行動モデルの応用：サンプル数が小さい時

名古屋大学　山本俊行

ビッグデータの時代にサンプル数が小さいとは？

- 個人間の異質性を突き詰めていくと
  個人毎のモデル推定

- 現時点で需要の小さい選択肢こそ
  需要予測が求められる

離散選択モデルにおける個人間異質性の表現

- 定数項を社会経済特性の関数にする
  - 社会経済特性ダミー（免許保有ダミー等）
- 交通サービス水準のパラメータを個別化
  - 社会経済特性の関数
  - 確率分布を仮定（連続／離散分布）
- モデル全体を個別化
  - 社会経済特性によるセグメント別モデル
  - 意思決定者毎のモデル

個人別モデル

これまで

- 主にマーケティング分野等で用いられてきた
- 交通行動分析では余り用いられてこなかった
  - PT調査では同一個人のトリップ数は数回

近年

- プローブパーソン調査では長期観測により同一個人の繰り返し選択行動が観測可能
- SP調査ではコンピュータ画面で繰り返し選択

需要の小さい選択肢

- タクシーやカーシェアリング，相乗り等の選択肢は鉄道や自動車（自分で運転）等の選択肢に比べて観測数が少ない

- 通常の対応
  - 選択肢別抽出によりサンプル数を増やす
  - 選択肢から除く

名古屋でのカーシェアリングへの加入による影響の調査結果(2005年)

平均保有台数

	会員	非会員	差
1年前	0.6	1.0	0.5**
現在	0.3	1.1	0.8**
変化	−0.3**	+0.0	

1年間の更新行動

	会員	非会員
1台増車	**0**	6
変化なし	19	170
1台減車	8	5
計	27	181

- 非会員の保有台数は名古屋市平均（1.1台）と一致
- 入会する人は元々保有台数が少ない
- 入会によって保有台数は減少する
- 入会して変化のない会員のうち7世帯は購入を見送っている

名古屋市名東区内の交通手段選択行動のモデリング

2011年中京ＰＴデータ

選択肢	サンプル数
鉄道	120
バス	97
タクシー	11
自家用車	1622
自転車	684
徒歩	1954
合計	4488

多項ロジットモデル推定結果

変数名	推定結果	t値
鉄道 定数項	0.466	2.50
バス 定数項	-0.258	-1.01
タクシー 定数項	0.170	0.45
自家用車 定数項	0.533	5.93
自転車 定数項	-0.255	-2.95
所要時間	-0.064	-16.85
費用	-0.003	-4.86
待ち時間	-0.236	-11.56
女性ダミー	-0.089	-1.05
年少者ダミー	-1.086	-12.88
老年者ダミー	-0.472	-1.85
主婦無職ダミー	0.331	3.99
補正済み尤度比		0.397

最尤推定法の特徴

- 一致性（consistency）：サンプル数を大きくしていけば推定値が真値に近づく
- 漸近的有効性（asymptotic efficiency）：サンプル数が十分大きければ推定値の分散は他のどの推定法より小さくなる
- 漸近的正規性（asymptotic normality）：サンプル数が大きくなれば推定値の分散は正規分布に従う

小サンプル時の問題

- 最尤推定法の望ましい性質はサンプル数が大きい時しか保証されない
- 特定の選択肢を選択するサンプルが少ない時，説明変数の組み合わせによってパラメータが発散し推定できないことも多い（separation）
- 二項ロジットモデルでは小サンプル時にパラメータ推定値がバイアスを持つことが示されてきた

パラメータが発散する時の解釈
（Frischknecht et al., 2014）

1. 確率的選択行動の仮定は正しいが，サンプル数が少ないために上手く推定できない

2. 辞書編纂型意思決定等の確定的な選択行動の証拠であり，確率的選択行動の仮定が間違っている

小サンプル時のパラメータのバイアス

- 医学分野等，小サンプルで二項ロジットモデルを推定し，オッズ比を算出したい場合に問題視されてきた

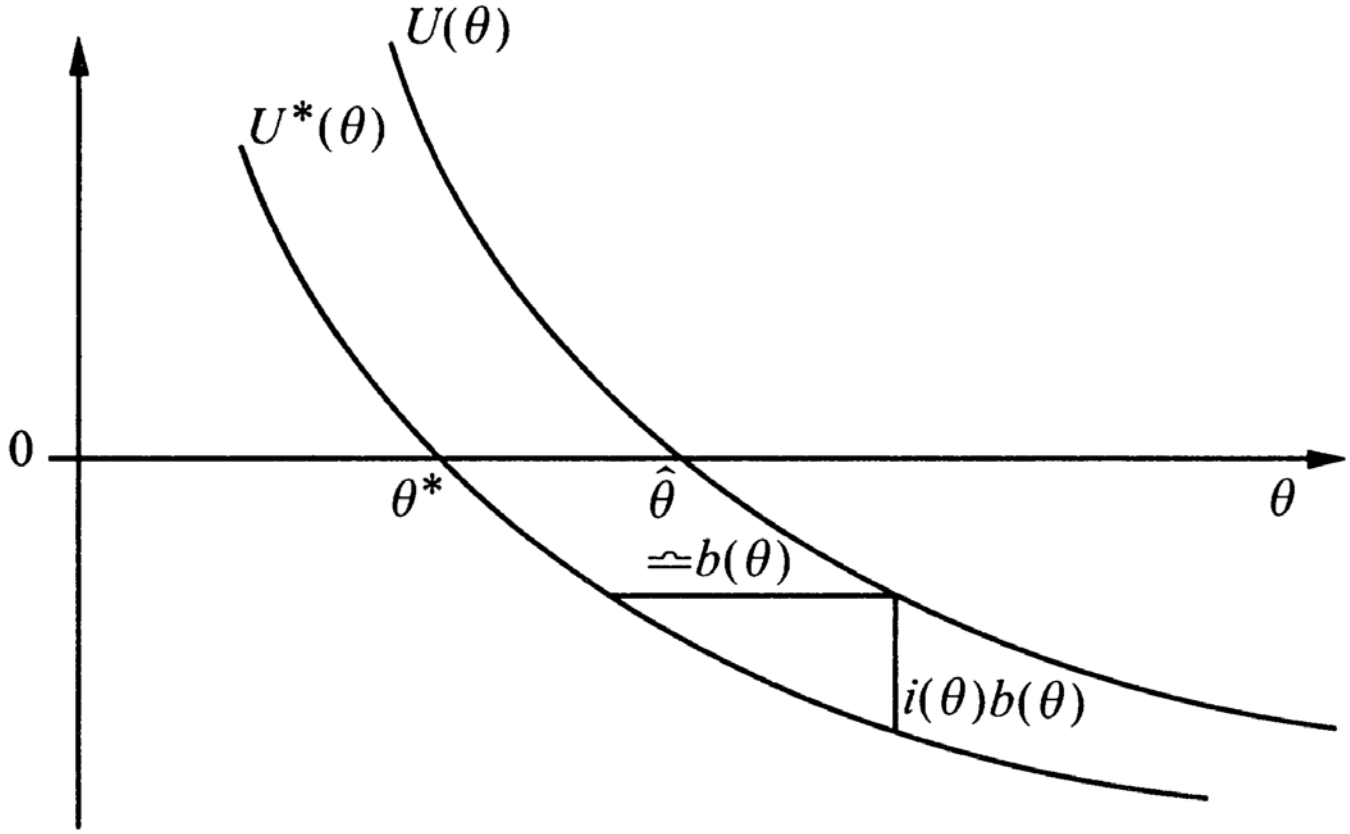
- 選択肢数が多かったり説明変数が多かったりするとパラメータ数に対するサンプル数が相対的に小さくなりバイアスが大きくなる（Bull et al., 2002）
  - 交通行動分野ではありがち？

バイアスの原因（Firth, 1993）

- スコア関数（対数尤度の一次微分$U(\theta)$）にバイアスがない　$E\{U(\theta)\} = 0$
- スコア関数がパラメータに対して非線形 $U''(\theta) \neq 0$

- *最尤推定ではスコア関数が0となるθを探索するが，スコア関数が非線形の時，サンプル平均が母集団平均と一致しない？*

Penalized maximum likelihood estimation (Firth, 1993)



$$U^*(\theta) = U(\theta) - i(\theta)b(\theta)$$

Penalized maximum likelihood estimation (Firth, 1993)

ペナルティ付き尤度　$L(\beta)^* = L(\beta)|I(\beta)|^{1/2}$,　　$I(\beta)$: 情報行列

ペナルティ付きスコア関数

$$U(\beta_r)^* \equiv U(\beta_r) + 1/2 \,\mathrm{trace}\,[I(\beta)^{-1}\{\partial I(\beta)/\partial \beta_r\}] = 0 \quad (r = 1, \ldots, k)$$

上式では β を使って補正する必要があるので，実際の推定は
以下の繰り返し計算となる

$$\beta^{(s+1)} = \beta^{(s)} + I^{-1}(\beta^{(s)})U(\beta^{(s)})^*$$

二項ロジットモデルについてはRのパッケージもあり（logistf）

Firth (1993)以降

- 多項ロジットモデルへの拡張（Bull et al., 2002）
- ベイズ推定との類似性の指摘（Gilbride et al., 2008; Evgeniou et al., 2007）
- 通常の情報行列を用いるより望ましいパラメータ信頼区間の推定法（Heinze and Schemper, 2002; Bull et al., 2007）

- 交通行動分析の分野でよく用いられる，より複雑なモデルでも有効なのか？

参考文献

- Bull, S.B., Mak, C., Greenwood, C.M.T. (2002): A modified score function estimator for multinomial logistic regression in small samples. Computational Statistics & Data Analysis 39, 57-74.
- Bull, S.B., Lewinger, J.P., Lee, S.S.F. (2007): Con□dence intervals for multinomial logistic regression in sparse data. Statistcs in Medicine 26, 903–918.
- Evgeniou, T., Pontil, M., Toubia, O. (2007): A convex optimization approach to modeling consumer heterogeneity in conjoint estimation. Marketing Science 26, 805–818.
- Firth, D. (1993): Bias reduction of maximum likelihood estimates. Biometrika 80, 27-38.
- Frischknecht, B.D., Eckert, C., Geweke, J., Louviere, J.J. (2014): A simple method for estimating preference parameters for individuals. International Journal of Research in Marketing 31, 35-48.
- Gilbride, T. J., Lenk, P. J., Brazell, J.D. (2008): Market share constraints and the loss function in choice-based conjoint analysis. Marketing Science 27, 995–1011.
- Heinze, G., Schemper, M. (2002): A solution to the problem of separation in logistic regression. Statistics in Medicine 21, 2409-2419.