第 3 章 3.4～3.5

二又航介

2017 年 5 月 11 日

説明変数を組み込んだモデル

前回まで

平均種子数 λ が全個体で共通であると仮定

今回

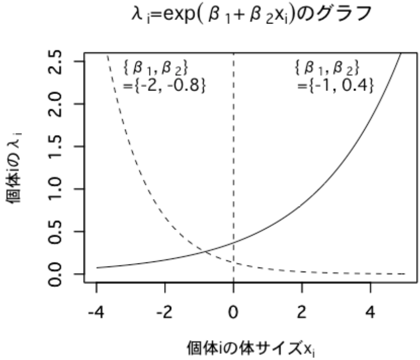
- 説明変数を組み込んだモデル
  - 個体ごとの平均種子数 λ_i を体サイズ x_i や施肥処理 f_i から推定

- ある個体 x_i において種子数が y_i である確率 $p(y_i|\lambda_i)$ はポアソン分布に従って
- $p(y_i|\lambda_i) = \frac{\lambda_{yi} \exp(-\lambda_i)}{y_i!}$

線形予測子とリンク関数

x_i による λ_i の関数

- 個体 x_i の差異によって種子数 λ_i を求める関数を定義
- ある個体 x_i の平均種子数 λ_i が
  - $\lambda_i = \exp(\beta_1 + \beta_2 x_i)$ として仮定



線形予測子、リンク関数

線形予測子

- $\log \lambda_i = \underline{\beta_1 + \beta_2 x_i}$
- 線形結合 (定数倍したパラメータ同士を和算したもの) で表される
- $\log \lambda_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2$ でも良い

リンク関数

- $\underline{(\lambda \text{ の関数})}$ = (線形予測子) の関係
- 対数リンク関数
- $\lambda_i = \exp(\text{線形予測子}) \geqq 0$

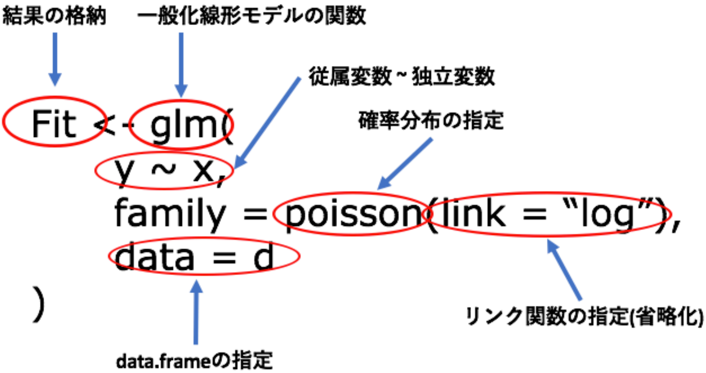
あてはめとあてはまりの良さ

ポアソン回帰

- 観測データに対するポアソン分布を用いた統計モデルのあてはめ
- 対数尤度 $\log L$ が最大となるパラメーター $\hat{\beta}_1$, $\hat{\beta}_2$ の推定値を決定

$$\log L(\beta_1, \beta_2) = \sum_i \log \frac{\lambda_i^{y_i} \exp(-\lambda_i)}{y_i!}$$

- λ_i が β_1 と β_2 の関数
- 線形予測子: $\log \lambda_i = \beta_1 + \beta_2 x_i$


結果の格納
一般化線形モデルの関数
従属変数 ~ 独立変数
確率分布の指定
Fit <- glm(
y ~ x,
family = poisson(link = “log”),
data = d
)
リンク関数の指定(省略化)
data.frameの指定

指定できる確率分布

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

確率分布 (family)	離散 or 連続	範囲
binomial	離散変数	$0 \sim +\infty$
gaussian	連続変数	$-\infty \sim +\infty$
Gamma	連続変数	$0 \sim +\infty$
inverse.gaussian	連続変数	$0 \sim +\infty$
poisson	離散変数	$0 \sim +\infty$
quasi	擬似尤度モデル	.

- 従属変数の分布を可視化して当てはまりそうなものを選択

指定できるリンク関数

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

リンク関数	名前	式
identity	恒等リンク	$\lambda = x$
log	対数リンク	$\log \lambda = x$
logit	ロジットリンク	$\log \frac{\lambda}{\lambda - 1} = x$
sqrt	平方根リンク	$\sqrt{\lambda} = x$
1/mu2̂	.	$\frac{1}{\lambda^2} = x$
inverse	逆数リンク	$\frac{1}{\lambda} = x$
power	べき乗リンク	$\lambda^n = x$

- 確率分布に線形予測子をうまく当てはめられそうなものを選択

確率分布とリンク関数の組み合わせ

Family	リンク関数
binomial	logit, probit, log, cloglog
gaussian	identity, log, inverse
Gamma	identity, inverse, log
inverse.gaussian	$\frac{1}{\mu^2}$, identity, inverse, log
poisson	identity, log, sqrt
quasi	logit, probit, cloglog, identity, inverse, log, $\frac{1}{\mu^2}$

結果の見かた 実行結果

Summary(fit)

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.29172	0.36369	3.552	0.000383	***
X	0.07566	0.03560	2.125	0.033580	*

```
Summary(fit)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.29172   0.36369  3.552   0.000383 ***
X           0.07566   0.03560  2.125   0.033580 *
```

$$\lambda_i = \exp(\underset{\text{切片}}{\beta_1} + \underset{\text{傾き}}{\beta_2} x_i)$$

- 最尤推定値
  - $\hat{\beta}_1 = 1.29$, $\hat{\beta}_2 = 0.0757$
  - $\log \lambda_i = 1.29 + 0.0757 x_i$

結果の見かた 標準誤差

Summary(fit)

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.29172	0.36369	3.552	0.000383	***
X	0.07566	0.03560	2.125	0.033580	*

Std. Error: 標準誤差

- 推定値 $\hat{\beta}_1, \hat{\beta}_2$ のばらつきを標準偏差で表したもの
- 推定値の精度についての指標となる

結果の見かた z 値

Summary(fit)

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.29172	0.36369	3.552	0.000383	***
X	0.07566	0.03560	2.125	0.033580	*

Z value: Z 値

- 最尤推定量を Std. Error で割った値 ($\frac{Estimate}{Std.Error}$)
- Wald 統計量
  - Wald 信頼区間を構成
  - 推定値が 0 から十分に離れているかの確認
  - 値が大きければ大きいほど離れている
- 0 から離れている $\simeq$ 信頼できる推定値

結果の見かた $Pr(>|z|)$

```
Summary(fit)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.29172   0.36369   3.552  0.000383 ***
X            0.07566   0.03560   2.125  0.033580 *
```

$Pr(>|z|)$: p 値

- 数値が大きいほど z 値が 0 に近い $\simeq$ 推定値が 0 に近い
- p 値として考えらるが、信頼区間として捉えるほうが良い
- 小さい値であるほど信頼区間が狭い $\simeq$ 推定値が信頼できる

結果の見かた 最大対数尤度

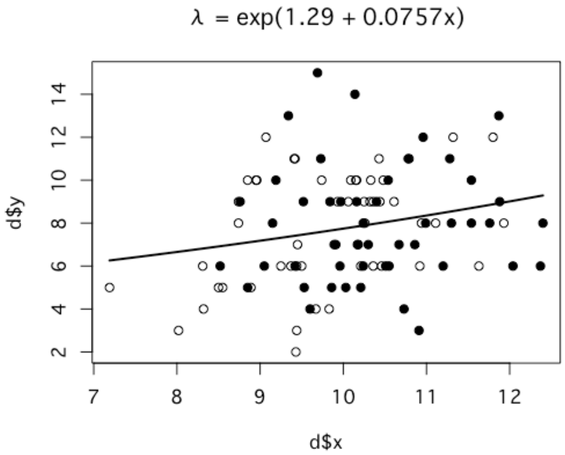
```
>logLik(fit)
'log Lik.' -235.3863 (df=2)
```

最大対数尤度

- $\log L(\beta_1, \beta_2)$ が最大
- モデルの当てはまりの良さの指標
- df:自由度
  - 最尤推定したパラメーターの個数
  - 今回は β_1 と β_2 の 2 個

モデルの可視化

```
> plot(d$x, d$y, pch = c(21, 19)[d$f])
> xx <- seq(min(d$x), max(d$x), length = 100)
> lines(xx, exp(1.29 + 0.0757 * xx), lwd = 2)
```



説明変数が因子型の統計モデル

施肥処理を説明変数に利用

- 因子型の説明変数はダミー変数で表される
- 種子数 y が施肥処理の有無 f に関係あると仮定
- $\lambda_i = \exp(\beta_1 + \beta_3 d_i)$ の関係

$$d_i = \begin{cases} 0 & f_i = \text{C の場合} \\ 1 & f_i = \text{T の場合} \end{cases}$$

- 個体 i が肥料なし ($f_i = C$) の場合
  - $\lambda_i = \exp(\beta_1)$
- 施肥処理した場合 ($f_i = T$)
  - $\lambda_i = \exp(\beta_1 + \beta_3)$

結果の見かた 実行結果

```
> fit.f <- glm(y ~ f, data=d, family = poisson)
> fit.f

Call:  glm(formula = y ~ f, family = poisson, data = d)

Coefficients:
(Intercept)           fT
    2.05156      0.01277
```

$$\lambda_i = \exp(\beta_1 + \beta_3 x_i)$$

切片　傾き

- 最尤推定値
  - $\hat{\beta}_1 = 2.05, \hat{\beta}_3 = 0.01277$
  - $\log \lambda_i = 2.05 + 0.0128 x_i$

結果の見かた, 実行結果

- $f_i = C$ の時,
  - $\lambda_i = \exp(2.05 + 0) = \exp(2.05) = 7.77$
- $f_i = T$ の時,
  - $\lambda_i = \exp(2.05 + 0.128) = \exp(2.0628) = 7.87$

「肥料をやると平均種子数が少しだけ増える」と予想

```
logLik(fit.f)
'log Lik.' -237.6273 (df=2)
```

- x_i だけのモデルの対数尤度-235.4 より当てはまりが悪い