統計モデリング入門 2015 (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

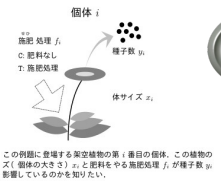


図 3.1 この例題に登場する架空植物の第 i 番目の個体．この植物の体サイズ（個体の大きさ）x_i と肥料をやる施肥処理 f_i が種子数 y_i にどう影響しているのかを知りたい．

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 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．登録のユーザーインターフェイスは日本語に変更できます．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

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．登録のユーザーインターフェイスは日本語に変更できます．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

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
- 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!

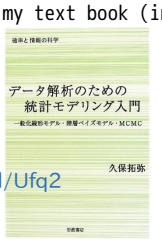
- Understand how to fit statistical models to your data データにあてはめられる統計モデルを作ろう
- Use the statistical software  to show your data structure

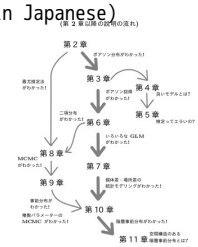
教科書とソフトウェア

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「講義のーと」もあります

授業 web page に「講義のーと」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

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

「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

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)

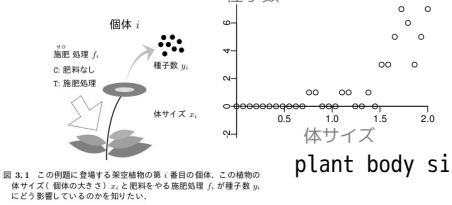


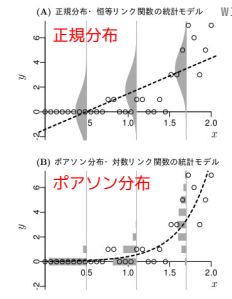
図 3.1 この例題に登場する架空植物の第 i 番目の個体．この植物の体サイズ（個体の大きさ）x_i と肥料をやる施肥処理 f_i が種子数 y_i にどう影響しているのかを知りたい．

number of seeds
種子数

体サイズ
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!



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

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

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!!

全体の流れ（1/2）

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（2/2）

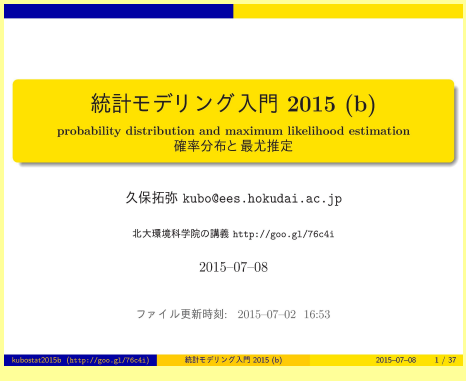
第 4 回: 7/15（水）モデル選択と検定
Model Selection and Statistical Test

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

7/8（水）



統計モデリング入門 2015 (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

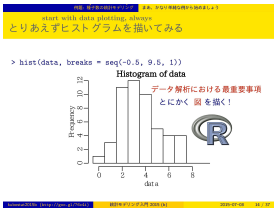
北大環境科学院の講義 http://goo.gl/76c4i

2015-07-08

ファイル更新時刻: 2015-07-02 16:53

kubostat2015b (http://goo.gl/76c4i) 統計モデリング入門 2015 (b) 2015-07-08 1 / 37

単純化した例題



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

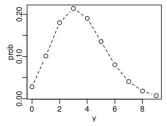


図 4　平均 λ = 3.56 のポアソン分布．種子数 y とその確率 prob の関係が示されている．図 5 の表を図にしたもの．R の plot() 関数の引数 type = "h" によって「丸と折れ線による図示」，lty = 2 によって「折れ線は破線で」と指示している．

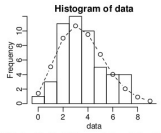


図 5　観測データと確率分布の対応をながめる．ヒストグラムは図 6 と同じ．それに重ねられている丸と破線は y 個の種子をもつ個体数の予測．平均 3.56 の図 4 のポアソン分布の確率分布に全個体数 50 をかけて得られる．

さいゆう 最尤推定という考えかたを説明します

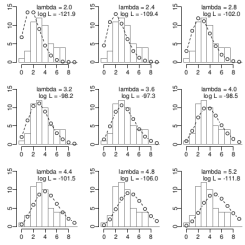
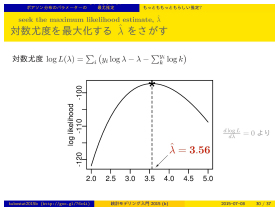


図 7　平均 λ（lambda）を変化させていった式ポアソン分布と，観測データへのあてはまりの良さ（対数尤度 logL）．すべてのヒストグラムは図 6 と同じ．



7/13（月）

統計モデリング入門 2015 (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

2015–07–13

ファイル更新時刻：2015–07–02 16:24

ここで登場する --- 「何でも正規分布」ではダメ！という発想

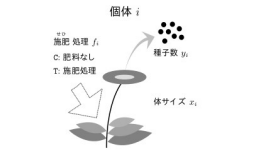


図 3.1　この例題に登場する架空植物の第 i 番目の個体．この植物の体サイズ（個体の大きさ）x_i と肥料をやる施肥処理 f_i が種子数 y_i にどう影響しているのかを知りたい．

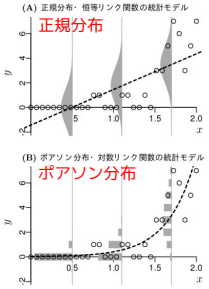


図 3.9　回帰モデルと確率分布の関係．また別の架空データに対して GLM をあてはめた例．破線は x とともに変化する平均値．グレイで

Free の統計ソフトウェア R で統計モデリング

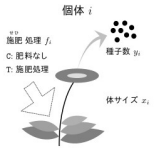


図 3.1　この例題に登場する架空植物の第 i 番目の個体サイズ（個体の大きさ）x_i と肥料をやる施肥処理にどう影響しているのかを知りたい．

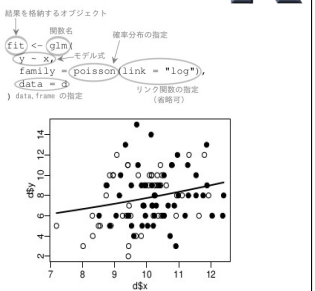


図 17　平均種子数 λ の予測．図 16 に λ の予測値（実線）を上がきしたもの．

7/15（水）

統計モデリング入門 2015 (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

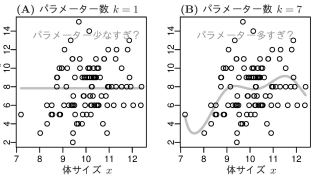
2015–07–15

ファイル更新時刻：2015–07–02 16:24

Q．モデル選択とは何か？

もくじ

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?

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 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?

「検定」ってエラいの?

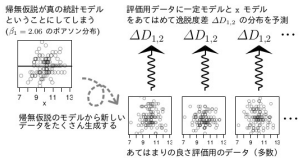
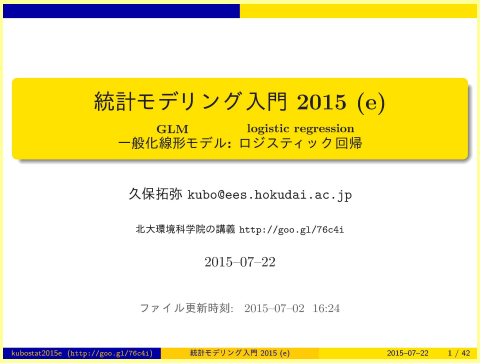


図 6　尤度比検定に必要な $\Delta D_{1,2}$ の分布の生成．まず帰無仮説である一定モデル（$\hat{p}_1 = 2.00$, p.[2] 参照）が既の統計モデルだと仮定し，そこから得られるデータを使って逸脱度差 $\Delta D_{1,2}$ がどのような分布になるかを調べる．

7/22（水）

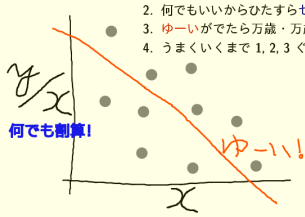


生物学のデータ解析は「割算」しまくり!!

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

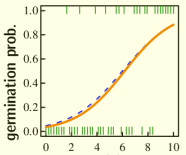


ちなみにこれは $\frac{m}{m}$ と $\frac{\circ}{m}$ を比較してるんだから，反応例みたいな偽「個の相関」がでるのはあたりまえ

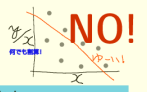
GLM のひとつ，ロジスティック回帰を使おう

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



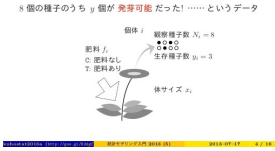
- むやみに 区画わけしない！
- 何でも 割り算するな！
- たくさん 図を描く
- 「観測データを説明する 確率分布は何か?」 を考える



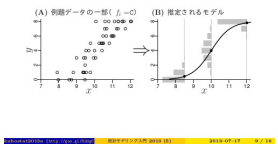
コツ: 不自然にデータをこねくりまわさない
データの性質・構造にあったモデリングを!

GLM のひとつ，ロジスティック回帰を使おう

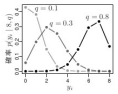
またいつもの例題? …… ちょっとちがう



ロジスティック回帰とは何なのか?



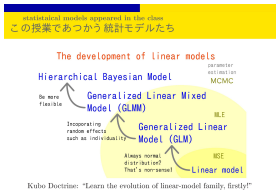
二項分布: N 回のうち y 回，となる確率



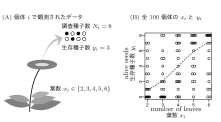
7/27（月）

GLM ではうまく対処できない問題

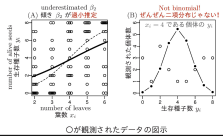
この授業であつかう統計モデルたち



今日の例題: 種子の生存確率……前回と同じ?!

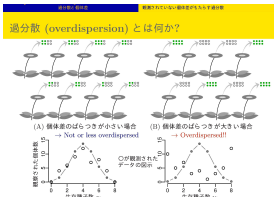


GLM では説明できないばらつき!

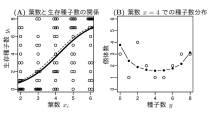


今まで「個体差」無視した統計モデル（GLM)を使っていた!

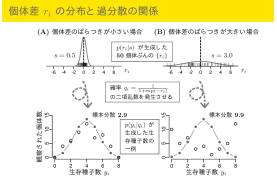
過分散 (overdispersion) とは何か?



推定された GLMM を使った 予測



個体差 r_i の分布と過分散の関係



一般化線形混合モデル
(Generalized Linear Model, GLMM) を使って問題解決

7/29（水）

統計モデリング入門 2015 (e)

階層ベイズモデル

久保拓弥 kubo@ees.hokudai.ac.jp

北大環境科学院の講義 http://goo.gl/76c4i

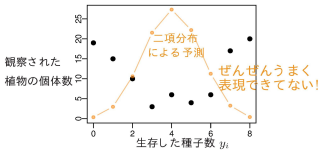
2015–07–29

ファイル更新時刻: 2015–07–02 16:24

GLM ではうまく説明できないデータ!?

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?

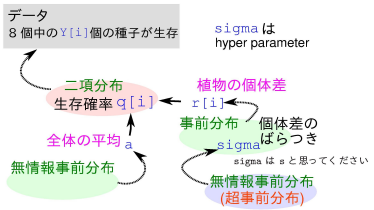
（「統計モデリング入門」第 10 章の最初の例題）

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

なぜ「階層」ベイズモデルと呼ばれるのか?

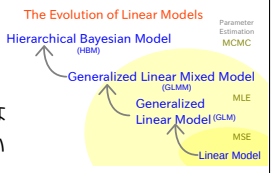
超事前分布 → 事前分布という階層があるから



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?

- 生態学や漁業のデータ解析は難しいから!
  - 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

全体の流れ（1/2）

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（2/2）

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