第３章 一般化線形モデル（ＧＬＭ）
―ポアソン回帰―

新納研究室

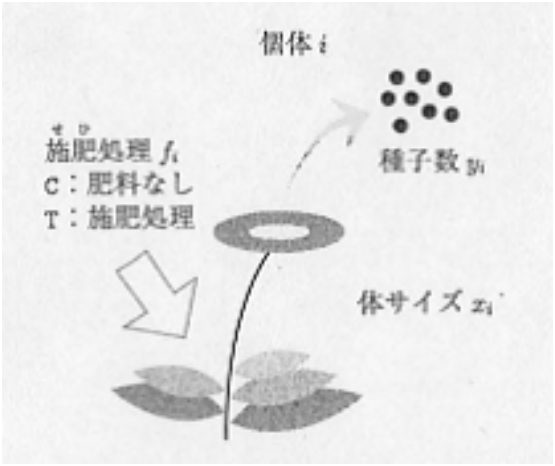
12T4069L 佐鳥 恭太郎

１．はじめに

- ポアソン回帰
  - 個体ごとに異なる説明変数（個体の属性）によって平均種子数が変化する統計モデルを観測データにあてはめること
- 一般線形モデル
  - ポアソン回帰と似た構造の統計モデルの総称

２．　例題：個体ごとに平均種子数が異なる場合

- 問題設定：
  - 個体数：１００
    - ＊個体iに対して … 体サイズ：x_i，種子数：y_i
    - ＊個体 i に対して以下の処理f_iがされている
    - ＊$i \in$ ｛１，２，…，５０｝は肥料なし：処理C
    - ＊$i \in$ ｛５１，５２，…，１００｝は施肥処理：処理T



３．　観測されたデータの概要を調べる

- 入力ファイル：data3a.csv
- 読み込み

```
> d <- read.csv("data3a.csv")
> d
     y     x f
1    6  8.31 C
2    6  9.44 C
3    6  9.50 C
 ...
98   8 10.24 T
99   7 10.86 T
100  9  9.97 T
```

４．　クラスの違い

- 数値型とファクター型

```
> class(d$y)
[1] "integer"
> d$y
  [1]  6  6  6 12 10  4  9  9  9 11  6 10  6 10 11  8
 [17]  3  8  5  5  4 11  5 10  6  6  7  9  3 10  2  9
...
> class(d$x)
[1] "numeric"
> d$x
  [1]  8.31  9.44  9.50  9.07 10.16  8.32 10.61 10.06
  [9]  9.93 10.43 10.36 10.15 10.92  8.85  9.42 11.11
...
> class(d$f)
[1] "factor"
> d$f
  [1] C C C C C C C C C C C C C C C C C C C C C C C C C
 [26] C C C C C C C C C C C C C C C C C C C C C C C C C
 [51] T T T T T T T T T T T T T T T T T T T T T T T T T
 [76] T T T T T T T T T T T T T T T T T T T T T T T T T
Levels: C T
```

５．データの概要

```
> summary(d)
       y                x           f
 Min.   : 2.00   Min.   : 7.190   C:50
 1st Qu.: 6.00   1st Qu.: 9.428   T:50
 Median : 8.00   Median :10.155
 Mean   : 7.83   Mean   :10.089
 3rd Qu.:10.00   3rd Qu.:10.685
 Max.   :15.00   Max.   :12.400
```

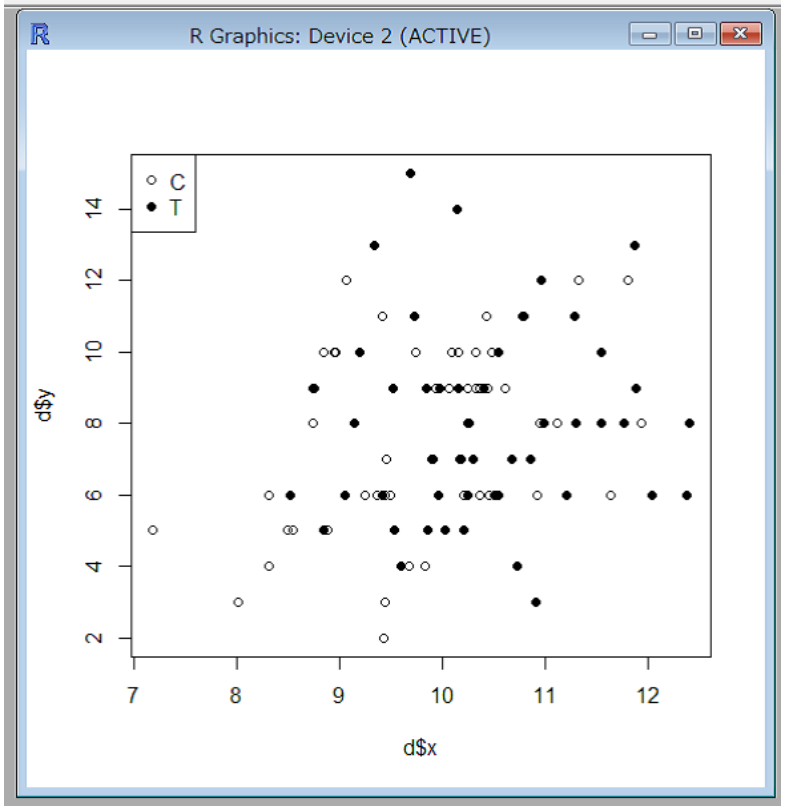
６．データの図示

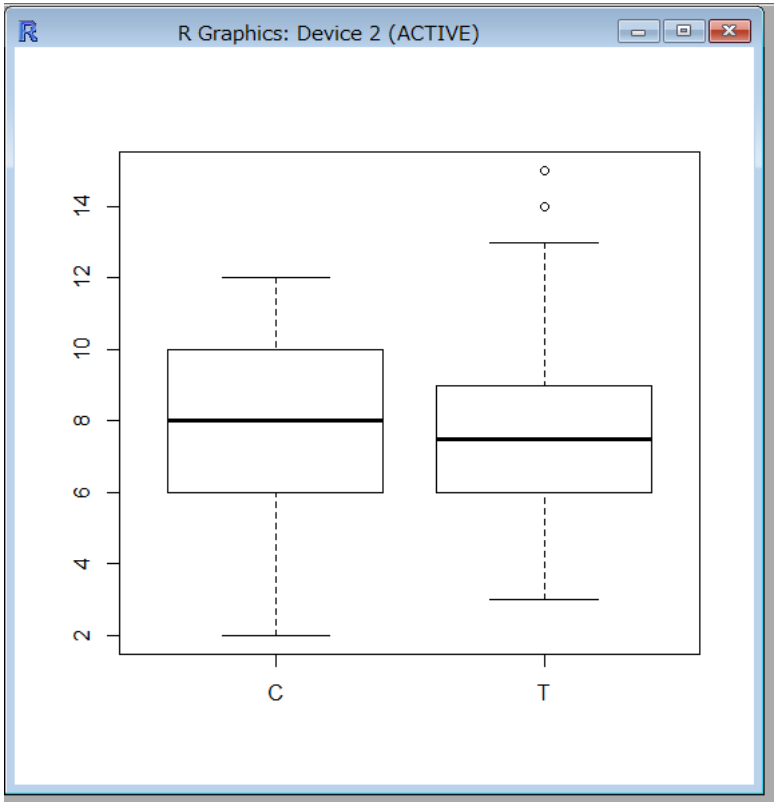
次ページ左図

```
> plot(d$x, d$y, pch = c(21, 19)[d$f])
> legend("topleft", legend = c("C", "T"), pch = c(21, 19))
```

次ページ右図

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


R Graphics: Device 2 (ACTIVE)
C
T
d$y
d$x
2
4
6
8
10
12
14
7
8
9
10
11
12


R Graphics: Device 2 (ACTIVE)
2
4
6
8
10
12
14
C
T

７．　ポアソン回帰の統計モデル

- 平均種子数 λ_i が体サイズ x や処理 f に影響されるモデルをつくる

最初に
個体 i の体サイズ x_i だけに依存する統計モデルを考える

- 説明変数：x_i，応答変数：y_i，施肥効果 f_i は無視
- ある個体 i において種子数が y_i である確率 $p(y_i|\ \lambda_i)$ はポアソン分布に従って

$$p(y_i|\ \lambda_i) = \frac{\lambda_i^{y_i} \exp(-\lambda_i)}{y_i!}$$

と仮定する。

８．　線形予測子と対数リンク関数

- 平均種子数 λ_i を x_i の関数として定義する

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

- β_1，β_2 … パラメータ（β_1：切片，β_2：傾き）
- この式の両辺に log をとると

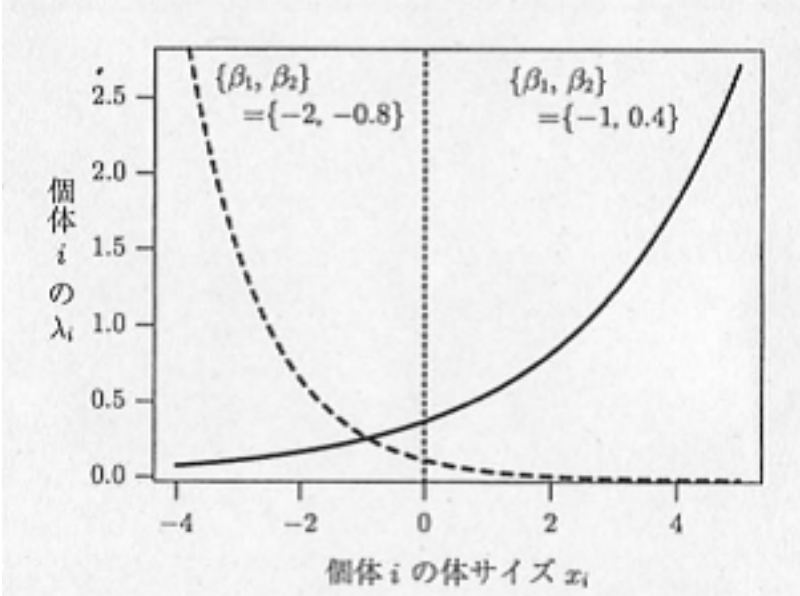
$$\log \lambda_i = \beta_1 + \beta_2 x_i$$

となり、左辺をリンク関数，右辺を線形予測子と呼ぶ

- ポアソン回帰では、たいていこの式のように対数をとった対数リンク関数を使用する

９． 対数リンク関数を使う理由

- 推定計算に都合がよい
  - $\lambda_i = \exp(\text{線形予測子}) \geq 0$でポアソン分布の平均は非負でなければならない
  - この条件によりRで最尤推定値を探索するときに便利



１０．　あてはめとあてはまりの良さ

- 対数尤度$logL$が最大になるパラメータ $\hat{\beta}_1$と $\hat{\beta}_2$の推定値をきめる
  - 統計モデルのあてはめ
- データYのもとでの、対数尤度

$$\log L(\beta_1, \beta_2) = \sum_i \log \frac{\lambda_i^{y_i} \exp(-\lambda_i)}{y_i!}$$

- Rで求める

```
> fit <- glm(y ~ x, data = d, family = poisson)
```

１１．パラメータの推定値

```
> summary(fit)
 ...
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.29172    0.36369   3.552 0.000383 ***
x            0.07566    0.03560   2.125 0.033580 *
 ...
```

- (Intercept)：切片 β_1，x：傾き β_2
- Estimate：推定値，Std.Error：パラメータの標準誤差の推定値

  z value：(最尤推定値をStd.Errorで除した値，Wald統計量)
  推定値たちがゼロから十分に離れているかの粗い目安

  `Pr(>|z|)`：推定値 $\hat{\beta}_1$や $\hat{\beta}_2$がゼロに近いことを表現する

１２． 最大対数尤度の評価

- この本では最大対数尤度を「あてはまりのよさ」と呼ぶ
- 最もあてはまりが良くなるのは対数尤度$\log L(\beta_1, \beta_2)$が最大のところ
- Rで最大対数尤度を評価する

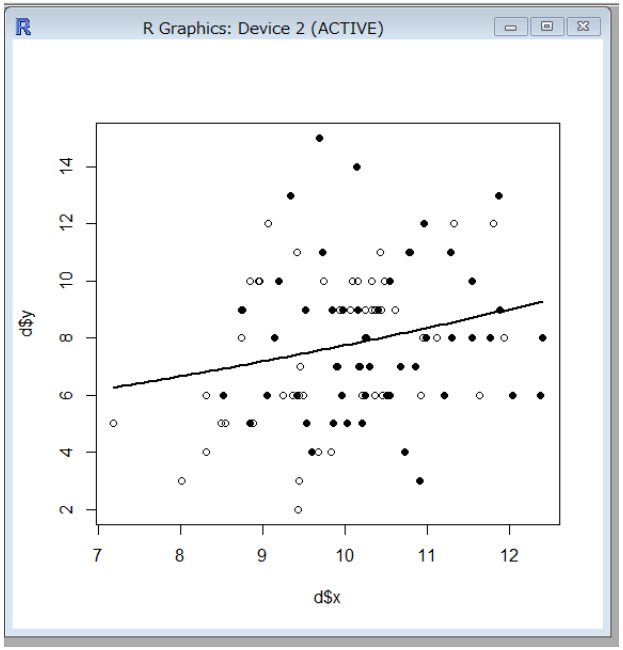
```
> logLik(fit)
'log Lik.' -235.3863 (df=2)
```

よって、最大対数尤度は約-235.4，自由度は2だと分かる
(自由度=最尤推定したパラメータ(β_1, β_2)の数)

１３．ポアソン回帰モデルによる予測

- 推定値が導かれたのでx-yグラフ上にλの予測値を図示する
- 平均種子数：$\lambda = \exp(1.29 + 0.0757x)$

```
> plot(d$x, d$y, pch = c(21, 19)[d$f])
> xx <- seq(min(d$x), max(d$x), length = 100)
> lines(xx, exp(1.29 + 0.0757 * xx), lwd = 2)
```



１４．　説明変数が因子型の統計モデル

- 因子型の説明変数をダミー変数におきかえる
- 体サイズx_iを無視して、施肥効果f_iだけが影響する

$$\lambda_i = \exp(\beta_1 + \beta_3 d_i)$$

- β_1：切片，　β_3：施肥効果

$$d_i = \begin{cases} 0(f_i = C\text{の場合}) \\ 1(f_i = T\text{の場合}) \end{cases} \qquad \lambda_i = \begin{cases} \exp(\beta_1)(f_i = C\text{の場合}) \\ \exp(\beta_1 + \beta_3 d_i)(f_i = T\text{の場合}) \end{cases}$$

１５．説明変数が因子型のパラメータ推定

- Rでパラメータ推定を行う
- Rでは、因子型の説明変数であってもglm()関数は動作する

```
> fit.f <- glm(y ~ f, data = d, family = poisson)
> summary(fit.f)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.05156    0.05070  40.463   <2e-16 ***
fT           0.01277    0.07148   0.179    0.858
```

- $fT = \beta_3$：説明変数f_iがT水準でとる値を示す。
  水準Cは基準値であるゼロをとる
- $\lambda_i = \exp(2.05 + 0) = 7.77$
- $\lambda_i = \exp(2.05 + 0.0128) = \exp(2.0628) = 7.87$

「肥料をやると平均種子数がほんの少し増える」という予測

１６． 説明変数が因子型のモデルのあてはまり

- 最大対数尤度

```
> logLik(fit.f)
'log Lik.' -237.6273 (df=2)
```

- よって、説明変数が体サイズx_iの場合の最大対数尤度-235.4と比べて小さい
- あてはまりが悪い

１７．説明変数が数量型＋因子型の統計モデル

- 説明変数が体サイズx_iの統計モデルと施肥効果f_iの統計モデルをあわせる
- 平均種子数 λ_iは

$$\lambda_i = \exp(\beta_1 + \beta_2 x_i + \beta_3 d_i)$$

となる。

１８．パラメータ推定

- Rでパラメータ推定を行う

```
> fit.all <- glm(y ~ x + f, data = d, family = poisson)
> fit.all
 ...
Coefficients:
(Intercept)                x                fT
    1.26311          0.08007         -0.03200
 ...
```

- 最大対数尤度

```
> logLik(fit.all)
'log Lik.' -235.2937 (df=3)
```

- よって、説明変数が体サイズx_iの場合の最大対数尤度-235.4と比べて少し大きい
- あてはまりが良くなっている

１９．対数リンク関数のわかりやすさ

- 平均種子数 $\lambda_i = \exp(1.26 + 0.08x_i - 0.032)$ $(f_i = T)$ のとき、
- λ_iは以下のように言える

$$\lambda_i = \exp(1.26) \times \exp(0.08x_i) \times \exp(-0.032)$$
$$= (\text{定数}) \times (\text{サイズの効果}) \times (\text{施肥処理の効果})$$

- $\exp(-0.032) = 0.969$なので、肥料をあげると平均種子数が0.969倍になる