

統計モデリング入門 新潟大 2015 (2)  
 ポアソン分布の一般化線形モデル (GLM)

久保拓弥 kubo@ees.hokudai.ac.jp, @KuboBook

新潟大学集中講義 <http://goo.gl/m8HSBM>

2015-05-26

ファイル更新時刻: 2015-05-18 16:30

この時間に説明したいこと

- ① ポアソン回帰の例題: 架空植物の種子数データ  
植物個体の属性, あるいは実験処理が種子数に影響?
- ② GLM の詳細を指定する  
確率分布・線形予測子・リンク関数
- ③ R で GLM のパラメーターを推定  
あてはまりの良さは対数尤度関数で評価
- ④ 処理をした・しなかった 効果も統計モデルに入れる  
GLM の因子型説明変数

一般化線形モデルって何だろう?

一般化線形モデル (GLM)

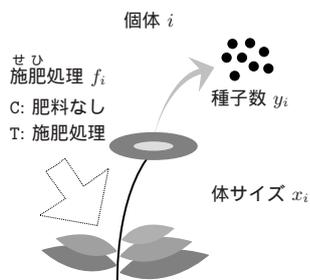
- **ポアソン回帰** (Poisson regression)
- ロジスティック回帰 (logistic regression)
- 直線回帰 (linear regression)
- .....

1. ポアソン回帰の例題: 架空植物の種子数データ

植物個体の属性, あるいは実験処理が種子数に影響?

まずはデータの概要を調べる

個体サイズと実験処理の効果を調べる例題



- 応答変数: 種子数  $\{y_i\}$
- 説明変数:
  - 体サイズ  $\{x_i\}$
  - 施肥処理  $\{f_i\}$

標本数

- 無処理 ( $f_i = C$ ): 50 sample ( $i \in \{1, 2, \dots, 50\}$ )
- 施肥処理 ( $f_i = T$ ): 50 sample ( $i \in \{51, 52, \dots, 100\}$ )

データファイルを読みこむ



data3a.csv は CSV (comma separated value) format file なので, R で読みこむには以下のようにする:

```
> d <- read.csv("data3a.csv")
```

データは d と名付けられた data frame (表みたいなもの) に格納される

とりあえず data frame d を表示

```
> d
  y     x f
1  6  8.31 C
2  6  9.44 C
3  6  9.50 C
... (中略) ...
99 7 10.86 T
100 9  9.97 T
```



## 2. GLM の詳細を指定する

確率分布・線形予測子・リンク関数

ポアソン回帰では log link 関数を使うのが便利

## 一般化線形モデルを作る

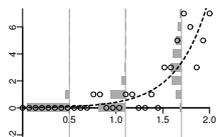
### 一般化線形モデル (GLM)

- 確率分布は?
- 線形予測子は?
- リンク関数は?

### GLM のひとつである **ポアソン回帰モデル** を指定する

#### ポアソン回帰のモデル

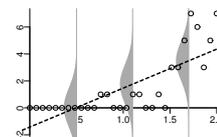
- 確率分布: **ポアソン分布**
- 線形予測子: e.g.,  $\beta_1 + \beta_2 x_i$
- リンク関数: **対数リンク関数**



### GLM のひとつである **直線回帰モデル** を指定する

#### 直線回帰のモデル

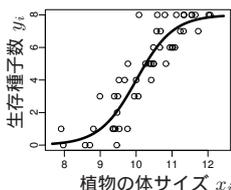
- 確率分布: **正規分布**
- 線形予測子: e.g.,  $\beta_1 + \beta_2 x_i$
- リンク関数: **恒等リンク関数**



### GLM のひとつである **logistic 回帰モデル** を指定する

#### ロジスティック回帰のモデル

- 確率分布: **二項分布**
- 線形予測子: e.g.,  $\beta_1 + \beta_2 x_i$
- リンク関数: **logit リンク関数**



### 「結果 ← 原因」関係を表現する線形モデル

- 結果: 応答変数
- 原因: 説明変数
- 線形予測子 (linear predictor):  
 (応答変数の平均) = 定数 (切片)  
 + (係数 1) × (説明変数 1)  
 + (係数 2) × (説明変数 2)  
 + (係数 3) × (説明変数 3)  
 + ...

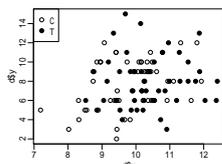
(交互作用項については粕谷さんが説明してくれます)

R で一般化線形モデル: glm() 関数

	確率分布	乱数生成	パラメーター推定
(離散)	ベルヌーイ分布	rbinom()	glm(family = binomial)
	二項分布	rbinom()	glm(family = binomial)
	ポアソン分布	rpois()	glm(family = poisson)
	負の二項分布	rnbinom()	glm.nb() in library(MASS)
(連続)	ガンマ分布	rgamma()	glm(family = gamma)
	正規分布	rnorm()	glm(family = gaussian)

- glm() で使える確率分布は上記以外もある
- GLM は直線回帰・重回帰・分散分析・ポアソン回帰・ロジスティック回帰その他の「よせあつめ」と考えてもよいかも
- 今日はポアソン回帰を使った GLM だけ紹介します

さてさて、この例題にもどって



種子数  $y_i$  は平均  $\lambda_i$  のポアソン分布にしたがうとしましょう

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

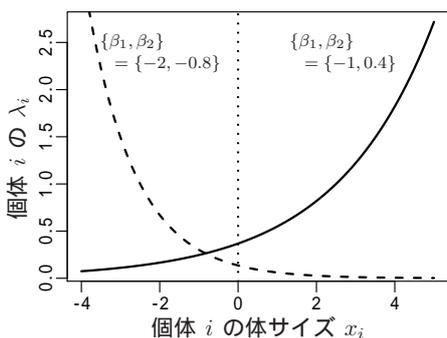
個体  $i$  の平均  $\lambda_i$  を以下のようにおいてみたらどうだろう.....?

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

- $\beta_1$  と  $\beta_2$  は係数 (パラメーター)
- $x_i$  は個体  $i$  の体サイズ,  $f_i$  はとりあえず無視

指数関数ってなんだっけ?

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



GLM のリンク関数と線形予測子

個体  $i$  の平均  $\lambda_i$

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

⇕

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

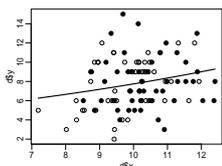
log(平均) = 線形予測子

log リンク関数とよばれる理由は、上のようにになっているから

この例題のための統計モデル

ポアソン回帰のモデル

- 確率分布: ポアソン分布
- 線形予測子:  $\beta_1 + \beta_2 x_i$
- リンク関数: 対数リンク関数



3. R で GLM のパラメーターを推定

あてはまりの良さは対数尤度関数で評価

推定計算はコンピューターにおまかせ

### glm() 関数の指定

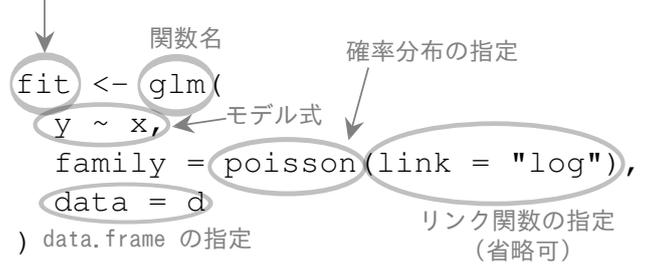
```
> d
  y   x f
1  6 8.31 C
2  6 9.44 C
3  6 9.50 C
... (中略) ...
99 7 10.86 T
100 9 9.97 T
```

これだけ!

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

### glm() 関数の指定の意味

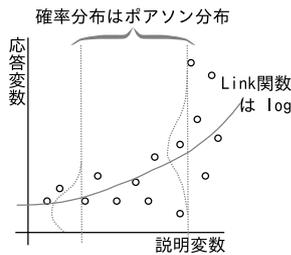
結果を格納するオブジェクト



- モデル式 (線形予測子  $z$ ): どの説明変数を使うか?
- link 関数:  $z$  と応答変数 ( $y$ ) 平均値 の関係は?
- family: どの確率分布を使うか?

### glm() 関数の指定を再確認

- family: poisson, ポアソン分布
- link 関数: "log"
- モデル式 (線形予測子  $z$ ): た  
たとえば  $y \sim x$  と指定したと  
する



- 線形予測子  $z = \beta_1 + \beta_2 x$   
 $\beta_1, \beta_2$  は推定すべきパラメーター
- 応答変数の平均値を  $\lambda$  とすると  $\log(\lambda) = z$   
つまり  $\lambda = \exp(z) = \exp(\beta_1 + \beta_2 x)$
- 応答変数 は平均  $\lambda$  のポアソン分布に従う:  $y \sim \text{Pois}(\lambda)$

### glm() 関数の出力

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

all: glm(formula = y ~ x, family = poisson, data = d)

Coefficients:
(Intercept)          x
      1.2917       0.0757

Degrees of Freedom: 99 Total (i.e. Null); 98 Residual
Null Deviance:      89.5
Residual Deviance: 85 AIC: 475
```

### glm() 関数のくわしい出力

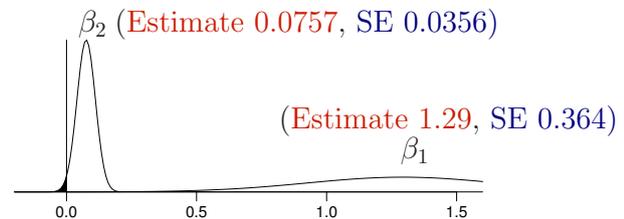
```
> summary(fit)
Call:
glm(formula = y ~ x, family = poisson, data = d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.368  -0.735  -0.177   0.699   2.376

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.2917    0.3637    3.55  0.00038
x            0.0757    0.0356    2.13  0.03358

..... (以下, 省略) .....
```

### 推定値と標準誤差



### この図の要点:

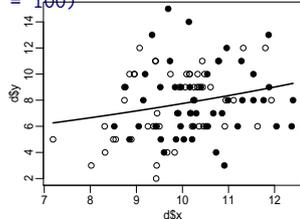
- 確率  $p$  は **ゼロからの距離** をあらわしている
- $p$  がゼロに近いほど **推定値  $\hat{\beta}$**  はゼロから離れている
- $p$  が 0.5 に近いほど **推定値  $\hat{\beta}$**  はゼロに近い

## モデルの予測

```
> fit <- glm(y ~ x, data = d, family = poisson)
...
Coefficients:
(Intercept)      x
      1.2917      0.0757
```

```
> plot(d$x, d$y, pch = c(21, 19)[d$f]) # data
> xp <- seq(min(d$x), max(d$x), length = 100)
> lines(xp, exp(1.2917 + 0.0757 * xp))
```

ここでは観測データと予測の関係  
を見ているだけ、なのだが

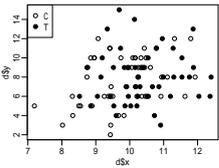


## 4. 処理をした・しなかった 効果も統計モデルに入れる

### GLM の因子型説明変数

数量型 + 因子型 という組み合わせで

## 肥料の効果 $f_i$ もいれましょう



種子数  $y_i$  は平均  $\lambda_i$  のポアソン分布にしたがうと  
しましょう

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

個体  $i$  の平均  $\lambda_i$  を次のようにする

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

- $\beta_3$  は施肥処理の効果の係数
- $f_i$  のダミー変数

$$d_i = \begin{cases} 0 & (f_i = C \text{ の場合}) \\ 1 & (f_i = T \text{ の場合}) \end{cases}$$

## glm(y ~ x + f, ...) の出力

```
> summary(glm(y ~ x + f, data = d, family = poisson))
...(略)...
```

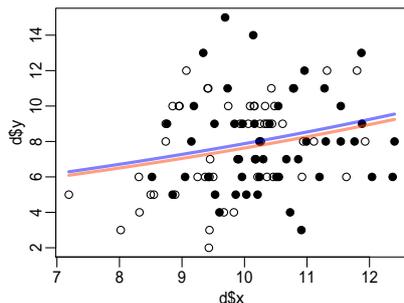
Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	1.2631	0.3696	3.42	0.00063
x	0.0801	0.0370	2.16	0.03062
fT	-0.0320	0.0744	-0.43	0.66703

..... (以下, 省略) .....

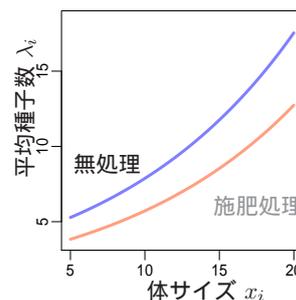
## x + f モデルの予測

```
> plot(d$x, d$y, pch = c(21, 19)[d$f]) # data
> xp <- seq(min(d$x), max(d$x), length = 100)
> lines(xp, exp(1.2631 + 0.0801 * xp), col = "blue", lwd = 3) # C
> lines(xp, exp(1.2631 + 0.0801 * xp - 0.032), col = "red", lwd = 3) # T
```



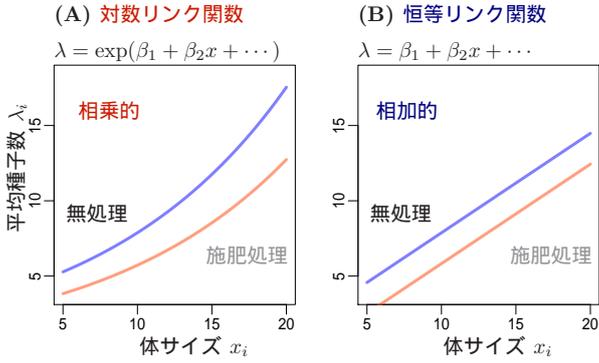
## 複数の説明変数をいれた場合の統計モデル

- $f_i = C: \lambda_i = \exp(1.26 + 0.0801x_i)$
- $f_i = T: \lambda_i = \exp(1.26 + 0.0801x_i - 0.032)$   
 $= \exp(1.26 + 0.0801x_i) \times \exp(-0.032)$

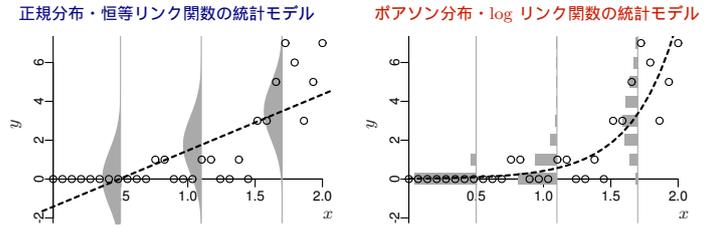


施肥効果である  $\exp(-0.032)$  は  
かけ算できくことに注意!

## リンク関数が違うとモデルの解釈が異なる



## GLM: 適切な確率分布 とリンク関数を選ぶ



## この授業であつかう統計モデルたち

