

統計モデリング入門 2016 (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

北大環境科学の講義 <http://goo.gl/76c4i>

2016-07-25

ファイル更新時刻: 2016-07-24 17:05

kubostat2016e (<http://goo.gl/76c4i>) 統計モデリング入門 2016 (e) 2016-07-25 1 / 42

もくじ

今日のハナシ I

- ① “ N 個のうち k 個が生きてる” タイプのデータ
count data or categorical data with upper bound
上限のあるカウントデータ

logistic regression
- ② ロジスティック回帰の部品
二項分布 binomial distribution と logit link function
- ③ ちょっとだけ 交互作用項 について
interaction term
complicate terms in linear predictor
線形予測子の中の複雑な項
- ④ 何でも「割算」するな!
NO data statistics!
use GLM with offset term
「脱」割算の offset 項わざ

kubostat2016e (<http://goo.gl/76c4i>) 統計モデリング入門 2016 (e) 2016-07-25 2 / 42


もくじ

今日の内容と「統計モデリング入門」との対応

<http://goo.gl/Ufq2>

今日はおもに「第6章 GLMの応用 範囲をひろげる」の内容を説明します。

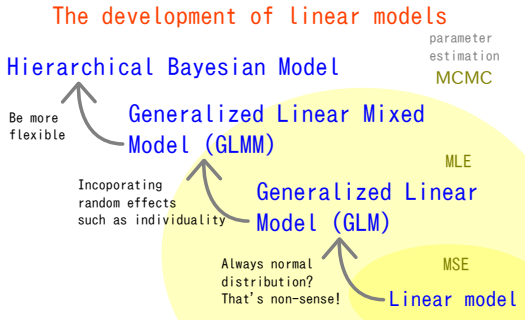
- 著者: 久保拓弥
- 出版社: 岩波書店
- 2012-05-18 刊行



kubostat2016e (<http://goo.gl/76c4i>) 統計モデリング入門 2016 (e) 2016-07-25 3 / 42

もくじ

statistical models appeared in the class
この授業であつかう統計モデルたち



Kubo Doctrine: “Learn the evolution of linear-model family, firstly!”

kubostat2016e (<http://goo.gl/76c4i>) 統計モデリング入門 2016 (e) 2016-07-25 4 / 42

もくじ

一般化線形モデルって何だろう?

Generalized Linear Model

一般化線形モデル (GLM)

- ポアソン回帰 (Poisson regression)
- **ロジスティック回帰 (logistic regression)**
- 直線回帰 (linear regression)
-

kubostat2016e (<http://goo.gl/76c4i>) 統計モデリング入門 2016 (e) 2016-07-25 5 / 42

もくじ

how to specify GLM
一般化線形モデルを作る

Generalized Linear Model

一般化線形モデル (GLM)

- probability distribution
- 確率分布は?
- linear predictor
- 線形予測子は?
- link function
- リンク関数は?

kubostat2016e (<http://goo.gl/76c4i>) 統計モデリング入門 2016 (e) 2016-07-25 6 / 42

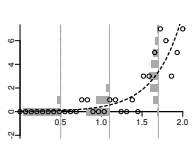
もくじ

how to specify Poisson regression model, a GLM
GLM のひとつである **ポアソン回帰モデル** を指定する

ポアソン回帰のモデル

probability distribution Poisson distribution

- 確率分布: **ポアソン分布**
- linear predictor
- 線形予測子: e.g., $\beta_1 + \beta_2 x_i$
- link function log link function
- リンク関数: **対数リンク関数**



kubostat2016e (http://goo.gl/76c4i) 統計モデリング入門 2016 (e) 2016-07-25 7 / 42

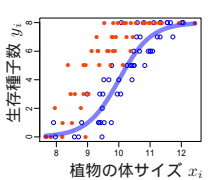
もくじ

how to specify logistic regression model, a GLM
GLM のひとつである **logistic 回帰モデル** を指定する

ロジスティック回帰のモデル

probability distribution binomial distribution

- 確率分布: **二項分布**
- linear predictor
- 線形予測子: e.g., $\beta_1 + \beta_2 x_i$
- link function
- リンク関数: **logit リンク関数**



kubostat2016e (http://goo.gl/76c4i) 統計モデリング入門 2016 (e) 2016-07-25 8 / 42

"N 個のうち k 個が生きてる" タイプのデータ 上限のあるカウントデータ

1. "N 個のうち k 個が生きてる" タイプのデータ

count data or categorical data with upper bound
上限のあるカウントデータ

$y_i \in \{0, 1, 2, \dots, 8\}$

kubostat2016e (http://goo.gl/76c4i) 統計モデリング入門 2016 (e) 2016-07-25 9 / 42

"N 個のうち k 個が生きてる" タイプのデータ 上限のあるカウントデータ

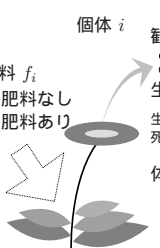
またいつもの例題? ちょっとちがう

8 個の種子のうち y 個が **発芽可能** だった! というデータ

肥料 f_i

C: 肥料なし

T: 肥料あり



個体 i

観察種子数 $N_i = 8$

生存種子数 $y_i = 3$


生存種子 (alive) は
死亡種子 (dead) は

体サイズ x_i

kubostat2016e (http://goo.gl/76c4i) 統計モデリング入門 2016 (e) 2016-07-25 10 / 42

"N 個のうち k 個が生きてる" タイプのデータ 上限のあるカウントデータ

Reading data file

データファイルを読みこむ 

data4a.csv は CSV (comma separated value) format file なので, R で読みこむには以下のようにする:

```
> d <- read.csv("data4a.csv")
```

or

```
> d <- read.csv(
+ "http://hosho.ees.hokudai.ac.jp/~kubo/stat/2015/Fig/binomial/data4a.csv")
```

データは d と名付けられた data frame (「表」みたいなもの) に格納される

kubostat2016e (http://goo.gl/76c4i) 統計モデリング入門 2016 (e) 2016-07-25 11 / 42

"N 個のうち k 個が生きてる" タイプのデータ 上限のあるカウントデータ

data frame d を調べる

```
> summary(d)
```

	N	y	x	f
Min.	:8	:0.00	: 7.660	C:50
1st Qu.	:8	:3.00	1st Qu. : 9.338	T:50
Median	:8	:6.00	Median : 9.965	
Mean	:8	:5.08	Mean : 9.967	
3rd Qu.	:8	:8.00	3rd Qu. :10.770	
Max.	:8	:8.00	Max. :12.440	

kubostat2016e (http://goo.gl/76c4i) 統計モデリング入門 2016 (e) 2016-07-25 12 / 42

「N 個のうち y 個が生きてる」タイプのデータ 上層のあるカウントデータ

まずはデータを図にしてみる

```
> plot(d$x, d$y, pch = c(21, 19)[d$f])
> legend("topleft", legend = c("C", "T"), pch = c(21, 19))
```

生存種子数 y_i

植物の体サイズ x_i

fertilization effective
今回は 施肥処理 がきいている?

kubostat2016e (http://goo.gl/76c4i) 統計モデリング入門 2016 (e) 2016-07-25 13 / 42

ロジスティック回帰 の部品 二項分布 binomial distribution と logit link function

logistic regression

2. ロジスティック回帰の部品

二項分布 binomial distribution と logit link function

kubostat2016e (http://goo.gl/76c4i) 統計モデリング入門 2016 (e) 2016-07-25 14 / 42

ロジスティック回帰 の部品 二項分布 binomial distribution と logit link function

binomial distribution

二項分布 : N 回のうち y 回, となる確率

$$p(y | N, q) = \binom{N}{y} q^y (1-q)^{N-y}$$

$\binom{N}{y}$ は「N 個の観察種子の中から y 個の生存種子を選ばずる場合の数」

確率 $p(y_i | 8, q)$

kubostat2016e (http://goo.gl/76c4i) 統計モデリング入門 2016 (e) 2016-07-25 15 / 42

ロジスティック回帰 の部品 二項分布 binomial distribution と logit link function

logistic curve

ロジスティック曲線 とはこういうもの

ロジスティック関数の関数形 (linear predictor z_i : 線形予測子, e.g. $z_i = \beta_1 + \beta_2 x_i$)

$$q_i = \text{logistic}(z_i) = \frac{1}{1 + \exp(-z_i)}$$

```
> logistic <- function(z) 1 / (1 + exp(-z)) # 関数の定義
> z <- seq(-6, 6, 0.1)
> plot(z, logistic(z), type = "l")
```

確率 q

線形予測子 z

kubostat2016e (http://goo.gl/76c4i) 統計モデリング入門 2016 (e) 2016-07-25 16 / 42

ロジスティック回帰 の部品 二項分布 binomial distribution と logit link function

β_1 and β_2 change logistic curve

パラメーターが変化すると.....

黒い曲線は $\{\beta_1, \beta_2\} = \{0, 2\}$. (A) $\beta_2 = 2$ と固定して β_1 を変化させた場合.
(B) $\beta_1 = 0$ と固定して β_2 を変化させた場合.

確率 q

説明変数 x

説明変数 x

パラメーター $\{\beta_1, \beta_2\}$ や説明変数 x がどんな値をとっても確率 q は $0 \leq q \leq 1$ となる便利な関数

kubostat2016e (http://goo.gl/76c4i) 統計モデリング入門 2016 (e) 2016-07-25 17 / 42

ロジスティック回帰 の部品 二項分布 binomial distribution と logit link function

logit link function

- logistic 関数

$$q = \frac{1}{1 + \exp(-(\beta_1 + \beta_2 x))} = \text{logistic}(\beta_1 + \beta_2 x)$$
- logit 変換

$$\text{logit}(q) = \log \frac{q}{1-q} = \beta_1 + \beta_2 x$$

logit は logistic の逆関数, logistic は logit の逆関数
logit is the inverse function of logistic function, vice versa

kubostat2016e (http://goo.gl/76c4i) 統計モデリング入門 2016 (e) 2016-07-25 18 / 42

ロジスティック回帰 の部 二項分布 binomial distribution と logit link function

logistic regression MLE for β_1 and β_2

R でロジスティック回帰 — β_1 と β_2 の最尤推定

(A) 例題データの一部 ($f_i = C$)

(B) 推定されるモデル

```

> glm(cbind(y, N - y) ~ x + f, data = d, family = binomial)
...
Coefficients:
(Intercept)          x          fT
    -19.536         1.952         2.022
    
```

kubostat2016e (http://goo.gl/76c4i) 統計モデリング入門 2016 (e) 2016-07-25 19 / 42

ロジスティック回帰 の部 二項分布 binomial distribution と logit link function

統計モデルの予測: 施肥処理によって応答が違う

(A) 施肥処理なし ($f_i = C$)

(B) 施肥処理あり ($f_i = T$)

kubostat2016e (http://goo.gl/76c4i) 統計モデリング入門 2016 (e) 2016-07-25 20 / 42

ちょっとだけ 交互作用項 について 線形予測子の中の複雑な項

interaction term

3. ちょっとだけ交互作用項 について

complicate terms in linear predictor
線形予測子の中の複雑な項

ロジスティック回帰を例に

kubostat2016e (http://goo.gl/76c4i) 統計モデリング入門 2016 (e) 2016-07-25 21 / 42

ちょっとだけ 交互作用項 について 線形予測子の中の複雑な項

交互作用項とは何か?

$$\text{logit}(q) = \log \frac{q}{1-q} = \beta_1 + \beta_2 x + \beta_3 f + \beta_4 x f$$

... in case that $\beta_4 < 0$, sometimes it predicts ...

kubostat2016e (http://goo.gl/76c4i) 統計モデリング入門 2016 (e) 2016-07-25 22 / 42

ちょっとだけ 交互作用項 について 線形予測子の中の複雑な項

in today's example no interaction effect

この例題データの場合, 交互作用はない

```

^^I glm(y ~ x + f, ...) glm(y ~ x + f + x:f, ...)
    
```

(A) 交互作用のないモデル

(B) 交互作用のあるモデル

little difference
差がほとんどない

kubostat2016e (http://goo.gl/76c4i) 統計モデリング入門 2016 (e) 2016-07-25 23 / 42

何でも「割算」するな! 「脱」割算の offset 項わざ

NO data statistics!

4. 何でも「割算」するな!

use GLM with offset term
「脱」割算の offset 項わざ

ポアソン回帰を強めてみる

kubostat2016e (http://goo.gl/76c4i) 統計モデリング入門 2016 (e) 2016-07-25 24 / 42

何でも「割算」するな! 「脱」割算の offset 頂わざ

割算値ひねくるデータ解析はなぜよくないのか?

- 観測値 / 観測値 がどんな確率分布にしたがうのか見とおしが悪く、さらに説明要因との対応づけが難しくなる
- 情報が失われる: 「10 打数 3 安打」と「200 打数 60 安打」, 「どちらも 3 割バッター」と言ってよいのか?
- 割算値を使わないほうが見とおしのよい, 合理的なデータ解析ができる (今回の授業の主題)
- したがって割算値を使ったデータ解析は不利な点ばかり, そんなことをする必要はどこにもない

kubostat2016e (<http://goo.gl/76c4i>) 統計モデリング入門 2016 (e) 2016-07-25 25 / 42

何でも「割算」するな! 「脱」割算の offset 頂わざ

How to avoid data/data? 避けられるわりざん

avoidable data/data values

- 避けられる割算値
 - 確率
 - 例: N 個のうち k 個にある事象が発生する確率
 - use statistical model with binomial distribution
 - 対策: ロジスティック回帰など二項分布モデルで
 - indices such as densities
 - 例: 人口密度, specific leaf area (SLA) など
 - use offset term! described later
 - 対策: offset 頂わざ — このあと解説!

kubostat2016e (<http://goo.gl/76c4i>) 統計モデリング入門 2016 (e) 2016-07-25 26 / 42

何でも「割算」するな! 「脱」割算の offset 頂わざ

unfortunately, sometimes fractions appear ... 避けにくいわりざん

hard to avoid ...

- 避けにくい割算値
 - outputs from some measuring machines
 - 測定機器が内部で割算した値を出力する場合
 - sometimes we have no choice but plot data/data values ...
 - 割算値で作図せざるをえない場合があるかも

kubostat2016e (<http://goo.gl/76c4i>) 統計モデリング入門 2016 (e) 2016-07-25 27 / 42

何でも「割算」するな! 「脱」割算の offset 頂わざ

example population densities in research plots offset 頂の 例題 : 調査区画内の個体密度

light intensity index

- 何か架空の植物個体の密度が「明るさ」 x に応じて どう変わるかを知りたい
- 明るさ は $\{0.1, 0.2, \dots, 1.0\}$ の 10 段階で観測した



これだけなら単純に `glm(..., family = poisson)` とすればよいのだが

kubostat2016e (<http://goo.gl/76c4i>) 統計モデリング入門 2016 (e) 2016-07-25 28 / 42

何でも「割算」するな! 「脱」割算の offset 頂わざ

What? Differences in plot size?! 「場所によって調査区の面積を変えました」?!



- 明るさ x と面積 A を同時に考慮する必要あり
- ただし「密度 = 個体数 / 面積」といった割算値解析はやらない!
- `glm()` の offset 頂わざでうまく対処できる
- ともあれその前に観測データを図にしてみる

kubostat2016e (<http://goo.gl/76c4i>) 統計モデリング入門 2016 (e) 2016-07-25 29 / 42

何でも「割算」するな! 「脱」割算の offset 頂わざ

R の data.frame: 面積 Area, 明るさ x , 個体数 y

```

> load("d2.RData")
> head(d, 8) # 先頭 8 行の表示
      Area x y
1 0.017249 0.5 0
2 1.217732 0.3 1
3 0.208422 0.4 0
4 2.256265 0.1 0
5 0.794061 0.7 1
6 0.396763 0.1 1
7 1.428059 0.6 1
8 0.791420 0.3 1
  
```

kubostat2016e (<http://goo.gl/76c4i>) 統計モデリング入門 2016 (e) 2016-07-25 30 / 42

何でも「割算」するな! 「脱」割算の offset 頂わざ

明るさ vs 割算値図の図

```
> plot(d$x, d$y / d$Area)
```

いまいちよくわからない

kubostat2016e (http://goo.gl/76c4i) 統計モデリング入門 2016 (e) 2016-07-25 31 / 42

何でも「割算」するな! 「脱」割算の offset 頂わざ

面積 A vs 個体数 y の図

```
> plot(d$Area, d$y)
```

面積 A とともに区画内の個体数 y が增大するようだ

kubostat2016e (http://goo.gl/76c4i) 統計モデリング入門 2016 (e) 2016-07-25 32 / 42

何でも「割算」するな! 「脱」割算の offset 頂わざ

明るさ x の情報 (マルの大きさ) も図に追加

```
> plot(d$Area, d$y, cex = d$x * 2)
```

同じ面積でも明るいほど個体数が多い?

kubostat2016e (http://goo.gl/76c4i) 統計モデリング入門 2016 (e) 2016-07-25 33 / 42

何でも「割算」するな! 「脱」割算の offset 頂わざ

密度が明るさ x に依存する統計モデル

- 区画内の個体数 y の平均は面積 × 密度
- 密度は明るさ x で変化する

kubostat2016e (http://goo.gl/76c4i) 統計モデリング入門 2016 (e) 2016-07-25 34 / 42

何でも「割算」するな! 「脱」割算の offset 頂わざ

「平均個体数 = 面積 × 密度」モデル

- ある区画 i の応答変数 y_i は平均 λ_i のポアソン分布にしたがうと仮定:
 $y_i \sim \text{Pois}(\lambda_i)$
- 平均値 λ_i は面積 A_i に比例し、密度は明るさ x_i に依存する
 $\lambda_i = A_i \exp(\beta_1 + \beta_2 x_i)$
つまり $\lambda_i = \exp(\beta_1 + \beta_2 x_i + \log(A_i))$ となるので
 $\log(\lambda_i) = \beta_1 + \beta_2 x_i + \log(A_i)$ 線形予測子は右辺のようになる
このとき $\log(A_i)$ を offset 項とよぶ (係数 β がない)

kubostat2016e (http://goo.gl/76c4i) 統計モデリング入門 2016 (e) 2016-07-25 35 / 42

何でも「割算」するな! 「脱」割算の offset 頂わざ

この問題は GLM であつかえる!

- family: poisson, ポアソン分布
- link 関数: "log"
- モデル式: $y \sim x$
- offset 項の指定: $\log(\text{Area})$
 - 線形予測子 $z = \beta_1 + \beta_2 x + \log(\text{Area})$
 a, b は推定すべきパラメーター
 - 応答変数の平均値を λ とすると $\log(\lambda) = z$
つまり $\lambda = \exp(z) = \exp(\beta_1 + \beta_2 x + \log(\text{Area}))$
 - 応答変数 は平均 λ のポアソン分布に従う:

kubostat2016e (http://goo.gl/76c4i) 統計モデリング入門 2016 (e) 2016-07-25 36 / 42

何でも「割算」するな! 「脱」割算の offset 頂わざ

glm() 関数の指定

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

結果を格納するオブジェクト: fit
 モデル式: y ~ x
 関数名: glm
 確率分布の指定: family = poisson(link = "log")
 offset の指定: offset = log(Area)
 リンク関数の指定 (省略可): link = "log"

kubostat2016e (http://goo.gl/76c4i) 統計モデリング入門 2016 (e) 2016-07-25 37 / 42

何でも「割算」するな! 「脱」割算の offset 頂わざ

R の glm() 関数による推定結果

```
> fit <- glm(y ~ x, family = poisson(link = "log"), data = d,
  offset = log(Area))
> print(summary(fit))
```

Call:
 glm(formula = y ~ x, family = poisson(link = "log"), data = d,
 offset = log(Area))

(... 略...)

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.321	0.160	2.01	0.044
x	1.090	0.227	4.80	1.6e-06

kubostat2016e (http://goo.gl/76c4i) 統計モデリング入門 2016 (e) 2016-07-25 38 / 42

何でも「割算」するな! 「脱」割算の offset 頂わざ

Plotting the model prediction based on estimation 推定結果にもとづく予測を図にしてみる

solid lines
 • 実線 は glm() の推定結果にもとづく 予測
 • 点線 lines "true" model
 • 破線 は データ生成時に指定した関係

kubostat2016e (http://goo.gl/76c4i) 統計モデリング入門 2016 (e) 2016-07-25 39 / 42

何でも「割算」するな! 「脱」割算の offset 頂わざ

まとめ: glm() の offset 頂わざで「脱」割算

- 平均値が面積などに比例する場合は、この面積などを **offset 項** として指定する
- 平均 = 面積 × 密度、というモデルの密度を exp(線形予測子) として定式化する

kubostat2016e (http://goo.gl/76c4i) 統計モデリング入門 2016 (e) 2016-07-25 40 / 42

何でも「割算」するな! 「脱」割算の offset 頂わざ

Improve your statistical model and remove data/data values! 統計モデルを工夫してわりざんやめよう

avoidable data/data values

- 避けられる割算値
 - probability
 - 確率

例: N 個のうち k 個にある事象が発生する確率

use statistical model with binomial distribution
 対策: ロジスティック回帰など**二項分布モデル**で
 - indices such as densities
 - 密度などの指数

例: 人口密度, specific leaf area (SLA) など

use offset term! Improve your statistical model!
 対策: **offset 頂わざ** — 統計モデリングの工夫!

kubostat2016e (http://goo.gl/76c4i) 統計モデリング入門 2016 (e) 2016-07-25 41 / 42

何でも「割算」するな! 「脱」割算の offset 頂わざ

次回予告 The next topic

種子数分布

N 個のうち y 個 という形式のデータなのに 二項分布ではまったく説明できない!

階層ベイズモデル Hierarchical Bayesian Model (HBM)

kubostat2016e (http://goo.gl/76c4i) 統計モデリング入門 2016 (e) 2016-07-25 42 / 42