

6.1 GLMの定義

GLMは2つの要素（指数関数属、リンク関数）から成り立っている。

● 指数関数属

- ・ 定義式：P115①

(注) θ : canonical parameter . . . location (位置?) を表す

φ : dispersion parameter . . . scale (スケール?) を表す

- ・ a, b, c を決めることで、様々なモデル族を定義できる。

正規分布 : $\theta = \mu$ 、 $\varphi = \sigma^2$ 、 $a(\varphi) = \varphi$ 、 $b(\theta) = \theta^2/2$ 、

$$c(y, \varphi) = -(y^2/\varphi + \log(2\pi\varphi))/2$$

ポワソン分布 : $\theta = \log(\mu)$ 、 $\varphi = 1$ 、 $a(\varphi) = 1$ 、 $b(\theta) = \exp(\theta)$ 、 $c(y, \varphi) = -\log y!$

二項分布 : $\theta = \log(\mu/(1-\mu))$ 、 $b(\theta) = -\log(1+\exp(\theta))$ 、 $c(y, \varphi) = \log(nCy)$

- ・ ガンマ分布、逆正規分布、他の使用頻度の低い分布は7章で。
- ・ 負の二項分布やワイブル分布のような密度関数は指数関数属に含まれないが、修正することでGLMとして扱うこともできる
- ・ 指数分布属はP116②式に示すように平均と分散を持っている。
- ・ 平均 : θ のみの関数、分散 : θ と φ 両方の関数。
- ・ $b''(\theta)$: variance function (分散関数) と呼ばれ、分散と平均の関係を示している。
- ・ また、 $a(\varphi) = \varphi/w$ w は重み を定義する。

● リンク関数

- ・ リンク関数 g とは、平均値 μ を線形式 η (P116④式) と結びつける関数 (P116⑤式)
- ・ $\eta = g(\mu) = \theta$ となっているので、指数分布属では canonical parameter がリンク関数 (canonical link) となっている。各属のリンク関数はTable6.1に示すとおり。

(例) 正規分布 : $\eta = \mu$ 、ポワソン分布 : $\eta = \log \mu$ 、

二項分布 : ロジット、プロビット、complementary log-log 関数

6.2 GLM へのあてはめ

- ・ GLM のパラメータ β は、最尤法によって導かれる。
- ・ 正規分布以外では β の解は定まらないのだが、Fisher のスコア法を用いた Newton-Raphson 法による繰り返し最小 2 乗法 (IRWLS) によって最適化することができる。(回帰を重み付けすることで、 β が得られ、それをさらに重み付けしてより適した β が得られ・・・これを繰り返すことで、最適な $\hat{\beta}$ が得られる)
- ・ この当てはめ処理には、 $\eta = g(\mu)$ と $V(\mu)$ (分散関数) のみを使っており、 y の分布を考慮しなくていい。

(2.7 章の Bliss data をつかって、このアルゴリズムを調べてみよう！)

- ・ 各パラメータは、P118 最終行のとおり
- ・ で、計算すると (P119 の上のプログラム) 切片 -2.3025、傾き 1.1536 となった (これは、P118 の結果とは異なる)
- ・ でもこれを 5 回繰り返してみると (P119 真ん中のプログラム)、P118 の下の結果になる。
- ・ この例ではあつという間に収束した。多くの例でそのような収束が速やかにおこる。
- ・ もし収束しないようだったら、モデルがおかしいか、データがおかしいかのどちらか。
- ・ 収束しない例は 2.8 章にある。

6.3 仮説の検定

- ・ 統計モデルは、データをどのようにして、系統的なものとランダムなものに分けるか、というもの
- ・ null model は最小のモデル (通常、 y =切片だけのモデル)。
- ・ saturated や full モデルは、データを系統的に表現したモデル。
- ・ ただ、saturated model は、データを完全に説明するモデル (n パラメータ n データの多項式モデル) →あまり情報価値のないもの
- ・ さまざまなモデルがある中で、モデル間でどちらが優れているかということは、対数尤度を用いた尤度比検定で調べる。(P120①式)
- ・ もし、各々の測定・観察が独立していて、データが指数分布属だとしたら、(P120②式) に簡易化できる：簡易化したモデルは、 $D(y, \hat{\mu})/\phi$
($D(y, \hat{\mu})$ は deviance、 $D(y, \hat{\mu})/\phi$ は scaled deviance と呼ばれる)
- ・ 各分布属の $D(y, \hat{\mu})$ は表 6.2 のとおり。
- ・ なお、検定には Pearson's X^2 も使われることもあるが (P121 上式)、このときの $V(\hat{\mu})$ は deviance の代わりのものである。

(2種類の仮説検証)

1. **goodness of fit test** : 現在のモデルがデータとフィットしているかを調べる。

- **goodness of fit test** では、モデルは正しいものとみなして、Deviance や Pearson's X^2 をパラメータ数の自由度を持つ X^2 分布に近似させて検定を行う
- 正規分布では **dispersion parameter** ϕ がないので、このテストを行うことができない
- また、データ数が少ないと、この近似が有効でなくなる。)

2. **comparing model test** : 2つのモデル (パラメータの少ないモデルと多いモデル) の比較

- 大きなモデル Ω とその入れ子状になっている小さなモデル ω を比較するとき、deviance の差 ($D_\omega - D_\Omega$) は、自由度が2つのモデルのパラメータ数の差の X^2 分布に近似できる
- **dispersion** ϕ が不明な正規分布などの場合、この検定は直接使えないが、 ϕ の推定値を導入すれば、(P121②式) の形で、F 分布に近似して検定することが出来る。

(Bliss データを使ってみたいよう！)

● **goodness of fit test**

- **goodness of fit test** は、**residual deviance** と **df** との値の考察で行うことができる
- 結果、有意ではないので、フィットしていないわけではない。
- この結果は、このモデルが正しくて、他にいいモデルがない、ということを意味しているわけではない。
- **null model** をみると、deviance が 64.7 で df が 4 ということで、**null model** はあまりよろしくないことが分かる。

● **comparing model test**

- **conc** を入れたモデルと **null model** との比較。すると、(P122 中ほどの結果) **concentration** 項は有意だということが分かった。
- また、より複雑なモデル (2乗の項を加えたモデル) を使って、**fit** するかどうか調べた (P122 下の結果)。すると、新たな項の必要性は認められなかった。

● **Wald test** (他のアプローチ)

- 推定したパラメータの標準誤差を使って、**z 統計量** ($\hat{\beta} / \text{se}(\hat{\beta})$) (normal null distribution の推定値 (??)) を定義する。
- Bliss データの場合、**z 統計量** は 6.4 であり、**P 値** は 0 に近かった (帰無仮説を棄却)。
- ただ、正規分布ではこの統計はできない。正規分布の場合は、**z 統計量** を **t 統計量** とし、**t 分布** を用いて調べる。

6.4 GLM 診断

- ・ 線形モデルと同様に推定をチェックすることが大事である。
- ・ 正規線形モデルのチェック方法と似ているが、GLM のタイプにあわせて診断方法を変えることが求められる。

Residuals

診断に使う GLM の残差を以下のように定義しよう

- ・ Pearson residual r_p (ピアソン残差) : 正規分布の場合に使われたが、(P123①) と定義する。 $V(\mu)$ を $b''(\theta)$ とする。 $\sum r_p^2 = X^2$ となる。これはノンパラのときにも使える。
- ・ Deviance residual r_D (尤離度残差) : $\sum r_D^2 = \text{Deviance} = \sum d_i$ と定義される。つまり、 $r_D = \text{sign}(y_i - \hat{\mu}_i) \sqrt{d_i}$ 。ポワソン分布の場合は、(P123③ 式) となる。モデルのチェックにもっともよく使われる。
- ・ Response residuals (応答残差) : モデルとデータの差。
- ・ working residual (作業残差) : IRWLS による残差

Leverage and influence (てこ比とあてはめへの影響)

(leverage について)

- ・ 線形モデルでは、 $\hat{y} = Hy$ として計算され、 H は hat matrix と呼ばれる。
- ・ ここで、 h_i (=leverage、てこ比) を H の対角要素とする。 h_i (leverage) はモデルのフィットにどれだけ影響を与えたかその能力を示している
- ・ 重みづけ行列 W と hat matrix (H) を P124 ③ のように定義すると、 H の対角要素から leverage が抽出される
- ・ leverage が大きいと、独立変数が適当なものではないことを示している。
- ・ 線形モデルでいう leverage との違いについて、GLM の leverage は X の関数ではなく、 W に依存するということだ。
- ・ GLM でも線形モデルと同様に、スチューデント化された残差 (P124⑤) もしくは、ジャックナイフ残差 (i データを含むモデルの残差と、 i データを除いた場合のモデルの残差) により、当てはまりをみることができる。
- ・ 外れ値があると、ジャックナイフ残差が大きくなる

(Influence について)

- ・ それぞれの要素を除いたときの適合具合のチェックを P125③ で行える
- ・ もちろん Cook 統計を使ってもわかる (P125④⑤) 。
- ・ いずれの手法でも最初の要素を除いたときの影響が一番大きいことが分かる

Model diagnostics

2種類の手法が挙げられる。

1. モデルにフィットしていないものを検出するもの（外れ値の検出など）
 2. モデルの推定を検査するもの（独立要因の選択や変換についてなどモデルの形のチェック、および平均値の分散などについて確率論的なチェック）
- ・ 線形モデルでは、実際のデータに対する予測値との関係をグラフにするのが一つの効果な手法だった。
 - ・ GLMの場合は、予測値としてふさわしいスケールとして、 $\hat{\mu}$ よりも $\hat{\eta}$ を用いるのがよい。

（このことを、ガラパゴスデータ（3-1章）をつかってみてみよう）

- ・ Fig6-1は $\hat{\mu}$ を使ったプロット P126①。この場合、多くのプロットが左側に集中しているため、residual と fitted value の関係を評価するのは難しい
- ・ 一方、 $\hat{\eta}$ を使ってみると P126②、（真ん中のグラフ）x軸方向に対して一様に分布した。

（診断方法について）

1. residual と fitted value の間に、非線形関係があるのかどうか。

→ 非線形関係があったとしたら、モデルにうまくフィットしていないということであり、モデルの修正が求められる

修正方法

独立変数を変えたり変換したりすることが可能なら、それがもっともよい方法である。リンク関数を変える場合もあるが、モデルの解釈を容易にさせるリンク関数は数種しかないため、あまり望ましい方法ではない。

2. 残差のばらつきはどうか。

→ GLM の推定には、プロット内で一定の分散（等分散？）が必要である。（ちなみに、この場合、等分散は成り立っている）

修正方法

7.4章で述べる擬似尤度 GLM を使うのがいい。ポワソン分布の場合は、代わりに負の二項分布を使ってみることも可能である。また、データの特徴を認識できるならば、weight を使って修正してもよい。なお、もし分散関数が正しいのならば、Deviance residual(尤離度残差)を用いてグラフを書くと、プロットのばらつきが均一になる。

（残差プロットが有用ではない場合）

- ・ 残差のプロットが曲線になる場合。例えば、binary なものや、グループサイズの小さい二項分布やポワソン分布で起こりうる。
- ・ また、二項分布で片方のクラスのサイズがもう片方に比べてすごく大きいときにもこの問題が生じる

<診断のもう一つの目的>

独立変数と従属変数の関係を明らかにすること（ガラパゴスデータを使ってそのことを見てみよう！）

- ・ 島の面積と種数との関係をプロットする (Fig6.2 左)
- ・ x も y もどちらも偏っている。
- ・ そこで、独立変数（島の面積）を対数変換してみると (Fig6.2 中)、曲線関係が認められる。
- ・ ただ、 y の値はもともと対数をとっているので（ポワソンのリンク関数は対数だから）、そのことを考慮に入れないといけない。
- ・ リンク関数の影響を考慮に入れるため、直線応答； z 値を y 軸にとってプロットしてみる (P128①) →すると直線関係が認められた (Fig6.3 左) つまりこれ以上変数を変換する必要はない。
- ・ 中には、モデルをつくるのに、すべての変数を対数変換することを考える人もいるかもしれない。実際、そのほうが deviance が小さくなる。 (P128③)

<他の独立要因の影響を見たいとき>

- ・ 線形モデルでは、偏残差プロット (partial residual plot) を使って、他の独立変数の効果を評価した。
- ・ GLM でも、 $z - \hat{\eta} - \beta_j x_j$ と x_j (=偏残差プロット) との関係から、同様のことが出来る。
- ・ この場合、非線形でなかったので、変数変換をしなくてもよいし、明らかな外れ値がないことがわかる。 (Fig.6.3 左)
- ・ グラフを書かずに変数変換の有無を検証することも出来るが詳しくは 12 章で。

<リンク関数のチェック>

- ・ リンク関数を変えることは非常に困難を伴うが、はたしてこのリンク関数で正しいのかチェックすることも大切。
- ・ チェックをする前に、外れ値の処理や変数変換をしておこう。
- ・ 処理が終わったら、線形化応答変数 z と線形予測子 η の関係をプロットしてみる（結果は Fig.6.3 右 この場合は問題ない）
- ・ リンク関数チェックの別の方法は、同じ属のリンク関数を用いてみて調べること。→現在のリンク関数との fit 具合を調べてみる。詳細は 7.4 章で。

Unusual Points

- ・ グラフを使って外れ値を調べてみよう→GLM の場合 half-normal plot を使う。
- ・ これは、残差と half-normal 分布の変異点を比較する。 (P129①)

(ガラパゴスの例)

- ・ ジャックナイフ残差をプロットする (P130Fig6.4 左) と・・・プロットの中に外れ値は認められなかった。
- ・ ところで、half-normal plot は leverages や Cook 統計の診断にも役に立つ
- ・ まず、leverage を見てみた (P130Fig6.4 右)
- ・ すると、Santa Cruz 島のデータが外れ値であり、leverage を持っていることがわかった → Santa Cruz 島の Scrutz パラメータが 0 であるためだろう (log 変換したときに問題が生じるため、Scrutz パラメータには 0.1 を加えているが、そのため結果の解釈が非常に難しい)
- ・ また、Cook 統計の結果を Fig6.5 に示す。やはり Santa Cruz 島の結果が影響を与えている。
- ・ Cook 統計による係数の変化をしてみる。Case25 (Santa Cruz 島のデータが) 大きな影響を与えていることが分かる。
- ・ そこで、Santa Cruz 島のデータを除いたときのモデルと比較する。 (P130④)
- ・ 結果をみると、除く前と除いた後で Scrutz の係数の値が大きく違う (P131①)
- ・ この問題の解決策として、0.1 でなく 0.5 を足すということが考えられる (どうして??)
- ・ overdispersion は残っているが、最終モデルは P131② のとおり。
- ・ 3.1 章で使ったモデルとは異なり、deviance は小さく、elevation term は入っていない。