チュートリアル：
Rを使った線形モデル入門

日高　昇平

北陸先端科学技術大学

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY
1990

日本認知科学会・対話と学習研究会・第46
回研究会 2012年3月22日

確率的モデル

- 確率的モデル～データをある確率分布からの標本とみなす
  - 確率分布
  - サンプリングに関する仮定(独立性など)
  - リンク関数(今回は主に線形関数)
- モデルの推定
  - 最尤推定

線形モデルの階層

定数項モデル
(確率分布)

(一般)線形モデル
(線形関数＋正規誤差)

一般化線形モデル
(非線形リンク関数+非正規誤差)

(一般)線形モデル (general) linear model

(※一般*化*線形モデル(Generalized linear model)とは異なる)

- モデル ~ 線形予測子＋正規誤差

- モデルの背景にある仮定 (中心極限定理)
  - 独立・同一分布(independent identical distribution)からの抽出
  - 線形性(linearity)
  - 正規性(normality)
  - 大標本または漸近性(asymptotic)

確率分布

確率変数 x が 平均μ, 標準偏差σの正規分布に従う。

確率分布(密度)関数

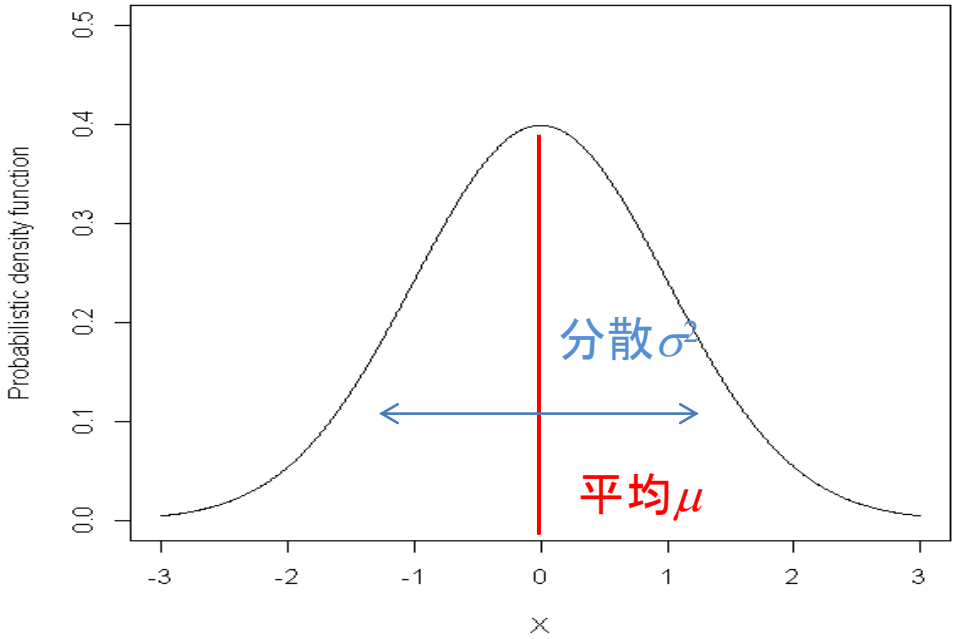
$$p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$$

確率変数　パラメータ(平均・分散)　正規化定数　指数関数

Xの定義域における、確率(密度)の和(積分) は1 (正規化)

正規分布

$$p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$$




```
##############################################################
## Plot probabilistic density function of standard normal distribution
##############################################################
X = seq( -3, 3, 0.01 );
PDF = dnorm( X, mean = 0, sd = 1 ); # pdf of normal distribution with mean 0 and sd 1.
plot( X, PDF, "l", xlim = c( -3, 3 ), ylim = c( 0, 0.5 ),
xlab = "X", ylab = "Probabilistic density function" );
```

線形モデル：線形性

- 以下の2つの性質を満たす関数fを線形関数という
  - $f(x+y) = f(x) + f(y)$　　(加法性)
  - $f(cx) = cf(x)$　　(斉次性)

    - **例えば,**
      - $f(1, x) = (2x + 4) = (2x) + (4)$
      - $f(1, x, y, z) = 2x + 3y + 4z + 5$
    - **以下は線形ではない(非線形)**
      - 2次関数: $g(x, y) = (x+ y)^2 \neq x^2 + y^2$
      - 指数関数: $g(x, y) = e^{x+y} \neq e^x + e^y$
    - **しかし、以下は(x, x^2またはx^3の) "線形"関数**
      - $f(1, x^2) = 2x^2 + 4 = 2x^2 + 4$
      - $f(1, x, x^2, e^x) = 2x + 3x^2 + 4\,e^x + 5$

線形関数の有用性

- 線形性を満たす関数は、計算が容易
  - なぜか?
    - 全体を部分に分解し、個別・独立に計算できるため
  - なので
    - 線形性の仮定は正規性の仮定(同じく独立性を要する)と相性がよい

$$f(x+y) = f(x) + f(y)$$
$$f(cx) = cf(x)$$

定数項モデル

- モデル: $y = c + \varepsilon$
- ある変数yがある平均cをもち、その誤差εが正規分布$N(0, \sigma^2)$に従うとするモデル
  - ある正規分布$N(c, \sigma^2)$のパラメータを推定するのと同じ

$$X = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

デザイン行列

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$$

応答ベクトル

$$A = c$$

パラメタ

$$L = \sum_{i=1}^{N} \{y_i - (c)\}^2$$

負の対数尤度関数 (deviance)

単回帰モデル

- モデル: $y = ax + c + \varepsilon$
- ある変数yがある平均($ax + c$)をもち、その誤差εが正規分布N(0, σ^2)に従うとするモデル
  - ある正規分布N($ax + c$, σ^2)のパラメータを推定するのと同じ

$$X = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_N & 1 \end{pmatrix} \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \quad A = \begin{pmatrix} a \\ c \end{pmatrix} \quad L = \sum_{i=1}^{N} \{y_i - (ax_i + c)\}^2$$

L: 負の尤度関数(deviance)

重回帰モデル

- モデル: $y = a_1x_1+a_1x_2+\ldots+a_Mx_M + c + \varepsilon$
- ある変数yがある平均($\Sigma a_i x_i + c$)をもち、その誤差εが正規分布$N(0, \sigma^2)$に従うとするモデル
  - ある正規分布$N(\Sigma a_i x_i + c, \sigma^2)$のパラメータを推定するのと同じ

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1M} & 1 \\ x_{21} & \cdots & x_{2M} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ x_{N1} & \cdots & x_{NM} & 1 \end{pmatrix} \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \quad A = \begin{pmatrix} a_1 \\ \vdots \\ a_M \\ c \end{pmatrix} \quad L = \sum_{i=1}^{N} \left\{ y_i - \left(\sum_{j=1}^{M} a_j x_{ij} + c \right) \right\}^2$$

L: 負の尤度関数(スカラー倍)

分散分析モデル(一元配置)

- モデル: $y = a_1x_1+(a_2-a_1)x_2+ \varepsilon$
- ある変数yがある平均($a_1x_1+(a_2-a_1)x_2$)をもち、その誤差εが正規分布N(0, σ^2)に従うとするモデル(x_1, x_2は群所属を表すダミー変数)
  - 多くの場合パラメタ$(a_2-a_1)\neq 0$を検定する。

$$X=\begin{pmatrix}1 & 0\\ 1 & 0\\ 1 & 1\\ 1 & 1\end{pmatrix} \quad Y=\begin{pmatrix}y_1\\ y_2\\ \vdots\\ y_N\end{pmatrix} \quad A=\begin{pmatrix}a_1\\ a_2\end{pmatrix} \quad L=\sum_{i=1}^{N}\left\{y_i-\left(a_1x_{i1}+\left(a_2-a_1\right)x_{i2}\right)\right\}^2$$

線形モデルの推定:最尤推定(最小2乗法)

Maximum likelihood estimator (Least square method)

- 線形モデル: $Y = XA + \varepsilon$

対数尤度関数

$$\varepsilon_i \approx N\left(0, \sigma^2\right) \propto \exp\left(-\frac{\varepsilon_i^2}{2\sigma^2}\right)$$

$$E = \log \prod_{i=1}^{N} N\left(\varepsilon_i^2 \mid 0, \sigma^2\right)$$

$$= \sum_{i=1}^{N} \frac{\left\{y_i - \left(ax_i + c\right)\right\}^2}{-2\sigma^2}$$

解(ベクトルによる微分・擬似逆行列)

$$E = \frac{1}{2}\left|Y - XA\right|^2$$

$$\frac{\partial E}{\partial A} = Y - XA = 0$$

$$A = \left(X^T X\right)^{-1} X^T Y$$

$$X = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_N & 1 \end{pmatrix} \qquad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$$

$$A = \begin{pmatrix} a \\ c \end{pmatrix}$$

尤度(likelihood)とは:
P(X|θ)～パラメタθが与えられたときのデータXの条件つき確率(Xの関数)
L(θ|X)～パラメタがデータXを生成する尤もらしさ(Θの関数)

```
# In script “LinearModel.R”
##########################################
## Estimation of the model's parameter (Atrue)
##########################################
XX <- solve( t( X ) %*% X ); # solve the Y ~ X %*% A
Aestimated <- XX %*% t( X ) %*% Y; #
```

R演習1: 単回帰モデルの解

LinearModel.R

- 用意されているスクリプトを利用して、(Rパッケージ関数を使わずに)単回帰モデルを解き、Rパッケージを使って解いた解と一致している事を確認せよ
  - 使用関数: lm, solve

```
# In script “LinearModel.R”
##############################################
## Estimation of the model's parameter (Atrue)
##############################################
XX <- solve( t( X ) %*% X ); # solve the Y ~ X %*% A
Aestimated <- XX %*% t( X ) %*% Y; #
```

線形方程式を解いた結果

```
> Atrue
[1] 2.0 0.5
> t( Aestimated )
          [,1]      [,2]
[1,] 1.976269 0.5364442
## Using the linear fitting function in R
> Results = lm( Y ~ 0 + X )
> coefficients( Results )
       X1        X2
1.9762689 0.5364442
```

“lm”関数の結果

結果の一例

```
> Results = lm( Y ~ 0+X )
> summary( Results )

Call:
lm(formula = Y ~ 0 + X)

Residuals:
     Min       1Q   Median       3Q      Max
-0.104843 -0.030977  0.000826  0.033624  0.082713

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
X1  2.00785    0.01433  140.13   <2e-16 ***
X2  0.49697    0.01267   39.23   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05588 on 18 degrees of freedom
Multiple R-squared: 0.9992,     Adjusted R-squared: 0.9992
F-statistic: 1.181e+04 on 2 and 18 DF,  p-value: < 2.2e-16
```

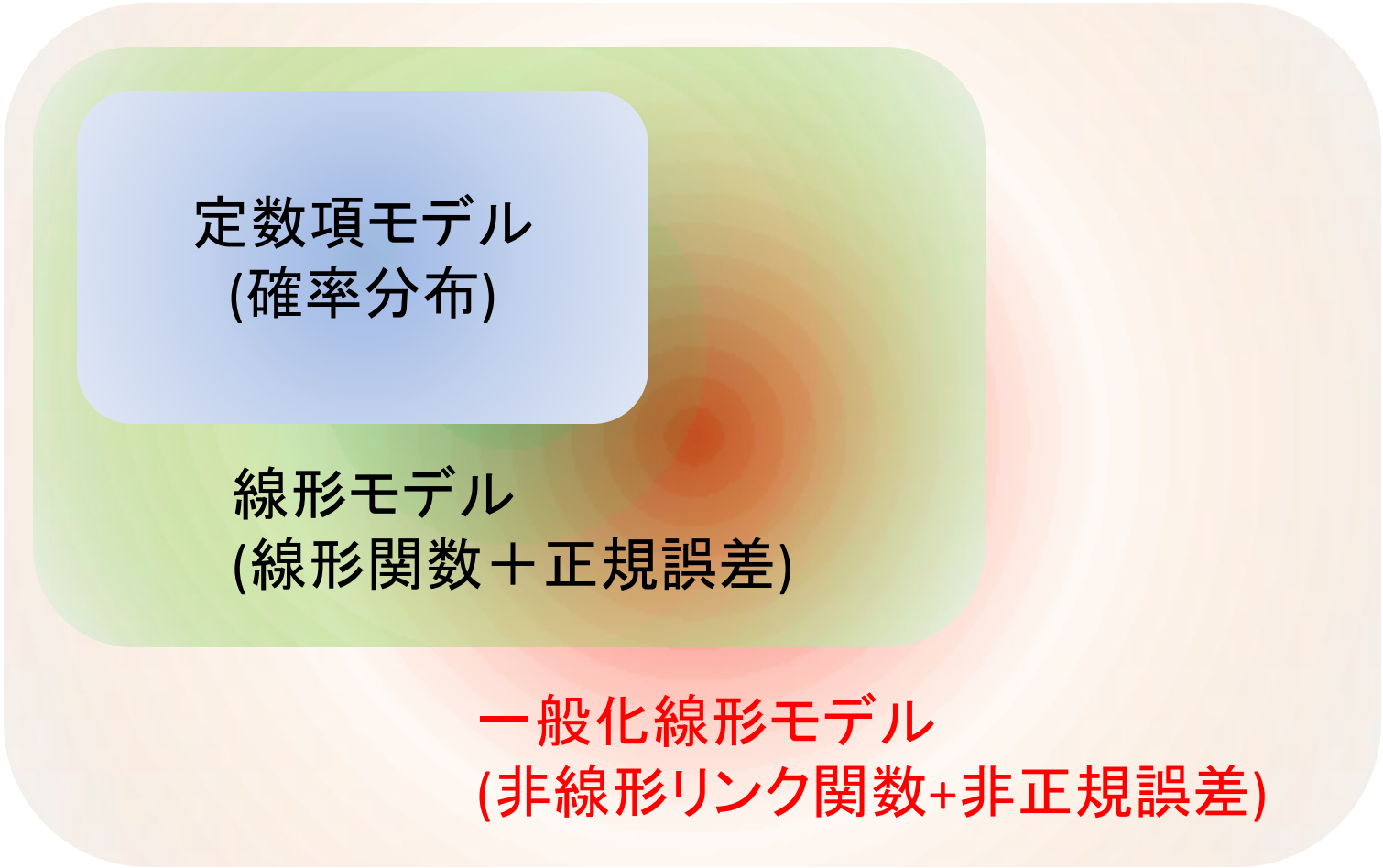
設定した式(モデル)

残差の代表値

推定された係数(パラメタ)

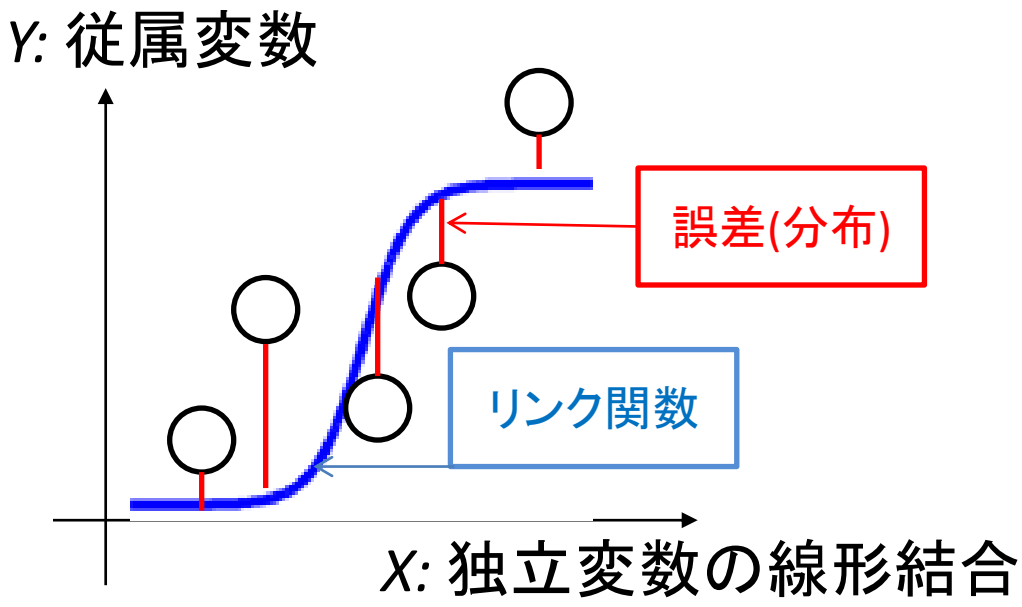
推定の誤差など統計量
- 残差の標準誤差と自由度
- 決定係数, 自由度調整済み決定係数
- F統計量(その平方根はt値)

線形モデルの階層



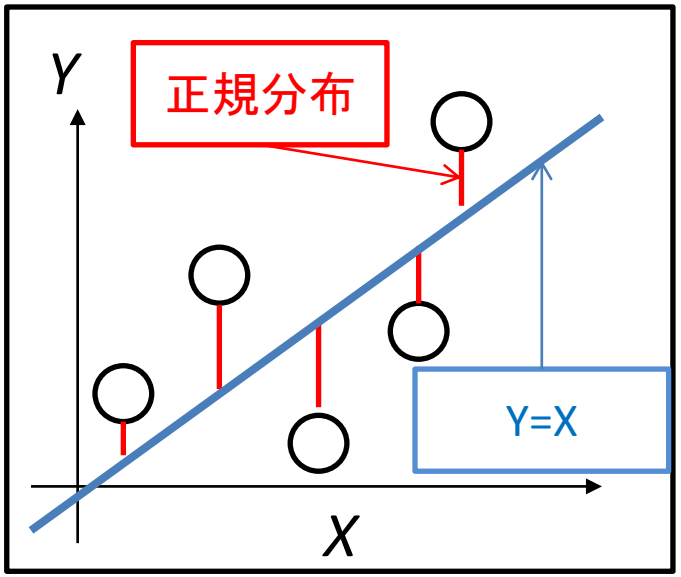
一般化線形モデル

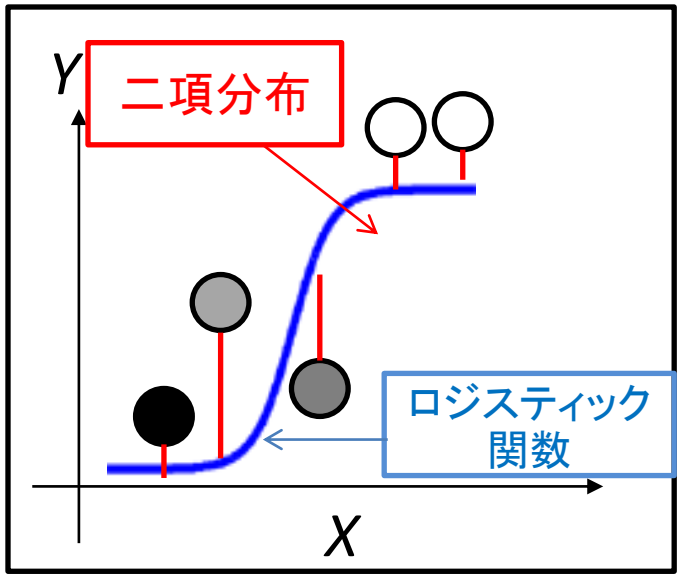
- 線形モデルの拡張：２つの組み合わせ
  - 指数族分布(正規分布・二項分布など)
  - 非線形リンク関数 (+線形予測子)



一般化線形モデルの下位モデル

- 例
  - 線形関数(Y=X)＋正規分布～　回帰分析・分散分析
  - ロジスティック関数＋二項分布 ～　ロジスティック回帰
  - 対数関数＋ポアソン分布～　対数回帰分析





応用例：回数データ(比データ)

- 同一条件(独立変数)で [0,1]~(失敗/成功)の1回/複数回測定した場合
  - e.g., 10人中7人、”類似”と答えた
- “比”の線形回帰？
  - 線形関数は、[0, 1]の範囲を超える可能性あり
  - 比によって分散が異なる→等分散性の欠如
    - 50%のときに最大, 0%,100%で最小→二項分布
  - →ロジスティック回帰(二項分布)が最適

ロジスティック回帰

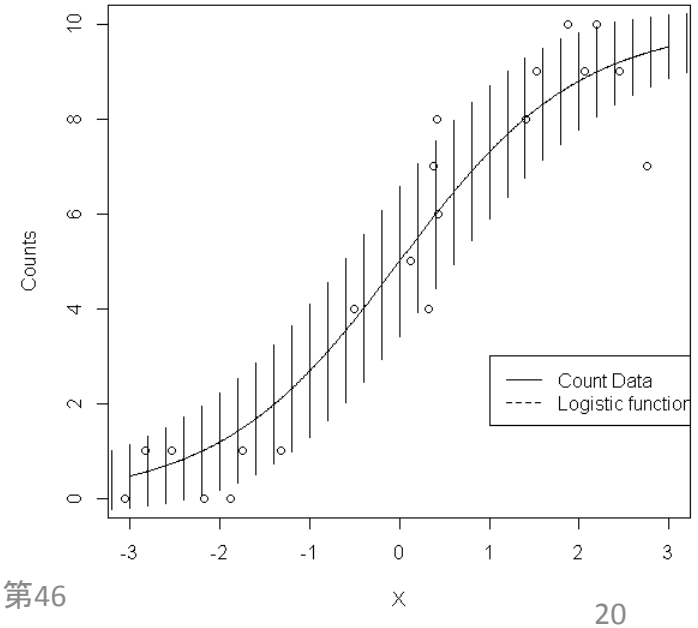
- モデル: y = logistic($ax + c$) + ε
- ある変数yがある確率p=logistic($ax + c$)の二項分布Binom(p, N)に従うとするモデル

$$Y \approx \mathrm{logistic}(XA) + \varepsilon$$

$$L_{\mathrm{binomial}} = \sum_{i=1}^{N} [y_i \log p_i + (1 - y_i) \log(1 - p_i)]$$

$$p_i = \mathrm{logit}(ax_i + c)$$

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1M} & 1 \\ x_{21} & \cdots & x_{2M} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ x_{N1} & \cdots & x_{NM} & 1 \end{pmatrix} \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \quad A = \begin{pmatrix} a_1 \\ \vdots \\ a_M \\ c \end{pmatrix}$$



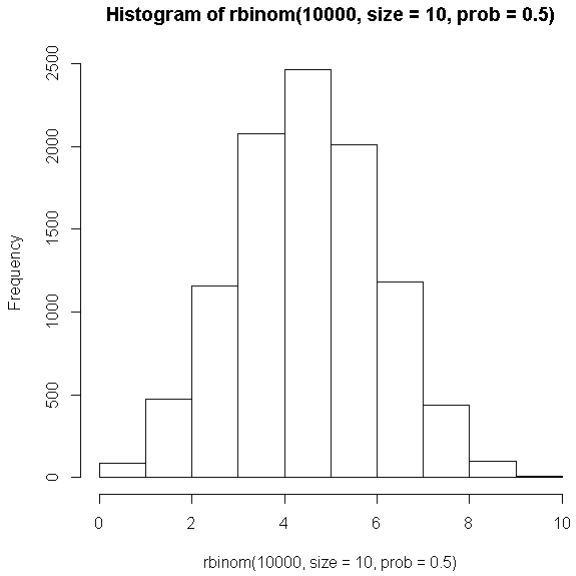
ロジスティック回帰

二項分布・ロジスティック関数

- 二項分布(binom)
  - M回の試行のうち、N回が”1,” (N-M)回が”0”となる確率

$$P\left(\sum_{i=1}^{N} X_i = M\right) = \frac{N!}{(N-M)!M!} p^M (1-p)^{N-M}$$

```
## Visualize binomial distribution
hist( rbinom( 10000, size = 10, prob = 0.5 ), breaks = seq( 0, 10 ) );
```



二項分布のヒストグラムの描画 (N=10, p = 0.5)

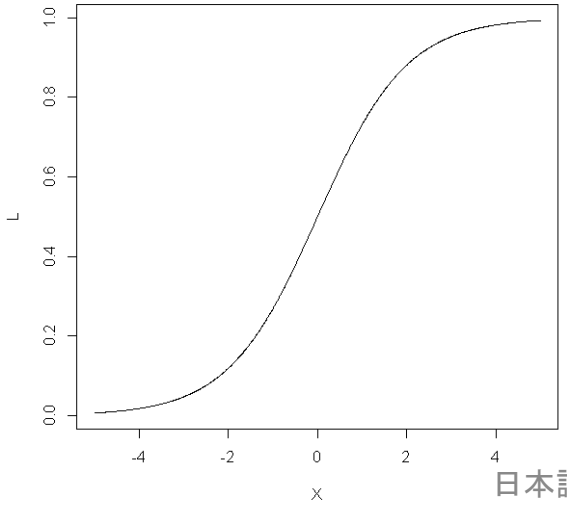
ロジスティック回帰

二項分布・ロジスティック関数

- ロジスティック関数
  - 指数的な加速(x~-∞) / 指数的な減速(x~+∞)

$$f(x) = \frac{1}{1+\exp(-x)}$$

$$\frac{\partial f(x)}{\partial x} = \frac{\exp(-x)}{(1+\exp(-x))^2} = f(x)(1-f(x))$$




```
# In script “GeneralizedLinearModel.R”
## Visualize logistic function
# define the logistic function
logistic <- function( x ) 1/ (1 + exp( - x ) )
X <- seq( -5, 5, by = 0.01 );
L <- logistic( X );
plot( X, L, "l" );
```

ロジスティック関数の描画

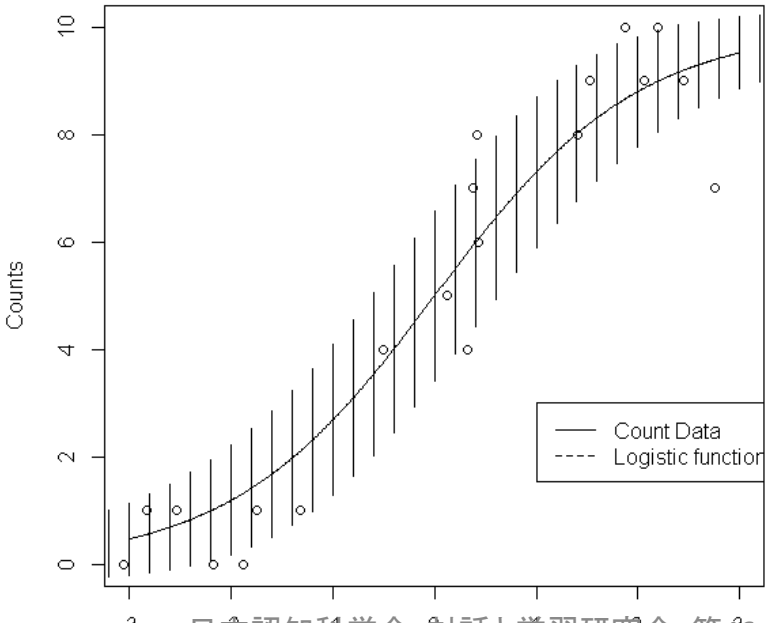
線形回帰との対比

	線形回帰分析	ロジスティック回帰分析
典型的なデータの種類	任意の連続値	非負整数
線形予測子	XA	XA
リンク関数(逆関数)	Y=X	Y=1/(1 + exp(-X)) X = log(Y/(1-Y))
尤度関数	$-1/(2\pi\sigma^2)^{1/2}\ (Y-X)^2$	$Y^M(1-Y)^{(N-M)}$
分散	全てのYで等分散σ	Yに応じて分散が変化 (Y(1-Y))

R演習2: 一般化線形モデル

GeneralizedLinearModel.R

- 用意されているスクリプトを利用して、ロジスティック回帰モデルによる分析をせよ
  - 使用関数: glm,



データ: 類似性判断

- 仮想心理実験
  - 10人の被験者がそれぞれ特徴1の類似性(X1), 特徴2の類似性(X2)を操作した物体対(50対)の類似性を判断(類似・非類似の２択)した。
  - 独立変数はX1, X2, C(“類似”バイアス)
  - 被験者のそれぞれの類似性に対する注意～係数A=(1.5, -0.8, 0.5)
  - P(類似反応) = logistic(X*A), Y~Binom(P, size=5)
- 手元のデータ　X, Yから選択的注意Aを推定

```
# In script “LinearModel.R”
## Data generation
NumData = 50; ## The number of object pairs to rate their similarity
Trials = 10; ## The number of obtained response for each object (subjects)

# generate an artificial data. The second column is the constant
Feature1Similarity = seq( -5, 4.8, 0.2 ) + rnorm( NumData )*0.01;
Feature2Similarity = rnorm( NumData );
Bias = rep( 1, NumData );
X <- cbind( Feature1Similarity, Feature2Similarity, Bias );
Atrue <- c( 1.5, -0.8, 0.5 ); # A set of coefficients
P <- logistic( X%*%Atrue ); # Generate the output Y by using "linear model".
Y <- rbinom( NumData, size = Trials, prob = P ); # Draw samples from binomial distribution
```

パラメタ推定

推定結果 = glm(式, family=分布族)

```
## Using the logistic fitting function in R
# Y is two-column matrix ( N, M-N )where M is the total number of trials
YwithCounts = cbind( Y, rep( Trials, NumData ) - Y );

# default link function is logistic for binomial family /the output is object "Results"
Results = glm( YwithCounts ~ 0+X, family = binomial );
```

結果例

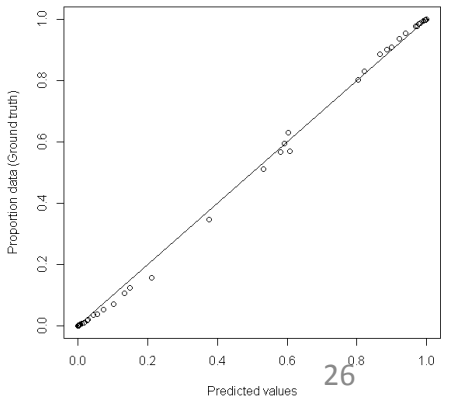
```
> summary( Results ) ## Print out the summary of GLM
Call:
glm(formula = YwithCounts ~ 0 + X, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.64677  -0.25308   0.01869   0.34570   1.66323

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
XFeature1Similarity   1.6791     0.1747   9.609  < 2e-16 ***
XFeature2Similarity  -0.8097     0.2155  -3.757 0.000172 ***
XBias                 0.4800     0.1921   2.499 0.012438 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Null deviance: 535.938  on 50  degrees of freedom
Residual deviance:  28.055  on 47  degrees of freedom
AIC: 72.3
```

<u>推定された結果</u>

XFeature1Similarity　1.6791
XFeature2Similarity　-0.8097
XBias　0.4800

<u>真の値</u>

類似性1: 1.5
類似性2:　-0.8
バイアス: 0.5



モデル選択

- 複数のモデルのデータに対する適合性(尤度・自由度)を比較する
  - データを生成する尤度の高さ
  - より単純なモデル(自由度)
- モデル選択基準
  - AIC (Akaike, 1973): AIC = -2log(尤度) + 2(自由度)
  - BIC (Schwarz, 1978): BIC = -2log(尤度) + log(標本数)(自由度)
  - AIC(またはBICなど)のより低いモデルを選ぶ
- 検定との違い
  - 検定は、効果の有無を調べるとき
  - モデルは、特定の仮説の妥当性を相対的に比較

R演習2-2: モデル選択

- 冗長モデルを正しく棄却できるか？
  - 類似性判断とは無関係な変数を加えた対立モデル

```
# In script “LinearModel.R”
## Model selection: a redundant model as an alternative hypothesis
## Data with irrelevant dimension (the model is redundant, and would have higher AIC)
XwithDummy <- cbind(X, rnorm( NumData ) );

# default link function is logistic for binomial family /the output is object "Results"
Results2 = glm( YwithCounts ~ 0+XwithDummy, family = binomial );

## AIC( Results ) < AIC( Results2 ) if it correctly reject the model2 with redundant variable.
```

```
> AIC( Results )
[1] 72.29962
> AIC( Results2 )
[1] 74.2347
```

まとめ2

- 定数項、線形、一般化線形モデルを階層的な関係の中で紹介した
  - 単純～複雑なモデルまでの関連性・連続性を意識
- 関連する話題
  - モデル選択・次元縮約
- より一般的なモデル(分散の分解)
  - 階層性/欠損データ: 一般化線形混合モデル(GLMM)
  - 非線形性: 非線形回帰、ニューラルネット
  - 複数vs複数: 正準相関分析、構造方程式モデル

参考文献

- McCullagh, P & Nelder, J. (1989). *Generalized Linear Models, Second Edition*. Boca Raton: Chapman and Hall/CRC.
- 船尾 (2005). The R Tips, 九天社
- 岡田 (2004). The R Book, 九天社
- Online materials
  - http://cse.naro.affrc.go.jp/takezawa/r-tips/r.html
  - http://mjin.doshisha.ac.jp/R/15.pdf
  - http://www.rd.dnc.ac.jp/~otsu/Komaba2005/Komaba05Oct.pdf