統計モデリング入門 2017（a）

An overview for this Statistical Modeling class

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

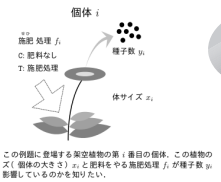


図 3.1 この例題に登場する架空植物の第 i 番目の個体．この植物の体サイズ（個体の大きさ）x_i と肥料をやる施肥処理 f_i が種子数 y_i にどう影響しているのかを知りたい．

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!

この統計モデリング授業

改訂

- web 上の「課題」に回答してください
  - 回答もメイルで送信してください
  - 課題 https://goo.gl/z9yCJY
- 成績評価は「課題」の回答
  - 出欠関係なし

subject: kubostat

to: kubo@ees.hokudai.ac.jp

Performance Rating

改訂

- E-mail assignment https://goo.gl/z9yCJY
  - That's ALL!
- Attendance? NOT care.

subject: kubostat

to: kubo@ees.hokudai.ac.jp

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?

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!

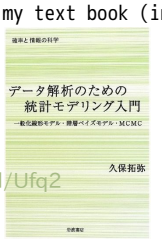
- Understand how to fit statistical models to your data データにあてはめられる統計モデルを作ろう
- Use the statistical software  to show your data structure

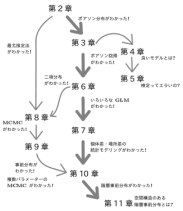
教科書とソフトウェア

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!!

Statistical software for this course 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

追記メモ：RStudio の紹介！

統計モデルとは何か?

What? statistical modeling?

「統計モデル」とは何か?

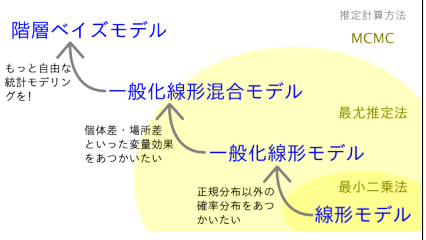
どんな統計解析においても
統計モデルが使用されている

- 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!

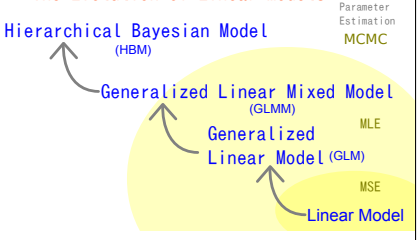
線形モデルの発展



GLM and extended GLMs!

a better statistica model for better data analysis!

The Evolution of Linear Models



たとえばこんなデータがあったしましょう

An example

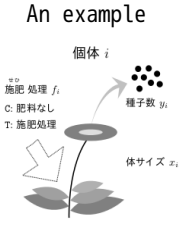
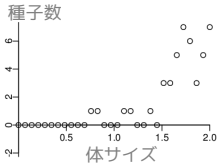


図 3.1 この例題に登場する架空植物の第 i 番目の個体．この植物の体サイズ（個体の大きさ）x_i と肥料をやる施肥処理 f_i が種子数 y_i にどう影響しているのかを知りたい．

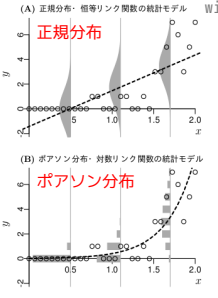
number of seeds

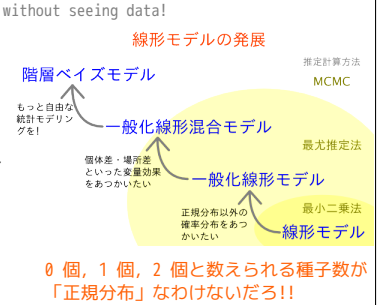


plant body size

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!





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
Course Overview

統計モデリング入門 2017 (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

京大霊長研の講義 https://goo.gl/z9yCJY

2017–11–14

ファイル更新時刻: 2017–11–11 16:02

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

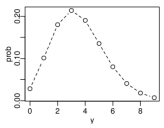


図 4 平均 $\lambda = 3.56$ のポアソン分布．種子数 y とその確率 prob の関係が示されている．図 3 の表を図にしたもの．R の plot() 関数の引数，type = "h" によって「丸と折れ線による図示」，lty = 2 によって「折れ線は破線で」と指示している．

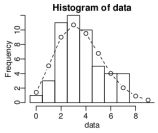
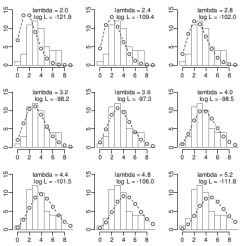


図 5 観測データと確率分布の対応をながめる．ヒストグラムは図 3 と同じ．それに重ねられている丸と縦線は y 個の種子をもつ個体数の予測．平均 3.56 の図 4 のポアソン分布の確率分布に全個体数 50 をかけて得られる．

さいゆう
最尤推定という考えかたを説明します



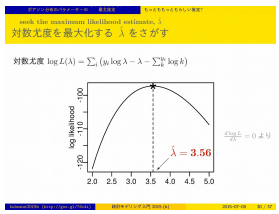


図 7 平均 λ（lambda）を変化させていったポアソン分布と，観測データへのあてはまりの良さ（対数尤度 logL）．すべてのヒストグラムは図 3 と同じ．

How to fit the distribution to the observation?

Maximum likelihood estimation!

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

京大霊長研の講義 https://goo.gl/z9yCJY

2017–11–14

ファイル更新時刻: 2017–11–11 16:02

ここで登場する ---
「何でも正規分布」ではダメ！という発想

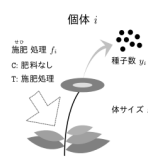


図 3.1 この例題に登場する架空植物の第 i 番目の個体．この植物の体サイズ（個体の大きさ）x_i と肥料をやる施肥処理 f_i が種子数 y_i にどう影響しているのかを知りたい．

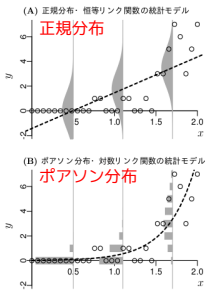


図 3.9 回帰モデルと確率分布の関係．また別の架空データに対して GLM をあてはめた例．破線は x とともに変化する平均値．グレイで

the “normal distirbution is NOT “normal”

Free の統計ソフトウェア R で統計モデリング

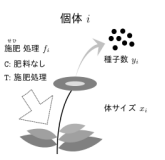


図 3.1 この例題に登場する架空植物の第 i 番目の個体サイズ（個体の大きさ）x_i と肥料をやる施肥処理にどう影響しているのかを知りたい．

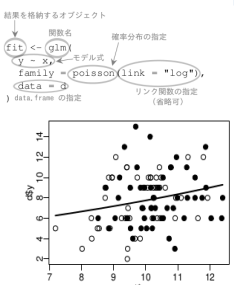


図 17 平均種子数 λ の予測．図 16 に λ の予測値（実線）を上がきしたもの．

統計モデリング入門 2017 (d)

model selection and statistical test
モデル選択と検定

久保拓弥 kubo@ees.hokudai.ac.jp

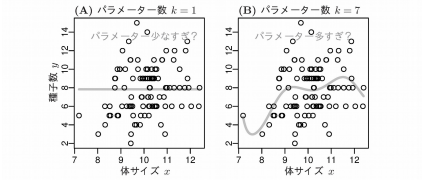
京大霊長研の講義 https://goo.gl/z9yCJY

2017–11–14

ファイル更新時刻: 2017–11–11 16:02

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?

model selection for better predictions
A. より良い予測をする統計モデルを探すこと

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 ←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 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

統計学って「検定」のこと?

「検定」って何なの?

fallacy of statistical significance?

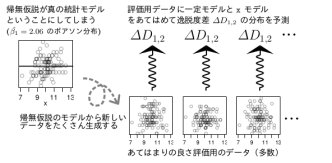


図 6 尤度比検定に必要な $\Delta D_{1,2}$ の分布の生成．まず帰無仮説である一定モデル（$\beta_1 = 2.00$, p.[illegible] 参照）が真の統計モデルだと仮定し，そこから得られるデータを使って逸脱度差 $\Delta D_{1,3}$ がどのような分布になるかを調べる．

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

京大霊長研の講義 https://goo.gl/z9yCJY

2017–11–14

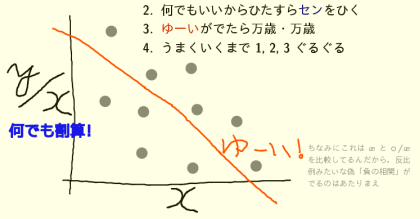
ファイル更新時刻: 2017–11–11 16:02

mesurement / mesurement?… sounds bad!
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

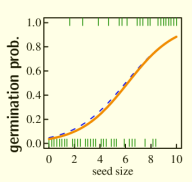


2012–11–02 k4 (2012–10–26 17.07 修正版) 14/ 44

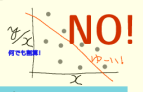
Use logistic regressions!
GLM のひとつ，ロジスティック回帰を使おう

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



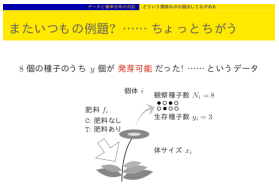
- むやみに 区画わけしない！
- 何でも 割り算するな！
- たくさん 図を描く
- 「観測データを説明する 確率分布は何か?」 を考える

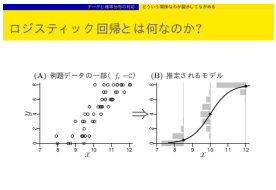


コツ: 不自然にデータをこねくりまわさない
データの性質・構造にあったモデリングを!

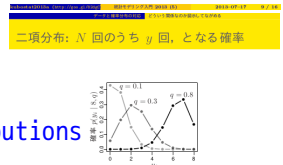
2012–11–02 k4 (2012–10–26 17.07 修正版) 43/ 44

GLM のひとつ，ロジスティック回帰を使おう





a statistical model
for fractions
using binomial distributions



統計モデリング入門 2017 (f)

Generalized Linear Mixed Model (GLMM)
一般化線形混合モデル

久保拓弥 kubo@ees.hokudai.ac.jp

京大霊長研の講義 https://goo.gl/z9yCJY

2017–11–14

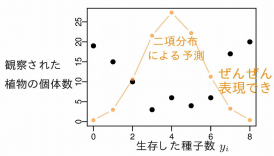
ファイル更新時刻: 2017–11–11 16:02

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……



GLM does NOT work?!

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

A solution: Hierarchical Bayesian GLM
GLM を階層ベイズモデル化して対処

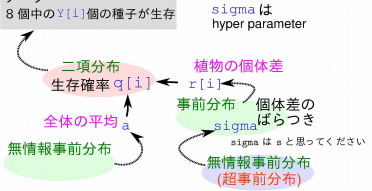
GLMM は階層ベイズモデルの一種　事前分布をどう選ぶかが重要

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

データ
8 個中の Y[i] 個の種子が生存

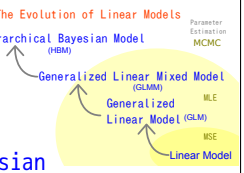
sigma は
hyper parameter



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



What for hierarchical Bayesian modeling? --- to detect interesting effects embedded in noisy & dirty data in the field of Ecology!

統計モデリング入門 2017 (g)

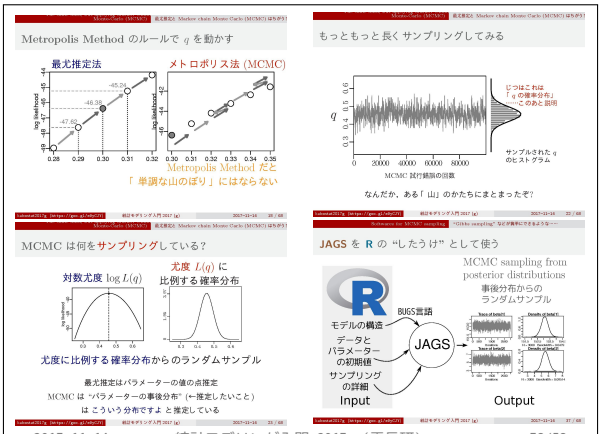
階層ベイズモデル Hierarchical Bayesian Model

久保拓弥 kubo@ees.hokudai.ac.jp

京大霊長研の講義 https://goo.gl/z9yCJY

2017–11–14

ファイル更新時刻: 2017–11–11 16:02



Applications of Hierarchical Bayesian Model
– nested data and ``paired data'' –

統計モデリング入門 2017 (h)

階層ベイズモデルの応用と "対応" のあるデータ解析

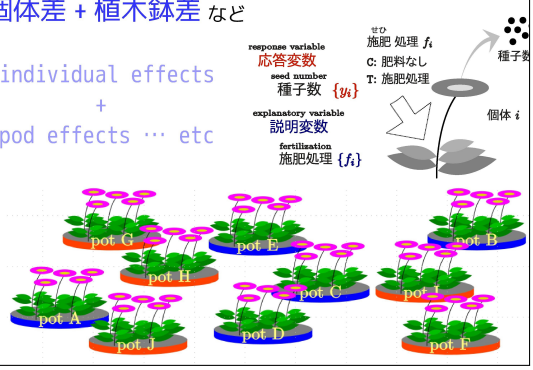
久保拓弥 kubo@ees.hokudai.ac.jp

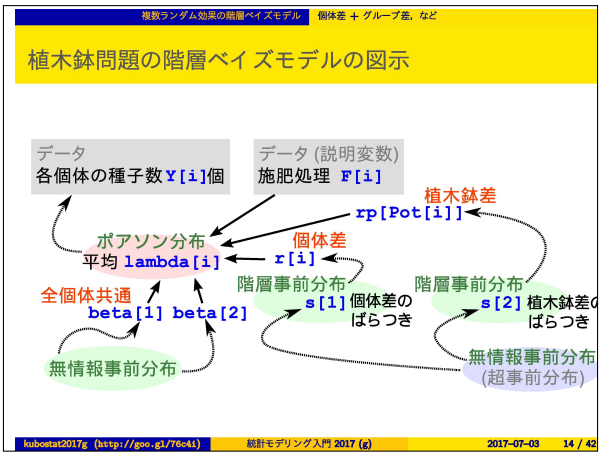
2017-11-15

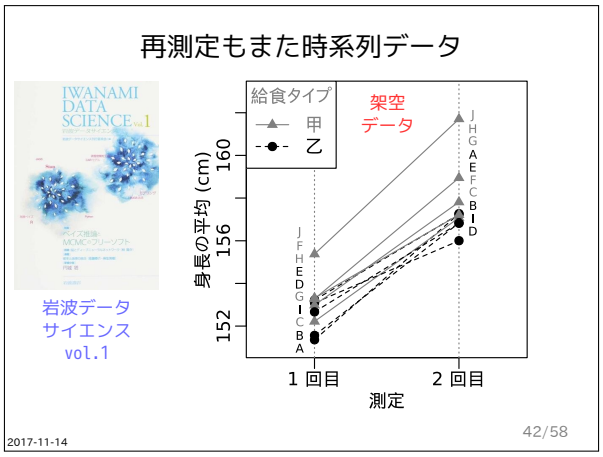
1. 複数ランダム効果の階層ベイズモデル

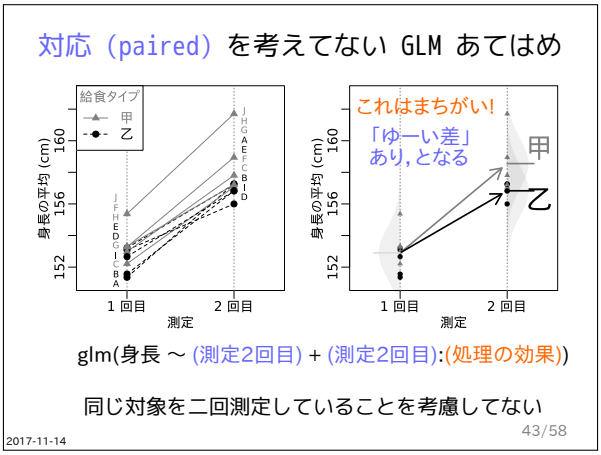
個体差 + 植木鉢差 など

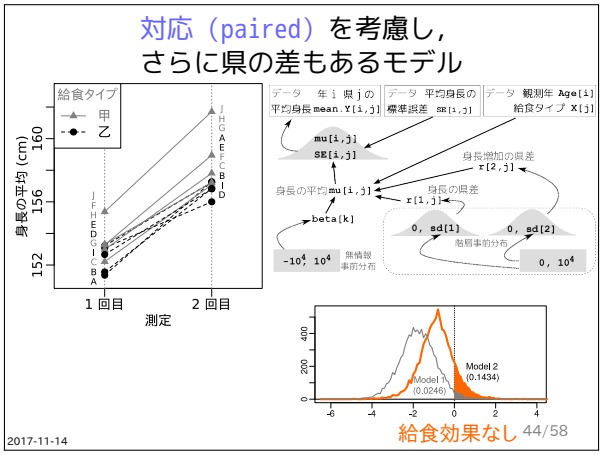
individual effects
+
pod effects ⋯ etc

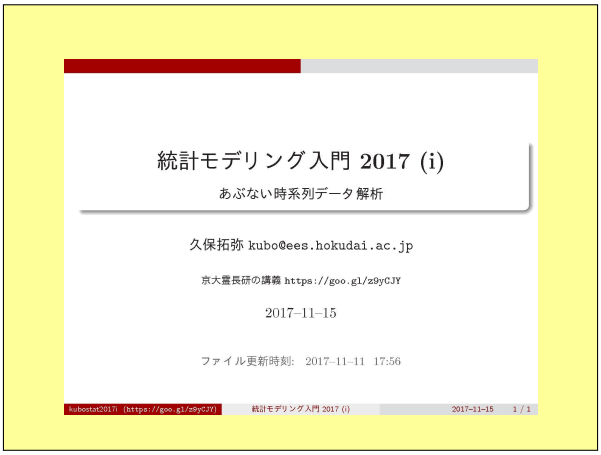


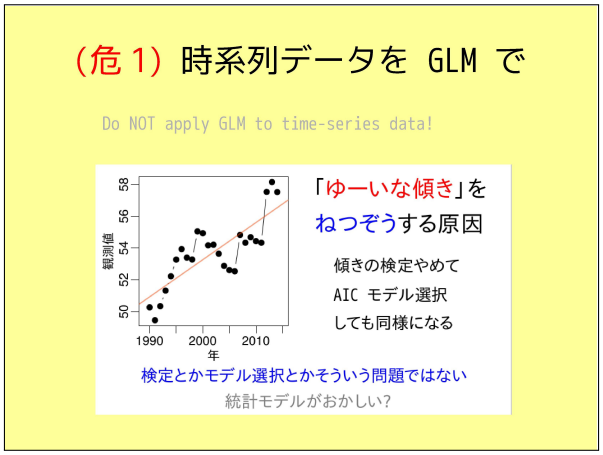

複数ランダム効果の階層ベイズモデル 個体差 + グループ差，など
植木鉢問題の階層ベイズモデルの図示
データ
各個体の種子数 Y[i]個
データ (説明変数)
施肥処理 F[i]
植木鉢差
rp[Pot[i]]
ポアソン分布
平均 lambda[i]
個体差
r[i]
全個体共通
beta[1] beta[2]
階層事前分布
s[1] 個体差のばらつき
階層事前分布
s[2] 植木鉢差のばらつき
無情報事前分布
無情報事前分布
(超事前分布)
kubostat2017g (http://goo.gl/76c41)
統計モデリング入門 2017 (g)
2017-07-03
14 / 42

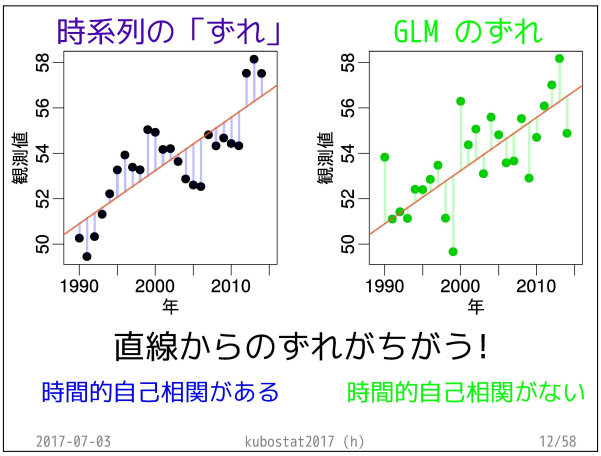

再測定もまた時系列データ
IWANAMI DATA SCIENCE vol.1
岩波データサイエンス vol.1
給食タイプ
甲
乙
架空データ
身長の平均 (cm)
152
156
160
1 回目
2 回目
測定
2017-11-14
42/58

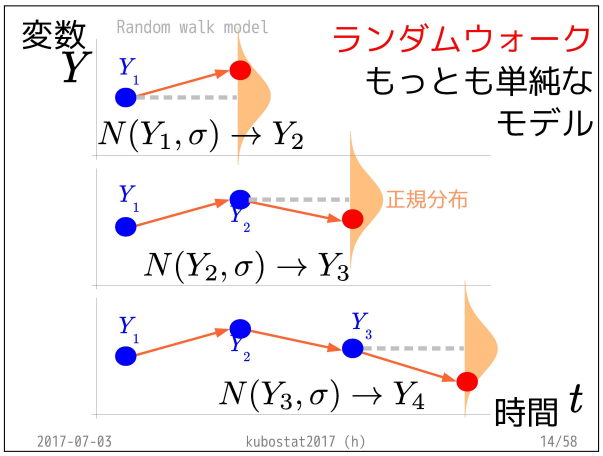

対応（paired）を考えてない GLM あてはめ
給食タイプ
甲
乙
身長の平均 (cm)
152
156
160
1 回目
2 回目
測定
これはまちがい!
「ゆーい差」あり，となる
glm(身長 ～ (測定2回目) + (測定2回目):(処理の効果))
同じ対象を二回測定していることを考慮してない
2017-11-14
43/58

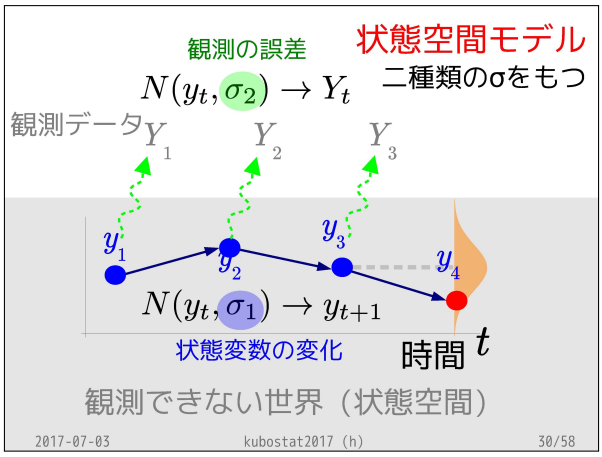

対応（paired）を考慮し，
さらに県の差もあるモデル
給食タイプ
甲
乙
身長の平均 (cm)
152
156
160
1 回目
2 回目
測定
mu[i,j]
SE[i,j]
身長増加の県差
r[2,j]
身長の平均
身長の県差
r[1,j]
beta[k]
0, sd[1]
0, sd[2]
階層事前分布
-10^4, 10^4
無情報事前分布
0, 10^4
Model 2
(0.1434)
Model 1
(0.0246)
給食効果なし
2017-11-14
44/58

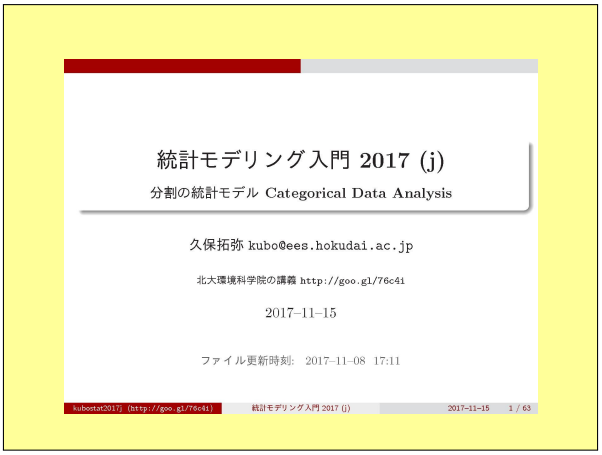

統計モデリング入門 2017 (i)
あぶない時系列データ解析
久保拓弥 kubo@ees.hokudai.ac.jp
京大農長研の講義 https://goo.gl/z9yCJY
2017–11–15
ファイル更新時刻: 2017–11–11 17:56

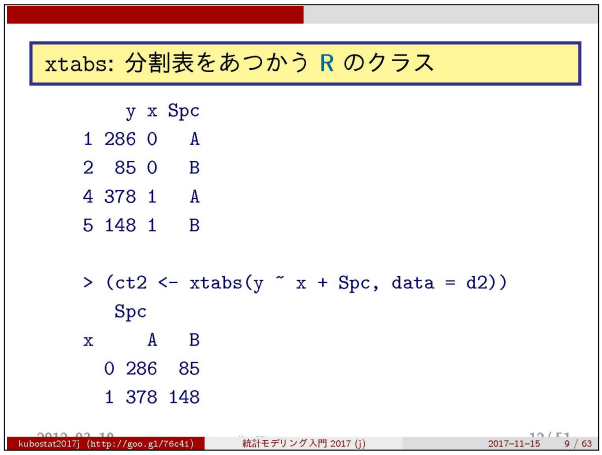

(危 1) 時系列データを GLM で
Do NOT apply GLM to time-series data!
観測値
50
52
54
56
58
1990
2000
2010
年
「ゆーいな傾き」を
ねつぞうする原因
傾きの検定やめて
AIC モデル選択
しても同様になる
検定とかモデル選択とかそういう問題ではない
統計モデルがおかしい?

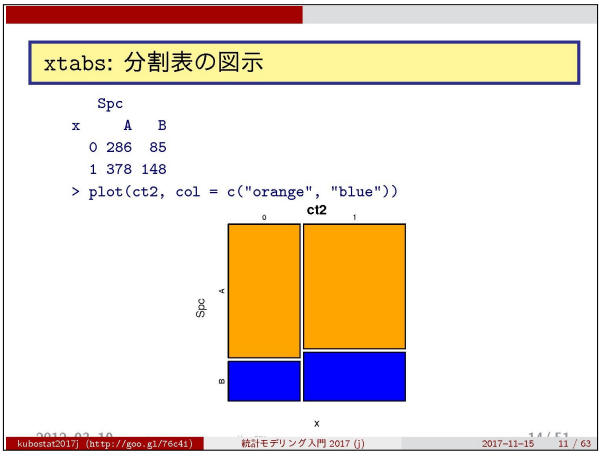

時系列の「ずれ」
GLM のずれ
観測値
50
52
54
56
58
1990
2000
2010
年
直線からのずれがちがう!
時間的自己相関がある
時間的自己相関がない
2017-07-03
kubostat2017 (h)
12/58

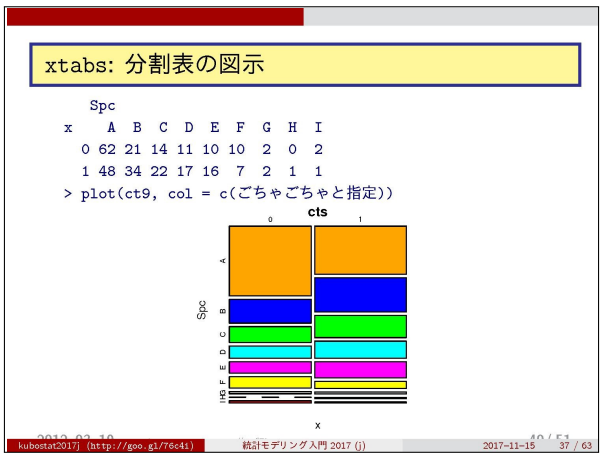

変数 Y
Random walk model
ランダムウォーク
もっとも単純な
モデル
Y_1
$N(Y_1, \sigma) \to Y_2$
Y_2
正規分布
$N(Y_2, \sigma) \to Y_3$
Y_3
$N(Y_3, \sigma) \to Y_4$
時間 t
2017-07-03
kubostat2017 (h)
14/58

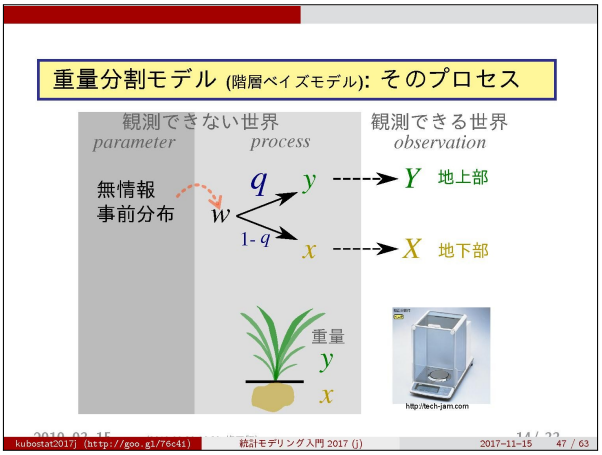

観測の誤差
状態空間モデル
二種類のσをもつ
$N(y_t, \sigma_2) \to Y_t$
観測データ
Y_1 Y_2 Y_3
y_1 y_2 y_3 y_4
$N(y_t, \sigma_1) \to y_{t+1}$
状態変数の変化
時間 t
観測できない世界（状態空間）
2017-07-03 kubostat2017 (h) 30/58

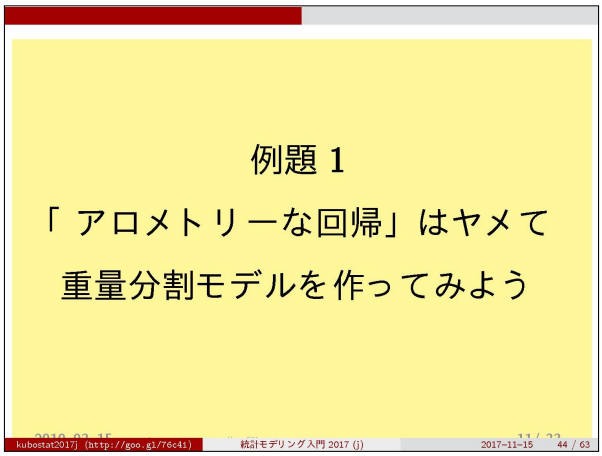

統計モデリング入門 2017 (j)
分割の統計モデル Categorical Data Analysis
久保拓弥 kubo@ees.hokudai.ac.jp
北大環境科学院の講義 http://goo.gl/76c41
2017–11–15
ファイル更新時刻: 2017–11–08 17:11

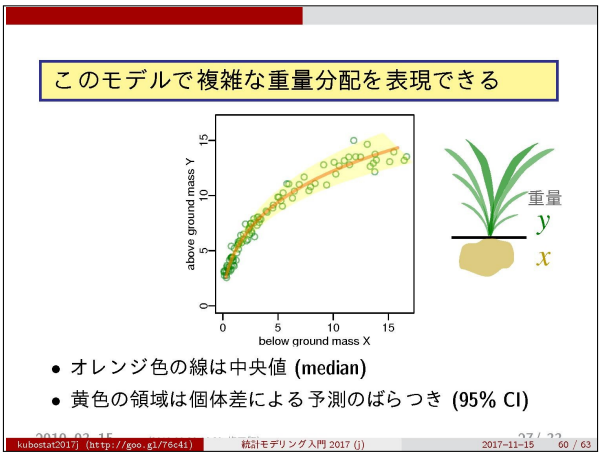

xtabs: 分割表をあつかう R のクラス
y x Spc
1 286 0 A
2 85 0 B
4 378 1 A
5 148 1 B
> (ct2 <- xtabs(y ~ x + Spc, data = d2))
Spc
x A B
0 286 85
1 378 148


xtabs: 分割表の図示
Spc
x A B
0 286 85
1 378 148
> plot(ct2, col = c("orange", "blue"))
ct2
Spc
x


xtabs: 分割表の図示
Spc
x A B C D E F G H I
0 62 21 14 11 10 10 2 0 2
1 48 34 22 17 16 7 2 1 1
> plot(ct9, col = c(ごちゃごちゃと指定))
cts
Spc
x


重量分割モデル (階層ベイズモデル): そのプロセス
観測できない世界
parameter
process
観測できる世界
observation
無情報
事前分布
q
w
$1-q$
y
x
Y 地上部
X 地下部
重量
y
x
http://tech-jam.com


例題 1
「アロメトリーな回帰」はヤメて
重量分割モデルを作ってみよう


このモデルで複雑な重量分配を表現できる
above ground mass Y
below ground mass X
0 5 10 15
重量
y
x
• オレンジ色の線は中央値 (median)
• 黄色の領域は個体差による予測のばらつき (95% CI)

A long trailer, done

any questions?

11/15（明日）午後は
「統計モデリングの個別相談」
を実施します（先着順）

https://goo.gl/z9yCJY の下のほう