一般化線形モデル入門の入門

第 6 回 Armitage 勉強会

土居正明

1　はじめに

1.1　本稿の内容

本稿では、「一般化線形モデルとは何か?」についてご説明します。モデルの具体例をいくつかご紹介した後、数値例を用いて推定の方法と、SAS による実行まで軽くご説明します。

1.2　用語の確認

本稿でよく出てくる用語を二つだけ確認しておきます。まず、パラメータ μ_1, μ_2, μ_3 の**一次結合** とは、定数 a_1, a_2, a_3 に対して、

$$a_1\mu_1 + a_2\mu_2 + a_3\mu_3$$

のように、μ_1, μ_2, μ_3 に**定数をかけて足し算したもの**です。μ_1, μ_2, μ_3 に関して二乗したり三乗したり、log を考えたりしてはいけませんが、a_1, a_2, a_3 はただの定数ですので二乗したりしてもよいです。つまり、

$$a_1\mu_1 + a_1^2\mu_2 + a_1^3\mu_3$$

のようなものも、μ_1, μ_2, μ_3 に関しては 一次結合になっています。この点は回帰分析モデルで重要となりますので、しっかり押さえておいてください。

次に、**応答変数**という言葉です。これは「結果変数」「反応変数」などと同義[*1]で、要は 評価項目のデータ のことです。今回「データ」と言うと応答変数以外にも共変量（説明変数）の値も含まれますので、区別するために言葉を分けます。

2　（一般）線形モデルと一般化線形モデル

2.1　一般線形モデルとその一般化

一般化線形モデルのお話をするためには、まず（一般）線形モデル についての解説が必要となります。具体例として、応答変数 y が正規分布に従う「t 検定のモデル」「回帰分析」「分散分析」「共分散分析」を頭に入れておいてください。特徴は、

- **応答変数 y が正規分布に従う**
- **y の平均がパラメータの一次結合で表わされる**

の 2 つです。そして、これを一般化させたのが 一般**化**線形モデル です。これは

- **応答変数 y が正規分布、二項分布、Poisson 分布など（厳密には「指数型分布族」に含まれる分布）に従う**
- **y の平均（や確率）をある関数で変化させたら、パラメータの一次結合で表わされる**

という特徴を持ちます。

[*1] まったく同じ意味かどうかは知りませんが、同じと思っておいてそれほど大きな害はないと思います。

まとめとして、比較の表を示しておきましょう。

	応答変数	平均（や確率）の**構造**
（一般）線形モデル	正規分布	パラメータの一次結合
一般化線形モデル	正規分布, 二項分布, Poisson 分布など（指数型分布族）	関数で変形するとパラメータの一次結合になる

というものです。今の段階ではよくわからないものもあるかもしれませんが、以下具体的にご説明していきます。

3 「構造」とは何か？

さて、上の表で「平均（や確率）の**構造**」という言葉に引っかかった方も多いかと思います。この点について、まずは整理していきましょう。

要は、平均（や確率）が「全員同じ」ではなくて、「性別・年齢などの（共変量の）値によって変わる」ようなモデルを考える、ということです。

3.1 例 1：「全員同じ」モデル（1 群の t 検定）

たとえば、1 群の t 検定を用いる場合、応答変数の分布は

$$y_1, \cdots, y_n \ \sim \ N(\mu, \sigma^2)$$

となります。

書き換えると

$$y_i = \mu + \epsilon_i \quad (\epsilon_i \ \sim \ N(0, \sigma^2))$$

となります。平均にだけ注目しますと、

$$E[y_i] = \mu$$

です。これは、**「平均は全員等しい」**という構造が入っていることを意味しています[*2]。

3.2 例 2：「平均が個人個人で違う」モデル

例 1 の「全員同じ」モデルが最も基本のモデルとなりますが、我々の扱うデータはそう簡単にすむ場合だけではありません。たとえば、一般には塩分摂取量が多くなると血圧が増加します。このような場合、個人個人の血圧の平均が全員同じと想定するのは無理があります。それよりも、**「個人個人の血圧は、塩分摂取量に依存して変化する」というモデルの方が妥当**になるでしょう。そのような場合、被験者 i のある測定時点での血圧を y_i、その人の平均的な 1 日の塩分摂取量を x_i とおくと、たとえば

$$y_i = \alpha + \beta x_i + \epsilon_i \quad (\epsilon_i \ \sim \ N(0, \sigma^2), \ \alpha, \beta \text{ は定数})$$

のように想定するほうが妥当でしょう[*3]。y_i の平均で考えますと、

$$E[y_i] = \alpha + \beta x_i$$

となります。これより、このモデルでは**「1 人 1 人の血圧の平均値はそれぞれの塩分摂取量によって変化する」**ということを表わしています。

[*2] つまり、「個人ごとに値が異なるのは全て誤差 ϵ_i のせい」ということです。

[*3] 別に x_i の一次関数ではなくて、二次関数や指数関数などを考えることもできますが、今回は一次関数に絞ってお話します。

3.3 例 3：「確率が個人個人で違う」モデル

では次に、確率が個人個人で違うモデルを考えて見ましょう。これは、ある病気の発生率が血圧に依存している場合 を考えてみましょう。被験者 i がある病気にかかる確率を p_i として、その人の血圧（共変量）を x_i としましょう。このとき、x_i が p_i に依存することをモデル化したいとします。このときたとえば、

$$p_i = \frac{1}{1+\exp(-\alpha-\beta x_i)} \qquad (\alpha, \beta \text{ は定数})$$

というモデルを考えることがよくあります。ここで、関数 $y = \frac{1}{1+\exp(-\alpha-\beta x)}$ には**ロジスティック関数**という名前がついています。ロジスティック関数は奇妙な形をしていますが、どうしてこのような関数を考えるかを見るために、試しに $y = \frac{1}{1+\exp(-x)}$ のグラフを描いてみましょう[*4]。

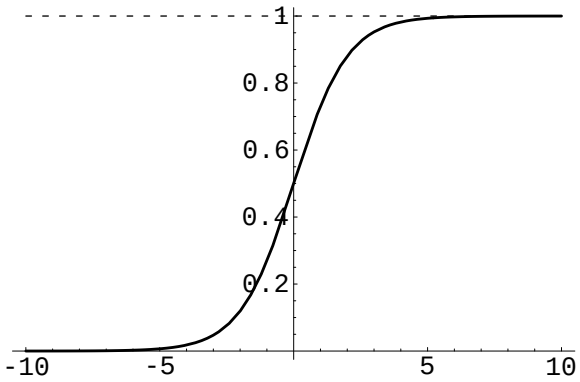

図 1　$y = \frac{1}{1+\exp(-x)}$ のグラフ

このように、血圧が増えれば増えるほど発生率が上がっていく様子が表せています。さらに、x の値がいくつであっても $y = \frac{1}{1+\exp(-x)}$ の値は常に 0 から 1 になります[*5]。確率 p_i は常に 0〜1 のどこかの値をとりますので、ロジスティック関数は連続的な共変量を用いて 確率 に構造を入れるのに適しています。このような、ロジスティック関数を用いた回帰分析のモデルは**ロジスティック回帰モデル**と呼ばれます[*6]。

4 （一般）線形モデルと一般化線形モデルの具体例

では、（一般）線形モデルと一般化線形モデルの 2 つの具体例をいくつか挙げていくことにしましょう。以下、共変量 x_i は連続値をとるデータ（既知）とし、$\alpha, \beta, \mu, \mu_i$ などをパラメータ（未知）とします。

4.1 （一般）線形モデル

以下の全てのモデルは、「応答変数 y が正規分布に従う」「平均がパラメータの一次結合」を満たします。つまり、（一般）線形モデルとなっています。

[*4] つまり、$\alpha = 0,\ \beta = 1$ のときです。

[*5] 一般形 $y = \frac{1}{1+\exp(-\alpha-\beta x)}$ は、$y = \frac{1}{1+\exp(-x)}$ を x 軸方向 に**「平行移動したり伸縮させたり」**しただけですので、y 軸方向については変わらず 0〜1 の値をとることなります。

[*6] なお、分子分母に $\exp(\alpha+\beta x)$ をかけると

$$\frac{1}{1+\exp(-\alpha-\beta x)} = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$$

となります。こちらの表記の方になじんでいる方も多いかもしれませんが、同じ式を指しています。

4.1.1 対応のない t 検定・分散分析モデル

最初に、対応のない t 検定と分散分析のモデルについて考えます。t 検定と分散分析（一元配置・fixed effect モデル）は、群の数が 2 群か 3 群以上かの違いしかありませんので、同じように扱います。i を投与群、j を各投与群における症例番号としますと、モデルは

$$y_{ij} \ \sim \ N(\mu_i, \sigma^2)$$

となります。書き換えると

$$y_{ij} = \mu_i + \epsilon_{ij} \quad (\epsilon_{ij} \ \sim \ N(0, \sigma^2))$$

です。y_{ij} の平均を考えますと、

$$E[y_{ij}] = \mu_i$$

となりますので、「平均はパラメータ μ_i（の一次結合）[*7]」「応答変数 y_{ij} は正規分布に従う」ので、これは**（一般）線形モデル**となります[*8]。

二元配置以降の場合も同様に成り立ちますが、特にもう一つだけ取り上げておきます。「二元配置・fixed effect モデル（交互作用あり）」のモデルです。

$$y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij} \quad (\epsilon_{ij} \ \sim \ N(0, \sigma^2))$$

とおきますと、平均は

$$E[y_{ij}] = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

より、「平均はパラメータ $\mu, \alpha_i, \beta_j, \gamma_{ij}$ の一次結合」で、「応答変数 y_{ij} は正規分布に従う」ので、これも（一般）線形モデルになります。

4.1.2 回帰分析モデル

次に、回帰分析のモデルを考えます。

$$y_i = \alpha + \beta x_i + \epsilon_i \quad (\epsilon_i \ \sim \ N(0, \sigma^2))$$

となります[*9]。平均に注目すると、

$$E[y_i] = \alpha + \beta x_i$$

ですので、「平均はパラメータ $\alpha,\ \beta$ の一次結合」「応答変数 y_i は正規分布に従う」ので、これも**（一般）線形モデル**の例です。

また、共変量が 2 つ以上に増えて

$$E[y_i] = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}$$

となっても、やはり（一般）線形モデルです。

次に、一次結合の定義のところで述べましたが、共変量のべき乗を含む

$$E[y_i] = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k$$

のような形も、（一般）線形モデルになります。先にも述べましたが、今回は**「主役はパラメータ $\alpha,\ \beta_1, \cdots, \beta_k$」**なので、$\alpha,\ \beta_1, \cdots, \beta_k$ について一次結合なら、(一般) 線形モデル になるのです。

[*7] 平均は $E[y_{ij}] = 1 \cdot \mu_i$ となり、これも一次結合に含まれます。

[*8] 平均の部分は、$\mu_i = \mu + \alpha_i$ と変形して、

$$E[y_{ij}] = \mu + \alpha_i$$

のように表現することもありますが、こちらにしても「平均はパラメータ μ, α_i の一次結合」つまり、$E[y_{ij}] = 1 \cdot \mu + 1 \cdot \alpha_i$ となっています。

[*9] 別の書き方では、$y_i \sim N(\alpha + \beta x_i, \sigma^2)$ とも書けます。

4.1.3 共分散分散分析モデル

では次に、共分散分析モデルです。共分散分析モデルは分散分析と回帰分析を足し合わせたものですので、この 2 つが（一般）線形モデルであることから共分散分析のモデルも（一般）線形モデルになるであろうことは容易に想像していただけると思います。とはいえ、きちんと見ていくことにします。第 i 群の被験者 j の共変量を x_{ij} としますと

$$y_{ij} = \mu + \alpha_i + \beta x_{ij} + \epsilon_{ij} \quad (\epsilon_{ij} \ \sim \ N(0, \sigma^2))$$

となります。ここで、薬剤の影響が α_i、共変量 x_{ij} の影響が β で表わされています。このとき、平均は

$$E[y_{ij}] = \mu + \alpha_i + \beta x_{ij}$$

となります。これより、「平均はパラメータ μ, α_i, β の一次結合」「応答変数 y_i は正規分布に従う」ので、**（一般）線形モデル**です。

4.2 一般化線形モデル

では次に、一般化線形モデルです。最初に整理したことをもう一度思い出しますと、

・応答変数を正規分布から、より一般の指数型分布族に含まれる分布へ

・平均（や確率）の構造を「パラメータの一次結合」から「関数で変換したらパラメータの一次結合」へ

拡張したものが、一般化線形モデルです。

4.2.1 ロジスティック回帰

例 3 で扱いました、ロジスティック回帰モデルについて考えます。応答変数 y_i は被験者 i の疾病の発生の有無を示す $0, 1$ の二値関数（発生は $y_i = 1$）で、発生確率は p_i（個人ごとに異なる）とします。このとき、y_i はの Bernoulli 分布 $Be(p_i)$ に従うとできます[*10]。このとき、$h(x) = \frac{1}{1+\exp(-x)}$ とおくと

$$\begin{aligned} y_i \ &\sim \ Be(p_i) \\ p_i &= h(\alpha + \beta x_i) \end{aligned} \tag{1}$$

のように表せます。つまり、関数 $h(x)$ の中身が一次関数 という形です。

さて、ここで (1) を「$= \alpha + \beta x_i$」の形になるように書き直してやります。$h(x) = \frac{1}{1+\exp(-x)}$ の逆関数は $h^{-1}(x) = \log\left(\frac{x}{1-x}\right)$ ですので[*11]、結局

$$h^{-1}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta x_i$$

となります。このとき、$h^{-1}(x) = \log\left(\frac{x}{1-x}\right)$ のように**確率（や平均）をパラメータの一次結合に変形する関数**を リンク関数 と呼びます。また、今回の関数 $\log\left(\frac{x}{1-x}\right)$ には ロジット関数 という名前がついています[*12]。これより、このモデルは

[*10] 具体例の計算のところで詳しく書きますが、Bernoulli 分布とは、「全体の人数が 1 人の二項分布」です。今回のように「1 人 1 人別々の分布に従う二値データ」に対して用いられます。

[*11] 逆関数の求め方は、まず $y = \frac{1}{1+\exp(-x)}$ を $x =$ になるように変形して、

$$\begin{aligned} y = \frac{1}{1+\exp(-x)} \quad &\Longleftrightarrow \quad \frac{1}{y} = 1 + \exp(-x) \quad \Longleftrightarrow \quad \frac{1}{y} - 1 = \exp(-x) \quad \Longleftrightarrow \quad \log\left(\frac{1}{y} - 1\right) = -x \\ &\Longleftrightarrow \quad \log\left(\frac{1-y}{y}\right) = -x \quad \Longleftrightarrow \quad \log\left(\frac{y}{1-y}\right) = x \quad \Longleftrightarrow \quad x = \log\left(\frac{y}{1-y}\right) \end{aligned}$$

とします。そして、最後の式の x と y を入れ替えた $y = \log\left(\frac{x}{1-x}\right)$ が、$y = \frac{1}{1+\exp(-x)}$ の逆関数となります。

[*12] $\mathrm{logit}(x) = \log\left(\frac{x}{1-x}\right)$ という書き方をする場合もあります。なお、「リンク関数」は「確率（や平均）をパラメータの一次結合に変形する関数」一般を指す名前であり、「ロジット関数」は $\mathrm{logit}(x) = \log\left(\frac{x}{1-x}\right)$ という一つの関数の名前です。ですので今回は「リンク関数がロジット関数である」という言い方になります。

・応答変数 y_i は Bernoulli 分布に従う[*13]

・リンク関数はロジット関数 $(\mathrm{logit}(p_i) = \alpha + \beta x_i)$

の一般化線形モデルとなります。

4.2.2 プロビット回帰

次に、同じく確率をモデル化する場合で、今度はロジスティック関数の代わりに標準正規分布の累積分布関数 $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt$ を持ってきます。つまり、応答変数の分布はそのままで、

$$\begin{aligned} y_i &\sim Be(p_i) \\ p_i &= \Phi(\alpha + \beta x_i) \end{aligned}$$

としたものを**プロビット回帰モデル**と呼びます。$\Phi(x)$ の逆関数 $\Phi^{-1}(x)$ を持ってきて[*14]、

$$\Phi^{-1}(p_i) = \alpha + \beta x_i$$

のように書きますとパラメータ α, β の一次結合になりますので、$\Phi^{-1}(x)$ がリンク関数です。従って、これもやはり一般化線形モデルになります。

・応答変数 y_i は Bernoulli 分布に従う

・リンク関数は $\Phi^{-1}(x)$

です。

4.2.3 Poisson 回帰モデル

次は、Poisson 回帰モデルです。上の 2 つでは、応答変数は発生のありなしの二値でしたが、今度は 0 から始まる計数データの場合です。ある臨床試験で、第 i 番目の施設において特定の有害事象が発生した被験者の数 y_i が平均 λ_i の Poisson 分布 $Po(\lambda_i)$ に従うとします。そしてこの λ_i が、ある共変量 x_i を用いて、$\lambda_i = e^{\alpha+\beta x_i}$ と表わせるとします。両辺の log をとると $\log \lambda_i = \alpha + \beta x_i$ となり、パラメータ α, β の一次結合になりますので、

$$\begin{aligned} y_i &\sim Po(\lambda_i) \\ \log \lambda_i &= \alpha + \beta x_i \end{aligned}$$

となり、一般化線形モデルであることがわかります。

・応答変数 y_i は Poisson 分布に従う[*15]

・リンク関数は $\log x$

です。

5 パラメータの推定：最尤法

ではロジスティック回帰モデルを用いた具体例を通して、パラメータの推定方法について見ていきましょう。最尤法を用いた計算方法をご紹介します[*16]。

5.1 データと統計モデル

5 人の被験者のある疾病の発生の有無を表わす二値応答変数 (y_i) と、共変量である収縮期血圧 (x_i) のデータです $(i = 1, \cdots, 5)$。

[*13] Bernoulli 分布は指数型分布族に入っています。

[*14] 面倒ですので式の形は省略します。

[*15] Poisson 分布も指数型分布族に入っています。

[*16] 最尤法について不安な方は、「計算特訓」に最尤法の問題が 3 ファイルありますので、それを済ませてから以下をお読みください。

疾病の発生：y_i （あり：1 , なし：0）	収縮期血圧：x_i (mmHg)
0	120
1	130
1	140
0	150
1	160

表 1　ある疾病の発生と収縮期血圧のデータ

ここで、疾病の発生の確率は収縮期血圧に依存して変わるものとします。y_i はそれぞれ 0 か 1 になりますが、全員の発生確率が異なりますので、「二項分布で 5 人いっぺんに」考えることはできません[*17]。そこで、1 人 1 人の発生確率を p_i とおくと、各人の発生確率が異なりますので、1 人 1 人別々の二項分布、つまり Bernoulli 分布を考えます。このとき、確率 p_i でこの疾病が発生することを示す Bernoulli 分布を $Be(p_i)$ で表すとしますと、

$$y_1 \sim Be(p_1),\ \ y_2 \sim Be(p_2),\ \ y_3 \sim Be(p_3),\ \ y_4 \sim Be(p_4),\ \ y_5 \sim Be(p_5)$$

となります。これより、各 y_i の確率関数は

$$f(y_i|p_i) = p_i^{y_i}(1-p_i)^{1-y_i} \qquad (i = 1, \cdots, 5)$$

となります。5 人全員分を書き下しますと、

$$f(y_1|p_1) = p_1^{y_1}(1-p_1)^{1-y_1},\ \ f(y_2|p_2) = p_2^{y_2}(1-p_2)^{1-y_2},\ \ f(y_3|p_3) = p_3^{y_3}(1-p_3)^{1-y_3},$$

$$f(y_4|p_4) = p_4^{y_4}(1-p_4)^{1-y_4},\ \ f(y_5|p_5) = p_5^{y_5}(1-p_5)^{1-y_5}$$

となります。

次に、確率 p_i に対して、ロジスティック回帰モデル

$$p_i = \frac{1}{1+\exp(-\alpha-\beta x_i)} = \frac{\exp(\alpha+\beta x_i)}{1+\exp(\alpha+\beta x_i)}$$

を考えます。5 人分を書き下し、表 1 の x_i の値を代入しますと、

$$p_1 = \frac{\exp(\alpha+\beta x_1)}{1+\exp(\alpha+\beta x_1)} = \frac{\exp(\alpha+120\beta)}{1+\exp(\alpha+120\beta)},\ \ p_2 = \frac{\exp(\alpha+\beta x_2)}{1+\exp(\alpha+\beta x_2)} = \frac{\exp(\alpha+130\beta)}{1+\exp(\alpha+130\beta)}$$

$$p_3 = \frac{\exp(\alpha+\beta x_3)}{1+\exp(\alpha+\beta x_3)} = \frac{\exp(\alpha+140\beta)}{1+\exp(\alpha+140\beta)},\ \ p_4 = \frac{\exp(\alpha+\beta x_4)}{1+\exp(\alpha+\beta x_4)} = \frac{\exp(\alpha+150\beta)}{1+\exp(\alpha+150\beta)}$$

$$p_5 = \frac{\exp(\alpha+\beta x_5)}{1+\exp(\alpha+\beta x_5)} = \frac{\exp(\alpha+160\beta)}{1+\exp(\alpha+160\beta)}$$

となります。モデルをまとめて書いておきますと、

$$\begin{aligned} y_i &\sim Be(p_i) \\ p_i &= \frac{\exp(\alpha+\beta x_i)}{1+\exp(\alpha+\beta x_i)} \end{aligned} \qquad \left(\iff \log\left(\frac{p_i}{1-p_i}\right) = \alpha+\beta x_i \iff \mathrm{logit}(p_i) = \alpha+\beta x_i\right)$$

となります。

[*17] 二項分布は「**同じ発生確率の人**が n 人いた場合、その中で x 人が発生する確率」を表現するものですので、今回のモデルでは 1 人 1 人別々にしか使えません。そして、1 人に対する二項分布を Bernoulli 分布と呼びます。

さて、我々の興味があるのは

・各人の発生率 p_i がどのくらいか？

です。いま、その p_i に今回ロジスティック回帰モデルを当てはめていますので、結局

・パラメータ $\alpha,\ \beta$ の値はいくつか？

が分かればよいことになります[*18]。つまり、これからやることは**「最尤法を用いて $\alpha,\ \beta$ を推定する」**ことです。

5.2 最尤法

では実際に最尤法を用いて、パラメータ $\alpha,\ \beta$ の値を推定していきましょう。

5.2.1 式による計算

まず最初に式による計算をしていきます。尤度関数をまず p_i で表わし、その後各 p_i にロジスティック関数の値を代入しますと、

$$
\begin{aligned}
&L(\alpha,\beta|y_1,\cdots,y_5)\\
&=\prod_{i=1}^{5} f(y_i|p_i)\\
&=f(y_1|p_1)\cdot f(y_2|p_2)\cdot f(y_3|p_3)\cdot f(y_4|p_4)\cdot f(y_5|p_5)\\
&=(1-p_1)\cdot p_2\cdot p_3\cdot(1-p_4)\cdot p_5 \qquad (\because y_1=0,\ y_2=1,\ y_3=1,\ y_4=0,\ y_5=1)\\
&=\left(1-\frac{\exp(\alpha+120\beta)}{1+\exp(\alpha+120\beta)}\right)\cdot\left(\frac{\exp(\alpha+130\beta)}{1+\exp(\alpha+130\beta)}\right)\\
&\quad\cdot\left(\frac{\exp(\alpha+140\beta)}{1+\exp(\alpha+140\beta)}\right)\cdot\left(1-\frac{\exp(\alpha+150\beta)}{1+\exp(\alpha+150\beta)}\right)\cdot\left(\frac{\exp(\alpha+160\beta)}{1+\exp(\alpha+160\beta)}\right)\\
&=\frac{1}{1+\exp(\alpha+120\beta)}\cdot\frac{\exp(\alpha+130\beta)}{1+\exp(\alpha+130\beta)}\cdot\frac{\exp(\alpha+140\beta)}{1+\exp(\alpha+140\beta)}\cdot\frac{1}{1+\exp(\alpha+150\beta)}\cdot\frac{\exp(\alpha+160\beta)}{1+\exp(\alpha+160\beta)}
\end{aligned}
$$

[*18] 特に、β の値は「収縮期血圧が発生率に与える影響」を表現します。従って、「血圧はこの疾病の発生に関係ないかどうか」を検定したければ、$H_0:\beta=0$ として検定すればよいことになります。

となります。これより、対数尤度関数は

$$
\begin{aligned}
& l(\alpha,\beta|y_1,\cdots,y_5) \\
&= \log L(\alpha,\beta|y_1,\cdots,y_5) \\
&= \log\left(\frac{1}{1+\exp(\alpha+120\beta)}\cdot\frac{\exp(\alpha+130\beta)}{1+\exp(\alpha+130\beta)}\cdot\frac{\exp(\alpha+140\beta)}{1+\exp(\alpha+140\beta)}\cdot\frac{1}{1+\exp(\alpha+150\beta)}\cdot\frac{\exp(\alpha+160\beta)}{1+\exp(\alpha+160\beta)}\right) \\
&= \log\left(\frac{1}{1+\exp(\alpha+120\beta)}\right)+\log\left(\frac{\exp(\alpha+130\beta)}{1+\exp(\alpha+130\beta)}\right)+\log\left(\frac{\exp(\alpha+140\beta)}{1+\exp(\alpha+140\beta)}\right) \\
&\quad +\log\left(\frac{1}{1+\exp(\alpha+150\beta)}\right)+\log\left(\frac{\exp(\alpha+160\beta)}{1+\exp(\alpha+160\beta)}\right) \\
&= -\log\left(1+\exp(\alpha+120\beta)\right)+\{\log\left(\exp(\alpha+130\beta)\right)-\log\left(1+\exp(\alpha+130\beta)\right)\}+\{\log\left(\exp(\alpha+140\beta)\right) \\
&\quad -\log\left(1+\exp(\alpha+140\beta)\right)\}-\log\left(1+\exp(\alpha+150\beta)\right)+\{\log\left(\exp(\alpha+160\beta)\right)-\log\left(1+\exp(\alpha+160\beta)\right)\} \\
&= -\log\left(1+\exp(\alpha+120\beta)\right)+(\alpha+130\beta)-\log\left(1+\exp(\alpha+130\beta)\right)+(\alpha+140\beta) \\
&\quad -\log\left(1+\exp(\alpha+140\beta)\right)-\log\left(1+\exp(\alpha+150\beta)\right)+(\alpha+160\beta)-\log\left(1+\exp(\alpha+160\beta)\right) \\
&= -\log\left(1+\exp(\alpha+120\beta)\right)-\log\left(1+\exp(\alpha+130\beta)\right)-\log\left(1+\exp(\alpha+140\beta)\right)-\log\left(1+\exp(\alpha+150\beta)\right) \\
&\quad -\log\left(1+\exp(\alpha+160\beta)\right)+3\alpha+430\beta
\end{aligned}
$$

となります。では、α, β で微分していきます。まず、α では、

$$
\begin{aligned}
& \frac{\partial l}{\partial\alpha}(\alpha,\beta|y_1,\cdots,y_5) \\
&= -\frac{\exp(\alpha+120\beta)}{1+\exp(\alpha+120\beta)}-\frac{\exp(\alpha+130\beta)}{1+\exp(\alpha+130\beta)}-\frac{\exp(\alpha+140\beta)}{1+\exp(\alpha+140\beta)}-\frac{\exp(\alpha+150\beta)}{1+\exp(\alpha+150\beta)}-\frac{\exp(\alpha+160\beta)}{1+\exp(\alpha+160\beta)}+3
\end{aligned}
$$

です。次に、β では、

$$
\begin{aligned}
& \frac{\partial l}{\partial\beta}(\alpha,\beta|y_1,\cdots,y_5) \\
&= -\frac{120\exp(\alpha+120\beta)}{1+\exp(\alpha+120\beta)}-\frac{130\exp(\alpha+130\beta)}{1+\exp(\alpha+130\beta)}-\frac{\exp(\alpha+140\beta)}{1+\exp(\alpha+140\beta)}-\frac{150\exp(\alpha+150\beta)}{1+\exp(\alpha+150\beta)}-\frac{160\exp(\alpha+160\beta)}{1+\exp(\alpha+160\beta)}+430
\end{aligned}
$$

となります。さて、最尤推定値を求めるには尤度方程式

$$
\begin{cases}
\frac{\partial l}{\partial\alpha}\left(\widehat{\alpha},\widehat{\beta}|y_1,\cdots,y_5\right) &= 0 \\
\frac{\partial l}{\partial\beta}\left(\widehat{\alpha},\widehat{\beta}|y_1,\cdots,y_5\right) &= 0
\end{cases}
$$

を解けばよいことになります。つまり、連立方程式

$$
\begin{aligned}
& -\frac{\exp(\widehat{\alpha}+120\widehat{\beta})}{1+\exp(\widehat{\alpha}+120\widehat{\beta})}-\frac{\exp(\widehat{\alpha}+130\widehat{\beta})}{1+\exp(\widehat{\alpha}+130\widehat{\beta})}-\frac{\exp(\widehat{\alpha}+140\widehat{\beta})}{1+\exp(\widehat{\alpha}+140\widehat{\beta})}-\frac{\exp(\widehat{\alpha}+150\widehat{\beta})}{1+\exp(\widehat{\alpha}+150\widehat{\beta})}-\frac{\exp(\widehat{\alpha}+160\widehat{\beta})}{1+\exp(\widehat{\alpha}+160\widehat{\beta})}+3=0 \\
& -\frac{120\exp(\widehat{\alpha}+120\widehat{\beta})}{1+\exp(\widehat{\alpha}+120\widehat{\beta})}-\frac{130\exp(\widehat{\alpha}+130\widehat{\beta})}{1+\exp(\widehat{\alpha}+130\widehat{\beta})}-\frac{\exp(\widehat{\alpha}+140\widehat{\beta})}{1+\exp(\widehat{\alpha}+140\widehat{\beta})}-\frac{150\exp(\widehat{\alpha}+150\widehat{\beta})}{1+\exp(\widehat{\alpha}+150\widehat{\beta})}-\frac{160\exp(\widehat{\alpha}+160\widehat{\beta})}{1+\exp(\widehat{\alpha}+160\widehat{\beta})}+430=0
\end{aligned}
$$

を解くことにとなります。これを解けば最尤推定量 $\widehat{\alpha}, \widehat{\beta}$ が求まります。しかし、これを数式で厳密に解くのは非常に大変なのです[*19]。そこで、**きっちり式で解くことはあきらめて数値的に（近似計算で）解く**ことにします。ですので、式での計算はここまでにして、続きは SAS に譲ることにします。

5.2.2 SAS による計算 1：proc logistic

では今の計算を SAS にやらせてみましょう。まずデータは

```
data d1;
  input y x;
  cards;
  0  120
  1  130
  1  140
  0  150
  1  160
run;
```

です。ロジスティック回帰モデルの解析には proc logistic と proc genmod の 2 つが使えます。proc logistic はロジスティック回帰に特化したものですが、proc genmod はプロビット回帰なども含む一般化線形モデルを統一的に扱うプロシジャです。本稿では両方ご紹介しますが、まずは proc logistic による解析プログラムと結果をご紹介します。プログラムは

```
proc logistic data=d1 descending;
  model y=x;
run;
```

です[*20]。結果の主要な部分は以下のようになります。

モデルの詳細	
データセット	WORK.D1
応答変数	y
応答の水準数	2
モデル	binary logit
最適化手法	Fisher's scoring

「モデル」のところの「binary logit」は、「応答変数 y_i の分布が二項分布（Bernoulli 分布も含みます）」「リンク関数はロジット関数」を指しています。また、「最適化手法」の「Fisher's scoring」というのは、先に示しました $\widehat{\alpha}, \widehat{\beta}$ を求めるための数値計算（近似計算）の方法の名前です[*21]。

[*19] 出来るのかどうかも分かりません。偶然解ける可能性もあるかもしれませんが、多分解けないのではないかと思います。5〜10 分くらい解こうと努力しておくことはお勧めします。

[*20] すぐ後に述べますが、"descending"は、$y = 0, 1$ の二値のときに、「$y = 1$（大きい方）の確率をモデル化する」ための指定です。デフォルトは「$y = 0$（小さい方）の確率をモデル化」します。次の、proc genmod でも同様です。

[*21] つまり、あくまで「推定のやり方」は**最尤法**です。そして、「最尤推定値を求めるための数値計算の方法」に Fisher's scoring 法を用いている、ということです。

モデルの確率基準は *y=1* です。

今回はある疾病の発生確率を考えていますので、「$y=1$（大きい方）となる確率」をモデル化する必要があります。デフォルトでは、「$y=0$（小さい方）となる確率」をモデル化してしまいますので、今回のように「発生：1、発生しない：0」の場合は、プログラムに **descending** と指定して調整してやる必要があります。

なお、この部分、英語版では "Probability modeled is $y=1$" と書かれています。個人的には「$y=1$ となる確率をモデル化しています」と訳す方が分かりやすいと思います。

モデル収束状態
収束基準 (GCONV=1E-8) **は満たされました。**

これは、先の「手で出来ない計算」を**「きちんと数値的に解けました」**というメッセージです。さりげないですが、大変重要なメッセージです。計算がうまくいかない例については、補足でご説明します。

モデルの適合度統計量		
基準	**切片のみ**	**切片と共変量**
AIC	8.730	10.302
SC	8.340	9.521
-2 Log L	6.730	6.302

ここは、モデルの当てはまり具合を見て、どのモデルが適当か（どの共変量を入れるべきか）などの参考にしますが、今回は深入りはしません[*22]。

H0: BETA=0 の検定			
検定	**カイ 2 乗**	**自由度**	**Pr > ChiSq**
尤度比	0.4277	1	0.5131
Score	0.4167	1	0.5186
Wald	0.3947	1	0.5298

先ほど注で少しだけ触れましたが、ロジスティック回帰モデル $p_i=\frac{1}{1+\exp(-\alpha-\beta x_i)}$ において、「$H_0:\beta=0$」の検定とは、**「共変量 x_i の影響があるかどうかの検定」**になります[*23]。今回はどの検定を用いても有意差がないですが、例数が少なすぎますので、この結果からは「共変量の影響があるとはいえない。けど、ないとも言えない。」くらいに、だいぶ謙虚（優柔不断？）な発言に留めておくのがよいでしょう[*24]。

最大尤度推定値の分析					
パラメータ	**自由度**	**推定値**	**標準誤差**	**Wald カイ 2 乗**	**Pr > ChiSq**
Intercept	1	-5.7094	9.7211	0.3450	0.5570
x	1	0.0439	0.0700	0.3947	0.5298

[*22] 今回はデータ数が極端に少ないですし、結構無理矢理ロジスティックモデルを当てはめているところもあります。今回ここに深入りすると「あまりあてはまりがよくない」ということがばれてしまいます…。とりあえず、あくまで「計算例」と考えて先に進むことにします。

[*23] $\beta=0$ なら $p_i=\frac{1}{1+\exp(-\alpha)}$ となり、共変量 x_i に依存しなくなります。

[*24] 有意差がつかない場合に「ないとも言えない」と判断するのは普通ですが、それをあえて口にするあたりで謙虚さを表現しているつもりです。

これはパラメータの推定値です。「推定値」の部分が先の尤度方程式における近似解です。ロジスティックモデル

$$p_i = \frac{1}{1+\exp(-\alpha-\beta x_i)} \quad \Longleftrightarrow \quad \mathrm{logit}(p_i) = \alpha + \beta x_i$$

の右側の表現から、Intercept（切片項）が α、x の影響が β で表わされていることが分かります。これより、表中の Intercept の行の推定値が $\widehat{\alpha}$ に、x の行の推定値が $\widehat{\beta}$ となり、

$$\begin{cases} \widehat{\alpha} &=& -5.7094 \\ \widehat{\beta} &=& 0.0439 \end{cases}$$

となります[*25]。そして、x の部分の **Pr** > **Chisq** の部分は、「$H_0 : \beta = 0$」の検定に対応しています。上の「$H_0 : \beta = 0$ の検定」の Wald の欄の確率と同じになっています。

これより、確率の推定値は

$$\widehat{p_i} = \frac{1}{1+\exp(5.7094-0.0439x_i)}$$

となります。血圧の値 $x_1 = 120, \cdots, x_5 = 160$ をそれぞれ代入していきますと、この疾病の被験者ごとの発生確率の推定値は

$$\widehat{p_1} = \frac{1}{1+\exp(5.7094-0.0439\cdot 120)} = 0.39, \quad \widehat{p_2} = \frac{1}{1+\exp(5.7094-0.0439\cdot 130)} = 0.50$$
$$\widehat{p_3} = \frac{1}{1+\exp(5.7094-0.0439\cdot 140)} = 0.61, \quad \widehat{p_4} = \frac{1}{1+\exp(5.7094-0.0439\cdot 150)} = 0.71$$
$$\widehat{p_5} = \frac{1}{1+\exp(5.7094-0.0439\cdot 160)} = 0.79$$

となります。では、血圧 x と発生確率の推定値の関係をプロットしてみましょう。

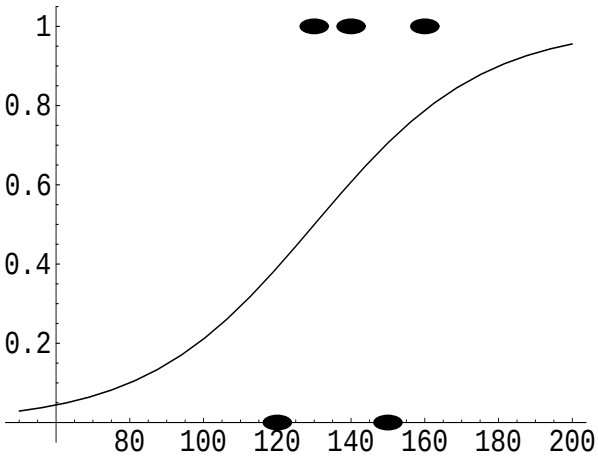

図 2　血圧と確率の推定値の関係 $y = \frac{1}{1+\exp(5.7094-0.0439x)}$ のグラフ（黒丸はデータ）

結構ずれていますね。あまりよいモデルではなさそうです。

オッズ比推定値			
変動因	**点推定値**	**95% Wald 信頼限界**	
x	1.045	0.911	1.198

[*25] 念のために断っておきますと、x の行の「推定値」は「x の推定値」ではなくて「x の影響（つまり β) の推定値」を意味しています。

これは、**「x が 1 増えたときと比べたオッズ比」**です。つまり、血圧が x のときの確率の推定値と $x+1$ のときの確率の推定値を持ってきて、

$$\widehat{p}(x)=\frac{1}{1+\exp(5.7094-0.0439x)},\quad \widehat{p}(x+1)=\frac{1}{1+\exp(5.7094-0.0439(x+1))}$$

とします。確率 p_i の構造 $\log\left(\frac{p_i}{1-p_i}\right)=\alpha+\beta x_i$ を思い出して、両辺の指数をとると $\frac{p_i}{1-p_i}=\exp(\alpha+\beta x_i)$ となり、オッズがでてきます。これよりそれぞれ

$$\widehat{\mathrm{odds}}(p(x))=\frac{\widehat{p}(x)}{1-\widehat{p}(x)}=\exp(-5.7094+0.0439x),\quad \widehat{\mathrm{odds}}(p(x+1))=\frac{\widehat{p}(x+1)}{1-\widehat{p}(x+1)}=\exp(-5.7094+0.0439(x+1))$$

となります[*26]。オッズ比 OR の推定値 $\widehat{OR}$ はオッズの推定値の比をとればよいですので、

$$\widehat{\mathrm{OR}}=\frac{\widehat{\mathrm{odds}}(p(x+1))}{\widehat{\mathrm{odds}}(p(x))}=\frac{\exp(-5.7094+0.0439(x+1))}{\exp(-5.7094+0.0439x)}=\frac{\exp(-5.7094+0.0439x)\cdot\exp(0.0439)}{\exp(-5.7094+0.0439x)}$$

$$=\exp(0.0439)=1.045$$

となります。計算の過程で x が消えてしまいますので、**「血圧がいくつであろうとも、1 増えたときのオッズ比は同じ」**となります。文字で書きますと、$\widehat{\mathrm{OR}}=\exp(\widehat{\beta})$ となります。

なお、あまり深入りはしませんが、「血圧が 1 上がるというのは、誤差の範囲であまり面白くない。例えば 10 上がったときのオッズ比はどうなる？」という風に考えられる方も多いかと思います。式の計算は、上の $x+1$ を $x+10$ にしていただくと、$\widehat{\mathrm{OR}}=\exp(10\widehat{\beta})$ になり、今回の場合は 1.552 になるのですが、SAS ではプログラムを、

```
proc logistic data=d1 descending;
   model y=x;
   units x=10;
run;
```

と変更すれば x が 10 増えたときのオッズ比 が出力されます。アウトプットは表示しませんが、「調整済みオッズ比」という欄が出力され、「単位」が 10、「推定値」が 1.552 となります。

5.2.3 SAS による計算 2：proc genmod

では次に、proc genmod を用いた解析です。プログラムは

```
proc genmod data=d1 descending;
   model y=x  /  link=logit dist=bin;
run;
```

です。model ステートメントで「リンク関数がロジット関数」「応答変数の分布が二項分布（Bernoulli 分布も含む）」を指定してやります。結果の主要な部分は、以下のようになります。

[*26] この部分、記号が煩雑になりますので、少し丁寧にご説明します。「オッズ比」は、「真の確率」p をベースにして求まる $\mathrm{odds}(p)=\frac{p}{1-p}$ を指します。ですので、これは「(p が分からない限り）絶対に知りえない値」です。**データを用いて計算できるのは、「オッズ比の推定値」**です。この推定値を $\widehat{\mathrm{odds}}(p)=\frac{\widehat{p}}{1-\widehat{p}}$ として求めるのです。このような記号の使い方は、本や資料ごとに異なる可能性がありますので、本や資料ごとに注意しながら読んでください。

モデルの詳細	
データセット	WORK.D1
分布	Binomial
リンク関数	Logit
従属変数	y

このあたりは、ほぼ proc logistic と同じです。

PROC GENMOD **によるモデルの確率基準は** *y='1'* **です。**

この部分も同じです。英語では ”PROC GENMOD is modeling the probability with $y =$’1’.” となります。やはり英語を直訳して「PROC GENMOD は $y = 1$ となる確率をモデル化しています」とする方が分かりやすいかと思います。なお、やはり **descending がなければ小さい方 (**$y = 0$**) をモデル化してしまいます**。

適合度の評価基準			
基準	**自由度**	**値**	**値/自由度**
残差	3	6.3024	2.1008
尺度化残差	3	6.3024	2.1008
Pearson カイ 2 乗	3	4.9694	1.6565
尺度化 Pearson X2	3	4.9694	1.6565
対数尤度		-3.1512	

これは、proc logistic における「モデルの適合度統計量」つまり、AIC などの部分に対応します。ここもやはり入り込むと厄介ですので詳しくはご説明しませんが、proc logistic の ”$-$2Log” は、ここの「対数尤度」の -2 倍であることと、**「残差」**のもとの英語は **deviance** であることくらいは書いておきます。

アルゴリズムは収束しました。

これは、「計算がきちんとできました」という意味です。これもうまくいかない例を補足で示します。

パラメータ推定値の分析							
パラメータ	**自由度**	**推定値**	**標準誤差**	**Wald 95% 信頼限界**		**カイ 2 乗値**	**Pr > ChiSq**
Intercept	1	-5.7094	9.7211	-24.7624	13.3435	0.34	0.5570
x	1	0.0439	0.0700	-0.0932	0.1811	0.39	0.5298
Scale	0	1.0000	0.0000	1.0000	1.0000		

Note:尺度パラメータは固定されています。

パラメータ推定値 $\widehat{\alpha},\ \widehat{\beta}$ は proc logistic の場合と全く同じです。「Scale」の部分（Scale）が増えていますが、これは overdispersion（過分散）を考慮したモデルを考える場合に必要になるところで、今回は無視していただいてかまいません。

6　補足 1：紛らわしい用語について

さて、本稿で見てきました

・一般線形モデル (General Linear Model)

・一般 化 線形モデル (Generalized Linear Model)

の 2 つは字面が大変よく似ています。たまに混同されることがありますので注意しておきます。

どちらも頭文字では GLM となりますが、一般に GLM と書きますと、一般 化 線形モデルの方を指します。しかし、SAS の ”Proc GLM” は、**（一般）線形モデル**を指しています。ですので、このような言葉遣いを用いる場合、**「Proc GLM では GLM は扱えない」**という、一見奇妙な現実に直面します。面倒ですが、慣用名なので仕方ありません。

文献によっては、一般線形モデルの「一般」を書かずに**線形モデル（Linear Model）**として、LM と略しているものもあります[*27]。そうすると、**「Proc GLM は LM を扱うもので GLM を扱うものではない」**という、また一見奇妙な文章が出来上がります。

7　補足 2：計算がうまくいかないとき

先の例では、proc logistic や proc genmod でそれぞれ「計算が収束しました」のようなメッセージがでました。これが「計算がきちんとうまくいきました」という意味だとはご説明しましたが、ここでは**計算がうまくいかないデータの例**についてご紹介しましょう。「完全分離」「擬似完全分離」の 2 つです[*28]。

7.1　完全分離

以下のようなデータを見てみます。

```
data d1;
  input y x;
  cards;
  0  120
  0  130
  0  140
  1  150
  1  160
run;
```

このデータの特徴は、たとえば $x = 145$ でデータを分けると $x < 145$ では必ず $y = 0$、$x \geq 145$ では必ず $y = 1$ となっています。このように、**共変量をある値以上とある値より小さいものの 2 つに分けたときに、y の値が片方は 1 のみ、もう片方は 0 のみとなる**ものを データが完全分離している といいます[*29]。

7.1.1　proc logistic による解析

このデータを proc logistic で解析すると、以下のメッセージが出ます。

[*27] 他の文献では（一般）線形モデルを GLM、一般化線形モデルを GLIM と略す場合もあるようですが、私は LM と GLM にする方が好みです。頭文字が同じなのが混乱の元だと思いますので。しかし、本によって異なる略称が使われることがある、ということを押さえていただくのは重要かもしれません。

[*28] といっても、「完全分離」は「擬似完全分離」の一部です。

[*29] 「ある値以下とある値より大きい」でも OK です。

モデル収束状態
データ点の完全分離が検出されました。

さらに、

WARNING：最尤推定量は存在しません。

WARNING：LOGISTIC プロシジャは上記の警告にもかかわらず継続します。最尤反復にもとづいて結果が表示されます。
モデルの当てはめの妥当性は疑わしいです。

というメッセージが出ます。要は**「推定値の欄に結果っぽい数値は表示されますが、信用しないでください」**ということです。ですので、推定値の値を見る前にこちらを必ずチェックしてください。

7.1.2 proc genmod による解析

次は、proc genmod の場合です。今度は

WARNING: 相対的な Hessian 収束基準 1.0868635408 は限界値 0.0001 を超えています。収束は疑わしいです。

となります。意味的には、proc logistic の出力と同じことです。やはり、**「推定値の欄に書いてある数値は信頼できません」**という意味です。

7.2 擬似完全分離

次に、擬似完全分離です。以下のようなデータです。

```
data d1;
  input y x;
  cards;
  0  120
  0  130
  1  130
  1  150
  1  160
run;
```

完全分離とかなり似ていますが、今度は $x = 130$ のときに $y = 0$ と $y = 1$ の両方があります。つまり、$x = 130$ で分けたときに、$x < 130$ では必ず $y = 0$、$x > 130$ では必ず $y = 1$。$x = 130$ のときは $y = 0, 1$ の両方あってよい という風になっています。このように、**「境界上では $y = 0, 1$ の両方を含んでよいけれど、それ以外は完全分離」**の場合、データが擬似完全分離している といいます[*30]。

7.2.1 proc logistic による解析

このとき、proc logistic では、

モデル収束状態
データ点の擬似完全分離が検出されました。

[*30] ですので、先の注でも見ましたとおり概念としては「完全分離」は「擬似完全分離」に含まれます。ですので、完全分離を擬似完全分離と言っても間違いではありません。しかし、通常「完全分離」の方が不都合が大きいので、完全分離の場合ははっきり完全分離と記載されます。つまり、SAS が「擬似完全分離」と言って来たら、「完全分離でない擬似完全分離」だと考えていただいて結構です。

というメッセージが出て、さらに、

WARNING：最尤推定量は存在していない可能性があります。

WARNING：LOGISTIC プロシジャは上記の警告にもかかわらず継続します。最尤反復にもとづいて結果が表示されます。
モデルの当てはめの妥当性は疑わしいです。

というメッセージが出ます。今度は最尤推定量が存在していない「可能性がある」と、完全分離よりはややマイルドな表現となっていますが、どういうときに最尤推定量が存在するのかについては、私はまだ知りません。ですので、現在の私にとっては「どちらも同じ」という悲しい結論になります[*31]。

7.2.2 proc genmod による解析

次に、proc genmod では、

WARNING: Hessian の負は正定値ではありません。

というメッセージがでます。このメッセージが出たら「擬似完全分離かも」と疑ってください[*32]。

[*31] ご存知の方、ぜひ教えてください。

[*32] 厳密に一致するのかどうかは、まだ知りません。今回の補足は歯切れが悪くてすみません。