『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・階層ベイズモデル・MCMC』
pp. 67~ 92

4章　GLMのモデル選択
ーAICとモデルの予測の良さー

2014/12/9

担当　日髙

4章　GLMのモデル選択 ―AICとモデルの予測の良さ―

今回のテーマ

「よい統計モデルとは何だろう？」
「よいモデルを選び出すには？」

よい統計モデルを選ぶこと　：　モデル選択
（model selection）

この章では･･･

モデル選択の規準を紹介

下準備

R

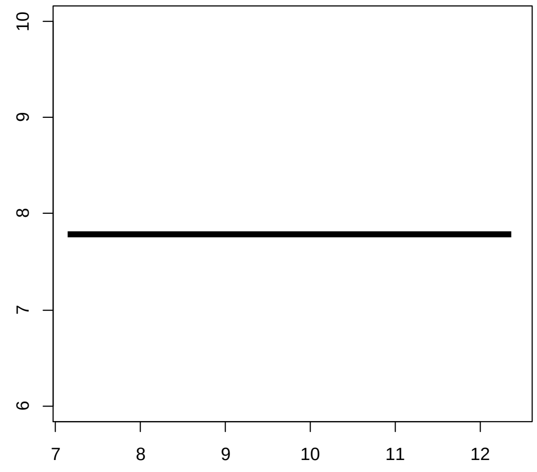
・使用データファイル：Kubo2012　chapter 3

☆ディレクトリはchapter 3のままでOK！

一番適しているのはどのモデル？

最大対数尤度（maximum log likelifood）、

つまり「あてはまりの良さ」をみればよい？



(A) 一定モデル(k=1)　：

説明変数のない、切片β_1だけのモデル
Rでは“null model”とよばれる

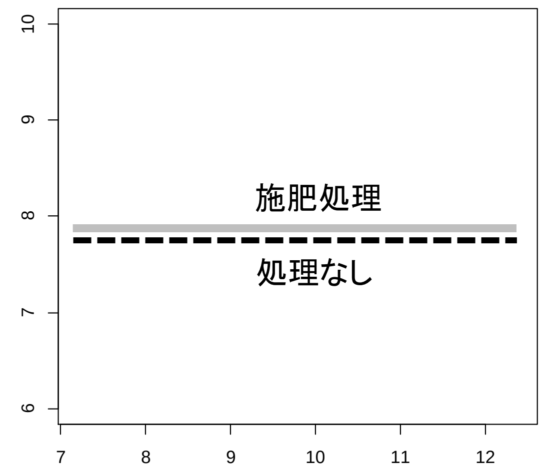
（第4章　4.2項　参照）

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

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

最大対数尤度

（B）fモデル（k＝2）　：

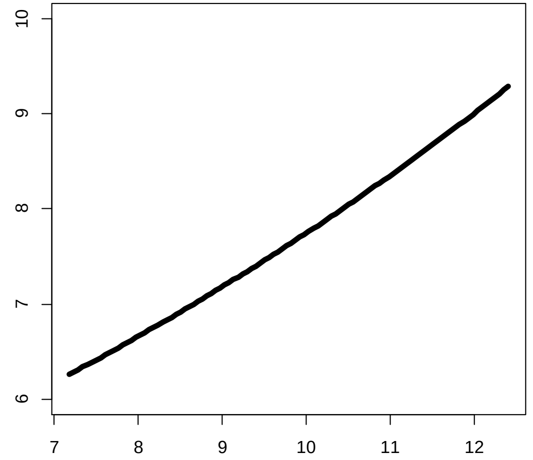


平均種子数λ_iが施肥処理f_iだけに依存するモデル

（第3章　3.5項　参照）

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

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



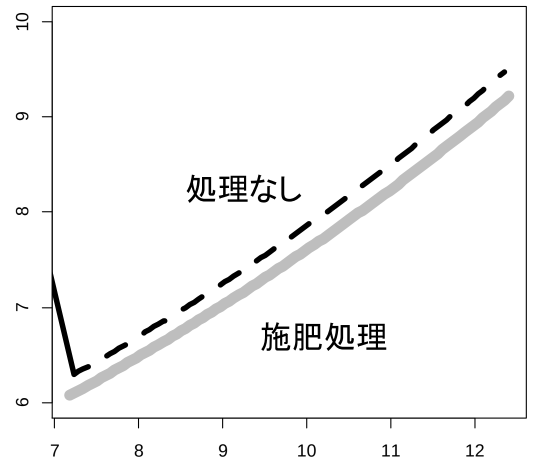
（C）Xモデル（k＝2）　：

平均種子数λ_iが植物の体サイズx_iだけに依存するモデル

（第3章　3.4.1項　参照）

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

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



（D）X+fモデル（k＝3）：

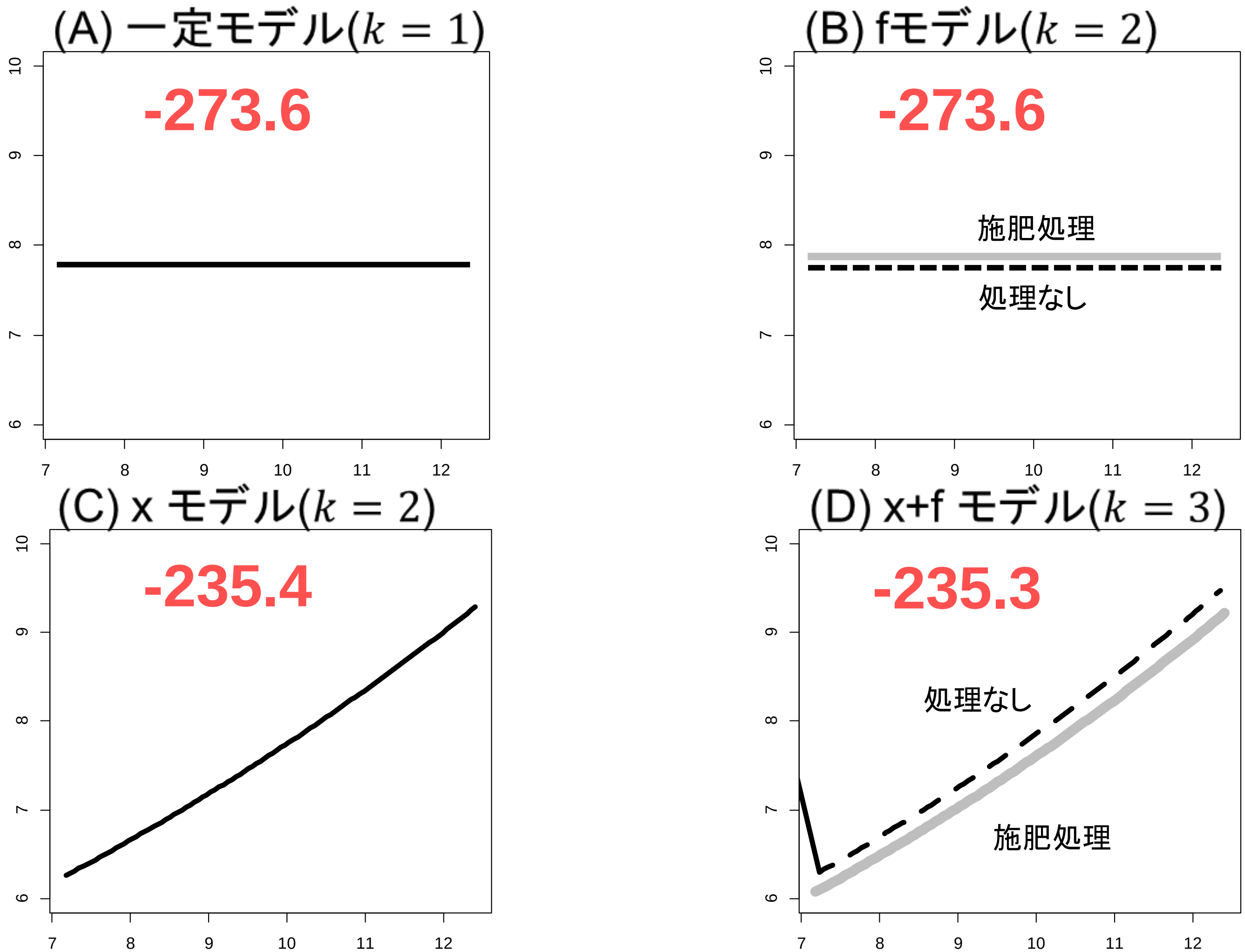
平均種子数λ_iが
施肥処理f_iと植物の体サイズx_iに依存するモデル

（第3章　3.6項　参照）

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

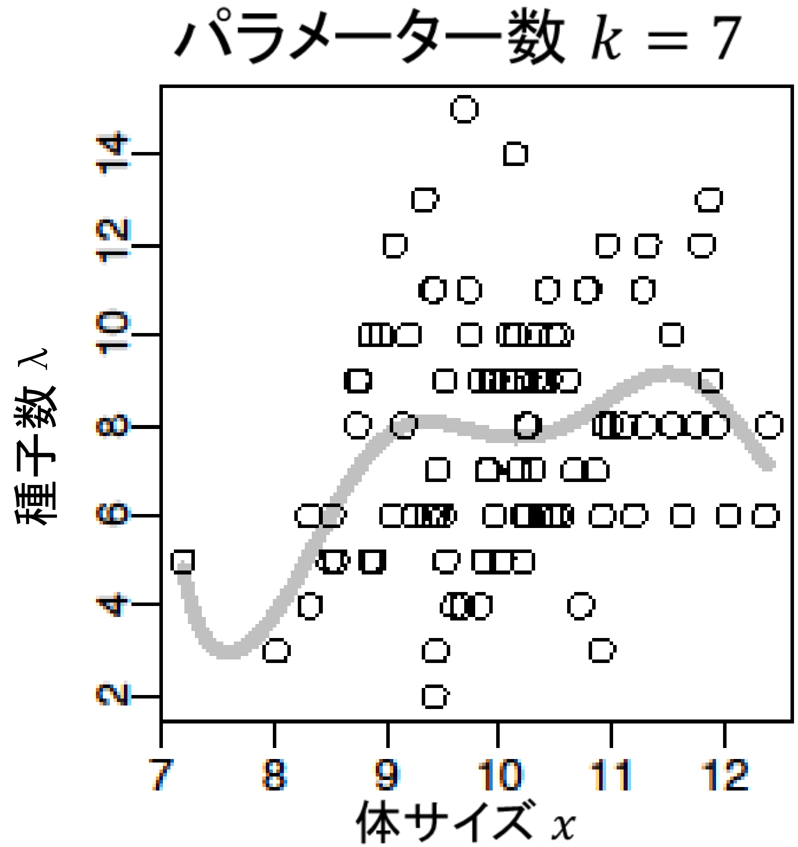
```
> logLik(glm(y ~ x + f, data = d, family = poisson))
'log Lik.' -235.29 (df=3)
```

パラメーター数が多いとあてはまりが良い



（横軸は個体の体サイズx_i，縦軸は平均種子数λ_i）

しかしパラメーター数を増やしすぎると･･･



なんかちがう気がするなあ･･･

一番適しているのはどのモデル？

最大対数尤度（maximum log likelifood）、
つまり「あてはまりの良さ」をみればよい？

ほかの指標をみる必要がある！

逸脱度（deviance）

統計モデルの当てはまりの悪さを表現する指標

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

※今回は対数尤度$\log L(\{\beta_1\})$を$\log L$と省略
最大対数尤度を$\log L^*$と表記

例えばxモデルでは・・・

$$\log L^* = -235.4$$

$$\therefore D = -2\log L^* = 470.8$$

しかしRのglm()では...

ここが関係してそう

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

・・・(中略)・・・

```
Null Deviance:      89.51
Residual Deviance: 84.99        AIC: 474.8
```

どこにも書かれていない・・・

さまざまな逸脱度

名前	定義
逸脱度（D）	$-2\log L^*$
最小の逸脱度	フルモデルをあてはめたときのD
残差逸脱度	D−最小のD
最大の逸脱度	NullモデルをあてはめたときのD
Null逸脱度	最大のD−最小のD

（$\log L^*$は最大対数尤度）

※フルモデル： データすべてをパラメータとして使ったモデル
（ここではデータ数100個を使ったモデル）

ちなみに…

RでのフルモデルのDの求め方

```
> sum(log(dpois(d$y,lambda = d$y)))
[1] -192.8898
```

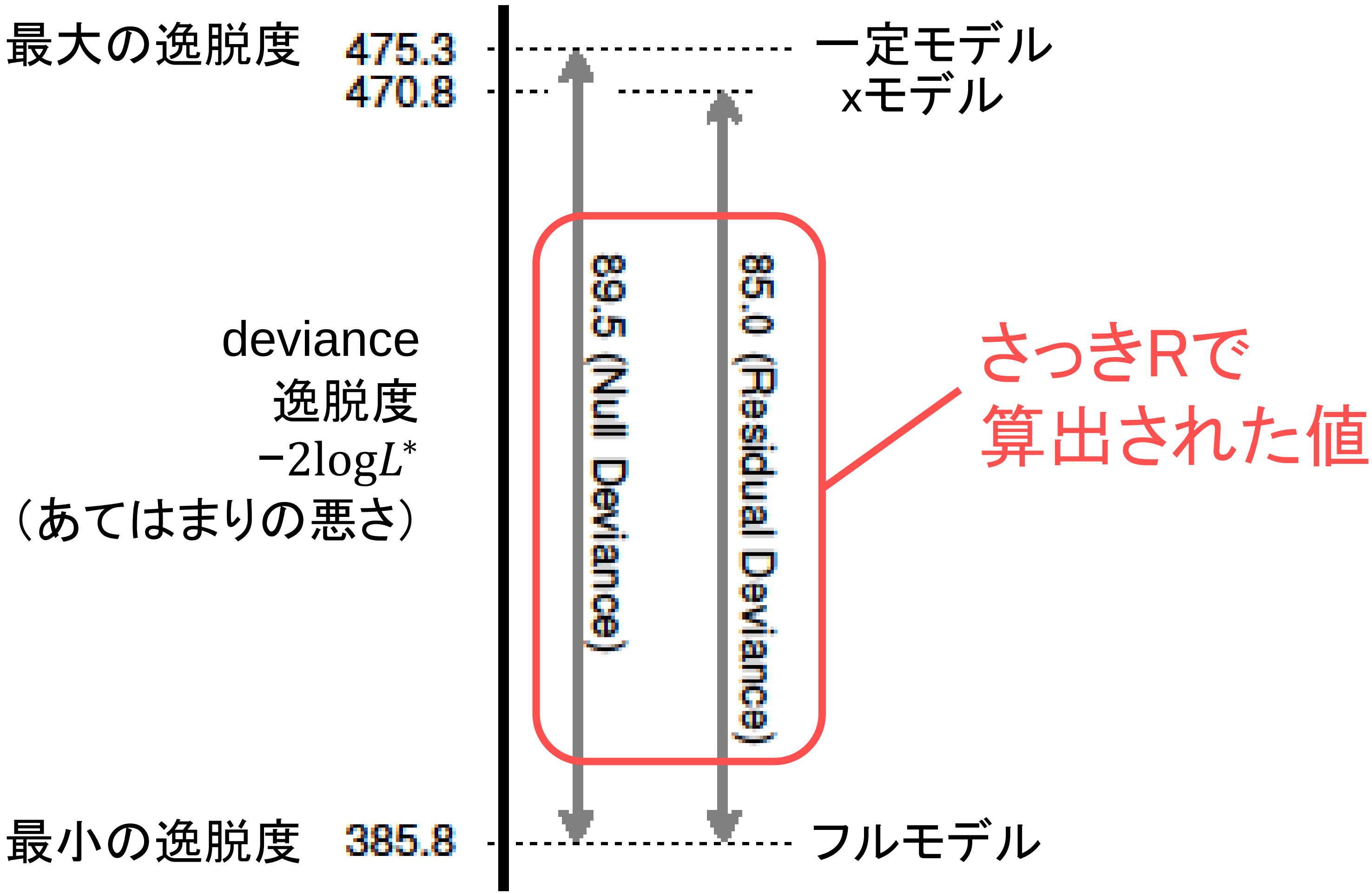
$\log L^* = -192.9$より

$D = -2\log L^* = 385.8$

つまり、このxモデルの残差逸脱度は

D−（最小のD）＝ 470.8 - 385.8 ＝**85.0**

4.2 統計モデルの当てはまりの悪さ：逸脱度



Null Deviance：最大逸脱度

Residual Deviance：残差逸脱度

種子数モデルの最大対数尤度と逸脱度

モデル	k	$\log L^*$	deviance $-2\log L^*$	residual deviance
一定	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	84.5
フル	100	-192.9	385.8	0.0

AIC（Akaike's information criterion）

予測の良さ(goodness of prediction)を重視する指標

AIC

=−2{(最大対数尤度)−(最尤推定したパラメーター数)}

$$= -2(\log L^{*} - k)$$

$$= \boldsymbol{D} + \boldsymbol{2k}$$

AICが一番小さいモデルが良いモデル

種子数モデルの最大対数尤度と逸脱度

モデル	k	$\log L^*$	deviance $-2\log L^*$	residual deviance	AIC
一定	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	84.5	476.6
フル	100	-192.9	385.8	0.0	585.8

xモデルが最も良いモデルと推定された！

一番適しているのはどのモデル？

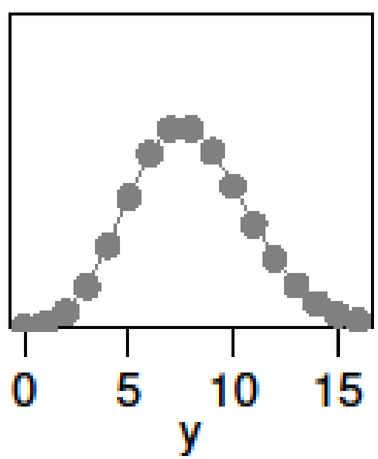
AIC、すなわち「予測の良さ」をみればよい

なぜそういいきれる？

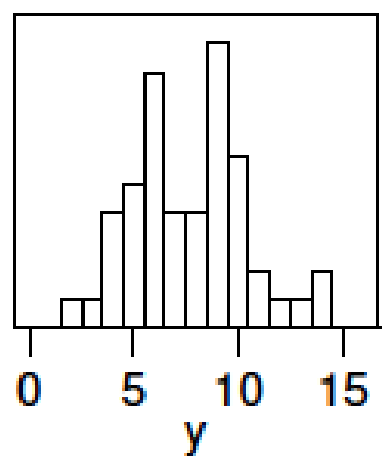
- AICで選ばれたモデルは何が「良いのか」
- AIC$= -2(\log L^{*} - k)$となるのはなぜか

統計モデルのパラメーターの最尤推定の流れ

（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

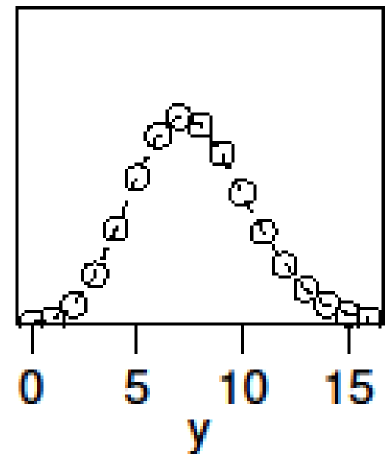


データをサンプリング



推定用の観測データ

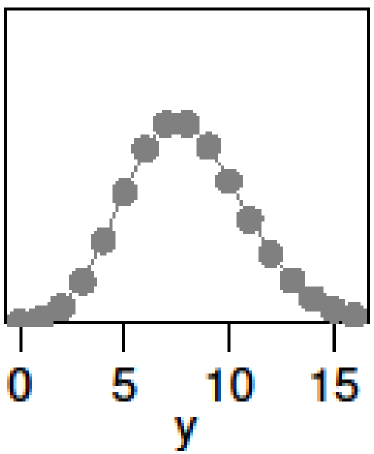
パラメーター推定



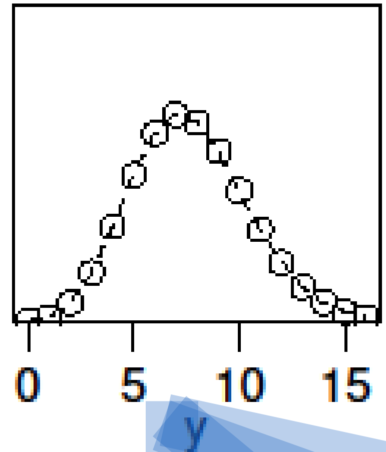
観測データから
推定された一定モデル
$\beta_1 = 2.04$のポアソン分布

4.5 なぜAICでモデル選択してよいのか？

（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

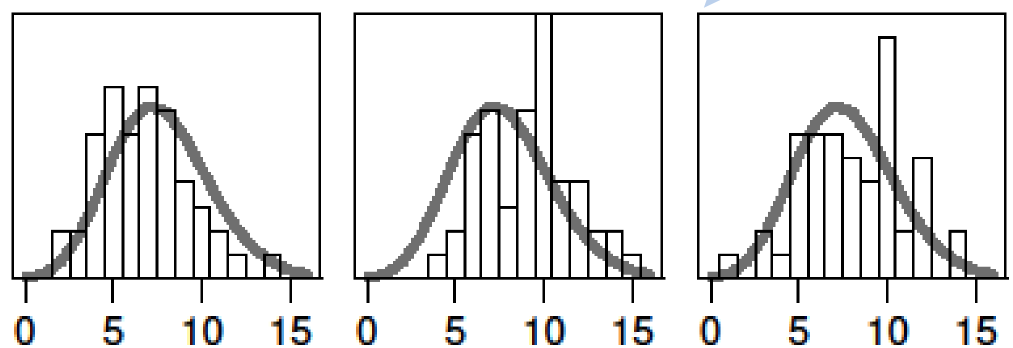


観測データから
推定された一定モデル
$\beta_1 = 2.04$のポアソン分布



データを
サンプリング
（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）

200個の対数尤度の平均 ⇨ 平均対数尤度 $E(\log L)$

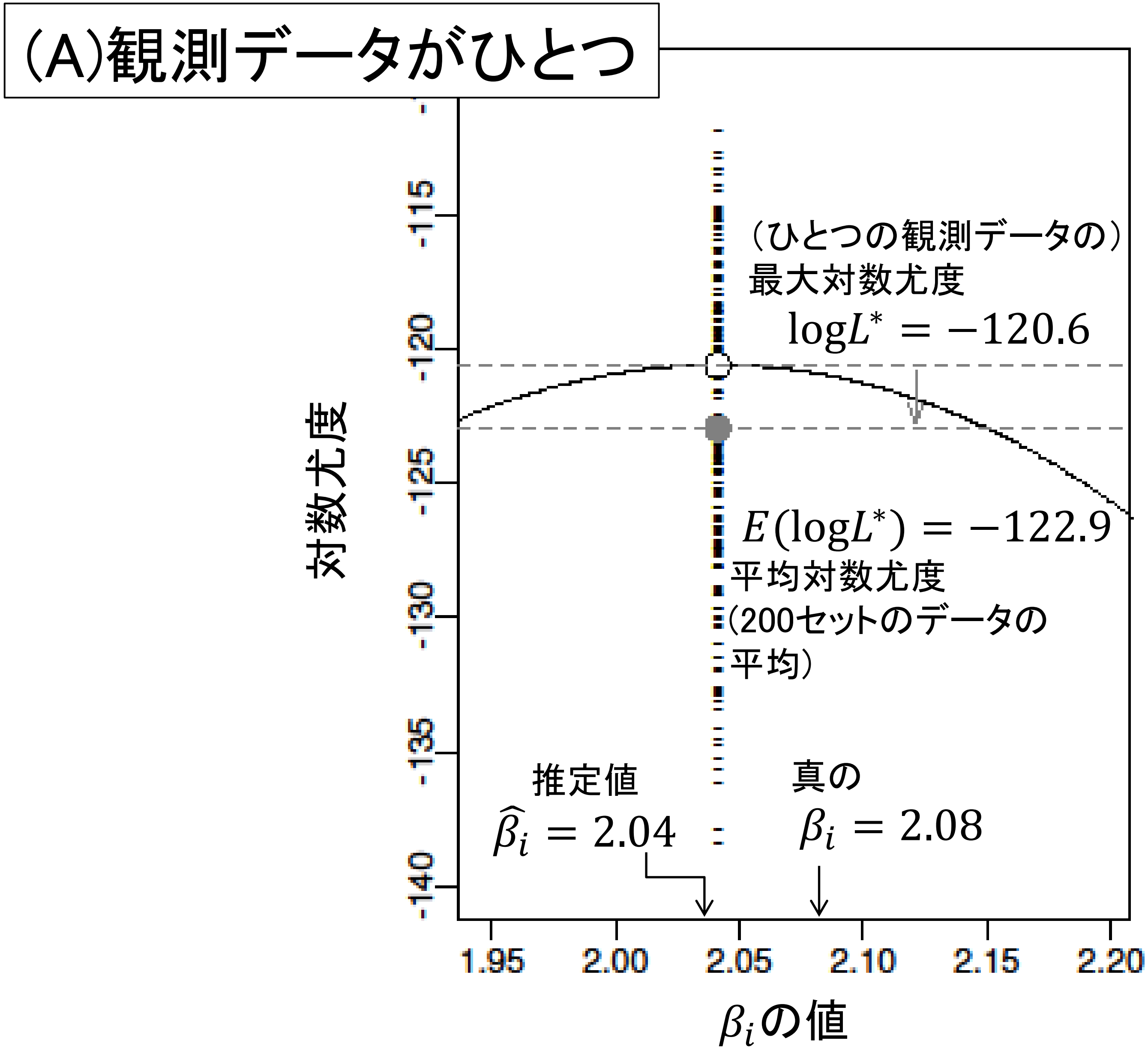
平均対数尤度$E(\log L)$は
統計モデルの良さの指標として適当！

しかしながら現実のデータ解析では

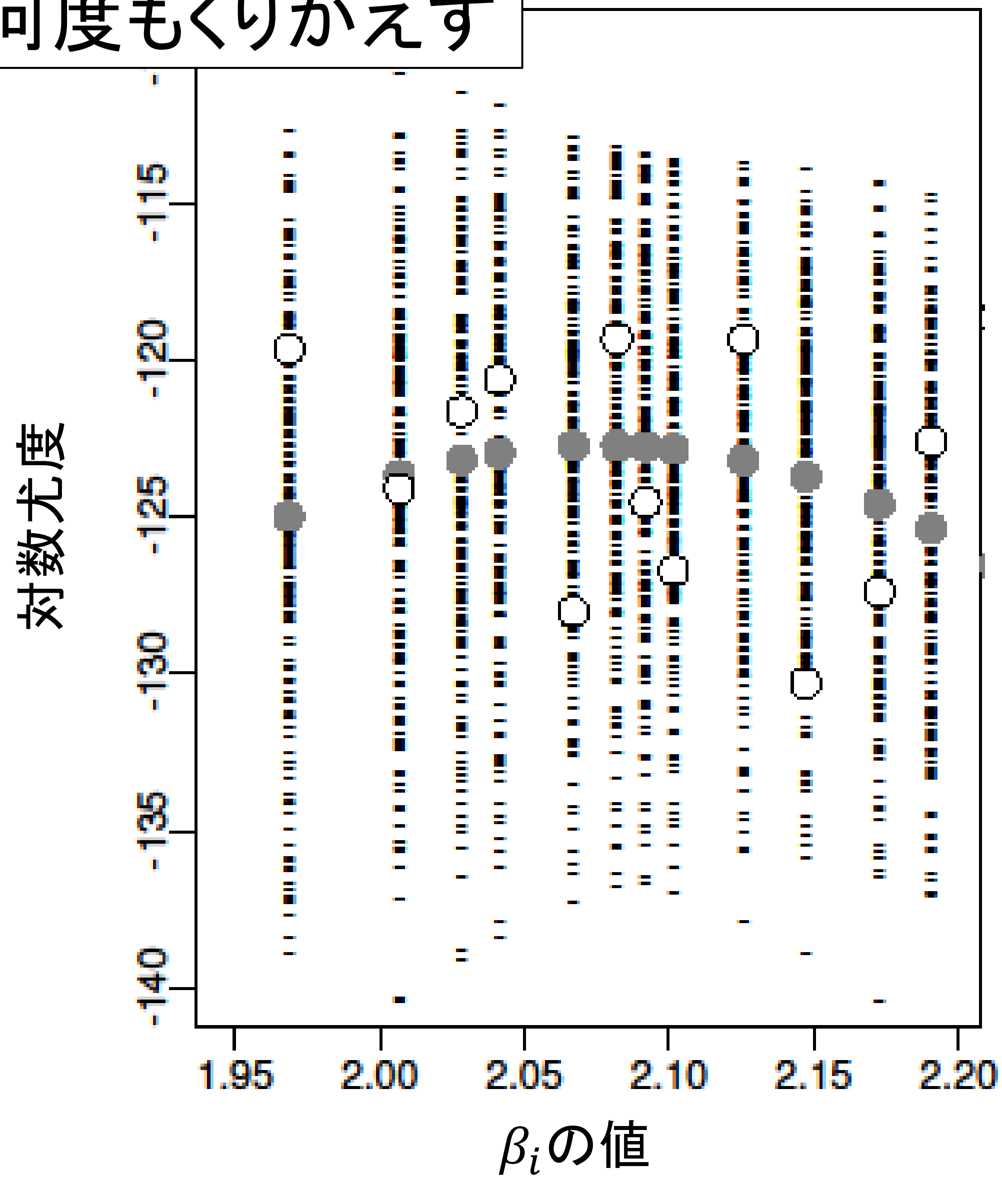
真の統計モデルが不明

⇒ 平均対数尤度もわからない…

なにか工夫が必要


(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
$\log L^* = -120.6$
$E(\log L^*) = -122.9$
平均対数尤度
(200セットのデータの
平均)
推定値
$\widehat{\beta_i} = 2.04$
真の
$\beta_i = 2.08$
対数尤度
-115
-120
-125
-130
-135
-140
1.95
2.00
2.05
2.10
2.15
2.20
β_iの値

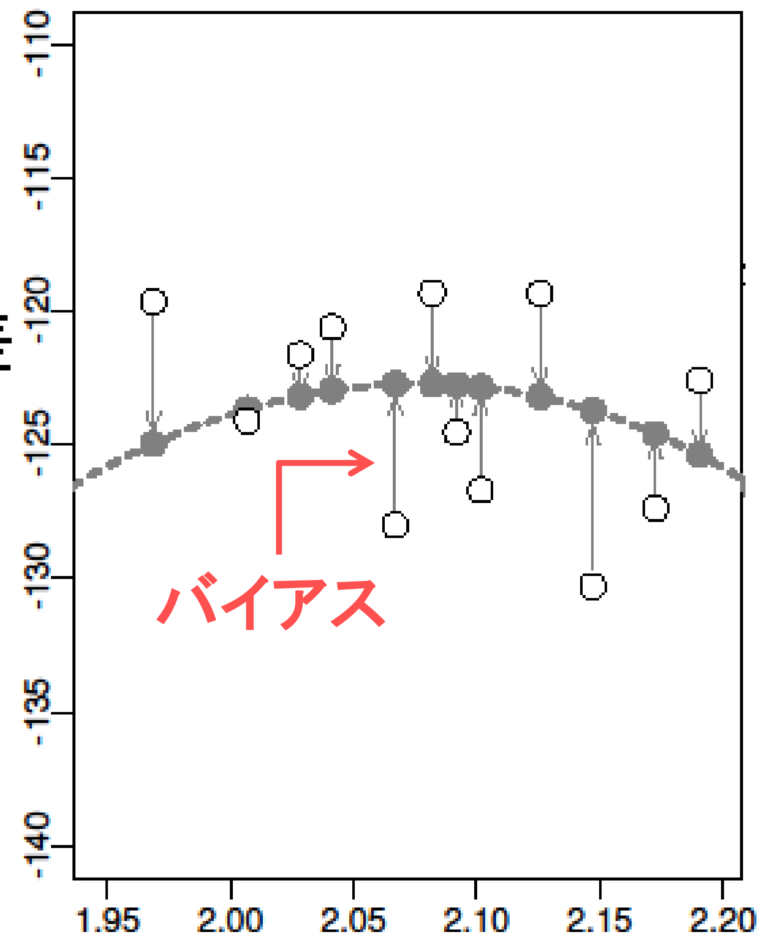
(B)(A)を何度もくりかえす



最大対数尤度と平均対数尤度の差

⇒ バイアス

$$b = \log L^{*} - E(\log L)$$



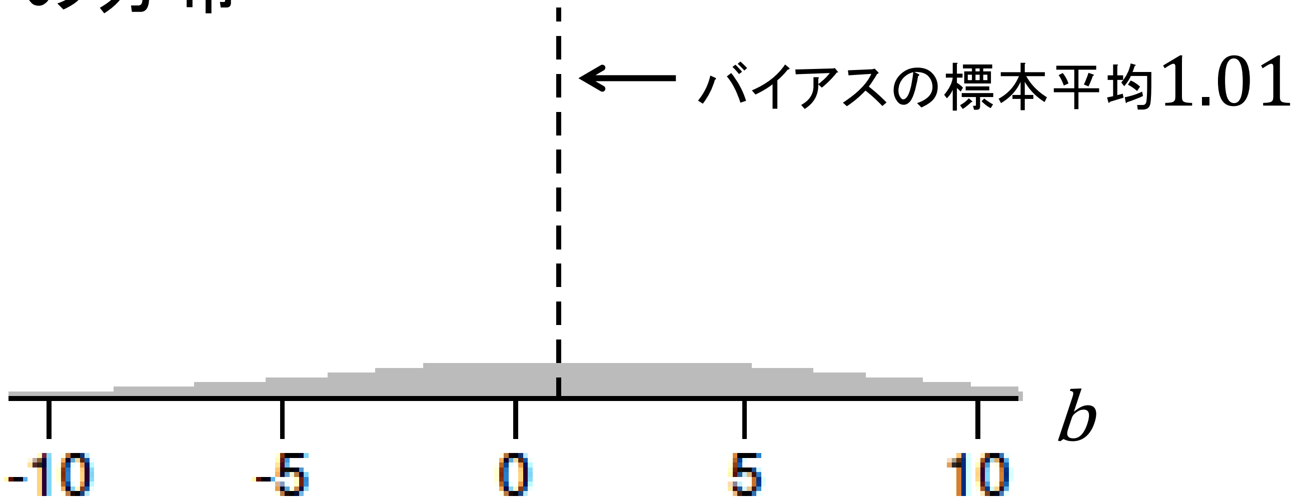
この式を変形すると　$E(\log L) = \log L^* - b$

すなわち…

平均的なbと最大対数尤度$\log L^*$の値がわかれば
平均対数尤度$E(\log L)$の推定値が得られるはず

⇨ バイアス補正

バイアスbの分布



・平均バイアスがおよそ1くらい

$$\therefore E(\log L) = \log L^{*} - 1$$

パラメーター数kのモデルの平均対数尤度は

$E(\log L) = \log L^* - k$ 　といわれている

※一定モデルでは$k = 1$なので

$$E(\log L) = \log L^* - 1$$

これに-2をかけたものがAIC！

$$\mathrm{AIC} = -2 \times (\log L^* - k)$$

4.5 なぜAICでモデル選択してよいのか？

一番適しているのはどのモデル？

AIC、すなわち「予測の良さ」をみればよい

・AICは平均対数尤度の推定値
（バイアス補正によって評価）
・モデルの複雑さを考慮している