統計学01

早稲田大学政治経済学部

第22回

西郷 浩

本日の目標

- 質的な被説明変数
  - 例：タイタニック号乗客の生存・死亡
  - 説明変数の影響の程度をロジット回帰分析（質的被説明変数のための回帰モデル）で分析する。
- 参考文献
  - 小暮厚之(2009)『Rによる統計データ分析入門』朝倉書店

データ

- タイタニック号の乗客の生存・死亡（1313人分）
  - 氏名
  - 客室の等級（1, 2, 3）
  - 性別（male, female）
  - 年齢（数量）
  - 生存・死亡（1, 0）
  - 参考文献に入手方法が記載されています。

データ

表1:基本統計

PClass	Age	Sex	Survived
1st:322	Min. : 0.17	female:462	Min. : 0.0000
2nd:280	1st Qu.: 21.00	male :851	1st Qu.: 0.0000
3rd:711	Median : 28.00		Median :0.0000
	Mean : 30.40		Mean : 0.3427
	3rd Qu.: 39.00		3rd Qu.: 1.0000
	Max. : 71.00		Max. : 1.0000
	NA's : 557.00		

データ

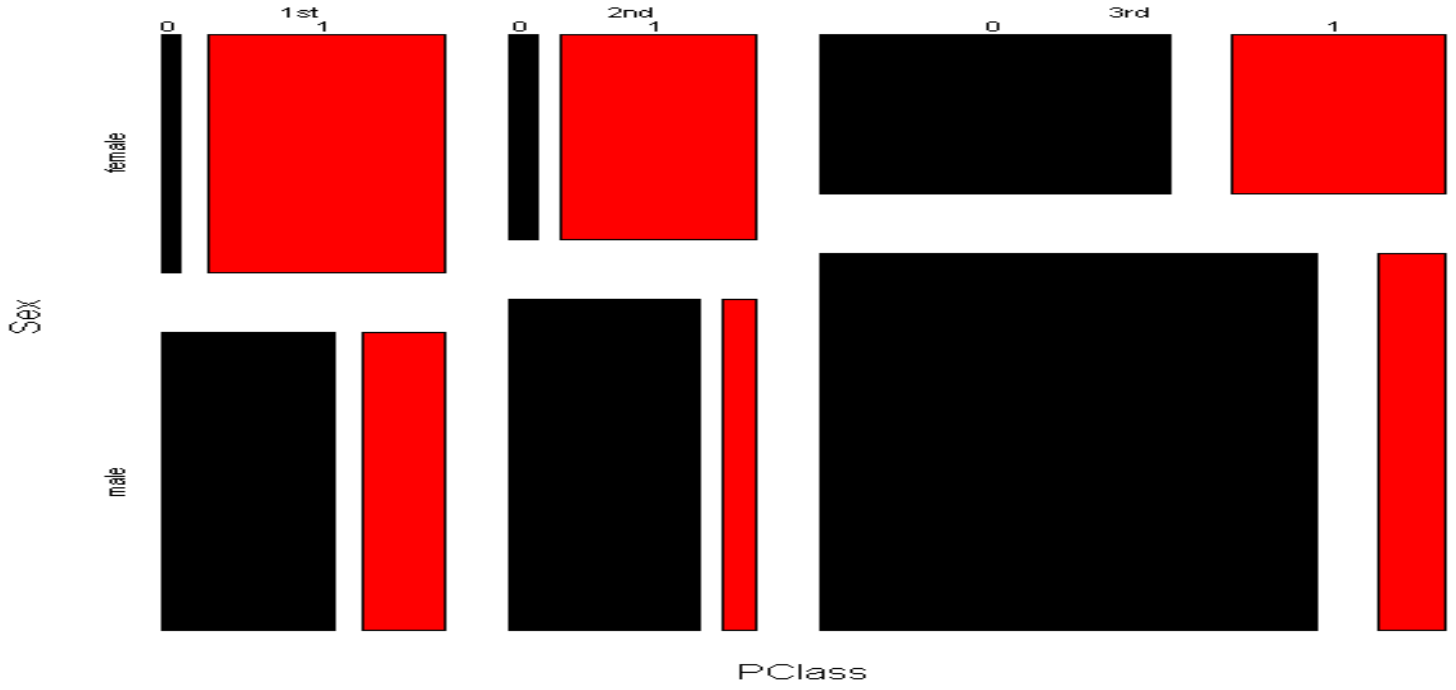

図1：性・客室等級別死亡・生存者数

データ

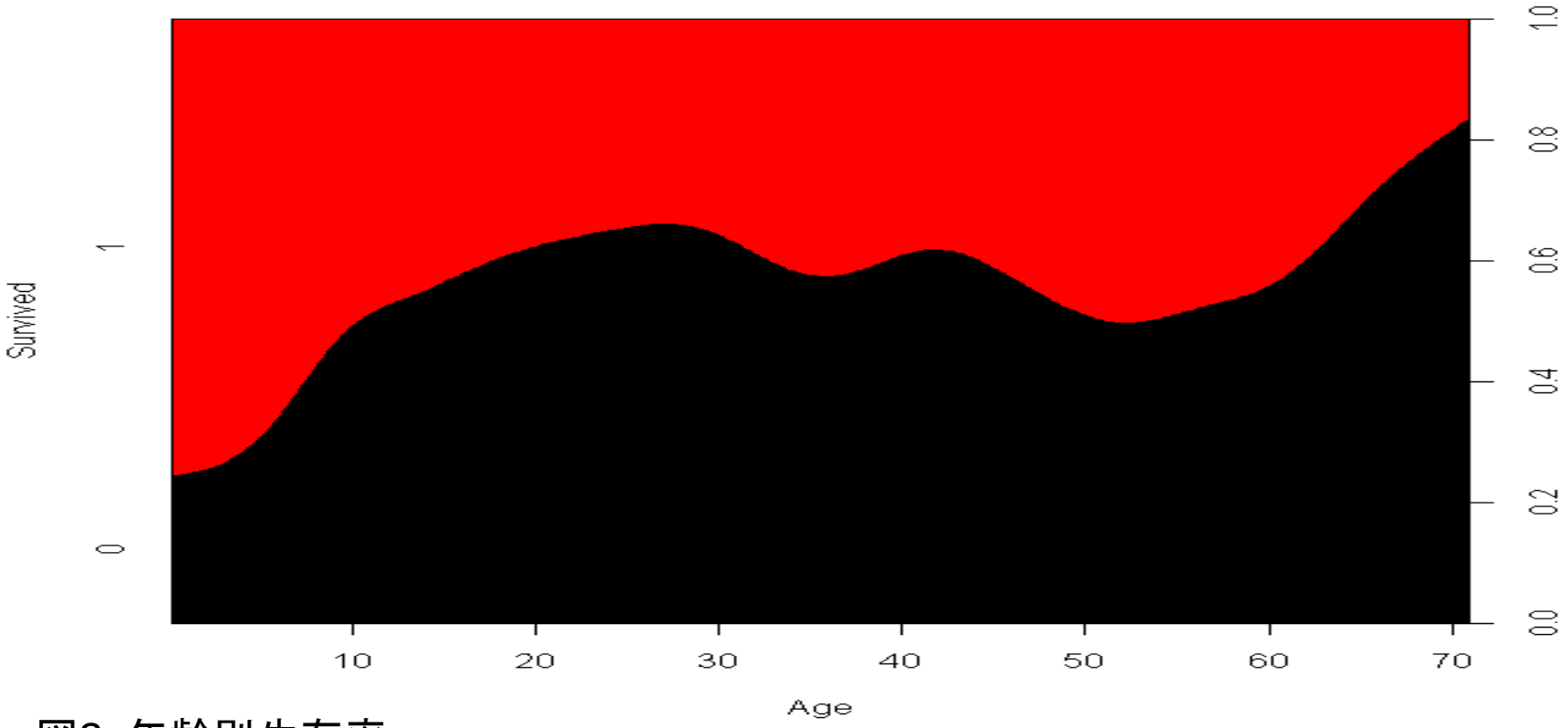

図2:年齢別生存率

生存率推定のモデル

- 被説明変数
  - Y:死亡・生存（0,1）
- 説明変数
  - X_2: 客室の等級（1, 2, 3）
  - X_3: 性別（male, female）
  - X_4: 年齢（数量）
- 回帰モデル?
  - $Y_i = \beta_1+\beta_2X_{2i}+\beta_3X_{3i}+\beta_4X_{4i}+\varepsilon_i$
    - 不適切。理由：Y_iの取りうる値が0-1に限られているから。

生存率推定のモデル

- 生存率
  - $p_i = Pr(Y_i = 1)$
  - 回帰モデル？
    - $p_i = \beta_1+\beta_2X_{2i}+\beta_3X_{3i}+\beta_4X_{4i}(+ \varepsilon_i)$
      - 不適切。理由：$0 < p_i < 1$ に限定されているから。
- オッズ
  - $p_i /(1- p_i)$
  - 回帰モデル？
    - $p_i /(1- p_i) = \beta_1+\beta_2X_{2i}+\beta_3X_{3i}+\beta_4X_{4i}(+ \varepsilon_i)$
      - 不適切。理由：$p_i /(1- p_i) > 0$ に限定されているから。

生存率推定のモデル

- 対数オッズ（logit）
  - $\log p_i /(1- p_i) = \beta_1+\beta_2 X_{2i}+\beta_3 X_{3i}+\beta_4 X_{4i}(+ \varepsilon_i)$
    - 適切。理由：$-\infty < \log p_i /(1- p_i) < \infty$
  - 書き換えると
    - $p_i = \exp(\beta_1+\beta_2 X_{2i}+\beta_3 X_{3i}+\beta_4 X_{4i})/\{1+\exp(\beta_1+\beta_2 X_{2i}+\beta_3 X_{3i}+\beta_4 X_{4i})\}$
  - 回帰係数の推定は？
    - 最尤法
      - 尤度関数の最大化

生存率モデルの推定

同時確率

$$\Pr(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n) \quad (y_i = 0 \text{ or } 1)$$

$$= \Pr(Y_1 = y_1)\Pr(Y_2 = y_2)\cdots\Pr(Y_n = y_n) \quad (\text{independence})$$

$$\text{Note: } \Pr(Y_i = y_i) = p_i^{y_i}(1-p_i)^{1-y_i}$$

$$= p_1^{y_1}(1-p_1)^{1-y_1}\, p_2^{y_2}(1-p_2)^{1-y_2}\cdots p_n^{y_n}(1-p_n)^{1-y_n}$$

$$\text{Note: } p_i = \frac{\exp(\beta_1 + \beta_2 X_{i2} + \cdots + \beta_p X_{ip})}{1+\exp(\beta_1 + \beta_2 X_{i2} + \cdots + \beta_p X_{ip})}$$

$$= L(\beta_1, \beta_2, \ldots, \beta_p; y_1, y_2, \ldots, y_n) \quad (\text{尤度関数})$$

生存率モデルの推定

最尤法

$$\max L(\beta_1, \beta_2, \ldots, \beta_p; y_1, y_2, \ldots, y_n)$$

最尤推定量

最尤法によってえられた

$\beta_1, \beta_2, \ldots, \beta_p$の値

(よい性質をもつことが知られている)

ソフトウェアRでは、回帰分析と同じ手間で実行できる。

生存率モデルの推定

表2:ロジスティック回帰の出力

```
> titanic. glm <- glm(Survived~PClass*Sex+Age, family=binomial,
data=titanic.df)
> summary(titanic.log.glm)

Call:
glm(formula = Survived ~ PClass * Sex + Age, family = binomial,
    data = titanic.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0869  -0.6453  -0.4643   0.4599   2.3346
```

生存率モデルの推定

表2：ロジスティック回帰の出力（つづき）

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	4.845505	0.598061	8.102	5.41e-16 ***
PClass2nd	-1.486038	0.587018	-2.532	0.011357 *
PClass3rd	-4.038030	0.544289	-7.419	1.18e-13 ***
Sexmale	-3.702774	0.507177	-7.301	2.86e-13 ***
Age	-0.044854	0.008179	-5.484	4.16e-08 ***
PClass2nd:Sexmale	-0.089869	0.656052	-0.137	0.891043
PClass3rd:Sexmale	2.256406	0.581805	3.878	0.000105 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

生存率モデルの推定（おまけ）

- 年齢の効果は直線的か？
  - 変数変換の可能性
  - 先験的に関数形を決めにくい。
    - 一般化加法モデルの利用
      - スプライン関数を利用して柔軟な当てはめを実行する。

生存率モデルの推定（おまけ）

表3：一般化加法モデルの出力

```
> titanic.gam <- gam(Survived~PClass*Sex+s(Age),
family=binomial,data=titanic.df)
> summary(titanic.gam)

Family: binomial
Link function: logit

Formula:
Survived ~ PClass * Sex + s(Age)
```

生存率モデルの推定（おまけ）

表3：一般化加法モデルの出力（つづき）

Parametric coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.3929	0.4865	6.974	3.09e-12 ***
PClass2nd	-1.4112	0.5966	-2.366	0.018005 *
PClass3rd	-3.9049	0.5473	-7.134	9.72e-13 ***
Sexmale	-3.6979	0.5115	-7.229	4.87e-13 ***
PClass2nd:Sexmale	-0.1029	0.6646	-0.155	0.876905
PClass3rd:Sexmale	2.2687	0.5858	3.873	0.000108 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:

	edf	Ref.df	Chi.sq	p-value
s(Age)	3.385	4.211	38.23	1.32e-07 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

生存率モデルの推定（おまけ）

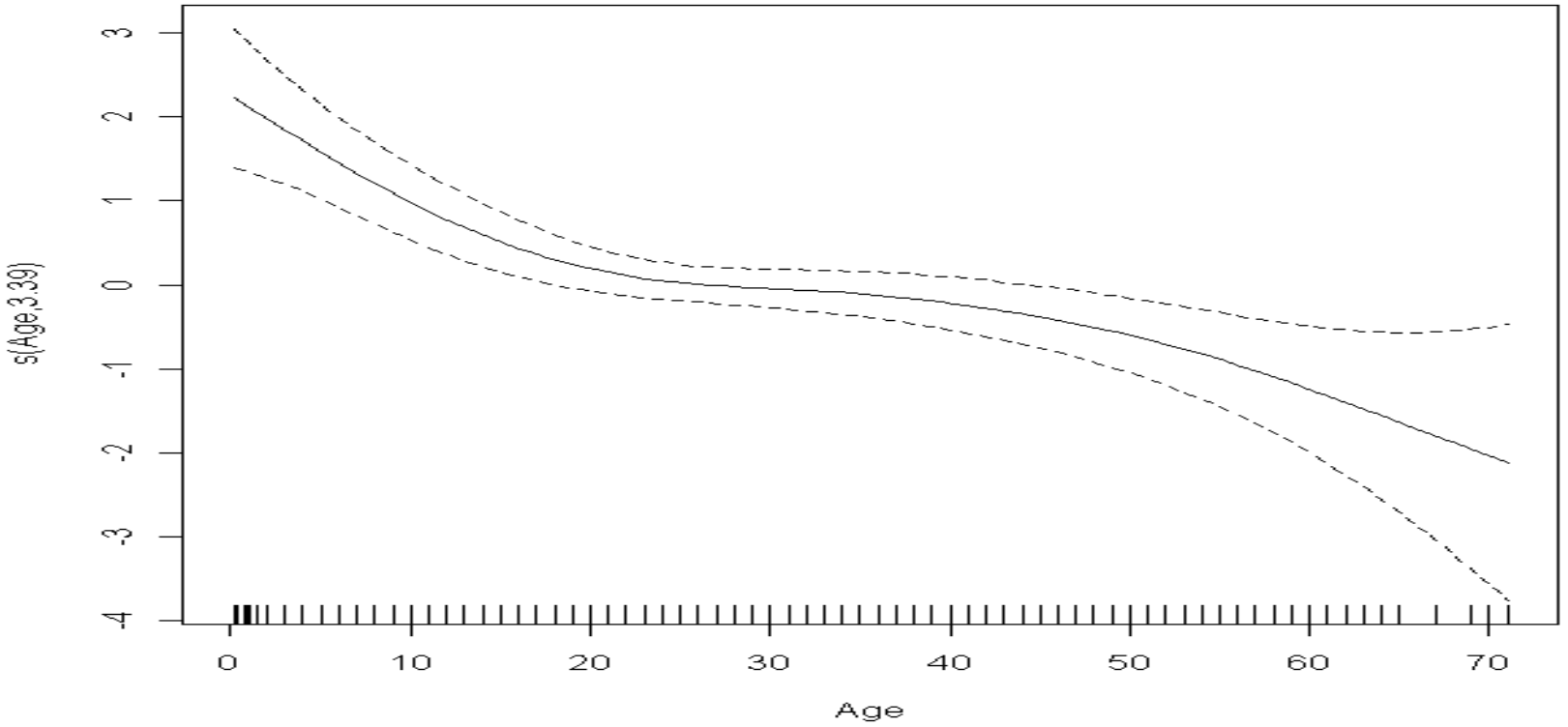

図3：一般化加法モデルにおける年齢の効果

まとめ

- 多変量解析（回帰分析系）
  - 単回帰分析：説明変数ひとつ
  - 重回帰分析：説明変数2つ以上
    - ダミー変数による回帰式のシフト
  - 分散分析：質的（離散的）な要因
    - 一元配置
    - 二元配置
      - 交互作用の導入
  - ロジスティック回帰：質的被説明変数

まとめ

- Rなどのソフトウェアでは
  - どの手法もほとんど同じコマンドで実行できる。
    - どんどん使ってみることが大事。