環境統計学ぷらす

# 第4回
# 一般化線形モデル・モデル選択

高木　俊
shun.takagi@sci.toho-u.ac.jp

2013/11/14

# 予定

- 第1回：　Rの基礎と仮説検定
- 第2回：　分散分析と回帰
- 第3回：　一般線形モデル・交互作用
- 第4.1回：　一般化線形モデル
- 第4.2回：　モデル選択
- 第5回：　一般化線形混合モデル
- 第6回：　多変量解析（12/5予定）

# 今日やること（第4.1回）

- 統計編
  - 一般化線形モデル
- 表現編
  - 一般化線形モデルの結果

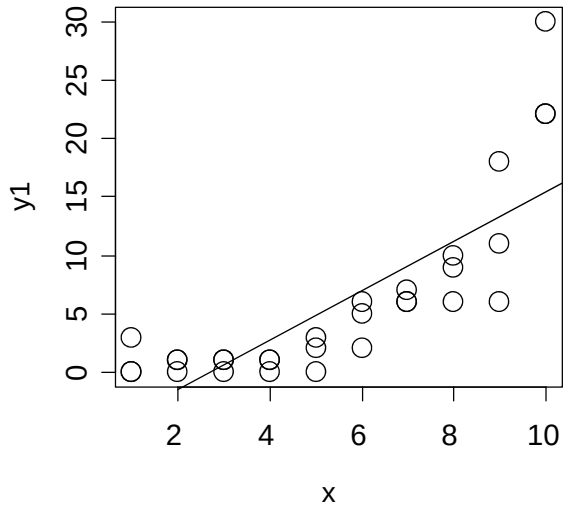
# 一般化線形モデル

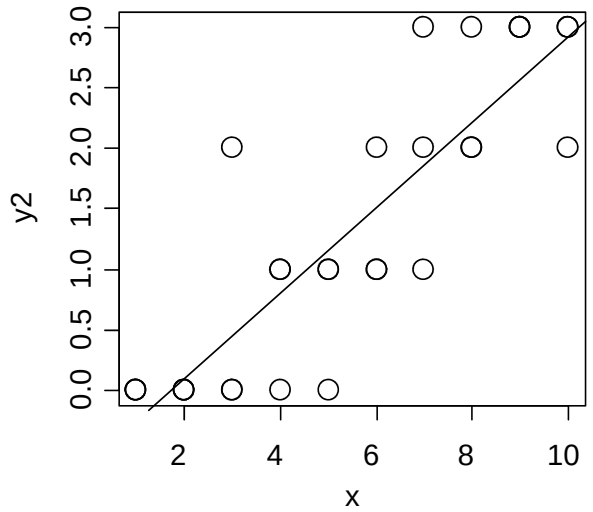
# 一般化線形モデルで扱うデータ

- 一般線形モデルでは、直線的関係＋等分散・正規誤差を仮定

だが、生物学で扱うデータはこれに当てはめられないものも多い

例） 個体数： 非負の整数値しかとらないデータ
個体の生死： 事象が起きるか起きないかの1/0データ




このような曲線的関係＋正規分布以外のデータを
一般化線形モデルで扱う（扱えない場合もある）

# おさらい：一般線形モデル

- 一般線形モデル

$$y_{ij} = \beta_0 + \sum \beta_i \times x_{ij} + \varepsilon_{ij}$$

観測値　線形予測式　誤差：
平均ゼロ、分散$\sigma_\varepsilon{}^2$の正規分布

書き換えると

$$\mu_{ij} = \beta_0 + \sum \beta_i \times x_{ij}$$

① 予測値$\mu_{ij}$は線形予測式で記述でき、

$$y_{ij} \sim \mathrm{Normal}(\mu_{ij}, \sigma_\varepsilon{}^2)$$

② 観測値$y_{ij}$は平均$\mu_{ij}$、分散$\sigma_\varepsilon{}^2$の正規分布に従う

③ パラメータ（$\beta_i$と$\sigma_\varepsilon{}^2$）を最小二乗法で推定

# 一般線形モデルと一般化線形モデル

- 一般線形モデルLM

$$\mu_{ij} = \beta_0 + \sum \beta_i \times x_{ij}$$

$$y_{ij} \sim \mathrm{Normal}(\mu_{ij}, {\sigma_\varepsilon}^2)$$

① 予測値$\mu_{ij}$は線形予測式で記述でき、

② 観測値$y_{ij}$は平均$\mu_{ij}$、分散${\sigma_\varepsilon}^2$の正規分布に従う

③ パラメータ($\beta_i$と${\sigma_\varepsilon}^2$)を最小二乗法で推定

- 一般化線形モデルGLM(ポアソン分布仮定の例)

$$\log \lambda_{ij} = \beta_0 + \sum \beta_i \times x_{ij}$$

$$y_{ij} \sim \mathrm{Poisson}(\lambda_{ij})$$

① 予測値$\lambda_{ij}$は線形予測式とリンク関数で記述でき、

② 観測値$y_{ij}$は平均$\lambda_{ij}$のポアソン分布に従う

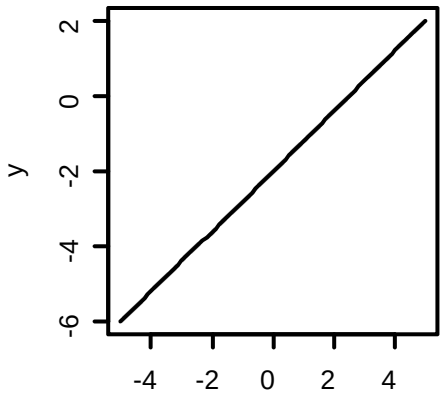
③ パラメータを最尤法で推定

# GLMの基本①:リンク関数

- リンク関数を指定することで、曲線的関係を扱うことができる

## LM

リンク関数なし(Identity)



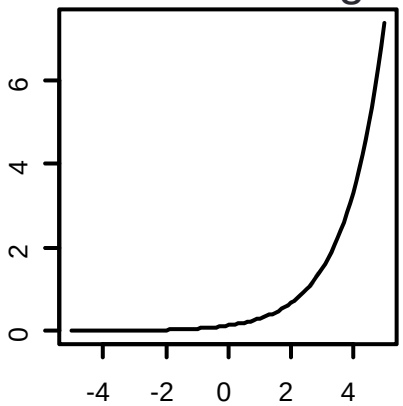

$$y = -2 + 0.8x$$

$$-\infty < y < +\infty$$

## GLM

$$\mathrm{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

対数リンク(log)

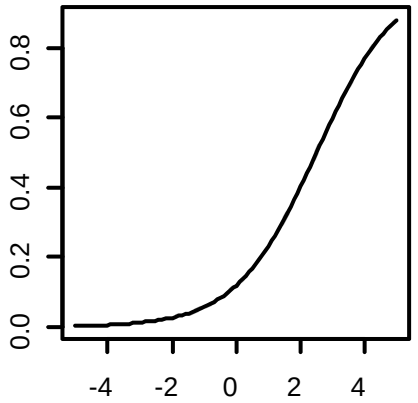



$$\log(y) = -2 + 0.8x$$
$$\Leftrightarrow \quad y = \exp(-2 + 0.8x)$$

$$0 < y < +\infty$$

個体数など

logitリンク(logit)




$$\mathrm{logit}(y) = -2 + 0.8x$$
$$\Leftrightarrow y = \frac{\exp(-2 + 0.8x)}{(1 + \exp(-2 + 0.8x))}$$

$$0 < y < 1$$

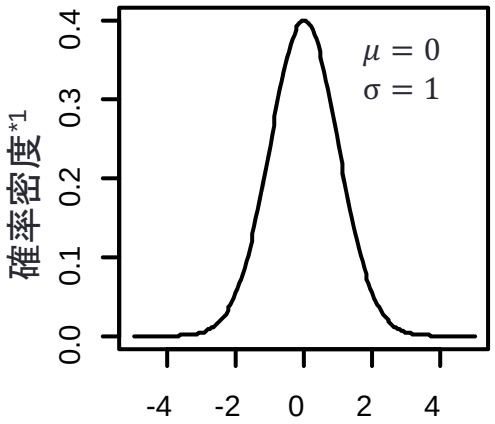
死亡率など

# GLMの基本②:確率分布

- 正規分布以外の確率分布を扱える

## LM

正規分布

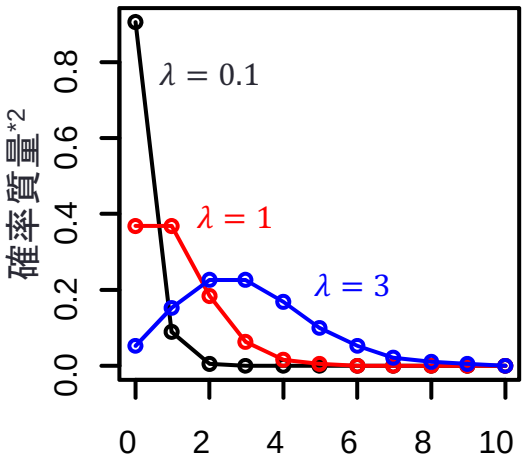



左右対称
平均と分散で形が決まる

## GLM

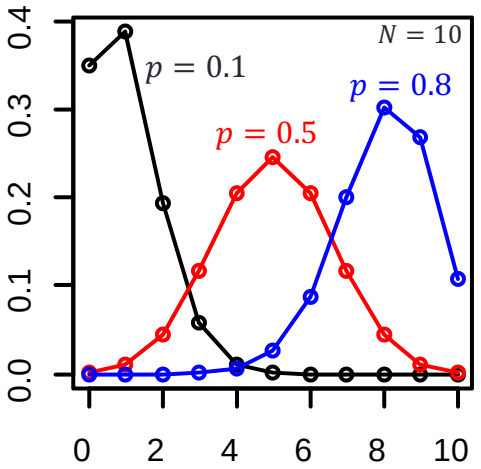
$N = 1$の時はベルヌーイ分布

ポアソン分布




大きい方に裾が広い
平均＝分散

二項分布




p＝0.5の時左右対称

*1 xがある区間の値をとる確率はその区間の確率密度の積分で表される
*2 xがある値（離散量）をとる確率はその値の確率質量で表される

# GLMの基本③:最尤法

- 最尤法によるパラメータ推定

最尤法:最も尤度が高くなるパラメータを推定する方法

あるパラメータ$\theta$のもとでデータ$D$が得られる確率 $P(D|\theta)$

例:コイントス3回で表2回・裏1回のデータが得られた。このコインの表が出る確率pは？

$p = 0.1$の時、表2裏1のデータ得られる確率は

$$P(D|p = 0.1) = 0.1^2 \times (1 - 0.1)^1 = 0.009$$

同様に、$p = 0.2, 0.3, \cdots$の時

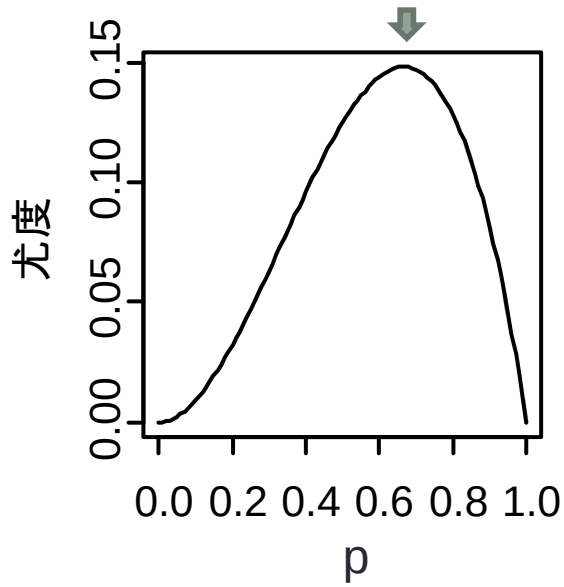
$$P(D|p = 0.2) = 0.2^2 \times (1 - 0.2)^1 = 0.032$$

$$P(D|p = 0.3) = 0.3^2 \times (1 - 0.3)^1 = 0.063$$

$$\cdots$$

尤度のグラフは右図のようになり、
p=0.067の時、表2裏1のデータが得られる確率が最も高い

最尤推定値: $p = 0.067$

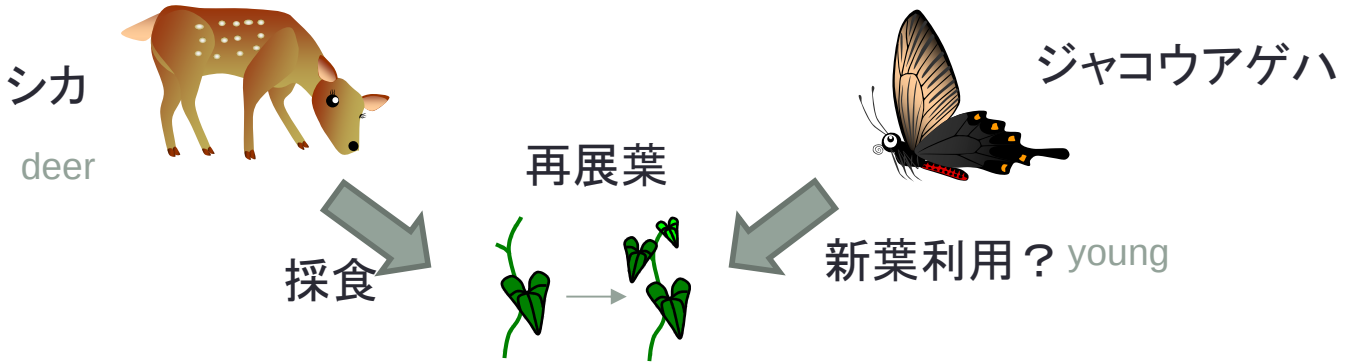


# 実例:二項分布を仮定したGLM

- 使い所:　生死・繁殖の有無・雌雄の別など1/0データ
- リンク関数は通常logitを用いる

$$\mathrm{logit}(p) = \log(\frac{p}{1-p})$$

例:　不定芽の発芽率は葉の大きさと関係があるか？
種子の発芽率は温度と関係があるか？
ジャコウアゲハの新葉利用率はシカ密度と関係があるか？



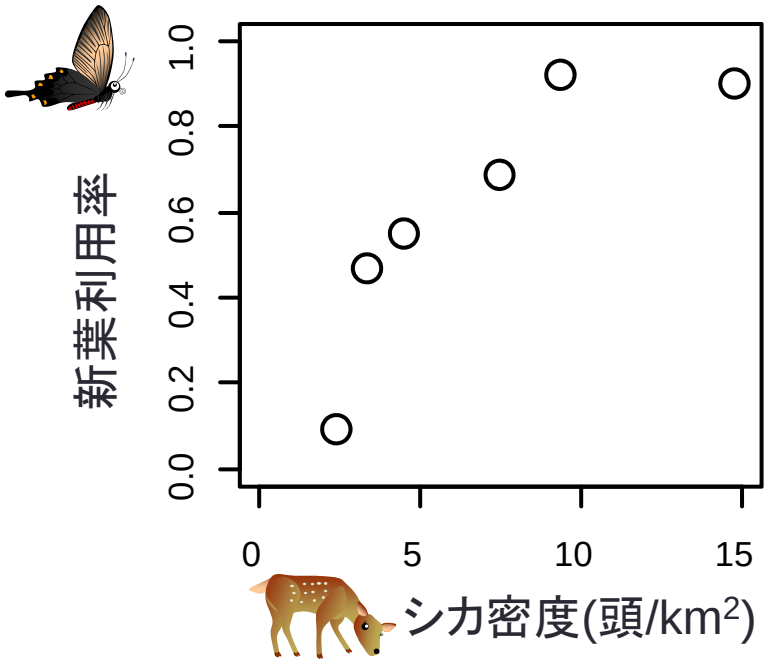

data4.1.csv

# まずはplot

- 縦軸を （新葉の利用数)/(新葉＋古葉の利用数） でプロット

```
plot(I(young/(young+old))~deer,data4.1,cex=2,xlim=c(0,15),ylim=c(0,1))
```

モデル式の中で数値を計算する場合はI()で表記




データは頭打ち, 0~1しか取り得ない。
直線回帰は不適。

↓

二項分布を仮定した一般化線形モデル
(ロジスティック回帰)で解析

# モデル式の構造とglm()による推定

一試行あたり

$$\text{logit}(\text{事象}A\text{が起きる確率}) = \text{切片} + \sum(\text{傾き} \times \text{説明変数})$$

$$\text{事象}A\text{が起きた回数} \sim \text{Binomial}(\text{事象}A\text{が起きる確率}, \text{Total試行回数})$$

Rでの記述

```
glm(cbind(事象A生じた回数,生じなかった回数)~説明変数,
                                    データ名, family="binomial")
```

この例では、

$$\text{logit}(p_i) = \beta_0 + \beta_1 \times \text{シカ密度}_i$$

$$\text{新葉への産卵数}_i \sim \text{Binomial}(p_i, \text{Total産卵数}_i)$$

```
model4.1<-
    glm(cbind(young,old)~deer,data4.1,family="binomial")
```

# 推定結果

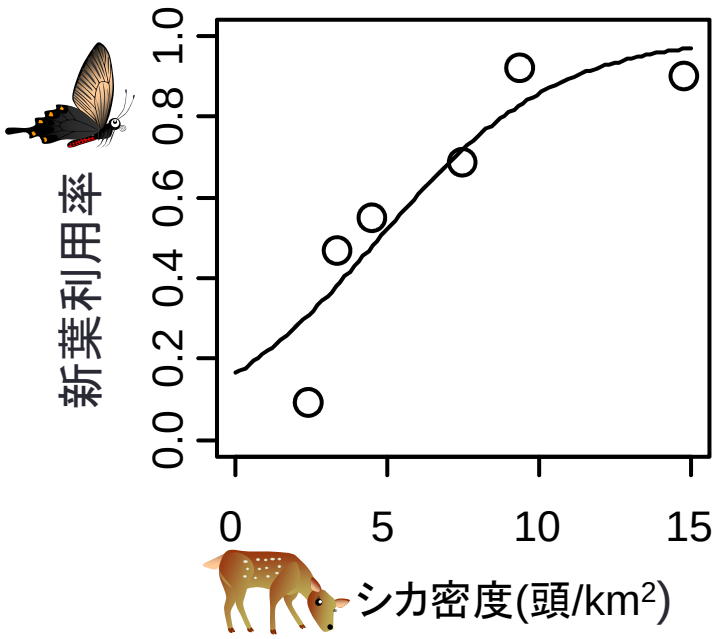
```
> model4.1<- glm(cbind(young,old)~deer,data4.1,family="binomial")
> summary(model4.1)

Call:
glm(formula = cbind(young, old) ~ deer, family = "binomial",
    data = data4.1)

Deviance Residuals:
      1       2       3       4       5       6
-1.0128  0.8555 -0.2908  0.6317  0.6268 -1.7459

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.62641    0.56678   -2.87  0.00411 **
deer         0.34311    0.09692    3.54  0.00040 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25.1667  on 5  degrees of freedom
Residual deviance:  5.6822  on 4  degrees of freedom
AIC: 25.161

Number of Fisher Scoring iterations: 5
```

係数表

モデルの当てはまりに関して（後述）

# 回帰曲線を引く

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.62641    0.56678   -2.87  0.00411 **
deer         0.34311    0.09692    3.54  0.00040 ***
```




$\mathrm{logit}(p) = \log\left(\frac{p}{1-p}\right) = \mathrm{A}$ とおくと、

$p = \frac{\exp(A)}{1+\exp(A)} = \frac{1}{1+\exp(-A)}$ なので

$$\text{新葉利用率} = \frac{1}{1+\exp(1.63 - 0.34 \times \text{シカ密度})}$$

```
curve(1/(1+exp(1.62641-0.34311*x)),add=T)
```

# モデルの当てはまりの評価

```
    Null deviance: 25.1667  on 5  degrees of freedom
Residual deviance:  5.6822  on 4  degrees of freedom
```

$Deviance$(逸脱度) $= -2 \times$ (モデルの対数尤度 $-$ フルモデルの対数尤度)

当てはまりの悪さ(説明できなかったばらつきみたいなもの)と考えてください

Null deviance: 説明変数なし(nullモデル)の逸脱度
Residual deviance: モデルの逸脱度

正規分布の時は
Null deviance＝$SS_{Total}$ (全体のばらつき)
Residual deviance＝$SS_{Residual}$ (モデルで説明できなかったばらつき)
となる

# GLMの検定：尤度比検定

- 一般線形モデルでは平均平方（≒分散）の比であるF比を用いて検定（分散分析）
- 一般化線形モデルではモデルの$-2\times$対数尤度の比（=逸脱度の差）を用いて検定（尤度比検定・逸脱度分析）

見たい説明変数を抜いたモデルの逸脱度－
見たい説明変数入りのモデルの逸脱度

← 説明変数の自由度の$\chi^2$分布で近似できる

*近似を使わない方法として、パラメトリックブートストラップ法による検定がある

Rでの記述

```
model1<- glm(y~x)                    #xの効果を見たい
model0<- glm(y~1)                    #右辺の1は切片のみのモデル（説明変数なし）
anova(model0,model1,test="Chi")
                #anova関数を使うが、やっているのは分散分析ではなく尤度比検定
```

# 尤度比検定の実行

```
> model4.1.0<- glm(cbind(young,old)~1,data4.1,family="binomial")   #nullモデル作る
> anova(model4.1.0,model4.1,test="Chi")

Analysis of Deviance Table          #Analysis of Varianceでないことに注意！

Model 1: cbind(young, old) ~ 1
Model 2: cbind(young, old) ~ deer
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1         5    25.1667
2         4     5.6822  1   19.485 1.014e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

#Model間でのDevianceの差がモデルで説明できたばらつき（みたいなもの）
#これが十分に大きければ、有意な効果あり

ジャコウアゲハの新葉利用率はシカ密度が高い場所で高い傾向を示した（$\chi_1^2 = 19.5, P < 0.001$）。

# 二項分布仮定のGLMが使えるデータ

| 個体id | survive | X |
|---|---|---|
| 1 | 0 | 14.8 |
| 2 | 1 | 9.4 |
| 3 | 0 | 7.5 |
| 4 | 1 | 4.5 |
| 5 | 1 | 3.4 |

事象が起きたかどうか1/0で入力されたデータ

```
glm(survive~X,
        data,family=”binomial”)
```

| 場所id | survive | death | X |
|---|---|---|---|
| 1 | 9 | 1 | 14.8 |
| 2 | 11 | 1 | 9.4 |
| 3 | 11 | 5 | 7.5 |
| 4 | 11 | 9 | 4.5 |
| 5 | 7 | 8 | 3.4 |

事象が起きた回数と起きなかった回数で入力されたデータ

```
glm(cbind(survive,death)~X,
        data,family=”binomial”)
```

# 実例:ポアソン分布を仮定したGLM

- 使い所:　生物の個体数など非負の整数値データ
- リンク関数には通常logを用いる

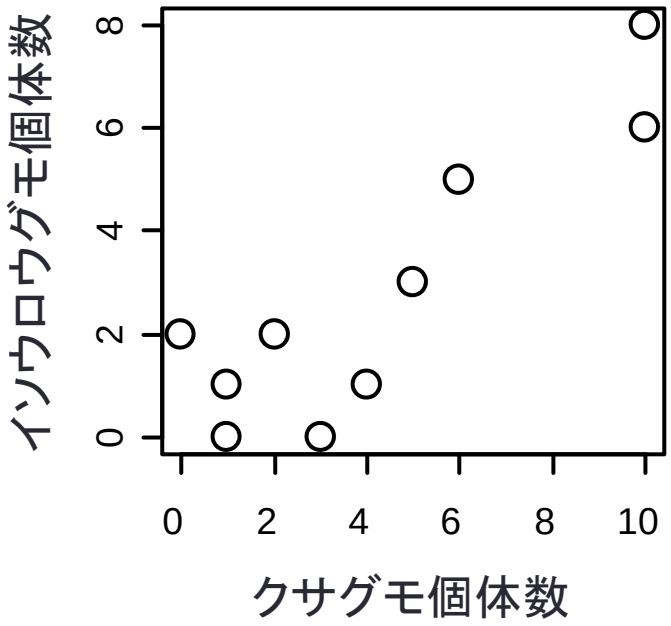
例:　個体密度は河川からの距離と関係があるか？
　　捕獲個体数は溶存酸素量と関係があるか？
　　クサグモとイソウロウグモの個体数は餌量や営巣場所に制限されているか？


data4.2.csv

# まずはプロット

イソウロウグモの個体数と宿主であるクサグモの個体数の関係をプロット

```
plot(isourou~kusa,data4.2,cex=2)
```




データは非負の整数値。
（そうでもないが）非直線的増加。

↓

**ポアソン分布**を仮定した一般化線形モデル（ポアソン回帰）で解析

# モデル式の構造とglm()による推定

$$\log(\text{期待値}) = \text{切片} + \sum(\text{傾き} \times \text{説明変数})$$

$$\text{観察値} \sim \text{Poisson}(\text{期待値})$$

Rでの記述

```
glm(観察値~説明変数, データ名, family="poisson")
```

この例では、

$$\log(\lambda_i) = \beta_0 + \beta_1 \times \text{クサグモ個体数}_i$$

$$\text{イソウロウ個体数}_i \sim \text{Poisson}(\lambda_i)$$

```
model4.2<- glm(isourou~kusa,data4.2,family="poisson")
```

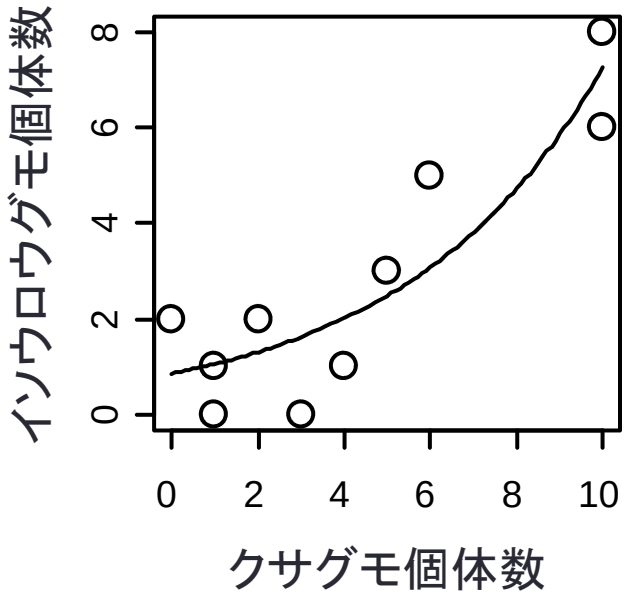
# 推定結果

```
> model4.2<- glm(isourou~kusa,data4.2,family="poisson")
> summary(model4.2)

Call:
glm(formula = isourou ~ kusa, family = "poisson", data = data4.2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7988  -0.7120   0.1063   0.5016   1.0603

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.16318    0.42168  -0.387    0.699
kusa         0.21476    0.05441   3.947 7.91e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 25.345  on 9  degrees of freedom
Residual deviance:  8.816  on 8  degrees of freedom
AIC: 36.111

Number of Fisher Scoring iterations: 5
```

係数表

# 回帰曲線

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.16318    0.42168  -0.387    0.699
kusa         0.21476    0.05441   3.947 7.91e-05 ***
```




$\log(\lambda) = \mathrm{A}$ とおくと、$\lambda = \exp(\mathrm{A})$ なので

イソウロウ個体数＝

$\exp(-0.16 + 0.21 \times$ クサグモ個体数$)$

```
curve(exp(-0.16318+0.21476*x),add=T)
```

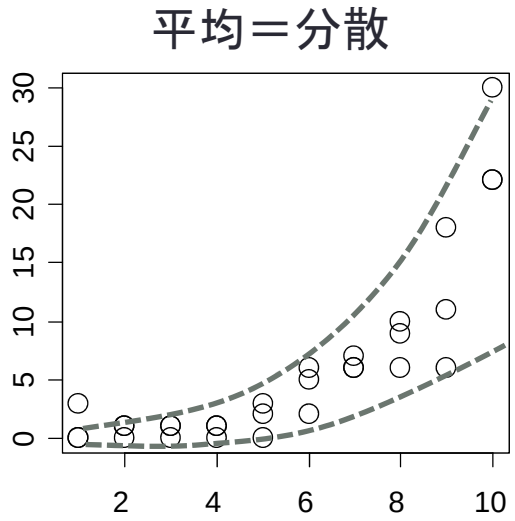
# 尤度比検定

- 尤度比検定は二項分布の時と同様

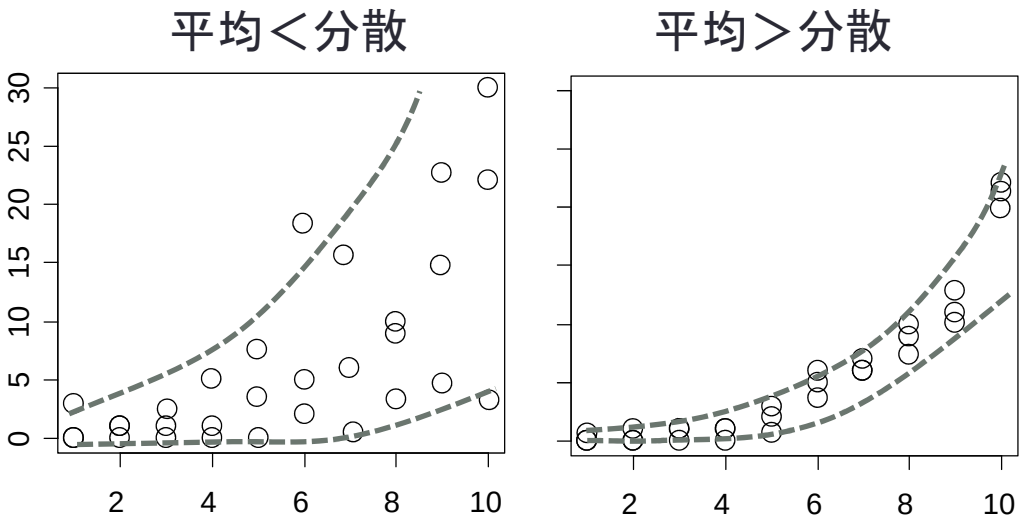
```
> model4.2.0<- glm(isourou~1,data4.2,family="poisson")
> anova(model4.2.0,model4.2,test="Chi")
Analysis of Deviance Table

Model 1: isourou ~ 1
Model 2: isourou ~ kusa
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1         9     25.345
2         8      8.816  1   16.529 4.792e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0
```

クサグモ個体数が多い場所ほどイソウロウグモ個体数が多くなる傾向が見られた（$\chi_1^2 = 16.5, P < 0.001$）。

# ポアソン分布は平均＝分散




ポアソン分布に当てはめられる




ポアソン分布に当てはめられない

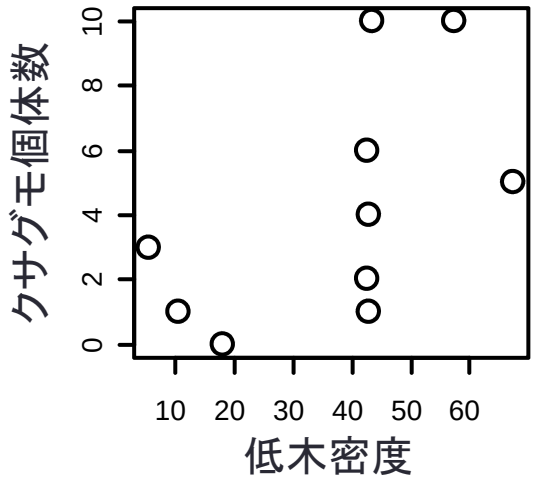
- 非負の整数値であっても、平均＝分散から大きく外れるデータ（過分散 overdispersion）はポアソン分布を仮定できない

→この場合、負の二項分布を用いる

# 過分散の判定

- Residual devianceとその自由度の比で判別

```
> plot(kusa~tree,cex=2,data4.2)
> pomodel4.2<- glm(kusa~tree,data4.2,family="poisson")
> summary(pomodel4.2)

Call:
glm(formula = kusa ~ tree, family = "poisson", data = data4.2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1015  -1.2577  -0.4083   0.9516   2.2887

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.306686   0.454876   0.674  0.50017
tree        0.026971   0.009314   2.896  0.00378 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 29.606  on 9  degrees of freedom
Residual deviance: 20.309  on 8  degrees of freedom
AIC: 52.633

Number of Fisher Scoring iterations: 5
```



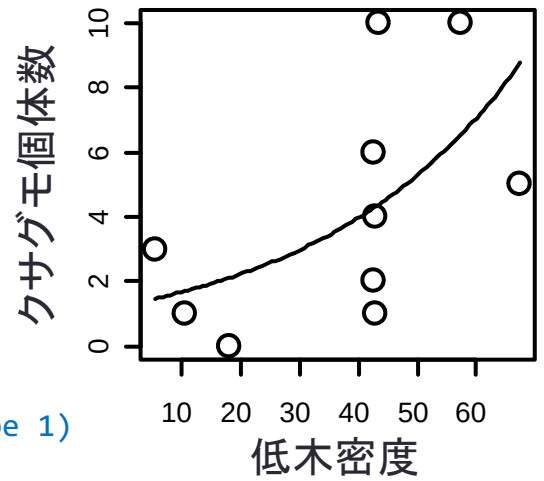

ポアソン分布で、この比は1になる
2を超えると過分散と言われる

20.3/8=2.54 →過分散

# 負の二項分布を仮定したGLM

- 推定にはパッケージMASSのglm.nb関数を用いる

```
> library(MASS)
> nbmodel4.2<- glm.nb(kusa~tree,data4.2)
> summary(nbmodel4.2)

Call:
glm.nb(formula = kusa ~ tree, data = data4.2, init.theta = 4.186874329,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8494  -0.8941  -0.3091   0.6884   1.4756

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.23013    0.59629   0.386   0.6995
tree         0.02875    0.01299   2.213   0.0269 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(4.1869) family taken to be 1)

    Null deviance: 16.308  on 9  degrees of freedom
Residual deviance: 11.281  on 8  degrees of freedom
AIC: 51.915
（省略）
```




```
curve(exp(0.23013+0.02875*x),add=T)
```

# 個体数でなく密度との関係を見たい場合

| 個体数 | 調査面積 | 密度 | X |
|---|---|---|---|
| 9 | 1.5 | 6.00 | 14.8 |
| 11 | 1.2 | 9.17 | 9.4 |
| 11 | 2.4 | 4.58 | 7.5 |
| 11 | 1.9 | 5.79 | 4.5 |
| 7 | 1.3 | 5.38 | 3.4 |

個体数は非負の整数値だが、密度にしたら小数になってしまう・・・

↓

ポアソン分布で扱えない？

説明変数に調査面積を入れる手もあるが、

$$\log(\text{個体数}) = \beta_0 + \beta_1 \times X + \beta_2 \times \text{面積}$$ ではなく、

$$\log\left(\frac{\text{個体数}}{\text{面積}}\right) = \beta_0 + \beta_1 \times X$$ の関係を知りたい

# offset項の活用

- offset項を用いるとその変数について傾きは推定されない

glm(Y~X1+X2,family="poisson") の場合、

$$\log(Y) = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2$$

glm(Y~X1+offset(X2),family="poisson") の場合、

$$\log(Y) = \beta_0 + \beta_1 \times X_1 + X_2$$

$$\log\left(\frac{\text{個体数}}{\text{面積}}\right) = \beta_0 + \beta_1 \times X$$

$\log\left(\frac{B}{A}\right) = \log(B) - \log(A)$より

$$\Leftrightarrow \log(\text{個体数}) = \beta_0 + \beta_1 \times X + \log(\text{面積})$$

なので

glm(個体数~X+offset(log(面積)),family="poisson") #logを忘れずに！
で密度との関係が推定できる

# まとめ：一般化線形モデル

- リンク関数と分布の指定で、一般線形モデルで扱いにくいカウントデータや比率のデータをうまく扱うことができる
- 最尤法による推定を行い、尤度（逸脱度）をもとにした検定を行う
- 一般線形モデルと同様、連続変数・カテゴリカル変数・二次項・交互作用等扱える

# 方法での表現

- GLMを用いる際には、誤差構造（分布）とリンク関数を明記すること。

- Yを目的変数、X1・X2を説明変数とする一般化線形モデルによる解析を行った。目的変数はポアソン分布に従うと仮定し、リンク関数はlogとした。
- 個体の生存または死亡を目的変数とした。リンク関数をlogitとし、誤差構造に二項分布を仮定した一般化線形モデルにより解析を行った。
- 要旨等でスペースがない場合は「一般化線形モデル（誤差構造：ポアソン、リンク関数：log）で解析した。」

# 結果での表現

- 尤度比検定（逸脱度検定）の結果は文章中に表記するか「～な傾向を示した（$\chi_1^2 = 19.5, P < 0.001$）」、表にまとめる

各説明変数に対する尤度比検定の結果

| 説明変数 | $\chi^2$ | P値 |
|---|---|---|
| 低木 | 4.116 | 0.04 |
| 餌昆虫 | 0.322 | 0.57 |
| 低木×餌昆虫 | 0.147 | 0.70 |

分散分析表とちがって、形式は決まってない（と思う）ので見やすいようにまとめればOK