

統計モデリング入門 2016 (a)

An Introduction to Statistical Modeling

観測されたパターンを説明する統計モデル

久保拓弥 (北海道大・環境科学)
kubo@ees.hokudai.ac.jp


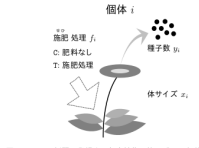





図 3.1 この問題に登場する架空植物の例：種子の個体。この植物の体サイズ(個体の大きさ) x_i と肥料をやる施肥処理 j が種子数 n_i にどう影響しているのかわかりたい。

2016-07-06
統計モデリング入門 2016a
1/60

The main language of this class is Japanese ... Sorry

- Why in Japanese? ... because even in Japanese, **statistics is difficult** for Japanese students to understand.
- I will **compensate for language disadvantages** in foreign students when I give grades.
- **Questions in English are always welcomed!**

2016-07-06
統計モデリング入門 2016a
2/60

統計モデリング授業の web page

http://goo.gl/76c4i

植物生態学特論 I (Advanced Course of Plant Ecology I)

生態学の統計モデリング 7月6日から
Statistical Modeling for Ecology, commence on July 6
13:00 - 14:30, Monday and Wednesday

担当: 久保拓弥 (KUBO Takuya) kubo@ees.hokudai.ac.jp

- 授業メーリングリスト (Course Mailing List)
 - <http://goo.gl/f0vCn8> で早めに登録してください。登録のユーザインターフェイスは日本語に必要できます。授業の資料ダウンロードの連絡などします。
 - subscribe at <http://goo.gl/f0vCn8> immediately, or you can not download course handouts.
- 授業 Web Site: <http://goo.gl/76c4i>

2016-07-06
統計モデリング入門 2016a
3/60

統計モデリング授業 Mailing List

http://goo.gl/f0vCn8

植物生態学特論 I (Advanced Course of Plant Ecology I)

生態学の統計モデリング 7月6日から
Statistical Modeling for Ecology, commence on July 6
13:00 - 14:30, Monday and Wednesday

担当: 久保拓弥 (KUBO Takuya) kubo@ees.hokudai.ac.jp

- 授業メーリングリスト (Course Mailing List)
 - <http://goo.gl/f0vCn8> で早めに登録してください。登録のユーザインターフェイスは日本語に必要できます。授業の資料ダウンロードの連絡などします。
 - subscribe at <http://goo.gl/f0vCn8> immediately, or you can not download course handouts.
- 授業 Web Site: <http://goo.gl/76c4i>

2016-07-06
統計モデリング入門 2016a
4/60

この統計モデリング授業の Mailing List (ML) **kubostat**

- ML を使って各回の「課題」を出します
 - 回答もメールで送信してください
- 成績評価は「課題」の回答
 - 出欠関係なし (欠席の連絡いりません)
- 単位とらない人も ML 登録してください
 - 講義資料のダウンロード案内などあります

2016-07-06
統計モデリング入門 2016a
5/60

Performance Rating

- E-mail assignment (via Mailing List)
 - **That's ALL!**
- Attendance? NOT care.

2016-07-06
統計モデリング入門 2016a
6/60

What for Statistical Modeling?

なぜデータ解析の方法を勉強しなければならないのか?

All you depend on statistics whenever you conclude something based on your data

- データ解析がおかしいと **結論もおかしい**
- Crazy data analysys → Crazy results
- 統計解析わからんと批判的に読めない
- A lack of statistical knowledge → no critical reading of papers

2016-07-06
統計モデリング入門 2016a
8/60


データ解析はあまり重視されてなかった

内容がわからなくてもソフトウェアにまるなげ

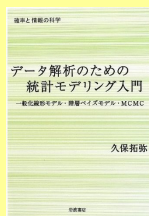
- ブラックボックス統計解析
- No “Blackbox” statistics!
- とにかく「ゆーい差」さえ出せばよいという発想になっている
- Don't blindly believe “Significance” !

この授業のねらい (aim)

できるだけ内容を理解して統計ソフトウェアを使おう!

- Understand how to fit statistical models to your data データにあてはめられる統計モデルを作ろう
- Use the statistical software  to show your data structure

教科書とソフトウェア



この授業は「統計モデリング入門」にそった内容を説明します

my text book (in Japanese)
 著者: 久保拓弥
 出版社: 岩波書店
 2012-05-18 刊行
 価格 3990 円

<http://goo.gl/Ufq2>

割引販売 3000 円!!

「統計モデリング入門」のもとになった「講義の一と」もあります



授業 web page に「講義の一と」へのリンクがあります! <http://goo.gl/82dgC>

統計ソフトウェア R

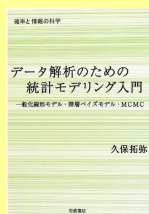
統計学の勉強には良い統計ソフトウェアが必要!

- 無料で入手できる
- 内容が完全に公開されている
- 多くの研究者が使っている
- 作図機能が強力

この教科書でも R を使って問題を解決する方法を説明しています



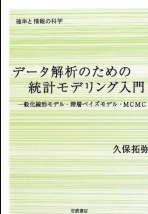
統計モデルとは何か?



「統計モデル」とは何か?

どんな統計解析においても統計モデルが使用されている

- 観察によってデータ化された現象を説明するために作られる
- 確率分布が基本的な部品であり、これはデータにみられるばらつきを表現する手段である
- データとモデルを対応づける手づきが準備されていて、モデルがデータにどれぐらい良くあてはまっているかを定量的に評価できる



「統計モデリング入門」の主張

「何でも正規分布」じゃないだろ!

線形モデルの発展

階層ベイズモデル
もっと自由な統計モデリングを!

一般化線形混合モデル
個体差・場所差といった変量効果をあつかいたい

一般化線形モデル
正規分布以外の確率分布をあつかいたい

線形モデル

推定計算方法
MCMC
最尤推定法
最小二乗法

2016-07-06 統計モデリング入門 2016a 17/60

「統計モデリング入門」の主張

right probability distribution
for right statistical modeling

The Evolution of Linear Models

Hierarchical Bayesian Model (HBM)

Generalized Linear Mixed Model (GLMM)

Generalized Linear Model (GLM)

Linear Model

Parameter Estimation
MCMC
MLE
MSE

2016-07-06 統計モデリング入門 2016a 18/60

たとえばこんなデータがあったしましょう

An example (次の時間の例題)

種子数 y_i

個体 i

体サイズ x_i

plant body size

number of seeds

種子数

体サイズ

plant body size

図 3.1 この初期に登場する東洋植物の属；番目の個体。この植物の体サイズ（個体の大きさ） x_i と肥料をやる施肥処理 f_i が種子数 y_i にどう影響しているのかを知りたい。

2016-07-06 統計モデリング入門 2016a 19/60

一般化線形モデル - ばらつきをよく見る

Don't use the normal distribution without seeing data!

(A) 正規分布・恒等リンク関数の統計モデル
正規分布

(B) ポアソン分布・対数リンク関数の統計モデル
ポアソン分布

階層ベイズモデル
もっと自由な統計モデリングを!

一般化線形混合モデル
個体差・場所差といった変量効果をあつかいたい

一般化線形モデル
正規分布以外の確率分布をあつかいたい

線形モデル

推定計算方法
MCMC
最尤推定法
最小二乗法

0 個、1 個、2 個と数えられる種子数が「正規分布」なわけないだろ!!

3.9 回帰モデルと確率分布の関係。また別の東洋データに対して GLM をあてはめたとき、誤差は μ ともに変化する平均値、グレイで

2016-07-06 統計モデリング入門 2016a 20/60

全体の流れ (1/3)

第 1 回: 7/06 (月) 観測されたパターンを説明する統計モデル
Introduction

第 2 回: 7/08 (水) 確率分布と最尤推定
Probability Distributions and Maximum Likelihood Estimation (MLE)

第 3 回: 7/13 (月) 一般化線形モデル: ポアソン回帰
Generalized Linear Model (GLM): Poisson Regression

全体の流れ (2/3)

第 4 回: 7/15 (水) モデル選択と検定
Model Selection and Statistical Test

第 5 回: 7/22 (水) 一般化線形モデル: ロジスティック回帰
GLM: Logistic Regression

第 6 回: 7/27 (月) 一般化線形混合モデル
Generalized Linear Mixed Model (GLMM)

第 7 回: 7/29 (水) 階層ベイズモデル
Bayesian GLMM and Markov Chain Monte Carlo

全体の流れ (3/3)

第 7 回: 8/1 (月) 時間変化データの統計モデル (1)
Time change data analysis: common mistakes

第 8 回: 8/3 (水) 時間変化データの統計モデル (2)
Time change data analysis using HBM

第 9 回: 8/9 (火) 時間変化データの統計モデル (3)
State space model: an application of HBM

(tentative)

7/11 (月)

統計モデリング入門 2016 (b)
probability distribution and maximum likelihood estimation
確率分布と最尤推定

久保拓弥 kubo@ees.hokudai.ac.jp

北大環境科学の講義 <http://goo.gl/76c41>

2016-07-11

ファイル更新時刻: 2016-06-30 16:47

kubostat2016a (<http://goo.gl/76c41>) 統計モデリング入門 2016 (b) 2016-07-11 1 / 52

単純化した例題

こんなデータ (葉空) があってほしいよ

まあ、なんだかこういへんな植物を測定してるとします

個体 i 種子数 y_i
この y_i が観測データ!
全 50 個体 $i = \{1, 2, 3, \dots, 50\}$ $y_i = \{0, 0, \dots, 10\}$

このデータ $\{y_i\}$ がすでに R という統計ソフトウェアに格納されていた、としましょう

```
> data
[1] 2 2 4 6 4 5 2 3 1 2 0 4 3 3 3 3 4 2 7 2 4 3 3 3 4
[26] 3 7 5 3 1 7 6 4 6 5 2 4 7 2 2 6 2 4 5 4 5 1 3 2 3
```

Histogram of data

データ解析における重要な事項とくに「調」を楽く!

カウントデータはポアソン分布を使って説明できないかを調べる

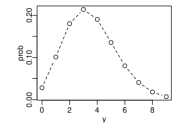


図 4 平均 $\lambda = 3.56$ のポアソン分布。種子数 y とその確率 $prob$ の関係が示されている。図 4 の表を詳しく見ても、表の $prob$ 欄の数値: $type = "n"$ によって「 n と対称線による関係」: 10^y をもとめて「対称線(対称度)」と書かれている。

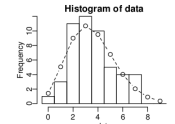


図 5 観測データと確率分布の対比をながめる。ヒストグラム(図 4)と同じ、それに重ねられていると対称線 y 軸の対称性を示す確率分布の線。平均 $\lambda = 3.56$ の観測データのポアソン分布の確率分布に全無関係(調) あるいは存在する。

さいゆう 最尤推定という考えかたを説明します

対数尤度を最大化する λ をさがす

対数尤度 $\log L(\lambda) = \sum (y_i \log \lambda - \lambda - \sum \log y_i)$

図 7 平均 λ (3.56) を変化させていたポアソン分布と、観測データへのあてはまりの良さ (対数尤度 $\log L$)。すべてのヒストグラムは図 4 と同じ。

7/13 (水)

統計モデリング入門 2016 (c)

Poisson regression, a generalized linear model (GLM)
一般化線形モデル: ポアソン回帰

久保拓弥 kubo@ees.hokudai.ac.jp

北大環境科学院の講義 <http://goo.gl/776c41>

2016-07-13

ファイル更新時刻: 2016-06-30 16:48

kubostat2016c (<http://goo.gl/776c41>) 統計モデリング入門 2016 (c) 2016-07-13 1 / 46

ここで登場する --- 「何でも正規分布」ではダメ! という発想

図 3.1 この例題に登場する某空植物の葉: 番目の個体、この植物の体サイズ(個体の大きさ) x_i と肥料をやる施肥処理 f_i が種子数 y_i にどう影響しているのかを知りたい。

(A) 正規分布・相関リンク関数の統計モデル
正規分布

(B) ポアソン分布・対数リンク関数の統計モデル
ポアソン分布

図 3.9 回帰モデルと確率分布の関係。また別の観測データに對して GLM をあてはめた例。破線は x とともに変化する平均値。グレイで

Free の統計ソフトウェア R で統計モデリング

結果を格納するオブジェクト

```
fit <- glm(y ~ x, family = poisson(link = "log"), data = data)
```

図 3.1 この例題に登場する某空植物の葉: 番目の個体サイズ(個体の大きさ) x_i と肥料をやる施肥処理 f_i にどう影響しているのかを知りたい。

図 17 平均種子数 λ の予測。図 17 に λ の予測値(実験)を上記したもの。

7/20 (水)

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/776c41>

2016-07-20

ファイル更新時刻: 2016-06-30 16:56

kubostat2016d (<http://goo.gl/776c41>) 統計モデリング入門 2016 (d) 2016-07-20 1 / 40

Q. モデル選択とは何か?

パラメーター数は多くても少なくてもへん?

(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 ?

2016-07-06 統計モデリング入門 2016a 32/60

A. より良い予測をする統計モデルを探すこと

統計学の伝統として、その詳細性
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 ← こっちだ!

検定はモデル選択じゃない! 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

2016-07-06 統計モデリング入門 2016a 33/60

統計学って「検定」のこと?

「検定」って何なの?

「検定」ってエライの?

帰無仮説が真の統計モデルということにしてしまう ($\beta_1 = 2.06$ のポアンソン分布)
↓
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
↓
評価用データに一定モデルと α モデルをあてはめて過剰度差 $\Delta D_{1,2}$ の分布を予測
↓
あてはまりの良い良き評価用データ (多数)

図 6 尤度比検定に必要な $\Delta D_{1,2}$ の分布の生成。まず帰無仮説である一定モデル ($\beta_1 = 2.06$; 1. 帰無) が真の統計モデルだと仮定し、そこから得られるデータを使って過剰度差 $\Delta D_{1,2}$ がどのような分布になるかを調べる。

2016-07-06 統計モデリング入門 2016a 34/60

7/25 (月)

統計モデリング入門 2015 (e)

GLM logistic regression
一般化線形モデル: ロジスティック回帰

久保拓弥 kubo@ees.hokudai.ac.jp
北大環境科学院の講義 http://goo.gl/76c44i
2015-07-22
ファイル更新時刻: 2015-07-02 16:24

2016-07-06 統計モデリング入門 2015 (e) 2015-07-22 1 / 42

生物学のデータ解析は「割算」しまくり!!

この悪しき「割算」な統計学

世間でよくみかけるおススメできない作法の例

1. データどんどん割算・割算
2. 何でもいからひたすらセンをひく
3. y/x がでたら万歳・万歳
4. うまくいくまで 1, 2, 3 ぐるぐる

何でも割算!

2012-11-02 k4 (2012-10-26 17:07 修正版) 14/44

2016-07-06 統計モデリング入門 2016a 36/60

GLM のひとつ, ロジスティック回帰を使おう

データにあわせたより良い統計モデリングを!

おススメできないデータ解析を回避するための注意点

- むやみに区画わけしない!
- 何でも割算するな!
- たくさん図を描く
- 「観測データを説明する確率分布は何か?」を考える

NO!

コツ: 不自然にデータをこねくりまわさない
データの性質・構造にあったモデリングを!

2012-11-02 k4 (2012-10-26 17:07 修正版) 43/44

2016-07-06 統計モデリング入門 2016a 37/60

GLM のひとつ, ロジスティック回帰を使おう

またいつもの例題? ちょっとちがう

ロジスティック回帰とは何なの?

8 個の種子のうち y 個が発芽可能だった! というデータ

(A) 観測データの一部 ($y = 0$) (B) 推定されるモデル

二項分布: N 回のうち y 回, となる確率

2016-07-06 統計モデリング入門 2016a 38/60

7/27 (水)

統計モデリング入門 2016 (f)

階層ベイズモデル Hierarchical Bayesian Model

久保拓弥 kubo@ees.hokudai.ac.jp
北大環境科学院の講義 http://goo.gl/76c44i
2016-07-27
ファイル更新時刻: 2016-06-30 17:01

2016-07-27 統計モデリング入門 2016 (f) 2016-07-27 1 / 67

GLM ではうまく説明できないデータ!?

GLMM は階層ベイズモデルの一種 事前分布をどう選ぶかが重要

また別の観測データ: 二項分布だめだめ?!

100 個体の植物の合計 800 種子中 403 個の生存が見られたので、平均生存確率は 0.50 と推定されたが.....

観測された植物の個体数 y_i 生存した種子数 y_{ij}

二項分布による予測
ぜんぜんうまく表現できてない!

さっきの例題と同じようなデータなの!?
(「統計モデリング入門」第 10 章の最初の例題)

第 6 回と同じような例題を、こんどはベイズモデルを使ってモデリングします

2016-07-06 統計モデリング入門 2016a 39/60

GLM を階層ベイズモデル化して対処

なぜ「階層」ベイズモデルと呼ばれるのか？

超事前分布 → 事前分布という階層があるから

データ: 8個中のY[i]個の種子が生存

二項分布: 生存確率 $q[i]$

植物の個体差: $r[i]$

事前分布: 個体差のばらつき σ

無情報事前分布 (超事前分布)

sigma は σ とってください

無情報事前分布 (超事前分布)

失印は手順ではなく、依存関係をあらわしている

2016-07-06 統計モデリング入門 2016a 41/60

なぜ階層ベイズモデルまで勉強するのか？

生態学!

- 個体差・エリア差・空間相関・時間相関・種差などめんどろなことをあつかわないといけない

The Evolution of Linear Models

- Linear Model
- Generalized Linear Model (GLM)
- Generalized Linear Mixed Model (GLMM)
- Hierarchical Bayesian Model (HBM)
- Parameter Estimation MCMC

そういう難しい状況では……

- ベイズモデル化
- そのパラメータの事後分布を MCMC 法を使って推定するのが無難

2016-07-06 統計モデリング入門 2016a 42/60

第 7, 8, 9 回は「時間変化」するデータの統計モデリング (階層ベイズモデルの応用)

時間変化のデータ解析

時系列データ解析

短い時系列データ

時系列の長短に関係なく「対応のある」データ点かどうか本質的な問題

再測定もまた時系列データ

岩波データサイエンス vol.1

2016-07-06 46/60

対応 (paired) を考えてない GLM あてはめ

これはまちがいが! 「ゆーい差」あり, となる

glm(身長 ~ (測定2回目) + (測定2回目):(処理の効果))

同じ対象を二回測定していることを考慮してない

2016-07-06 47/60

対応 (paired) を考えてない GLM あてはめ

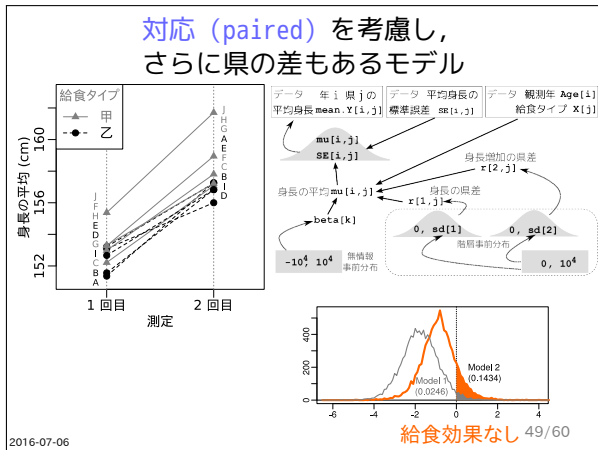
「ゆーい」になりやすい

これはまちがいが! 「ゆーい差」あり, となる

glm(身長 ~ (測定2回目) + (測定2回目):(処理の効果))

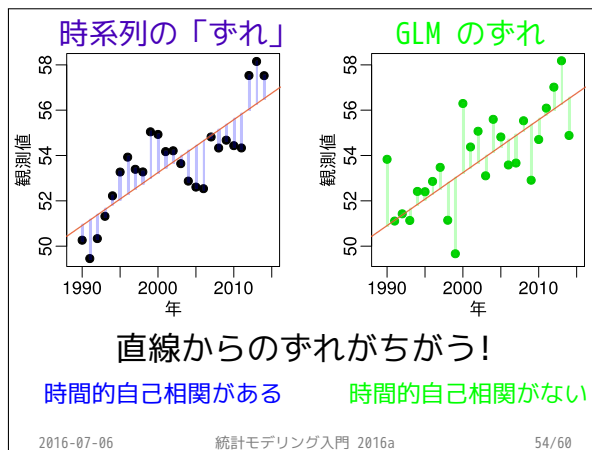
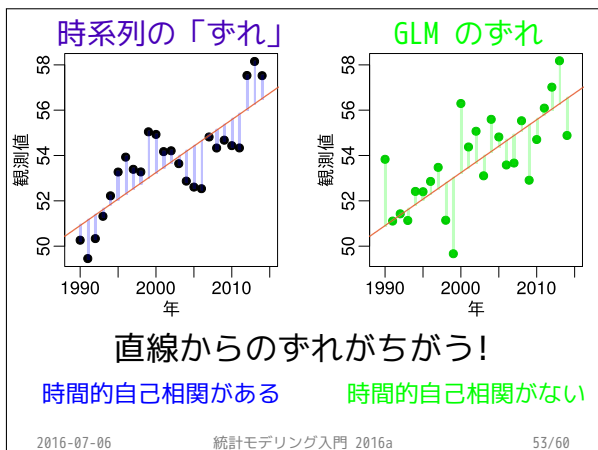
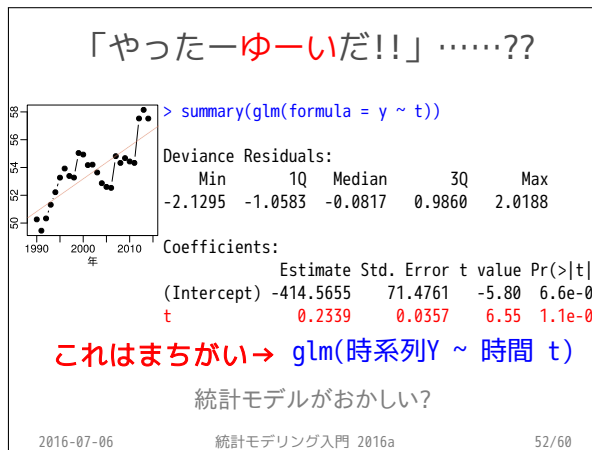
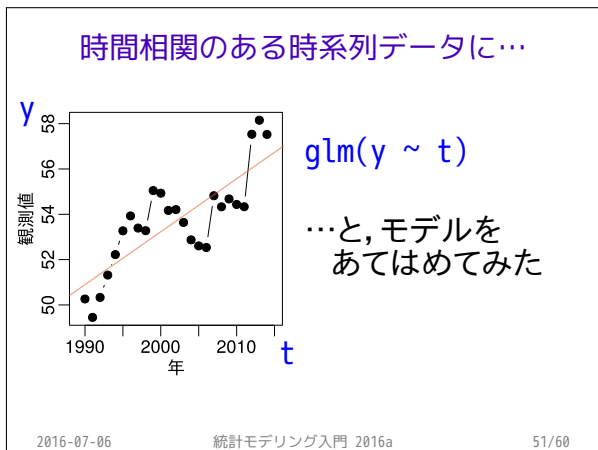
同じ対象を二回測定していることを考慮してない

2016-07-06 48/60



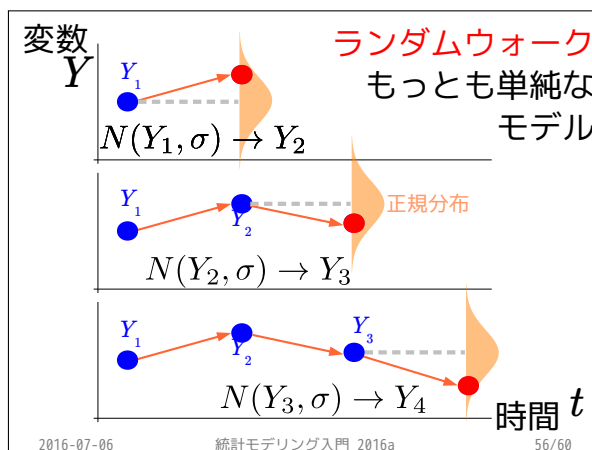
長い時系列データ

データ上で「時間相関」が見える
「時間相関」のモデリングが必要



統計モデルづくりの要点

時系列データの解析は
階層ベイズモデル化した
状態空間モデルを使うのが便利



状態空間モデル + 観測モデル

この部分にポアソン分布や二項分布をいれる 誤差 **状態空間モデル**

$N(y_t, \sigma_2) \rightarrow Y_t$ 二種類の σ をもつ

観測データ Y_1, Y_2, Y_3

状態変数の変化 時間 t

観測できない世界 (状態空間)

2016-07-06 統計モデリング入門 2016a 57/60

状態空間モデル + 観測モデル

他にも季節変動などを入れることができます

観測データ Y_1, Y_2, Y_3

状態変数の変化 時間 t

観測できない世界 (状態空間)

2016-07-06 統計モデリング入門 2016a 58/60

説明してみたいこと

- 時系列データ: 単純な回帰はダメ(続)
- 状態空間モデル: 乱歩と雑音の分離
- 差分と時間的自己相関係数
- 欠測と不等間隔
- 時系列と「対応のある」データ
- 説明しないこと - 因果推定など

2016-07-06 59/60

今日はここまで