統計モデリング入門 2017 (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

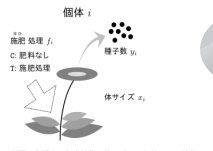


図 3.1 この例題に登場する架空植物の第 i 番目の個体．この植物の体サイズ（個体の大きさ）x_i と肥料をやる施肥処理 f_i が種子数 y_i にどう影響しているのかを知りたい．

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!

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 6 月 5 日から

Statistical Modeling for Ecology, commence on June 5

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 に登録してください．授業の資料ダウンロードの連絡などします．単位を取得する院生は必須．登録のユーザーインターフェイスは日本語に変更できます．
  - subscribe at http://goo.gl/f0vCn8 immediately, or you can not download course handouts.
- 授業 Web Site: http://goo.gl/76c4i

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 6 月 5 日から

Statistical Modeling for Ecology, commence on June 5

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 に登録してください．授業の資料ダウンロードの連絡などします．単位を取得する院生は必須．登録のユーザーインターフェイスは日本語に変更できます．
  - subscribe at http://goo.gl/f0vCn8 immediately, or you can not download course handouts.
- 授業 Web Site: http://goo.gl/76c4i

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
  - 回答もメイルで送信してください
  - Send your assignment via the ML
- 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

Performance Rating

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

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 analsys → 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

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 R to show your data structure

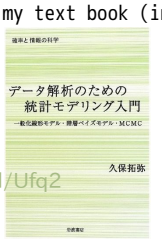
教科書とソフトウェア

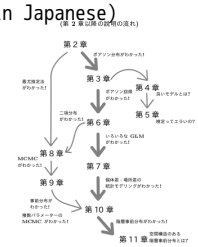
R

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!!

Statistical software for this course

統計ソフトウェア R

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?

What? statistical modeling?

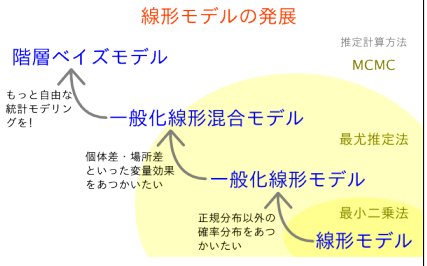
「統計モデル」とは何か?

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つ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!



「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) — Parameter Estimation: MCMC
Generalized Linear Mixed Model (GLMM) — MLE
Generalized Linear Model (GLM) — MSE
Linear Model

たとえばこんなデータがあったしましょう

An example

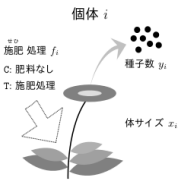
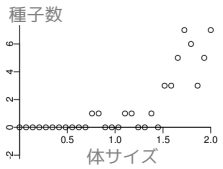


図 3.1 この例題に登場する架空植物の第 i 番目の個体．この植物の体サイズ（個体の大きさ）x_i と肥料をやる施肥処理 f_i が種子数 y_i にどう影響しているのかを知りたい．

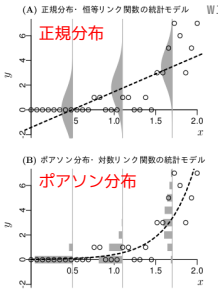
number of seeds



plant body size

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!



線形モデルの発展

階層ベイズモデル — 推定計算方法: MCMC
もっと自由な統計モデリングを！
一般化線形混合モデル — 最尤推定法
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 — 最小二乗法
正規分布以外の確率分布をあつかいたい
線形モデル

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!!

図 3.9 回帰モデルと確率分布の関係．また別の架空データに対して 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

全体の流れ（1/3）

第 1 回: 6/05（月）観測されたパターンを説明する統計モデル
Introduction
6/07（水）（授業なし no class）← 奇怪な学年暦
第 2 回: 6/12（月）確率分布と最尤推定
Probability Distributions and Maximum Likelihood Estimation (MLE)
第 3 回: 6/14（水）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）

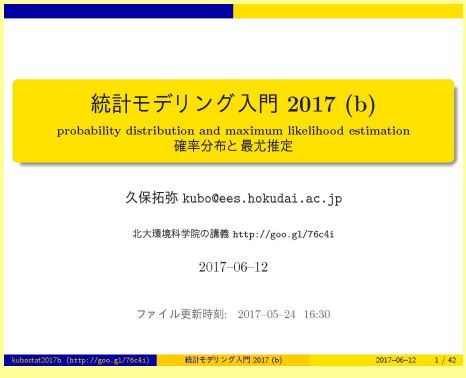
第 4 回: 6/19（月）モデル選択と検定
Model Selection and Statistical Test
第 5 回: 6/21（水）一般化線形モデル: ロジスティック回帰
GLM: Logistic Regression
第 6 回: 6/26（月）一般化線形混合モデル
Generalized Linear Mixed Model (GLMM)

全体の流れ（3/3）

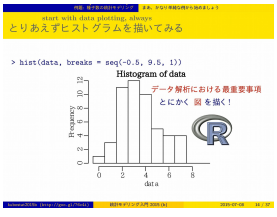
第 7 回: 6/28（水）階層ベイズモデル
Bayesian GLMM and Markov Chain Monte Carlo
第 8 回: 7/03（月）時間変化データの統計モデル
Time change data analysis: common mistakes

(tentative)

6/12（月）



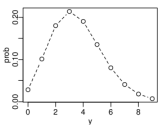
単純化した例題

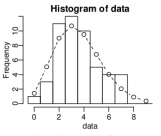


Simplified examples to learn statistical modeling

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

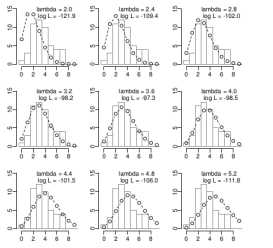
Find some appropriate probability distributions to fit the observed distributions

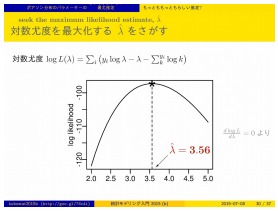




2017-06-05 統計モデ!

さいゆう 最尤推定という考えかたを説明します





How to fit the distribution to the observation?

Maximum likelihood estimation!

2017-06-05 統計モデリング入門 2017a 26/59

6/14（水）

統計モデリング入門 2017 (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

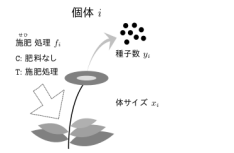
北大環境科学院の講義 http://goo.gl/76c4i

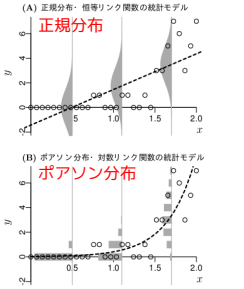
2017–06–14

ファイル更新時刻: 2017–05–24 16:30

kubostat2017c (http://goo.gl/76c4i) 統計モデリング入門 2017 (c) 2017–06–14 1 / 47

ここで登場する --- 「何でも正規分布」ではダメ！という発想

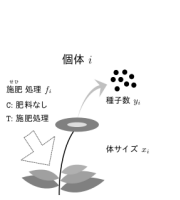




the “normal distirbution is NOT “normal”

2017-06-05 統計モデリング入門 2017a 28/59

Free の統計ソフトウェア R で統計モデリング



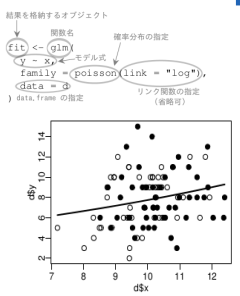


図 17 平均種子数 λ の予測．図 16 に λ の予測値（実線）を上がきしたもの．

2017-06-05 統計モデリング入門 2017a 29/59

6/19（月）

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

2017–06–19

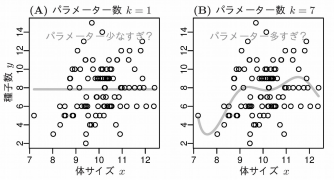
ファイル更新時刻: 2017–05–24 16:32

kubostat2017d (http://goo.gl/76c4i) 統計モデリング入門 2017 (d) 2017–06–19 1 / 44

statistical model selection

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ヘン？



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

2017-06-05 統計モデリング入門 2017a 31/59

A. より良い予測をする統計モデルを探すこと

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ーの 最尤(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

2017-06-05 統計モデリング入門 2017a 32/59

統計学って「検定」のこと?

「検定」って何なの?

「検定」ってエラいの?

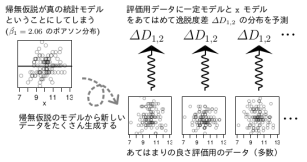


図 6　尤度比検定に必要な $\Delta D_{1,2}$ の分布の生成．まず帰無仮説である一定モデル（$\hat{p}_i = 2.00$, p.[2] 参照）が既の統計モデルだと仮定し，そこから得られるデータを使って逸脱度差 $\Delta D_{1,2}$ がどのような分布になるかを調べる．

6/21（水）

統計モデリング入門 2017 (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/76c4i

2017-06-21

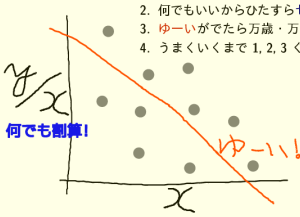
ファイル更新時刻: 2017-05-24 16:32

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めできない作法の例

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

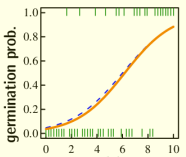


ちなみにこれは $\frac{y}{x}$ と $\frac{\circ}{\infty}$ を比較してるんだから，反応併みたいな偽「集の相関」ができるのはあたりまえ

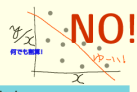
GLM のひとつ，ロジスティック回帰を使おう

データにあわせたより良い統計モデリングを!

おススめ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 割り算するな！
- 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

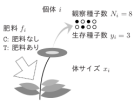


コツ: 不自然にデータをこねくりまわさない
データの性質・構造にあったモデリングを!

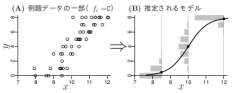
GLM のひとつ，ロジスティック回帰を使おう

またいつもの例題? …… ちょっとちがう

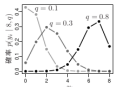
8 個の種子のうち y 個が 発芽可能 だった! …… というデータ



ロジスティック回帰とは何なのか?



二項分布: N 回のうち y 回，となる確率



6/26（月）

統計モデリング入門 2017 (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/76c4i

2017-06-26

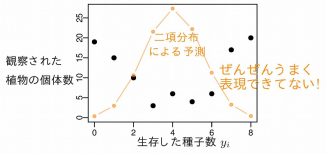
ファイル更新時刻: 2017-05-24 16:32

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……



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

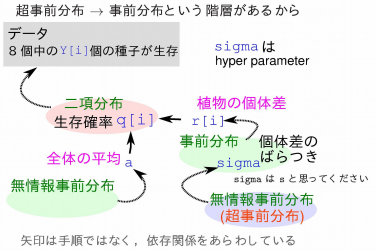
第 6 回と同じような例題を，こんどはベイズモデルを使ってモデリングします

GLM を階層ベイズモデル化して対処

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

なぜ「階層」ベイズモデルと呼ばれるのか?

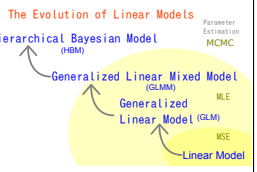
超事前分布 → 事前分布という階層があるから



矢印は手順ではなく，依存関係をあらわしている

なぜ階層ベイズモデルまで勉強するの?

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



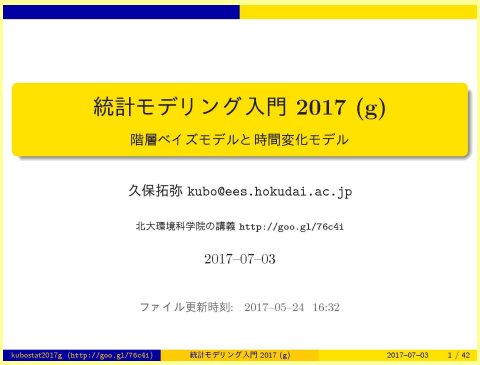
そういう難しい状況では……
- ベイズモデル化
- そのパラメーターの事後分布を MCMC 法を使って推定するのが無難

2017-06-05 統計モデリング入門 2017a 41/59

第 7, 8 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)

6/ (水)


統計モデリング入門 2017 (g)

階層ベイズモデルと時間変化モデル

久保拓弥 kubo@ees.hokudai.ac.jp

北大環境科学院の講義 http://goo.gl/76c4i

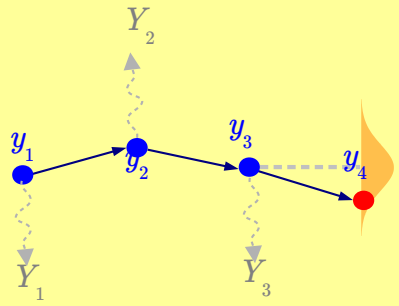
2017–07–03

ファイル更新時刻: 2017–05–24 16:32

kubostat2017g (http://goo.gl/76c4i) 統計モデリング入門 2017 (g) 2017–07–03 1 / 42

時間変化のデータ解析

時系列データ解析

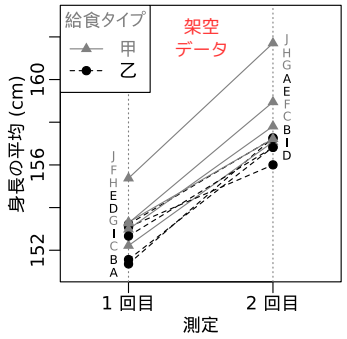


短い時系列データ

時系列の長短に関係なく「対応のある」データ点かどうかが本質的な問題

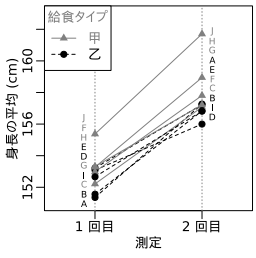
再測定もまた時系列データ

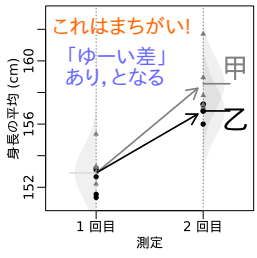
岩波データサイエンス vol.1



2017-06-05 46/59

対応 (paired) を考えてない GLM あてはめ



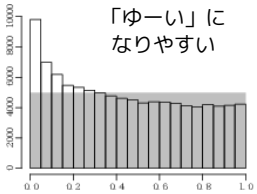


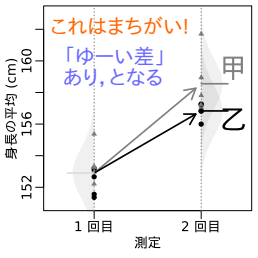
glm(身長 ~ (測定2回目) + (測定2回目):(処理の効果))

同じ対象を二回測定していることを考慮してない

2017-06-05 47/59

対応 (paired) を考えてない GLM あてはめ



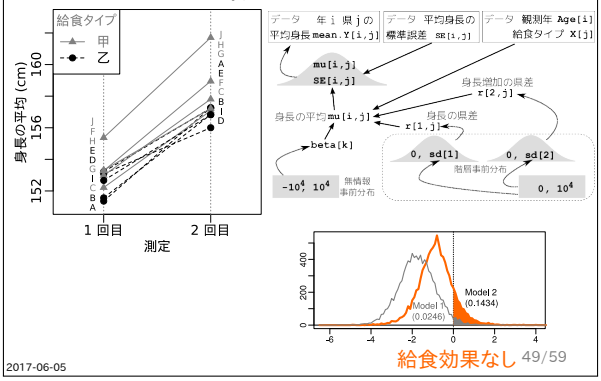


glm(身長 ~ (測定2回目) + (測定2回目):(処理の効果))

同じ対象を二回測定していることを考慮してない

2017-06-05 48/59

対応（paired）を考慮し，
さらに県の差もあるモデル

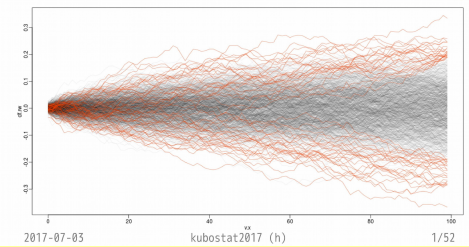


給食効果なし

7/27（水）

生態学の時系列データ解析でよく見る
『あぶない』モデリング

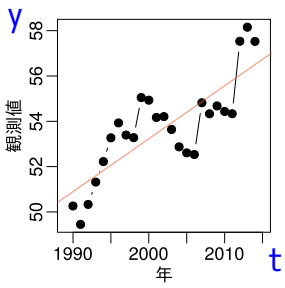
久保拓弥 mailto:kubo@ees.hokudai.ac.jp



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

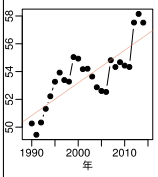
時間相関のある時系列データに…



glm(y ~ t)

…と，モデルを
あてはめてみた

「やったーゆーいだ!!」……??

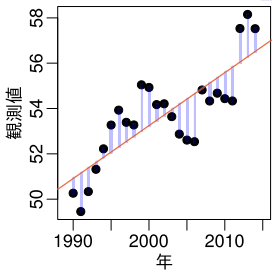


```
> summary(glm(formula = y ~ t))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1295  -1.0583  -0.0817   0.9860   2.0188

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -414.5655    71.4761   -5.80  6.6e-06
t              0.2339     0.0357    6.55  1.1e-06
```

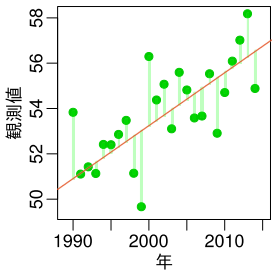
これはまちがい→ glm(時系列Y ~ 時間 t)

統計モデルがおかしい?

時系列の「ずれ」



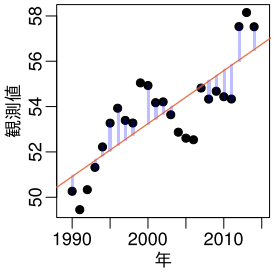
GLM のずれ



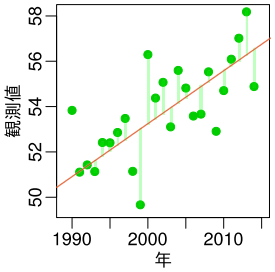
直線からのずれがちがう!

時間的自己相関がある　　時間的自己相関がない

時系列の「ずれ」



GLM のずれ



直線からのずれがちがう!

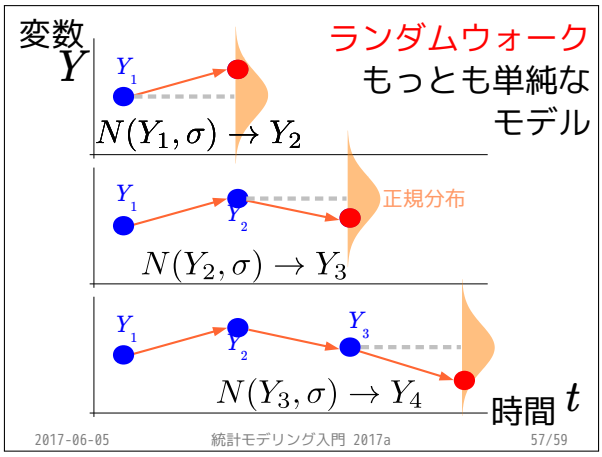
時間的自己相関がある　　時間的自己相関がない

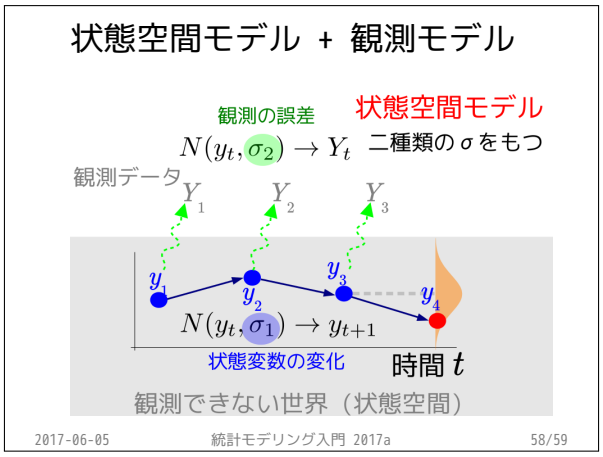
統計モデルづくりの要点

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





今日はここまで

any questions?