《研究ノート》

労働科学データへの一般化線形モデルの適用とリンク関数選択問題

奥　　喜　正

1．はじめに

一般化線形モデル（Generalized Linear Model; GLIM）は，Nelder & Wedderburn (1972)により，正規線形モデルの枠組みを緩和して，より広い「モデルの族」に対して統一的に線形推測が可能になるように拡張されたものである。その初期において，GLIMは指数分布族に従うデータに対してのみ適応可能とみなされてきた。その後，Wedderburn (1974)による擬似尤度（Quasi-liklihood）の導入により，観測変量Yの期待値E(Y) = uと分散Var(Y)との間に関数関係，すなわち

$$\mathrm{Var}(Y) = a(\phi)V(u)$$

(ここで，V(u)を分散関数，ϕをちらばり母数，mを既知の定数としてa(ϕ) = ϕ/mである)が成立するならば，同様な線形推測が可能であることが判明した。GLIMの発展は，一般化に伴う理論上・計算上の複雑化よりも，その枠組みによりもたらされる一般性のほうが大きかったことによるといえよう。

ところで，GLIMの問題点はPregibon(1984)や椿 (1988) によって数項目，指摘されている。Pregibonは，非線形模型の取扱い，誤差分布を指数分布族から拡張すること，散らばり母数の取扱いなどを記している。その他，分散関数，リンク関数等を誤って想定した場合の推定への影響評価の問題などがある。

本稿の目的は，労働科学データを用いて，リンク関数の選択が要因解析の結果に与える影響を検討することにある。GLIMにおけるリンク関数に関してはMallic & Gelfand (1994) によって行われた最近の研究がある。元来，GLIMにおけるリンク関数の意味あいは，線形モデルに適合するためにデータ変換をすることなく(Box & Cox, 1964)，観測変量の期待値をリンク関数で変換して対処するということにある(Pregibon, 1984)。そこで，本稿では，最初にGLIMの基本的枠組みを見直す。続いて，佐伯・野尻ら (1994) により実施された調査研究により集計された大規模なアンケート調査データを用いて，リンク関数の選択が要因解析に与える影響について，一つの例証を与える。すなわち，選択されたリンク関数の違いにより，データの解析結果として有意な説明変数の組合せが異なることを示し，要因解析に及ぼす程度，そして実際のデータ解析結果に対する解釈への影響を考察する。

2．一般化線形モデルと分析方法

2．1　一般化線形モデル

最初に，Nelder & Wedderburn (1972) によるGLIMの定義を与える。確率変数Yの従う確率分布を指数型分布族であるとして，分布型を完全に規定するもので，指数型一般化線形モデル（EGLIM；Exponential Family Generalized Linear Models）と呼ばれることもある。GLIMは3個の成分から構成される。

(A)　ランダム成分

確率変数Yは指数型確率分布族に従う。すなわち，

$$Y \sim f(y;\theta,\phi) =$$

$$\exp\{\left(y\theta - b(\theta)\right)/a(\phi) + c(y, \phi)\}$$

ここで，$a(\phi) = \phi/m$ (m は既知の定数)，ϕ を散らばり母数 (dispersion parameter) という。また，

$$\mu = E(Y) = b'(\theta)$$
$$Var(Y) = a(\phi)b''(\theta)$$

ここで，$b''(\theta)$ を μ の関数として $b''(\theta) = V(\mu)$ と書くとき $V(\mu)$ を分散関数という。本稿で扱う通常の二項データ y の場合の分散関数は $nY \sim B(n, p)$ とすると次のようになる。

$$\mu = E(Y) = p$$
$$\therefore \quad Var(Y) = p(1-p) = \mu(1-\mu) = V(\mu)$$

(B) 系統成分

説明変数 $x_1, \cdots, x_p$ を用いて系統成分を次のように構成する。

$$\eta = \sum_{j=1}^{p} \beta_j x_j$$

を構成する。

(C) ランダム成分と系統成分のリンク

$$\eta = g(\mu)$$

ここで，$g(\cdot)$ をリンク関数といい，Y の期待値 μ の単調な微分可能関数である。

EGLIM には，古典的線形モデル，社会学や経営学になじみの深いロジットモデル，対数線形モデルが属することは言うまでもない。

さて，Wedderburn (1974) により擬似尤度の概念が導入され，GLIM は EGLIM からさらにモデルの拡張がなされた。擬似尤度 ql とは，平均 μ，分散 $\phi V(\mu)$ である確率変数 Y に対して，

$$\partial\left(ql(\mu, \phi; y)\right)/\partial\mu = (y - \mu)/(\phi V(\mu))$$

を満たすもの，として定義される。つまり，確率変数の 2 次モーメントまでの規定による推定法である。そして，GLIM は，上記 (A) から次の条件 (A)′，

(A)′

$E(Y) = \mu$ のとき，$Var(Y) = a(\phi)V(\mu)$ (ここで，$V(u)$ を既知の関数，m を既知の定数として $a(\phi) = \phi/m$ とする)

を満たすモデルというように定義が拡張される。

こうして確率分布を完全には規定しないという形式で，GLIM のモデルの拡張がなされたことになる。この拡張された定義により，二項データに対する分散関数が

$$V(\mu) = \phi\mu(1 - \mu) \quad (\phi > 1)$$

であるようなモデル，すなわち Over-dispersion (Cox, 1983) に対処するモデルが GLIM に属することになる。Over-dispersion に対処する第一のモデルは，擬似尤度によるものである。現実のデータ解析では，Over-dispersion はブロック効果により生ずるよく見かける現象である。

2.2 パラメータの推定

パラメータベクトル $\boldsymbol{\beta}$ の推定は最尤法により行われる。対数尤度関数を $l(\boldsymbol{\beta}; \boldsymbol{y})$ とすると，$\boldsymbol{\beta}$ の推定は，尤度方程式

$$\partial\left(l(\boldsymbol{\beta}; \boldsymbol{y})\right)/\partial\beta_j = 0 \quad (j=1, \cdots, p) \quad (2\text{-}2\text{-}1)$$

という，非線形方程式を反復的に解くことになる。第 m 近似解 $\hat{\boldsymbol{\beta}}^{(m)}$ が求まっているとすると，(2-2-1) の左辺を $\hat{\boldsymbol{\beta}}^{(m)}$ のまわりでテーラー展開すると

$$\boldsymbol{0} = \partial\{l(\boldsymbol{\beta}; \boldsymbol{y})\}/\partial\boldsymbol{\beta} = \partial\{l(\hat{\boldsymbol{\beta}}^{(m)}; \boldsymbol{y})\}/\partial\boldsymbol{\beta}$$
$$+ [\sum_{k=1}^{p} \partial^2 l/\partial\beta_j\partial\beta_k \left(l(\hat{\boldsymbol{\beta}}^{(m)}; \boldsymbol{y})\right)(\beta_k - \hat{\beta}_k^{(m)})]$$
$$+ [o(|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(m)}|)] \quad (\boldsymbol{\beta} \to \hat{\boldsymbol{\beta}}^{(m)}) \quad (2\text{-}2\text{-}2)$$

となる。ここで，(2-2-2) の右辺の第二項，第三項は $p \times 1$ 型の縦ベクトルである。右辺の第一項を $\boldsymbol{u}^{(m)}$，第二項の前部を H (ヘシアン行列) とおくと (2-2-2) は

$$\boldsymbol{0} = \boldsymbol{u}^{(m)} + H^{(m)}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(m)})$$

スコア法を用いて，J を情報行列とすると

$$\boldsymbol{0} = \boldsymbol{u}^{(m)} - J^{(m)}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(m)})$$
$$\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(m)} = J^{(m)-1}\boldsymbol{u}^{(m)}$$

$$\therefore \quad \hat{\boldsymbol{\beta}}^{(m+1)} = \hat{\boldsymbol{\beta}}^{(m)} + \mathrm{J}^{(m)-1}\,\boldsymbol{u}^{(m)} \qquad (2\text{-}2\text{-}3)$$

と変形され，反復公式 (2-2-3) が得られる。反復公式 (2-2-3) はさらに $\boldsymbol{\delta} = \mathrm{J}^{(m)-1}\,\boldsymbol{u}^{(m)}$ とおくと

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{(m+1)} &= \hat{\boldsymbol{\beta}}^{(m)} + \boldsymbol{\delta} \\ &= (\mathrm{X'W}^{(m)}\mathrm{X})^{-1}\mathrm{X'W}^{(m)}\boldsymbol{z}^{(m)} \qquad (2\text{-}2\text{-}4)\end{aligned}$$

のように表されることを示すことができる。ここで，X はデータ行列，W は重み行列，z は調整従属変数ベクトルである。詳しい解説については Dobson (1990) を参照されたい。(2-2-4) の右辺を一見すれば，それが W を重みとする線形回帰になっていることがわかる。すなわち，ベクトル $\boldsymbol{\beta}$ の最尤推定値が重み付き最小二乗法の反復 (Iterative Weighted Least Squares Method) により求められることが GLIM アルゴリズムの特徴である。

ところで，最尤推定量 $\hat{\boldsymbol{\beta}}$ は，漸近的に不偏である。すなわち

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + O(\mathrm{n}^{-1}) \qquad (\mathrm{n} \to \infty)$$

である。また，次の関係が成立する。

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\mathrm{X'WX})^{-1}\{1 + O(\mathrm{n}^{-1})\} \qquad (\mathrm{n} \to \infty)$$

さらに，最尤推定量 $\hat{\boldsymbol{\beta}}$ は，漸近的に多変量正規分布 $\mathrm{N}(\boldsymbol{\beta}, (\mathrm{X'WX})^{-1})$ に従う。これより，回帰係数に関する仮説 $\mathrm{H}_0 : \beta_j = 0$ $(\mathrm{j} = 1, \cdots, \mathrm{p})$ に対する Wald 検定が得られることになる。なお，Wald 検定統計量，尤度比検定統計量，スコア検定統計量は，漸近的には等しい確率分布に従う (Buse, 1982)。

2．3　モデル適合度の検討

GLIM のデータに対するモデル適合度の検定においては対数尤度比基準が基本となる。いま，フルモデルのパラメータベクトル $\boldsymbol{\beta}_{max}$ に対する最尤推定量ベクトルを $\boldsymbol{b}_{max}$，対数尤度関数を $l(\boldsymbol{b}_{max}; \boldsymbol{y})$ とする。また，関心のあるモデルのパラメータベクトル $\boldsymbol{\beta}$ に対する最尤推定量ベクトルを $\boldsymbol{b}$，対数尤度関数を $l(\boldsymbol{b}; \boldsymbol{y})$ とし，モデル適合度検定統計量を次のように定義し

$$\mathrm{D} = 2\{l(\boldsymbol{b}_{max}; \boldsymbol{y}) - l(\boldsymbol{b}; \boldsymbol{y})\} \qquad (2\text{-}3\text{-}1)$$

逸脱度関数（厳密には，scaled deviance）と呼ぶ (Nelder & Wedderburn, 1972)。ところで，パラメータ数が p のとき，対数尤度比検定統計量の漸近分布は

$$2\{l(\boldsymbol{b}; \boldsymbol{y}) - l(\boldsymbol{\beta}; \boldsymbol{y})\} \sim \chi^2_p \qquad (2\text{-}3\text{-}2)$$

となる（(2-3-2)式については例えば竹村 (1991, p.306) の説明がある）。(2-3-1) を変形する。

$$\begin{aligned}\mathrm{D} = {} & 2\{l(\boldsymbol{b}_{max}; \boldsymbol{y}) - l(\boldsymbol{\beta}_{max}; \boldsymbol{y})\} \\ & - 2\{l(\boldsymbol{b}; \boldsymbol{y}) - l(\boldsymbol{\beta}; \boldsymbol{y})\} \\ & + 2\{l(\boldsymbol{\beta}_{max}; \boldsymbol{y}) - l(\boldsymbol{\beta}; \boldsymbol{y})\} \qquad (2\text{-}3\text{-}3)\end{aligned}$$

(2-3-4) の第一項，第二項は (2-3-2) より，それぞれ自由度 N（ただし，N はデータ数），p のカイ二乗分布に従う。一方，現在関心のあるデータに対するモデルの適合度が良ければ，第三項の値はゼロに近くなる。よって，関心のあるモデルのデータに対するモデル適合度が良好であれば，D は漸近的に自由度 N－p のカイ二乗分布に従い，標本分布 D を検定統計量に利用できる。一般に，自由度 p のカイ二乗分布に従う確率変数を Y とすると，$\mathrm{E}(\mathrm{Y}) = \mathrm{p}$ であるから，モデルが良好であれば，

$$\mathrm{D} \simeq \mathrm{N} - \mathrm{p} \qquad (2\text{-}3\text{-}4)$$

がおおざっぱに成立する。なお，検定統計量 D は，grouped data のときに利用可能で，ungrouped data の場合は利用できないが，この事項についての検討は 4 節において行う。逸脱度関数 D は，GLIM においてモデル適合度の検定に最も用いられる検定統計量である。ただし，確率分布の規定が前提となることは自明であるから，確率変数の二次モーメントまでしか仮定しないような拡張された一般化線形モデルでは，このような適合度の評価は不可能である。ところで，取り扱うデータが二項データの場合，D は離散的になるが，離散分布 D の極限分布が連続型確率分布のカイ二乗分布をするための条

件は，(1)サンプルが互いに独立に二項分布に従うこと，(2)部分母集団の大きさ m_i が十分に大きいこと，である(MacCullagh & Nelder, 1989, p.118)。

2.4 分析方法

本稿で利用するデータは佐伯・野尻ら（1994）が実行した調査研究により得られたものである。

この調査研究では，長距離トラック事故発生率とトラック運転者の身体的要因及び労働条件を直接リンクさせ，事故発生に関与する要因群の抽出や影響度合いを探求するアンケート調査である。アンケート調査の結果得られた労働科学データに対しては，第一ステップとして，過去5年以内に事故を起こした事故経験者と未経験者とを対比させ集計するという，ケース・コントロール的なデータ粗集計が行われた。そして，第二ステップでは事故経験者と未経験者との間で結果に有意差が認められる質問項目群を説明変数とし，事故発生率を応答変数とした，GLIM に属するロジスティク回帰分析によるデータ解析を行った。すなわち，事故を起こしたか否かという二値データ(Binary data)の解析である。

本稿は，この第二ステップの二値データ解析をGLIMの観点から再検討を行うものである。リンク関数としては，基本的な関数であるロジット関数$\left(\eta = \log\{\mu/(1-\mu)\}\right)$，プロビット関数$\left(\eta = \Phi^{-1}(\mu)\right)$，補対数対数関数$\left(\eta = \log\{-\log(1-\mu)\}\right)$を選択した。応答変数には期待事故率を，そして説明変数には，第一ステップから吟味されたもの，すなわち，年齢，家族構成，仮眠・休息時間，主観的過労度，狭心症，輸血歴，心身症，抑鬱性，軽そう感情，強迫感，恐怖感，運動動作能力，モチベーション，運転歴の14変数を選んだ。カテゴリー型説明変数をはじめとした，それらに対する得点化は表1のと

表1 諸説明変数に対する数値の割付

1) 目的変数（反応変数）		
事故経験者		1点
未経験者		0点
2) 説明変数		
① 家族構成	独身	1点
	それ以外	0点
② 長距離運行時の休息・仮眠状態	とれていない	1点
	それ以外	0点
③ 主観的過労感	ある	1点
	それ以外	0点
④ 狭心症，輸血暦，心身症	既往暦あり	1点
	なし	0点
⑤ 抑鬱性	0点（抑鬱性なし）	0点
	1～3点	1点
	4点以上（やや傾向あり）	2点
⑥ 軽そう感情	0点	0点
	1～3点	1点
	4点以上（やや傾向あり）	2点
⑦ 強迫感	0点	0点
	1～3点	1点
	4点以上（やや傾向あり）	2点
⑧ 恐怖感	0点	0点
	1～2点	1点
	3点以上（やや傾向あり）	2点

年齢，運動動作能力，モチベーションの項目では，データの数値をそのまま活用した。

おりである。なお，欠損値を含まない完全データ件数は665であった。また，変数選択においては，上記の説明変数群はどれも第一ステップで，十分吟味され抽出されたものであるので，全ての変数を同時に対象とする変数固定法を使用した。また，本稿ではデータ数が十分大きいので，grouped data とみなせるとしデータに対するモデル適合度は逸脱度関数Dの実現値により検討した。なお，AIC(赤池の情報量基準, Akaike, 1973) の値も付記することとした。

3．結　果

データ解析の結果，データに対するモデル適合度の良好度は，表2の逸脱度関数の実現値により示される。本稿の3種のリンク関数の中では，補対数対数関数の選択が最良の結果を与え，ついでプロビット関数，ロジット関数の順であったが，Dの値をみる限りでは，リンク関数の選択の違いによる大きな差異は認められなかった。対応するカイ二乗分布の自由度は df = 665 − 15 = 650 であり，一方，表2のDの値が650前後であることから上側確率は50%前後と推定され，モデル適合度はどのリンク関数を選択してもほとんど同じ程度に良好と言える。ところで，一般化ピアソン統計量を X^2 とすると，ちらばり母数 ϕ の推定量は，$\hat{\phi} = X^2/(N-p)$ で与えられる (McCullagh & Nelder, 1989, p.127)。$X^2 \fallingdotseq D$ が成立するので(Dobson, 1990, p.116)，本稿では $D/(N-p)$ によりちらばり母数を推定する。どのリンク関数を選択した場合でも，ちらばり母数の推定値は1前後の値に計算されるので本稿のデータ解析では Overdispersion は認められなかった。よって，本稿のデータ解析では擬似尤度を用いたモデル (Quasi-likelihood models) を利用する必要はなかろう。

次に，回帰係数に関する解析結果を表3～5について Wald 検定に基づき検討する。モデル適合度は同程度であるにもかかわらず，回帰係数に対する Wald 検定の結果として有意となる

表2　リンク関数の違いによるモデル適合度の変動

リンク関数	逸脱度関数の実現値	AIC
ロジット関数	653.248	683.248
プロビット関数	651.825	681.825
補対数対数関数	649.416	679.416

表3　解析結果1（ロジット関数）

説明変数	パラメータ推定値	標準誤差	Wald 統計量	P-value	有意性
切片	−0.4089	0.7327	0.3114	0.5768	
年齢	−0.0587	0.0219	7.1629	0.0074	＊＊
家族構成	−0.0291	0.2706	0.0116	0.9142	
仮眠・休息時間	−0.0876	0.255	0.1173	0.7319	
主観的過労度	0.6879	0.247	7.7564	0.0054	＊＊
抑鬱性	0.3122	0.1624	3.6952	0.0546	＊
軽そう感情	0.2735	0.1714	2.546	0.1106	
強迫感	−0.7712	0.189	16.6534	0.0001	＊＊
恐怖感	−0.1485	0.1601	0.8605	0.3536	
狭心症	1.4608	0.841	3.0167	0.0824	
輸血歴	0.1472	0.5514	4.328	0.0375	＊
心身症	0.6023	0.6436	0.876	0.3493	
運動動作能力	0.2562	0.0823	9.6897	0.0019	＊＊
モチベーション	0.0467	0.098	0.227	0.6338	
運転歴	0.0427	0.022	3.7649	0.0523	＊

＊＊：$p<0.01$，＊：$p<0.05$

表4 解析結果2（プロビット関数）

説明変数	パラメータ推定値	標準誤差	Wald 統計量	P-value	有意性
切片	−0.2713	0.4239	0.04095	0.5222	
年齢	−0.0339	0.0125	7.2946	0.0069	＊＊
家族構成	−0.00686	0.1594	0.0019	0.9657	
仮眠・休息時間	−0.0597	0.1484	0.1622	0.6872	
主観的過労度	0.3881	0.1448	7.1888	0.0073	＊＊
抑鬱性	0.179	0.0955	3.5124	0.0609	
軽そう感情	0.1683	0.1007	2.7919	0.0947	
強迫感	−0.4612	0.1105	17.4194	0.0001	＊＊
恐怖感	−0.0863	0.0941	0.8409	0.3592	
狭心症	0.8849	0.4967	3.1741	0.0748	
輸血歴	0.6693	0.3348	4.0029	0.0454	＊
心身症	0.3824	0.3801	1.0123	0.3144	
運動動作能力	0.1517	0.0468	10.4884	0.0012	＊＊
モチベーション	0.0227	0.0565	0.1618	0.6875	
運転歴	0.0251	0.0126	3.998	0.0456	＊

＊＊：$p<0.01$，＊：$p<0.05$

表5 解析結果3（補対数対数関数）

説明変数	パラメータ推定値	標準誤差	Wald 統計量	P-value	有意性
切片	−0.00762	0.3842	0.0004	0.9842	
年齢	−0.0293	0.0112	6.8187	0.009	＊＊
家族構成	0.0259	0.1497	0.03	0.8625	
仮眠・休息時間	−0.0783	0.1337	0.3432	0.558	
主観的過労度	0.3367	0.135	6.2225	0.0126	＊
抑鬱性	0.161	0.0891	3.2639	0.0708	
軽そう感情	0.1785	0.0944	3.5758	0.0586	
強迫感	−0.4418	0.1042	17.9782	0.0001	＊＊
恐怖感	−0.0777	0.0881	0.7773	0.378	
狭心症	0.9694	0.5968	2.6386	0.1043	
輸血歴	0.6013	0.3529	2.9038	0.0884	
心身症	0.4819	0.3904	1.5238	0.21771	
運動動作能力	0.1440	0.0414	12.0859	0.0005	＊＊
モチベーション	0.0148	0.0509	0.0843	0.7715	
運転歴	0.0227	0.0122	4.1361	0.042	＊

＊＊：$p<0.01$，＊：$p<0.05$

説明変数の組合せ（セット）がリンク関数の選択により異なることが，本稿のデータ解析結果では認められた。ロジット関数では，年齢，主観的過労度，抑鬱性，強迫感，輸血歴，運動動作能力，運転歴という7個の説明変数が有意となったことに対し，プロビット関数では，抑鬱性が，さらに，補対数対数では輸血歴が有意な説明変数から欠落したことがそれらの表よりわかる。すなわち，本稿のデータ解析はGLIMにおけるリンク関数の違いが要因解析に及ぼす影

響について一つの例証を与えたことになる。もちろん，説明変数間の共線関係の影響も見逃せない。

リンク関数の選択によって有意な説明変数の組合せ（セット）が異なってしまうことは，実際の調査研究における解析結果のインプリケーションに非常に大きな影響を与える。観測変量の期待値をリンク関数により変換することによって，データを変形することなく線形回帰を実行することがGLIMにおけるリンク関数の位置づけである。Cox (1984) が言うように計算アルゴリズムのうえで必要のない概念であるという見解もあるが，しかし解析結果のインプリケーションを省みると，リンク関数の選択には明確な基準がないのにもかかわらず，やはり見逃せない現実的な問題であることが示された。言うまでもなく，2項データに対する連結関数としてロジット関数がCanonical linkであるからといって，応用面で適切な選択であるという必然性はない。

なお，連結関数としてロジット関数を選んだ場合，回帰係数からオッズ比が計算できるので，7個の有意な説明変数についてのそれらを列挙する。

① 年齢： オッズ比は，exp(−0.0587)＝0.94（95%信頼区間0.90〜0.99）である。一歳加齢すると，事故率は，0.94倍低下する。

② 主観的過労度： オッズ比はexp(0.688)＝1.99（同1.23〜3.23）で，業務がきついと感じている運転者は，そうでない者に比べ，1.99倍，事故率が高い。

③ 輸血歴： オッズ比は，exp(0.1472)＝1.16（同1.07〜9.28）であった。

④ 抑鬱性： オッズ比は，exp(0.3122×2)＝1.87（同0.988〜3.53）で，抑鬱性の傾向のある運転者は，ない者に比べ，1.87倍，事故率が大きい。

⑤ 強迫感： オッズ比は，exp(−0.7712×2)＝0.214（同0.102〜0.448）で，強迫感のある者は，0.214倍事故率が低下する。

⑥ 運動動作能力： オッズ比は，exp(0.2562)＝1.29（同1.10〜1.52）で，一点上昇するごとに事故率は，1.29倍上昇する。

⑦ 運転歴： オッズ比は，exp(0.0427)＝1.04（1.00〜1.09）となる。

交通事故率に影響を与える有意な説明変数に関するさらに詳細な考察については，佐伯・野尻 (1994) による解釈を参照されたい。

4．考　察

逸脱度関数DによるGLIMのモデル適合度検定の妥当性は，sparsenessの時に問題となる事柄である。その極端な場合であるungrouped dataに対しては逸脱度関数は完全にモデルの評価に利用できないことを本稿の最後に考察する（McCullagh & Nelder, 1989）。

本稿は，二項データの解析であるので，それに呼応するDを検討する。データ y_i が二項分布 $B(m_i, \pi_i)$ に従う確率変数 Y_i の実現値とする。この時，尤度関数Lは，定数項を除外すると

$$L = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{m_i-y_i} \qquad (4\text{-}1)$$

となる。ここで，nは部分母集団の数である。ゆえに，対数尤度関数は

$$\begin{aligned} & l(\pi_1, \pi_2, \cdots, \pi_n ; y_1, y_2, \cdots, y_n) \\ &= l(\boldsymbol{\pi}, \boldsymbol{y}) \\ &= \log L \\ &= \log \Pi \pi_i^{y_i}(1-\pi_i)^{m_i-y_i} \\ &= \Sigma \log \pi_i^{y_i}(1-\pi_i)^{m_i-y_i} \\ &= \Sigma\{y_i \log \pi_i + (m_i - y_i)\log(1-\pi_i)\} \end{aligned} \qquad (4\text{-}2)$$

となる。ところで，フルモデルの最尤推定量は，当然 $\hat{\pi}_{imax} = y_i/m_i$ である。一方，関心のあるモデルのもとでの最尤推定量を，$\hat{\pi}_i = \hat{u}_i/m_i$ とすると，逸脱度関数Dは，対数尤度関数が $l(\boldsymbol{b}; \boldsymbol{y}) = l(\boldsymbol{\pi}; \boldsymbol{y})$ と表現できることに注意すれば，Dの定義から

$$D = 2\{l(\boldsymbol{b}_{max}; \boldsymbol{y}) - l(\boldsymbol{b}; \boldsymbol{y})\}$$

$$= 2\{l(\hat{\boldsymbol{\pi}}_{max}; \boldsymbol{y}) - l(\hat{\boldsymbol{\pi}}; \boldsymbol{y})\}$$
$$= 2\Sigma\{y_i \log(y_i/\hat{u}_i) + (m_i - y_i)\log(m_i - y_i)/(m_i - \hat{u}_i)\} \quad (4\text{-}3)$$

となる。これが，二値変数の場合の逸脱度関数Dの具体的な型式である。ここで，ungrouped dataとは，$m_i=1$である場合であるから，

$$D = 2\Sigma\{y_i \log(y_i/\hat{u}_i) + (1 - y_i)\log(1 - y_i)/(1 - \hat{u}_i)\}$$
$$= 2\Sigma\{y_i \log(y_i) + (1 - y_i)\log(1 - y_i) - y_i \log \hat{u}_i/(1 - \hat{u}_i) - \log(1 - \hat{u}_i)\} \quad (4\text{-}4)$$

となる。ところで，$\log \hat{u}_i/(1 - \hat{u}_i) = \boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}$, $\lim_{y\to 0}(y \log y) = \lim(\log y/y^{-1}) = \lim(y^{-1}/(-1)y^{-2}) = \lim(-y) = 0$である。よって，(4-4)の第一項，第二項は消えて

$$D = -2\boldsymbol{y}^{\mathrm{T}}(X\hat{\boldsymbol{\beta}}) - 2\Sigma \log(1 - \hat{u}_i)$$
$$(\boldsymbol{y}^{\mathrm{T}}(X\hat{\boldsymbol{\beta}}) = (X\hat{\boldsymbol{\beta}})^{\mathrm{T}}\boldsymbol{y} = \hat{\boldsymbol{\beta}}^{\mathrm{T}}X^{\mathrm{T}}\boldsymbol{y} = \hat{\boldsymbol{\beta}}^{\mathrm{T}}X^{\mathrm{T}}\hat{\boldsymbol{u}} = \boldsymbol{\eta}^{\mathrm{T}}\hat{\boldsymbol{u}})$$
$$= -2\hat{\boldsymbol{\eta}}^{\mathrm{T}}\hat{\boldsymbol{u}} - 2\Sigma \log(1 - \hat{u}_i) \quad (4\text{-}5)$$

となる。ゆえにungrouped dataに対しては，逸脱度関数Dはパラメータ推定量$\hat{\boldsymbol{\beta}}$だけの関数になり，適合度検定統計量として使用できない。すなわち，確率分布が退化し，モデル適合度検定統計量として利用できない。

本稿は，一般化線形モデルの枠組み，擬似尤度の概念，パラメータの推定方法，モデル適合度の検定法を説明した。続いて，リンク関数の選択が要因解析に与える影響についてデータ解析を行い考察した。すなわち，大規模な労働科学データを解析して，データを変形することなく線形回帰を実行することを目的とする，リンク関数の選択の違いにより，実際には有意な説明変数の組合せが異なることを示すことで，一つの例証を与えたわけである。

謝　辞

稿を終えるにあたり，データの全面的な提供及び終始調査研究の御指導を頂きました流通経済大学流通問題研究所，佐伯弘治学長，野尻俊明教授に深甚なる謝意を捧げます。またGLIMの研究指導を承りました学習院大学経済学部の新居玄武教授に感謝致します。

参考文献

Akaike, H. Information Theory and an Extension of the Maximum Likelihood Principle. 2nd International Symposium on Information Theory. Budapest, Académiai Kiado, 267-81, 1973.

Buse, A. The Likelihood Ratio, Wald and Lagrange Multiplier Test. *Am. Statistn.* **36**, 153-57(1982).

Box, G. E. P. & Cox, D. R. An Analysis of Transformations. *J. R. Stat.* B. **26**, 211-52(1964).

Cox, D. R. Some Remarks on Overdispersion. *Biom.* **70**, 269-74(1983).

Cox, D. R. Generalized Linear Models-The Missing Link. *Appl. Statist.* **33**, 18-24(1984).

Dobson, A. J. *An Introduction to Generalized Linear Models.* London, Chapman & Hall, 1990.

Mallick, B. K. & Gelfand, A. E. Generalized linear Models with Unknown Link Fuctions. *Biom.* **81**, 2, 237-45(1994).

McCullagh, P. Quasi-Likelihood Functions. *Ann. Statist.* **11**, 59-67(1983).

McCullagh, P. & Nelder, J. A. *Generalized Linear Models,* 2nd Ed. London, Chapman & Hall, 1989.

Nelder, J. A. & Wedderburn, R. W. M. Generalized Linear Models. *J. R. Stat.* A. **135**, 370-84(1972).

Pregibon, D. Book Review: Generalized Linear Models. *Ann. Statist.* **12**, 4, 1589-96(1984).

Wedderburn, R. W. M. Quasilikelihood Functions, Generalized Linear Models and Gauss-Newton Method. *Biom.* **61**, 439-47(1974).

佐伯　弘治，野尻　俊明　他　物流における長距離運転者の年齢的・身体的限界に関する調査研究．流通問題研究．24，12-80(1994)．

竹村　彰道　現代数理統計学．東京，創文社(1991)．

椿　広計　一般化線形模型の問題点と擬似尤度の一般化．応用統計学．17，1-12(1988)．