

統計モデリング入門 2017 (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>

2019-07-29

ファイル更新時刻: 2019-07-20 08:16

kubostat2017d (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (d) 2019-07-29 1 / 44

もくじ

今日のハナシ I

- 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 AIC を使ったモデル選択
あてはまりの悪さ: deviance
- ③ 統計学的な検定
そして,その非対称性
- ④ モデル選択 と 統計学的な検定
のさまざまな誤解

kubostat2017d (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (d) 2019-07-29 2 / 44


もくじ

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



データ解析のための
統計モデリング入門
— 一般化線形モデル・隠層ベイズモデル・MCMC —
久保拓弥

kubostat2017d (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (d) 2019-07-29 3 / 44

もくじ

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 $k = 7$

How many parameters do you need for the best prediction?

kubostat2017d (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (d) 2019-07-29 4 / 44

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?

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

kubostat2017d (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (d) 2019-07-29 5 / 44

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?

パラメーター数 k 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 $k = 7$

“良いモデル”? number of parameters k ?

kubostat2017d (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (d) 2019-07-29 6 / 44

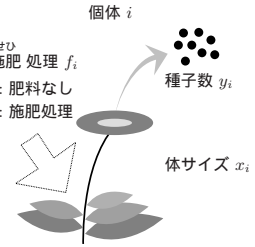
前回と同じ例題: 種子数データ 植物個体の属性, あるいは実験処理が種子数に影響?

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



個体 i
種子数 y_i
体サイズ x_i
せひ 施肥 処理 f_i
C: 肥料なし
T: 施肥処理

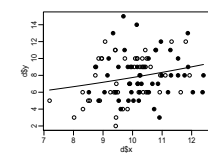
kubostat2017d (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (d) 2019-07-29 7 / 44

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?

この例題のための統計モデル

ポアソン回帰のモデル

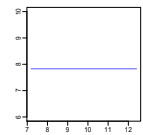
- 確率分布: ポアソン分布
- 線形予測子: $\beta_1 + \beta_2 x_i + \beta_3 f_i$
- リンク関数: 対数リンク関数



kubostat2017d (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (d) 2019-07-29 8 / 44

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 つの可能なモデル候補: (A) constant λ

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


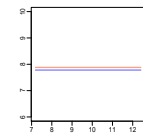
あてはまりの良さを対数尤度 (log likelihood) で評価する

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

kubostat2017d (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (d) 2019-07-29 9 / 44

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 つの可能なモデル候補: (B) f model

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


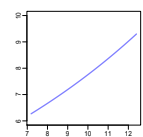
あてはまりの良さを対数尤度 (log likelihood) で評価する

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

kubostat2017d (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (d) 2019-07-29 10 / 44

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 つの可能なモデル候補: (C) x model

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


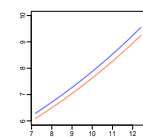
あてはまりの良さを対数尤度 (log likelihood) で評価する

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

kubostat2017d (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (d) 2019-07-29 11 / 44

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 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)
```

kubostat2017d (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (d) 2019-07-29 12 / 44

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

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 ($k = 3$)

kubostat2017d (http://goo.gl/76c4i) 統計モデリング入門 2017 (d) 2019-07-29 13 / 44

AIC を使ったモデル選択 あてはまりの悪さ: deviance

2. 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

kubostat2017d (http://goo.gl/76c4i) 統計モデリング入門 2017 (d) 2019-07-29 14 / 44

AIC を使ったモデル選択 あてはまりの悪さ: deviance

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?

kubostat2017d (http://goo.gl/76c4i) 統計モデリング入門 2017 (d) 2019-07-29 15 / 44

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

kubostat2017d (http://goo.gl/76c4i) 統計モデリング入門 2017 (d) 2019-07-29 16 / 44

AIC を使ったモデル選択 あてはまりの悪さ: deviance

Null deviance, Residual deviance, ...

kubostat2017d (http://goo.gl/76c4i) 統計モデリング入門 2017 (d) 2019-07-29 17 / 44

AIC を使ったモデル選択 あてはまりの悪さ: 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

kubostat2017d (http://goo.gl/76c4i) 統計モデリング入門 2017 (d) 2019-07-29 18 / 44

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 (estimation) 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

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

parameter estimation
パラメーター推定

データをサンプル

推定用の観測データ

kubostat2017d (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (d) 2019-07-29 19 / 44

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であてはまりの悪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) **biased "goodness of fit"!**

kubostat2017d (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (d) 2019-07-29 20 / 44

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

(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 (実際のデータ解析では不可能)

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)

kubostat2017d (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (d) 2019-07-29 21 / 44

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 $\log L^* = -120$

(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) $E(\log L) = -122.9$

(C) バイアス補正

log likelihood β_1 の値

推定値 $\beta_1 = 2.04$ 真の $\beta_1 = 2.08$

kubostat2017d (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (d) 2019-07-29 22 / 44

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

kubostat2017d (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (d) 2019-07-29 23 / 44

統計学的な検定 そして、その非対称性

3. 統計学的な検定

そして、その非対称性

ここでは 尤度比検定 を紹介

kubostat2017d (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (d) 2019-07-29 24 / 44

統計学的な検定 そして、その非対称性

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

kubostat2017d (http://goo.gl/76c4i) 統計モデリング入門 2017 (d) 2019-07-29 25 / 44

統計学的な検定 そして、その非対称性

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

kubostat2017d (http://goo.gl/76c4i) 統計モデリング入門 2017 (d) 2019-07-29 26 / 44

統計学的な検定 そして、その非対称性

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

kubostat2017d (http://goo.gl/76c4i) 統計モデリング入門 2017 (d) 2019-07-29 27 / 44

統計学的な検定 そして、その非対称性

統計学的な検定 (Neyman-Pearson framework)

statistical test



Null hypothesis
帰無仮説
 $glm(y \sim 1)$
is better!

どうでもいい
… 興味ない…

VS



Alternative hypothesis
対立仮説
 $glm(y \sim x)$
is better!

重要! これを
主張したい!


非対称性 asymmetry?

kubostat2017d (http://goo.gl/76c4i) 統計モデリング入門 2017 (d) 2019-07-29 28 / 44

統計学的な検定 そして、その非対称性


統計学的な検定 (Neyman-Pearson framework)

statistical test



Null hypothesis
帰無仮説
 $glm(y \sim 1)$
is better!

VS



Alternative hypothesis
対立仮説
 $glm(y \sim x)$
is better!

test! ↓

(if ...) reject 棄却 ----- support 支持

非対称性 asymmetry?

kubostat2017d (http://goo.gl/76c4i) 統計モデリング入門 2017 (d) 2019-07-29 29 / 44

統計学的な検定 そして、その非対称性

統計学的な検定 (Neyman-Pearson framework)

statistical test



Null hypothesis
帰無仮説
 $glm(y \sim 1)$
is better!

VS



Alternative hypothesis
対立仮説
 $glm(y \sim x)$
is better!

test! ↓

(if ...) NOT reject ----- Say Nothing!?

非対称性 asymmetry?

kubostat2017d (http://goo.gl/76c4i) 統計モデリング入門 2017 (d) 2019-07-29 30 / 44

統計学的な検定 そして、その非対称性

また同じ例題 The same example, again

個体 i

種子数 y_i

体サイズ x_i

(施肥処理は無視!)

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

kubostat2017d (http://goo.gl/76c4i) 統計モデリング入門 2017 (d) 2019-07-29 31 / 44

統計学的な検定 そして、その非対称性

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	$D_1 = 475.3$ 帰無仮説
x	2	-235.4	$D_2 = 470.8$ 対立仮説

検定の非対称性: 帰無仮説はゴミあつかい
にもかかわらず、帰無仮説だけをしつこく調べる

kubostat2017d (http://goo.gl/76c4i) 統計モデリング入門 2017 (d) 2019-07-29 32 / 44

統計学的な検定 そして、その非対称性

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
- 対立仮説の一例: $\log \lambda_i = \beta_1 + \beta_2 x_i$
- ネストした 帰無仮説: $\log \lambda_i = \beta_1$ (切片だけのモデル)

kubostat2017d (http://goo.gl/76c4i) 統計モデリング入門 2017 (d) 2019-07-29 33 / 44

統計学的な検定 そして、その非対称性

検定の目的: 帰無仮説の棄却

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)	(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

検定の非対称性: 第一種の過誤だけに注目

kubostat2017d (http://goo.gl/76c4i) 統計モデリング入門 2017 (d) 2019-07-29 34 / 44

統計学的な検定 そして、その非対称性

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

帰無仮説 が真のモデルであるとして!

帰無仮説が真の統計モデルということにしてしまう ($\beta_1 = 2.06$ のポアソン分布)

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

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)

kubostat2017d (http://goo.gl/76c4i) 統計モデリング入門 2017 (d) 2019-07-29 35 / 44

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
    
```

- `rpois()` によるポアソン乱数の生成 (架空データ)
- 架空データを使って `glm()` あてはめ

kubostat2017d (http://goo.gl/76c4i) 統計モデリング入門 2017 (d) 2019-07-29 36 / 44

統計学的な検定 そして、その非対称性

あらかじめ棄却域を決めておく

たとえば 5% とか?

3500
2500
1500
500
0

0 5 10 15

NOT significant ←

→ significant (5%)

kubostat2017d (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (d) 2019-07-29 37 / 44

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))
}
```

kubostat2017d (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (d) 2019-07-29 38 / 44

統計学的な検定 そして、その非対称性

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

3500
2500
1500
500
0

0 5 10 15

観察された逸脱度差

$\Delta D_{1,2} = 4.5$

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

(R code is in the next page)

kubostat2017d (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (d) 2019-07-29 39 / 44

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.

kubostat2017d (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (d) 2019-07-29 40 / 44

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

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

個体 i

種子数 y_i

体サイズ x_i

seed number y_i

body size x_i

D : deviance

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

kubostat2017d (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (d) 2019-07-29 41 / 44

統計学的な検定 そして、その非対称性

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

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

kubostat2017d (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (d) 2019-07-29 42 / 44

モデル選択 と 統計学的な検定 のさまざまな誤解

4. モデル選択 と 統計学的な検定

のさまざまな誤解

kubostat2017d (<http://geo.g1/76c41>) 統計モデリング入門 2017 (d) 2019-07-29 43 / 44

モデル選択 と 統計学的な検定 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>

kubostat2017d (<http://geo.g1/76c41>) 統計モデリング入門 2017 (d) 2019-07-29 44 / 44