Test of independence for categories with nonignorable nonresponse

狩野 裕，高井啓二
大阪大学 大学院基礎工学研究科

はじめに　カテゴリカル変数における欠測値問題は多くの文献で議論されており，研究を概観できる論文や成書も出版されている (e.g., Little and Rubun, 1987, 2002; Molenberghs, et al., 1999)．それらの文献では，識別性・推定可能性に加えて，セル確率や回帰パラメータをバイアスなく推定すること，モデル適合の吟味，不適解の処理等が議論されている．Ma, Geng and Li (2003) は，2×2 分割表において，観測変数 (Y_1, Y_2) と欠測指標 (M_1, M_2) の関係を分解可能モデルで記述し，識別可能性等を網羅的に議論している．本発表では，無視できない欠測を含む 2×2 分割表データにおいて，Y_1 と Y_2 の独立性の検定問題を考える．

問題　図 1 に 2×2 分割表データにおける無視できない欠測を表す分解可能モデル (グラフ) を示す．Ma, et al. (2003) は，無視できない欠測を含むモデルが識別可能であるための必要条件として Y_1 と Y_2 が独立でないことを挙げている (図 1 の (d) 以外)．それは，Y_i と M_i $(i = 1, 2)$ は単独では識別可能でないが，Y_1 と Y_2 とが繋がることによって，識別できる可能性が出てくると解釈できる．

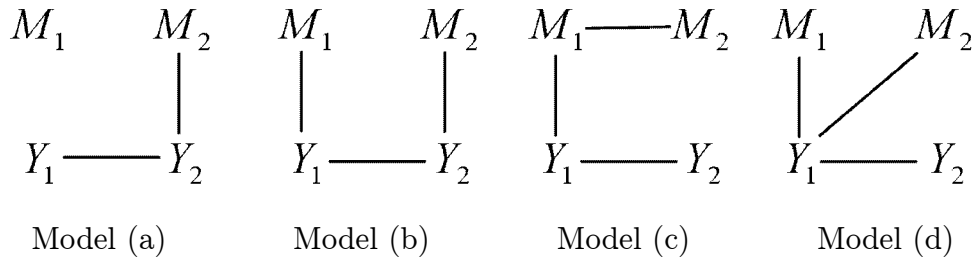

図 1: 無視可能でない欠測メカニズムを表す分解可能グラフ

独立性の検定はカテゴリカルデータの分析としてきわめて基本的なものである．しかし，独立性の仮定は Ma, et al. (2003) の条件に反することから，対数線形モデルによる尤度比検定をナイーブに適用することができない．そこで，識別性を保証するための適当な条件をおいた上で，無視できない欠測がある場合のモデル分析と完全データのみによる分析についてその適用可能性と有用性を比較検討する．

解析　取り扱うデータは表 1(左) の形をしている．supplemental margine には Y_1 か Y_2 のどちらか，もしくは両者共欠測しているデータが与えられている．対応するセル確率は表 1(右) に与えられている．$n_{00,ij}$, $n_{10,+j}$, $n_{01,i+}$, $n_{11,++}$ は 9 つのカテゴリーからなる多項分布にしたがうと仮定する．このときの尤度を以下に示す．

$$L(\boldsymbol{\pi}|\boldsymbol{n}) \propto \prod_{i,j=1}^{2} \pi_{00,ij}^{n_{00,ij}} \times \prod_{i=1}^{2} \pi_{01,i+}^{n_{01,i+}} \times \prod_{j=1}^{2} \pi_{01,+j}^{n_{10,+j}} \times \pi_{11,++}^{n_{11,++}} \tag{1}$$

		$M_2=0$		$M_2=1$
		$Y_2=1$	$Y_2=2$	$Y_2=?$
$M_1=0$	$Y_1=1$	$n_{00,11}$	$n_{00,12}$	$n_{01,1+}$
	$Y_1=2$	$n_{00,21}$	$n_{00,22}$	$n_{01,2+}$
$M_1=1$	$Y_1=?$	$n_{10,+1}$	$n_{10,+2}$	$n_{11,++}$

		$M_2=0$		$M_2=1$
		$Y_2=1$	$Y_2=2$	$Y_2=?$
$M_1=0$	$Y_1=1$	$\pi_{00,11}$	$\pi_{00,12}$	$\pi_{01,1+}$
	$Y_1=2$	$\pi_{00,21}$	$n_{00,22}$	$\pi_{01,2+}$
$M_1=1$	$Y_1=?$	$\pi_{10,+1}$	$\pi_{10,+2}$	$\pi_{11,++}$

表 1: Cell frequencies and probabilities with supplemental margins

図 1 での各モデルはセル確率間に (異なった) 構造を導く．ここでは，Y_1 と Y_2 が独立のときに問題が生じるモデルである (a)-(c) を扱う．

独立性の検定は

$$H_0 : P(Y_1=1,Y_2=1)P(Y_1=2,Y_2=2)-P(Y_1=1,Y_2=2)P(Y_1=2,Y_2=1)=0$$

versus

$$H_1 : P(Y_1=1,Y_2=1)P(Y_1=2,Y_2=2)-P(Y_1=1,Y_2=2)P(Y_1=2,Y_2=1)\neq 0$$

である．一般に欠測によってセル確率にバイアスが生じ $P(Y_1=i,Y_2=j)\not\propto \pi_{00,ij}$ である．しかし，モデル (a)-(c) の下では，上記の独立性の仮説は完全データのみを用いた独立性の検定

$$H_0 : \pi_{00,11}\pi_{00,22}-\pi_{00,21}\pi_{00,12}=0 \quad \text{versus} \quad H_1 : \pi_{00,11}\pi_{00,22}-\pi_{00,21}\pi_{00,12}\neq 0 \tag{2}$$

と同等であることが示される．したがって，完全データにもとづく通常の尤度比検定

$$G_{(1)}=-2\sum_{i,j=1}^{2} n_{00,ij}\log\frac{\hat{n}_{00,ij}}{n_{00,ij}} \quad \text{with} \quad \hat{n}_{00,ij}=\frac{n_{00,i\bullet}n_{00,\bullet j}}{n_{00,\bullet\bullet}} \tag{3}$$

を適用することができる．一方，全データを用いた尤度比検定統計量は

$$G_{(2)}=-2\log\frac{\max_{H_0} L(\boldsymbol{\pi}|\boldsymbol{n})}{\max_{H_1} L(\boldsymbol{\pi}|\boldsymbol{n})} \tag{4}$$

である．モデルの識別性を保証するためにパラメータ間に簡単な等式制約をおく．

$G_{(1)}$ と $G_{(2)}$ のパフォーマンスを理論的にそして数値実験によって比較する．詳細は当日示す．

参考文献

[1] Kano, Y. and Takai, K. (2005). Test of independence for categorical variables with nonignorable nonresponse. (Technical Report).

[2] Little, R. J. A. and Rubin, D. B. (1987, 2002). *Statistical Analysis with Missing Data (2nd edition)*. New York, Wiley.

[3] Ma, W.-Q., Geng, Z. and Li, X.-T. (2003). Identification of nonresponse mechanisms for two-way contingency tables. *Behaviormetrika*, **30**, 125-144.

[4] Molenberghs, G., Goetghebeur, E., Lipsitz, S. R. and Kenward, M. G. (1999). Nonrandom missingness in categorical data: strengths and limitations. *The American Statistician*, **53**, 110-118.