

## L04 線形回帰モデル - 重回帰

樋口さぶろお <https://hig3.net>

龍谷大学 先端理工学部 数理・情報科学課程

多変量解析☆演習 L04(2021-10-21 Thu)

最終更新: Time-stamp: "2021-10-20 Wed 14:17 JST hig"

### 今日の目標

- 重回帰と偏回帰係数の意味を説明できる
- 重回帰を statsmodels で実行できる
- 重回帰をベクトルと行列を使って説明できる
- 多重共線性の危険性と除去方法を説明できる
- 重回帰の変数選択の方法を説明できる



## L02-Q1

## Quiz 解答:正規分布のモーメントを利用した定積分の計算

- $$C = \sqrt{2\pi(1/2)^2} \cdot \int_{-\infty}^{+\infty} x^0 \cdot \frac{1}{\sqrt{2\pi(1/2)^2}} e^{-\frac{1}{2(1/2)^2}(x-6)^2} e^{72} dx = \sqrt{2\pi(1/2)^2} e^{72}.$$
- $$\sqrt{2\pi(1/2)^2} \cdot \int_{-\infty}^{+\infty} x^1 \cdot \frac{1}{\sqrt{2\pi(1/2)^2}} e^{-\frac{1}{2(1/2)^2}(x-6)^2} e^{72} dx = \sqrt{2\pi(1/2)^2} e^{72} = 6C.$$
- $$\sqrt{2\pi(1/2)^2} \cdot \int_{-\infty}^{+\infty} x^2 \cdot \frac{1}{\sqrt{2\pi(1/2)^2}} e^{-\frac{1}{2(1/2)^2}(x-6)^2} e^{72} dx = \sqrt{2\pi(1/2)^2} e^{72} = ((1/2)^2 + 6^2)C.$$

## L02-Q2

## Quiz 解答:2次元正規分布

## ① 指数関数の引数を平方完成する.

$$\begin{aligned}
 & -x^2 - 4x - 2y^2 + 12y - 5 \\
 &= -\frac{1}{2}(2x^2 + 8x + 4y^2 - 24y + 10) \\
 &= -\frac{1}{2}\left(\frac{(x+2)^2}{(1/\sqrt{2})^2} - 8 + \frac{(y-3)^2}{(1/2)^2} - 36 + 10\right)
 \end{aligned}$$

より,  $E[X] = -2, E[Y] = 3, V[X] = 1/2, V[Y] = 1/4$ .  $f(x, y)$  は  $x$  部分と  $y$  部分の積だから  $X, Y$  は独立であり,  $\text{Cov}[X, Y] = 0$ .

②  $f(x, y)$  は次の2つの形に書ける.

$$C \cdot e^{-\frac{(x+2)^2}{2(1/\sqrt{2})^2} - \frac{(y-3)^2}{2(1/2)^2}} e^{-\frac{1}{2}(-34)} = \frac{1}{\sqrt{2\pi(1/2)}} \frac{1}{\sqrt{2\pi(1/4)}} e^{-\frac{(x+2)^2}{2(1/\sqrt{2})^2} - \frac{(y-3)^2}{2(1/2)^2}}$$

$$\text{よって, } C = \frac{1}{\sqrt{2\pi(1/2)}} \frac{1}{\sqrt{2\pi(1/4)}} e^{-17}.$$

## L02-Q3

## Quiz 解答:2次元正規分布

- ① 指数関数の引数を平方完成する.

$$\begin{aligned} & -4x^2 - \frac{1}{6}y^2 + 2y \\ &= -\frac{1}{2}(8x^2 + \frac{1}{3}y^2 - 4y) \\ &= -\frac{1}{2}\left(\frac{(x-0)^2}{(1/2\sqrt{2})^2} + \frac{(y-6)^2}{(\sqrt{3})^2} - 12\right) \end{aligned}$$

より,  $E[X] = 0, E[Y] = 6, V[X] = \frac{1}{8}, V[Y] = 3, \text{Cov}[X, Y] = 0$ .

- ②  $f(x, y)$  は次の2つの形に書ける.

$$C \cdot e^{-\frac{(x-0)^2}{2(1/2\sqrt{2})^2} - \frac{(y-6)^2}{2(\sqrt{3})^2}} e^{-\frac{1}{2}(-12)} = \frac{1}{\sqrt{2\pi(1/8)}} \frac{1}{\sqrt{2\pi \cdot 3}} e^{-\frac{(x-0)^2}{2(1/2\sqrt{2})^2} - \frac{(y-6)^2}{2(\sqrt{3})^2}}$$

であるべきだから,  $C = \frac{1}{\sqrt{2\pi(1/8)}} \frac{1}{\sqrt{2\pi \cdot 3}} e^{-6}$ .

## L02-Q4

## Quiz 解答:2次元正規分布の確率密度関数

- ①  $E[X] = -4, E[Y] = 5.$   
 $V[X] = 2, V[Y] = 3, \text{Cov}[X, Y] = -1.$
- ②  $\det \Sigma = 5, \Sigma^{-1} = \frac{1}{5} \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}$  より,

$$f(x, y) = \frac{1}{(2\pi)^{2/2} \cdot 5} e^{-\frac{1}{2} \frac{3}{5} (x+4)^2 - \frac{1}{2} \cdot 2 \cdot \frac{2}{5} (x+4)(y-5) - \frac{1}{2} \frac{2}{5} (y-5)^2}$$

L02-Q5

Quiz 解答:2次元正規分布

$$E[X] = 2.$$

$$E[Y] = -3.$$

$$\Sigma = (\Sigma^{-1})^{-1} = \begin{pmatrix} 4 & -6 \\ -6 & 14 \end{pmatrix}^{-1} = \frac{1}{20} \begin{pmatrix} 14 & +6 \\ +6 & 4 \end{pmatrix}$$

$$V[X] = \frac{14}{20}, \text{Cov}[X, Y] = \frac{6}{20}, V[Y] = \frac{4}{20}.$$

L02-Q6

Quiz 解答:2次元正規分布

- ① 指数関数の引数の2次式を対称行列を用いて書く.  $x, y$  の1次の項はないので,  $X, Y$  の母平均値は0であり,

$$-x^2 + 4xy - 7y^2 = -\frac{1}{2} \begin{pmatrix} x-0 & y-0 \end{pmatrix} \begin{pmatrix} 2 & -4 \\ -4 & 14 \end{pmatrix} \begin{pmatrix} x-0 \\ y-0 \end{pmatrix}$$

よって,

$$\Sigma^{-1} = \begin{pmatrix} 2 & -4 \\ -4 & 14 \end{pmatrix}, \quad \Sigma = \frac{1}{2 \cdot 14 - (-4)(-4)} \begin{pmatrix} 14 & 4 \\ 4 & 2 \end{pmatrix}.$$

- ②  $f(x, y) = Ce^{-x^2+4xy-7y^2} = \frac{1}{2\pi\sqrt{\det \Sigma}} e^{-\frac{1}{2} \mathbf{x} \Sigma^{-1} \mathbf{x}}$ . よって  
 $C = \frac{1}{2\pi\sqrt{1/12}}$ .

L02-Q7

Quiz 解答:2次元正規分布

- ① 指数関数の引数の2次式を対称行列を用いて書く.  $x, y$  の1次の項はないので,  $X, Y$  の母平均値は0であり,

$$-2x^2 - xy - \frac{1}{2}y^2 = -\frac{1}{2} \begin{pmatrix} x-0 & y-0 \end{pmatrix} \begin{pmatrix} 4 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x-0 \\ y-0 \end{pmatrix}$$

よって,

$$\Sigma^{-1} = \begin{pmatrix} 4 & 1 \\ 1 & 1 \end{pmatrix}, \quad \Sigma = \frac{1}{4 \cdot 1 - 1 \cdot 1} \begin{pmatrix} 1 & -1 \\ -1 & 4 \end{pmatrix}.$$

- ②  $f(x, y) = Ce^{-2x^2 - xy - \frac{1}{2}y^2} = \frac{1}{2\pi\sqrt{\det \Sigma}} e^{-\frac{1}{2} \mathbf{x}^t \mathbf{V}^{-1} \mathbf{x}}$ . よって  
 $C = \frac{1}{2\pi\sqrt{1/3}}$ .



## ここまで来たよ

### 2 次元正規分布

### 4 線形回帰モデル - 重回帰

- 重回帰
- 線形回帰の行列とベクトルによる定式化
- 多重共線性
- 重回帰での変数選択

## 線形回帰 (重回帰) linear multiple regression モデルとは

永田棟方 多変量解析法入門 §5

このドーナツ製造機で作るドーナツの重さ  $Y$  は、温度  $x_1$  と水の量  $x_2$  によるらしい。次の**線形回帰モデル** (重回帰モデル) を仮定する。

$$Y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

$Y, \epsilon$ : 連続型確率変数,  $\beta_0, \beta_1, \beta_2$ : **偏回帰係数**,  $\sigma > 0$ : 定数 = パラメータ = 母数

$Y$ : **目的変数** (従属変数) 確率変数

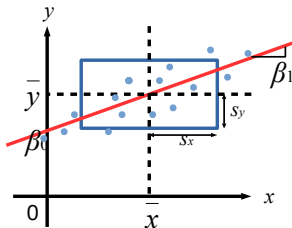
$x_1, x_2, \dots, x_p$ : **説明変数** (独立変数) 確率変数でない

一般の  $p$ : **重回帰** multiple regression

$p = 1$ : **単回帰** simple regression

**重回帰は、ローテクだがとりあえず使えて、超強力な機械学習の予測手法**

- $\beta_j$ :  $x_j$  を 1 増やしたとき (単位による)  $y$  が増える量.
- $x_2 = x_1^2$  とか  $x_2 = e^{x_1}$  とか  $x_3 = x_1 x_2$  でもよい (多項式回帰). 関数の意味で線形独立であれば  $\rightsquigarrow$  多重共線性  $\times x_2 = 3x_1$
- $x_2 = \text{if(性別==女) return 1; else return 0}$  でもよい (真偽値に対するダミー変数). 2 群で定数項  $\beta_0$  に差をつけることに相当.
- $x_1, x_2$  両方が  $\text{○○}$  なとき  $\dots \rightarrow$  交互作用



## ソフトウェアで偏回帰係数を推定

```
import statsmodels.formula.api as smf
```

重回帰なら formula='y ~ x1 + x2', formula='y ~ x1 + I(x1\*x1)'

青い表の行で,  $\beta_0, \beta_1$  に  $\beta_2, \dots, \beta_p$  と加わる

```
[ ] 1 formula='height ~ weight' # formula 表記. ~ の左辺を目的変数, 右辺を説明変数として線形回
2 result=smf.ols(formula, body).fit() # ols = 普通の最小二乗法で fit せよ
3 result.summary() # result に書き込まれた結果を取り出す
```

目的変数は height		OLS Regression Results		決定係数R (Adjusted 自由度調整済)	
Dep. Variable:	height	R-squared:	0.905		
Model:	OLS	Adj. R-squared:	0.904		
Method:	Least Squares	F-statistic:	935.3		
Date:	Sat, 09 Oct 2021	Prob (F-statistic):	6.31e-52		
Time:	00:59:55	Log-Likelihood:	-255.30		
No. Observations:	100	AIC:	514.6		
Df Residuals:	98	BIC:	519.8		
Df Model:	1				

	coef	std err	t	P> t	[0.025	0.975]
Intercept	129.4555	1.122	115.412	0.000	127.230	131.681
weight	0.8233	0.027	30.582	0.000	0.770	0.877

Omnibus: 1.147 Durbin-Watson: 2.218  
 Prob(Omnibus): 0.564 Jarque-Bera (JB): 1.053  
 Skew: -0.063 Prob(JB): 0.591  
 Kurtosis: 2.514 Cond. No. 149.

## 単回帰の $\beta_1$ と重回帰の偏回帰係数 $\beta_1$ は等しい?

データ  $(y_i, x_{i1}, x_{i2}) \xrightarrow{\text{推定}} (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$

データ  $(y_i, x_{i1}, x_{i2}) \xrightarrow{x_{i2}\text{捨てる}} (y_i, x_{i1}) \xrightarrow{\text{推定}} (\hat{\beta}_0, \hat{\beta}_1)$  .

同じ  $\hat{\beta}_1$  ?

前者の  $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) \xrightarrow{\hat{\beta}_2\text{捨てる}} (\hat{\beta}_0, \hat{\beta}_1)$  ってやっていい?  $\rightarrow$  No.

## ここまで来たよ

### 2 次元正規分布

### 4 線形回帰モデル - 重回帰

- 重回帰
- 線形回帰の行列とベクトルによる定式化
- 多重共線性
- 重回帰での変数選択

# 行列とベクトルによる定式化

永田棟方 多変量解析法入門 §5.4

$x_{ij}$  データ番号  $i = 1, \dots, n$ , 説明変数番号  $j = 1, \dots, p$ .

$$y_1 = \alpha_0 + \beta_1(x_{11} - \bar{x}_1) + \beta_2(x_{12} - \bar{x}_2) + \epsilon_1$$

$$\vdots$$

$$y_n = \alpha_0 + \beta_1(x_{n1} - \bar{x}_1) + \beta_2(x_{n2} - \bar{x}_2) + \epsilon_n$$

は,  $\epsilon_i$  は独立より  $\text{Cov}[\epsilon_i, \epsilon_j] = 0$  ( $i \neq j$ ) で,

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I_n) \quad (n \text{次元正規分布})$$

とまとめられる.

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} - \bar{x}_1 & x_{21} - \bar{x}_2 \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} - \bar{x}_1 & x_{2n} - \bar{x}_2 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \alpha_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$\hat{\beta}$  を推定値, 残差 (縦) ベクトル  $e = y - X\hat{\beta}$  とおく.  
 行列  $M$ , ベクトル  $y$  の転置 (transpose) を  ${}^tM$ ,  ${}^ty$  と書く.  
 残差の 2 乗和は,

$$\begin{aligned} S(\beta_0, \beta_1, \beta_2) &= |e|^2 = {}^te e = {}^t(y - X\hat{\beta})(y - X\hat{\beta}) \\ &= {}^ty y - 2 {}^t\hat{\beta} {}^tX y + {}^t\hat{\beta} {}^tX X \hat{\beta} \end{aligned}$$

この最小化.  $\frac{\partial S}{\partial \beta} = \nabla_{\beta} S = 0$

ベクトル解析

$$\frac{\partial S}{\partial \alpha_0} = 0 \rightsquigarrow 0 - 2[1 \cdots 1]y + 2[1 \cdots 1]X\hat{\beta} = 0$$

$$\frac{\partial S}{\partial \beta_1} = 0 \rightsquigarrow 0 - 2[x_{11} - \bar{x}_1 \cdots x_{n1} - \bar{x}_2]y + 2[x_{11} - \bar{x}_1 \cdots x_{n1} - \bar{x}_1]X\hat{\beta} = 0$$

$$\frac{\partial S}{\partial \beta_2} = 0 \rightsquigarrow 0 - 2[x_{12} - \bar{x}_1 \cdots x_{n2} - \bar{x}_2]y + 2[x_{12} - \bar{x}_2 \cdots x_{n2} - \bar{x}_2]X\hat{\beta} = 0$$

あわせて

$$-2 {}^tX y + 2 {}^tX X \hat{\beta} = \mathbf{0}.$$



$$-2 {}^tX\mathbf{y} + 2 {}^tXX\hat{\boldsymbol{\beta}} = \mathbf{0}.$$

$$\hat{\boldsymbol{\beta}} = ({}^tXX)^{-1} {}^tX\mathbf{y}.$$

$${}^tXX = \begin{bmatrix} n & 0 & 0 \\ 0 & ns_{11} & ns_{12} \\ 0 & ns_{12} & ns_{22} \end{bmatrix}, s_{jk} = \frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) \text{ から, } \hat{\boldsymbol{\beta}} \text{ を}$$

具体的に求められる.

## ここまで来たよ

### 2 次元正規分布

### 4 線形回帰モデル - 重回帰

- 重回帰
- 線形回帰の行列とベクトルによる定式化
- **多重共線性**
- 重回帰での変数選択

## 多重共線性 multicollinearity

永田棟方 多変量解析法入門 p.67

- $X^T X$  の逆行列がなかったら?
- $\Leftrightarrow \hat{\beta}$  の式で分母が zero だったら?
- $\Leftrightarrow$  データ  $(x_1, x_2, y)$  を  $(x_1, x_2)$  平面上に射影したとき、データが一直線上にあったら?

$\rightsquigarrow \hat{\beta}$  (平面) が決まらない

典型的なケース  $(x_1, x_2, y) = (10, 20, ?), (20, 40, ?), (16, 32, ?), \dots$

$x_1 = x_2$  なら,  $y = 5x_1 + 0x_2, y = 0x_1 + 5x_2, y = 2x_1 + 3x_2, \dots$  を区別する理由がない.

- $n$  次元ベクトル  $\mathbf{x}_1, \mathbf{x}_2$  が比例してる… 直線上の情報しかない.
- $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p$  が  $p$  次元部分空間を張ってない.
- 身長と足のサイズを説明変数にいたけど, 実質的に比例してた, とか.
- 数値計算的には「行列式が zero に近い」だけで問題になる.

## ここまで来たよ

### 2 次元正規分布

### 4 線形回帰モデル - 重回帰

- 重回帰
- 線形回帰の行列とベクトルによる定式化
- 多重共線性
- 重回帰での変数選択

## 決定係数, 寄与率

岩薩林 確率・統計 §9.2

永田棟方 多変量解析法入門 p.69

残差 平面 (予測値) からの上下  $y$  方向のずれ,

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i)$$

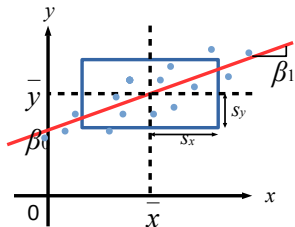
とすると, 単回帰と同じ式.

決定係数

$$R^2 = 1 - \frac{\sum_i e_i^2}{\sum_i (y_i - \bar{y})^2}$$

$0 \leq R^2 \leq 1$  で,  $1 - R^2$  が 0 に近いほど, あてはまりがよい.

→ 同じ  $p$  なら,  $R^2$  が大きくなる説明変数を選ぶのがよい.



## $p$ はいくつに取ったらいい?

永田棟方 多変量解析法入門 pp.71,81

多いほどいい. 説明変数を追加すると, 決定係数  $R^2$  も増加する.  
 それでいいの?

$p$  を 1 個増やすごとに, 1 個のデータ点を平面上に載せられる.  $p = n$  でぜんぶ載る?

機械学習のりて言えば「 $n = p$  としたら, 教師データは完全に再現できるが過学習. 汎化性能はよくない」

安上がりに (いい感じに小さい  $p$  で) いい結果 (小さい  $e^2$ , 大きい  $R^2$ ) を出したたい.

自由度調整済決定係数  $R^{*2}$  の大きさで見ると

永田棟方 多変量解析法入門 (5.39)

$$1 - R^{*2} = (1 - R^2) \times \frac{n-1}{n-1-p}.$$

× 変数を 1 個ずつ単回帰して,  $|\beta_j|$  が大きい  $x_j$  を残す

× ぜんぶの変数をまとめて重回帰して,  $|\beta_j|$  が大きい  $x_j$  を残す

○ 単回帰でいちばん  $R^{*2}$  の大きい変数をまず採用, 別の変数も加えて  $R^{*2}$  がある程度改善するなら採用. 繰り返す.

他の基準: AIC, BIC, 分散比  $F_0$ .