隠れた母集団に対する逆推定のシミュレーション

前田豊 *(立教大学)*・朝岡誠 *(立教大学)*

2017/8/6

問題関心

社会学では伝統的に周辺的な集団に対して一定の関心を払ってきた (e.g. シカゴ学派の系譜を継ぐ都市社の方々)．しかし，一般的にこうした周辺的な集団の母集団における構成比率は極めて小さく，またその存在についても必ずしも明確ではないため，ランダムサンプリングを理想とする従来型の抽出方法からは捉えることが難しい．加えて，例えば違法薬物使用者や売買春従事者，同性愛者などのように，属性に違法性やスティグマが付随している場合は，その属性であるという事実を回答者が秘匿するため，調査に非協力であったり，もしくは正確な回答を忌避する可能性がある．こうした従来型の標本調査から十分に捉えることが難しい集団を「隠れた母集団」(e.g. Heckathorn 1997) と呼び，彼（女）らの情報を十分に，そして正確に捉える抽出方法として，Snowball sampling(e.g. Goodman 1961) や Key informant sampling(Deaux and Callaghan 1985)，Respondent Driven sampling(e.g. Heckathorn 1997) といった回答者のネットワークを利用した種々の抽出方法 (Chain-referral sampling) が提唱され[*1]，例えば売春 (McCarthy et.al.2014)，自殺未遂者 (Clements-Nolle.et.al 2006), 薬物使用 (Heckathorn et.al. 2002) といった場面での適用が進められている．

いま，隠れた母集団を特徴づける属性，例えば売買春従事者や薬物使用者であるといった属性 $Y = 1$（$Y = 0$: その他）と共変量 X がどの程度関連しているのかに問題関心があるとしよう．つまり，$Pr(Y|X)$ に関心があり，何らかの統計量・回帰モデルから評価することを想定する．そして，上述の回答者ネットワークを利用した抽出方法を適用して標本を得ることを考える．このとき，典型的には $Y = 1$ である任意の初期回答者を起点として，彼（女）が持つネットワークを経由し，一定のサンプルサイズに到達するまで $Y = 1$ である回答者の情報を収集する手順が取られるので (e.g.McCarthy et.al.2014)，結果として，$Y = 1$ であることに条件づけられた X の標本が得られる（以下，$\{X, Y = 1\}$）．

こうした従属変数 Y の値に基づく抽出は，計量経済学において “choice-based sampling” や “endogenous sampling” といった名称で呼ばれ，たとえ従属変数の値に基づいて抽出された標本であっても，従属変数の値が完備されている場合，つまり，$\{X, Y = 1\}$ に加えて $Y = 0$ であることに条件づけられた X の標本（以下，$\{X, Y = 0\}$）も同時に得ることができれば，尤度に基づいて一致推定量を導けることが知られている (c.f.

[*1] 例えば，Respondent Driven sampling の基本的な手順は以下の通りである (Heckathorn 1997:179)．1) 調査者が起点 “seed” となる回答者をリクルートする．2)“seed” になった回答者は自身のネットワークを介してほかの回答者をリクルートする．リクルートのインセンティブとして，紹介した回答者が無事に調査を受けた場合に紹介元の回答者に謝金を支払う．3) 紹介された回答者も “seed” と同様に自身のネットワークから紹介を行い，自身が調査に協力にしたこと，そして紹介に成功したことに対する 2 つの謝金を受け取る．4) 焦点が当てられていた集団を十分に補足できた場合，もしくは計画したサンプルサイズに達するか，標本構成が安定的な水準に達した場合に，調査を終了する．また，ネットワーク構造の特性も考慮した研究として Salganik and Heckathorn(2004) を参照.

Amemiya 1985:319-338)．また，この問題は医療統計学において “case control study” として定式化されており，条件への不変性を持つオッズ比による評価が確立している (e.g.Breslow 1996)[*2]．

しかし，隠れた母集団の特性，とくに属性に付随する違法性やスティグマに起因した調査協力傾向を考慮した場合，これらの尤度・オッズ比による評価に必要な $\{X, Y = 0\}$ の標本をどのように収取するのかも問題となる．実際に売買春に従事していない個人や薬物を使用していない個人であれば，その経験を尋ねられたとしても，わざわざ虚偽の回答をするインセンティブは存在しないが[*3]，実際に売買春に従事している個人や薬物を使用している個人の場合，その属性であることを表明することに忌避感を覚えると予測できるため，調査協力へのインセンティブとの比較考量から正確に回答するか否かを決定すると考えられる．もし，この理解が正しいのであれば，本来の属性を Y^* とし，調査協力を行うか否かがランダムに決まっていると仮定すると，実際の標本 $\{X, Y = 1\}$ には本来の属性 $Y^* = 1$ の個人のみが含まれているが，一方の標本 $\{X, Y = 0\}$ には $Y^* = 0$ である個人に加えて，虚偽の回答を行った $Y^* = 1$ の個人が含まれることになる．また，回答を拒否した場合は，$\{X, Y = 0\}$ における欠測が，欠測している真の値に依存して欠測する NMAR になってしまう．それゆえ，結果に生じる標本 $\{X, Y = 0\}$ の歪みに起因して，例えば尤度関数の特定ミスが生じ，また $Pr(X|Y^* = 0)$ の推定が不確かなものとなるため，尤度，およびオッズ比からのアプローチを適用することが難しくなる．こうした $\{X, Y = 0\}$ の代表性に関わる実践上の対処方法として，例えば McCarthy et.al.(2014) の売春従事者を対象にした分析では，スタイリストや飲食店サービス従業者を具体的な $Y^* = 0$ の個人として定義したうえで，それらの職業に従事する個人と売春従事者とをそれぞれを対象にした 2 つの独立した Respondent Driven sampling を行なっている．しかし，この方法ではどのような属性が $Y^* = 0$ であるかについての調査者の恣意性が介在してしまうため，本質的な解決策にはなってはいない．

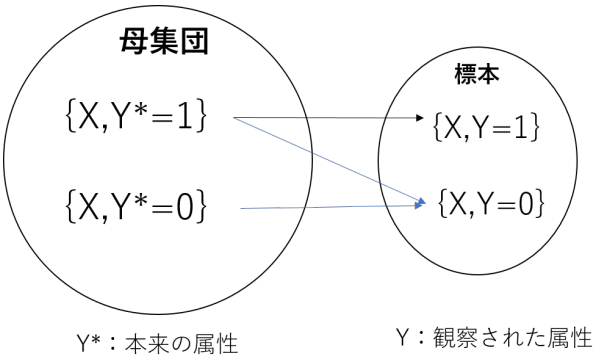

図1

だが，近年，必ずしも隠れた母集団との対応は明示的ではないものの，代表性が疑わしい $\{X, Y = 0\}$ の標本を用いず，代わりに Y の値は欠損しているが X を完備している標本（以下，$\{X\}$）を用いて，$\{X, Y = 1\}$ の標本から $Pr(X|Y)$ にアプローチするいくつかの方法が理論的に提唱されている．独自に抽出が必要な $\{X, Y = 0\}$

[*2]
$$\frac{Pr(X=1|Y=1)Pr(X=0|Y=0)}{Pr(X=1|Y=0)Pr(X=0|Y=1)} = \frac{Pr(Y=1|X=1)Pr(Y=0|X=0)}{Pr(Y=1|X=0)Pr(Y=0|X=1)}$$

[*3] ただし実際には，例えば Respondent Driven sampling の場合，リクルート (紹介した回答者が回答した場合) と回答協力にそれぞれ謝金を払うプロセスを取っており，隠れた母集団の属性ではない個人を紹介する可能性が存在している．それゆえ，回答者属性の確認が抽出のステップにおいて必須になっている (Heckathorn 1997:179)．

の標本に比べて，$\{X\}$ は，例えばセンサスや無作為抽出による大規模社会調査の個票データを利用することができるため，より簡便に入手することが可能で，また調査者の恣意性が介入しづらい点から，より信頼性が認められる手法として評価できるだろう．しかし，これら提唱された手法を実際のデータに適用したケースは，管見の限り，あまり存在しておらず，選択した手法のタイプに起因して応用上でどのような差異が生じるのかについては十分には理解されていない．そこで，本研究では，比較的実装が簡単な疑似尤度を用いたアプローチとして，Steinberg and Cardell(1992)（以下，SC 推定）と Tang, Little and Raghunathan(2003)（以下，TLR 推定）を取り上げ，この 2 つの手法の概要を簡単に述べたのちに，モンテカルロシミュレーションを行い，これらの手法の特徴を簡単に議論する[*4]．また，加えて，本研究では，統計量として理解する手法として，Manski(2003) の相対リスクに基づくアイディアをログリニアモデルから拡張した方法を提唱する．以下，$\{X, Y = 1\}$ を**完全データ**，$\{X\}$ を**補足データ**と呼び，それぞれ対応する母集団の特性 $Pr(X|Y = 1)$ と $Pr(X)$ を正しく推定することが可能であると仮定する．

ログリニアモデルによる相対リスクの拡張

検討を行う一つ目の方法は，相対リスク (Manski 2003) をログリニアモデルから拡張する方法である．いま簡単に X を二値変数とすると，相対リスクは以下のように示すことができる．

$$\frac{Pr(Y = 1|X = 1)}{Pr(Y = 1|X = 0)}$$

相対リスクを完全データに基づき評価しようとしても，X の値に依存せずにすべての個人が $Y = 1$ になっているため，必ず 1 になってしまう．また，補足データを用いたとしても，Y が欠測しているため，計算することができない．しかし，相対リスクはベイズの公式から

$$\frac{Pr(Y = 1|X = 1)}{Pr(Y = 1|X = 0)} = \frac{Pr(X = 1|Y = 1)\frac{Pr(Y=1)}{Pr(X=1)}}{Pr(X = 0|Y = 1)\frac{Pr(Y=1)}{Pr(X=0)}} = \frac{Pr(X = 1|Y = 1)}{Pr(X = 0|Y = 1)}\frac{Pr(X = 0)}{Pr(X = 1)}$$

と変形することができる．最右辺の第一項は完全データから計算することができ，また第二項は補足データより計算可能なので，関心のあった X に条件づけられた Y の相対リスクを，これらの 2 つのデータより得ることが可能である．この事実は，case control study におけるオッズ比のように，$Y = 1$ に基づき抽出された標本の利用においては（補足データとの併用により）相対リスクが適用可能な統計量となり得ていることを示唆する．

しかし，相対リスクの問題点として，相関のある複数の独立変数が存在する場合にそのまま適用することができないという問題がある．いま，それぞれ二値変数である X_1 と X_2 の 2 つの独立変数が存在し，X_1 の影響に関心があるとする．しかし，この場合は X_2 の値による以下の 2 つの相対リスクが存在する．

[*4] こうしたシミュレーション研究の試みとしては，特に生物統計学で確認できる (Phillips and Elith 2011; Keating et.al 2004)．しかし，SC 推定と TLR 推定の対比はされていない．また，これら 2 つの手法以外にも，例えば Lancaster and Imbens(1996) による GMM 推定，Ward et.al(2009) の EM アルゴリズムによる推定が提唱されているが，前者は GMM の設定が煩雑で，後者は後述するが NMAR にはナイーブに適用できないために割愛する．また，Tang, Little and Raghunathan(2003) の発想に近い研究として，Ramalho and Smith(2013) があるが，この研究を知ったのはごく最近で正直読んでおらず，また GMM を用いているので実装が難しい．なお，管見の限り，日本語で読める統計学の教科書で，これらの問題に言及しているのは高井ら (2016) をのぞいてみあたらない（のは探していないから）．

$$\frac{Pr(Y=1|X_1=1,X_2=0)}{Pr(Y=1|X_1=0,X_2=0)} = \frac{Pr(X_1=1,X_2=0|Y=1)}{Pr(X_1=0,X_2=0|Y=1)}\frac{Pr(X_1=0,X_2=0)}{Pr(X_1=1,X_2=0)}$$

$$\frac{Pr(Y=1|X_1=1,X_2=1)}{Pr(Y=1|X_1=0,X_2=1)} = \frac{Pr(X_1=1,X_2=1|Y=1)}{Pr(X_1=0,X_2=1|Y=1)}\frac{Pr(X_1=0,X_2=1)}{Pr(X_1=1,X_2=1)}$$

X_1 と X_2 が独立であれば 2 つの相対リスクは同じ値を示すが，X_1 と X_2 に相関がある場合は，この 2 つの相対リスクは異なる値を示すため，どちらを選択するのかで結果が異なる．複数の独立変数に相関がある場合，オッズ比では統制変数を導入した調整オッズ比から一意に関心のある独立変数の影響を検討できるが (c.f. 高井ら 2016)，相対リスクに同様のロジックを導入した場合，統制変数の分布による周辺化が必要となり[*5]，また一つの独立変数ではなく，複数の独立変数が存在する場合は，複数回の推定が必要となってしまい煩雑になる．

そこで，ここではログリニアモデルから拡張し，Purge Method(e.g.Clogg 1978) のアイディアを用いて相対リスクの定式化を行う．いま完全データにおける X_1 と X_2 の同時分布 $Pr(X_1, X_2|Y=1)$ と補足データにおける X_1 と X_2 の同時分布 $Pr(X_1, X_2)$ をそれぞれ以下の式で表す（実際には頻度になるけど）．

$$Pr(X_1, X_2|Y=1) = \tau\tau_i^1\tau_j^2\tau_{ij}^{12}$$

$$Pr(X_1, X_2) = \mu\mu_i^1\mu_j^2\mu_{ij}^{12}$$

ログリニアモデルによる表現から，懸念の 2 つの相対リスクはそれぞれ以下のように理解できる．

$$\frac{Pr(Y=1|X_1=1,X_2=0)}{Pr(Y=1|X_1=0,X_2=0)} = \frac{\tau_1^1\tau_{10}^{12}}{\tau_0^1\tau_{00}^{12}}\frac{\mu_0^1\mu_{00}^{12}}{\mu_1^1\mu_{10}^{12}}$$

$$\frac{Pr(Y=1|X_1=1,X_2=1)}{Pr(Y=1|X_1=0,X_2=1)} = \frac{\tau_1^1\tau_{11}^{12}}{\tau_0^1\tau_{01}^{12}}\frac{\mu_0^1\mu_{01}^{12}}{\mu_1^1\mu_{11}^{12}}$$

このログリニアモデルによる表現から，2 つの相対リスクの差異は交互作用項 μ_{ij}^{12} と τ_{ij}^{12} から理解することが可能である．それゆえ，関心が X_1 と X_2 との交互作用項にないのであれば，交互作用項で観察された頻度を調整し，交互作用の効果を “追放” する Purge method(e.g. Clogg 1978) のアイディアをそのまま導入することが可能である[*6]．この Purge Method により調整した場合，X_1 と X_2 は独立になるので，相対リスクは X_2 の条件に依存せず，以下の式で一意に得ることができる．

$$\frac{Pr(Y=1|X_1=1,X_2)}{Pr(Y=1|X_1=0,X_2)} = \frac{\tau_1^1}{\tau_0^1}\frac{\mu_0^1}{\mu_1^1}$$

[*5] 調整オッズ (e.g. 高井ら 2016) の考えを相対リスクに適用した場合を示す．いま，X_1 に関心があり，X_2 を統制として利用することを想定する．つまり，$\frac{Pr(Y=1|X_1=1,X_2)}{Pr(Y=1|X_1=0,X_2)}$ に関心があるとする．この式を展開すれば

$$\frac{Pr(Y=1|X_1=1,X_2)}{Pr(Y=1|X_1=0,X_2)} = \frac{Pr(X_1=1|Y=1,X_2)}{Pr(X_1=0|Y=1,X_2)}\frac{Pr(X_1=0)}{Pr(X_1=1)}$$

となる．ここで例えばロジットモデル $logit(Pr(X=1)) = \beta_0 + \beta_1 X_2$ の利用を考えれば，

$$\frac{Pr(X_1=1|Y=1,X_2)}{Pr(X_1=0|Y=1,X_2)}\frac{Pr(X_1=0)}{Pr(X_1=1)} = exp(\beta_0+\beta_1X_2)\frac{Pr(X_1=0)}{Pr(X_1=1)}$$

となる．それゆえ X_2 の分布に依存して値が変化するため周辺化をする必要がある．

$$\frac{Pr(X_1=0)}{Pr(X_1=1)}\int exp(\beta_0+\beta_1X_2)dF(X_2)$$

[*6] Purge Method は，単純に観察された度数 F_{ij}^{12} を，交互作用項，例えば μ_{ij}^{12} で割るだけの話．

このログリニアモデルで拡張し Purge Method を応用した相対リスクの評価は，ログリニアモデルの推定が可能な統計ソフトで簡単に行うことができる．具体的には，補足データと完全データでそれぞれ関心のある独立変数を対象にした飽和モデルを推定し，得られたパラメータの推定結果を上述の式に当てはめることで簡単に計算できる．また，上の議論では 2 つの独立変数を想定し，片方の独立変数の影響のみに関心を限定してきたが，たとえ複数の独立変数が存在していても，関心に応じた交互作用項の選択を行えば良く，また複数の独立変数に同時に関心があっても，一回のログリニアモデルの推定結果から計算することができる．

ただし，この手法の応用には，ログリニアモデルはいったん飽和モデルを想定する必要がある．というのも，あくまでこの手法は複数の独立変数を独立にすることが目的になっているが，事前にどの変数間で相関があるのかについてはアプリオリには分からないため，誤ったモデル，例えば独立モデル，を使用した場合は，omitted variable bias が発生する危険性があるためである．また，しばしログリニアモデルの応用では対数尤度に基づく種々の (検定) 統計量からより「節約的な」モデル選択が行われているが，相対リスクに適用した場合は，モデル全体での当てはまりというよりは個々のパラメータの値に関心があるため，重回帰分析の t 検定と F 検定の違いのように，対数尤度に基づくモデル選択と個々のパラメータの推定結果による理解は必ずしも一致しない可能性があり，対数尤度に基づくモデル選択の結果を所与とすることは，関心の相違から肯定することが難しい．よって，飽和モデルの適用はパラメータの「推定」ではなく，観察された値をパラメータに「分解」するものであると理解し，得られた相対リスクも観察された結果を記述する統計量であると理解したほうがよい．また，この注意点から派生する限界として，相対リスクの標準誤差は推定できないという点があり[*7]，加えて独立変数はカテゴリカル変数しか許容できない点も挙げられる．

疑似尤度からのアプローチ

Manski（2003）の相対リスクのアイディアをログリニアモデル，および Purge Method で拡張した方法は，標準的な統計ソフトがあれば簡単に計算することができる．しかし，その射程は飽和モデルの利用から記述統計レベルの議論に限定され，また可能な独立変数もカテゴリカル変数のみに限定されてしまうという問題がある．これに対して，近年（なのかしらないけど），回帰モデルを想定した疑似尤度ベースの推定がいくつか提唱されている．ここでは比較的簡便に実装することができる Steinberg and Cardell(1992) による推定方法（SC 推定）と，Tang, Little and Raghunathan(2003) による推定方法（TLR 推定）を取り上げ，モンテカルロシミュレーションの結果からそれぞれの手法の特性について簡単に議論する．なお，以下では簡便のため，回帰モデルはすべて二項ロジット $Pr(Y=1|X,\beta)=\frac{exp(X\beta)}{1+exp(X\beta)}, Pr(Y=0|X,\beta)=\frac{1}{1+exp(X\beta)}$ を想定する．

[*7] Manski(2003) の bound，もしくはブートストラップならいける？

Steinberg and Cardell の方法

Steinberg and Cardell（1992）が提唱した疑似尤度は，基本的には以下の式で定義される二項ロジットの対数尤度関数を分解することで理解できる．

$$\begin{aligned} LL &= \sum_{i=1}^{N} Y_i logPr(Y=1|X_i,\beta) + \sum_{i=1}^{N}(1-Y_i)logPr(Y=0|X_i,\beta) \\ &= \sum_{Y_i=1} logPr(Y=1|X_i,\beta) + \sum_{Y_i=0} logPr(Y=0|X_i,\beta) \end{aligned}$$

$\sum_{Y_i=j}$ は $Y=j$ であるサンプルのみで総和することを意味する．いま，手元に完全データと補足データしかない状況において，上の対数尤度を構成する 2 つの項のうち，前者の $Y_i=1$ であるサンプルに適用される項は完全データより得ることができるが，後者の $Y_i=0$ に適用される項については，2 つのデータから得ることができない．しかし，対数尤度の別表現

$$\begin{aligned} LL &= \sum_{i=1}^{N} Y_i logPr(Y=1|X_i,\beta) + \sum_{i=1}^{N}(1-Y_i)logPr(Y=0|X_i,\beta) \\ &= \sum_{i=1}^{N} logPr(Y=0|X_i,\beta) + \sum_{i=1}^{N} Y_i logPr(Y=1|X_i,\beta) - \sum_{i=1}^{N} Y_i logPr(Y=0|X_i,\beta) \\ &= \sum_{i=1}^{N} logPr(Y=0|X_i,\beta) + \sum_{Y_i=1} logPr(Y=1|X_i,\beta) - \sum_{Y_i=1} logPr(Y=0|X_i,\beta) \end{aligned}$$

より，

$$\sum_{Y_i=0} logPr(Y=0|X_i,\beta) = \sum_{i=1}^{N} logPr(Y=0|X_i,\beta) - \sum_{Y_i=1} logPr(Y=0|X_i,\beta)$$

と理解することができるので，完全データ・補足データより得ることができなかった $Y_i=0$ に適用される尤度関数の項は，いったんすべてのサンプルが「疑似的」に $Y_i=0$ であると想定した対数尤度の値から，実際に $Y_i=1$ であるサンプルに誤って $Y_i=0$ と想定した分を差し引くことで得られると理解できる．それゆえ，実際のデータ上で $Y_i=0$ が観察されていなくても，X を母集団より歪みなく抽出した標本，つまり補足データと，$Y_i=1$ である母集団から X を抽出した標本，つまり完全データがあれば，観察されない $Y_i=0$ に対応する対数尤度の項を再現すること可能である．

これらを踏まえ，Steinberg and Cardell（1992）は，上の「疑似的」な従属変数を生成し得られる疑似対数尤度関数を提唱している．具体的には，サンプリング率を考慮し，$s=Pr(Y=1)$ が既知であるという条件のもと，以下の疑似対数尤度関数 LL_{sc} を提唱した．

$$LL_{sc} = \sum_{R_i=0} logPr(Y=0|X_i,\beta) + s\frac{N_{sup}}{N_{comp}} \sum_{R_i=1} log\frac{Pr(Y=1|X_i,\beta)}{Pr(Y=0|X_i,\beta)}$$

ここで $R_i=0,1$ はそれぞれ補足データと完全データに当該サンプルが含まれていることを示すインデックスで，N_{sup} は補足データのサンプルサイズ，N_{comp} は完全データのサンプルサイズをそれぞれ指す．この疑似対数尤度 LL_{sc} を最大化する $\hat{\beta}$ は，一致推定量であり，かつ漸近的に正規分布に従う[*8]．

[*8] ただし，分散は単純なヘッシアンの逆行列の対角線上で得られるわけではなさそう．

この疑似的な従属変数を想定した SC 推定は，ネガティブな重みづけを許す統計ソフトが利用可能なら，独自で最大化のプログラムを構築することなく推定することができる．具体的には，補足データと，複製した２つの完全データを用意し，実際の従属変数に代わり，補足データには $Y_i = 0$ を，一つ目の完全データには $Y_i = 1$，二つ目の完全データには $Y_i = 0$ を疑似的に外挿し，補足データには 1，$Y_i = 1$ が外挿された完全データには $s\frac{N_{sup}}{N_{sup}}$，$Y_i = 0$ が外挿された完全データには $-s\frac{N_{sup}}{N_{sup}}$ をそれぞれウェイトとして尤度関数に課せば推定することができる[*9]．

Tang, Little and Raghunathan の方法

SC 推定は，スタンダードな対数尤度の適用において観察されない $Y_i = 0$ に相当するパートを，「疑似的」な従属変数の生成という方法から，完全データと補足データを用いて補完した方法として理解できる．これに対して，Tang, Little and Raghunathan（2003）が提唱した疑似尤度[*10]を用いた方法は，基本的には $Y_i = 1$ に相当する対数尤度のパートを操作したものとして理解できる．

良く知られた通り，従属変数に欠測がある場合，その欠測がどのように決まっているのかよって，推定の方法が大きく変わる (e.g. Little and Rubin 2002)．最も煩雑なのは，いわゆる「欠測した値に（も）欠測するか否かが決まっている」という NMAR タイプの欠測で，欠測メカニズムに関する関数を特定することが一致推定量を得るために必要な作業になる．いま，この議論を完全データと補足データをマージしたデータから理解すると，マージデータの従属変数において，完全データの箇所に対応するところは観察されているが，補足データの箇所に対応するところは欠測している状況になる．そして，このマージデータが，隠れた母集団の一つの例である違法薬物を使用している $(Y_i = 1)$ か否か $(Y_i = 0)$ を従属変数とするデータだった場合，違法薬物を使用している個人は，その違法性の暴露によるディスインセンティブと報酬などのインセンティブの比較から，調査協力を行うか否かが決まり，使用していない個人はその属性により回答に協力するか否かが変わらず正確に回答をしてくれる（と期待できる）ので，マージデータでの欠測には，本来違法薬物を使用していない個人に加えて，本当は違法薬物を使用しているが回答を拒否した個人の 2 つのパターンが想定される．それゆえ，こういう状況は，「欠測した値 (= 違法薬物を使用しているか否か) そのものにより欠測するか否かが決まる」典型的な NMAR になっていると理解でき，バイアスのない推定量を得るためには，より正確な関数として欠測メカニズムを特定する必要がある．

しかし，Tang, Little and Raghunathan（2003）は，欠測メカニズムに関わるパラメータと，関心のある $Pr(Y|X;\beta)$ のパラメータ β が分離可能であるという仮定のもとでは，特に欠測メカニズムを特定することなく，完全データのみを対象にした直接尤度の変形から関心のあるパラメータ β を推定することができることを証明している．証明の詳細については割愛するが，基本的なアイディアは以下の通りである．まず，Chain-referral sampling を前提とした場合の完全データは，従属変数 Y に基づいて X が抽出されているので，ストレートな尤度関数の表現は

$$L = \prod_{R_i=1} p(X_i|Y_i;\beta,\alpha)$$

となる．ここで β は Y と X の同時分布のみに関わるパラメータで，回帰モデルを想定した場合は回帰係数に

[*9] しかし，ネガティブなウェイトが可能な統計ソフトなんて見たことがない．

[*10] TLR 推定について日本語で解説した書籍としては高井ら (2016) がある．が，さらっと書きすぎ．

なる．また α は X の分布を規定するパラメータを指し，Y の分布とは無関係であると仮定する．ベイズの公式から，上の尤度関数は

$$\begin{aligned} L &= \prod_{R_i=1} p(X_i|Y_i;\beta,\alpha) = \prod_{R_i=1} \frac{Pr(Y_i|X_i;\beta)f(X_i;\alpha)}{Pr(Y_i)} \\ &= \prod_{R_i=1} \frac{Pr(Y_i|X_i;\beta)f(X_i;\alpha)}{\int Pr(Y_i|X;\beta)f(X;\alpha)dX} \end{aligned}$$

と表現することができ，いま，$f(X;\alpha)$ を経験分布 $F(X)$ で置き換えたうえで，対数を取り冗長な項をのぞけば，以下の疑似対数尤度関数 LL_{TLR} が得られる．この LL_{TLR} を最大化する $\hat{\beta}$ は一致推定量であり，かつ漸近的に正規分布に従うことが証明されている．

$$LL_{TLR} = \sum_{R_i=1} logPr(Y_i|X_i,\beta) - \sum_{R_i=1} log \int Pr(Y_i|X;\beta)dF(X)$$

Tang, Little and Raghunathan（2003）では特定の Y の値のみに基づく抽出を想定していなかったが，いま，補足データと完全データの利用を踏まえると，LL_{TLR} の右辺の第一項・第二項の $Pr(Y_i|X_i,\beta)$ は $Pr(Y=1|X_i,\beta)$ と置き換えられ具体的なロジットの式が導入されて，また，$Pr(Y_i|X_i,\beta)$ の期待値部分は，母集団における X の分布と対応する補足データの経験分布を利用して計算することができる．これらの条件を反映した完全データの標本に基づく疑似尤度から推定を行えば良い[*11]．ただし，管見の限り，この TLR 推定を実装している，もしくは簡便に実装可能な統計ソフトは存在しておらず，実際に推定したいときは，自分でプログラムを構築する必要がある．

モンテカルロシミュレーション

疑似的に従属変数を発生させるのか，それとも従属変数が備わっているパートのみで尤度を構築するのかという違いがあるものの，SC 推定，TLR 推定はどちらも $Y=0$ がかけている標本でいかに推定するのかという問題に，疑似尤度関数の構築からアプローチしたものである．そして，完全データと補足データの併用から，回帰モデルの係数をバイアスなく推定できることを明らかにした．しかし，統計ユーザーの視点に立てば，この 2 つの手法のいずれを，どのような条件のときに用いればよいのか，という問いにも関心がある．特に，疑似対数尤度関数の形をみれば，SC 推定ではそれぞれ補足データと完全データのサンプルサイズ（の比），そして $Y=1$ の周辺分布が既知であるという条件が含まれているが，TLR 推定ではそれらの条件は必要ではないので，これらの条件によっては 2 つの推定の結果が大きく変わる可能性もあると予測できる．そこで，以下では Lancaster and Imbens(1996) の枠組みを用いて，SC 推定と TLR 推定のモンテカルロシミュレーションを行う．

設定

【母集団の設定】

[*11] はず．

回帰モデル:

$$Pr(Y=1|X,\beta)=\frac{exp(\beta_0+\beta_1X_1+\beta_2X_2)}{1+exp(\beta_0+\beta_1X_1+\beta_2X_2)}$$

独立変数の密度関数:

$$X_1\sim N(0,1),X_2\sim N(0,1)$$

【完全データの抽出】

Step1 ：$X_1\sim N(0,1),X_2\sim N(0,1)$ から抽出し，上のロジットの式から $Pr(Y=1)$ を計算して，ベルヌーイ試行で $Y\in(0,1)$ の値を抽出する．

Step2 ：$Y=1$ であるときに，抽出確率 $h=0.5$ で完全データに含める．

Step3 ：既定のサンプルサイズ (N_{comp}) に到達するまで，Step1 ～ 2 を繰り返す．

【補足データの抽出】

$X_1\sim N(0,1),X_2\sim N(0,1)$ から規定のサンプルサイズ (N_{sup}) のデータを抽出．

【条件】

N_{comp} は 3000 にフィックスし，N_{sup} は 200,400,600 の 3 パターンを想定する．

結果

以下の表 1 は，それぞれ $\beta_0=0.5,\beta_1=1,\beta_2=-1$ を設定し，$s=Pr(Y=1)$ が既知であるという条件のもとで[*12]，1000 回の試行に基づいた各推定法の結果を示したものである．“Method” は推定法を示す列で，“SC” と “TLR” はそれぞれ SC 推定と TLR 推定，“IF” は参考のため，補足データに $Y=0$ を外挿して通常の二項ロジットで最尤推定を行なった結果を指している．また “SIZE” の列は完全データのサンプルサイズ，“XX_mean” と “XX_median” は XX に対応する独立変数の回帰係数の平均と中央値を示した列である．

表1: 平均と中央値

METHOD	SIZE	IC_mean	X1_mean	X2_mean	IC_median	X1_median	X2_median
IF	200	-2.798	0.306	-0.303	-2.795	0.304	-0.301
IF	400	-2.108	0.310	-0.311	-2.107	0.311	-0.312
IF	600	-1.703	0.314	-0.315	-1.703	0.315	-0.315
SC	200	0.525	1.072	-1.060	0.509	1.007	-0.990
SC	400	0.515	1.039	-1.045	0.506	1.009	-1.020
SC	600	0.510	1.022	-1.028	0.505	1.007	-1.007
TLR	200	0.704	1.244	-1.194	0.523	1.024	-1.033
TLR	400	0.479	1.045	-1.048	0.525	1.023	-1.019

[*12] 実際には補足データから計算している．

METHOD	SIZE	IC_mean	X1_mean	X2_mean	IC_median	X1_median	X2_median
TLR	600	0.503	1.033	-1.037	0.518	1.025	-1.026

この表 1 より，まず補足データに $Y = 0$ を外挿した場合の結果には，大きなバイアスが生じていることが分かる．SC 推定と TLR 推定の結果を比較すると，どの完全データのサンプルサイズでも，概ね SC 推定のほうが良好な推定値を与えていることが伺えるが，平均値に注目すると，完全データのサンプルサイズが大きくなるにつれて，どちらの推定でも徐々にバイアスが少なくなっていくことが分かる．

また，推定の誤差を検討するため，以下の表 2 に標準誤差 (XX_psd) と四分位範囲 (XX_IR) を示した．ここから完全データのサンプルサイズが大きくなるにつれて，どちらの推定も標準誤差が減少していくことが分かる．ただし，概ね SC 推定のほうが標準誤差が少なく，とくに小さなサンプルサイズのときの TLR 推定の誤差は深刻である．

表2: 標準誤差と四分位範囲

METHOD	SIZE	IC_psd	X1_psd	X2_psd	IC_IR	X1_IR	X2_IR
IF	200	0.027	0.069	0.069	0.036	0.092	0.094
IF	400	0.020	0.052	0.053	0.028	0.072	0.071
IF	600	0.016	0.047	0.044	0.023	0.064	0.058
SC	200	0.104	0.398	0.384	0.111	0.439	0.421
SC	400	0.077	0.264	0.271	0.090	0.336	0.336
SC	600	0.067	0.219	0.214	0.089	0.282	0.283
TLR	200	5.895	4.099	2.585	0.963	0.533	0.499
TLR	400	1.027	0.285	0.283	0.683	0.342	0.371
TLR	600	0.698	0.226	0.230	0.570	0.290	0.307

以上の簡単な確認から，ポイントでみたときと推定の誤差の程度で判断したとき，どちらの場合も概ね SC 推定のほうが良い推定方法であると判断できよう．しかし，この結果は，SC 推定において $s = Pr(Y = 1)$ が正確に分かっているという条件のもとで得られる結果ある．以下の図は，$s = Pr(Y = 1)$ とのかい離程度によりどの程度のバイアスがかかるのかを示した図である．横軸は真値との差で縦軸はそのかい離した値で 1000 回志向したときの係数の平均を示している（完全データのサンプルサイズは 600）．

図 2 より明らかな通り，SC 推定において誤って $s = Pr(Y = 1)$ を設定することは大きなバイアスを生み危険性がある．特に，そのかい離が大きければ大きいほど，バイアスの程度は非線形的に増大していく傾向にあると理解できる．また，バイアスの方向性も，真の $s = Pr(Y = 1)$ を境にして正負が逆転していることがうかがえる．それゆえ，バイアスの方向性が一定であれば，$s = Pr(Y = 1)$ の値を変化させたときの係数差を確認することで，現在設定している $s = Pr(Y = 1)$ と真値の大小関係を把握することができ，翻っては真値の特定，そしてバイアスの少ない係数を推定することがで可能だが，バイアスの方向性が一定でない以上は，こ

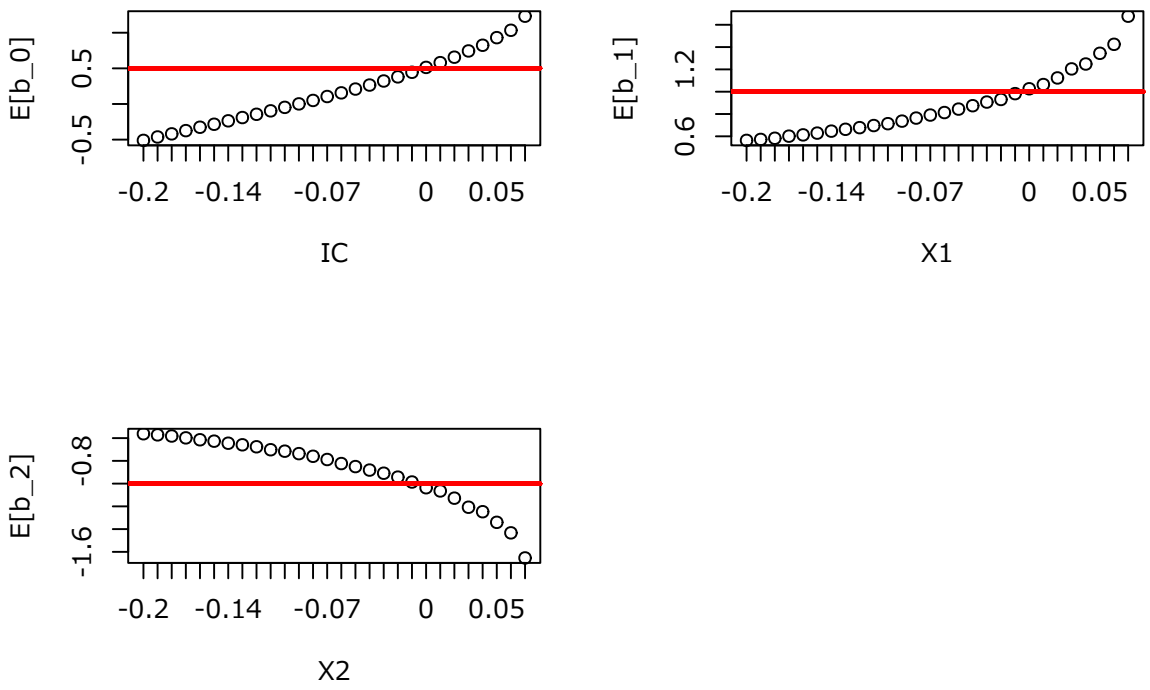

図2　s のかい離による影響　赤線は真値

の手法は使えない[*13].

以上の簡単なシミュレーション結果から，少なくとも SC 推定と TLR 推定のどちらを使うのかという判断においては，$s = Pr(Y = 1)$ が既知か否かが決定的な基準になっていると理解できる．まず，$s = Pr(Y = 1)$ が既知であれば，サンプルサイズを問わず SC 推定を利用したほうが良い．もし，$s = Pr(Y = 1)$ が未知であれば，十分な完全データのサンプルサイズがある場合には，TLR 推定を代用的に用いることができるが，サンプルサイズが十分でない場合は，得られた結果の解釈については慎重に行った方が良いと思われる．

応用例：赤線への参入

概要

以上の手法を応用した分析として，ここでは戦後の赤線（特殊飲食店）への参入に適用した結果を示す．まず，簡単に時代背景を述べておくと，日本では 1957 年に売春防止法が施行されるまで，性病の拡大予防を一貫した目的とする国家管理体制のもと，売春は実質的に合法な職業であった．実際に，1955 年 4 月 30 日現在で，

[*13] 実は Lancaster and Imbens(1996) の GMM 推定は，この s が分からなくても解ける方法になっている．ただし，バイアスは少ないが，標準誤差が桁違いに大きいことがシミュレーションからわかっている．というか GMM を実装するのは本当にきつい．

特殊飲食店（赤線）では 80205 人が，三業地・駐留軍基地・自衛隊付近のそれを含めれば 129019 人が従事していた (労働省婦人少年局 1955)．しかし，当時の赤線を扱った研究はかなり蓄積されているものの (e.g 藤野 XXXX, 平井 XXXX)，合法的な職業である売春に，どのような女性が参入したのかという問題はいまだ経験的には十分には議論されていない．

そこで，以下では，「特殊飲食店女子組合員調査」[*14]の結果を用いて，この問題にアプローチしていく．この調査は，全国性病予防自治会と呼ばれる，特殊飲食店業界関係者による業界調査で，当時盛り上がっていた廃娼運動の機運に対して，経験的に実態を示すことを目的とした調査である．特殊飲食店に従事した全従業員を対象にした行われ，そのうち約 300 の回答結果に今日アクセスすることができる．このデータを完全データとし，また，補足データとして，JGSS2000-2003 の累積データのうち，特殊飲食店女子組合員調査に含まれるサンプルと同じ出生コーホート（1917 年から 1937 年）に限定したサンプルを用いる．

独立変数として，人的資本をプロキシとして学歴（初等教育以下 =0，以上 =1），教育達成の要因，もしくは態度形成の要因として男兄弟の存在（兄もしくは弟がいない =0，いる =1），困窮度の変数として 15 歳時点での家族構成 (父母がいる =0，いずれか，もしくはいずれもいない =1) を用いる．また，$s = Pr(Y = 1)$ については，従業員数 80205 人を，出生コーホートが 1917 年から 1937 年である女性人口 15373325 人で除したものを利用する（$s = 0.0052$）．

結果

まず初めに記述統計として，ログリニアモデルと Purge Method を用いた，相対リスクの結果をしめす．

表3: 相対リスク

variable	Relative Risk
EDU	0.529
BRO_TRUE	0.592
PAR_notFULL	5.532

ここから，学歴が高いこと，男兄弟がいることで相対的に売春への参入が抑止され，対して出身家庭に父母が揃っていない場合に，逆に促進されている傾向がうかがえる．また，ブートストラップを使用した SC 推定と TLR 推定による推定結果を以下に示す（試行回数 =500，CI_x は x パーセンタイル）．

表4: SC

	coef	s.e	CI.05	CI.95
IC	-4.390	0.101	-4.560	-4.234
EDU	-1.233	0.157	-1.490	-0.992

[*14] http://hdl.handle.net/10577/178

	coef	s.e	CI.05	CI.95
BRO_TRUE	-1.005	0.130	-1.212	-0.787
PAR_notFULL	0.979	0.124	0.777	1.193

表5: TLR

	coef	s.e	CI.05	CI.95
IC	-15.197	0.609	-15.986	-14.373
EDU	-1.229	0.163	-1.482	-0.967
BRO_TRUE	-0.995	0.118	-1.194	-0.812
PAR_notFULL	0.969	0.125	0.759	1.173

切片を除いて，ほぼ同じ係数を SC 推定，TLR 推定ともに導いており，得られる知見も相対リスクで確認できた知見と同様である．

議論というかメモ

- 別に $X, Y = 1$ だけでのデータでも，X をしっかり母集団からとっていれば問題なく行ける．ただ，サンプルサイズと母集団での s の情報によって，推定の場合やり方が少し変わる．
- X そのものの二次分析というか，X を新たな調査の設計枠の基準に使う（＝独立変数を共通にして，$X, Y = 1$ の抽出をする）という方法を提唱できる．

Appendix1 SC と TLR 推定の R コード

SC と TLR 推定用の R コードを以下に示す．

```
pml<-function(formula,data_comp,data_sup,method="TLR",
              weight_comp=NULL,weight_sup=NULL,
              share=NULL,tol=1e-6,boot=FALSE,bootnum=1000){

  # formula：~ x1 + x2 の形でモデルを設定
  # data_comp：完全データの指定
  # data_sup：補足データの指定
  # method："SC" で SC 推定，"TLR" で TLR 推定を指定
  # weight_comp：完全データの指定 stratification weight（オプション）
  # weight_sup：補足データの Post stratification weight（オプション）
```

```
 # share：母集団における s=Pr(y=1) の指定（SC 選択時のみ必須）
 # tol：収束条件（オプション）
 # boot：ブートストラップを行うか（デフォルトは FALSE. FALSE の場合ヘッシアンから（未調整））
 # bootnum：ブートストラップの試行回数


options(warn = 0)

DC <- model.matrix(formula,data_comp)
ds <- model.matrix(formula,data_sup)

if(!is.null(weight_comp)){
  weight_comp<-weight_comp[as.numeric(dimnames(DC)[[1]])]
  DC<-DC*weight_comp
}

if(!is.null(weight_sup)){
  weight_sup<-weight_sup[as.numeric(dimnames(ds)[[1]])]
  ds<-ds*weight_sup
}

if(!method == "TLR"){

  weight <- share*nrow(ds)/nrow(DC)

  LL <- function(beta) {
    sum(log(1+exp(ds%*%beta)))-
      weight*sum(dc%*%beta)
  }

}else{

  N<-nrow(DC)

  LL <- function(beta) {
    p1<-sum(dc%*%beta-
                log(1+exp(dc%*%beta)))

    p2<-log(mean(exp(ds%*%beta)/
```

```
                    (1+exp(ds%*%beta))))

    -p1+p2*N
  }
}

if(boot == TRUE){

  HAKO<-NULL
  j<-0
  n <- nrow(DC)

  for(j in 1:bootnum){
    dc<-DC[sample(1:n,n,replace =T),]
    kekka<-nlm(LL, p = rep(0, ncol(ds)),hessian = T,gradtol = tol)
    HAKO<-rbind(HAKO,kekka$estimate)
  }

  options(warn = 0)

  coef <- apply(HAKO,2,mean)
  s.e <- apply(HAKO,2,function(x)sd(x)*sqrt(length(x)-1)/sqrt(length(x)))
  CI.05 <- apply(HAKO,2,function(x){quantile(x,probs = 0.05)})
  CI.95 <- apply(HAKO,2,function(x){quantile(x,probs = 0.95)})
  p.value <- round(sapply(abs(coef/s.e),function(x){2*(1-pnorm(x))}),4)

  result<-cbind(coef,s.e,CI.05,CI.95,p.value)
  dimnames(result)[[1]]<-dimnames(ds)[[2]]

}else{

  options(warn = 1)

  DC->dc

  kekka<-nlm(LL, p = rep(0, ncol(ds)),hessian = T,gradtol = tol)

  coef <- kekka$estimate
  s.e <- sqrt(diag(solve(kekka$hessian)))
```

```
   p.value <- round(sapply(abs(coef/s.e),function(x){2*(1-pnorm(x))}),4)

   result<-cbind(coef,s.e,p.value)
   dimnames(result)[[1]]<-dimnames(dc)[[2]]

    }

 result

}
```