

統計モデリング入門 BSJ2015 (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

日本行動計量学会 春セミナー <http://goo.gl/vIdtcv>

2015-03-07

ファイル更新時刻: 2015-03-07 08:16

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
- ⑤ **AIC を使ったモデル選択**
あてはまりの悪さ: deviance
- ⑥ **統計学的な検定**
そして,その非対称性

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)
-

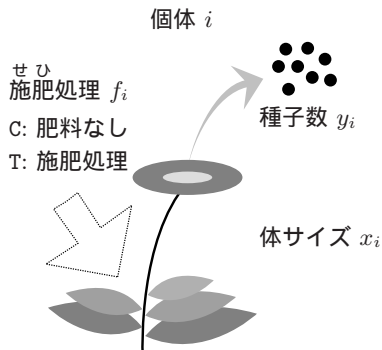
1. ポアソン回帰の例題: 架空植物の種子数データ

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
```

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"
```

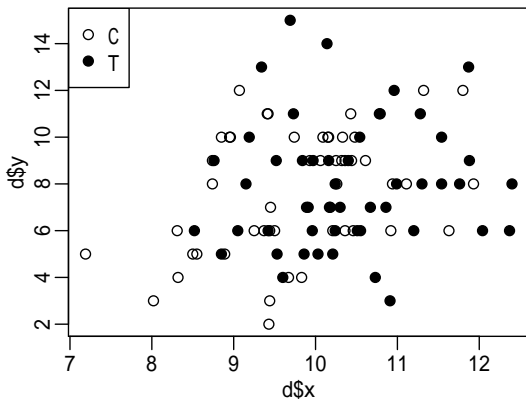
data frame の summary()

```
> 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	

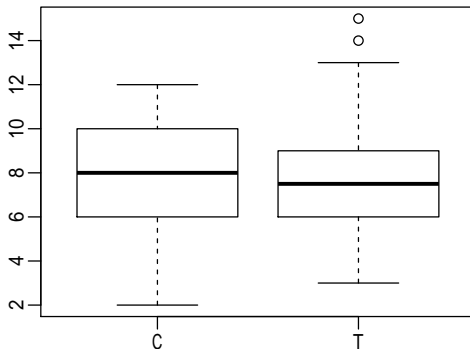
データはとにかく図示する!

```
> 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)
```



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

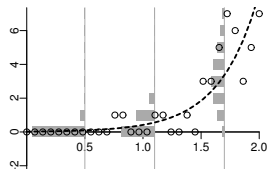
一般化線形モデル (GLM)

- 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

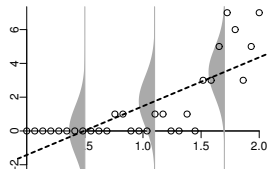
- 確率分布: **ポアソン分布**
- 線形予測子: e.g., $\beta_1 + \beta_2 x_i$
- リンク関数: **対数リンク関数**



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

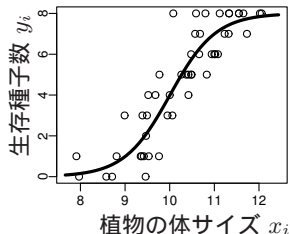
- 確率分布: 正規分布
- 線形予測子: e.g., $\beta_1 + \beta_2 x_i$
- リンク関数: 恒等リンク関数



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):

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

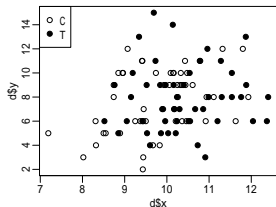
(交互作用項については粕谷さんが説明してくれます)

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 λ_i のポアソン分布にしたがうと
 しましょう

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

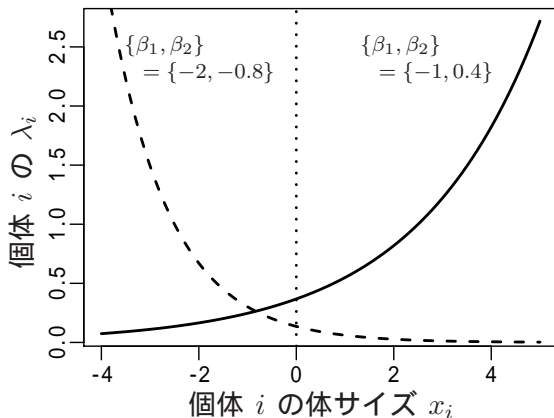
個体 i の平均 λ_i を以下のようにおいてみたらどうだろう.....?

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

- 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

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



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

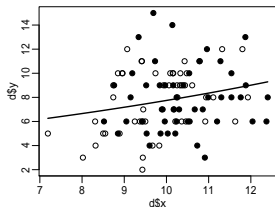
$$\log(\text{平均}) = \text{線形予測子}$$

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

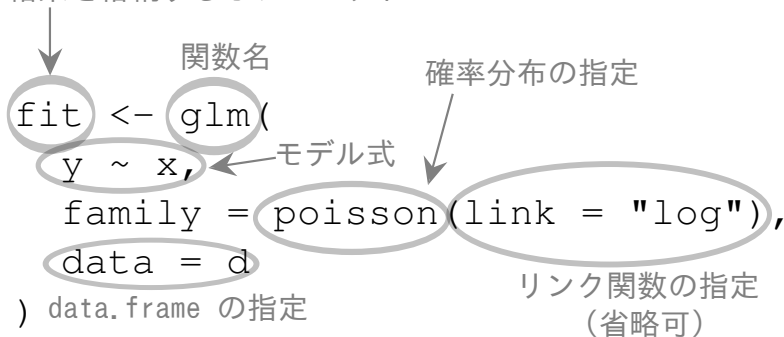
```
> d
      y      x f
1     6  8.31 C
2     6  9.44 C
3     6  9.50 C
... (中略) ...
99    7 10.86 T
100   9  9.97 T
```

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

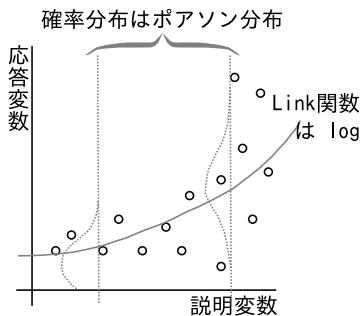


- モデル式 (線形予測子 z): どの説明変数を使うか?
- link 関数: z と応答変数 (y) **平均値** の関係は?
- family: どの確率分布を使うか?

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 $z = \beta_1 + \beta_2 x$
 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:
```

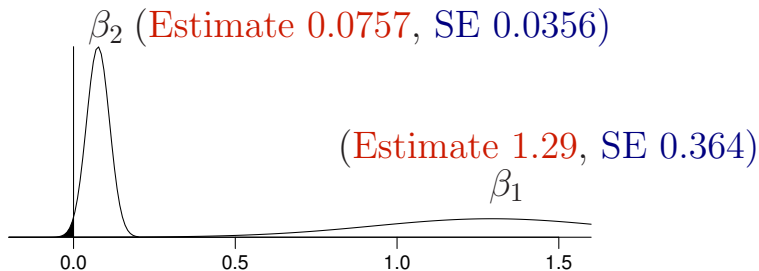
Min	1Q	Median	3Q	Max
-2.368	-0.735	-0.177	0.699	2.376

```
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



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:
```

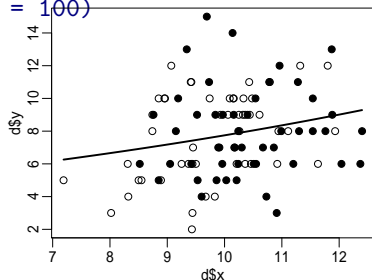
```
(Intercept)          x  
    1.2917         0.0757
```

```
> 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

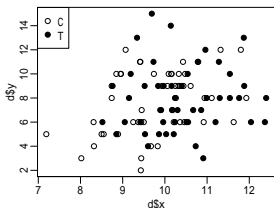


4. 処理をした・しなかった 効果も統計モデルに入れる

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 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)...
```

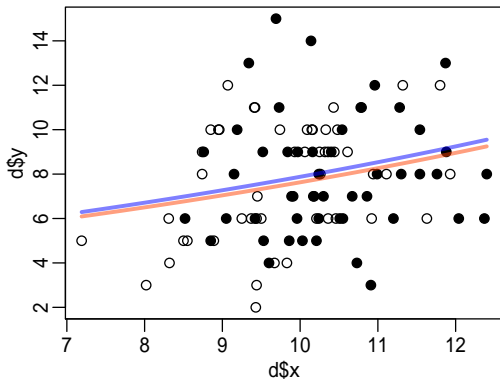
Coefficients:

	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)

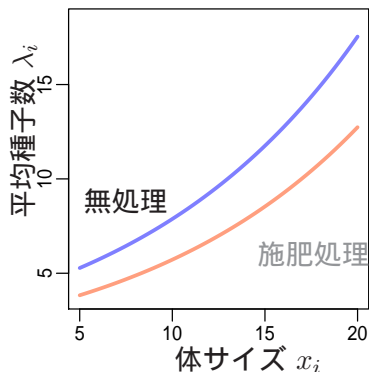
X + f モデルの予測

```
> 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

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

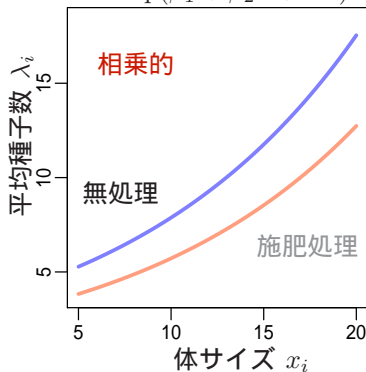


施肥効果である $\exp(-0.032)$ は
かけ算できくことに注意!

リンク関数が違うとモデルの解釈が異なる

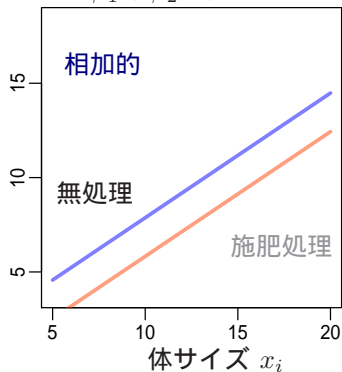
(A) 対数リンク関数

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



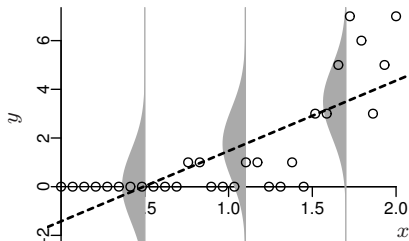
(B) 恒等リンク関数

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

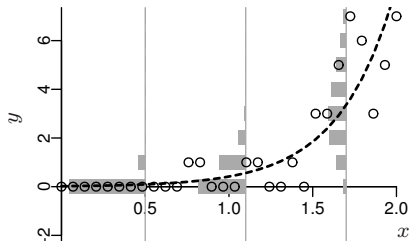


GLM: 適切な確率分布 とリンク関数を選ぶ

正規分布・恒等リンク関数の統計モデル

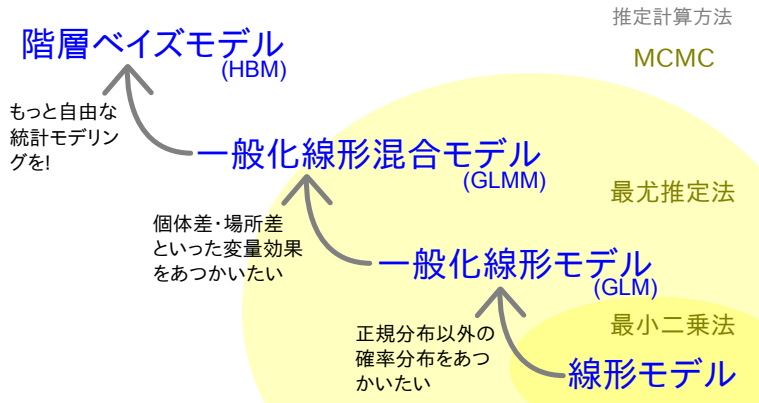


ポアソン分布・log リンク関数の統計モデル

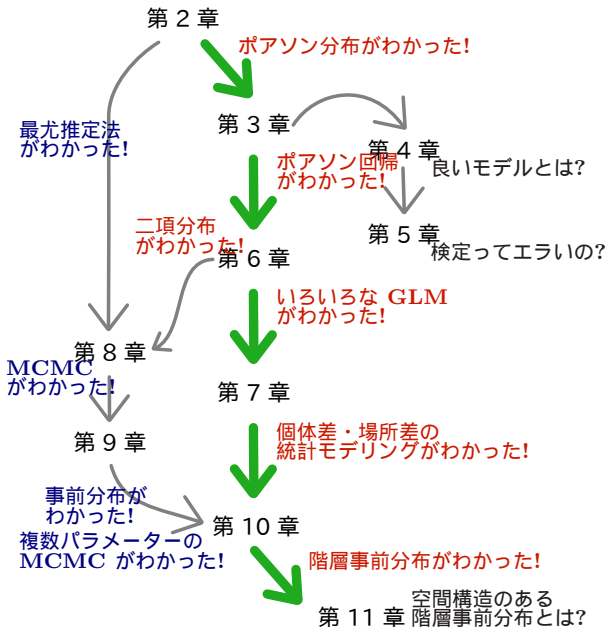


この授業であつかう統計モデルたち

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

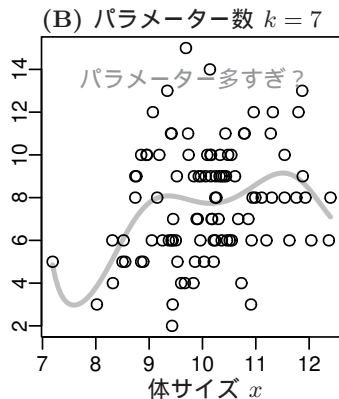
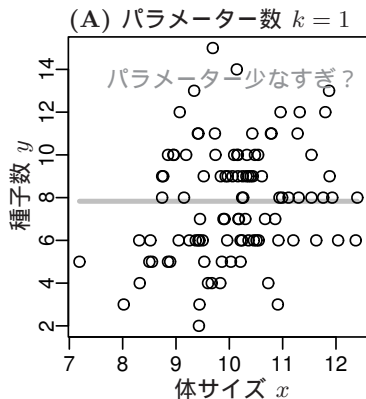


5. AIC を使ったモデル選択

あてはまりの悪さ: deviance

そして予測の悪さ: AIC

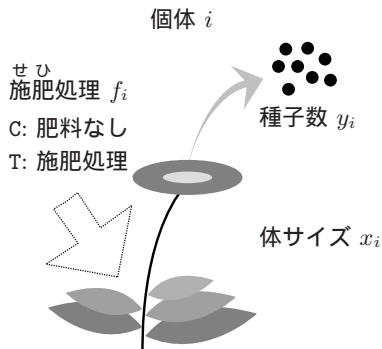
パラメーター数 k は多くても少なくてもヘン?



“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知る例題

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



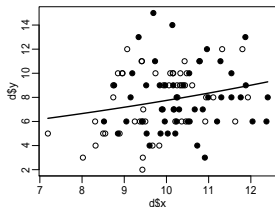
標本数

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

この例題のための統計モデル

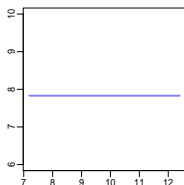
ポアソン回帰のモデル

- 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 λ

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

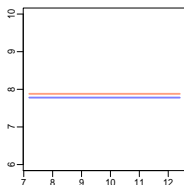


あてはまりの良さを対数尤度 (log likelihood) で評価する

```
> 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

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

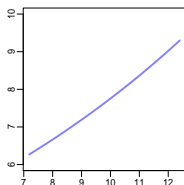


あてはまりの良さを対数尤度 (log likelihood) で評価する

```
> 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

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

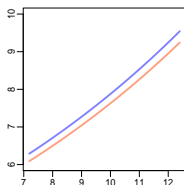


あてはまりの良さを対数尤度 (log likelihood) で評価する

```
> 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

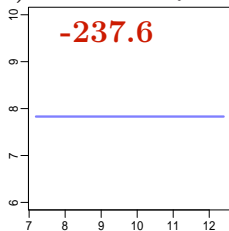
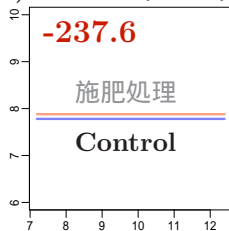
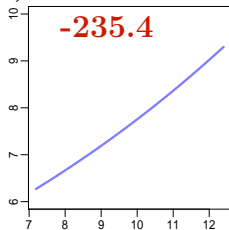
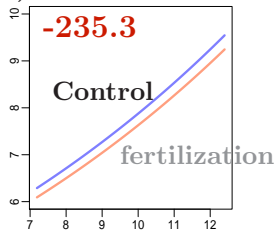
$$\lambda_i = \exp(\beta_1 + \beta_2 x_i + \beta_3 f_i)$$



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 λ ($k = 1$)(B) f model ($k = 2$)(C) x model ($k = 2$)(D) x + f model ($k = 3$)

「あてはまりの良さ」で
モデルを選んではいけない?

このあと、
あてはまりの悪さ: deviance
を紹介し
そして**予測の悪さ: AIC**
の使いかたを説明します

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
```

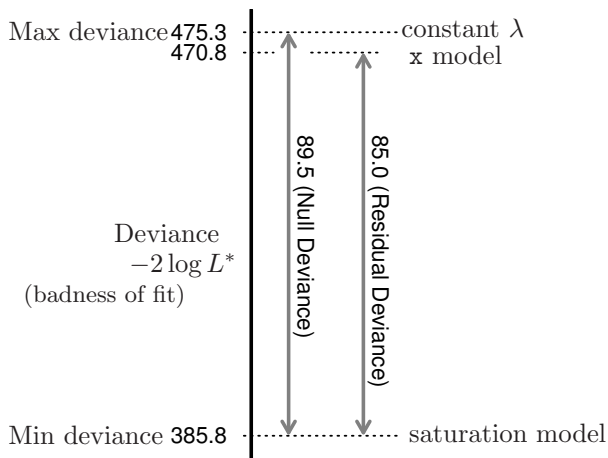
Residual Deviance? Null Deviance? AIC?

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

- 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 λ	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: $AIC = -2 \log L^* + 2k$

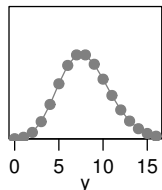
AIC 最小のモデルを選ぶ

model	k	$\log L^*$	Deviance $-2 \log L^*$	Residual deviance	AIC
constant λ	1	-237.6	475.3	89.5	477.3
f	2	-237.6	475.3	89.5	479.3
x	2	-235.4	470.8	85.0	474.8
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

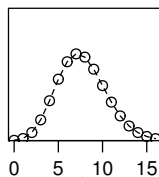
統計モデルによる推測って何だっけ？

(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



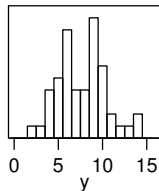
データをサンプル

推定用の観測データ

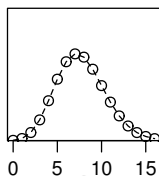


観測データから
推定された constant λ
 $\hat{\beta}_1 = 2.04$ のポアソン分布

パラメーター推定



推定に使ったデータであてはまりを評価している?

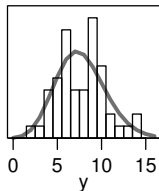


観測データから
推定された constant λ
 $\hat{\beta}_1 = 2.04$ のポアソン分布



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

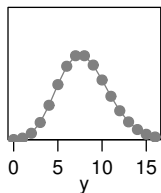


パラメーター推定に使った
データなのであてはまりの
良さにバイアスが生じる
(過大評価)

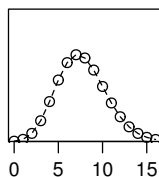
推定用の観測データ

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

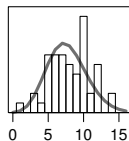
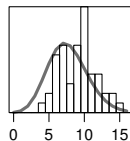
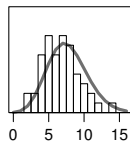


データ
をサンプル
(実際のデータ解析
では不可能)



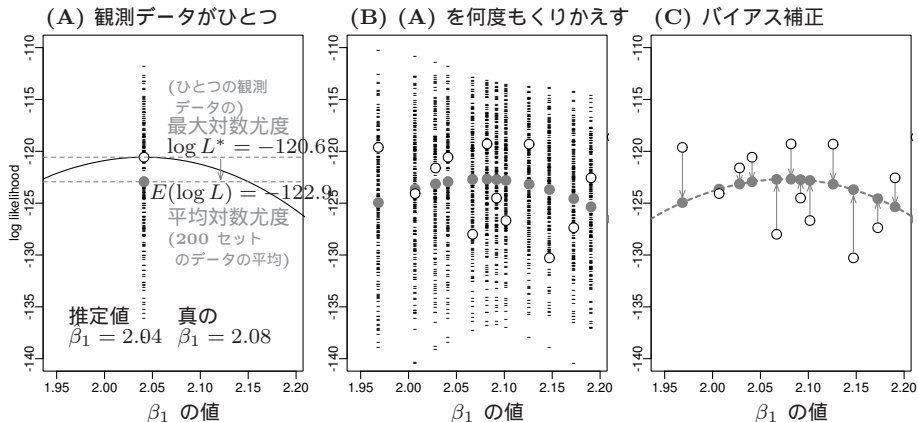
観測データから
推定された constant λ
 $\hat{\beta}_1 = 2.04$ のポアソン分布

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)

AIC の“罰金”って、なぜ $2 \times$ パラメーター数 (k)?

とりあえず「統計モデリング入門」第 4 章でも参照してください

6. 統計学的な検定

そして、その非対称性

これも今日のナガれの中では“わき道”なので

モデル選択 と統計学的検定 は その目的がぜんぜんちがう

目的?

モデル選択: よい予測をするモデルの
探索

統計学的検定: 帰無仮説の排除 (あとで
説明)

しかしモデル選択と検定の手順は途中まで同じ

統計モデルの検定

AIC によるモデル選択

解析対象のデータを確定



データを説明できるような統計モデルを設計

(帰無仮説・対立仮説)

(単純モデル・複雑モデル)



ネストした統計モデルたちのパラメーターの^{さいゆう}最尤推定計算



帰無仮説棄却の危険率を評価



モデル選択規準 AIC の評価



もし検定の結果が $P > 0.05$ だったら？

何も結論できない

λ 一定のモデルが良いとは言えない

検定の非対称性: **帰無仮説** はけっして
受容されない