

統計モデリング入門 2015 (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>

2015-07-15

ファイル更新時刻: 2015-07-14 16:48

kubostat2015d (<http://goo.gl/76c4i>) 統計モデリング入門 2015 (d) 2015-07-15 1 / 37

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?
- ② model selection using AIC
AIC を使ったモデル選択
badness of fit
あてはまりの悪さ: deviance
- ③ statistical test
統計学的な検定
and its asymmetry
そして, その非対称性

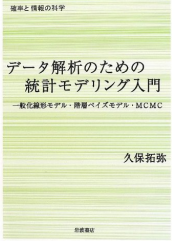
kubostat2015d (<http://goo.gl/76c4i>) 統計モデリング入門 2015 (d) 2015-07-15 2 / 37

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。

<http://goo.gl/Ufq2>

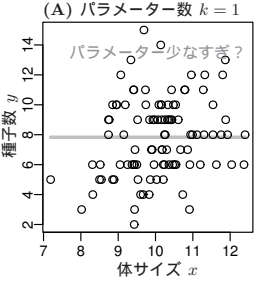
- 著者: 久保拓弥
- 出版社: 岩波書店
- 2012-05-18 刊行



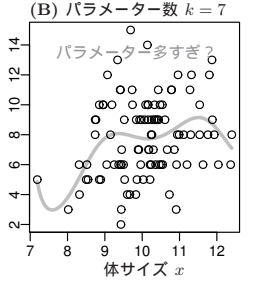
kubostat2015d (<http://goo.gl/76c4i>) 統計モデリング入門 2015 (d) 2015-07-15 3 / 37

パラメーター数は多くても少なくてもヘン?

(A) パラメーター数 $k=1$



(B) パラメーター数 $k=7$



What is the “best?” parameter number k ?

kubostat2015d (<http://goo.gl/76c4i>) 統計モデリング入門 2015 (d) 2015-07-15 4 / 37

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?

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

kubostat2015d (<http://goo.gl/76c4i>) 統計モデリング入門 2015 (d) 2015-07-15 5 / 37

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?

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

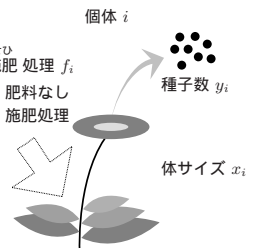
標本数

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 ($i \in \{51, 52, \dots, 100\}$)



kubostat2015d (<http://goo.gl/76c4i>) 統計モデリング入門 2015 (d) 2015-07-15 6 / 37

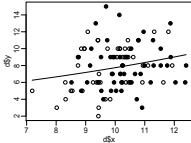
前回と同じ例題: 種子数データ 植物個体の属性, あるいは実験処理が種子数に影響?

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

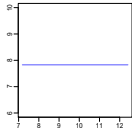
- 確率分布: **ポアソン分布**
- linear predictor
- 線形予測子: $\beta_1 + \beta_2 x_i + \beta_3 f_i$
- link function log link function
- リンク関数: **対数リンク関数**



kubostat2015d (<http://goo.gl/76c4i>) 統計モデリング入門 2015 (d) 2015-07-15 7 / 37

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?

4 candidate models
4 つの可能なモデル候補: (A) constant λ

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


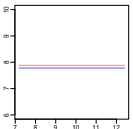
あてはまりの良さを対数尤度 (log likelihood) で評価する

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

kubostat2015d (<http://goo.gl/76c4i>) 統計モデリング入門 2015 (d) 2015-07-15 8 / 37

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?

4 candidate models
4 つの可能なモデル候補: (B) f model

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


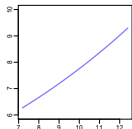
あてはまりの良さを対数尤度 (log likelihood) で評価する

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

kubostat2015d (<http://goo.gl/76c4i>) 統計モデリング入門 2015 (d) 2015-07-15 9 / 37

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?

4 candidate models
4 つの可能なモデル候補: (C) x model

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


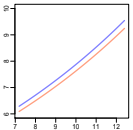
あてはまりの良さを対数尤度 (log likelihood) で評価する

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

kubostat2015d (<http://goo.gl/76c4i>) 統計モデリング入門 2015 (d) 2015-07-15 10 / 37

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?

4 candidate models
4 つの可能なモデル候補: (D) x + f model

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


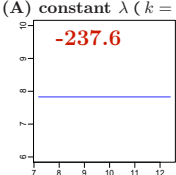
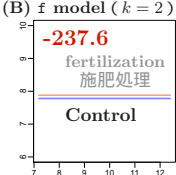
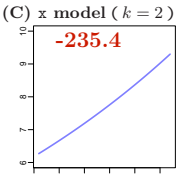
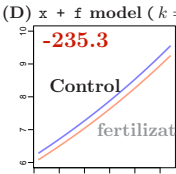
あてはまりの良さを対数尤度 (log likelihood) で評価する

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

kubostat2015d (<http://goo.gl/76c4i>) 統計モデリング入門 2015 (d) 2015-07-15 11 / 37

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?

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 ($k=3$)
	

kubostat2015d (<http://goo.gl/76c4i>) 統計モデリング入門 2015 (d) 2015-07-15 12 / 37

AIC を使ったモデル選択 あてはまりの悪さ : deviance

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

kubostat2015d (http://goo.gl/76c4i) 統計モデリング入門 2015 (d) 2015-07-15 13 / 37

AIC を使ったモデル選択 あてはまりの悪さ : deviance

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?

kubostat2015d (http://goo.gl/76c4i) 統計モデリング入門 2015 (d) 2015-07-15 14 / 37

AIC を使ったモデル選択 あてはまりの悪さ : deviance

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

kubostat2015d (http://goo.gl/76c4i) 統計モデリング入門 2015 (d) 2015-07-15 15 / 37

AIC を使ったモデル選択 あてはまりの悪さ : deviance

Null deviance, Residual deviance, ...

Max deviance 475.3
constant λ x model 470.8
Deviance $-2 \log L^*$ (badness of fit) 470.6
Min deviance 385.8
saturation model 0.0

89.5 (Null Deviance)
85.0 (Residual Deviance)

kubostat2015d (http://goo.gl/76c4i) 統計モデリング入門 2015 (d) 2015-07-15 16 / 37

AIC を使ったモデル選択 あてはまりの悪さ : deviance

badness of prediction

予測の悪さ : $AIC = -2 \log L^* + 2k$

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

kubostat2015d (http://goo.gl/76c4i) 統計モデリング入門 2015 (d) 2015-07-15 17 / 37

AIC を使ったモデル選択 あてはまりの悪さ : deviance

統計モデルによる推測って何だったけ?

(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

観測データから推定された constant λ $\beta_1 = 2.04$ のポアソン分布

データをサンプル

推定用の観測データ

パラメータ推定

kubostat2015d (http://goo.gl/76c4i) 統計モデリング入門 2015 (d) 2015-07-15 18 / 37

AICを使ったモデル選択 あてはまりの悪さ : deviance

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 $\beta_1 = 2.04$ 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

kubostat2015d (<http://goo.gl/76c4i>) 統計モデリング入門 2015 (d) 2015-07-15 19 / 37

AICを使ったモデル選択 あてはまりの悪さ : deviance

重要なこと: 「新データ」があてはまるかどうか

観測データから推定された constant λ $\beta_1 = 2.04$ のポアソン分布

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

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

データをサンプル (実際のデータ解析では不可能)

予測の良さ評価用のデータ (200 セット)

kubostat2015d (<http://goo.gl/76c4i>) 統計モデリング入門 2015 (d) 2015-07-15 20 / 37

AICを使ったモデル選択 あてはまりの悪さ : deviance

シミュレーションで予測の良さを調べる

(A) 観測データがひとつ (ひとつの観測データの最大対数尤度 $\log L^* = -120.6$)

(B) (A) を何度もくりかえす (平均対数尤度 (200 セットのデータの平均) $E(\log L) = -122.6$)

(C) バイアス補正 (推定値 $\beta_1 = 2.04$, 真の $\beta_1 = 2.08$)

kubostat2015d (<http://goo.gl/76c4i>) 統計モデリング入門 2015 (d) 2015-07-15 21 / 37

AICを使ったモデル選択 あてはまりの悪さ : deviance

バイアス補正を図示してみる

効果のあるパラメーター追加

無意味なパラメーター追加

最大対数尤度

平均対数尤度

パラメーター数 1 2 2

kubostat2015d (<http://goo.gl/76c4i>) 統計モデリング入門 2015 (d) 2015-07-15 22 / 37

統計学的な検定 そして、その非対称性

3. 統計学的な検定

and its asymmetricity
そして、その非対称性

likelihood ratio test
ここでは 尤度比検定 を紹介

kubostat2015d (<http://goo.gl/76c4i>) 統計モデリング入門 2015 (d) 2015-07-15 23 / 37

統計学的な検定 そして、その非対称性

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

kubostat2015d (<http://goo.gl/76c4i>) 統計モデリング入門 2015 (d) 2015-07-15 24 / 37

統計学的な検定 そして、その非対称性

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
described later
(あとで説明)

kubostat2015d (http://goo.gl/76c4i) 統計モデリング入門 2015 (d) 2015-07-15 25 / 37

統計学的な検定 そして、その非対称性

But their procedures are similar
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の **さいゆう** 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**

kubostat2015d (http://goo.gl/76c4i) 統計モデリング入門 2015 (d) 2015-07-15 26 / 37

統計学的な検定 そして、その非対称性

The same example, again
また同じ例題

個体 i 種子数 y_i
体サイズ x_i

D : deviance

seed number y_i
body size x_i

x model
 $D_2 = 470.8$
constant λ
 $D_1 = 475.3$
帰無仮説

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

kubostat2015d (http://goo.gl/76c4i) 統計モデリング入門 2015 (d) 2015-07-15 27 / 37

統計学的な検定 そして、その非対称性

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

asymmetry 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

kubostat2015d (http://goo.gl/76c4i) 統計モデリング入門 2015 (d) 2015-07-15 28 / 37

統計学的な検定 そして、その非対称性

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

observed
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」 「よくある差」
(帰無仮説を棄却) (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 H_0)	not significant (Not reject H_0)
TRUE	Type I error	(no problem)
NOT true	(no problem)	Type II error

asymmetry 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

kubostat2015d (http://goo.gl/76c4i) 統計モデリング入門 2015 (d) 2015-07-15 29 / 37

統計学的な検定 そして、その非対称性

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! 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 ($\beta_1 = 2.06$ のポアソン分布)

$\Delta D_{1,2}$ $\Delta D_{1,2}$ $\Delta D_{1,2}$...

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 あてはまりの良さ評価用のデータ (多数)

kubostat2015d (http://goo.gl/76c4i) 統計モデリング入門 2015 (d) 2015-07-15 30 / 37

統計学的な検定 そして、その非対称性

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

kubostat2015d (<http://goo.gl/76c4i>) 統計モデリング入門 2015 (d) 2015-07-15 31 / 37

統計学的な検定 そして、その非対称性

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?

kubostat2015d (<http://goo.gl/76c4i>) 統計モデリング入門 2015 (d) 2015-07-15 32 / 37

統計学的な検定 そして、その非対称性

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))
}
    
```

kubostat2015d (<http://goo.gl/76c4i>) 統計モデリング入門 2015 (d) 2015-07-15 33 / 37

統計学的な検定 そして、その非対称性

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

constant λ と x model の逸脱度の差 $\Delta D_{1,2}$

(R code is in the next page)

kubostat2015d (<http://goo.gl/76c4i>) 統計モデリング入門 2015 (d) 2015-07-15 34 / 37

統計学的な検定 そして、その非対称性

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.

kubostat2015d (<http://goo.gl/76c4i>) 統計モデリング入門 2015 (d) 2015-07-15 35 / 37

統計学的な検定 そして、その非対称性

In this case, 帰無仮説 is rejected

alternative hypothesis
 So we can state that 対立仮説 can be accepted.
 x model is better than constant λ .

D: deviance

x model $D_2 = 470.8$
 constant λ $D_1 = 475.3$
 帰無仮説

個体 i 種子数 y_i 体サイズ x_i

kubostat2015d (<http://goo.gl/76c4i>) 統計モデリング入門 2015 (d) 2015-07-15 36 / 37

統計学的な検定 そして、その非対称性

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

kubostat2015d (<http://geo.g1/76c41>) 統計モデリング入門 2015 (d) 2015-07-15 37 / 37