計数データと回帰分析
—中国地域の交通事故発生モデルの展開—

洲　浜　源　一

計数データ処理のための代表的なモデルである、ポアソン回帰モデルを交通事故発生の説明に適用する。使用するデータは中国地域5県の人口1万人以上の113市区町である。モデルの適合度については必ずしも満足な結果は得られていないが、検討した複数の独立変数のなかで事業所数の変数および瀬戸内側と山陰側（山陽山間部を含む）を区別するダミー変数の二つが良好な結果を示した。瀬戸内側と山陰側でそれぞれ事業所が1000箇所増加すると、前者では交通事故が5.8件発生し、後者では5.4件発生すると推定される。

キーワード：計数データ、回帰分析、ポアソン分布、交通事故件数

目　次

Ⅰ　はしがき

応用経済学で取り上げる回帰モデルの従属変数のなかには、整数値（0, 1、・・・）のみを取る離散的変数がある。このような値に対応するデータは他のデータと区別して計数データ（count data）またそのようなモデルは「計数モデル」と呼ぶれている。例えば企業が獲得した特許の数を研究開発費に回帰させるモデルではその特許の数が計数データである(Hausman, et al. 1984)。研究者が学位取得までに公表した論文数を研究者の個人的条件を表す変数（性別、子供数あるいは既婚・未婚等）に回帰させるモデルではその公表した論文数

(Long, 1997, Chp.8)、また船舶事故を船舶の種類とか建造年数等に回帰させるモデルではその船舶事故数（Green, 2000, Chp.19）はそれぞれ計数データである。また最近では Yoshida and Takagi（2002）が計数モデルを利用して健康保険制度の改革が診療回数等におよぼす影響を計数モデルで、また西本、駿河（2002）では育児休業制度を利用する企業数と企業が提供する諸支援施策との関係がそれぞれ計数モデルを利用して分析されている。

これらの計数モデルに共通して適用されている基本分析モデルは「ポアソン回帰モデル」(poisson regression models) およびこのモデルの欠陥を改良した「負の二項回帰モデル」(negative binomial regression models) である。ともに整数値をとる計数データの分析には適している。しかしポアソン回帰モデルの最大の欠陥は、独立変数Xのもとでの従属変数Yの条件付平均$E(y \mid x)$がその条件付分散$V(y \mid x)$に等しいという前提である。これは基礎となるポアソン分布そのものに由来する。しかし実際に多くの実証分析ではこの前提は満たれず、条件付分散が同平均を超える超過分散（overdispersion）の状態が多い（しかし本報告で扱う交通事故の計数データでは逆に過少分散が存在する）。そこでこの乖離を明示的に導入するため、通常ポアソン回帰モデルの条件付平均にガンマー分布に従う確率変数を加えて、負の二項回帰モデルが導かれる（Gourieroux, et al. 1984, Cameron、et al. 1986）。さて、このように単純なポアソン回帰モデルには条件付分散が同平均と等しいという最大の欠陥がある。しかしポアソン回帰モデルは、ガンマー分布との複合分布として導かれる負の二項回帰モデルよりも簡明である。さらに最尤推定法に基づく一階の条件より回帰係数の推定値は繰り返し加重最小２乗法（Hausman、et al. 1984, p.911）によって求めることができる、またその二階の条件は最大値の存在を保証する。しかし以下本報告で取り上げる交通事故発生数の計数データには０値が存在しない。したがってデータに対数変換を適用するより簡便な方法を用いてモデルを推定することができる。本報告のポアソンモデルの推定は、後の注（2）で示すように、この簡便法にしたがっている。この推定法の長所はこれまでの古典的回帰モデルの推定法に従うことができるということである。

この報告の後半では、中国地域５県における人口１万人以上の市区町の交通事故発生件数をポアソン回帰モデルで推定している。もともと交通事故発生に関するモデルの最終的目標は各種の事故防止施策（例えば、シートベルト着用等）の事故防止効果を量的な数値で評価することにある。その意味でこの報告のねらいは、そのための第一段階として安定した交通事故発生モデルを試論的に求めることである。

Ⅱ　ポアソン回帰モデル

本節ではポアソン回帰モデルを概説する。変数 y_i（$i = 1、2、\cdots、N$）を離散的変

数の観察値すなわち計数データとする。ここでiは観察個体の番号である。ポアソン回帰モデルではy_iが独立にパラメターμ_iのポアソン分布にしたがって分布すると仮定する。すなわちその密度関数は

$$f(y_i)=\frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!}$$

である。パラメターμ_iはポアソン分布の平均で、これはこの分布の分散に等しい。ポアソン回帰モデルではこれをK個の外生変数$x_i=(x_{i1},x_{i2},\cdots,x_{iK})$に関連付け、

$$\mu_i=\exp(x_ib)=\exp\left(\sum_{j=1}^{K}x_{ij}b_j\right)$$

と仮定する。ここで$b=(b_1,b_2,\cdots,b_K)$未知の回帰係数である。これよりパラメターμ_iが外生変数の指数関数として仮定されるためμ_iの正値性が保証される。またこの仮定は回帰係数bの推定を簡明にする。そして変数yの平均および分散はそれぞれ条件付$E(y \mid x)$, 同$V(y \mid x)$で表される。

上式の対数尤度関数は、

$$L(b)=-\sum_{i=1}^{N}\mu_i+\sum_{i-1}^{N}y_i\log\mu_i-\sum_{i=1}^{N}\log(y_i!)$$

$$=const.-\sum_{i=1}^{N}\exp(x_ib)+\sum_{i=1}^{N}y_ix_ib \qquad (1)$$

対数尤度関数$L(b)$を最大にする一階の条件は、

$$\frac{dL}{db}=-\sum_{i=1}^{N}x_i'[\exp(x_ib)-y_i]=0$$

$$-\sum_{i=1}^{N}x_i'(\mu_i-y_i)=0 \qquad (2)$$

となる。したがって回帰係数bの最尤推定量（b_{ML}）は上の非線形方程式（2）の解として与えられる（1)。またその二階の条件を表すヘシアン行列は

$$\frac{d^2L}{dbdb'}=-\sum_{i=1}^{N}x_i'x_i\mu_i \qquad (3)$$

である。ゆえに行列$X'X=\sum_{i=1}^{N}x_i'x_i$の階数が$K$（full rank）であれば、ヘシアン行列は負定符号行列となる。したがってが推定量が存在すれば、最尤推定量は一意に決定する。また

最尤法の通常の議論にしたがって、推定量 b_{ML} の漸近的な分散共分散行列は $b = b_{ML}$ で評価されたヘシアン行列から計算される。

ポアソン回帰モデルに関する推定結果の適合度指標が複数提案され、検討されている(Cameron, et al. ,1996)。本報告では2変数の単純相関係数に対応した次式を利用する。

$$R^2_{COR} = \frac{\left(\sum_{i=1}^{N}(y_i - \bar{y})(\hat{\mu}_i - \bar{\mu})\right)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2 \cdot \sum_{i=1}^{N}(\hat{\mu}_i - \bar{\mu})^2} \qquad (4)$$

ここで $\hat{\mu}_i$ は μ_i の推定量である。上式による適合度の定義は0と1の間にある。しかし独立変数の追加により、必ずしも指標は増加しない(Cameron, et al. , 1996、p.211)。

次にポアソン回帰モデルにおける過剰分散の検定統計量として、次のラグランジュ乗数統計量LMを利用する。

$$LM = \frac{\left(\sum_{i=1}^{N}(y_i - \bar{\mu}_i)^2 - N \cdot \bar{y}\right)^2}{\sum_{i=1}^{N}\hat{\mu}_i^{\,2}} \qquad (5)$$

対立仮説として「負の2項分布」を仮定しておくと、LM統計量の極限分布は自由度1のカイ2乗分布にしたがって分布する (Grenn, 2000, p.886)。

Ⅲ データ：交通事故発生件数等

本報告で取り上げるデータは中国地方五県（鳥取，島根，岡山，広島，山口）の人口一万人以上の市区町村である。利用する諸データの出所はすべて『 統計でみる市区町村のすがた2003 』(総務庁統計局 2002年)で、本書（以下、原表と呼ぶ）では2000年までに調査した最新の資料が掲載されている。本報告で取り上げるデータの総数は113件でその県別内訳は下表のとおり。なお人口を一万以上の市区町村に限定したために、対象となる村は存在しなかった。

表1 市区町の内訳（人口1万人以上）

	鳥取県	島根県	岡山県	広島県	山口県	合計
市	4	8	10	12	14	48
区				8		8
町	4	6	18	20	9	57
合計	8	14	28	40	23	113

（注）広島市は8区として分割 (2000年10月1日現在)

従属変数（計数データ）

交通事故発生件数（人口1000人当たり）：2000年1月1日～12月31日の間に発生した交通事故を各市区町（以下各地域と呼ぶ）の人口（2000年10月1日現在）で割り、これを1000倍して四捨五入した値である。車輌が関連した道路・踏み切り上の事故で、物損のみの事故は含まれない。そのカウント数および対応する度数は下表に示す。

表2　交通事故発生件数の分布表

カウント	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
度　数	0	2	1	9	15	18	20	20	15	8	4	0	0	1	0

独立変数（constant以外に五変数）

道路実延長（人口1000人当たり）：原表の道路実延長（1999年）を人口1000人あたりのkm数に換算した。この数値が高いほど交通事故発生の誘引になると考えられる。なお道路実延長は各地域における一級国道（高速自動車国道を除く）および市町村道路の総延長から重複使用・未使用・渡船の各部分を除いたｋｍ数である。

人口密度：各地の人口をその各地総面積（1999年）で割り、1km^{2}あたりの人口密度を作成し、人口集中度指標とする。人口集中度は交通事故を誘発すると考える。

事業所数（人口1000人当たり）：企業による車輌の所有台数を表す代理変数とする。また企業活動の多寡と事故発生件数の関係を検討するための補助変数とも考えられる。なお事業数は1996年現在の数値である。

地域要因：交通事故の発生には自然的要因が強く影響していると考えられる。気象条件あるいは地理的条件等の自然要因がそれである。この報告ではこれらの影響を地域要因と呼び、次のダミー変数D（1,0）を割り当てる。

山陽側の瀬戸内に面する市区町　；　D=1
山陽側の山間部に存在する市区町　；　D=0
山陰側の市区町　；　D=0

これによってD=1に分類される地域は73である。

表3　各変数の記述統計量

	交通事故(1)発生件数	道路実延長(1)	人口密度(2)	事業所数(1)	地域要因
平　均	6.03	14.11	718.42	52.88	73（D=1） 40（D=0）
標準偏差	2.06	8.93	1159.81	16.19	0.23
最大値	13	50.54	8314.6	158.73	1
最小値	1	2.35	33.98	25.48	0

注(1) 人口1000人当り　　注(2) 面積1 km^{2}当り

Ⅳ 推定結果

推定結果は表4に掲載している。この表の第2列の線形回帰ＯＬＳは通常の線形回帰モデルによる最小2乗推定結果である。また第3列から6列までがポアソン回帰モデルによる四モデルの推定結果である。(2) そのうち第3列は全変数によるポアソン回帰モデルの推定結果にあたる。線形回帰モデルとポアソン回帰モデルを直接比較することはできないが、回帰係数はそれぞれの独立変数について相異している。しかし道路実延長は両モデルともにマイナスであり予想した結果を示していない。また人口密度の効果は両モデルともに正であるが、その値は極めてわずかであり、しかも統計的に有意ではない。なお人口密度については、山林・湖沼面積等を除いた可住人口密度を用いて、同様な推定を試みたが期待した結果はえられなかった。

表4 交通事故発生の推定結果

変数	線形回帰*OLS*	ポアソン回帰モデル			
Constant	0.359 (4.827)	1.398 (9.107)	1.270 (8.727)	1.208 (8.670)	1.455 (26.161)
道路実延長	-0.040 (-1.911)	-0.010 (-2.275)			
人口密度	0.000 (0.934)	0.000 (0.278)	0.000 (1.404)		
事業所	0.030 (2.663)	0.004 (1.748)	0.003 (1.239)	0.004 (1.928)	
地方要因	2.154 (5.754)	0.387 (4.939)	0.427 (5.478)	0.467 (6.483)	0.463 (8.116)
R^2	0.385				
$L(b)$		-237	-236	-235	-233
$R^2_{(COR)}$		0.383	0.366	0.351	0.279
LM		18.368	17.666	17.163	14.173
$\bar{\mu}$		5.77	5.759	5.749	6.035

(注) 1. 括弧の数字は線形回帰ＯＬＳについてはt値、他はz値を表す。
2. $\bar{\mu}$ は $\hat{\mu}_i$ の平均、R^2 は*OLS*の決定係数を表す。他の記号は本文Ⅱ節参照。

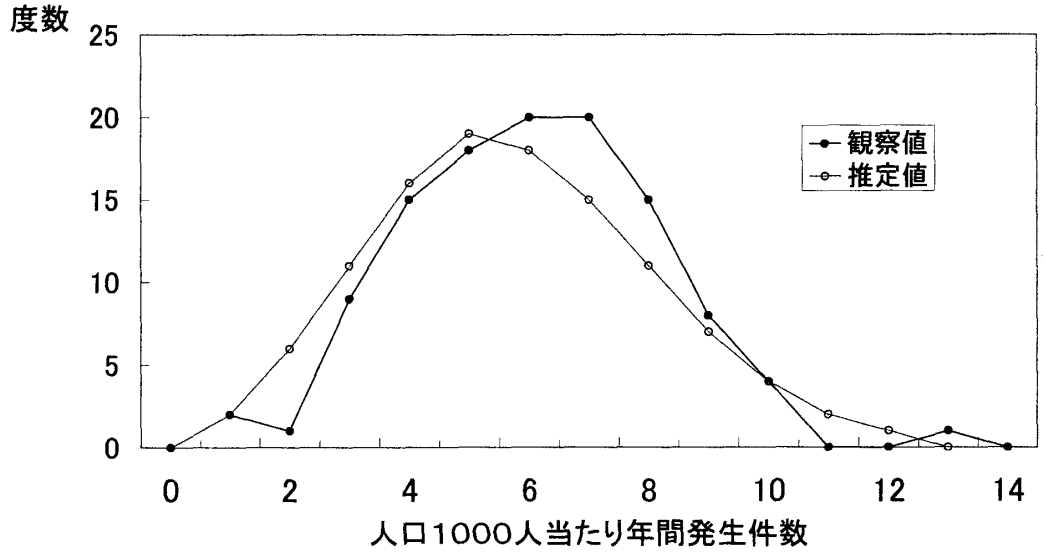
事業所変数については、期待した符号条件が満たされている。しかしポアソン回帰モデルでは統計的に有意な結果がえられていない。事業所変数は企業の車輌保有数の大小あるいは企業活動の多寡を交通事故発生に関連付ける代理変数である。本来は市町村別保有車輌数の記録あるいは企業活動を表す適切な指標に置換えるべきである。なお個人の保有車輌数の代理変数として地域の世帯数あるいは世帯当たりの家族数を検討したが符号条件に合わず、統計的に有意ではなかった。

地域の個別要因を表すダミー変数はすべてのモデルにおいて統計的にかなり有意な結果を

示しいる。特にポアソン回帰モデルではその推定値がほぼ同じ大きさ（0.387～0.463）である。検討した独立変数のうちでは最も安定度の高い結果がえられた。この結果、中国地方の瀬戸内側と山陰側（ただし山陽側の山間部を含む）の差異が，交通事故発生に大きな影響をあたえることをを表している。たとえば他の変数の影響を一定とすれば、瀬戸内側は山陰側の1.28～1.32倍の交通事故が発生する。ところでこの報告ではダミー変数として自然条件（気象条件、地理的条件等）を対応させた。しかし両地域を区別する諸点は、これらの自然条件のみにとどまらないであろう。経済的あるいは社会的な要因をさらに明示する必要がある。

さて推定結果全般からわかるように、適合度の指標（R^2,$R^2_{(COR)}$ ）はあまり高くない。特にポアソン回帰モデル（表4の第6列）では極度に低下している。またポアソンモデルに関する超分散の検定統計量LMは、すべてのモデルで14～18である。自由度1のカイ2乗分布より5％点は3.841であるから、ポアソン回帰モデルに不利な結果が示されている。最後に事故発生の観察値とポアソン回帰（表4の第3列のモデル）による推定値との比較は図1に掲載した。

図1　ポアソン分布による交通事故の推定



V　むすび

ポアソン回帰モデルは従属変数の平均がその分散に等しいという「等値性の仮定」と同変数の「確率的独立性の仮定」が、応用計量経済学分野での利用の前提となっている。前節の

最後に述べたごとく、本報告の交通事故発生モデルにおいても、前者の等値性の仮定が満たされていない。採用したLM統計量による検定は数ある検定法の一つである。他の検定法を採用するか、あらかじめ「負の2項分布モデル」の適用も考慮する必要がある。他方、ある地域の事故にその地域の住民・事業者のみが関係するとは限らない。地域間のクロス的事故発生も考慮するならば、本報告のような市区町単位の交通事故発生モデルでも前述の独立性の仮定を検証する必要がある。

交通事故のうち死亡事故に至る事故は、シートベルト類着用の普及等により毎年減少している。しかし全国の事故発生総数については、昨年（2002年）は前年よりも減少したものの、依然増加傾向にある。中国地域においても一部の例外はあるが、同様に増加傾向にある。有効な事故防止施策の選択のためにも、安定した交通事故発生モデルの構築が必要である。

注

(1) (本文3頁) 非線形方程式（2）式の解法について。繰返し法にしたがって解く。この式に含まれる非線型項 $\mu_i = exp(x_i b)$ の $b = b^{(0)}$ における近似展開式

$$\mu_i = exp(x_i b^{(0)}) + x_i exp(x_i b^{(0)})(b - b^{(0)})$$

を（2）式に代入して整理すると,第1段階の推定量

$$b(1) = b^{(0)} - [F'(\mu_i^{(0)})]^{-1} F(\mu_i^{(0)}) \qquad (6)$$

がえられる。ただし $F(\ ) = dL/db$ および $F'(\ ) = dL/dbdb'$ とし、μ_i を $\mu_i^{(0)} = exp(x_i b^{(0)})$ の点で評価したものである。以下同様に繰り返しの第2段階に進む。あきらかに上式（6）は高次方程式の根を求めるNewton近似法の拡張である（Winkelmann 2003, p.76)。さらに（6）式を本文（2)、(3）式を利用して次のように変形する。

$$b^{(1)} = b^{(0)} - \left[\sum_{i=1}^{N} x_i' x_i \mu_i^{(0)}\right]^{-1} \sum_{i=1}^{n} x_i (\mu_i^{(0)} - y_i)$$

上式の右辺第2項は加重された従属変数 $\{1/[1/(\mu_i)^{1/2}]\}(\mu_i - y_i)$ を同独立変数 $\{\mu_i\}^{1/2} x_i$ に回帰させる最小2乗推定にほかならない。ゆえに上式は加重最小2乗法の繰り返し推定である。いま $b^{(0)} = 0$ とすれば $b^{(1)} \fallingdotseq b_{ols}$（回帰係数の単純最小2乗推定量）となることが予想される。この結果は $b^{(0)} = b_{ols}$ を指示している。

(2) 本文（2）式より $\mu_i = exp(x_i b) \fallingdotseq y_i$ であれば、この一階の条件式は近似的に成立する。本報告では非線型回帰 $y_i = exp(x_i b)$ を推定することにより、ポアソン回帰モデルを近似推定した。

参考文献

Cameron, A. C. and P. K. Tribedi（1986）: "Econometric Models Based on Count Data: Comparisons and Application of some Estimators and Tests," *Journal of Applied Econometrics* 1, 29-53.

Cameron, A. C. and F. G. Windmeijer(1996) : "R-Squared Measure for Count Data Regression Models with Applications to Health-Care Utilization," *Journal of Business & Economic Statistics* 14 , No.2, American Statistical Association, 209-200.

Gourieroux, C., A. Monfort and A. Trognon : "Pseudo maximum Likelihood Methods : Applications to Poisson Models(1984)," *Econometrica* 52, No.3, May, 701-720.

Green, W. H. (2000): *Econometric Analysis*, fourth Edition, Prentice Hall, New Jersey.

Hausman , J., B H, Hall and Z, Griliches, (1984) : " Econometric models For Count Data with an Application to the Patents-R&D Relationship," *Econometrica* 52, No.4, etr July, 909-938.

Long, J. S. (1997) : *Regression Models for Categorical and Limited Dependent Variables*, Thousand Oaks, London.

Winkelmann, R. (2003): *Econometrics Analysis of Count Data*, fourth Edition, Springer, Berlin.

Yoshida, A. and S. Takagi (2002): "Effects of the Reform of the Social Medical Insurance System in Japan," *The Japanese Economic Review* 53,No.4,December、444-455.

西本真弓、駿河輝和(2002)：「ゼロ可変カウントモデルを用いた育児休業制度に関する実証分析」『日本統計学会誌』、第23巻、第3号、315-326.

資料・参考HP

『統計でみる市区町村のすがた　2003』総務庁統計局　平成13年12月.

警察庁交通局HP（2003.9）

「交通関係指標の推移」、「都道府県別交通事故発生状況」（『平成14年中の交通事故の発生状況』）

中国地域各県警HP(2003.10)