

統計モデリング入門 2016 (d)

モデル選択と検定

久保拓弥 kubo@ees.hokudai.ac.jp

北大環境科学院の講義 <http://goo.gl/76c4i>

2016-07-20

ファイル更新時刻: 2016-07-25 12:48

今日のハナシ I

today's example: seed number data, again

① 前回と同じ例題: 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?

model selection using AIC

② AIC を使ったモデル選択

badness of fit

あてはまりの悪さ: deviance

statistical test

③ 統計学的な検定

and its asymmetry

そして, その非対称性

model selection

statistical test

④ モデル選択 と 統計学的な検定

misunderstanding

のさまざまな

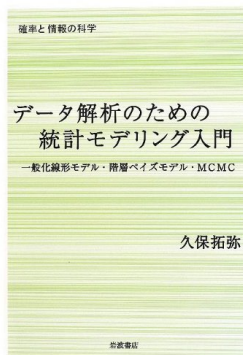
誤解

今日の内容と「統計モデリング入門」との対応

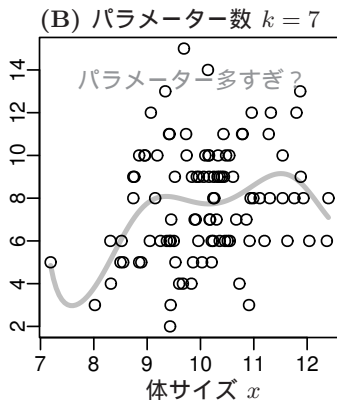
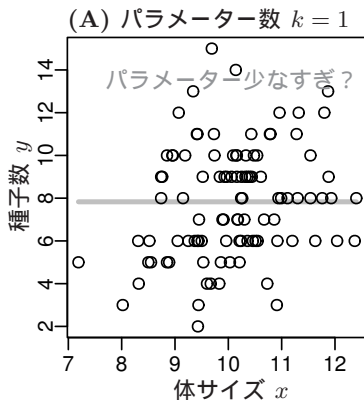
今日はおもに「**第4章 GLMのモデル選択**」と「**第5章 GLMの尤度比検定と検定の非対称性**」の内容を説明します。

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

<http://goo.gl/Ufq2>



パラメーター数は多くても少なくてもヘン?



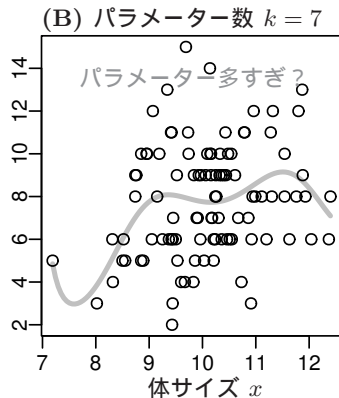
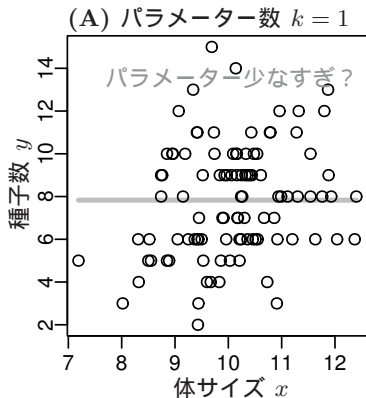
What is the “best?” parameter number k ?

today's example: seed number data, again

1. 前回と同じ例題: 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

パラメーター数 k は多くても少なくてもヘン?

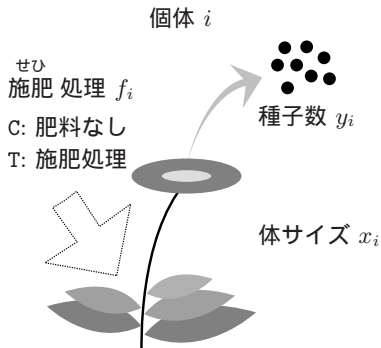
“良いモデル” とはなにか? k も重要なのか?

body size x and fertilization f change seed number y ? 個体サイズと実験処理の効果を調べる例題

- response variable seed number
 ● **応答変数** : 種子数 $\{y_i\}$
- explanatory variable
 ● **説明変数** :
 - body size
 ● 体サイズ $\{x_i\}$
 - fertilization
 ● 施肥処理 $\{f_i\}$

sample size
 標本数

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



a statistical model for this example

この例題のための統計モデル

ポアソン回帰のモデル

probability distribution Poisson distribution

- 確率分布: **ポアソン分布**

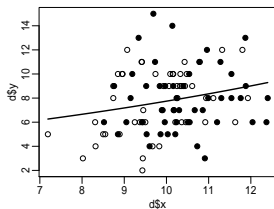
linear predictor

- 線形予測子: $\beta_1 + \beta_2 x_i + \beta_3 f_i$

link function

log link function

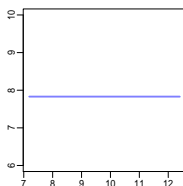
- リンク関数: **対数リンク関数**



4 candidate models

4 つの可能なモデル候補: (A) constant λ

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



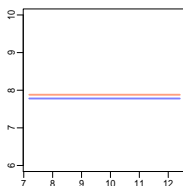
あてはまりの良さを対数尤度 (log likelihood) で評価する

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

4 candidate models

4 つの可能なモデル候補: (B) f model

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



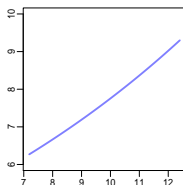
あてはまりの良さを対数尤度 (log likelihood) で評価する

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

4 candidate models

4 つの可能なモデル候補: (C) x model

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



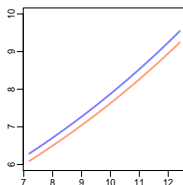
あてはまりの良さを対数尤度 (log likelihood) で評価する

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

4 candidate models

4 つの可能なモデル候補: (D) $x + f$ model

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

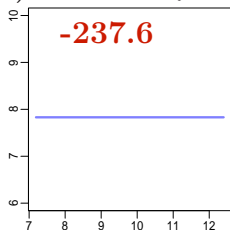
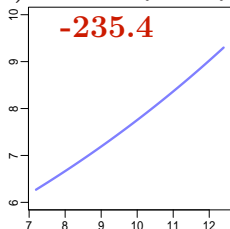
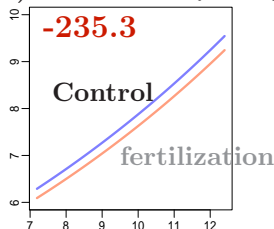


あてはまりの良さを対数尤度 (log likelihood) で評価する

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

k increases \rightarrow $\log L^*$ increases

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 ($k = 3$)

model selection using AIC

2. AIC を使ったモデル選択

badness of fit

あてはまりの悪さ: deviance

badness of prediction

そして予測の悪さ: AIC

output

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
```

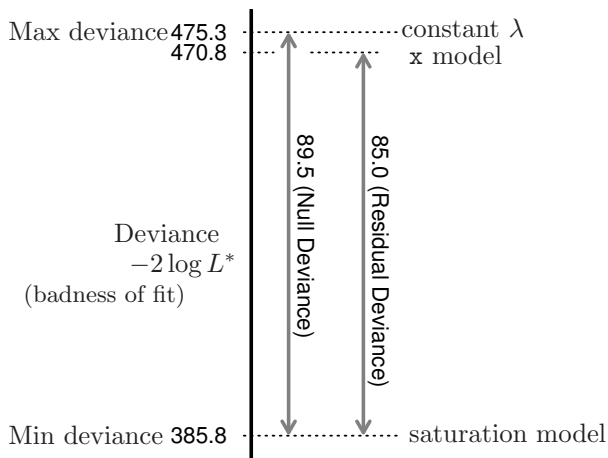
Residual Deviance? Null Deviance? AIC?

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

- 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, ...



badness of prediction

$$\text{予測の悪さ} : \text{AIC} = -2 \log L^* + 2k$$

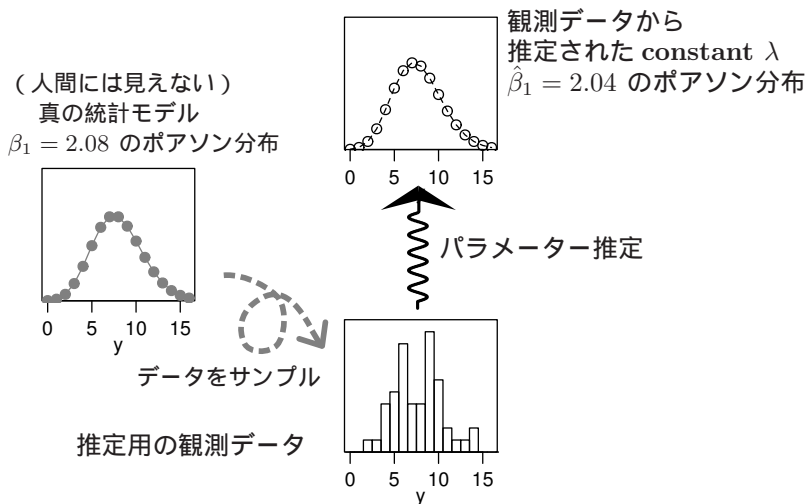
Look for a model of the smallest AIC

AIC 最小のモデルを選ぶ

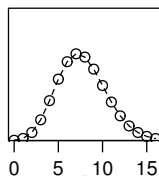
model	k	$\log L^*$	Deviance $-2 \log L^*$	Residual deviance	AIC
constant λ	1	-237.6	475.3	89.5	477.3
f	2	-237.6	475.3	89.5	479.3
x	2	-235.4	470.8	85.0	474.8
x + f	3	-235.3	470.6	84.8	476.6
saturation	100	-192.9	385.8	0.0	585.8

AIC: A (or Akaike) information criterion

統計モデルによる推測って何だっけ？



推定に使ったデータであてはまりを評価している？

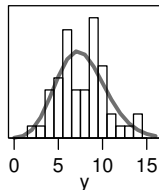


観測データから
推定された constant λ
 $\hat{\beta}_1 = 2.04$ のポアソン分布



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

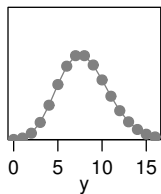


推定用の観測データ

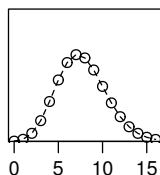
パラメーター推定に使った
データなのであてはまりの
良さにバイアスが生じる
(過大評価)

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

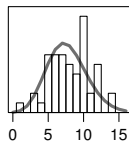
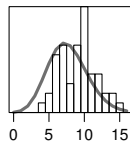
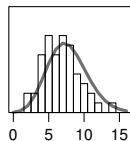


データ
をサンプル
(実際のデータ解析
では不可能)



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

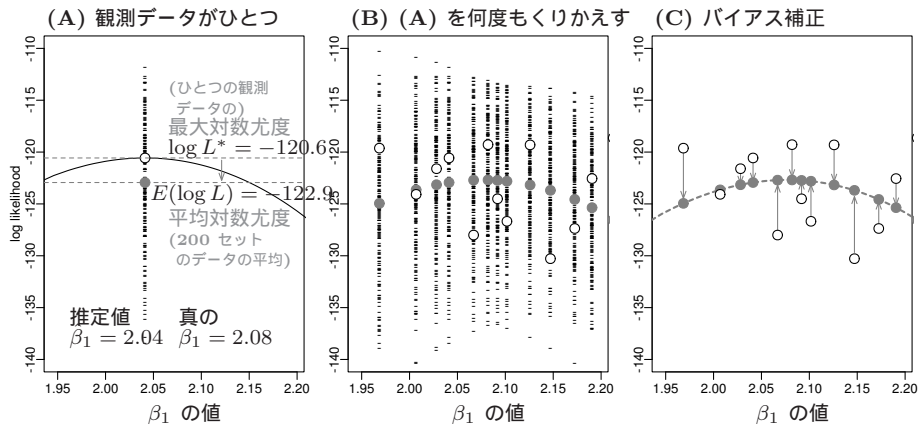
評価用のデータに
あてはめてみる
すると平均対数尤度
 $E(\log L)$ が得られる



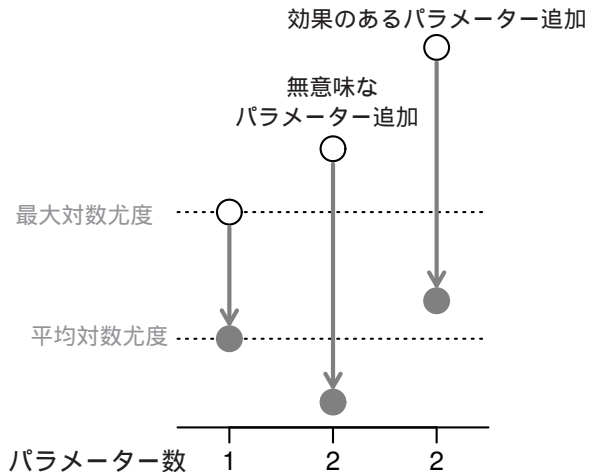
...

予測の良さ評価用のデータ (200 セット)

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



statistical test

3. 統計学的な検定

and its asymmetry

そして、その非対称性

likelihood ratio test

ここでは 尤度比検定 を紹介

But their procedures are similar

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



帰無仮説棄却の可否を判断

予測の良いモデルを選ぶ

model selection

statistical test

モデル選択 と 統計学的検定 は

totally different in their objectives

その目的がぜんぜんちがう

Objective

目的?

model selection

モデル選択:

Look for a model of better prediction

よい予測をするモデルの探索

statistical test

rejection of null hypothesis

統計学的検定: 帰無仮説の排除

統計学的な検定 (Neyman-Pearson framework)

statistical test



Null hypothesis

帰無仮説

$\text{glm}(y \sim 1)$
is better!

VS



Alternative hypothesis

対立仮説

$\text{glm}(y \sim x)$
is better!

どうでもいい
… 興味ない…

重要！これを
主張したい！

非対称性 asymmetry?

統計学的な検定 (Neyman-Pearson framework)

statistical
testNull
hypothesis

帰無仮説

 $\text{glm}(y \sim 1)$
is better!Alternative
hypothesis

対立仮説

 $\text{glm}(y \sim x)$
is better!

VS

test!



(if ...)

reject 棄却



support 支持

非対称性 asymmetricity?

統計学的な検定 (Neyman-Pearson framework)

statistical
testNull
hypothesis

帰無仮説

 $\text{glm}(y \sim 1)$
is better!Alternative
hypothesis

対立仮説

 $\text{glm}(y \sim x)$
is better!

VS

test!



(if ...)

NOT reject



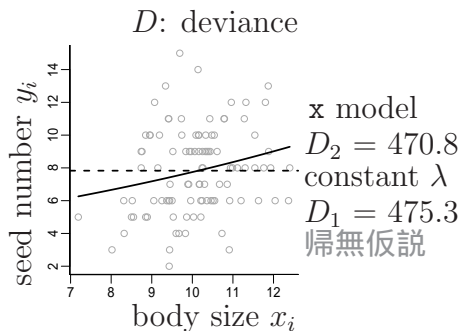
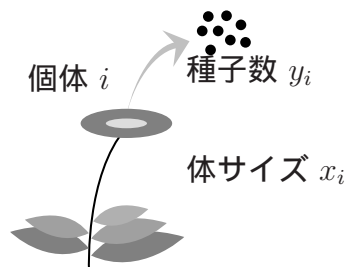
Say

Nothing!?

非対称性 asymmetry?

The same example, again



また同じ例題



neglect fertilization treatment
(施肥処理は無視!)

test statistics

検定統計量 $\Delta D_{1,2}$ difference in deviance $\Delta D_{1,2} = D_1 - D_2 = 4.51 \approx 4.5$ likelihood ratio? $-\log \frac{L_1^*}{L_2^*} = \log L_1^* - \log L_2^*$

model	k	$\log L^*$	Deviance $-2\log L^*$	
constant λ	1	-237.6	$D_1 = 475.3$	null hypothesis 帰無仮説 
x	2	-235.4	$D_2 = 470.8$	alternative hypothesis 対立仮説 

asymmetricity in test Null hypothesis is junk
検定の非対称性：帰無仮説はゴミあつかい
 ... yet we are focusing only on null hypothesis
にもかかわらず、**帰無仮説**だけをしつこく調べる




How to make null model 帰無仮説のつくりかた

対立仮説の中に帰無仮説がある (ネストした関係)

- カウントデータ $\{y_i\}$ は平均である λ_i のポアソン分布に従う
alternative hypothesis
- **対立仮説** の一例: $\log \lambda_i = \beta_1 + \beta_2 x_i$
- **帰無仮説** : $\log \lambda_i = \beta_1$ (切片だけのモデル)

objective null hypothesis rejection
 検定の目的: 帰無仮説  の 棄却

	observerd 観察された逸脱度差 $\Delta D_{1,2} = 4.5$ は.....	
帰無仮説は	「めったにない差」 (帰無仮説を棄却)	「よくある差」 (棄却できない)
真のモデルである	第一種の過誤	(問題なし)
真のモデルではない	(問題なし)	第二種の過誤

 is ...	significant (Reject )	not significant (Not reject )
TRUE	Type I error	(no problem)
NOT true	(no problem)	Type II error


asymmetricity in test evaluating only Type-I error
 検定の非対称性: 第一種の過誤だけに注目

generate $\Delta D_{1,2}$ distribution

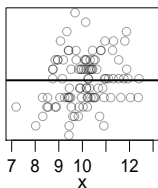
bootstrap likelihood test

 $\Delta D_{1,2}$ の分布を生成 : ブートストラップ尤度比検定

Suppose null hypothesis is TRUE!

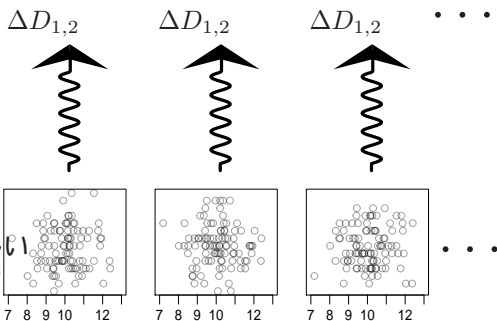
帰無仮説  が真のモデルであるとしてしよう!

帰無仮説が真の統計モデル
ということにしてしまう
($\hat{\beta}_1 = 2.06$ のポアソン分布)



帰無仮説のモデルから新しい
データをたくさん生成する

評価用データに constant λ と x model
をあてはめて逸脱度差 $\Delta D_{1,2}$ の分布を予測



あてはまりの良さ評価用のデータ (多数)



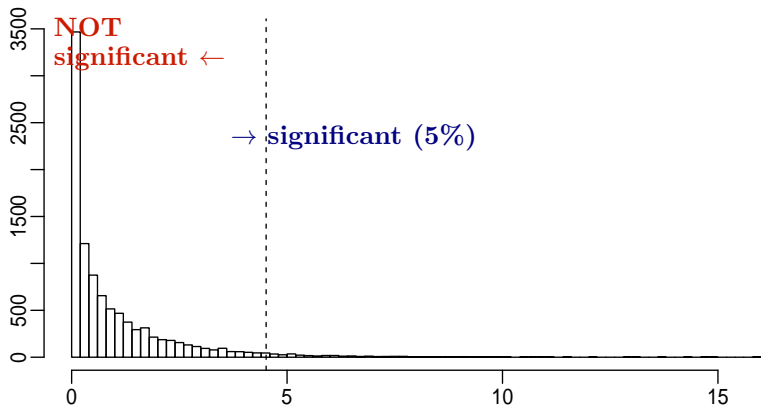
How to generate $\Delta D_{1,2}$ under is TRUE?

```
> d$y.rnd <- rpois(100, lambda = mean(d$y))
> fit1 <- glm(y.rnd ~ 1, data = d, family = poisson)
> fit2 <- glm(y.rnd ~ x, data = d, family = poisson)
> fit1$deviance - fit2$deviance
```

- generation of random numbers virtual data
• `rpois()` による ポアソン乱数の生成 (架空データ)
- fitting GLM to the virtual data
• 架空データを使って `glm()` あてはめ

You must define “rejection region” in advance
あらかじめ棄却域を決めておく

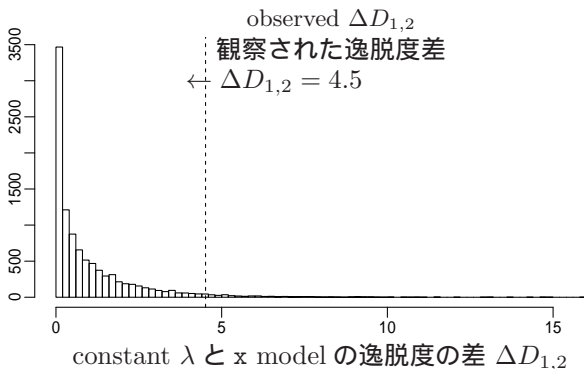
say, 5%?
たとえば 5% とか?



A random $\Delta D_{1,2}$ generator in R

```
get.dd <- function(d) # データの生成と逸脱度差の評価
{
  n.sample <- nrow(d) # データ数
  y.mean <- mean(d$y) # 標本平均
  d$y.rnd <- rpois(n.sample, lambda = y.mean)
  fit1 <- glm(y.rnd ~ 1, data = d, family = poisson)
  fit2 <- glm(y.rnd ~ x, data = d, family = poisson)
  fit1$deviance - fit2$deviance # 逸脱度の差を返す
}

pb <- function(d, n.bootstrap)
{
  replicate(n.bootstrap, get.dd(d))
}
```

Generated distribution of $\Delta D_{1,2} = D_1 - D_2$ 

(R code is in the next page)

$$\text{Probability}\{\Delta D_{1,2} \geq 4.5\} = \frac{38}{1000} = 0.038$$


```
> source("pb.R") # reading "pb.R" text file
> dd12 <- pb(d, n.bootstrap = 1000)
> hist(dd12, 100) # to plot histogram
> abline(v = 4.5, lty = 2)
> sum(dd12 >= 4.5)

[1] 38
```

so-called “ P -value” is 0.038.

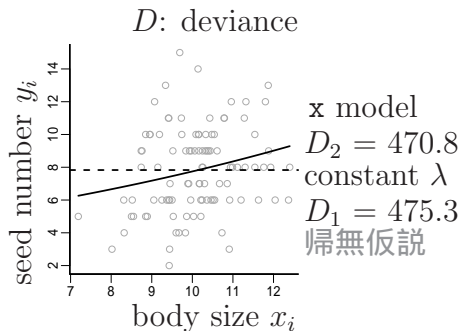
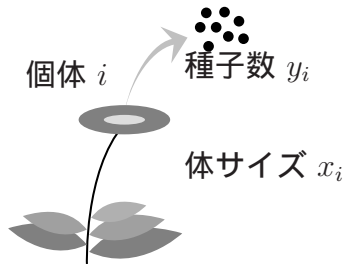
In this case, **帰無仮説**  **is rejected**

null hypothesis

So we can state that **対立仮説**  can be accepted.

alternative hypothesis

x model is better than constant λ .



In case that $P > 0.05$...?

No conclusion
何も結論できない

You can NOT state that constant λ is better
 λ 一定のモデルが良いとは言えない

Null hypothesis is never accepted

asymmetry in test

検定の非対称性 : 帰無仮説  はけっして受容されない

4. model selection statistical test
モデル選択 と 統計学的な検定

misunderstanding
のさまざまな 誤解

とりあえず FAQ モデル選択

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