データ解析のための統計モデリング (2008 年 10-11 月)

全 5 (+2) 回中の第 4 回 (2008–11–10)

# 一般化線形モデル (GLM) 2

久保拓弥 kubo@ees.hokudai.ac.jp

http://hosho.ees.hokudai.ac.jp/~kubo/ce/EesLecture2008.html

この講義のーとが「データ解析のための統計モデリング入門」として出版されました!

http://hosho.ees.hokudai.ac.jp/~kubo/ce/IwanamiBook.html

まちがいを修正し，詳しい解説・新しい内容をたくさん追加したものです

今日のもくじ



GLM 普及活動まっしぐらに推進中の「生態学の統計モデリング」，第 4 回目です．前回はポアソン分布を使った統計モデリング，ということで

- (線形) ポアソン回帰・(線形) ロジスティック回帰その他をまとめたものを**一般化線形モデル** (GLM) とよぶ
- GLM は**確率分布**と **link 関数**をえらび，要因をくみあわせて**線形予測子**を構築，最尤法によってパラメーター推定する
- どういう線形予測子の構成がよいのか**モデル選択**できる; これは「あてはまりのよさ」 (deviance) と「モデルの複雑さ」 (パラメーター数) のかねあいで決まるモデル選択規準 (AIC など) の比較による

といった内容を説明してみました．

今回は「もうひとつのよく使われる GLM」であるロジスティック回帰について説明したいと思います．これは観察された現象が**二項分布** (binomial distribution) で説明できそうなときに使う統計モデルです．

前回のポアソン回帰では，応答変数 $y$ が $y \in \{0, 1, 2, \cdots\}$ というように「0 以上だけど上限がどこにあるのかわからないカウントデータ」をあつかいました．

今回は $y \in \{0, 1, 2, \cdots N\}$ というふうに「カウントデータだけど，0 以上 $N$ 以下」の値をとる現象をあつかいます．コイン投げでいえば,「$N$ 個中の $y$ 個がオモテ，$N - y$ 個がウラ」となるような現象です．これはポアソン分布ではあつかえませんし，それゆえにポアソン回帰でパラメーター推定はできません．しかしながら，二項分布を使った **ロジスティック回帰**ならばこの現象を統計モデル化できます．

また,「$N$ 個中の $y$ 個」といったデータがあると生態学研究者などはすぐに $y/N$ といった割算をする経口があります．割算するだけならいいのですが，$\frac{\text{観測データ}}{\text{観測データ}}$ などといった **割算値を使ったデータは最悪**なものになりがちです．今日はそういった悪しきお作法を回避する方法についても検討したいと思います．

今日の講義で強調したポイントを列挙してみましょう:

- 今日の確率分布: **二項分布**
- 今日の link 関数: **logit link 関数**
- 今日のお作法: **何でも割算するな!**

# 1. 今日の例題: $N$ 個の胚珠のうち $y$ 個が結実

今回もまた架空データを使って統計モデルによるデータ解析を解説します．前回と同じような架空植物を使っていますが，今回は**最大種子数が全個体おなじ**というところが異なっています． [1]

個体は $i$ という記号であらわされ ($i = 1, 2, 3, \cdots, 100$, つまり 100 個体います)，その胚珠 (はいしゅ) [2] 数は 8 個 (全個体共通)，結実した [3] 種子数は $y_i$ とします．胚珠数が $N_i = 8$ 個なので全部結実した場合には種子数 $y_i = 8$ 個となり，これが最大種子数，最小種子数はもちろん全胚珠が結実に失敗して種子数ゼロ個の場合です．つまり，$y_i \in \{0, 1, 2, 3, \cdots, 8\}$ ということです．

ここではこの架空植物の個体ごとの (種子数ではなく) **結実確率**がどのように決まるか (統計モデルでどう表現するのがよいのか) をあつかいたいと

1. 前回の架空植物は最大種子数が不明な植物でした．

2. 種子のモトになる植物の器管，と考えてください．

3. つまりちゃんと種子のカタチになること．

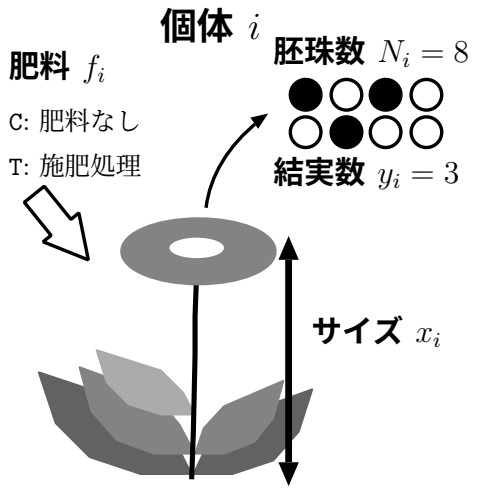



**図 1: 架空植物の第 $i$ 番目の個体 ($i = 1, 2, \cdots, 100$)**

します．結実確率は「ある胚珠が種子になる確率」です．この結実確率は個体 $i$ の中では共通の値 $q_i$ をとると仮定します．

あとは前回の架空植物の例題と同じで,「サイズ」[4] は $x_i$; また全個体のうち 50 個体 ($i = 1, 2, \cdots, 50$) は特に何もしてないけど (C 個体)，残り 50 個体 ($i = 51, 52, \cdots, 100$) には施肥処理と称して肥料をあげた (T 個体)，とします．

4. 前回と同じく，どうせ架空植物なので具体的には何も考えてなくて……「高さ」でも「重量」でも何でもいいです．

そしてこの例題で**調べたいこと**は「ある個体の結実確率 $q_i$ がサイズ $x_i$ や施肥処理によってどう変わるのか (あるいは変わらないのか)，それを統計モデルを使ったデータ解析，つまりパラメーター推定やモデル選択で明らかにしていこう，というものです．

前回の例題解析と同じく，R を起動してまずはデータファイルを読みこみます．今回は data4a.csv というファイル[5] にデータが格納されています．

5. 例によって講義 web page からダウンロードできます．

```
> d <- read.csv("data4a.csv")
```

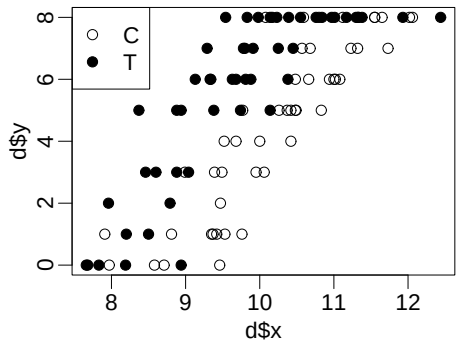
読みこんだデータを data.frame である d [6] に格納しました．その概要を調べましょう．

6. とまた安直な名前をつけた

```
> head(d)
  N y     x f
1 8 1  9.76 C
2 8 6 10.48 C
3 8 5 10.83 C
4 8 6 10.94 C
5 8 1  9.37 C
6 8 1  8.81 C
```

前回の種子数統計モデリングのデータとくらべると，今回は「N: 胚珠数」の列がひとつ増えています．あとは前回と同じで y: 結実種子数，x: 植物のサイズ，f: 施肥処理，という構造になっています．summary(d) で概況を表示させ，図であらわすと以下のようになります．

```
> summary(d)
       N           y               x          f
 Min.   :8   Min.   :0.00   Min.   : 7.660   C:50
 1st Qu.:8   1st Qu.:3.00   1st Qu.: 9.338   T:50
 Median :8   Median :6.00   Median : 9.965
 Mean   :8   Mean   :5.08   Mean   : 9.967
 3rd Qu.:8   3rd Qu.:8.00   3rd Qu.:10.770
 Max.   :8   Max.   :8.00   Max.   :12.440

> plot(d$x, d$y, pch = c(21, 19)[d$f])
```



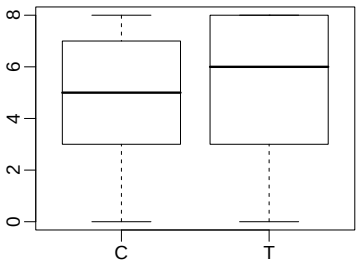

これをみると，

- サイズ $x_i$ が大きくなると結実種子数 $y_i$ が多くなるらしい
- 肥料をやると $(f_i = \mathtt{T})$ 結実種子数 $y_i$ が多くなるらしい

ということがわかります．ここでは各個体の胚珠数 $N_i = 8$ と一定の値に固定されていますから，結実種子数 $y_i$ が増大するということは結実確率 $q_i$ が増大しているため，と考えてよいでしょう．

ただし肥料の効果だけ見ようとしても，

```
> plot(d$f, d$y)
```




サイズがばらばらなので $y_i$ の range はかなり重複していますね．つまりさっきの図のほうが肥料の効果はよく「見える」といえます．

## 2. 二項分布のロジスティックモデルとは何か?

この架空植物の種子データのように「$N$ 個のうち $y$ 個が結実した」といった「した・しなかった」「あり・なし」データの統計モデリングの基本となるのは**二項分布** (binomial distribution) です．二項分布については第 2 回でちょっとだけふれました．ここでもう一度復習してみましょう．ポアソン分布の $y$ はゼロ以上のどんな値でもとることができますが，二項分布の $y$ が $\{0, 1, 2, \cdots, N\}$ の値しかとることができません．このときには二項分布を使うと $y$ が $\{0, 1, 2, \cdots, N\}$ となる確率を計算できます．

$$p(y \mid N, q) = \binom{N}{y} q^y (1-q)^{N-y}$$

ここで $q$ は注目している事象が成立する確率，この例題ですと胚珠が種子になる結実確率です．また

$$\binom{N}{y} = \frac{N!}{y!(N-y)!}$$

で[7] この例題ですと「$N$ 個の胚珠の中から $y$ 個の結実種子を選びだす場合の数」です．R では `choose(N, y)` を使って計算できます．[8]

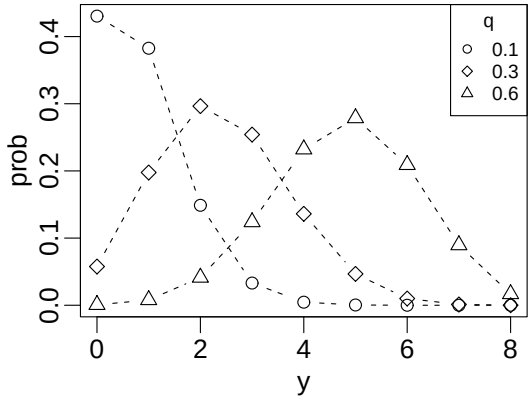
```
> choose(8, 0) # 8 個のうち 0 個が結実というパターンは 1 とおり
[1] 1
> choose(8, 1) # 8 個のうち 1 個が結実というパターンは 8 とおり
[1] 8
> choose(8, 4) # 8 個のうち 4 個が結実というパターンは 70 とおり
[1] 70
```

第 2 回の講義のーとでも示しましたが，二項分布の確率密度関数の例を R の `dbinom()` 関数を使って作図してみましょう．

7. $N!$ は N の階乗，$N! = 1 \times 2 \times \cdots \times N$

8. この例でいう「パターン」は「どの胚珠が結実したか」まで考慮するもので，たとえば `choose(8, 1)` だと 1 番目の胚珠だけが結実した場合，2 番目の胚珠だけが結実した場合，……，8 番目の胚珠だけが結実した場合の 8 とおり，という意味です．

```
> N <- 8
> y <- 0:N
> plot(y, dbinom(y, N, prob = 0.1), type = "b", lty = 2, pch = 21,
+ ylab = "prob")
> lines(y, dbinom(y, N, prob = 0.3), type = "b", lty = 2, pch = 23)
> lines(y, dbinom(y, N, prob = 0.6), type = "b", lty = 2, pch = 24)
> legend("topright", legend = c(0.1, 0.3, 0.6), pch = c(21, 23, 24),
+ title = "q", cex = 0.7)
```




さて，この例題のある個体 $i$ で $y_i$ 個の結実種子が観測された，という確率は二項分布を使って，[9]

$$p(y_i \mid N_i, q_i) = \binom{N_i}{y_i} q_i^{y_i} (1 - q_i)^{N_i - y_i}$$

と書けます．このときに個体 $i$ の胚珠の結実確率 $q_i$ がサイズ $x_i$ や施肥処理 $f_i$ によって上下してしまう (つまり依存している) という統計モデルを考えたい，とします．

このときに結実確率 $q_i$ をサイズや施肥処理の関数としてどのように書けばよいのでしょうか? いちばんよく使われるのはロジスティック関数 (logistic function) を使う方法です．これは GLM で言えば **logit link 関数**を使っていることになります．たとえば $q_i$ が**線形予測子** $z_i$ [10] の関数であり link 関数が logit である場合，

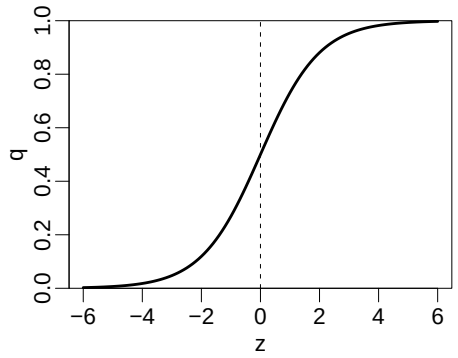
$$q_i = \frac{1}{1 + \exp(-z_i)}$$

と定義されます．

9. 今回の例題では説明をラクにするために $N_i = 8$ 個と固定しています．しかしながら，R の `glm()` 使ったロジスティック回帰では $N_i$ が個体によって異なる場合でも問題なくパラメーター推定できます．

10. 前回の講義を思いだしてください……線形予測子 $z$ は $z_i = \beta_1 + \beta_2 x_i + \cdots$ というふうにパラメーター $\beta_j$ に $x_i$ など要因の値をかけて，線形結合したものです．

```
> logistic <- function(z) 1 / (1 + exp(-z)) # 関数の定義
> z <- seq(-6, 6, 0.1)
> plot(z, logistic(z), type = "l", lwd = 3,
+ ylim = c(0, 1), yaxs = "i", ylab = "q")
> abline(v = 0, lty = 3)
```

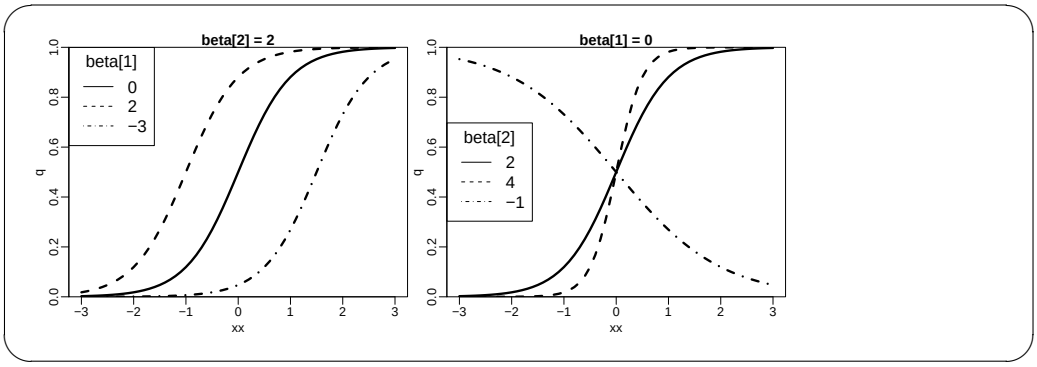



このように logistic 関数は $z_i$ が小さくなるとゼロに近づき，また $z_i$ が大きくなると 1 に近づきます．$z_i = 0$ のときに $q_i = \frac{1}{1+\exp(-0)} = 0.5$ となります．

さてここでたとえば結実確率 $q_i$ がサイズ $x_i$ だけに依存している，と仮定してみましょう (「サイズだけ」モデル)．線形予測子 $z_i = \beta_1 + \beta_2 x_i$ となります．このときに結実確率とパラメーター $\beta_1$ と $\beta_2$ の関係を図に描くと (すでに先ほどの `logistic()` 関数が定義されてるとして) ……

```
> par(mfrow = c(1, 2)) # 作図画面を 1 行 2 列に分割
> ## panel 1
> xx <- seq(-3, 3, 0.1)
> plot(xx, logistic(0 + 2 * xx), type = "l", lwd = 3,
+ ylim = c(0, 1), yaxs = "i", ylab = "q",
+ main = expression(beta[2]==2))
> lines(xx, logistic(2 + 2 * xx), lwd = 3, lty = 2)
> lines(xx, logistic(-3 + 2 * xx), lwd = 3, lty = 4)
> legend("topleft", legend = c(0, 2, -3), lty = c(1, 2, 4),
+ title = expression(beta[1]))
> ## panel 2
> plot(xx, logistic(0 + 2 * xx), type = "l", lwd = 3,
+ ylim = c(0, 1), yaxs = "i", ylab = "q",
+ main = expression(beta[1]==0))
> lines(xx, logistic(0 + 4 * xx), lwd = 3, lty = 2)
> lines(xx, logistic(0 - 1 * xx), lwd = 3, lty = 4)
> legend("left", legend = c(2, 4, -1), lty = c(1, 2, 4),
+ title = expression(beta[2]))
```

(次のページに続きます)




つまり

$$q_i = \frac{1}{1+\exp(-(\beta_1+\beta_2 x_i))}$$

と定義されてるモデルは

- $\beta_1$ を変えると曲線の「位置」が左右に動く $\Leftrightarrow$ ある $x_i$ における確率が上下する
- $\beta_2$ を変えると「傾き」みたいなものがきつくなったりゆるくなったりする [11]

といった関係があります.

線形予測子 $z_i = \beta_1 + \beta_2 x_i + \cdots$ の部分は必要とあらば施肥処理など他の要因もくみこむことができます. そして, 結実確率 $q_i$ の定義が決まったらあとはいつもと同じで, 尤度関数

$$L(q_i \mid y_i, N_i) = \prod_{i=1}^{100} \binom{N_i}{y_i} q_i^{y_i} (1-q_i)^{N_i - y_i}$$

から対数尤度関数

$$\log L(q_i \mid y_i, N_i) = \sum_{i=1}^{100} \left\{ \log \binom{N_i}{y_i} + y_i \log(q_i) + (N_i - y_i)\log(1-q_i) \right\}$$

を最大化するような推定値のセット $\{\hat{\beta}_j\}$ を最尤推定しなさい……という計算問題は R の `glm()` が全部やってくれているわけです.

11. ただし $\beta_1 + \beta_2 x_i$ という定式化のもとでは「傾き」が変わるだけでなく「位置」も左右に動きます. 上の右の図では $\beta_1 = 0$ としているので,「位置」は左右に動きません.

## 3. ロジスティック回帰: `glm(..., family = binomial)`

GLM の一部である logistic 回帰するところから始めてみましょう. あえて言えば, 何もわからなくてもできてしまう `glm(..., family = binomial)` といったところでしょうか.

さてさて……それではサイズ $x_i$ と施肥処理 $f_i$ が (結実確率 $q_i$ をつうじて) 結実種子数 $y_i$ をどのように変えているのか? それを R の glm() でロジスティック回帰してみましょう. 前回のポアソン回帰のときとほとんど書きかたは同じです. [12]

```
> fit.xf <- glm(cbind(y, N - y) ~ x + f, data = d,
+ family = binomial) # cbind(結実した胚珠数, 結実しなかった胚珠数)
```

ポアソン回帰とちがうところは, 応答変数の指定方法, cbind(y, N - y), です.

これは何を意味しているのか? データフレイム d の y 列と N 列を使って, 1 列目に結実した胚珠数 y を, 2 列目に結実しかなった胚珠数 N - y をもつ行列が cbind(y, N - y) によって作られています.

R でロジスティック回帰, つまり glm(..., family = "binomial") でモデル式「応答変数 ~ 説明変数」を指定している部分では, このように「cbind(結実した数, 結実しなかった数)」[13] というふうに指定してやる必要があります.

さて, このように R の glm() に推定計算させてみると,

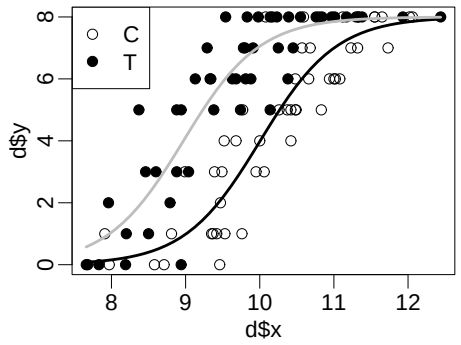
```
> fit.xf <- glm(cbind(y, N - y) ~ x + f, data = d,
+ family = binomial) # cbind(結実した胚珠数, 結実しなかった胚珠数)
> fit.xf
Call:  glm(formula = cbind(y, N - y) ~ x + f, # 略

Coefficients:
(Intercept)            x           fT
    -19.536        1.952        2.022

Degrees of Freedom: 99 Total (i.e. Null);  97 Residual
Null Deviance:      499.2
Residual Deviance: 123  AIC: 272.2
```

……なんとも簡単にパラメーター推定値が得られましたね. これはいったいどういう意味なのか? これはいったん**あとまわし**にして,「推定結果は『あてはまってる』のか?」を図示してみましょう.

12. このときに family = binomial と指定したときは default link 関数である"logit" が指定されている, つまり family = binomial(link = "logit") と指定しているのと同じことになります.

13. あるいは cbind(事象が発生した数, しなかった数), cbind(成功した数, 失敗した数) などなど.

```
> plot(d$x, d$y, pch = c(21, 19)[d$f])
> xx <- seq(min(d$x), max(d$x), length = 50)
> ff <- factor("C", levels = c("C", "T"))
> q <- predict(fit.xf, newdata = data.frame(x = xx, f = ff),
+ type = "response") # predict() を使ってモデルの予測計算
> lines(xx, q * 8, lwd = 3)
> ff <- factor("T", levels = c("C", "T"))
> q <- predict(fit.xf, newdata = data.frame(x = xx, f = ff),
+ type = "response")
> lines(xx, q * 8, col = "gray", lwd = 3) # 灰色の曲線
> legend("topleft", legend = c("C", "T"), pch = c(21, 19))
```




予測 (prediction) を描画させる R コードは上のごとくちょっとごたごたしてますけど，まあ何だかうまい推定ができているみたいです.

## 4. リンク関数 "logit" の意味・解釈

さて得られた推定結果をみながら，今回の講義で理解してほしいことのひとつ，logit リンク関数についてさらに説明してみましょう.

ある個体 $i$ のサイズ $x_i$ と施肥処理 $f_i$ が結実確率 $q_i$ を変えている (「全部いり」モデル)，というロジスティック + 二項分布モデルは

$$q_i = \frac{1}{1 + \exp(-(\beta_1 + \beta_2 x_i + \beta_3 f_i))}$$

と $\beta_3$ も加えた 3 パラメーターモデルで，追加された説明変数 $f_i$ は「肥料なし」 (C) なら $f_i = 0$,「施肥処理」 (T) なら $f_i = 1$ となっている変数，と「とりあえず」そう考えてもらってかまいません. [14]

この「全部いり」モデルを使ってパラメーター推定して,

14. 実際には R の中では対比行列という定式化の手法をつかって，このような factor な因子を巧妙にあつかっています. 水準が 3 以上の場合，たとえば「肥料なし」「旧型肥料」「新型肥料」の 3 水準の $f_i$ でも問題なく自動的に GLM にくみこんでくれます.

```
> fit.xf
Call:  glm(formula = cbind(y, N - y) ~ x + f, # 略

Coefficients:
(Intercept)            x           fT
    -19.536        1.952        2.022
... (略) ...
```

こように最尤推定値を得ました: 上の $q_i$ の定義と対応させると，$\beta_1 \approx 19.5$, $\beta_2 \approx 1.95$, $\beta_3 \approx 2.02$ となります．これらのパラメーターの値は何を意味しているのでしょうか?

確率あらわすロジスティックモデルが意味するところを理解するために，上の $q_i$ の式を変形してみましょう:

$$\begin{aligned}
\frac{1}{q_i} &= 1 + \exp(-(\beta_1 + \beta_2 x_i + \beta_3 f_i)) \\
\frac{1}{q_i} - 1 &= \exp(-(\beta_1 + \beta_2 x_i + \beta_3 f_i)) \\
\frac{1 - q_i}{q_i} &= \exp(-(\beta_1 + \beta_2 x_i + \beta_3 f_i)) \\
\frac{q_i}{1 - q_i} &= \exp(\beta_1 + \beta_2 x_i + \beta_3 f_i) = \exp(\beta_1)\exp(\beta_2 x_i)\exp(\beta_3 f_i)
\end{aligned}$$

……これで少しわかりやすいカタチになりました．左辺の $\frac{q_i}{1-q_i}$ は **オッズ** (odds) とよばれる量で，この場合だと (結実する確率) / (結実しない確率) と解釈できます．たとえば $q_i = 0.5$ のときにはオッズは 1 倍，$q_i = 0.8$ のときにはオッズは 4 倍である，といったりします．そしてこのオッズは $\exp$(パラメーター $\times$ 要因) と比例関係にあります．

R が推定した推定値 $\{\hat{\beta}_j\}$ を代入してみると，定数である $\exp(\hat{\beta}_1) = \exp(19.5)$ を省略して [15]

$$\frac{q_i}{1 - q_i} \propto \exp(1.95 x_i)\exp(2.02 f_i)$$

という関係が成立していることがわかります．

15. 記号 $\propto$ は「比例する」という意味です．

まず植物のサイズ $x_i$ の影響をみます．いま注目している個体 $i$ のサイズが「1 単位」増大したら結実確率のオッズはどう変化するでしょうか?

$$\begin{aligned}
\frac{q_i}{1 - q_i} &\propto \exp(1.95(x_i + 1))\exp(2.02 f_i) \\
&\propto \exp(1.95 x_i)\exp(1.95)\exp(2.02 f_i)
\end{aligned}$$

となるのでオッズは $\exp(1.95) \approx 7$ 倍になる，とわかります．同様に「肥料なし」 (C, $f_i = 0$) に比べて「施肥処理する」 (T, $f_i = 1$) とオッズが

$\exp(2.02) \approx 7.5$ 倍ふえることがわかります．このように $\exp(\beta_j$の推定値$)$ はオッズを (かけ算によって) 変えてしまうということです．

よく世間で「タバコを吸うとナントカ病の『リスク』が 7 倍になります」といった報道がなされます．ここでいう「リスク」とは (近似的には) **オッズ比** (odds ratio) のことです．[16] ある人間集団におけるナントカ病の発病と生活習慣のデータにもとづいたデータ解析でロジスティック回帰なんかを使っていたとすると，個人 $i$ が「タバコを吸う」という説明変数 $s_i$ につくパラメーター $\beta_s$ の最尤推定値は $\hat{\beta}_s = 1.95$ ぐらいだったときに，オッズ比は

$$\begin{aligned} \frac{(\text{喫煙者 odds})}{(\text{非喫煙者 odds})} &= \frac{\exp(\text{喫煙者・非喫煙者共通部分}) \times \exp(1.95 \times 1)}{\exp(\text{喫煙者・非喫煙者共通部分}) \times \exp(1.95 \times 0)} \\ &= \exp(1.95) \end{aligned}$$

ということで病気になるオッズ比 ($\approx$ 「リスク」) は $\exp(1.95) \approx 7$ 倍になりました，と発表されることになるのです．

ついでにハナシを上の種子の結実確率のオッズの定義にもどして，その対数をとってみると，

$$\log \frac{q_i}{1-q_i} = \beta_1 + \beta_2 x_i + \beta_3 f_i$$

となっていることがわかります．右辺の $\beta_1 + \beta_2 x_i + \beta_3 f_i$ は線形予測子 $z_i$ の定義そのものですね．そして左辺の $\log \frac{q_i}{1-q_i}$ は**対数オッズ**と呼ばれていて，かつこれは $q_i$ の **logit 関数**の定義になっています．つまり [17]

$$\text{logit}(q_i) = \log \frac{q_i}{1-q_i} = \beta_1 + \beta_2 x_i + \beta_3 f_i = z_i$$

という関係が成立しています．これは logit 関数が logistic 関数 $q_i = \dfrac{1}{1+\exp(-z_i)}$ の逆関数になっていることがわかりますね．

16. この架空報道例で言いたいのは，おそらく正確にはリスク比もしくは相対危険率とよばれる指標なのでしょうが……これは厳密にはオッズ比とは異なるものです．しかしながら，発症確率が低い疾病の場合にはリスク比 $\approx$ オッズ比となります．

17. logit 関数の逆関数は logistic 関数で，logistic 関数の逆関数は logit 関数である，ということです．

## 5. ロジスティックモデルのモデル選択

ある個体 $i$ のサイズ $x_i$ と施肥処理 $f_i$ が結実確率 $q_i$ を変えている「全部いり」モデル，R の `glm()` による推定結果をもうちょっと調べてみましょう．

```
> fit.xf <- glm(cbind(y, N - y) ~ x + f, data = d,
+ family = binomial) # cbind(結実した胚珠数, 結実しなかった胚珠数)
> fit.xf
Call:  glm(formula = cbind(y, N - y) ~ x + f, # 略

Coefficients:
(Intercept)            x           fT
    -19.536        1.952        2.022

Degrees of Freedom: 99 Total (i.e. Null);  97 Residual
Null Deviance:      499.2
Residual Deviance: 123  AIC: 272.2
```

Deviance まわりに関する前回の復習です．データ数 100 に対して null model は $\beta_1$ だけつまり上の表示だと（Intercept）だけモデルの自由度は 99 で，これに対して「全部いり」モデルは $\{\beta_1, \beta_2, \beta_3\}$ の 3 パラメーターを使っているので自由度は 97 残っている，ということです．二項分布を仮定しているときに一番単純な [18] null model の residual deviance (あてはまりの悪さ) は 499 だったのに対して，3 パラメーター使ったモデルでは 123 まで減少しました．

18. つまり結実確率はどんな個体でも一定．

前回のポアソン回帰と同じく，R の stepAIC() 関数を使うとモデル選択規準 AIC をつかったモデル選択ができます． [19]

```
> library(MASS) # stepAIC を定義する MASS package よみこみ
> stepAIC(fit.xf) # 「x + f モデル」を使う
```

19. サイズと施肥処理の「交互作用」項をどうしてもいれたい，というヒトは stepAIC(glm(y ~ x * f, ...)) とでもしてみてください．

この結果はここには示しませんが，前回と同じく最大化対数尤度 ($\log L^*$)・deviance・AIC の table をここに示してみると， [20] [21]

| **Model** | $k$ | $\log L^*$ | **Deviance** $-2\log L^*$ | **Residual deviance** | **AIC** |
|---|---|---|---|---|---|
| NULL | 1 | -321.2 | 642.4 | 499.2 | 644.4 |
| f | 2 | -316.9 | 633.8 | 490.6 | 637.8 |
| x | 2 | -180.2 | 360.3 | 217.2 | 364.3 |
| x + f | 3 | -133.1 | 266.2 | 123.0 | 272.2 |
| FULL | 100 | -71.6 | 143.2 | 0.0 | 343.2 |

20. モデル選択をするときには，このように「AIC の table」を作って，最良モデルだけでなく二番手・三番手のモデルが何なのかもいっしょに検討してみることをおススメします．

21. ここでいう FULL model とは，個体 $i$ ごとに $q_i = y_i/N_i$ といった具合に 100 個体ぶん 100 個の $q_i$ を与えてしまうモデルのことです．

サイズと施肥処理を同時に組みこんだ x + f モデルが AIC の観点からは「最良」ということがわかります．

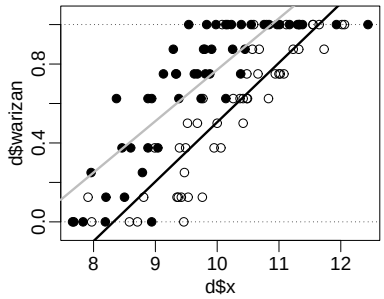
# 6. 本日のお作法: 何でも割算するな!

今回の GLM 紹介で述べたかったことは，$Y =$ (観測データ) / (観測データ) みたいな割り算値を使った解析をするな，ということです．理由はいろいろありますが，たとえば分母にも分子にも誤差が入ってる値を割って作った $Y$ はどんな確率分布にしたがうんだ? ふたつの値をそういう方法でひとつの値にしてよい根拠は何? などなどです．

そもそも，統計モデリングで (観測データ) / (観測データ) という数量を作りだして，それを何かの統計モデルにあてはめる，とういことはほとんどありえない，と断定してよいかと思います．つまりデータ解析にはこういう割算値は**不要**です．[22] こんな割り算値を使うぐらいなら，それより **マシ**な，より理解しやすい統計モデリングの方法が**必ずあります**．[23]

しかしながら，下のごとく**ダメ**なデータ解析がまだまだ横行しているのが現状です．たとえば今回つかっている種子結実データに対して割り算による「結実確率」をでっちあげて[24] ……

```
> d$warizan <- d$y / d$N
> plot(d$x, d$warizan, pch = c(21, 19)[d$f], ylim = c(-0.05, 1.05))
> dC <- d[d$f == "C",] # f == C だけ抽出
> abline(lm(warizan ~ x, data = dC), lwd = 3)
> dT <- d[d$f == "T",] # f == T だけ抽出
> abline(lm(warizan ~ x, data = dT), lwd = 3, col = "gray") # 灰色
> abline(h = c(0, 1), lty = 3) # q = 0, 1 の点線
```




データをぶちぶちと小分けして,「なんでも等分散正規分布」[25] とばかりに「結実確率」のセンをひいてしまう，といった悪しき作法です．批判すべき点は多々ありますが，たとえばサイズがある範囲をこえるとその「結実確率」がゼロより小さくなったり 1 より大きくなったりする意味不明な挙動だけでも十分にキモチ悪いものだと言えます．人間には意味不明な「予測」を吐きだす統計モデルはつくづく無価値なものです．[26]

今回紹介した GLM の一部，ロジスティックモデルは「$N$ 個のうち $y$ 個が応答した」現象の統計モデリングの基礎となるものです．もちろん私たちが

22. 作図のときには，使わざるをえない場合もあります.

23. 自然は割算をきらう，いや自然は割算をしない，といったところでしょうか.

24. このデータの場合は全個体の胚珠数が 8 に固定されているので，割り算値を図にしたときのみかけはそれほど悲惨ではありません……しかし個体ごとに胚珠数が異なる場合，その割り算値の図はめちゃくちゃ & 意味不明なものになりがちです.

25. 「等分散」だの「正規分布」だのという意識がないヒトも多いのですが.

26. ということで「センをひく」といったてきとーなる認識ではなく，どんなときでも「観測データを現象論的または機構論的に説明できるような統計モデルを構築する」という意識をもつことが重要なのです.

あつかう観測データの中には，この一番単純なモデルではうまく説明できない場合もあります．そういうときであっても，この二項分布や logistic 関数を部品とする拡張された統計モデルによって現象が説明されます． [27]

27. このようなモデルの改善方法については次の次の回からあつかいます．

# 7. 本日のわざ: offset 項わざで割算回避

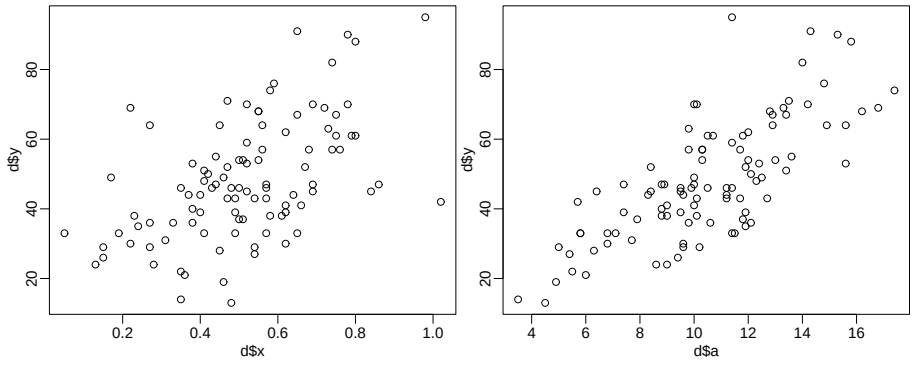
さきほど，割り算値を統計モデリングするようなことはありえない，と書きましたが，特に現象を 0 個 (回), 1 個, 2 個と数えられるようなカウントデータにおいては，(データ) / (データ) なる割り算値はまったく不要なものだと言えます．今回の一連の講義ではあつかいませんけれど，分割表のような複雑な table 形式のデータでもそれは同じことです．このあたりに関する教科書としては，Alan Agresti の入門的教科書の翻訳版「**カテゴリカルデータ解析入門**」(サイエンティスト社, 2003) をおすすめしておきます． [28]

28. より上級の教科書としては，同著者による Agresti (2002) *Categorical Data analysis*. John Wiley & Sons Inc. があります．

カウントデータの解析に割り算値はいらない，の補足のために前回あつかったポアソンモデルの「offset 項わざ」について説明してみます．たとえばこんな架空データがあったとします．

- 森林のあちこちに調査地 100 箇所を設定した ($i \in \{1, 2, \cdots, 100\}$)
- 調査地の面積 $a_i$ はふぞろいである
- また調査地の「明るさ」 $x_i$ を測っている
- 調査地 $i$ における植物個体数 $y_i$
- (解析の目的) 調査地 $i$ における植物個体の「密度」が「明るさ」 $x_i$ にどう影響されてるか知りたい

```
> d <- read.csv("data4b.csv") # データよみこみ
> par(mfrow = c(1, 2)) # 作図領域を二分割
> plot(d$x, d$y) # 明るさ vs 個体数
> plot(d$a, d$y) # 面積 vs 個体数
```



ここで密度の定義は個体数 / 面積，だからといって d$y / d$a みたいな割り算値をこしらえる**必要はありません**．このような問題は GLM の offset 項オプションを使って解決します．[29]

個体数が面積 $a_i$ に比例する，というのが個体密度の考えかたなので，ある場所における個体数の平均 $\lambda_i$ が

$$\lambda_i = a_i \times \text{密度}$$

と考えているわけですが，密度 $\geq 0$ でありかつ明るさ $x_i$ に依存しているので，

$$\lambda_i = a_i \exp(\beta_1 + \beta_2 x_i)$$

とします．この定義を変形すると，

$$\lambda_i = \exp(\beta_1 + \beta_2 x_i + \log(a_i))$$

となり線形予測子 $z_i = \beta_1 + \beta_2 x_i + \log(a_i)$ で log link 関数のポアソンモデルを適用できそうです．ただし，線形予測子の中の $\log(a_i)$ には (密度の定義から言って) $\beta_?$ といったパラメーターはつきません．このように線形予測子の中でパラメーターがつかない $+\log(a_i)$ のような項を **offset 項**と呼びます．[30]

R の glm() では offset 項を下のようにあつかいます．

```
> fit <- glm(y ~ x, offset = log(a), data = d, family = poisson)
Call:  glm(formula = y ~ x, family = poisson, ...(略)...)

Coefficients:
(Intercept)            x
     0.9749       1.0345

Degrees of Freedom: 99 Total (i.e. Null);  98 Residual
Null Deviance:      261.5
Residual Deviance: 81.66        AIC: 650.4
```

よく考えずにデータ解析してると，うっかり割り算してしまいそうな状況ですが，このような offset 項，あるいは別の状況では二項分布といった統計モデリングによって，**すっきりわかりやすい解析**ができるようになる，ということでした．

割算値を使った統計モデリングやデータ解析ってのは，まあ**「いらないもの」**なんだろう，とわかっていただけたでしょうか？[31]

29. じつは offset 項わざはポアソンモデルでは有効なのですが，二項分布モデルなどではあまり使いどころがありません．

30. 線形予測子に $\log(a_i)$ という「げた」をはかせているかんぢです．

31. これぐらいしつこく言っても，それでも割算してしまうところが生態学研究者の「すばらしい」なところですが……とある生態学研究者の生態観察者の意見．

# 8. GLM で統計モデリングを始めよう

前回・今回の 2 回にわたって GLM のとりかかりを説明してみました．データ解析初心者のみなさんには，まずは GLM で統計モデリングをやってみる，という方針をおすすめします．これによって「データはどんな確率分布にしたがうのだろうか」「確率分布のパラメーター [32] はどう指定するのか，といったことを考えるようになるからです．

32. GLM の場合は平均値 (のような量) を線形予測子と link 関数で与えましたね．

また R の `glm()` は `family` 指定で正規分布 (`gaussian`) も指定できるので，最小二乗法的な方法も GLM の世界で統一的にあつかえます．つまり直

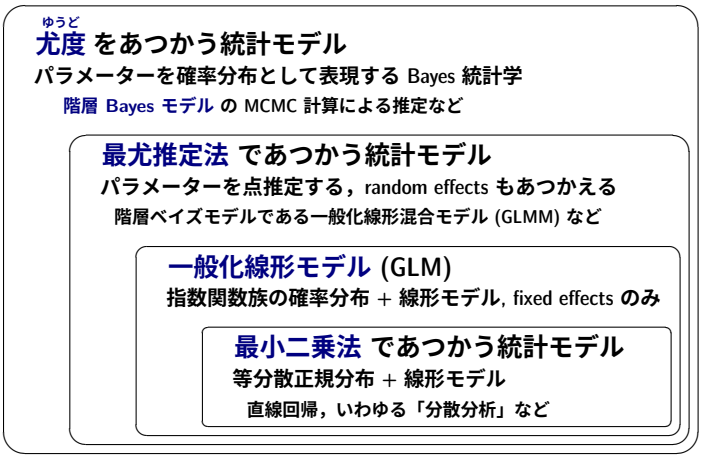



**図 2: 統計モデルの「地図」**

線回帰だのいわゆる「分散分析 (ANOVA)」だのも `glm()` であつかえます．

やたらと濫用されているこの「分散分析」，統計モデルとしては単なる

```
glm(y ~ x, family = gaussian, ...)
```

にすぎません．[33] そしてこれに「検定」をごちゃごちゃと組みあわせて，「モデル選択みたいなこと」をやります．この講義では「もっとすっきり AIC でモデル選択 (説明変数の選択) をやりましょう」という立場で説明してきました．

33. ただし (なぜかしら) $x$ は `factor` 型の説明変数のみ，となります．

最終回である次回はモデル選択と検定について検討してみることにします．