統計モデリング入門 2018 (a)

生物多様性学特論

An overview: 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

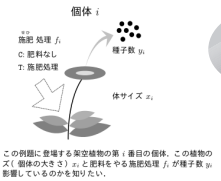


図 3.1 この例題に登場する架空植物の第 i 番目の個体．この植物の体サイズ（個体の大きさ）x_i と肥料をやる施肥処理 f_i が種子数 y_i にどう影響しているのかを知りたい．

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!

Performance Rating

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

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 class 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

統計モデリング授業の web page

http://goo.gl/76c4i

mailing list

http://goo.gl/f0vCn8

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!

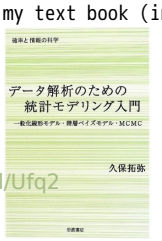
- Understand how to fit statistical models to your data データにあてはめられる統計モデルを作ろう
- Use the statistical software R to show your data structure

教科書とソフトウェア

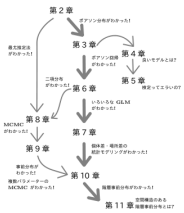
この授業は「統計モデリング入門」にそった内容を説明します

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?

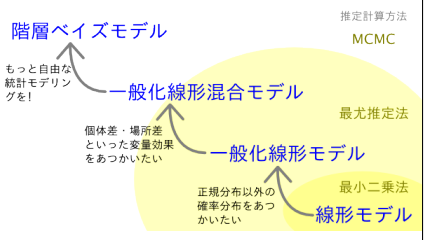
どんな統計解析においても
統計モデルが使用されている

- 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!

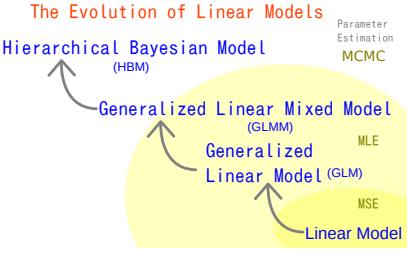
線形モデルの発展



GLM and extended GLMs!

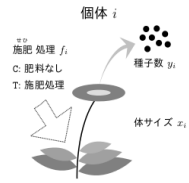
a better statistica model for better data analysis!

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



number of seeds

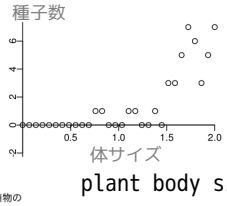
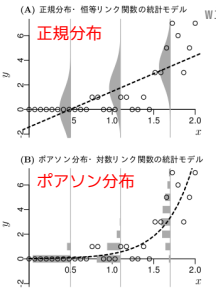
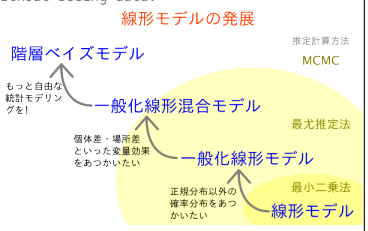


図 3.1 この例題に登場する架空植物の第 i 番目の個体．この植物の体サイズ（個体の大きさ）x_i と肥料をやる施肥処理 f_i が種子数 y_i にどう影響しているのかを知りたい．

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（1/3）

第 1 回: 6/18（月）観測されたパターンを説明する統計モデル
Introduction

第 2 回: 6/18（月）確率分布と最尤推定
Probability Distributions and Maximum Likelihood Estimation (MLE)

第 3 回: 6/25（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: 6/25（月）モデル選択と検定
Model Selection and Statistical Test

第 5 回: 7/02（水）一般化線形モデル: ロジスティック回帰
GLM: Logistic Regression

第 6 回: 7/02（月）階層ベイズモデル 1
Hierarchical Bayesian Models (HBM) 1

全体の流れ（3/3）

第 7 回: 7/09（月）繰り返し測定の階層ベイズモデル
Bayesian models for repeated measures

第 8 回: 7/09（月）時間変化データのベイズ統計モデル
Bayesian models for Time series data

next: Kohyama-san’ s Lecture
Plant Ecolgy

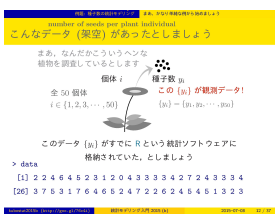
6/18

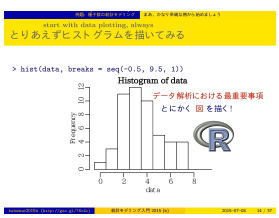
Overview Statistical Modeling 2018 (b)

Probability distributions and maximum likelihood estimation

さまざまな確率分布と最尤推定

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

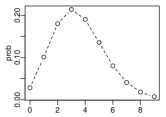


図 4 平均 $\lambda = 3.56$ のポアソン分布，種子数 y とその確率 prob の関係が示されている．図 3 の表を図にしたもの．R の plot() 関数の引数，type = "h" によって「丸と折れ線による図示」，lty = 2 によって「折れ線は破線で」と指示している．

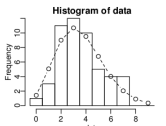


図 5 観測データと確率分布の対応をながめる．ヒストグラム法は図 3 と同じ．それに重ねられている丸と破線は y 値の種子をもつ個体数の予測．平均 3.56 の図 4 のポアソン分布の確率分布に全個体数 50 をかけて得られる．

さいゆう
最尤推定という考えかたを説明します

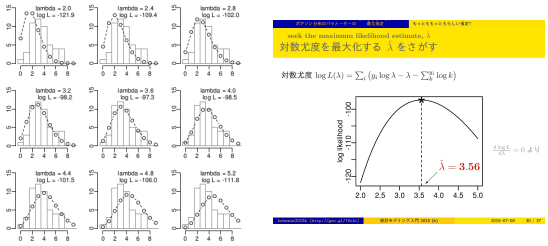


図 7 平均 λ（lambda）を変化させていったポアソン分布と，観測データへのあてはまりの良さ（対数尤度 logL）．すべてのヒストグラムは図 3 と同じ．

How to fit the distribution to the observation?

Maximum likelihood estimation!

6/25

Overview
Statistical Modeling 2018 (c)

Poisson regression
and generalized linear model

ポアソン回帰と GLM

ここで登場する ---
「何でも正規分布」ではダメ！という発想

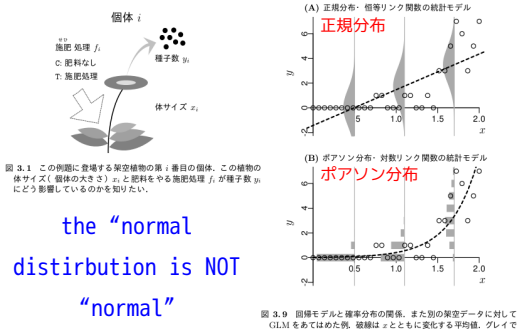


図 3.1 この例題に登場する架空植物の第 i 番目の個体．この植物の体サイズ（個体の大きさ）x_i と肥料をやる施肥処理 f_i が種子数 y_i にどう影響しているのかを知りたい．

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

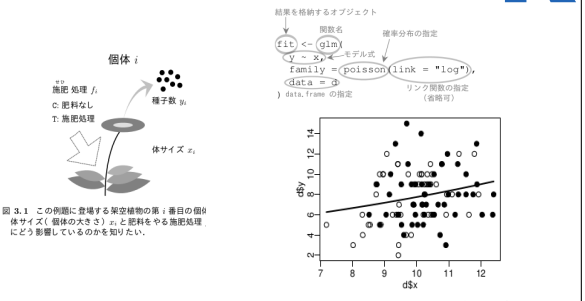


図 3.1 この例題に登場する架空植物の第 i 番目の個体（個体の大きさ）x_i と肥料をやる施肥処理 f_i が種子数 y_i にどう影響しているのかを知りたい．

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

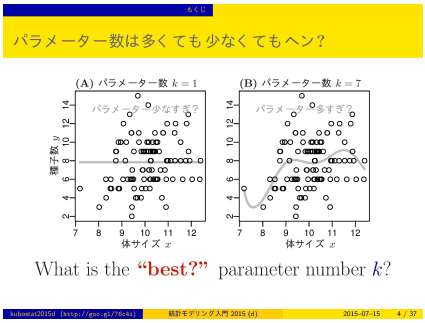
6/25

Overview
Statistical Modeling 2018 (d)

Model Selection
and Statistical Test

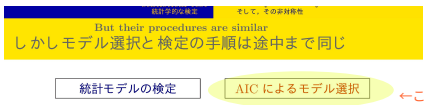
モデル選択と統計学的検定

statistical model selection
Q. モデル選択とは何か？



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

統計学って「検定」のこと？

「検定」って何なの？

fallacy of statistical significance?

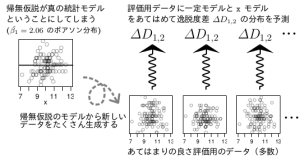


図 6 尤度比検定に必要な $\Delta D_{1,2}$ の分布の生成．まず帰無仮説である一定モデル（$\hat{\beta}_1 = 2.06$, p. [redacted] 参照）が真の統計モデルだと仮定し，そこから得られるデータを使って逸脱度差 $\Delta D_{1,2}$ がどのような分布になるかを調べる．

7/2

Overview
Statistical Modeling 2018 (e)

Logistic regression,
a generalized linear model

ロジスティック回帰

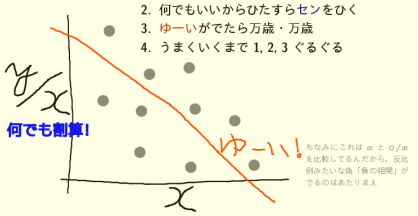
---

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

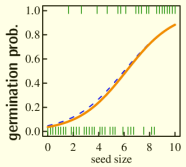


2012-11-02 k4 (2012-10-26 17:07 修正版) 14/ 44

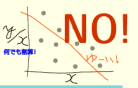
Use logistic regressions!
GLM のひとつ，ロジスティック回帰を使おう

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



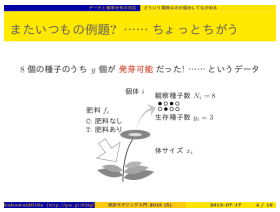
- むやみに 区画わけしない！
- 何でも 割り算するな！
- たくさん 図を描く
- 「観測データを説明する 確率分布は何か?」 を考える

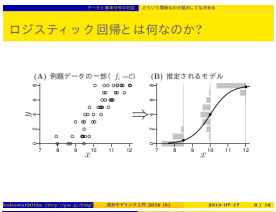


コツ: 不自然にデータをこねくりまわさない
データの性質・構造にあったモデリングを!

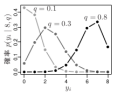
2012-11-02 k4 (2012-10-26 17:07 修正版) 43/ 44

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



二項分布: N 回のうち y 回，となる確率

7/2

Overview
Statistical Modeling 2018 (f)

Hierarchical Bayesian model
and MCMC sampling

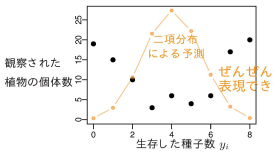
階層ベイズモデルと MCMC

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)

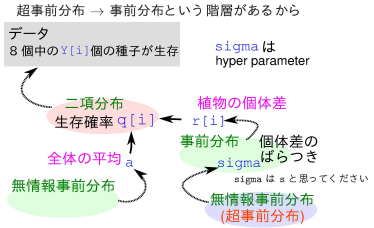
第 6 回と同じような例題を，こんどはベイズモデルを使ってモデリングします

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

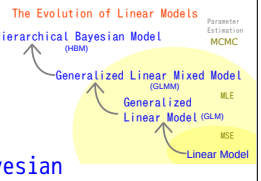


矢印は手順ではなく，依存関係をあらわしている

kubostat2015e (http://goo.gl/76C42) 統計モデリング入門 2015 (e) 2015-07-29 56 / 87

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!

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）
Modeling of time-series data as an application of hierarchical Bayesian modeling!

7/9

Overview
Statistical Modeling 2018 (g)

Modeling time change data
(short term)

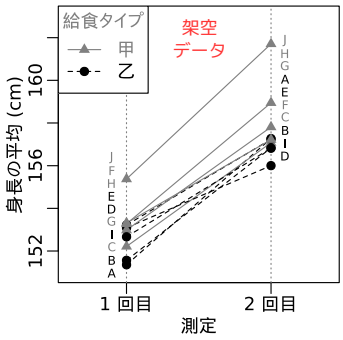
短い時系列データの統計モデル

A Time series model
for single step data
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

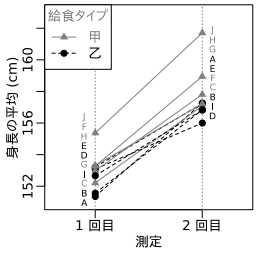
再測定もまた時系列データ

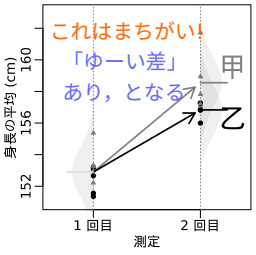
岩波データサイエンス vol.1



2018-06-18 44/56

対応（paired）を考えてない GLM あてはめ

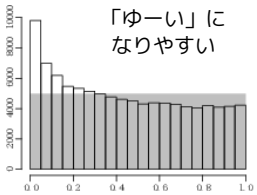


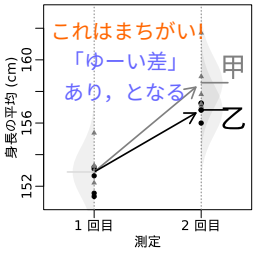


glm(身長 ~ (測定2回目) + (測定2回目):(処理の効果))
同じ対象を二回測定していることを考慮してない

2018-06-18 45/56

対応（paired）を考えてない GLM あてはめ

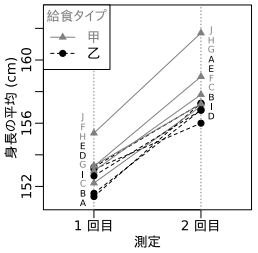


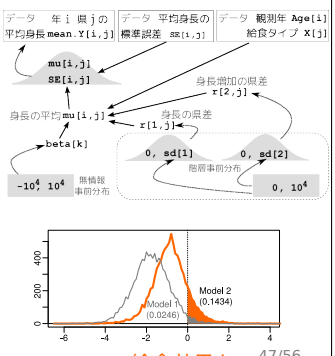


glm(身長 ~ (測定2回目) + (測定2回目):(処理の効果))
同じ対象を二回測定していることを考慮してない

2018-06-18 46/56

対応（paired）を考慮し，
さらに県の差もあるモデル





給食効果な

2018-06-18 47/56

7/9

Overview
Statistical Modeling 2018 (h)

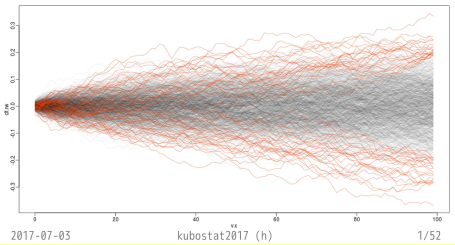
Modeling time series data
(long term)

長い時系列データの統計モデル

7/27（水）

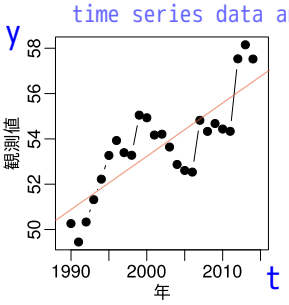
生態学の時系列データ解析でよく見る
『あぶない』モデリング

久保拓弥 mailto:kubo@ees.hokudai.ac.jp



時間相関のある時系列データに…

time series data and autocorrelation

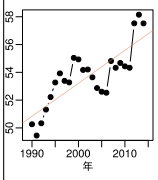


glm(y ~ t)

…と，モデルを
あてはめてみた

「やったーゆーいだ!!」……??

A fake significance

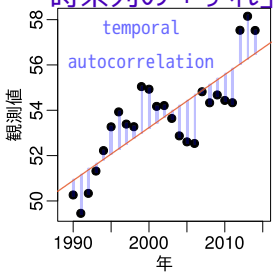



```
> summary(glm(formula = y ~ t))

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
```

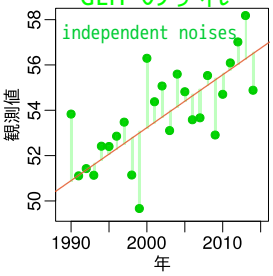
これはまちがい→ glm(時系列Y ~ 時間 t)

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

統計モデルづくりの要点

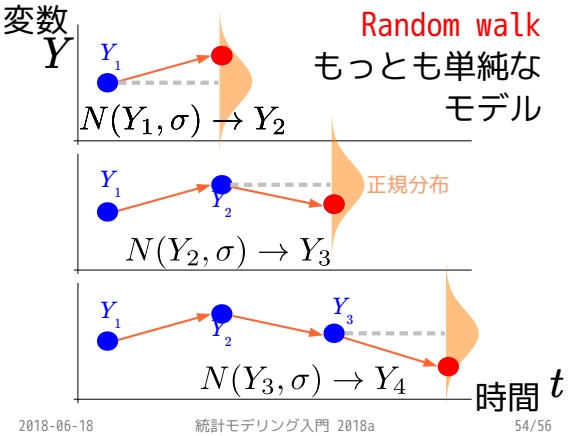
時系列データの解析は

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

Latent state model is a better model to know the characteristics of time-series data

Random walk
もっとも単純な
モデル

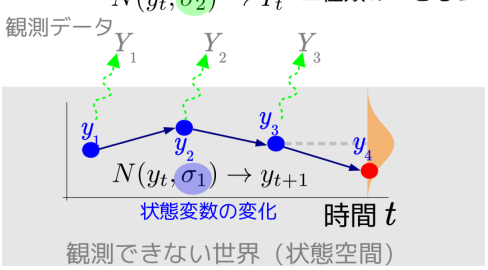


状態空間モデル + 観測モデル

Latent state variables + observation model

観測の誤差　状態空間モデル

$N(y_t, \sigma_2) \to Y_t$　二種類の σ をもつ



観測できない世界（状態空間）

今日はここまで

any questions?