

CHAPTER 7 Other GLMs

前半：ガンマ分布と逆正規分布について

後半：dual GLM と疑似尤度について

7.1 Gamma GLM

ガンマ分布の確率密度関数： $f(y) = \frac{1}{\Gamma(v)} \lambda^v y^{v-1} e^{-\lambda y}$ $y > 0$

v : 形状母数 (shape), λ : 尺度母数 (scale)

$\lambda = v / \mu$ を代入すると

$$f(y) = \frac{1}{\Gamma(v)} \frac{v}{\mu} y^{v-1} e^{-\frac{vy}{\mu}} \quad y > 0$$

- ガンマ分布は形状母数 v によって分布形が変化する (Figure 7.1)

$\lambda = 1/2, v = df/2$ の時、ガンマ分布は χ^2 分布となる。

- Canonical parameter: $\theta = -1/\mu \rightarrow$ canonical link: $\eta = -1/\mu$

- $b(\theta) = \log(1/\mu) = -\log(-\theta)$
- $b''(\theta) = \mu^2$ (← variance function) (p117 6.1 の表参照)
- deviance: $D(y, \hat{\mu}) = -2 \sum_i \{ \log y_i / \hat{\mu} - (y_i - \hat{\mu}) / \hat{\mu}_i \}$ (p121 6.2 の表も参照)

- Gamma GLM の使用

1. response 変数がガンマ分布になることが分かっている時

2. response の平均と分散の関係を推測できそうな時

→ 分布を確かめるには大量のデータが必要であるが、平均と分散の関係把握は比較的、少量のデータをグラフ化してもわかる。

- $\text{var } Y \propto EY$ の関係にあり、離散値 → ポアソン分布
- $\text{var } Y \propto (EY)^2$ の関係にあり、連続値 → ガンマ分布 ($\text{var } Y \propto (EY)^2$)
response を log 変換 (対数正規分布)

- Gamma GLM のリンク関数

1. $\eta = \mu^{-1}$ (default)

... mean response に限界値があるときに (??)

2. $\eta = \log \mu$

... predictor の効果が平均に対して二乗であるか疑わしい時に用いる。分散が小さい時、このアプローチは log 変換した response に対する Gaussian model に類似する。

3. $\eta = \mu$

... 分散成分や平方和をモデリングするときに有効。ガンマ分布の特別なケースである。

例 1：半導体の製造過程（Myers & Montgomery 1997）

- 4 要因（ x_1, x_2, x_3, x_4 ）が抵抗率（skewed distribution）に及ぼす影響を、full factorial デザインによる実験を行った

1. log 変換 → gaussian GLM

- ・ Box-Cox 変換より log 変換が妥当であると。そこで、抵抗率を log 変換して、AIC を用いたモデル選択を行った。

結果: 3 つの相互作用が選択された。

2. gamma GLM

- ・ gamma GLM による AIC によるモデル選択を行った（リンク関数は log を使用）

結果: 3 つの相互作用

1 と 2 の結果は非常に類似。係数や SD など類似

例 2：保険金の請求に対する支払い（Hallin & Ingenbleek 1983）

- スウェーデンの各地域（Make） 走行距離、ボーナス（車のタイプと過去の請求額）と「保険金の請求に対する支払い」との関係
 - ・ 掛けた保険が多いほど請求が多くなることが予想される → offset を用いる。

例 1 と同様に、log 変換した線形モデルと gamma GLM を比べてみると・・・

Kilometers（走行距離）が gamma GLM では有意であるが、log 変換の法では有意ではない。係数についても違いが見られる。

- ・ 二つのモデルを直接 AIC を使って比較はできない（入れ子になってないし、response の分布形が異なる）
 - ・ 両モデルの null deviance はほぼ同じであるが、residual deviance は gamma GLM の方が小さい → gamma GLM の方が好まれる？（residual を用いての診断は危険）
- 両モデルから得られた dispersion parameter を用いて、分布形を書いてみる（Figure 7.2）
 - ・ gamma GLM よりも log 変換した線形モデルの方が、少ない支払いに偏っている（尖度が大きい）
 - ・ 両モデルから得られる値を実際に算出してみる（条件、Make = 1, Kilometres = 1, Bonus = 1, Insured = 100）→ gamma GLM（63061）の方が log 変換線形モデル（59754.5）よりも高い。

7.2 Inverse Gaussian GLM

逆正規分布の確率密度関数: $f(y|\mu, \lambda) = (\lambda/2\pi y^3)^{1/2} \exp[-\lambda(y-\mu)^2/2\mu^2 y]$ $y, \mu, \lambda > 0$

平均: μ , 分散: μ^3/λ

- Canonical link: $\eta = 1/\mu^2$
- 分散関数: $V(\mu) = \mu^3$
- deviance: $D = \sum_i (y_i - \hat{\mu}_i)^2 / (\mu_i^2 y_i)$

- 逆正規分布も λ によって分布形が変化する (Figure 7.3 → 全て $\mu = 1$, λ が右から 0.5, 1, 5)

$\mu = 1$ の時の分布形は Wald 分布と言われる。

- inverse Gaussian GLM の使用

- 一定ではない故障率を持つ耐用年数の分布
- ブラウン運動の first passage time の分布 (??)
- ガンマ分布よりも急激に分散が大きくなるような時

例 売上高 (Whitmore 1986)

- 見積もり売上高 (projected) と実際の売上高 (actual): $y_i = \beta x_i$
 - 売上高の範囲が非常に広く (114 ~ 5959) 正の値しか取らない (正規分布 ×)
 - とりあえず、Gaussian GLM と inverse Gaussian GLM との比較
- 二つのモデルから得られる近似曲線に明らかな違いが見られた
(Figure 7.4 の左図 inverse Gaussian の方が傾きが急)
- Residual deviance を用いて、inverse Gaussian GLM モデルの逸脱度を求める (Figure 7.3)
Residual deviance が徐々に減少傾向にあることから inverse Gaussian では "too strong".
ここでは gamma GLM がより良い選択。

7.3 Joint Modeling of the Mean and Dispersion (初見のため? 多し)

- 共変量 X に伴って dispersion parameter: ϕ_i が変化する時について考えてみる (これまでは一定).

例: どんなレシピで作っても美味しいケーキが作れるようなケーキミックスを製造したい

(= 常に平均に近く、分散が小さいものを製造したい。)

- 分量や焼き時間・温度が多少間違えても、美味しいものを焼けるようにしたい

Table 7.1 :

- Response : 評点 ?
- Predictor : 小麦 (F), バター (S), 卵 (E), オープンの温度 (T), 焼き時間 (t)
- X (predictor) つまり分量や時間、温度に伴って dispersion parameter が変化するモデルを考えたい。

Joint Model Speciation (? ?)

$$\bullet \quad \mathbb{E}Y_i = \mu_i \qquad \eta_i = g(\mu_i) = \sum_j (x_{ij}\beta_j) \qquad \text{var } Y_i = \phi V(\mu_i) \qquad w_i = 1/\phi_i$$

- ϕ を fixed として考えない

$$\begin{array}{lll} \rightarrow \text{Ed}_i = \phi & \xi_i = \log(\phi_i) = \sum_j (z_{ij} \gamma_j) & \text{var } d_i = \tau \phi_i^2 \\ & \text{? ? ?} & \end{array}$$

例：溶接

- response: 溶接の強さ
 - predict: 9 要因 (どれも 0,1 データ)
-
- 線形モデルで考えてみると・・・Drying と Material が有意であり他の要因は有意でない。
 - 重みづけした解析を行うと・・・Drying と Material と Preheating が有意になった

7.4 Quasi-Likelihood (疑似尤度)

- 新しいモデルを作るためには、リンク関数と分散関数があれば良く、response の分布形が何なののかについての確固たる情報は必要ない。そのため、標準的な GLM に対応していないリンク関数と分散関数の組み合わせがあったとしても対応が可能である、分布形の情報は不要である。
- ただし、Deviance を求めるには尤度が必要であり、尤度を求めるには分布形が必要である。そのため、分布を想定しないので、尤度の代用が必要となる（疑似尤度）。

- Y_i (平均: μ_i 分散: $\Phi V(\mu_i)$) とすると U_i が次のように定義される

$$U_i = \frac{(Y_i - \mu_i)}{\phi V(\mu_i)}$$

$$EU_i = 0$$

$$\text{var } U_i = \frac{1}{\phi V(\mu_i)}$$

???

- 疑似尤度と対数尤度は対応関係にある。しかし、二項分布もしくはポアソン分布と対応した分散関数をもつモデルに対して疑似尤度を用いるには次の利点がある。
- 通常の GLM は $\phi = 1$ としているが、quasi-binomial や quasi-Poisson GLM では free parameter。
- dispersion parameter (ϕ)以外の全ての parameter は通常の GLM と同じ方法で求められるが、 ϕ は次の式が推奨されている。

$$\phi = \frac{X^2}{n - p}$$

- 疑似尤度は仮定が少ないという利点があるが、最尤推定法よりも劣っている（疑似尤度使うな）

- $D(y, \mu) = -2 \phi \sum (l(\mu|y_i) - l(y_i|y_i))$
- 疑似尤度: $Q = -2 \sum \int_{y_i}^{\mu_i} (y_i - t) / V(t)$

例 眠っている間の夢見る割合

- response : 62 種類の哺乳類の夢見て過ごす睡眠の割合
- predictor : 体重、脳重量、寿命、妊娠期間、捕食者からの攻撃されやすさ (predation)、見つかりやすさ (exposure)、全体の危険度? (danger)
- 夢見時間 / 睡眠時間 (= μ) を求めると
 - ゼロから睡眠の半分と変異が大きい (通常の GLM モデル ×)
中程度で分散が大きい
 - 分散関数が $\mu(1-\mu)$ の近似形であることを示唆 (binomial に対応)
quasi-binomial を使用

結果 : drop1 を用いて一つずつ落としていくと、残るのは log(body), log(lifespan), danger の 3 つ。重い体重、短い寿命、危険にさらされない哺乳類ほど割合が大きくなる。

疑似尤度は AIC を算出できないため、step は使えない。

null deviance に対して residual deviance が大きいため、モデルに上手く fit していない。

- モデルの診断を行うと・・・ (Figure 7.5)

Figure 7.5 左図 → アジア象が完全に影響している。

Figure 7.5 右図 → 分散が一定 (?) なので、分散関数の選択が reasonable。