

統計モデリング入門 2017 (e)
 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>
 2019-08-05

ファイル更新時刻: 2019-07-18 17:28

kubostat2017e (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (e) 2019-08-05 1 / 46

もくじ

今日のハナシ I

- ① “ N 個のうち y 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
- 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
 線形予測子の中の複雑な項 complicate terms in linear predictor
- ④ 何でも「割算」するな!
 「脱」割算の offset 項わざ

kubostat2017e (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (e) 2019-08-05 2 / 46


もくじ

今日の内容と「統計モデリング入門」との対応

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。

<http://goo.gl/Ufq2>

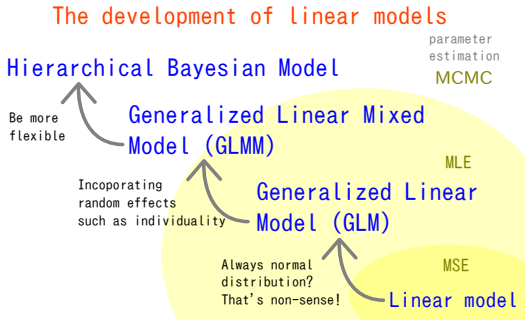
- 著者: 久保拓弥
- 出版社: 岩波書店
- 2012-05-18 刊行



kubostat2017e (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (e) 2019-08-05 3 / 46

もくじ

この授業であつかう統計モデルたち



kubostat2017e (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (e) 2019-08-05 4 / 46

もくじ

一般化線形モデルって何だろう?

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)
-

kubostat2017e (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (e) 2019-08-05 5 / 46

もくじ

一般化線形モデルを作る

一般化線形モデル (GLM)

- 確率分布は?
- 線形予測子は?
- リンク関数は?

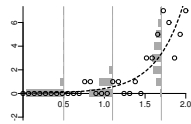
kubostat2017e (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (e) 2019-08-05 6 / 46

もくじ

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

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



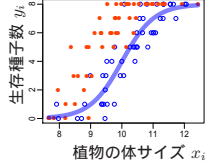
kubostat2017e (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (e) 2019-08-05 7 / 46

もくじ

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

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



kubostat2017e (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (e) 2019-08-05 8 / 46

"N 個のうち y 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 y 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

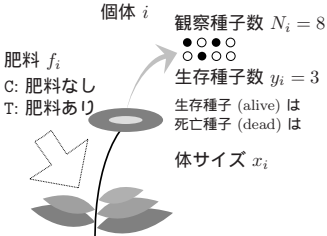
$$y_i \in \{0, 1, 2, \dots, 8\}$$

kubostat2017e (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (e) 2019-08-05 9 / 46

"N 個のうち y 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



kubostat2017e (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (e) 2019-08-05 10 / 46

"N 個のうち y 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

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

kubostat2017e (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (e) 2019-08-05 11 / 46

"N 個のうち y 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	

kubostat2017e (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (e) 2019-08-05 12 / 46

「N 個のうち y 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

今回は施肥処理 がきいている?

kubostat2017e (http://goo.gl/76c4i) 統計モデリング入門 2017 (e) 2019-08-05 13 / 46

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

2. ロジスティック回帰の部品

二項分布 binomial distribution と logit link function

kubostat2017e (http://goo.gl/76c4i) 統計モデリング入門 2017 (e) 2019-08-05 14 / 46

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

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 $p(y_i | 8, q)$

kubostat2017e (http://goo.gl/76c4i) 統計モデリング入門 2017 (e) 2019-08-05 15 / 46

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

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 (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

kubostat2017e (http://goo.gl/76c4i) 統計モデリング入門 2017 (e) 2019-08-05 16 / 46

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

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

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

kubostat2017e (http://goo.gl/76c4i) 統計モデリング入門 2017 (e) 2019-08-05 17 / 46

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

kubostat2017e (http://goo.gl/76c4i) 統計モデリング入門 2017 (e) 2019-08-05 18 / 46

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

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
    
```

kubostat2017e (http://goo.gl/76c4i) 統計モデリング入門 2017 (e) 2019-08-05 19 / 46

ロジスティック回帰の部品 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 ($f_i = T$)

kubostat2017e (http://goo.gl/76c4i) 統計モデリング入門 2017 (e) 2019-08-05 20 / 46

ちょっとだけ 交互作用項 について predictor

3. 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 complicate terms in linear predictor

ロジスティック回帰を例に

kubostat2017e (http://goo.gl/76c4i) 統計モデリング入門 2017 (e) 2019-08-05 21 / 46

ちょっとだけ 交互作用項 について predictor

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 ...

kubostat2017e (http://goo.gl/76c4i) 統計モデリング入門 2017 (e) 2019-08-05 22 / 46

何でも「割算」するな! 「脱」割算の offset 頂わざ

4. 何でも「割算」するな!

「脱」割算の offset 頂わざ

ポアソン回帰を強めてみる

kubostat2017e (http://goo.gl/76c4i) 統計モデリング入門 2017 (e) 2019-08-05 23 / 46

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?

- **data / data** 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

kubostat2017e (http://goo.gl/76c4i) 統計モデリング入門 2017 (e) 2019-08-05 24 / 46

何でも「割算」するな! 「脱」割算の offset 頂わざ

避けられるわりざん

- 避けられる割算値
 - 確率
 - 例: N 個のうち y 個にある事象が発生する確率
 - 対策: ロジスティック回帰など**二項分布モデル**で
 - 密度などの指数
 - 例: 人口密度, specific leaf area (SLA) など
 - 対策: **offset 頂わざ** — このあと解説!

kubostat2017e (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (e) 2019-08-05 25 / 46

何でも「割算」するな! 「脱」割算の offset 頂わざ

避けにくいわりざん

- 避けにくい割算値
 - 測定機器が内部で割算した値を出力する場合
 - 割算値で作図せざるをえない場合があるかも

kubostat2017e (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (e) 2019-08-05 26 / 46

何でも「割算」するな! 「脱」割算の offset 頂わざ

offset 項の例題: 調査区画内の個体密度

- 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



x 大
明るい



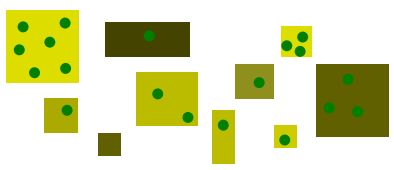
x 小
暗い

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

kubostat2017e (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (e) 2019-08-05 27 / 46

何でも「割算」するな! 「脱」割算の offset 頂わざ

「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

kubostat2017e (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (e) 2019-08-05 28 / 46

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

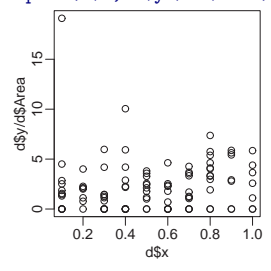
```

kubostat2017e (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (e) 2019-08-05 29 / 46

何でも「割算」するな! 「脱」割算の offset 頂わざ

明るさ vs 割算値 の図

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



いまいちよくわからない

kubostat2017e (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (e) 2019-08-05 30 / 46

何でも「割算」するな! 「脱」割算の offset 環わさ

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 が増大するようだ

kubostat2017c (http://goo.gl/76c4i) 統計モデリング入門 2017 (e) 2019-08-05 31 / 46

何でも「割算」するな! 「脱」割算の offset 環わさ

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?

kubostat2017c (http://goo.gl/76c4i) 統計モデリング入門 2017 (e) 2019-08-05 32 / 46

何でも「割算」するな! 「脱」割算の offset 環わさ

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 \times 密度
- 密度は明るさ x で変化する

kubostat2017c (http://goo.gl/76c4i) 統計モデリング入門 2017 (e) 2019-08-05 33 / 46

何でも「割算」するな! 「脱」割算の offset 環わさ

「平均個体数 = 面積 \times 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 が無い)

kubostat2017c (http://goo.gl/76c4i) 統計モデリング入門 2017 (e) 2019-08-05 34 / 46

何でも「割算」するな! 「脱」割算の offset 環わさ

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:

kubostat2017c (http://goo.gl/76c4i) 統計モデリング入門 2017 (e) 2019-08-05 35 / 46

何でも「割算」するな! 「脱」割算の offset 環わさ

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
 関数名: glm
 モデル式: y ~ x
 確率分布の指定: poisson
 リンク関数の指定 (省略可): link = "log"
 offset の指定: log(Area)

kubostat2017c (http://goo.gl/76c4i) 統計モデリング入門 2017 (e) 2019-08-05 36 / 46

何でも「割算」するな! 「脱」割算の offset 項わざ

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

kubostat2017e (http://goo.gl/76c4i) 統計モデリング入門 2017 (e) 2019-08-05 37 / 46

何でも「割算」するな! 「脱」割算の offset 項わざ

推定結果にもとづく予測を図にしてみる

• 実線は glm() の推定結果にもとづく予測
• 破線はデータ生成時に指定した関係

kubostat2017e (http://goo.gl/76c4i) 統計モデリング入門 2017 (e) 2019-08-05 38 / 46

何でも「割算」するな! 「脱」割算の offset 項わざ

まとめ: glm() の offset 項わざで「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

kubostat2017e (http://goo.gl/76c4i) 統計モデリング入門 2017 (e) 2019-08-05 39 / 46

何でも「割算」するな! 「脱」割算の offset 項わざ

統計モデルを工夫してわりざんやめよう

- 避けられる割算値
 - 確率
 - 例: N 個のうち y 個にある事象が発生する確率
 - 対策: ロジスティック回帰など**二項分布モデル**で
 - 密度などの指数
 - 例: 人口密度, specific leaf area (SLA) など
 - 対策: **offset 項わざ** — 統計モデリングの工夫!

kubostat2017e (http://goo.gl/76c4i) 統計モデリング入門 2017 (e) 2019-08-05 40 / 46

何でも「割算」するな! 「脱」割算の offset 項わざ

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)

kubostat2017e (http://goo.gl/76c4i) 統計モデリング入門 2017 (e) 2019-08-05 41 / 46

何でも「割算」するな! 「脱」割算の offset 項わざ

予習:

階層ベイズモデルで使う連続確率分布

A preview of continuous probability distributions to construct Hierarchical Bayesian Models

kubostat2017e (http://goo.gl/76c4i) 統計モデリング入門 2017 (e) 2019-08-05 42 / 46

何でも「割算」するな! 「脱」割算の offset 頂わざ

離散確率分布 ?

discrete probability distributions ?

連続確率分布 ?

continuous probability distributions ?

kubostat2017e (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (e) 2019-08-05 43 / 46

何でも「割算」するな! 「脱」割算の offset 頂わざ

離散確率分布 discrete probability distributions

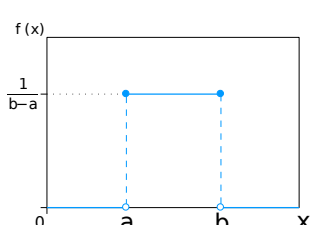
Poisson distribution Binomial distribution

kubostat2017e (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (e) 2019-08-05 44 / 46

何でも「割算」するな! 「脱」割算の offset 頂わざ

(連続) 一様分布 – 階層ベイズモデルの重要な部品

Uniform distribution (continuous) – an important “device” for HBM
parameter: min (a) and max (b)



The graph shows a coordinate system with the x-axis labeled 'x' and the y-axis labeled 'f(x)'. A horizontal blue line is drawn at the height of $\frac{1}{b-a}$ on the y-axis. This line is bounded by vertical dashed blue lines at $x=a$ and $x=b$ on the x-axis. The area under this line between $x=a$ and $x=b$ is shaded light blue, representing the uniform distribution's probability density function.

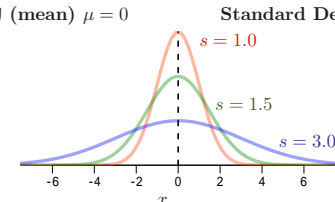
kubostat2017e (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (e) 2019-08-05 45 / 46

何でも「割算」するな! 「脱」割算の offset 頂わざ

正規分布あるいはガウス分布 – 階層ベイズモデルの重要な部品

the normal or Gaussian distribution – an important “device” for HBM
parameter: mean (μ) and SD ($s > 0$)

平均 (mean) $\mu = 0$ Standard Deviation (SD) s



The graph shows three bell-shaped curves centered at $x=0$ on the x-axis. The curves are colored red, green, and blue from left to right. The red curve is the tallest and narrowest, labeled $s = 1.0$. The green curve is shorter and wider, labeled $s = 1.5$. The blue curve is the shortest and widest, labeled $s = 3.0$. The x-axis is labeled with values from -6 to 6.

$$p(x | s) = \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{x^2}{2s^2}\right)$$

kubostat2017e (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (e) 2019-08-05 46 / 46