基 礎 統 計 学

第 6 回　多変量解析
重回帰とロジスティック回帰

2017. 11. 24

回帰分析

多変量解析でよく使われるものとして

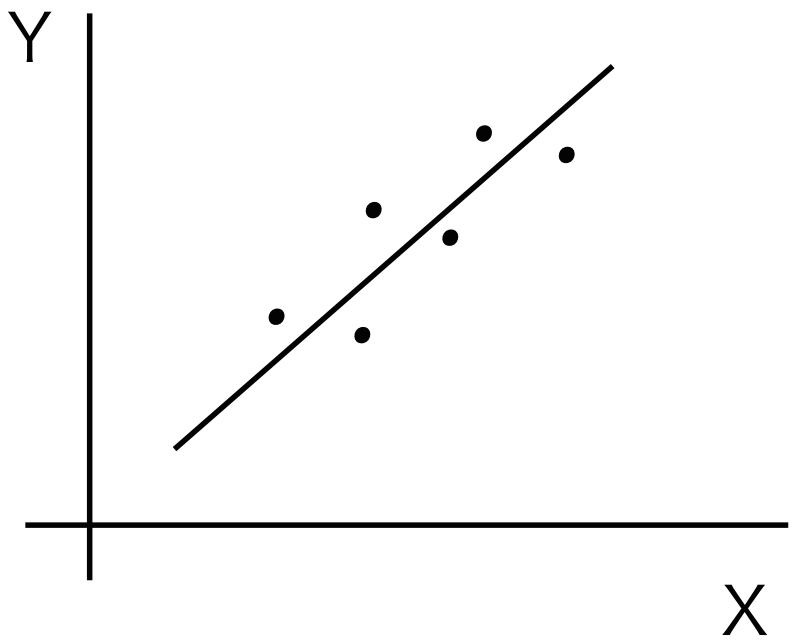
1. 重回帰分析
2. ロジスティック回帰分析
3. Cox 比例ハザード回帰（生存分析）

などがある.

単回帰分析

X, Y を連続変数とし，X, Y の散布図は直線的な分布であるとする.

散布の様子をもっともよく表す直線を回帰直線という.



回帰直線

回帰直線により Y の値を X の値から推定できる.

Table: 2 変数のデータ

No	1	2	$\cdots$	n
X	x_1	x_2	$\cdots$	x_n
Y	y_1	y_2	$\cdots$	y_n

標本回帰直線の式 $y = \alpha + \beta x$

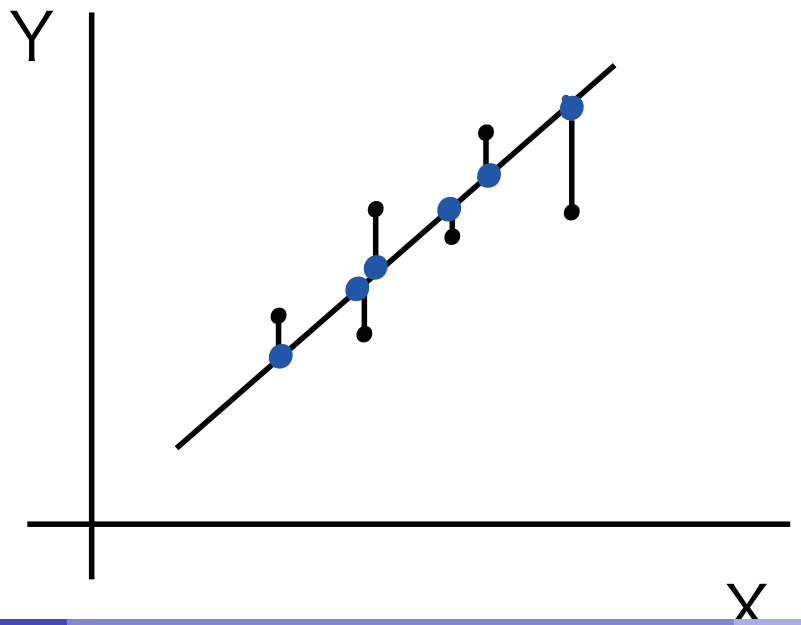
$$\alpha = \bar{Y} - \frac{s_{X,Y}}{s_X^2}\bar{X}, \qquad \beta = \frac{s_{X,Y}}{s_X^2} \text{ (標本回帰係数)}$$

$s_{X,Y}$ は X, Y の共分散

回帰直線の性質

$\hat{y}_i = \alpha + \beta x_i$ (予測値)，$e_i = y_i - \hat{y}_i$ (残差) とおく

標本回帰直線は残差平方和 $\sum_{i=1}^{n} e_i^2$ を最小にするように定めた直線である．このような方法を最小自乗法という．



回帰直線の検定

$y = A + Bx$ を母集団の回帰直線（母回帰直線）とする.

$B = 0$ (直線の傾き 0) ならば, X は Y を説明する変数ではないといえる.

そこで $B = 0$ かどうか検定したい.

仮定

各残差 e_i は正規分布 $N(0, \sigma^2)$ に従うとする.

検定の手順

1. 帰無仮説 (H_0) : 母回帰係数 $B = 0$

   対立仮説 (H_1) : 母回帰係数 $B \neq 0$

2. 帰無仮説のもとで,

統計量 $T = \dfrac{\beta}{\dfrac{S}{\sqrt{S_{XX}}}}$ は自由度 $n-2$ の t 分布に従う.

ここで、$S = \sqrt{\dfrac{1}{n-2}\sum_{i=1}^{n} e_i^2}$,

S_{XX} は X の偏差平方和, β は標本回帰係数

判定

- $|T| \geq t_{n-2}(\alpha)$ (p 値 $\leq \alpha$) ならば帰無仮説は棄却. したがって $B \neq 0$.
- $|T| < t_{n-2}(\alpha)$ (p 値 $> \alpha$) ならば $B \neq 0$ とはいえない.

回帰係数の信頼区間

$$\beta - t_{n-2}(\alpha)\frac{S}{\sqrt{S_{XX}}} \leq B \leq \beta + t_{n-2}(\alpha)\frac{S}{\sqrt{S_{XX}}}$$

信頼区間に 0 が含まれていなければ有意に $B \neq 0$ といえる.

EZR にて

regression.csv を開く.

EZR: 統計解析 → 連続変数の解析 →

線形回帰

R: summary(lm(data$weight ~ data$height))

R Commander: 統計量 – > モデルへの適合 – > 線形回帰...

重回帰分析

今までは説明変数 X はひとつであった（単回帰分析）

複数の説明変数 $X_1, X_2, \ldots, X_r$ を考える場合の回帰分析を重回帰分析という．重回帰分析では

多重線形モデル

$$Y = A + B_1 X_1 + \cdots + B_r X_r + \epsilon$$

を考える．

ここで ϵ は残差を表す変数で正規分布 $N(0, \sigma)$ に従うと仮定する．

重回帰係数

$X_1, \cdots, X_r$ の標本 x_{ij} から最小自乗法を用いて標本重回帰係数 $\alpha, \beta_1, \ldots, \beta_r$ が求まる. （実際の計算は複雑であるので統計ソフトが必要）

$$\hat{y}_i = \alpha + \beta_1 x_{i1} + \cdots + \beta_r x_{ir}$$

とおく（予測値）

決定係数

y_i と $\hat{y}_i$ の相関係数 R を重相関係数という.

重相関係数

$$R = \frac{\sum_i (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_i (y_i - \bar{y})^2}\sqrt{\sum_i (y_i - \bar{\hat{y}})^2}}$$

ここで, $\bar{y}$ は y_i の平均, $\bar{\hat{y}}$ は $\hat{y}_i$ の平均

決定係数（寄与率）

- $0 \leq R \leq 1$ であり,
- R^2 の値を回帰モデルの決定係数または寄与率といい, 回帰モデルの当てはまりの良さを表す.
- $X_1, \ldots, X_r$ により Y の $100R^2\%$ を説明していると解釈できる.
- R^2 が 1（100%）に近いほどよいモデルである.

重回帰係数の検定

- 重回帰係数 B_i（or β_i）は X_i の変化が Y に与える影響を表している.
- $B_i = 0$ なら，X_i は Y を説明する変数ではない（Y と関連のない変数）
- $B_i = 0$ かどうかの検定が可能
- 検定の結果，$B_i \neq 0$ と結論される変数が Y を説明する変数と考えられる．(不必要な変数を減らす)

EZR にて

再度，regression.csv をひらく．

EZR: 統計解析 → 連続変数の解析 →

線形回帰

R: summary(lm(formula = record..min. ~

body.fat.ratio + height + max.VO2 + subcutaneous.fat +

weight, data = data))

ロジスティック回帰

1948 年から開始されたフラミンガム研究

（Framingham Heart Study = 冠状動脈性疾患に関する大規模なコホート研究）でロジスティック回帰分析が使われ，多重リスクファクターの概念が成立した.

注

Cox 回帰と違いロジスティック回帰は時間経過に関する情報は利用されない．当時はまだ Cox 回帰の手法が一般的ではなかった.

リスク比（復習）

コホート研究など		結果（転帰）	
		あり (X)	なし (Y)
リスクファクター	あり (A)	a	b
	なし (B)	c	d

- リスク：$p = P(X|A) = \dfrac{a}{a+b}, \quad q = P(X|B) = \dfrac{c}{c+d}$
- リスク比（相対リスク）：$RR = \dfrac{p}{q} = \dfrac{a/(a+b)}{c/(c+d)}$
- 95%信頼区間

$$\exp\left(\log RR \pm 1.96\sqrt{\frac{1}{a}+\frac{1}{b}-\frac{1}{a+c}-\frac{1}{b+d}}\right)$$

オッズ比（復習）

ケース・コントロール研究など		結果（転帰）	
		case 群	control 群
リスクファクター	あり (A)	a	b
	なし (B)	c	d

- オッズ：$odds = \dfrac{p}{1-p}$
- p が小さいとき，$odds \fallingdotseq p$
- オッズ比：　$OR = \dfrac{ad}{bc}$

オッズ比（復習）

- ケース・コントロール研究ではRRを直接計算できないが，ORは計算できる.

  p が小さいとき，$OR \fallingdotseq RR$ となるので，リスク比の代用としてORが使える.

- 95%信頼区間　$\exp\left(\log RR \pm 1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right)$

ロジスティック回帰分析

(多重) ロジスティック回帰

- 応答変数 Y が 2 値のときに用いられる.

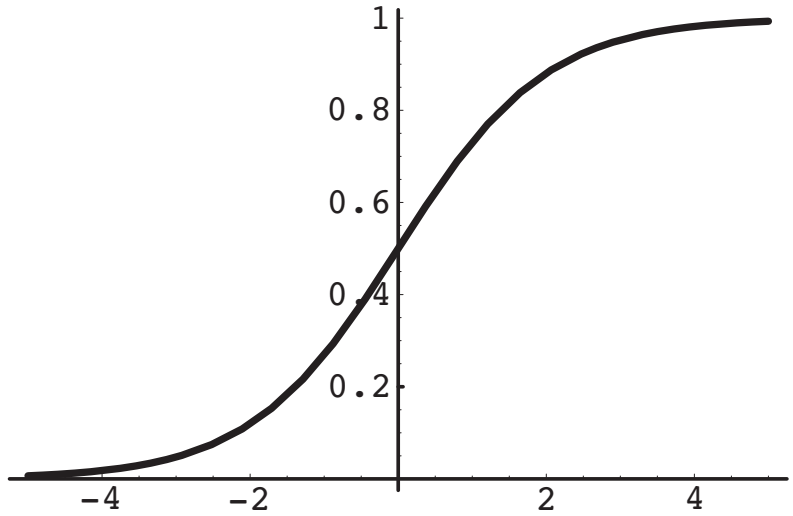
  （$Y = 1$ (あり)，$Y = 0$ (なし）)

  説明変数 X_i は連続，離散どちらでもよい．また交互作用項があってもよい.

- $p : Y = 1$ である確率
- モデル：$\log(\frac{p}{1-p}) = \alpha + \beta_1 X_1 + \cdots + \beta_r X_r$

  したがって $p = \frac{1}{1+\exp(-(\alpha+\beta_1 X_1+\cdots+\beta_r X_r))}$

  （変数は 1 つでもよい．）



ロジスティック関数 $y = \dfrac{1}{1+e^{-z}}$

- $\frac{p}{1-p}$ はオッズ，$\log(\frac{p}{1-p})$ はロジットといわれる.
- 標本から最尤法により α, β_i が（統計ソフトにより）求められる.

係数β_iの解釈

- たとえばX_1が1増加したときを考える.

$$\log \frac{q}{1-q} = \alpha + \beta_1(x_1 + 1) + \cdots + \beta_r x_r.$$

$$\log \frac{q}{1-q} - \log \frac{p}{1-p} = \beta_1 \quad \Longrightarrow \quad \log \frac{q/(1-q)}{p/(1-p)} = \beta_1$$

$$\Longrightarrow \quad \frac{q/(1-q)}{p/(1-p)} = \exp(\beta_1)$$

- $\exp(\beta_i)$を調整オッズ比という（x_iのみが1単位増加したときのオッズ比）.
- β_iは（調整）対数オッズ比.

モデル構築に対する注意

変量 X_1, X_2 の間に強い相関があるとき，多変量モデルに X_1 と X_2 を入れると正しい結果が得られない場合がある．これはオーバーマッチングと言われ，避けなければならない．そのために，X_1 または X_2 の一方をモデルに組み込めばよい．

変数選択の一般的な手順

1. 変数間の関連性を調べ，関連性が弱い変数の組を選ぶ.
2. モデルの式に当てはめ，係数を推定する.
3. 得られた係数を吟味し，不必要な変数は取り除く.
4. 最終的に得られたモデルを用いて，各種の推定，検定を行う.

モデルの評価

統計モデルの良さを評価する基準として AIC や BIC などがある.

AIC：赤池情報量基準（Akaike's Information Criterion)

BIC：ベイズ情報量基準（Bayesian Information Criterion)

$$AIC = -2\log L + 2p$$

$$BIC = -2\log L + p\log n$$

L: 最大尤度, p: 自由パラメーターの数

AIC or BIC が小さいほどよいモデル.

EZR にて

logistic.csv を開く.

低体重児のリスクファクターに関するケース・コントロール研究 Hosmer ー Lemeshow（1989）のデータの一部

統計解析　→ 名義変数の解析 → 二値変数に対する多変量解析（ロジスティック回帰）

レポートについて

成績評価のためのレポート問題３題がホームページ
http://iku3.webcrow.jp/Med_Stat/ にあります.

提出期限：2018 年 1 月 9 日
長崎宛（nagasaki@koto.kpu-m.ac.jp）