2008 年 1 月 16 日-18 日，広島大学理学部, 科研費研究集会「統計的データ解析手法の評価と開発」

観察データの推測の限界－揺らぎモデルアプローチ－

江口真透 （統計数理研究所, 総合大学院大学統計科学専攻）

1. はじめに

観察研究から得られたデータに対して標準的な統計的方法を単純に適用することは不適切な結論を導く恐れがある．このことは観察データは標準的な統計的方法が正当化される暗黙の仮定を満たさないことから生じる問題である．このようにデータ適用性を巡って統計学の基本的なパラダイムにこの重大な問題が議論された長い歴史がある [1, 3, 6, 7, 11]．選択バイアスの問題が多くの文献で指摘され，バイアスに対する方法が考察されている．しばしばミッシング，打ち切り，未観測交絡など不完全観測を伴うデータにはランダム・サンプルの仮定への乖離が生じる可能性があり，しかもほとんどの場合その選択バイアスの有無はデータからは検証できない [9].

バイアスに対する研究は膨大な文献に著され，様々なアプローチが展開されている．その中で私たちのアプローチは，仮定された統計モデルのミススペシケイションを表現するために，統計モデル包含する管状の近傍を考察した枠組み [4] を利用して，モデルからの乖離を記述するパラメータを導き，そのパラメータによる感受性解析をすることから始まった [2]．例題としてヒロシマ被爆者の生存解析において遅延登録によるバイアスの問題を考察しました [10]．この研究を発展させて，全ての選択バイアスの影響を許容した信頼領域の構成法を提案した [3]．メタ解析における出版バイアスの問題については，選択関数のグローバルな影響を考察して P 値と信頼量区間の限界を求めた [8]. その補正公式は簡明で最尤推定量の漸近分散を 2 倍にすることによって得られる．間接喫煙から肺がんの危険性を評価する問題が、動機を与えている例として考察された．

2. モデルの不確定性

必ずしも全てが観測できない仮想的な確率変数 Z に対して統計モデル $M_Z = \{f_Z = f_Z(z,\theta) : \theta \in \Theta\}$ が考えられたとする．統計モデル M_Z から仮想的な観測ベクトル $z_1, \cdots, z_n$ が得られたと考える．実際には観測ベクトルの分布が正確に統計モデル M_Z にあると考えるよりも，管状近傍

$$\mathcal{N}_\epsilon(M_Z) = \{g_Z : \min_{f_Z \in M_Z} D(g_Z, f_Z) < \epsilon\}$$

にあると考えた方が多くの場合で正確である．ここで D は分布の隔たりを表す尺度を表す．以下の議論では θ の推測には尤度法を考えることから D はクルバック・ライブラーのダイバージェンスを考えることが自然である [12]．実際，確率変数 Z が連続分布を持つならば，$\mathcal{N}_\epsilon(M_Z)$ は関数空間となるが，十分小さい ϵ を選べば，

$$\mathcal{L}_\epsilon(f_Z) = \{g_Z \in \mathcal{N}_\epsilon(M_Z) : D(g_Z, f_Z) = \min_{f_Z^* \in M_Z} D(g_Z, f_Z^*)\}$$

は分布 f_Z を貫くリーフとなり，統計モデル M_Z 上の全ての f_Z をとってリーフ $\mathcal{L}_\epsilon(f_Z)$ を重ねると Z の可能な分布全体の中で葉層を成すことが分かる．このようなリーフ $\mathcal{L}_\epsilon(f_Z)$ は

$$g_Z = g_Z(z, \epsilon, u_Z) = f_Z(z,\theta) \exp\{\epsilon u_Z(z,\theta) - \tfrac{1}{2}\epsilon^2\} \tag{1}$$

と書ける [2]．ここで $u_Z(z,\theta)$ は分布 f_Z の下で平均 0, 分散 1 を持つとする．以下の議論では, u_Z が更にモデル M_Z のスコアーベクトル s_Z と無相関であると仮定する．このような仮定からデータ分布 (1) を持つ完全データ $z_1, \cdots, z_n$ の最尤推定量 $\hat{\theta}_Z$ はバイアスをまたないことが判る：

$$\mathbb{E}_{f_Z}(\hat{\theta}_Z) = O(n^{-1}) + O(\epsilon^2) \tag{2}$$

以下では現実には不完全データしか得られない状況を考える．このとき観測の不完全性によるモデルの不確定性によってもはや最尤推定量は (2) のような不偏性が保たれなくなり，選択バイアスが現われるが分かる．

3. 不完全な観測

前節で考察したように完全データのモデルの不確定性は理想的に最尤推定量のバイアスは漸近的に無視可能となった．この節ではミッシング，打ち切り，未観測交絡，競合リスクなど不完全観測を伴う不完全な観察によるデータによって仮定されたモデルから微小な乖離が生じた状況を考えよう．データの不完全性がもたらす推定量の分散とバイアスの影響を表す漸近的な公式を与える．分散の影響はバイアスの影響に比べ無視可能である．これより古典的な α %信頼領域の形状は保たれるがバイアスの影響によって中心が揺らい

でしまう．この選択バイアスを許容するときのα%信頼領域の補正を提案した．この補正された信頼領域は検出不可能な選択バイアスによって揺らいだ古典的な信頼領域を常に包含することが示された．

不完全データ確率ベクトル $\boldsymbol{Y}$ が完全データ確率ベクトル $\boldsymbol{Z}$ に対して $\boldsymbol{Y}=\boldsymbol{h}(\boldsymbol{Z})$ と書かれたとする．ここで $\boldsymbol{h}$ は多対1の関数とする．$\boldsymbol{Y}$ のモデル分布は, $\boldsymbol{X}$ のモデル分布 $f_{\boldsymbol{Z}}(\boldsymbol{z},\boldsymbol{\theta})$ から誘導されて，形式的に

$$f_{\boldsymbol{Y}}(\boldsymbol{y},\boldsymbol{\theta})=\int_{\boldsymbol{h}^{-1}(\boldsymbol{y})} f_{\boldsymbol{Z}}(\boldsymbol{z},\boldsymbol{\theta})\mathrm{d}\boldsymbol{z} \tag{3}$$

と書かれる．ここで $\boldsymbol{h}^{-1}(\boldsymbol{y})$ は $\boldsymbol{h}$ の逆像を表す．このようにして完全データ $\boldsymbol{Z}$ のフィッシャー情報行列 $I_{\boldsymbol{Z}}$ は $\boldsymbol{h}$ によって不完全データ $\boldsymbol{Y}$ のフィッシャー情報行列 $I_{\boldsymbol{Y}}$ は情報損失を受ける．実際，$\boldsymbol{Y}$ のスコアーベクトル $\boldsymbol{s}_{\boldsymbol{Y}}$ は $\boldsymbol{s}_{\boldsymbol{Z}}$ の $\boldsymbol{Y}=\boldsymbol{y}$ の条件付期待値で表され，情報損失は $I_{\boldsymbol{Z}}-I_{\boldsymbol{Y}}=\mathrm{var}(\boldsymbol{s}_{\boldsymbol{Z}}-\boldsymbol{s}_{\boldsymbol{Y}})$ となる．もし不完全データ $\boldsymbol{y}_1,\cdots,\boldsymbol{y}_n$ がモデル分布 $f_{\boldsymbol{Y}}$ に従っているならば，$\boldsymbol{\theta}$ の最尤推定量 $\hat{\boldsymbol{\theta}}$ は $\boldsymbol{\theta}$ の漸近的不偏推定量で漸近正規性

$$\mathrm{Prob}\{n^{-\frac{1}{2}}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})\le\boldsymbol{\delta}\}\longrightarrow\int_{-\boldsymbol{\infty}}^{\boldsymbol{\delta}} dN(0,I_{\boldsymbol{Y}}^{-1}) \tag{4}$$

を持つ．ここで $dN(0,I_{\boldsymbol{Y}}^{-1})$ は平均ベクトル $\mathbf{0}$, 分散行列 $I_{\boldsymbol{Y}}^{-1}$ を持つ正規分布を表す．この性質より，点推定として $\hat{\boldsymbol{\theta}}$ を採れば，水準 α を持つ信頼領域は

$$\mathrm{CR}(\alpha)=\{\boldsymbol{\delta}:n(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^{\mathrm{T}}I_{\boldsymbol{Y}}^{-1}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})\le r_\alpha\} \tag{5}$$

として与えられる．ここで r_α はカイ2乗分布の α 分位点とする．これらの統計的方法の漸近有効性も標準的な考察から得られる．しかしながら，これらの性質は，不完全データがモデル分布に従っているとする仮定よって支持されるものである．観察データがミッシングなどの不完全観測から得られたときに，この仮定が成立している保証はない．

これから，モデルの不確定性を考慮して完全データの分布を (1) で定義した $g_{\boldsymbol{Z}}$ であると仮定する．上の式 (3) の考察と同様に，$\boldsymbol{Y}$ のデータ分布は

$$g_{\boldsymbol{Y}}(\boldsymbol{y},\boldsymbol{\theta},\epsilon,u_{\boldsymbol{Y}})=f_{\boldsymbol{Y}}(\boldsymbol{y},\boldsymbol{\theta})\exp\{\epsilon u_{\boldsymbol{Y}}(\boldsymbol{y},\boldsymbol{\theta})-\tfrac{1}{2}\epsilon^2\rho_{\boldsymbol{Y}}\} \tag{6}$$

となる．ここで $u_{\boldsymbol{Y}}$ は $u_{\boldsymbol{Z}}$ の条件 $\boldsymbol{Y}=\boldsymbol{y}$ の条件付期待値で，$\rho_{\boldsymbol{Y}}$ は $u_{\boldsymbol{Z}}$ の分散を表す．$\boldsymbol{\theta}$ の最尤推定量 $\hat{\boldsymbol{\theta}}$ はモデル分布 $f_{\boldsymbol{Y}}$ の仮定の下では，上の議論のように標準的な漸近性を持つが，データ分布 $g_{\boldsymbol{Y}}$ であれば，どうなるだろうか．実は，完全データの場合に仮定された綺麗な構造 (2) は不完全関数 $\boldsymbol{h}$ によって崩される．

以上の考察から不完全データによる最尤推定量 $\hat{\theta}_Y$ のバイアスは

$$b_\theta=\mathbb{E}_{f_Z}(\hat{\theta}_Y-\hat{\theta}_Z)=\epsilon^2\mathbb{E}_{f_Y}\{u_Y I_Y^{-1}s_Y\} \tag{7}$$

と表現される．

この考察を進めると選択バイアスの問題は2重の困難が横たわることが分かる．観察データを標準的な方法で解析すると無視できないバイアスが生じて標準的な推測が壊れてしまう困難さ，しかもその最悪となるケースに近づくと，そのバイアスが推定不能となるという困難さに会う．すなわち，観察データの解析者は，いつも，大きな誤りを導くかもしれない既存の統計方法を恐る恐る使わなければならない現実がある．

実際的な対処として，選択バイアスを推定する代わりに可能な値を与えてデータ分布 (6) の下で $\boldsymbol{\theta}$ を推定する方法が提案されている [2]．このようなアプローチを一般に感度分析と呼ぶ [9]. しかしながら感度分析で分かることは，統計方法が，仮想的な選択バイアスを与えたときにどの位の影響を受けるかをモニターできることが精一杯である．

更にこの感度分析をある反事実仮想を行うことによって，自動的に選択バイアスの影響の限界を計る方法を紹介する．これによって，選択バイアスがあったとしても信頼水準やP値を保証するように統計量を補正する簡単な方法が導かれることが紹介された．

参考文献

[1] Cochran, W. G. *J. Royal Statist. Soc.* A, 128 (1965) 234-255
[2] Copas, J. and Eguchi, S. *J. Royal Statist. Soc.* B, 63 (2001) 871-895.
[3] Copas, J. and Eguchi, S. *J. Royal Statist. Soc.* B, 67, 459-512 (2005).
[4] Eguchi, S. and Copas, J. *Biometrika*, 89, 1-22 (2002).
[5] Eguchi, S. and Copas, J. *J. Royal Statist. Soc.* B 60, (1998) 709-724.
[6] Greenland, S. (2005) *J. Royal Statist. Soc.* A, 168, 267-306.
[7] Heckman, J.J. *Econometrica* 47, 1 (1979) 153-162.
[8] Henmi, M., Copas, J. and Eguchi, S. *Biometrics* 63 (2007) 475-482.
[9] Little, R. J. A. and Rubin, D. B. (2002) *Statistical Analysis with Missing Data.* 2nd edn. New York: Wiley.
[10] Matsuura, M. and Eguchi, S. *Biometrics*, 61, 559-566 (2005).
[11] Rosenbaum, P. R. and Rubin, D. B. *J. Amer. Statist. Assoc.* 79, 387 (1984) 516-524.
[12] White, H. (1982) *Econometrica*, 50, 1-25.