

# 新潟大集中講義 2016 (c)

一般化線形モデル (ポアソン回帰) とモデル選択

久保拓弥 `kubo@ees.hokudai.ac.jp`

新潟大の講義 <http://goo.gl/aFLLHZ>

2016-11-15

ファイル更新時刻: 2016-11-10 17:15

# 今日のハナシ I

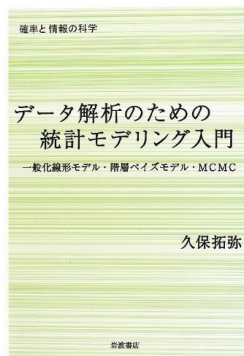
- ① ポアソン回帰の統計モデル  
応答変数  $y$  と説明変数  $x$
- ② ポアソン回帰の例題: 架空植物の種子数データ  
植物個体の属性, あるいは実験処理が種子数に影響?
- ③ GLM の詳細を指定する  
確率分布・線形予測子・リンク関数
- ④ R で GLM のパラメーターを推定  
あてはまりの良さは対数尤度関数で評価
- ⑤ 処理をした・しなかった 効果も統計モデルに入れる  
GLM の因子型説明変数
- ⑥ モデル選択  
予測力のよい統計モデルはどれか?
- ⑦ AIC を使ったモデル選択  
あてはまりの悪さ: deviance

# 今日の内容と「統計モデリング入門」との対応

<http://goo.gl/Ufq2>

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



## 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)
- .....

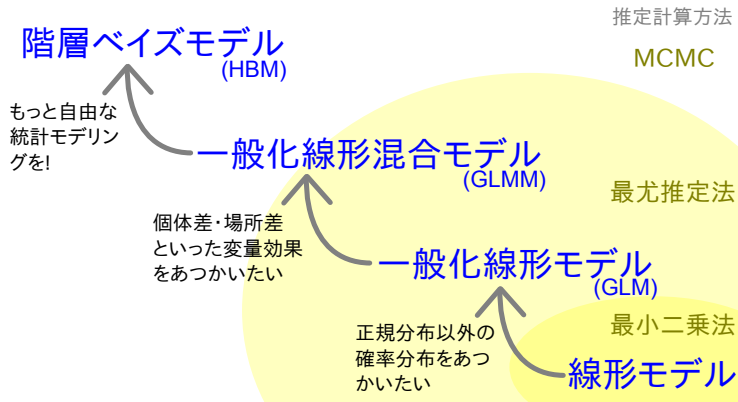
# 1. ポアソン回帰の統計モデル

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

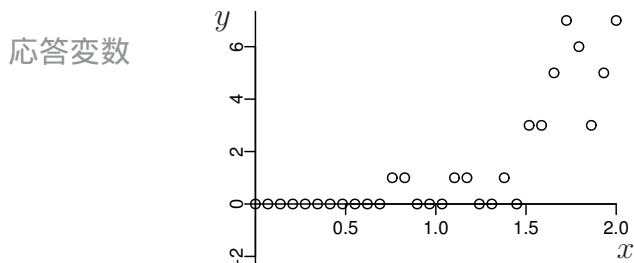
## この授業であつかう統計モデルたち

## 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

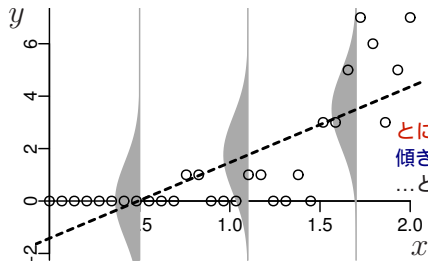
カウントデータ ( $y \in \{0, 1, 2, 3, \dots\}$  なデータ)

- 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

応答変数



説明変数

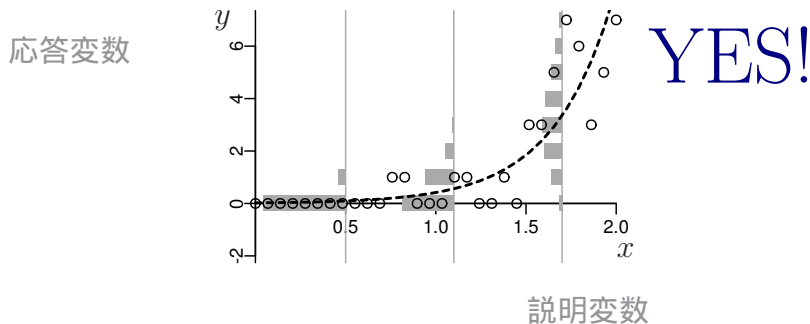
とにかくセンひきゃいいんでしょ  
傾き「ゆーい」ならいいんでしょ  
...という安易な発想のデータ解析

- 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

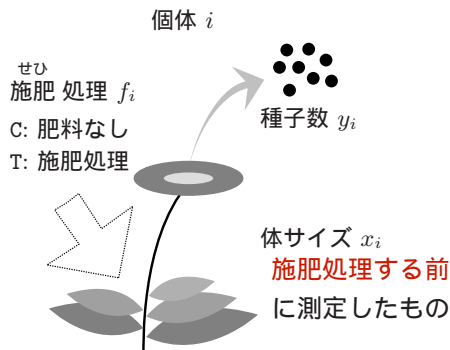
## 7. ポアソン回帰の例題: 架空植物の種子数データ

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 ( $f_i = \text{C}$ ): 50 sample ( $i \in \{1, 2, \dots, 50\}$ )
- 施肥処理 ( $f_i = \text{T}$ ): 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: 連続値と整数値

```
> 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: “因子型” のデータ

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
 [26] 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
 [76] 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)
```

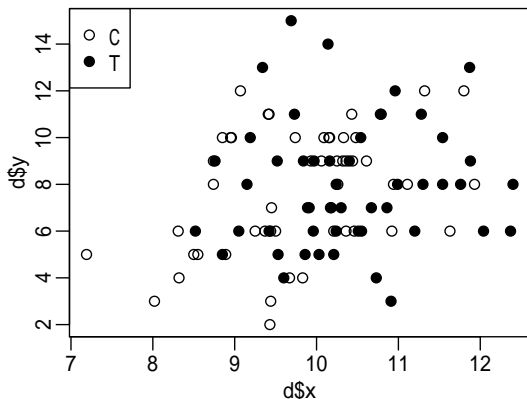
	y	x	f
Min.	: 2.00	Min. : 7.190	C:50
1st Qu.:	6.00	1st Qu.: 9.428	T:50
Median :	8.00	Median :10.155	
Mean :	7.83	Mean :10.089	
3rd Qu.:	10.00	3rd Qu.:10.685	
Max. :	15.00	Max. :12.400	



# データはとにかく図示する!

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

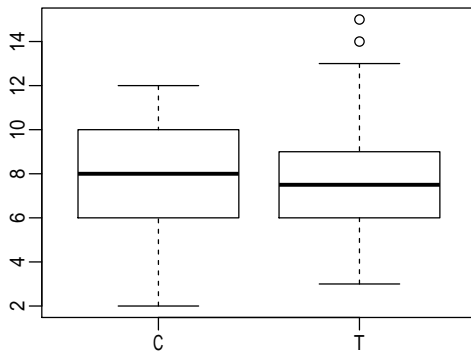
```
> legend("topleft", legend = c("C", "T"), pch = c(21, 19))
```



散布図

施肥処理  $f$  を横軸とした図

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



箱ひげ図

### 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?

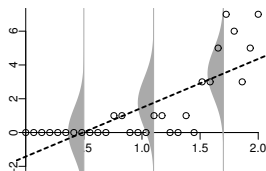
## GLM のひとつである直線回帰モデルを指定する

## 直線回帰のモデル

- 確率分布: 正規分布
- 線形予測子: e.g.,  $\beta_1 + \beta_2 x_i$

直線の式: (切片) + (傾き)  $\times x_i$

- リンク関数: 恒等リンク関数



## 結果 ← 原因 (かも?) を表現する線形モデル

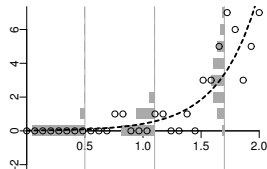
- 結果: 応答変数
- 原因: 説明変数
- 線形予測子 (linear predictor):

$$\begin{aligned} \text{(応答変数の平均)} &= \text{定数 (切片)} \\ &+ \text{(係数 1)} \times \text{(説明変数 1)} \\ &+ \text{(係数 2)} \times \text{(説明変数 2)} \\ &+ \text{(係数 3)} \times \text{(説明変数 3)} \\ &+ \dots \end{aligned}$$

## 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

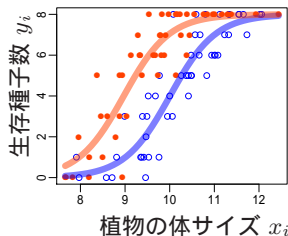
- 確率分布: ポアソン分布
- 線形予測子: e.g.,  $\beta_1 + \beta_2 x_i$
- リンク関数: 対数リンク関数



GLM のひとつである **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



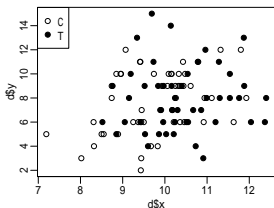


# R で一般化線形モデル (GLM) の推定を.....

	確率分布	乱数発生	GLM あてはめ
(離散)	ベルヌーイ分布	<code>rbinom()</code>	<code>glm(family = binomial)</code>
	二項分布	<code>rbinom()</code>	<code>glm(family = binomial)</code>
	ポアソン分布	<code>rpois()</code>	<code>glm(family = poisson)</code>
	負の二項分布	<code>rnbinom()</code>	<code>glm.nb()</code> in <code>library(MASS)</code>
(連続)	ガンマ分布	<code>rgamma()</code>	<code>glm(family = gamma)</code>
	正規分布	<code>rnorm()</code>	<code>glm(family = gaussian)</code>

- `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

## さてさて、種子数の例題にもどって



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

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

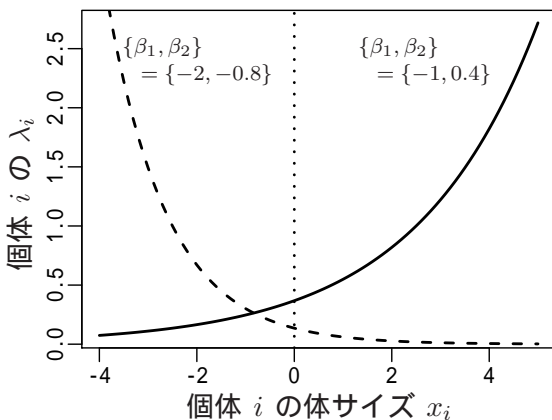
個体  $i$  の平均  $\lambda_i$  を以下のようにおいてみたらどうだろう.....?

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

- $\beta_1$  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 ← (直線の式)

個体  $i$  の平均  $\lambda_i$

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



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

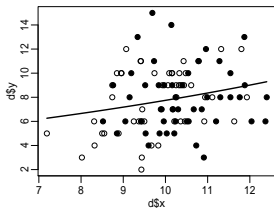
$$\log(\text{平均}) = \text{線形予測子}$$

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 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

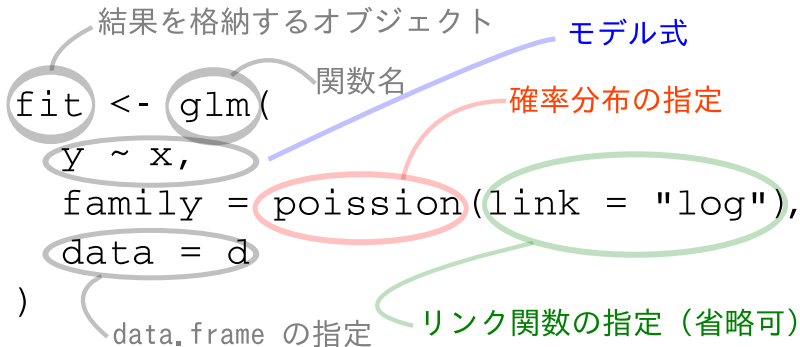
## glm() 関数の指定

```
> 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



- 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:  $\chi^2$  89.5

Residual Deviance: 85  $\chi^2$  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:
```

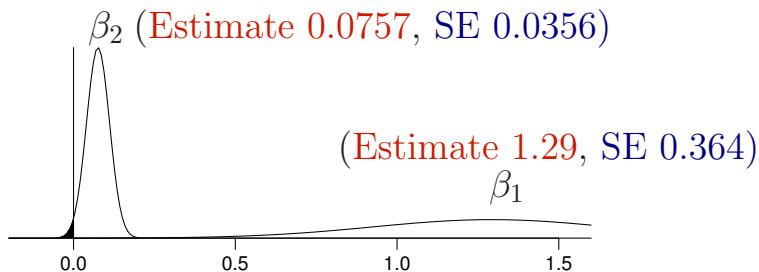
Min	1Q	Median	3Q	Max
-2.368	-0.735	-0.177	0.699	2.376

```
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) .....
```

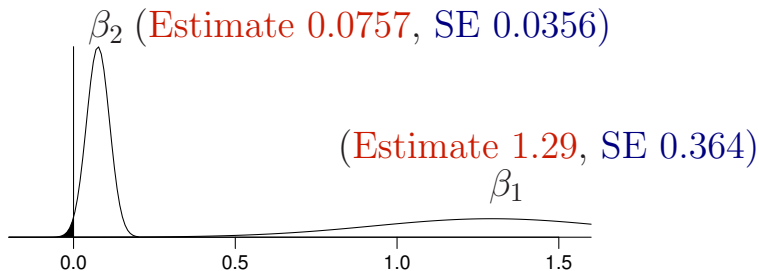
## 推定値と標準誤差のいめーじ (かなりいいかげんな説明)



- 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

(注: 頻度主義的な信頼区間の正しい解釈はもっとめんどくさい)

## 推定値と標準誤差のいめーじ (何がめんどくさいの?)



- 区間 95% 内に「ゼロ」があるとしよう → 「だから何？」
- 多数のパラメーターがある場合には？
- 授業の後半であつかうベイズ統計モデルでの解釈は **簡単** .....になるはず.....

## モデルの予測

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

```
...
```

```
Coefficients:
```

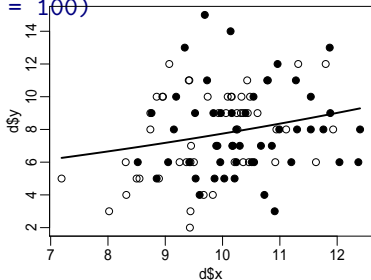
```
(Intercept)          x  
    1.2917         0.0757
```

```
> 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

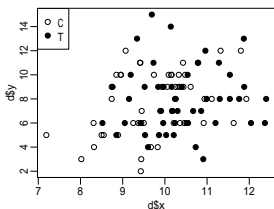


## 5. 処理をした・しなかった 効果も統計モデルに入れる

### 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

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

## glm(y ~ x + f, ...) の出力

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

Coefficients:

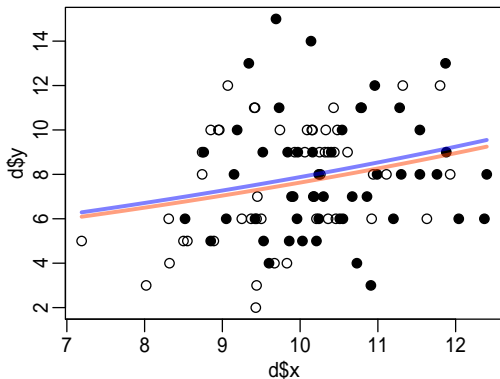
	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	1.2631	0.3696	3.42	0.00063
x	0.0801	0.0370	2.16	0.03062
fT	-0.0320	0.0744	-0.43	0.66703

..... (以下, 省略) .....



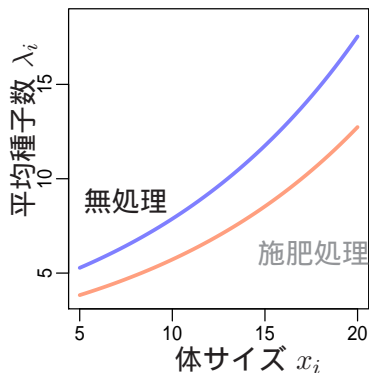
## x + f モデルの予測

```
> 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

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

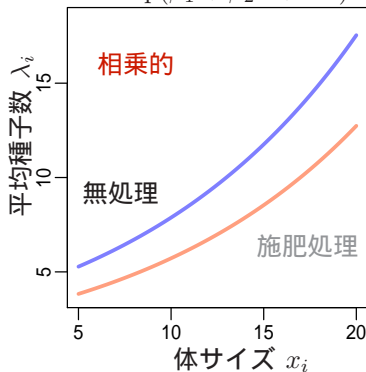


施肥効果である  $\exp(-0.032)$  は  
かけ算できくことに注意!

# リンク関数が違うとモデルの解釈が異なる

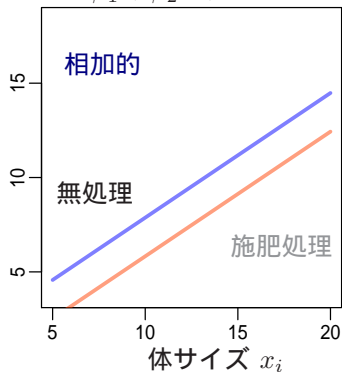
(A) 対数リンク関数

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



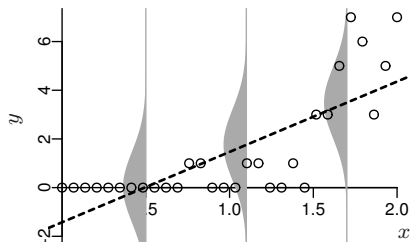
(B) 恒等リンク関数

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

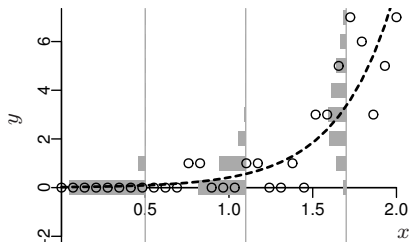


## GLM: 適切な確率分布 とリンク関数を選ぶ

正規分布・恒等リンク関数の統計モデル

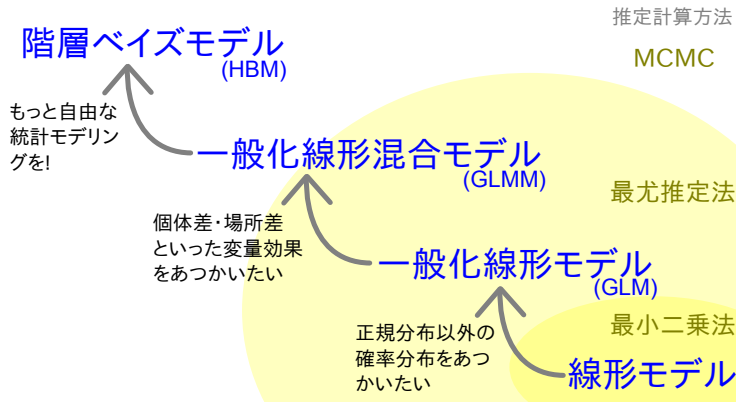


ポアソン分布・log リンク関数の統計モデル



# この講義であつかう統計モデルたち

## 線形モデルの発展



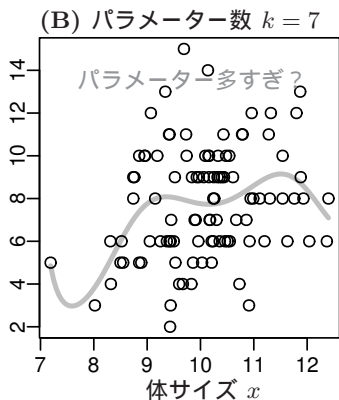
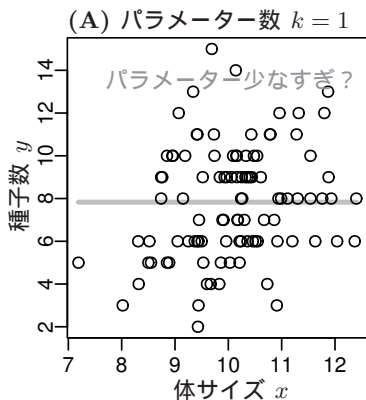
データの特徴にあわせて線形モデルを改良・発展させる

## 6. モデル選択

予測力のよい統計モデルはどれか？

予測の悪さの基準 AIC

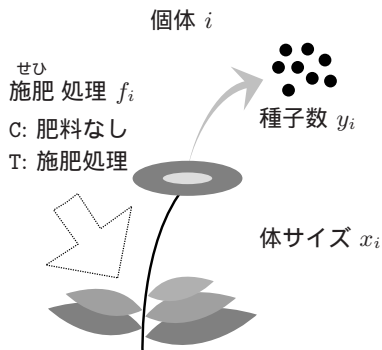
# パラメーター数 $k$ は多くても少なくてもヘン?



“**良いモデル**” とはなにか?  $k$  も重要なのか?

# 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

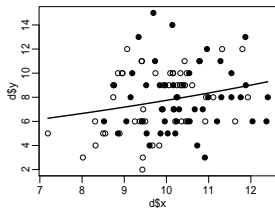
- 無処理 ( $f_i = \text{C}$ ): 50 sample ( $i \in \{1, 2, \dots, 50\}$ )
- 施肥処理 ( $f_i = \text{T}$ ): 50 sample ( $i \in \{51, 52, \dots, 100\}$ )



## この例題のための統計モデル

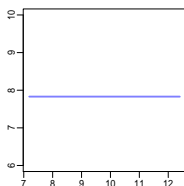
### ポアソン回帰のモデル

- 確率分布: **ポアソン分布**
- 線形予測子:  $\beta_1 + \beta_2 x_i + \beta_3 f_i$
- リンク関数: **対数リンク関数**



4 つの可能なモデル候補: (A) constant  $\lambda$ 

$$\lambda_i = \exp(\beta_1)$$

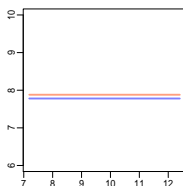


あてはまりの良さを対数尤度 (log likelihood) で評価する

```
> logLik(glm(y ~ 1, data = d, family = poisson))  
'log Lik.' -237.64 (df=1)
```

## 4 つの可能なモデル候補: (B) f model

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

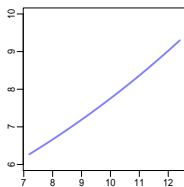


あてはまりの良さを対数尤度 (log likelihood) で評価する

```
> logLik(glm(y ~ f, data = d, family = poisson))  
'log Lik.' -237.63 (df=2)
```

## 4 つの可能なモデル候補: (C) x model

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

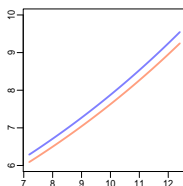


あてはまりの良さを対数尤度 (log likelihood) で評価する

```
> logLik(glm(y ~ x, data = d, family = poisson))  
'log Lik.' -235.39 (df=2)
```

4 つの可能なモデル候補: (D)  $x + f$  model

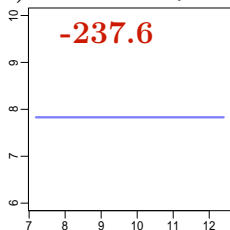
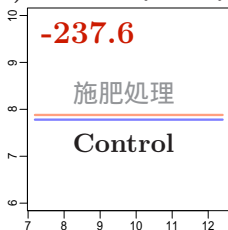
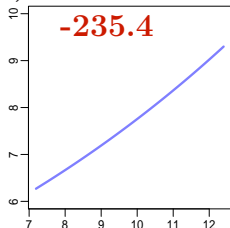
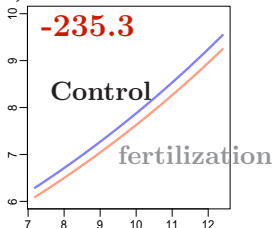
$$\lambda_i = \exp(\beta_1 + \beta_2 x_i + \beta_3 f_i)$$



あてはまりの良さを対数尤度 (log likelihood) で評価する

```
> logLik(glm(y ~ x + f, data = d, family = poisson))  
'log Lik.' -235.29 (df=3)
```

# パラメーター数が多いとあてはまりが良い

(A) constant  $\lambda$  ( $k = 1$ )(B) f model ( $k = 2$ )(C) x model ( $k = 2$ )(D) x + f model ( $k = 3$ )

## 7. AIC を使ったモデル選択

あてはまりの悪さ: deviance

そして予測の悪さ: AIC

## R の `glm()` は deviance を出力

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

```
Call:  glm(formula = y ~ x + f, family = poisson, data = d)
```

Coefficients:

(Intercept)	x	fT
1.2631	0.0801	-0.0320

```
Degrees of Freedom: 99 Total (i.e. Null); 97 Residual
```

```
Null Deviance: 89.5
```

```
Residual Deviance: 84.8 AIC: 477
```

Residual Deviance? Null Deviance? AIC?

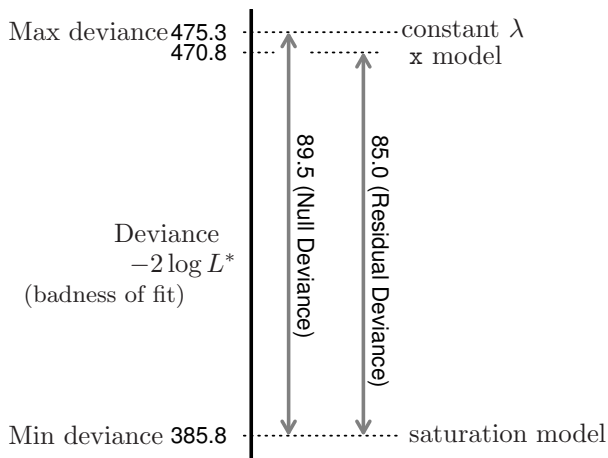


deviance  $D = -2 \times \log L^*$

- Maximum log likelihood  $\log L^*$ : goodness of fit
- Deviance  $D = -2 \log L^*$ : badness of fit

model	$k$	$\log L^*$	Deviance $-2 \log L^*$	Residual deviance
constant $\lambda$	1	-237.6	475.3	89.5
f	2	-237.6	475.3	89.5
x	2	-235.4	470.8	85.0
x + f	3	-235.3	470.6	84.8
saturation	100	-192.9	385.8	0.0

# Null deviance, Residual deviance, ...



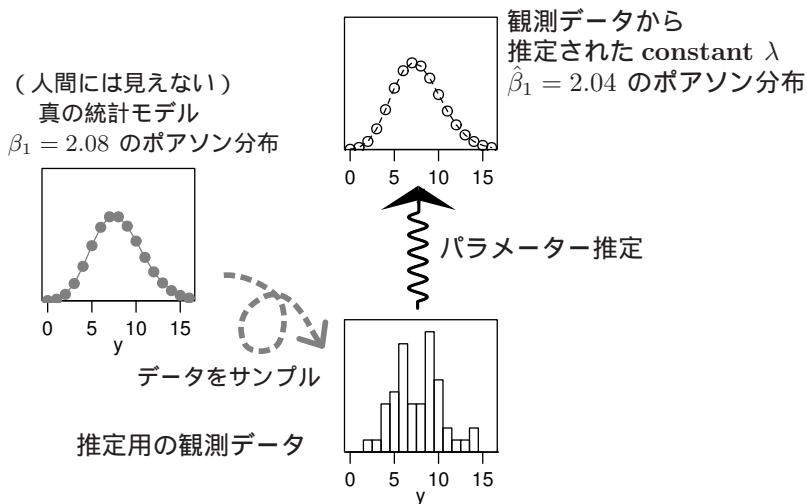
予測の悪さ:  $AIC = -2 \log L^* + 2k$

AIC 最小のモデルを選ぶ

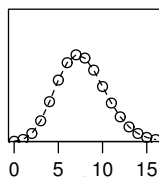
model	$k$	$\log L^*$	Deviance $-2 \log L^*$	Residual deviance	AIC
constant $\lambda$	1	-237.6	475.3	89.5	477.3
f	2	-237.6	475.3	89.5	479.3
<b>x</b>	<b>2</b>	<b>-235.4</b>	<b>470.8</b>	<b>85.0</b>	<b>474.8</b>
x + f	3	-235.3	470.6	84.8	476.6
saturation	100	-192.9	385.8	0.0	585.8

AIC: A (or Akaike) information criterion

## 統計モデルによる推測って何だっけ？



## 推定に使ったデータであてはまりを評価している?

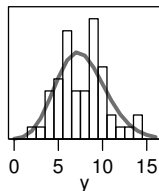


観測データから  
推定された constant  $\lambda$   
 $\hat{\beta}_1 = 2.04$  のポアソン分布



推定用の観測データを使って  
あてはまりの良さを評価

すると最大対数尤度  
 $\log L^*$  が得られる

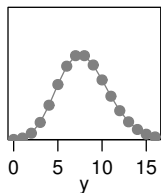


推定用の観測データ

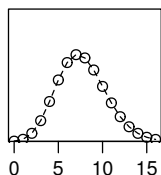
パラメーター推定に使った  
データなのであてはまりの  
良さにバイアスが生じる  
(過大評価)

# 重要なこと: 新データがあてはまるかどうか

(人間には見えない)  
真の統計モデル  
 $\beta_1 = 2.08$  のポアソン分布

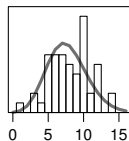
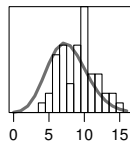
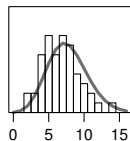


データ  
をサンプル  
(実際のデータ解析  
では不可能)



観測データから  
推定された constant  $\lambda$   
 $\hat{\beta}_1 = 2.04$  のポアソン分布

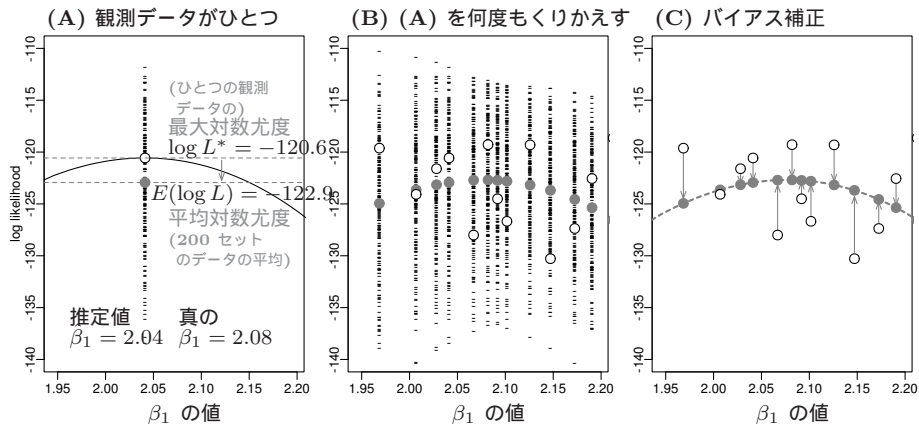
評価用のデータに  
あてはめてみる  
すると平均対数尤度  
 $E(\log L)$  が得られる



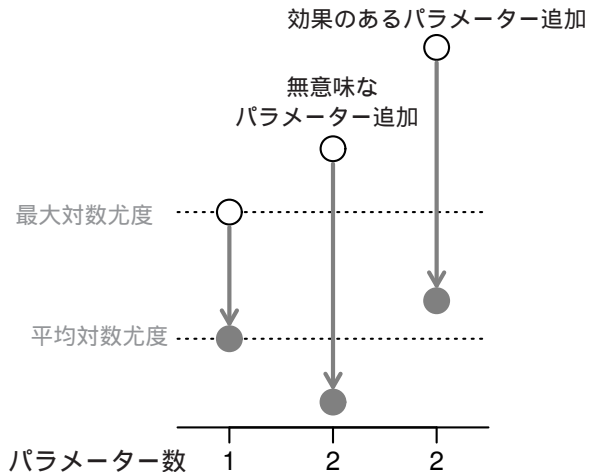
...

予測の良さ評価用のデータ (200 セット)

## シミュレーションで予測の良さを調べる



# バイアス補正を図示してみる





# FAQ モデル選択

(FAQ: frequently asked questions; よくある質問)

<http://hosho.ees.hokudai.ac.jp/~kubo/ce/FaqModelSelection.html>

# 次回予告

## 統計学的検定 と

## ロジスティック回帰 (GLM)