Biased samplingでの母集団モーメントや母数の推定

慶應義塾大学経済学研究科　清水祐弥　　　　　慶應義塾大学・理研 AIP　星野崇宏

1　問題意識

偏りのある抽出が行われたデータから, 母集団モーメントや母数 θ の推定を行う方法を考える. 関心のある変数を y, 共変量を x とする. このようなデータは, もし仮想的に無作為抽出標本 (サンプルサイズ $= N$) が得られていたならば, 無作為抽出標本のうち n ユニットでは観測がなされ ($r = 1$), 残りの $N - n$ ユニットでは観測がされない ($r = 0$) と考えることが可能である. しかし, 一般のバイアスのある抽出では通常 N や $N - n$ 自体が不明であり, あくまで n ユニットの偏りのある標本のみが観測される (図 1 参照). この状況では通常の欠測データ解析の枠組みは利用できないが, もし共変量に関して別の無作為抽出標本 (サイズ $= M >> N$) が得られていれば, バイアスの補正が可能となる.

同様の状況下での分析は, 機械学習分野における PU 分類 (Elkan and Noto (2008)) や, 生物統計学分野における Presence-Only データの分析 (Ward *et al.* (2009)) にみられる.

既存の研究の精度は, バイアス補正のための傾向スコアのモデリングに依存しており, もしモデルを誤設定すると一致推定量が得られない. そこで本研究では, たとえ傾向スコアが正しく設定されていなくても, 共変量を条件付けたとき変数の分布が正しく設定されていれば, 一致推定量を得られる二重にロバストな推定方法を提案する.

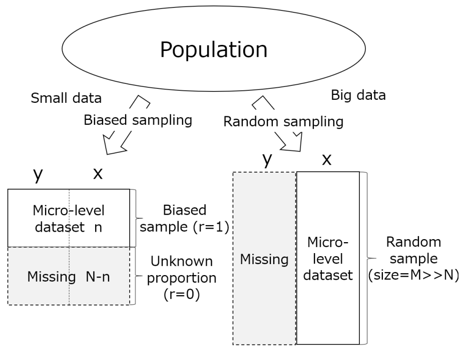

図 1: 本研究で扱うデータのイメージ

2　二重にロバストな推定

x を条件付けたときに y と r が独立であると仮定する.

推定したいパラメータ θ に対して不偏推定方程式 $\psi(y|\theta)$, s.t. $E[\psi(y|\theta)] = 0$ を考える. このとき, ① 傾向スコア $p(r = 1|x)$[1]又は密度比 $\dfrac{p(x)}{p(x|r = 1)}$ が正しく推定される場合, もしくは ②biased sampling されたデータから推定した条件付き分布 $\hat{p}(y|x)$ が真の分布 $p(y|x)$ と一致する場合のどちらかであれば, 次式の θ についての解がその一致推定量となる.

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{p}(r=1)}{\hat{p}(r=1|x_i)}\left\{\psi(y_i|\theta)-\frac{1}{L}\sum_{l=1}^{L}\psi(y_i^l|\theta)\right\}+\frac{1}{M}\sum_{m=1}^{M}\frac{1}{L}\sum_{l=1}^{L}\psi(y_m^l|\theta)=0,$$

ただし, $y_i^l \sim \hat{p}(y|x_i),\ l = 1,\ldots,L,\ i = 1,\ldots,n,$ また, $y_m^l \sim \hat{p}(y|x_m),\ l = 1,\ldots,L,\ m = 1,\ldots,M$ である.

引用文献

- Elkan, Charles, and Keith Noto. ”Learning classifiers from only positive and unlabeled data.” Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008.
- Nevo, Aviv. ”Using weights to adjust for sample selection when auxiliary information is available.” Journal of Business Economic Statistics 21.1 (2003): 43-52.
- Ward, Gill, et al. ”Presence-only data and the EM algorithm.” Biometrics 65.2 (2009): 554-563.

[1] この状況では傾向スコアは通常の方法で推定できないので, Nevo(2003) の方法を用いて推定する.