

統計モデリング入門

3章 一般化線形モデル (GLM)

3.1～3.3(p.40～p.45)

3 章の概要

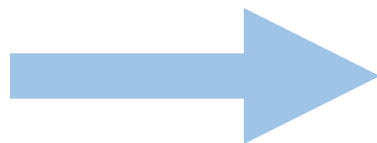
2 章：どの個体の種子数も平均 λ のポアソン分布にしたがい、平均種子数 λ は全個体で共通



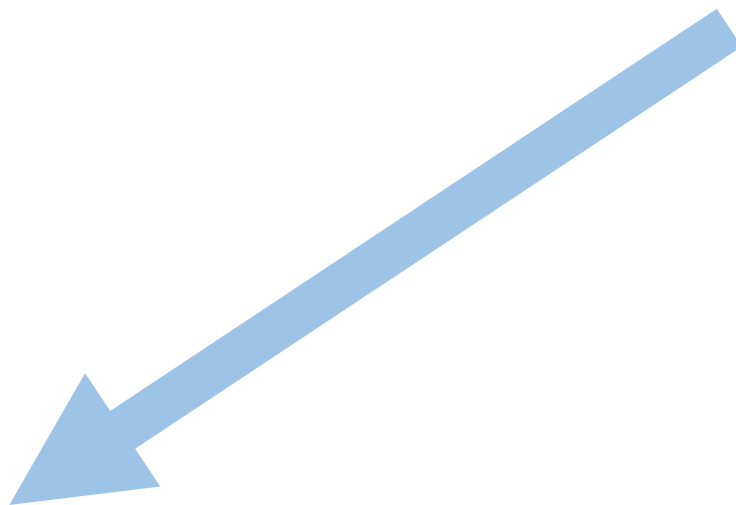
3 章：個体ごとに異なる説明変数によって平均種子数が変化する統計モデル

説明変数

- 個体のサイズ x_i
- 施肥処理 f_i



- 種子数 y_i



平均種子数 λ が変化

例題

- 架空植物100個体の種子数データを用意
- 植物個体 i
- 種子数 y_i
- 個体のサイズ x_i
- 施肥処理 f_i

50個体 ($i \in \{1, 2, \dots, 50\}$) は肥料なし (処理C)

残り 50個体 ($i \in \{51, 52, \dots, 100\}$) は施肥処理 (処理T)
(体サイズとは無関係)

データフレームとは

データフレーム=表 (table) のように扱えるデータ

データフレームの各列は一つの変数（項目）として認識され、それぞれの列の型（数値、文字、因子）は異なってもよい

データフレームの各行は、1つの観測値（人・個体・サンプル）として認識される

	性別	身長	体重
1	F	158	51
2	F	162	55
3	M	177	72
4	M	173	57
5	M	166	64

<https://goo.gl/0dxHJD>

異なるデータ型を一つの変数に格納できる

\$マークを使えば、「データフレーム名\$呼び出したい列」で列を取得できる

class関数とsapply関数

```
> class(d$y)
[1] "integer"
> class(d$x)
[1] "numeric"
> class(d$f)
[1] "factor"
```

```
> sapply(d, class)
      y      x      f
"integer" "numeric" "factor"
```

データ型(クラス)

- 整数 → integer型
- 実数 → numeric型
- 因子(文字列がある場合) → factor型

統計モデリングにおける大切なこと

- データを視覚的に捉える

