{V})$ に対するパラメトリックモデル(ロジスティック回帰モデル等)である．

4.5 二重頑健推定量 —2 水準の欠測の場合—

完全データ影響関数を固定するということは，実際には例えば，完全データに基づく推定関数(完全データ推定関数)を 1 つ与えることに対応する．すなわち，興味あるパラメータ $\boldsymbol{\beta}$ に対する完全データ推定関数 $\boldsymbol{u}^F(\boldsymbol{Z},\boldsymbol{\beta})$ が与えられると，(推定方程式を通して得られる RAL 推定量の影響関数として)完全データ影響関数

$$\boldsymbol{\phi}_{\boldsymbol{u}}^F(\boldsymbol{Z},\boldsymbol{\beta},\eta)=\left[E\left\{\frac{\partial \boldsymbol{u}^F}{\partial \boldsymbol{\beta}}(\boldsymbol{Z},\boldsymbol{\beta})\right\}\right]^{-1}\boldsymbol{u}^F(\boldsymbol{Z},\boldsymbol{\beta}) \tag{4.15}$$

が定まるが，このとき(4.14)の代わりに

$$\boldsymbol{u}^O(\boldsymbol{Z}^o,R,\boldsymbol{\beta})=\frac{R\boldsymbol{u}^F(\boldsymbol{Z},\boldsymbol{\beta})}{e(\boldsymbol{V};\boldsymbol{\alpha})}-\frac{R-e(\boldsymbol{V};\boldsymbol{\alpha})}{e(\boldsymbol{V};\boldsymbol{\alpha})}E\{\boldsymbol{u}^F(\boldsymbol{Z},\boldsymbol{\beta})|\boldsymbol{V}\} \tag{4.16}$$

によって $\boldsymbol{\beta}$ に対する観測データ推定関数を与えると，これから得られる $\boldsymbol{\beta}$ の RAL 推定量の影響関数は

$$\boldsymbol{\phi}_{\boldsymbol{u}}^O(\boldsymbol{Z}^o,R,\boldsymbol{\beta},\boldsymbol{\alpha},\eta)=\frac{R\boldsymbol{\phi}_{\boldsymbol{u}}^F(\boldsymbol{Z},\boldsymbol{\beta},\eta)}{e(\boldsymbol{V};\boldsymbol{\alpha})}-\frac{R-e(\boldsymbol{V};\boldsymbol{\alpha})}{e(\boldsymbol{V};\boldsymbol{\alpha})}E\{\boldsymbol{\phi}_{\boldsymbol{u}}^F(\boldsymbol{Z},\boldsymbol{\beta},\eta)|\boldsymbol{V}\} \tag{4.17}$$

となる． 但しこれは(4.16)において，($E\{\boldsymbol{u}^F(\boldsymbol{Z},\boldsymbol{\beta})|\boldsymbol{V}\}$ の部分の計算に必要な)$\boldsymbol{V}$ が与えられた下での $\boldsymbol{W}$ の条件付き分布とパラメータ $\boldsymbol{\alpha}$ が既知であると仮定した場合のものである．しかしながら実際にはこれらは未知なので，推定関数(4.16)を用いるためにはこれらを観測データから推定する必要があるが，そのために $\boldsymbol{V}$ が与えられた下での $\boldsymbol{W}$ の条件付き分布(密度関数)に対しても，何らかのパラメトリックモデル $p(\boldsymbol{w}|\boldsymbol{v};\boldsymbol{\xi})$ $(\boldsymbol{\xi}\in\Xi\subset\boldsymbol{R}^d)$ を想定する． ここで，$\boldsymbol{W}$ は常に観測されるとは限らないが，その欠測メカニズムは MAR であると仮定しているので，$\boldsymbol{W}$ と R は $\boldsymbol{V}$ が与えられた下で条件付き独立であり，その結果，$\boldsymbol{V}$ が与えられた下での $\boldsymbol{W}$ の条件付き分布は $\boldsymbol{V}$ と R $(R=1)$ が与えられた下での条件付き分布と等しくなる．すなわち，パラメータ $\boldsymbol{\xi}$ は $\boldsymbol{V}$ と $\boldsymbol{W}$ がともに観測された場合(complete-case)のデータのみから推定可能である． このとき，観測データによる 2 つのパラメータ $\boldsymbol{\alpha},\boldsymbol{\xi}$ の最尤推定量をそれぞれ $\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\xi}}$ とすると，推定方程式

$$\sum_{i=1}^n\left[\frac{R_i\boldsymbol{u}^F(\boldsymbol{Z}_i,\boldsymbol{\beta})}{e(\boldsymbol{V}_i;\hat{\boldsymbol{\alpha}})}-\frac{R_i-e(\boldsymbol{V}_i;\hat{\boldsymbol{\alpha}})}{e(\boldsymbol{V}_i;\hat{\boldsymbol{\alpha}})}E\{\boldsymbol{u}^F(\boldsymbol{Z}_i,\boldsymbol{\beta})|\boldsymbol{V}_i;\hat{\boldsymbol{\xi}}\}\right]=\boldsymbol{0} \tag{4.18}$$

$$\left(\text{但し}\quad E\{\boldsymbol{u}^F(\boldsymbol{Z}_i,\boldsymbol{\beta})|\boldsymbol{V}_i;\hat{\boldsymbol{\xi}}\}=\int\boldsymbol{u}^F(\boldsymbol{V}_i,\boldsymbol{w},\boldsymbol{\beta})p(\boldsymbol{w}|\boldsymbol{V}_i;\hat{\boldsymbol{\xi}})d\boldsymbol{w}\right)$$

の解 $\hat{\boldsymbol{\beta}}_{DR}$ について，以下のことが成り立つ．

命題 6.(二重頑健推定量 —2 水準の欠測の場合—)

(1) 傾向スコア P $(R=1|\boldsymbol{V})$ に対するモデル $e(\boldsymbol{V};\boldsymbol{\alpha})$ と，$\boldsymbol{V}$ が与えられた下での $\boldsymbol{W}$ の条件付き分布に対するモデル $p(\boldsymbol{w}|\boldsymbol{v};\boldsymbol{\xi})$ のうち，少なくともどちらか一方が正しく特定されていれば，$\hat{\boldsymbol{\beta}}_{DR}$ は $\boldsymbol{\beta}$ に対する RAL 推定量となる．

(2) どちらのモデルも正しく特定されていれば，$\boldsymbol{\beta}$ に対する RAL 推定量 $\hat{\boldsymbol{\beta}}_{DR}$ の影響関数は(4.17)で与えられる．

さて，推定方程式(4.18)の左辺第 1 項だけに着目すると，(完全データ推定関数 $\boldsymbol{u}^F(\boldsymbol{Z},\boldsymbol{\beta})$ に基づく)$\boldsymbol{\beta}$ の IPW 推定方程式

$$\sum_{i=1}^n\frac{R_i\boldsymbol{u}^F(\boldsymbol{Z}_i,\boldsymbol{\beta})}{e(\boldsymbol{V}_i;\hat{\boldsymbol{\alpha}})}=\boldsymbol{0} \tag{4.19}$$

が得られるが，この解として定まる IPW 推定量 $\hat{\boldsymbol{\beta}}_{IPW}$ の影響関数は，(傾向スコアに対するモデルが正しく特定されていれば)観測データ影響関数の一般形(4.12)において

$$\boldsymbol{\phi}^F(\boldsymbol{Z},\boldsymbol{\beta},\eta)=\boldsymbol{\phi}^F_{\boldsymbol{u}}(\boldsymbol{Z},\boldsymbol{\beta},\eta),\ \boldsymbol{k}(\boldsymbol{Z}^o,\boldsymbol{R},\boldsymbol{\alpha})=\boldsymbol{0}$$

としたものになる． 但し，$\boldsymbol{\phi}^F_{\boldsymbol{u}}(\boldsymbol{Z},\boldsymbol{\beta},\eta)$ は(4.15)で与えられる，完全データ推定関数 $\boldsymbol{u}^F(\boldsymbol{Z},\boldsymbol{\beta})$ に対応する影響関数である． よって，命題 5 と命題 6 の(2)より，傾向スコアに対するモデルと $\boldsymbol{V}$ が与えられた下での $\boldsymbol{W}$ の条件付き分布に対するモデルのどちらも正しく特定されていれば，DR 推定量 $\hat{\boldsymbol{\beta}}_{DR}$ の漸近分散共分散行列は(少なくとも)IPW 推定量 $\hat{\boldsymbol{\beta}}_{IPW}$ の漸近分散共分散行列よりも小さくなる．

第 3 節でも述べたように，パラメータ $\boldsymbol{\beta}$ に対する(漸近的な)信頼領域などを得るためには，DR 推定量 $\hat{\boldsymbol{\beta}}_{DR}$ の漸近分散共分散行列とその推定量が必要となるが，推定方程式(4.18)において，パラメータ $\boldsymbol{\alpha}$ と $\boldsymbol{\xi}$ の推定の影響を考慮しながら DR 推定量 $\hat{\boldsymbol{\beta}}_{DR}$ の漸近分散共分散行列 $\mathrm{Avar}(\hat{\boldsymbol{\beta}}_{DR})$ を計算し，(それに対して)サンドイッチ推定量(3.5)を適用すれば，傾向スコアに対するモデルと $\boldsymbol{V}$ が与えられた下での $\boldsymbol{W}$ の条件付き分布に対するモデルのうち，少なくともどちらか一方が正しく特定されているという状況の下で，$\mathrm{Avar}(\hat{\boldsymbol{\beta}}_{DR})$ の一致推定量が得られる．それについて詳しくは，Tsiatis(2006)の第 10 章を参照されたい．

2.4 節において，最も単純な 2 水準の欠測問題の設定で IPW 推定量と DR 推定量を紹介したが，それらはそれぞれ完全データ推定関数 $\boldsymbol{u}^F(Y,\mu)=Y-\mu$ に対する IPW 推定方程式，DR 推定方程式の解として得られる推定量となっており，そこで述べた DR 推定量の(1)二重頑健性と(2)漸近有効性は命題 6 の特別な場合に当たる． 但し，その問題設定では，MAR と傾向スコアモデルのみが仮定され，完全データモデル S^F に対しては何も特別な制約を加えていない(つまり完全データに対してはノンパラメトリックモデルが想定されている)ので，興味あるパラメータ μ に対する完全データに基づく RAL 推定量の影響関数は唯一つしか存在しない．したがって，この場合においては，影響関数(4.17)は有効影響関数(漸近分散が最小となる RAL 推定量の影響関数)であり，命題 6 の(2)は，DR 推定量が(2 つのモデルが正しく特定されていれば)セミパラメトリック漸近有効な推定量であることを意味する． しかしながら一般には，影響関数(4.17)は有効影響関数であるとは限らず，DR 推定量も(2 つのモデルが正しく特定されていたとしても)セミパラメトリック漸近有効な推定量になるとは限らない．次節の例は，そのような場合についてのものである．

4.6 ロジスティック回帰モデルにおける共変量欠測

2 水準の欠測の例として，ここではロジスティック回帰モデルにおける共変量の欠測問題について考え，それを通して二重頑健推定量の構成法をもう少し具体的に見てみることにする．

Y を(1 または 0 の値を取る)2 値の結果変数とし，共変量(説明変数)については簡単のために実数値を取る 2 つの連続変数 X_1, X_2 を考え，これらについてロジスティック回帰モデル

$$\mathrm{logit}\{P(Y=1|X_1,X_2)\}=\beta_0+\beta_1X_1+\beta_2X_2 \tag{4.20}$$

が成り立っているものとする． 但し，Y と X_1 は常に観測されるが，X_2 については観測されるとは限らないという場合を想定し，X_2 の観測指示変数を R として(つまり X_2 が観測されれば $R=1$, 欠測すれば $R=0$)，その欠測メカニズムは MAR であると仮定する． 完全データ (Y,X_1,X_2) のランダムサンプルを $(Y_1,X_{11},X_{21}),\ldots,(Y_n,X_{1n},X_{2n})$ とすると，これらがもし全て観測されたならば，(4.20)における回帰パラメータ $\boldsymbol{\beta}=(\beta_0,\beta_1,\beta_2)^T$ の最尤推定量は推定方程式(スコア方程式)

$$\sum_{i=1}^{n}\{Y_i - e_Y(\boldsymbol{X}_i;\boldsymbol{\beta})\}\overline{\boldsymbol{X}}_i = \boldsymbol{0} \tag{4.21}$$

の解として与えられる．但し，$\boldsymbol{X}_i = (X_{1i}, X_{2i})^T$, $\overline{\boldsymbol{X}}_i = (1, X_{1i}, X_{2i})^T$, $e_Y(\boldsymbol{X}_i;\boldsymbol{\beta}) = 1/\{1 + \exp(-\boldsymbol{\beta}^T\overline{\boldsymbol{X}}_i)\}$ $(i = 1, \ldots, n)$ である．しかしながら，データ (Y_i, X_{1i}, X_{2i}) のうち X_{2i} は n 個体全てで観測されるとは限らないので，この(完全データ)推定方程式をベースとして，観測データに基づく DR 推定方程式(4.18)を構成すると，それは以下のように与えられる．

$$\sum_{i=1}^{n}\left[\frac{R_i\boldsymbol{u}^F(Y_i,\boldsymbol{X}_i;\boldsymbol{\beta})}{e_R(Y_i,X_{1i};\hat{\boldsymbol{\alpha}})} - \frac{R_i - e_R(Y_i,X_{1i};\hat{\boldsymbol{\alpha}})}{e_R(Y_i,X_{1i};\hat{\boldsymbol{\alpha}})}E\{\boldsymbol{u}^F(Y_i,\boldsymbol{X}_i;\boldsymbol{\beta})|Y_i,X_{1i};\hat{\boldsymbol{\xi}}\}\right] = \boldsymbol{0} \tag{4.22}$$

ここで，$\boldsymbol{u}^F(Y,\boldsymbol{X};\boldsymbol{\beta}) = \{Y - e_Y(\boldsymbol{X};\boldsymbol{\beta})\}\overline{\boldsymbol{X}}$, $e_R(Y,X_1;\boldsymbol{\alpha})$ は傾向スコア $P(R=1|Y,X_1)$ に対するパラメトリックモデル(ロジスティック回帰モデル等)で，$\hat{\boldsymbol{\alpha}}$ はそのモデルにおけるパラメータ $\boldsymbol{\alpha}$ の最尤推定量である．また $\hat{\boldsymbol{\xi}}$ は，Y と X_1 が与えられた下での X_2 の条件付き分布(密度関数)に対するあるパラメトリックモデル $p(x_2|y,x_1;\boldsymbol{\xi})$ におけるパラメータ $\boldsymbol{\xi}$ の最尤推定量であり，

$$E\{\boldsymbol{u}^F(Y_i,\boldsymbol{X}_i;\boldsymbol{\beta})|Y_i,X_{1i};\hat{\boldsymbol{\xi}}\} = \int \boldsymbol{u}^F(Y_i,X_{1i},x_2;\boldsymbol{\beta})p(x_2|Y_i,X_{1i};\hat{\boldsymbol{\xi}})dx_2 \tag{4.23}$$

である．但しここで重要なのは，パラメトリックモデル $p(x_2|y,x_1;\boldsymbol{\xi})$ は，完全データに対して(元々)仮定されているロジスティック回帰モデル(4.20)と整合するように設定されなければならないということである．例えば，$Y=0, Y=1$ の下での $\boldsymbol{X}$ の条件付き分布が，それぞれ平均ベクトルは異なるが，分散共分散行列は共通の二変量正規分布であると想定すると，ロジスティック回帰モデル(4.20)が成立し，またこのとき，Y と X_1 が与えられた下での X_2 の条件付き分布(密度関数)は以下のように与えられる．

$$p(x_2|y,x_1;\boldsymbol{\xi}) = n(x_2;\xi_0 + \xi_1x_1 + \xi_2y, \xi_{\sigma^2}) \ \ (\boldsymbol{\xi} = (\xi_0,\xi_1,\xi_2,\xi_{\sigma^2})^T). \tag{4.24}$$

但し，右辺は平均 $\xi_0 + \xi_1x_1 + \xi_2y$, 分散 ξ_{σ^2} の正規分布の密度関数を表すものとする．つまり，二重頑健推定のための，Y と X_1 が与えられた下での X_2 の条件付き分布(密度関数)に対する「作業モデル」として，例えば(4.24)のように設定しておけば，特に「$Y=0, Y=1$ の下での $\boldsymbol{X}$ の条件付き分布が，それぞれ平均ベクトルは異なるが，分散共分散行列は共通の二変量正規分布である」という場合には，元々のロジスティック回帰モデル(4.20)と両立し，矛盾がない．モデル(4.24)のパラメータ $\boldsymbol{\xi}$ は，MAR の下では Complete-Case(Y_i, X_{1i}, X_{2i} の全てが観測されている個体 i)のデータのみを用いて最尤法で推定されるが，今の場合は正規分布を仮定しているので，標準的な線型回帰の方法(最小二乗法)によって推定できる．一方，(4.23)の積分計算は簡単ではないが，数値積分やモンテカルロ法などにより実行可能である．

さて，推定方程式(4.21)は最尤推定量を導くスコア方程式であり，完全データに対してはモデル(4.20)の下で最適なものであるので，それをベースとして観測データに対する DR 推定方程式を構成するのは自然に思われるが，その解として得られる DR 推定量は，(推定方程式の構成に必要な 2 つのモデルが正しく特定されていたとしても)最適，つまりセミパラメトリック漸近有効になるとは一般には限らない．DR 推定方程式(4.22)はあくまで，完全データ推定関数を $\boldsymbol{u}^F(Y,\boldsymbol{X};\boldsymbol{\beta})$ に固定したときに最適な(漸近分散共分散行列が最小の)推定量を与えるものであり，それがセミパラメトリック漸近有効になる保証はなく，実際，そのようにはなっていない．このことはつまり，漸近有効な推定量を与えるような完全データ推定関数(あるいは影響関数)が別に存在する可能性を示唆しているが，(この場合も含めて)一般にそれは複雑なものとなる．この問題については，ここではこれ以上述べないが，詳しくは Tsiatis(2006)の第 11 章以降を参照されたい．

5. おわりに

前節では,「2 水準の欠測(two-level missingness)」という特別な欠測パターンの下で二重頑健推定量を導いたが,応用上もう 1 つ重要な欠測のパターンとして,「単調な欠測(monotone missingness)」と呼ばれるものがある. これは例えば経時データ解析において,一度脱落(dropout)した対象者は二度と対象とならない(つまり,脱落した時点の直前までしかデータが観測されず,以後の時点では全て欠測する)というような欠測パターンのことであるが,この場合も,観測データ影響関数(4.13)の右辺第 2 項(Ker ($\mathcal{L}^*$) への直交射影の部分)を明示的に表すことが可能であり,その表現に基づいて二重頑健推定量が導出される.(完全データ推定方程式として一般化推定方程式(GEE)を用いる場合が,その典型例である.)しかしながら,欠測のパターンが非単調な場合(一般の場合)には,原理的には二重頑健推定量を考えることはできるが,既に述べたように,観測データ影響関数(4.13)の右辺第 2 項は一般には明示的には表せず計算も単純ではなく,またデータの全ての要素が観測される確率(Complete-Case の確率)$\omega(\boldsymbol{Z}, \mathbf{1}; \boldsymbol{\alpha}) = P(\boldsymbol{R} = \mathbf{1}|\boldsymbol{Z})$ の推定も一筋縄には行かないため,応用上はまだ困難な問題が残っている.

「二重頑健推定量」という名称は,その推定量を導く推定方程式を構成するのに必要な 2 つのパラメトリックモデルのうち,どちらか一方が正しく特定されていればその推定量は一致性を持つという性質(二重頑健性)に由来するが,本稿で扱ったセミパラメトリック推測理論の立場から見た場合は,その性質はどちらかと言えば副次的なものであり,理論ではむしろ推定量の漸近的な推定精度の向上(漸近分散を可能な限り小さくするということ)の方に焦点が置かれている. 欠測データ解析法におけるこの「二重頑健性」については他の立場からも考察がなされているが,それについては例えば,Bang and Robins(2005),Kang and Schafer(2007)等を参照されたい. また,二重頑健性という性質自身は一般的なものであり,欠測データ解析法に限らず他の場面にも現れ得るものである. Robins and Rotnitzky(2001)では,そのための条件について,セミパラメトリック推測理論の立場から一般的な考察がなされている.

二重頑健推定量は,二重頑健性と漸近有効性(あるいは漸近的な推定精度の向上)の双方の面から,逆確率重み付け推定量(IPW 推定量)よりは優れていると言えるが,その構成に必要な 2 つのパラメトリックモデルのどちらも誤特定している場合には,必ずしも良い挙動を示すとは言えない. そこで近年,それらのモデルの誤特定に対してより頑健性の強い「多重頑健推定量」がいくつか提案されている(Han and Wang, 2013; Chan, 2013; Hattori and Henmi, 2014). これは,推定対象となる 2 つの要素は変わらないが,その推定のためのモデルとしてそれぞれ複数の候補を許容し,それらのうちのどれか 1 つが正しく特定されていれば,興味あるパラメータの一致推定量が得られるという方法である. また,モデルの誤特定の問題以外にも実用上いくつかの問題点が指摘されており,それらに対処するための改良版二重頑健推定量もいくつか提案されている(Kang and Schafer, 2007; Cao et al., 2009; Tan, 2010; Rotnitzky et al., 2012).

本稿では,欠測のメカニズムが MAR であるという仮定の下で議論を行ったが,2.1 節でも述べたように,その仮定は原理的に観測データからは検証できないものなので,データ取得のためのデザインなどから予め MAR であることが保証されているような場合を除いては,MAR からの乖離の影響を調べる感度解析などを行うことが望ましい. IPW 推定や DR 推定(二重頑健推定)に対するこの種の感度解析法については,例えば Scharfstein et al.(1999),Li et al.(2011),Shen et al. (2011)等がある.

IPW 推定量や DR 推定量は,(直接的な)欠測の問題だけでなくしばしば因果推論の問題でも扱われるが,これは,因果推論の問題が潜在的結果変数(potential outcome),あるいは反事実結果変数(counterfactual variable)と呼ばれるデータの欠測問題として見られることによる. この点を通して両者の問題は類似した構造を有しているが,例えば本稿で用いた「傾向スコア」とい

う用語は，もともとは因果推論の文脈で提案されたものである(Rosenbaum and Rubin, 1983)．因果推論における IPW 推定量や DR 推定量(特に後者)については，Tsiatis(2006)の第 13 章や Bang and Robins(2005)等を参照されたい．

欠測データに対する二重頑健推定法は，まだ応用上の課題は残っているものの，完全データの分布全体に対するパラメトリックなモデル化を必要としないセミパラメトリックな解析法としていくつかの有用な性質を持ち，今後も有力な方法として発展していくと思われる．本稿ではセミパラメトリック推測の一般論の立場から，その理論的な側面を中心に解説を行ったが，応用上の観点からは例えば，ロジスティック回帰モデルの共変量欠測については Tchetgen Tchetgen(2009)，一般化推定方程式(GEE)における経時データの欠測については Seaman and Copas(2009)等が参考になる．

参 考 文 献

Bang, H. and Robins J. M. (2005). Doubly robust estimation in missing data and causal inference models, *Biometrics*, **61**, 962–972.

Boos, D. D. and Stefanski, L. A. (2013). *Essential Statistical Inference*, Springer, New York.

Cao, W., Tsiatis, A. A. and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data, *Biometrika*, **96**, 723–734.

Carpenter, J. and Kenward, M. (2013). *Multiple Imputation and Its Application*, Wiley, Chichester.

Chan, K. C. G. (2013). A simple multiply robust estimator for missing response problem, *Stat*, **2**, 143–149.

Eekhout, I., de Boer, M. R., Twisk, J. W. R., de Vet, H. C. W. and Heymans, M. W. (2012). Missing data: A systematic review of how they are reported and handled, *Epidemiology*, **23**, 729–732.

Han, P. and Wang, L. (2013). Estimation with missing data: beyond double robustness, *Biometrika*, **100**, 417–430.

Hattori, S. and Henmi, M. (2014). Stratified doubly robust estimators for the average causal effect, *Biometrics* (to appear).

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, **47**, 663–685.

Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data, *Statistical Science*, **22**, 523–539.

Li, L., Shen, C., Wu, A. C. and Li, X. (2011). Propensity score-based sensitivity analysis method for uncontrolled confounding, *American Journal of Epidemiology*, **174**, 345–353.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed., Wiley, New Jersey.

Robins, J. M. and Rotnitzky, A. (2001). Comment on the Bickel and Kwon article, "On double robustness," *Statistica Sinica*, **11**, 920–936.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed, *Journal of the American Statistical Association*, **89**, 846–866.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika*, **70**, 41–55.

Rotnitzky, A., Lei, Q., Sued, M. and Robins, J. M. (2012). Improved double-robust estimation in missing data and causal inference models, *Biometrika*, **99**, 439–456.

Rubin, D. B. (1976). Inference and missing data, *Biometrika*, **63**, 581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York.

Scharfstein, D. O., Rotnitzky, A. and Robins, J. M. (1999). Adjusting nonignorable drop-out using semiparametric nonresponse models, *Journal of the American Statistical Association*, **94**, 1096–1120.

Seaman, S. and Copas, A. (2009). Doubly robust generalized estimating equations for longitudinal data, *Statistics in Medicine*, **28**, 937–955.

Seaman, S. R. and White I. R. (2011). Review of inverse probability weighting for dealing with missing data, *Statistical Methods in Medical Research*, **22**, 278–295.

Shen, C., Li, X., Li, L. and Were, M. C. (2011). Sensitivity analysis for causal inference using inverse probability weighting, *Biometrical Journal*, **53**, 822–837.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting, *Biometrika*, **97**, 661–682.

Tchetgen Tchetgen, E. J. (2009). A simple implementation of doubly robust estimation in logistic regression with covariates missing at random, *Epidemiology*, **20**, 391–394.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*, Springer, New York.

van der Vaart, A. W. (2000). *Asymptotic Statistics*, Cambridge University Press, Cambridge.

122 Proceedings of the Institute of Statistical Mathematics Vol. 62, No. 1, 103–122 (2014)

Semiparametric Statistical Methods for Missing Data

Masayuki Henmi

The Institute of Statistical Mathematics

In observational studies such as cohort studies in epidemiology, results of statistical analysis can be biased for various reasons such as missing data. There are two kinds of statistical methods for missing data. One is a parametric method such as the maximum likelihood method and multiple imputation. The other is a semiparametric method such as inverse probability weighting and doubly robust estimation. This paper focuses on the latter, which has been rapidly developed in resent years especially in biostatistics. Although there have been few applications of the semiparametric method in practice, it is expected as a powerful method for missing data and its demand should increase in the future. Semiparametric statistical methods are widely used for various problems in biostatistics. The argument in this paper is based on general theory of semiparametric inference in order to help our theoretical understanding of other cases as well as missing data problems.

Key words: MAR, influence functions, estimating functions, nuisance tangent spaces, asymptotic efficiency.