L04 重回帰

樋口さぶろお

龍谷大学 先端理工学部 数理・情報科学課程

理論物理学特論　L04(2021-10-19 Tue)

最終更新: Time-stamp: "2021-10-20 Wed 08:42 JST hig"

今日の目標

- 重回帰をベクトルと行列を使って説明できる
- 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
- 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

L03-Q1

Quiz 解答:2 次元正規分布の確率密度関数

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

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

L03-Q2

Quiz 解答:2 次元正規分布

$\mathrm{E}[X] = 2.$

$\mathrm{E}[Y] = -3.$

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

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

L03-Q3

Quiz 解答:回帰係数と回帰直線

$y + 4 = \frac{-25}{49} \times (x - 9).$

ここまで来たよ

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の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.

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

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

Y: 目的変数 (従属変数) 確率変数
$x_1, x_2, \ldots, x_p$: 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

- β_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, $x_2 = \mathrm{e}^{x_1}$ 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
- $x_2 =$`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 β_0 に差をつけることに相当.
- x_1, x_2 両方が◯◯なとき $\cdots \rightarrow$ 交互作用

しばらく $p = 2$ で. 永田棟方 多変量解析法入門 §5.2

p:説明変数の個数 $(j = 1, \ldots, p)$. n:データの個数 $(i = 1, \ldots, n)$. x_{ij}.

$\beta_0, \beta_1, \beta_2$: 本当の値

$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$: データ $(x_{i1}, x_{i2}, y_i)(i = 1, 2, \ldots, n)$ から求まる推定値.

回帰平面

(x_1, x_2, y) 空間内の平面をデータ点のなるべく近くにとりたい.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_1 + \hat{\beta}_2 \cdot x_2$$

ソフトウェアで推定，何がわかったか?

`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，β_0, β_1 に $\beta_2, \ldots, \beta_p$ 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 に蓄えられた結果を取り出す
```

目的変数は height

決定係数R (Adjusted 自由度調整済)

OLS Regression Results

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		
Covariance Type:	nonrobust		

データの個数 自由度

F統計量 F検定のp値 (有意確率)

t検定のp値 (有意確率)

95%信頼区間

値　標準誤差

	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

切片β0
説明変数weightの係数β1

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.

決定係数，寄与率

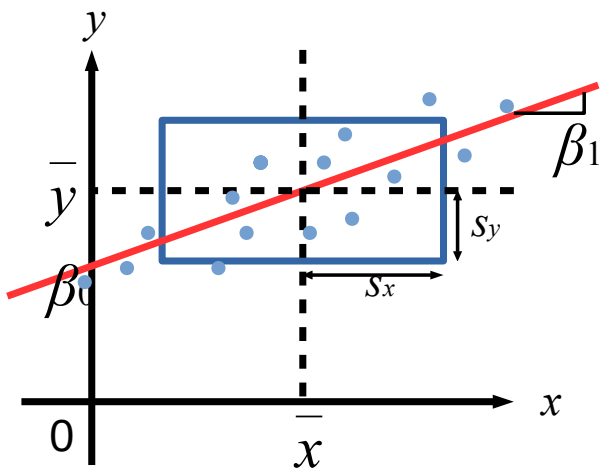
岩薩林 確率・統計 §9.2 | 永田棟方 多変量解析法入門 p.69

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

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

$0 \leq R^2 \leq 1$ で, $1 - R^2$ が 0 に近いほど, あてはまりがよい. 同じ p ならこの値で比較できる.
実は R^2 は, 2 次元データ $(\hat{y}_i - \overline{y}, y_i - \overline{y})$ の相関係数の二乗.



ここまで来たよ

単回帰の β_1 と重回帰の β_1 は等しい?

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

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

同じ $\hat{\beta}_1$?

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

ここまで来たよ

3 2 次元正規分布・単回帰

3 重回帰
- 重回帰
- 偏回帰係数と単回帰の回帰係数の違いを説明できる
- 線形回帰の行列とベクトルによる定式化
- 多重共線性
- 重回帰の変数選択の方法を説明できる

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 $i = 1, \ldots, n$, 説明変数番号 $j = 1, \ldots, p$.

$$
\begin{aligned}
y_1 =& \alpha_0 + \beta_1(x_{11} - \overline{x}_1) + \beta_2(x_{12} - \overline{x}_2) + \epsilon_1 \\
&\vdots \\
y_n =& \alpha_0 + \beta_1(x_{n1} - \overline{x}_1) + \beta_2(x_{n2} - \overline{x}_2) + \epsilon_n
\end{aligned}
$$

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

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

とまとめられる.

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

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

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

これの最小化.

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

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

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

あわせて

$$-2\,{}^{\mathrm{t}}X\boldsymbol{y} + 2\,{}^{\mathrm{t}}X X\hat{\boldsymbol{\beta}} = \boldsymbol{0}.$$

$$-2\,{}^{\mathrm{t}}X\boldsymbol{y} + 2\,{}^{\mathrm{t}}X X\hat{\boldsymbol{\beta}} = \mathbf{0}.$$

$$\hat{\boldsymbol{\beta}} = ({}^{\mathrm{t}}X X)^{-1}\,{}^{\mathrm{t}}X\boldsymbol{y}.$$

${}^{\mathrm{t}}X X = \begin{bmatrix} n & 0 & 0 \\ 0 & ns_{11} & ns_{12} \\ 0 & ns_{12} & ns_{22} \end{bmatrix}$ より, $\hat{\boldsymbol{\beta}}$ の表式を与える.

ここまで来たよ

多重共線性 multicolinearity

永田棟方 多変量解析法入門 p.67

- ${}^{\mathrm{t}}XX$ の逆行列がなかったら?
- $\Leftrightarrow \hat{\boldsymbol{\beta}}$ 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 $(x_1, x_2, y) = (10, 20, ?), (20, 40, ?), (16, 32, ?), \ldots$.
$x_1 = x_2$ なら, $y = 5x_1 + 0x_2$, $y = 0x_1 + 5x_2$, $y = 2x_1 + 3x_2$,.. を区別する理由がない.

- n 次元ベクトル $\boldsymbol{x}_1, \boldsymbol{x}_2$ が比例してる… 直線上の情報しかない.
- $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ 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れたけど, 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

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?
機械学習のりで言えば「$n = p$ 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) を出したい.
基準のひとつ 自由度調整済決定係数 永田棟方 多変量解析法入門 (5.39 R^{*2} を最大に.
$1 - R^{*2} = (1 - R^2) \times \frac{n-1}{n-1-p}$.
その他に, AIC, BIC など.