

統計モデリング入門 2014 (3)

一般化線形モデル: ポアソン回帰

久保拓弥 kubo@ees.hokudai.ac.jp

北大環境科学院の講義 <http://goo.gl/XeBR2x>

2014-07-09

ファイル更新時刻: 2014-07-08 22:26

今日のハナシ I

① ポアソン回帰の統計モデル

応答変数 y と説明変数 x

② 例題: 少し生態学っぽい種子数データ

植物個体の属性, あるいは実験処理が種子数に影響?

③ GLM の詳細を指定する

確率分布・線形予測子・リンク関数

④ R で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

⑤ 推定されたモデルを使って予測

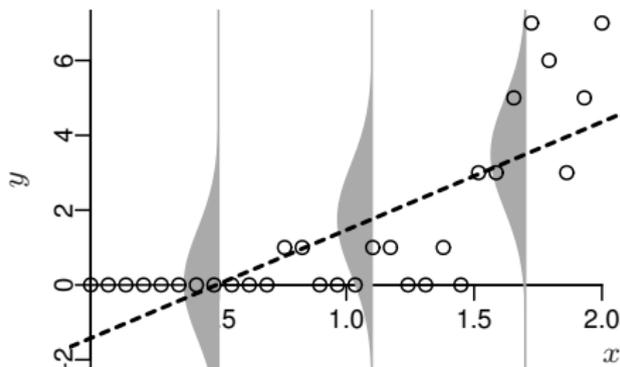
推定された結果とデータを比較する

今日のハナシ II

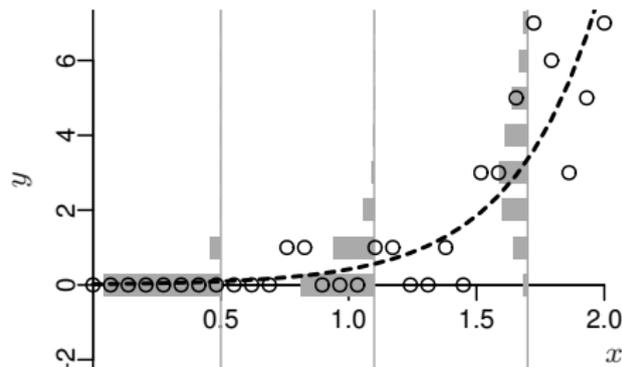
⑥ 処理をした・しなかった 効果も統計モデルに入れる

GLM の因子型説明変数

正規分布・恒等リンク関数の統計モデル



ポアソン分布・log リンク関数の統計モデル

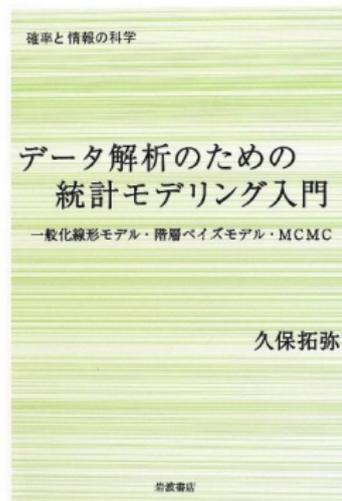


今日の内容と統計モデリング入門との対応

今日はおもに「**第3章 一般化線形モデル (GLM)**」の内容を説明します。

- 著者: 久保拓弥
- 出版社: 岩波書店
- 2012-05-18 刊行

<http://goo.gl/Ufq2>



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)
-

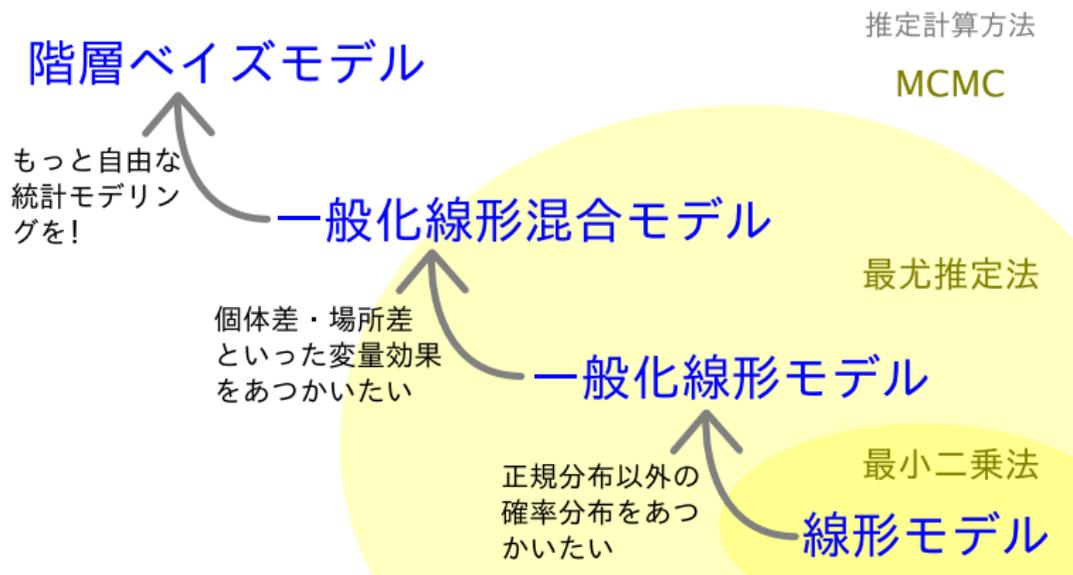
1. ポアソン回帰の統計モデル

応答変数 y と説明変数 x

一般化線形モデルにとりくんでみる

この授業であつかう統計モデルたち

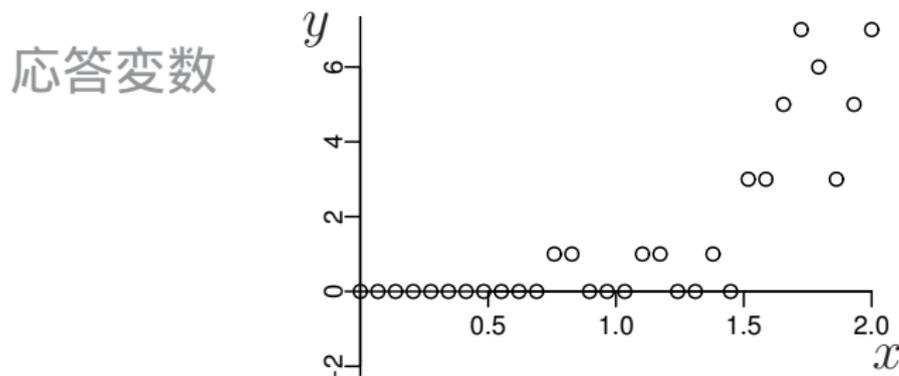
線形モデルの発展



データの特徴にあわせて線形モデルを改良・発展させる

0 個, 1 個, 2 個と数えられるデータ

カウントデータ ($y \in \{0, 1, 2, 3, \dots\}$ なデータ)

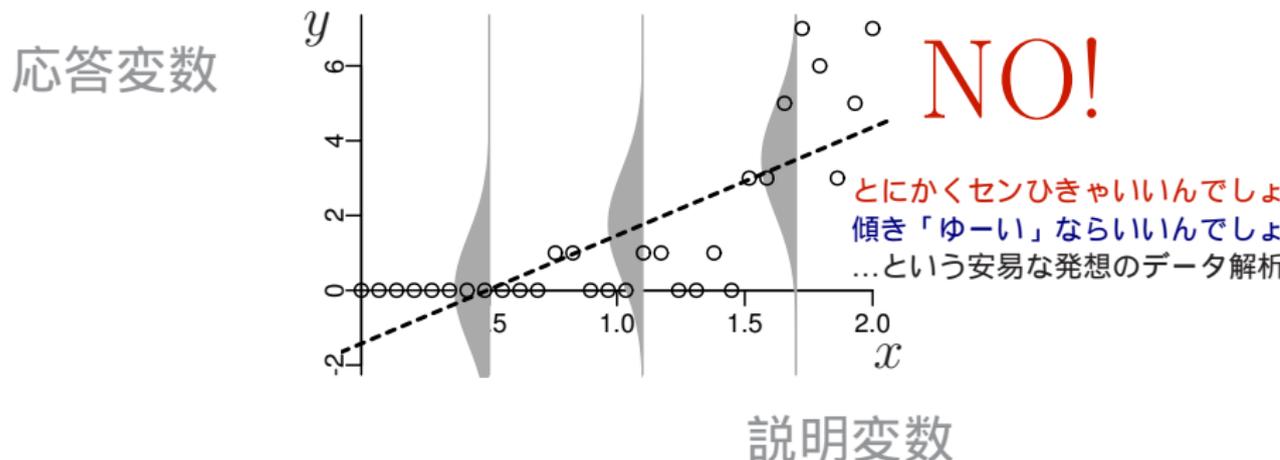


説明変数

- たとえば x は植物個体の大きさ, y はその個体の花数
- 体サイズが大きくなると花数が増えるように見えるが.....
- この現象を表現する統計モデルは?

正規分布を使った統計モデル ムリがある？

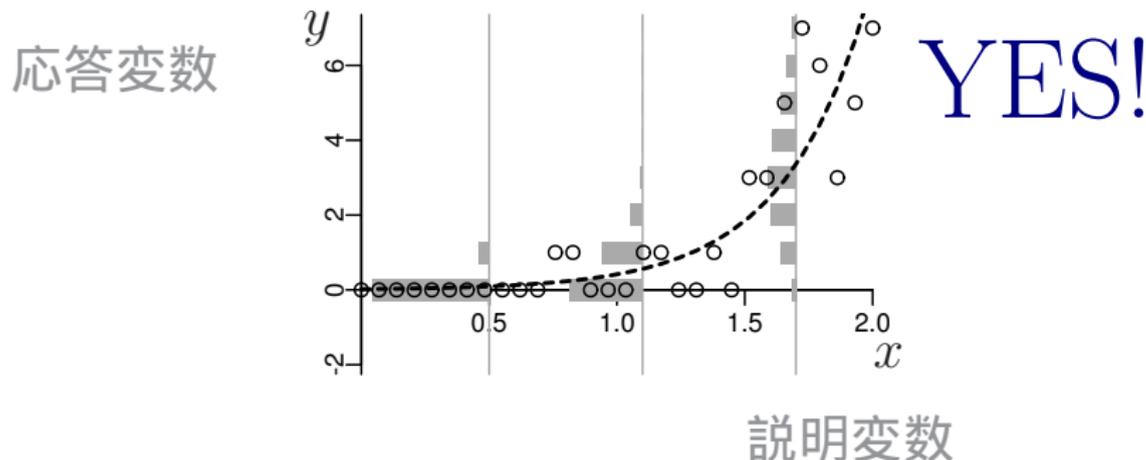
正規分布・恒等リンク関数の統計モデル



- タテ軸のばらつきは「正規分布」なのか？
- y の値は 0 以上なのに
- 平均値がマイナス？

ポアソン分布を使った統計モデルなら良さそう?!

ポアソン分布・対数リンク関数の統計モデル



- タテ軸に対応する「ばらつき」
- 負の値にならない「平均値」
- 正規分布を使ってるモデルよりましだね

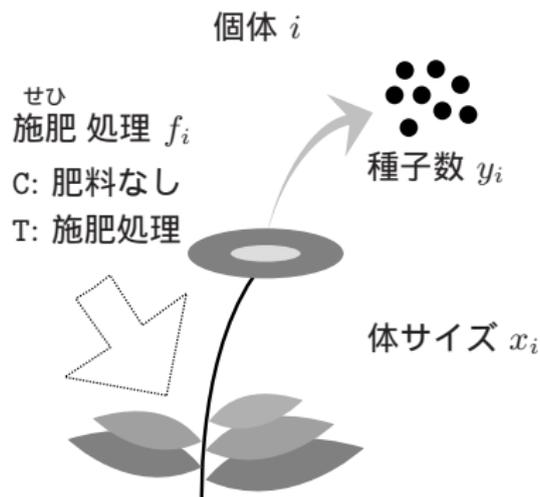
2. 例題: 少し生態学っぽい種子数データ

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

data frame `d` を調べる: `d$x`, `d$y`

```
> 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
... (中略) ...
[97]  8.52 10.24 10.86  9.97

> 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
... (中略) ...
[97]  6  8  7  9
```

data frame d を調べる: d\$f — factor type!

施肥処理の有無をあらわす f 列はちょっと様子がちがう

```
> 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
 [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
 [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
 [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
Levels: C T
```

因子型データ: いくつかの水準をもつデータ
ここでは C と T の 2 水準

Rのデータのクラスとタイプ

```
> class(d) # d は data.frame クラス
[1] "data.frame"
> class(d$y) # y 列は整数だけの integer クラス
[1] "integer"
> class(d$x) # x 列は実数も含むので numeric クラス
[1] "numeric"
> class(d$f) # そして f 列は factor クラス
[1] "factor"
```

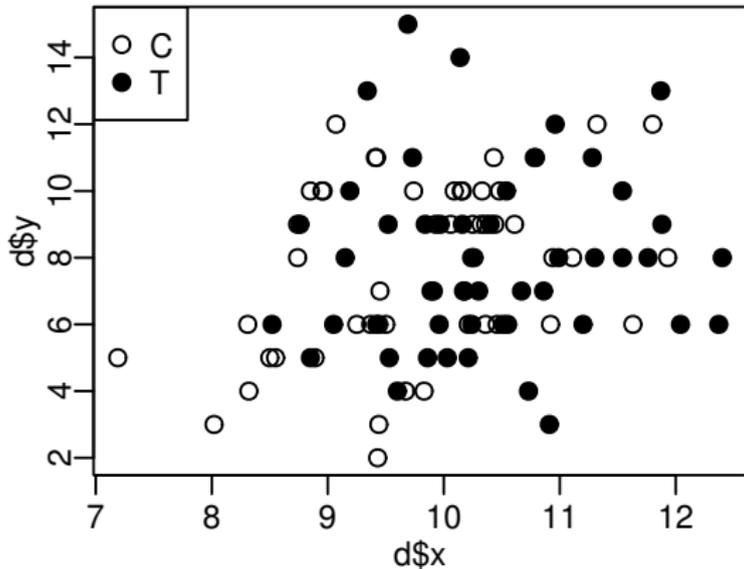
data frame の summary()

```
> 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	

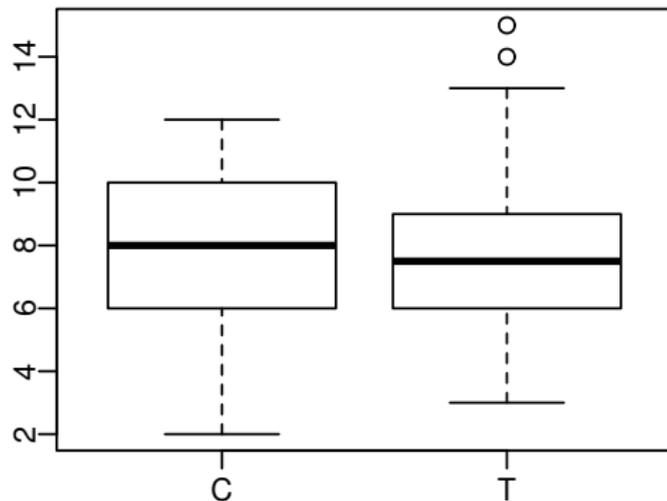
データはとにかく図示する!

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



施肥処理 f を横軸とした図

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



3. 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

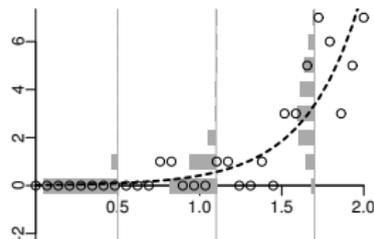
一般化線形モデル (GLM)

- 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

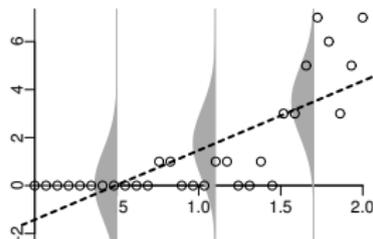
- 確率分布: ポアソン分布
- 線形予測子: e.g., $\beta_1 + \beta_2 x_i$
- リンク関数: 対数リンク関数



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

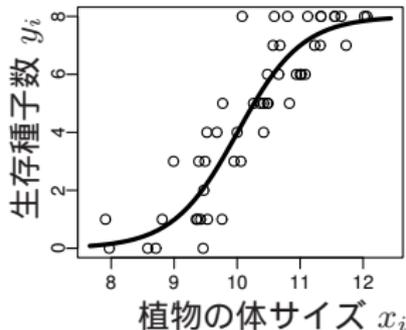
- 確率分布: 正規分布
- 線形予測子: e.g., $\beta_1 + \beta_2 x_i$
- リンク関数: 恒等リンク関数



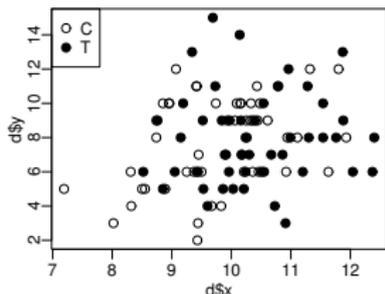
GLM のひとつである **logistic 回帰モデル** を指定する

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



さてさて，この例題にもどって



種子数 y_i は平均 λ_i のポアソン分布にしたが
うとしましょう

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

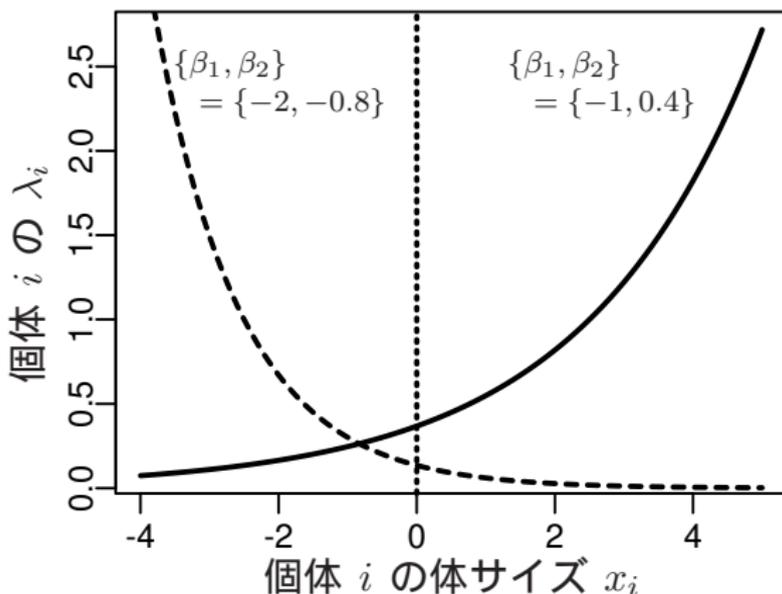
個体 i の平均 λ_i を以下のようにおいてみたらどうだろう……?

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

- β_1 と β_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 λ_i

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



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

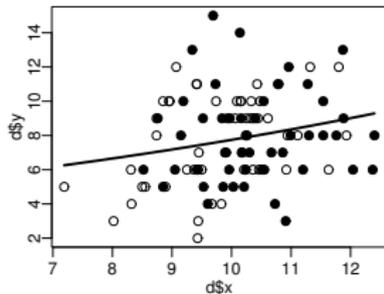
$$\log(\text{平均}) = \text{線形予測子}$$

log リンク関数とよばれる理由は、上のようになっているから

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**



4. R で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

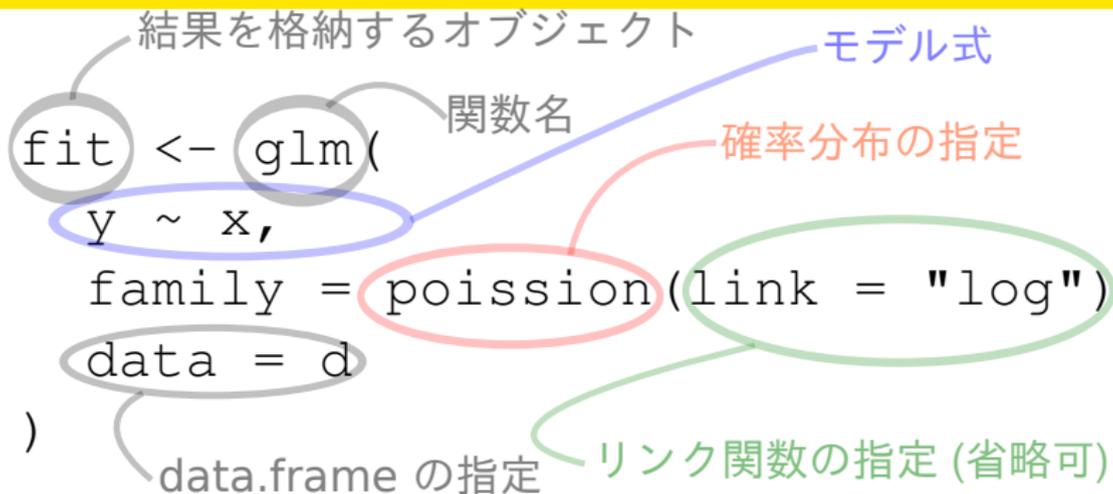
glm() 関数の指定

```
> 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)
```

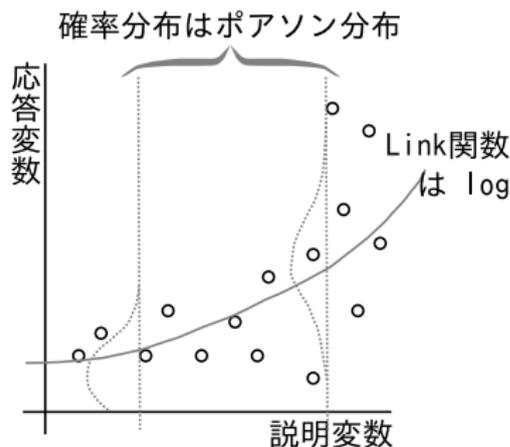
glm() 関数の指定の意味



- 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** $z = \beta_1 + \beta_2 x$
 β_1, β_2 は推定すべきパラメーター
 - **応答変数の平均値**を λ 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 λ 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:
```

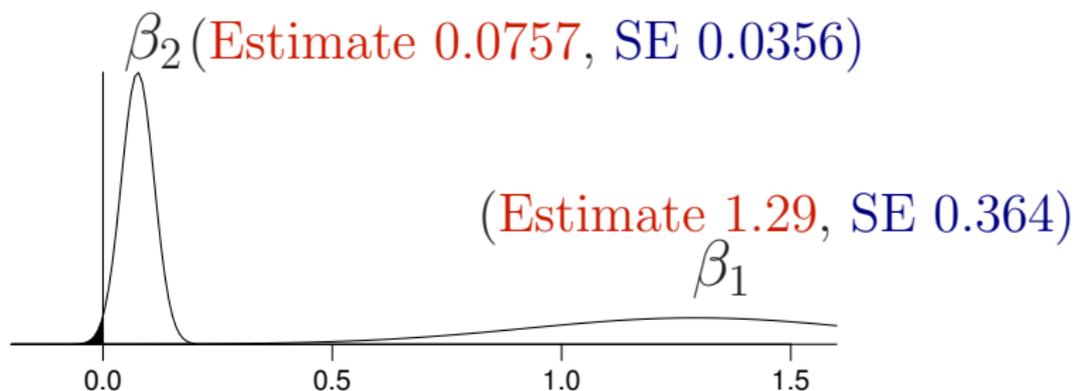
Min	1Q	Median	3Q	Max
-2.368	-0.735	-0.177	0.699	2.376

```
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

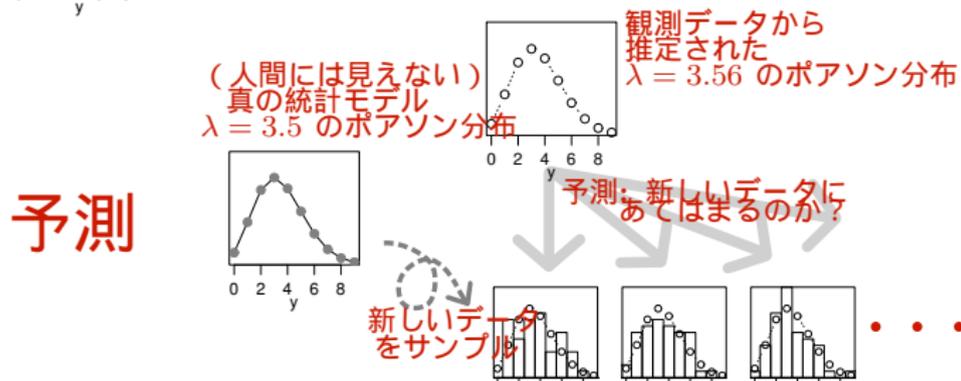
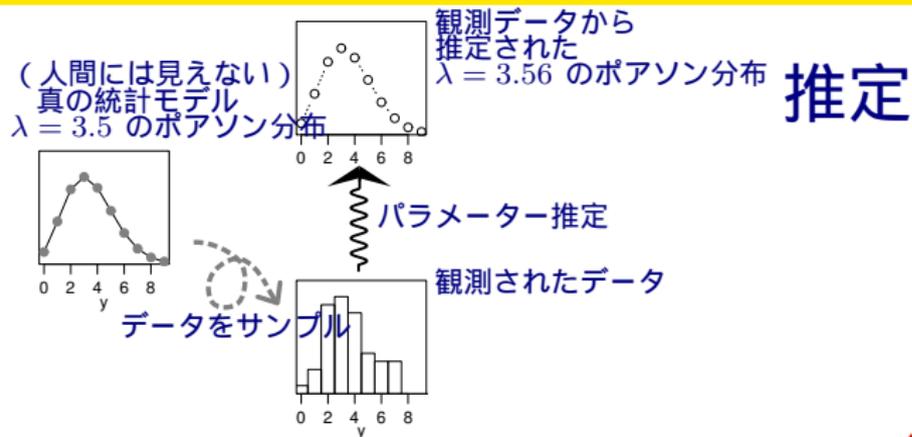


5. 推定されたモデルを使って予測

推定された結果とデータを比較する

ここでも作図が重要!

統計学における推定と予測



モデルの予測

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

```
...
```

```
Coefficients:
```

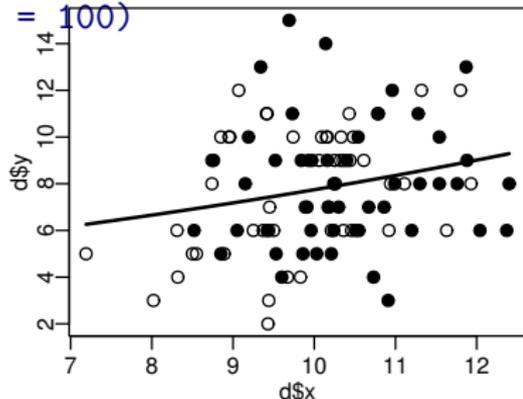
```
(Intercept)          x
    1.2917         0.0757
```

```
> 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

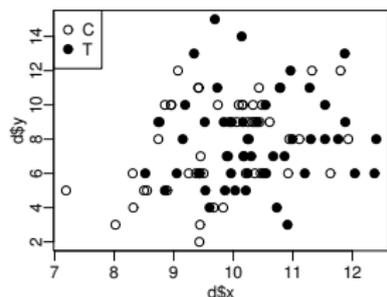


6. 処理をした・しなかった 効果も統計モデル に入れる

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 λ_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 λ_i を次のようにする

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

- β_3 は施肥処理の効果の係数
- f_i のダミー変数

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

glm(y ~ x + f, ...) の出力

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

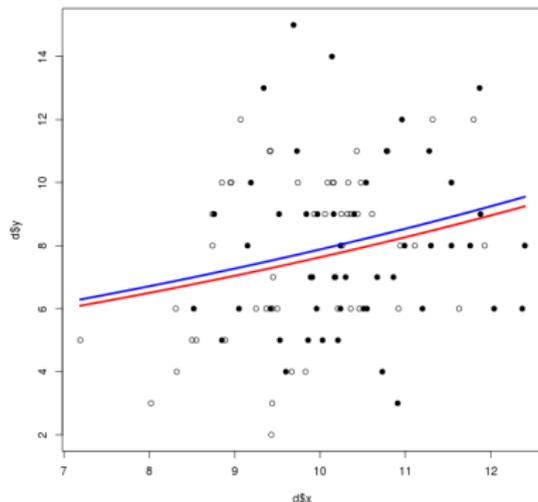
Coefficients:

	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)

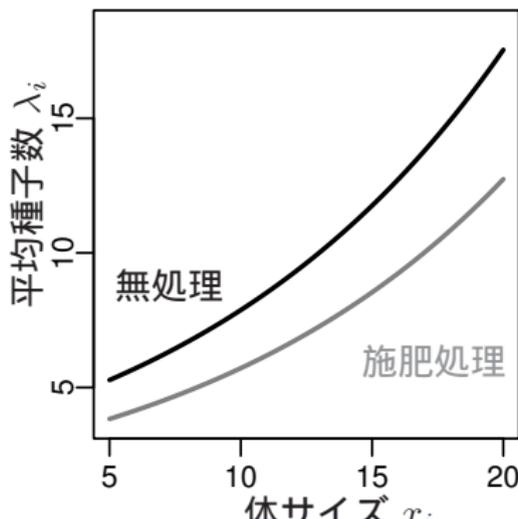
X + f モデルの予測

```
> 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

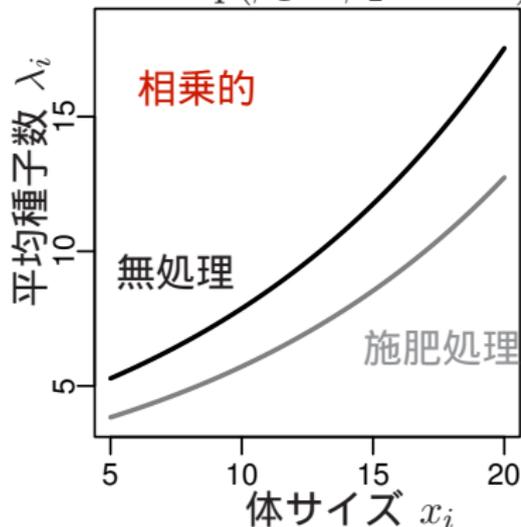
- $f_i = \text{C}$: $\lambda_i = \exp(1.26 + 0.0801x_i)$
- $f_i = \text{T}$: $\lambda_i = \exp(1.26 + 0.0801x_i - 0.032)$
 $= \exp(1.26 + 0.0801x_i) \times \exp(-0.032)$



リンク関数が違うとモデルの解釈が異なる

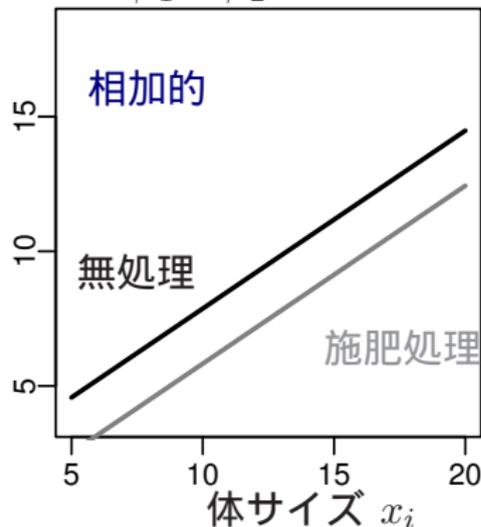
(A) 対数リンク関数

$$\lambda = \exp(\beta_1 + \beta_2 x + \dots)$$



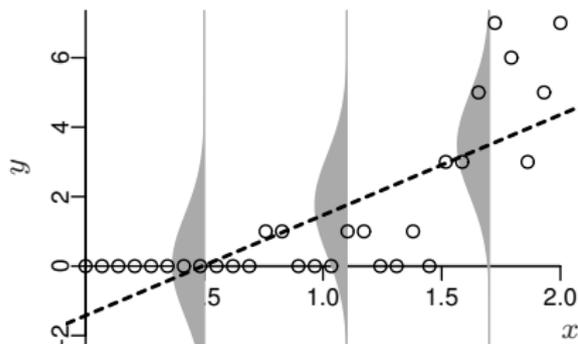
(B) 恒等リンク関数

$$\lambda = \beta_1 + \beta_2 x + \dots$$

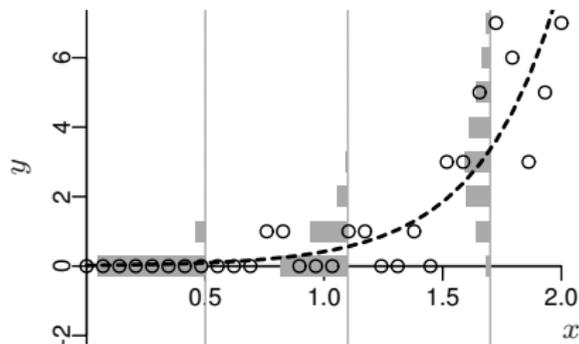


GLM: 適切な確率分布 とリンク関数を選ぶ

正規分布・恒等リンク関数の統計モデル

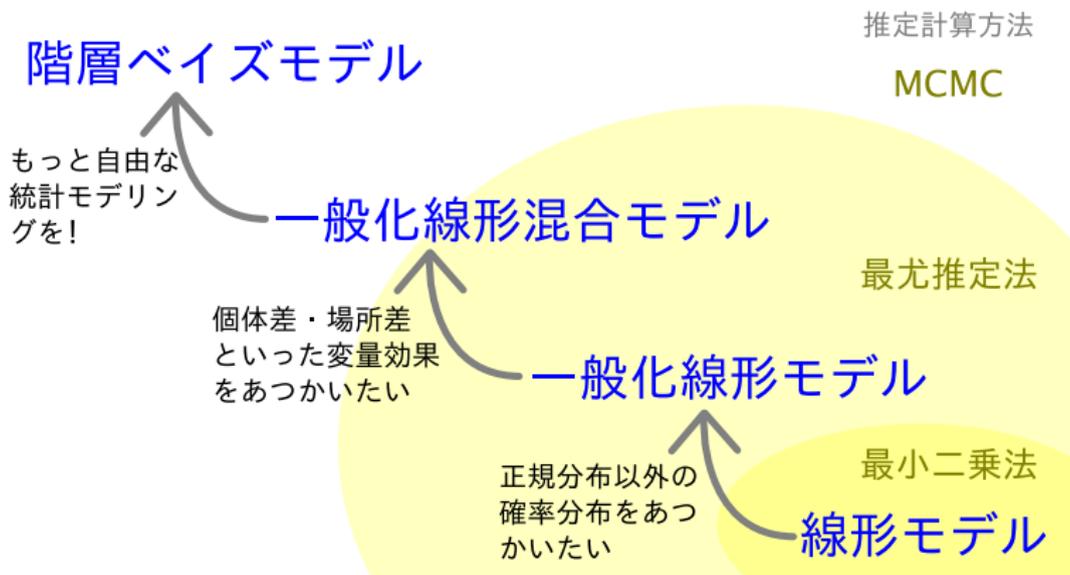


ポアソン分布・log リンク関数の統計モデル



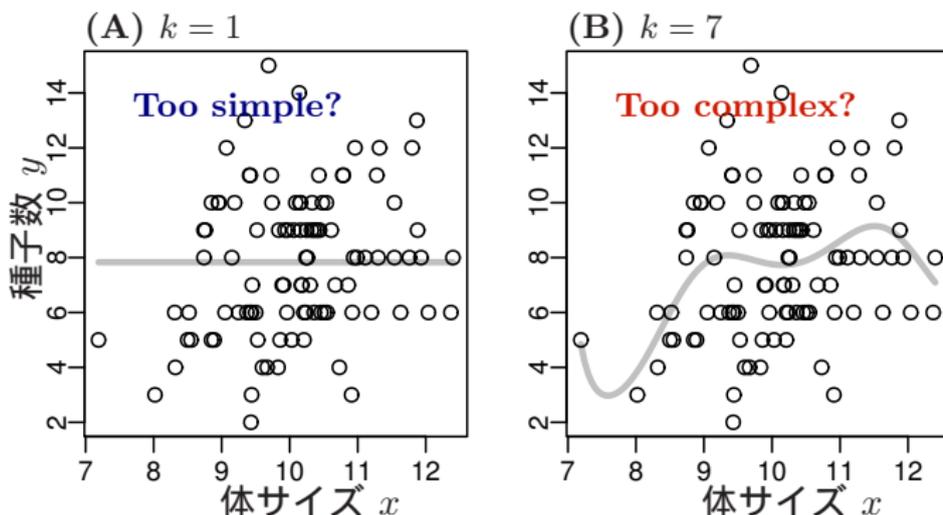
この授業であつかう統計モデルたち

線形モデルの発展



データの特徴にあわせて線形モデルを改良・発展させる

次回予告



モデル選択と統計学的検定

Model selection and statistical test