

# 数理統計 補助資料 ～ロジスティック回帰モデル～

2023年度2学期: 月曜1限, 水曜3限  
担当教員: 石垣 司

1

## ロジスティック回帰モデル #1

### 目的変数が2値変数(0 or 1など)の回帰モデル

【説明変数】

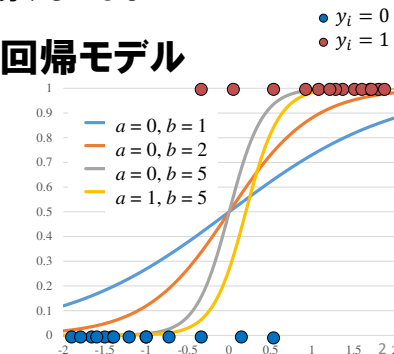
毎日の喫煙量  
広告視聴時間  
毎日の学習時間  
気温の累計

【目的変数】

肺がんの発病 or 非発病  
商品を買う or 買わない  
志望校への合格 or 不合格  
植物が発芽する or しない

### 説明変数が1つのロジスティック回帰モデル

- データ:  $\{x_i, y_i\} (i = 1, \dots, N)$
- 目的変数:  $y \in \{0, 1\}$
- 説明変数:  $x \in \mathbb{R}$
- パラメータ:  $\{a, b\}$
- 関数形:  $f(x) = \frac{1}{1+e^{-(a+bx)}}$



## ロジスティック回帰モデル #2

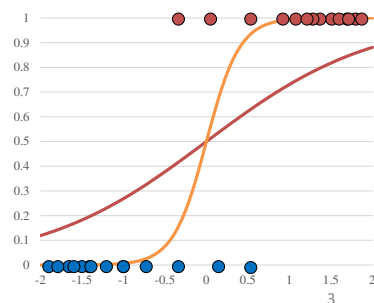
### ロジスティック回帰モデルの特徴

- $f(x) = \Pr(y = 1)$   
関数  $f(x)$  の値は  $y = 1$  となる確率  $\Pr(y = 1)$  の値
  - $\Pr(y = 0) = 1 - f(x)$   
 $y = 0$  となる確率  $\Pr(y_i = 0)$  の値は  $1 - f(x)$
- S字型の関数形はシグモイド関数とも呼ばれる

### ロジスティック回帰モデルの推定

- $f(x) = \frac{1}{1+e^{-(a+bx)}}$
- パラメータ  $\theta = \{a, b\}$  を推定

# 深層学習(ディープラーニング)はこのシグモイド関数(様々な亜種を含む)を多数組み合わせさせた予測・判別手法



3

## ロジスティック回帰モデル #3

### 説明変数が複数あるロジスティック回帰モデル

- 目的変数: 変数  $y$
- 説明変数: 変数  $x = [x_1, x_2, \dots, x_p]^T$
- データ:  $D = \{y_i, x_i\} (i = 1, \dots, N)$   
 $x_i = [1 \ x_{i1} \ \dots \ x_{ip}]^T$  (行列  $X$  の  $i$  番目の行)
- 回帰係数: 係数  $b_0, b_1, \dots, b_p$  (パラメータ)

$$b = [b_0 \ b_1 \ \dots \ b_p]^T$$

$$\Pr(y = 1|x) = \frac{1}{1+e^{-(b_0+b_1x_1+\dots+b_px_p)}} = \frac{1}{1+e^{-x^T b}}$$

4

# ロジスティック回帰分析をやってみよう

## ポルトガルの金融機関のダイレクトマーケティングデータ

- 2008年から2013年の実データ (N = 45,211)
- 消費者にマーケティングプロモーションを実施
- 目的変数: 長期定期預金への加入 or 非加入
- 説明変数: 年齢(連続変数), 預金残高(連続変数), 住宅ローンの有無(2値変数), ローンの有無(2値変数), 債務不履行の経験の有無(2値変数), 結婚(未婚, 既婚, 離婚の3カテゴリ), 教育(初等, 中等, 高等, 不明の4カテゴリ)

#メモ UCI Machine Learning Repository (<http://archive.ics.uci.edu/ml>)でBank Marketing Data Setとして無料でダウンロード可能

# Rによるロジスティック回帰分析

## Rのコードと出力

```
Rのコード
Data = read.csv("bank-full.csv",header=T)
Data$y = factor(Data$y,label=c("no","yes"))
logistic =
glm(y~age+balance+housing+loan+default+marital+education, family = binomial, data=Data)
summary(logistic)

データ
> head(Data)
  age  job marital education default balance housing loan contact day
1  58 management married tertiary no 2143 yes no unknown 5
2  44 technician single secondary no 29 yes no unknown 5
3  33 entrepreneur married secondary no 2 yes yes unknown 5
4  47 blue-collar married unknown no 1506 yes no unknown 5
5  33 unknown single unknown no 1 no no unknown 5
6  35 management married tertiary no 231 yes no unknown 5
  month duration campaign pdays previous poutcome y
1 may 261 1 -1 0 unknown no
2 may 151 1 -1 0 unknown no
3 may 76 1 -1 0 unknown no
4 may 92 1 -1 0 unknown no
5 may 198 1 -1 0 unknown no
6 may 139 1 -1 0 unknown no

ロジスティック回帰分析の結果
Call:
glm(formula = y ~ age + balance + housing + loan + default + marital + education, family = binomial, data = Data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8055  -0.5550  -0.4309  -0.3516   2.8196

Coefficients:
(Intercept)          -2.352e+00  1.032e-01 -22.797 < 2e-16 ***
age                 1.053e-02  1.563e-03  6.734 1.65e-11 ***
balance            2.358e-05  3.937e-06  5.991 2.08e-09 ***
housing            -7.888e-01  3.124e-02 -25.249 < 2e-16 ***
loan               -5.920e-01  5.033e-02 -11.761 < 2e-16 ***
default            -5.167e-01  1.458e-01 -3.543 0.000395 ***
maritalmarried    -1.695e-01  4.814e-02 -3.522 0.000428 ***
maritalsingle     3.128e-01  5.424e-02  5.767 8.07e-09 ***
educationsecondary 2.678e-01  4.972e-02  5.386 7.20e-08 ***
educationtertiary 5.060e-01  5.161e-02  9.805 < 2e-16 ***
educationunknown 3.188e-01  8.166e-02  3.904 9.44e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 32631 on 45210 degrees of freedom
Residual deviance: 31177 on 45200 degrees of freedom
AIC: 31199

Number of Fisher Scoring iterations: 5
```

## ロジスティック回帰係数の解釈 #1

### ロジスティック回帰係数の推定結果(最尤推定)

	Estimate (推定値)	Std.Error (標準誤差)	t value (t値)	Pr(> t ) (p値)
切片 (b <sub>0</sub> )	-2.35244	0.10319	-22.79711	0.00000***
年齢 (b <sub>1</sub> )	0.01053	0.00156	6.73369	0.00000***
預金残高 (b <sub>2</sub> )	0.00002	0.00000	5.99123	0.00000***
住宅ローンあり (b <sub>3</sub> )	-0.78878	0.03124	-25.24911	0.00000***
ローンあり (b <sub>4</sub> )	-0.59198	0.05033	-11.76101	0.00000***
既婚 (b <sub>5</sub> )	-0.51667	0.14582	-3.54310	0.00040***
未婚 (b <sub>6</sub> )	-0.16954	0.04814	-3.52216	0.00043***
中等教育 (b <sub>7</sub> )	0.31283	0.05424	5.76702	0.00000***
高等教育 (b <sub>8</sub> )	0.26781	0.04972	5.38621	0.00000***
教育歴不明 (b <sub>9</sub> )	0.50603	0.05161	9.80508	0.00000***

- ロジスティック回帰係数の最尤推定量は一致推定量
- 線形回帰モデルと同様に t 検定(自由度 N - (P + 1))が可能

#メモ1 一致推定量と t 検定については分布収束の知識が必要なため、証明はこの授業では扱わない  
#メモ2 どのように推定されたのかは次回以降の授業で紹介

## ロジスティック回帰係数の解釈 #2

### 回帰係数と対数オッズ比

$$- \Pr(y = 1|x) = \frac{1}{1+e^{-x^T b}}$$

$$- \Pr(y = 0|x) = 1 - \Pr(y = 1|x) = 1 - \frac{1}{1+e^{-x^T b}} = \frac{e^{-x^T b}}{1+e^{-x^T b}}$$

$$- \text{オッズ比: } \frac{\Pr(y=1|x)}{\Pr(y=0|x)} = e^{x^T b}$$

$$- \text{対数オッズ比: } \log\left(\frac{\Pr(y=1|x)}{\Pr(y=0|x)}\right) = x^T b$$

### ロジスティック回帰係数の解釈

- 説明変数 x<sub>p</sub> が1単位増減すると対数オッズ比がロジスティック回帰係数 b<sub>p</sub> の分だけ増減する

※線形回帰モデルと解釈が異なるので注意

## ロジスティック回帰係数の解釈 #3

### オッズ比 $\frac{\Pr(y=1|x)}{\Pr(y=0|x)}$ の意味

- 説明変数  $x$  が与えられた時の  $y = 1$  の出やすさの度合い

$$\frac{\Pr(y=1|x)}{\Pr(y=0|x)} > 1 \text{ なら } y = 1 \text{ が出やすい}$$

$$\frac{\Pr(y=1|x)}{\Pr(y=0|x)} < 1 \text{ なら } y = 0 \text{ が出やすい}$$

$$\frac{\Pr(y=1|x)}{\Pr(y=0|x)} = 1 \text{ なら } y = 1 \text{ と } y = 0 \text{ の出やすさは等確率}$$

例:  $\frac{\Pr(y=1|x)}{\Pr(y=0|x)} = 2$  のとき,  $\Pr(y = 1|x) : \Pr(y = 0|x) = 2 : 1$

例:  $\frac{\Pr(y=1|x)}{\Pr(y=0|x)} = 0.5$  のとき,  $\Pr(y = 1|x) : \Pr(y = 0|x) = 1 : 2$

9

## ロジスティック回帰係数の解釈 #4

### オッズ比とロジスティック回帰係数の関係

$$\frac{\Pr(y=1|x)}{\Pr(y=0|x)} = e^{a+bx} = e^a e^{bx}$$

$e^a$  : 説明変数  $x$  に関係しない元々の  $y = 1$  の出やすさ

$e^{bx}$  : 説明変数  $x$  が与えられた時のオッズ比の増減

線形回帰モデル:  $y = a + bx$  のとき  $a + b(x + \Delta) \rightarrow y + b\Delta$

ロジスティック回帰モデル:

$$\frac{\Pr(y=1|x)}{\Pr(y=0|x)} = e^a e^{bx} \text{ のとき } e^a e^{b(x+\Delta)} \rightarrow \frac{\Pr(y=1|x)}{\Pr(y=0|x)} e^{b\Delta}$$

### ロジスティック回帰係数のオッズ比での解釈

- 説明変数が1単位増減するとオッズ比が  $e^b$  倍になる
  - $b > 0$  なら  $e^{b\Delta} > 1$  なので  $x$  が増加で  $y = 1$  が出やすくなる
  - $b < 0$  なら  $e^{b\Delta} < 1$  なので  $x$  が増加で  $y = 0$  が出やすくなる

10

## ロジスティック回帰係数の解釈 #5

### ロジスティック回帰係数の推定結果(抜粋)

	Estimate (推定値)		Pr(> t ) (p値)
年齢 ( $b_1$ )	0.01053	$\exp(0.01053) = 1.01$	0.00000***
住宅ローンあり ( $b_3$ )	-0.78878	$\exp(-0.78878) = 0.45$	0.00000***
既婚 ( $b_5$ )	-0.51667	$\exp(-0.51667) = 0.60$	0.00040***
未婚 ( $b_6$ )	-0.16954	$\exp(-0.16954) = 0.84$	0.00043***

### 推定されたロジスティック回帰係数の解釈

- 年齢が1歳増えるとオッズ比が 1.01 倍  
(10歳増える  $\Rightarrow 1.01^{10} = 1.11$ , 30歳増える  $\Rightarrow 1.01^{30} = 1.37$ )
- 住宅ローンありと比べて, 無しはオッズ比が 0.45 倍
- 離婚者と比べて, 既婚の人はオッズ比が 0.6 倍, 未婚の人は 0.84 倍

11

## 統計モデルのモデル選択 #1

ある観測データに対する複数の統計モデルの当てはまりの良さを比較し, より良いモデルを選び出すこと

観測データ  $D$

比較したいモデル群

説明変数が4種類の場合のロジスティック回帰モデル  $\Pr(y = 1|x) = \frac{1}{1+e^{-x^T b}}$

Model 1:  $x^T b = b_0 + b_1 x_1$

Model 2:  $x^T b = b_0 + b_2 x_2$

Model 3:  $x^T b = b_0 + b_1 x_3$

Model 4:  $x^T b = b_0 + b_1 x_4$

Model 5:  $x^T b = b_0 + b_1 x_1 + b_2 x_2$

Model 6:  $x^T b = b_0 + b_1 x_1 + b_2 x_3$

Model 7:  $x^T b = b_0 + b_1 x_1 + b_2 x_4$

Model 8:  $x^T b = b_0 + b_1 x_2 + b_2 x_3$

Model 9:  $x^T b = b_0 + b_1 x_2 + b_2 x_4$

Model 10:  $x^T b = b_0 + b_1 x_3 + b_2 x_4$

Model 11:  $x^T b = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$

Model 12:  $x^T b = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_4$

Model 13:  $x^T b = b_0 + b_1 x_1 + b_2 x_3 + b_3 x_4$

Model 14:  $x^T b = b_0 + b_1 x_2 + b_2 x_3 + b_3 x_4$

Model 15:  $x^T b = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4$

当てはまりの良さの指標に基づいてモデル選択  
 $\Rightarrow$  どのような意味で「良い」のか?

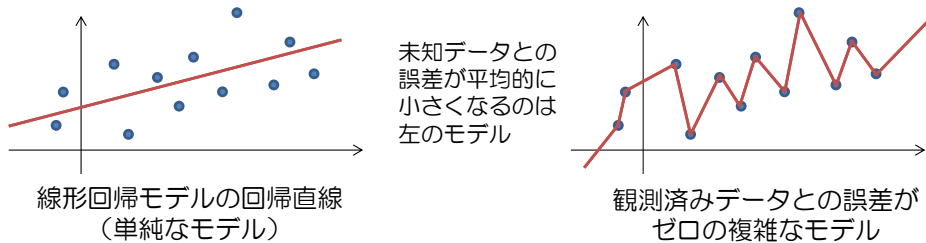
12

# 統計モデルのモデル選択 #2

予測の意義: 将来得られる未観測データを予測をしたい

実証の意義: データから母集団の性質を知りたい

- 既に手元にある観測データの目的変数を高精度に予測できて、多くの場合に意味はない。過学習 (overfitting) を回避
- 過学習: 手元にある観測済みのデータのみで過剰に適合してしまい、未観測データの予測や実証の性能が低い統計モデルを作成すること



複雑なモデルの構築や余計な説明変数の追加は、統計モデルの性能を悪化させることがある

# モデル選択規準

過学習を回避するモデル選択の規準

- 代表例: AIC(Akaike's Information Criterion), BIC(Bayesian Information Criterion), MDL(Minimum Description Length), WAIC(Widely Applicable Information Criterion), 変数選択法, 交差検証法等による予測精度の検証も一つ的手段

AIC 赤池情報量規準

- 想定したモデルが真のモデルの近くにある場合、将来に得られる未知データの予測の平均的な良さの評価

$$AIC = -2 \times \text{対数尤度} + 2 \times \text{パラメータ数}$$

(値が小さい方が良い指標)



#メモ 具体的な数式の導出は本授業の範囲を大きく超えるため紹介できない。詳しくは小西・北川「情報量規準」, 朝倉書店, 2004 を参照

Google を検索または URL を入力

2017年11月5日のGoogle 14

# AICでモデル選択を試みよう

ポルトガルの金融機関のダイレクトマーケティングデータ

- 目的変数: 長期定期預金への加入 or 非加入
- 説明変数: 年齢, 預金残高, 住宅ローンの有無, ローンの有無, 債務不履行の経験の有無, 結婚, 教育

$$Pr(y = \text{加入} | x) = \frac{1}{1 + e^{-x^T b}}$$

$$x^T b = b_0 + b_1 \text{年齢} + b_2 \text{預金} + b_3 \text{住宅} + b_4 \text{ローン} + b_5 \text{不履行} + b_6 \text{結婚} + b_7 \text{教育}$$

```
> ll = step(logistic)
Start: AIC=31199.32
y ~ age + balance + housing +
```

<none>	Df	Deviance	AIC
- none	1	31177	31199
- default	1	31192	31212
- balance	1	31211	31231
- age	1	31222	31242
- education	3	31289	31305
- loan	1	31334	31354
- marital	2	31349	31367
- housing	1	31835	31855

AICの値から、すべての説明変数を利用したモデルが「良いモデル」と判断

$$Pr(y = \text{加入} | x) = \frac{1}{1 + e^{-x^T b}}$$

$$x^T b = b_0 + b_1 \text{年齢} + b_2 \text{預金} + b_3 \text{住宅} + b_4 \text{ローン} + b_5 \text{乱数} + b_6 \text{結婚} + b_7 \text{教育}$$

```
> ll = step(logistic)
Start: AIC=31213.77
y ~ age + balance + housing +
```

<none>	Df	Deviance	AIC
- Random	1	31192	31212
<none>		31192	31214
- balance	1	31228	31248
- age	1	31237	31257
- education	3	31305	31321
- loan	1	31357	31377
- marital	2	31363	31381
- housing	1	31848	31868

AICの値から、変数「乱数」を取り除いたモデルが「良いモデル」と判断

# 演習問題

1. 目的変数が2値変数の場合に、線形回帰モデルを適用してデータ分析を行うことを考える。このとき、予測と実証の両方の観点から、その問題点を指摘しなさい
2. ①決定係数, ②尤度(または、対数尤度)の両方は、モデル選択の基準として不適格となることがある。その理由をそれぞれについて答えなさい