試験のためだけの損保数理
一般化線形モデル

pseudomathematician

平成 29 年 5 月 6 日

1　はじめに

線形回帰モデルにおいて，説明変数を複数必要とするような場合がある．モデリングの教科書から抜粋する．

例 1（モデリング教科書より引用）

コンビニ事業を営む企業が，時期 i $(i = n$ 年 4 月, 5 月, $\cdots)$ において，おでんの売上を分析したいと考えている．ある部署にアクチュアリーがいたので分析をお願いすることとした．まず，x_i をコンビニ利用者数，y_i をおでんの売上とする．回帰式 $y = \alpha + \beta x$ を考えたいが，季節的な影響によって，四半期単位で傾向が変わっていると考えられるため，それも加味して分析したいと考えている．この社員はモデリングを勉強したときの記憶があり，以下のような方法を取ることとした：ダミー変数 d_{1i}, d_{2i}, d_{3i} を，

$$d_{1i} = \begin{cases} 1 & (i = \text{第 1 四半期に属する月}) \\ 0 & (i \neq \text{第 1 四半期に属する月}) \end{cases}$$

$$d_{2i} = \begin{cases} 1 & (i = \text{第 2 四半期に属する月}) \\ 0 & (i \neq \text{第 2 四半期に属する月}) \end{cases}$$

$$d_{3i} = \begin{cases} 1 & (i = \text{第 3 四半期に属する月}) \\ 0 & (i \neq \text{第 3 四半期に属する月}) \end{cases}$$

として定義し，このダミー変数を用いて

$$y = \alpha + \beta_1 x + \beta_2 d_{1i} + \beta_3 d_{2i} + \beta_4 d_{3i}$$

という説明変数が 4 つである重回帰式を考える．この回帰式によって推定を行えば，

$$\text{第 1 四半期} \Longrightarrow y = \widehat{\alpha} + \widehat{\beta}_2 + \widehat{\beta}_1 x$$
$$\text{第 2 四半期} \Longrightarrow y = \widehat{\alpha} + \widehat{\beta}_3 + \widehat{\beta}_1 x$$
$$\text{第 3 四半期} \Longrightarrow y = \widehat{\alpha} + \widehat{\beta}_4 + \widehat{\beta}_1 x$$
$$\text{第 4 四半期} \Longrightarrow y = \widehat{\alpha} + \widehat{\beta}_1 x$$

を得る．

こうして得た値と実績値の二乗誤差を最小にするような $\widehat{\alpha}$, $\widehat{\beta_i}$ を求めることで回帰直線を得ることができる．

二乗誤差を評価して直線を求める理由を簡単に説明する．

線形回帰モデルでは，実績データと回帰直線から求めたデータの誤差分布が正規分布に従うと仮定する．そのことからおでんの売上を表す確率変数 Y も自然と正規分布に従うこととなる．その確率密度関数を $f(y, \mu) = \dfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$ とする．実績データ (x_i, y_i) に対し，$\mu_i = \alpha + \beta_1 x_i + \beta_2 d_{1i} + \beta_3 d_{2i} + \beta_4 d_{3i}$ と書き，$L = \prod f(y_i, \mu_i)$ に最尤法を適用する．対数尤度関数は以下のようになる．

$$l = \log L = \sum \left\{ -\frac{(y_i - \mu_i)^2}{2\sigma^2} - \text{Constant} \right\}.$$

この結果から二乗誤差を評価する理由がわかった．

一般化線形モデルは，データの確率分布は正規分布であるという回帰分析の前提において，正規分布ではなくポアソン分布やガンマ分布などの場合に拡張するものである．のちほど説明するがモデリングの教科書で紹介されている非線形モデル（指数関数モデル，ロジスティック関数モデルなど）は一般化線形モデルと解釈できる．

2　一般化線形モデル導入

一般化線形モデルを勉強しようとする動機は先述の通りであるが，損保数理の教科書の記号にあわせる必要があるため，先述の例を教科書の例に当てはめる．

[教科書 4.3.5] おでんの売上を季節ごとに分析したい．実績データとして以下が与えられている状態である．

季節	i	売上
第 1 四半期	1	1,000
第 2 四半期	2	800
第 3 四半期	3	600
第 4 四半期	4	500

（情報不足は否めないが）この情報から，各季節において，以下の表の“売上平均”を経営者に報告したい．

季節	i	売上 Y	売上平均 μ
第 1 四半期	1	Y_1	$E(Y_1) = \mu_1$
第 2 四半期	2	Y_2	$E(Y_2) = \mu_2$
第 3 四半期	3	Y_3	$E(Y_3) = \mu_3$
第 4 四半期	4	Y_4	$E(Y_4) = \mu_4$

[一般化線形モデル導入] まずはじめに注意として，回帰直線を求めることが目的ではないので，これからの説明では，説明変数（x）と定数項（α）は無視することとする．

ダミー変数 x_{ij} を以下のように定義し（教科書の記号に従う），

季節	i	x_{i1}	x_{i2}	x_{i3}
第 1 四半期	1	1	0	1
第 2 四半期	2	1	0	0
第 3 四半期	3	0	1	1
第 4 四半期	4	0	1	0

μ_i を求める準備として以下のような線形和 η_i を考える：

$$\eta_i := \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} \, .$$

これは以下のように表される：

$$\eta_i = \begin{cases} \beta_1 + \beta_3 & (i=1) \\ \beta_1 & (i=2) \\ \beta_2 + \beta_3 & (i=3) \\ \beta_2 & (i=4) \end{cases} .$$

ここからが一般化線形モデルである．求めたい μ_i は，リンク関数と呼ばれる関数 g にて，

$$\mu_i = g^{-1}(\eta_i)$$

と書かれていると仮定する．また，売上 Y の確率密度関数を $f(y) = f(y, \mu)$ とする．

一般化線形モデル

一般化線形モデルとは，以下の表からわかるとおり，Y の分布とリンク関数を一般化させることで，線形回帰モデルの拡張を与えるものである．

	リンク関数 $g(x)$			
Y の分布	x	$\log x$	$\log \frac{x}{1-x}$	$\frac{1}{x}$
正規分布	①	③	④	-
ポアソン分布	-	②	-	-
ガンマ分布	-	-	-	-
対数正規分布	-	⑤	-	-

以上の組み合わせの元で最尤法を適用する．

①は通常の線形回帰モデルを意味し，②は教科書に計算例があるものである．③④はそれぞれモデリング教科書にある指数関数モデル，ロジスティック関数モデルである．「-」としているのは，計算例が無いものである．恥ずかしながら筆者は①②以外の計算例は過去問にあるもの以外知らない．また，どういう例において，どの組み合わせを選ぶのかについても何も紹介が無いのでわからない．その時点で教科書としての価値を疑うが，粛々と計算例だけ覚えておくことが試験合格という意味で無難なのであろう．

なお，⑤であるが，これはモデリングの教科書にある「対数線形モデル」だと思われる．筆者はこれを確認していないので，推測を記載するのはモラル上問題はあるが，この文書はあくまで筆者の趣味による落書きと思ってもらって見過ごしてもらいたい．解釈に誤りがあれば指摘していただけると幸いである．

また，例 1 にあった $\alpha + \beta_1 x$ に該当する部分は μ_i を求める目的には不要な項目であるため，であればいっそのこと削除してしまい，$\eta_i := \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$ としたことは注意しておく．

それでは実際に計算してみる．

[解説] 一般化線形モデルには①を選択する．すなわち通常の線形回帰モデルである．Y_i の確率密度関数を $f(y_i, \mu_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\mu_i)^2}{2\sigma^2}}$ とする．

対数尤度関数 l を考える：

$$\begin{aligned} l &= \log \prod_{i=1}^{4} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\mu_i)^2}{2\sigma^2}} \\ &= \sum_{i=1}^{4} \left\{ -\frac{(y_i-\mu_i)^2}{2\sigma^2} - \text{Constant} \right\} . \end{aligned}$$

これについて，

$$\frac{\partial l}{\partial \beta_i} = 0 \quad (i = 1, 2, 3)$$

を解くことで μ_i が求まる：

$$\begin{aligned} \frac{\partial l}{\partial \beta_1} &= \frac{1,800 - (2\beta_1 + \beta_3)}{\sigma^2} = 0 \, , \\ \frac{\partial l}{\partial \beta_2} &= \frac{1,100 - (2\beta_2 + \beta_3)}{\sigma^2} = 0 \, , \\ \frac{\partial l}{\partial \beta_3} &= \frac{1,600 - (\beta_1 + \beta_2 + 2\beta_3)}{\sigma^2} = 0 \end{aligned}$$

より，

$$\beta_1 = 825, \ \beta_2 = 475, \ \beta_3 = 150,$$

$$\mu_1 = 975, \ \mu_2 = 825, \ \mu_3 = 625, \ \mu_4 = 475$$

を得る．なお，教科書に記載の解答と異なるが，教科書の解答は誤っているように思われる． ■

[**教科書 4.3.5**] 先述の計算問題を，リンク関数 $g(x) = \log x$ かつ Y の分布をポアソン分布として解け．（すなわち一般化線形モデル②で解け．）

[解説]

$$\mu_i = \begin{cases} e^{\beta_1 + \beta_3} & (i=1) \\ e^{\beta_1} & (i=2) \\ e^{\beta_2 + \beta_3} & (i=3) \\ e^{\beta_2} & (i=4) \end{cases}$$

と，

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

より，対数尤度関数 l は

$$l = \log \prod_{i=1}^{4} e^{-\mu_i} \frac{{\mu_i}^{y_i}}{y_i!} = \sum_{i=1}^{4} \{-\mu_i + y_i \log \mu_i - \text{Constant}\}$$

となる．これについて，

$$\frac{\partial l}{\partial \beta_i} = 0 \quad (i = 1, 2, 3)$$

を解く：

$$\frac{\partial l}{\partial \beta_1} = e^{\beta_1} \cdot (e^{\beta_1} + 1) = 1,800 \ ,$$
$$\frac{\partial l}{\partial \beta_2} = e^{\beta_2} \cdot (e^{\beta_1} + 1) = 1,100 \ ,$$
$$\frac{\partial l}{\partial \beta_3} = e^{\beta_3} \cdot (e^{\beta_1} + e^{\beta_2}) = 1,600$$

より，$\beta_1, \beta_2, \beta_3$ が求まり，結果，$\mu_1, \mu_2, \mu_3, \mu_4$ がわかる．ただし，教科書の解答が怪しいので計算は各自確認してもらいたい． ■

[過去問 H25] 危険標識を地域（都市か郊外か）および構造（木造か非木造か）の 2 区分で設定している火災保険があり，その実績クレーム単価のデータが下表の通りであったとする．

＜クレーム単価＞

	木造	非木造
都市	300	400
郊外	400	500

地域・構造別のクレーム単価 Y_i $(i = 1, 2, 3, 4)$ を一般化線形モデル，すなわち，Y_i の従う指数型分布族をポアソン分布 $P(Y_i = y_i) = e^{-\mu_i} \dfrac{\mu_i^{y_i}}{y_i!}$ （ここで $\mu_i = E(Y_i)$ である），リンク関数を $g(x) = x$ とし，次のとおり定義される説明変数 $x_{ij} (i = 1, 2, 3, 4, j = 1, 2, 3)$ を用いて，$\mu_i = g^{-1}(\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})$ と表されるモデルを用いて分析する．

区分	i	x_{i1}	x_{i2}	x_{i3}
[都市, 木造]	1	1	0	1
[都市, 非木造]	2	1	0	0
[郊外, 木造]	3	0	1	1
[郊外, 非木造]	4	0	1	0

ここで $\beta_1, \beta_2, \beta_3$ はパラメータであり，最尤法で推定する．このとき次の（1），(2)，(3）の問に答えなさい．

（1）対数尤度関数を表せ．

（2）$\beta_1, \beta_2, \beta_3$ が満たす連立方程式を表せ．

（3）「都市かつ非木造」のクレーム単価の期待値 μ_2 を求めよ．

[解説] 粛々と計算するだけである．

[過去問 H26] 危険標識を年齢（26 歳異常か 26 歳未満か）と用途（自家用か営業用か）の 2 区分で設定している自動車保険があり，その実績クレーム単価のデータが下表の通りであったとする．

＜クレーム単価＞

	自家用	営業用
26 歳以上	300	400
26 歳未満	400	500

年齢・用途別のクレーム単価 Y_i $(i = 1, 2, 3, 4)$ を一般化線形モデル，すなわち，Y_i の従う指数型分布族をガンマ分布 $f(y_i; \mu_i, \phi) = \dfrac{y_i^{-1}}{\Gamma(1/\phi)} \left(\dfrac{y_i}{\mu_i \phi} \right) \exp \left(-\dfrac{y_i}{\mu_i \phi} \right)$，リンク関数を $g(x) = 1/x$ とする．計算過程を明記し，各 μ_i を求めよ．

[解説]

区分	i	x_{i1}	x_{i2}	x_{i3}
[26 歳以上, 自家用]	1	1	0	1
[26 歳以上, 営業用]	2	1	0	0
[26 歳未満, 自家用]	3	0	1	1
[26 歳未満, 営業用]	4	0	1	0

$$\mu_i = g^{-1}(\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}) = \begin{cases} \frac{1}{\beta_1 + \beta_3} & (i = 1) \\ \frac{1}{\beta_1} & (i = 2) \\ \frac{1}{\beta_2 + \beta_3} & (i = 3) \\ \frac{1}{\beta_2} & (i = 4) \end{cases}$$

より，粛々と計算するだけである． ■

このほかにも問題例はあるが（H22），計算ルールに従うだけで，難しい要素は何も無い．各自確認してもらいたい．

なお，諸条件の元では結果が Minimum Bias 法と一致するらしい．詳しくは教科書を参照されたい．

以上