線形重回帰モデルの変数選択

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　L05(2026-10-08 Thu)

最終更新: Time-stamp: "2026-10-25 Sun 10:43 JST hig"

今日の目標

- 重回帰の多重共線性を説明できる
- 重回帰で説明変数選択の方針を説明できる
- 連続-離散同時分布，混合ガウス分布を説明できる

ここまで来たよ

5. 線形重回帰モデルの変数選択
   - 残差分析
   - 線形重回帰で使えるいろんな説明変数
   - 多重共線性
   - 重回帰の変数選択
   - 混合ガウス分布

残差

残差ベクトル $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ = 実測値 − 予測値

$$\text{残差二乗和} S_{\mathrm{e}} = \sum_i e_i e_i = \boldsymbol{e}^\top \boldsymbol{e}$$

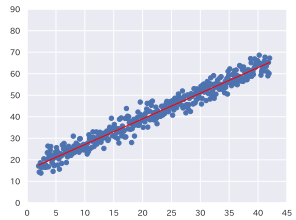
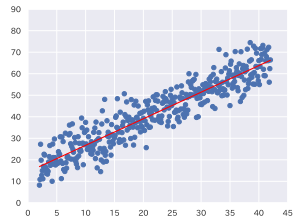
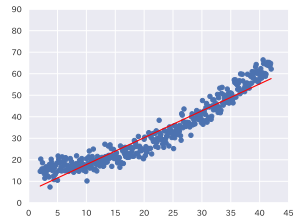
$$\begin{aligned} S_{\mathrm{t}} &= |\boldsymbol{y} - \overline{\boldsymbol{y}}|^2 \\ &= |(\boldsymbol{y} - \hat{\boldsymbol{y}}) + (\hat{\boldsymbol{y}} - \overline{\boldsymbol{y}})|^2 \\ &\overset{!}{=} |\boldsymbol{y} - \hat{\boldsymbol{y}}|^2 + |\hat{\boldsymbol{y}} - \overline{\boldsymbol{y}}|^2 \\ &= S_{\mathrm{e}} + S_{\mathrm{r}} \end{aligned}$$

- S_{t}: 全平方和（すべての y のちらばり)total
- S_{r}: 回帰平方和（ちらばりのうち，回帰平面で説明できる部分）regression
- S_{e}: 残差平方和（残り，予測値と実測値のずれ）error

残差分析　久保川 統計学入門 §12.2

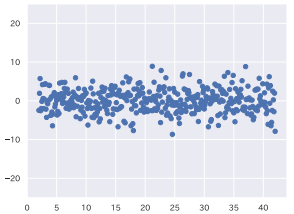
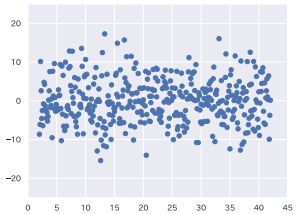
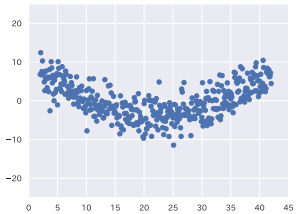
単回帰 $p = 1$ で.
$\epsilon_i = y_i - (\beta_0 + \beta_1 x_i) \sim N(0, \sigma^2)$.
残差 $e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$.

1. x 対実測値 y

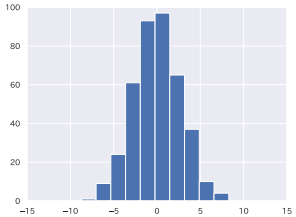
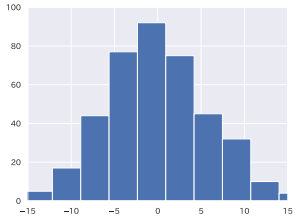
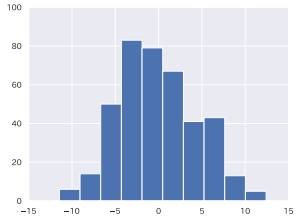




決定係数 $R^2 = 1 - \dfrac{\sum \text{残差}^2}{\sum \text{水平な直線からのずれ}^2} = 0.962$　　0.860　　0.914

2. (x, e) の散布図. x 依存性があるならモデルに改善余地.

3. 残差 e_i のヒストグラム. 正規でないならモデルに改善余地.

ここまで来たよ

5. 線形重回帰モデルの変数選択
   - 残差分析
   - 線形重回帰で使えるいろんな説明変数
   - 多重共線性
   - 重回帰の変数選択
   - 混合ガウス分布

線形重回帰で使えるいろんな説明変数

線形重回帰モデルとして次を考えた．

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

β_j の元々の意味 x_j を 1 増やしたとき y が増える量．
「線形」＝ 右辺が，$\beta_j x_j$ の形の項の和．実は説明変数 ~~x_j の~~回帰係数 β_j の線形結合 石村畑 §5.3 ．
次のような量を β_j にかけて右辺に加えても線形重回帰が可能

- $x_2 = x_1^2$ とか $x_3 = x_1 x_2$ でもよい (多項式回帰)．
- $x_4 = \frac{1}{x_1}$，$x_5 = \mathrm{e}^{x_1}$．
- $x_2 = 3x_1$ とかはだめ．関数の意味で線形独立である必要．⇝ 多重共線性

名義変数のダミー変数

(男, 女), (A,B,AB,O) のようなデータは，様々な方法確率統計 II,III で扱えるが，線形重回帰にも取り込める．

- 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 両方が真の時だけ … $\to$ 交互作用 $x_1 x_2$．
- 名義変数である血液型を，$x_2 = 0, 1, 2, 3$ (A,B,AB,O) などとするのはおかしい．ダミー変数は $x_{\mathrm{A}} = 0, 1$, $x_{\mathrm{B}} = 0, 1$, $x_{\mathrm{AB}} = 0, 1$, $x_{\mathrm{O}} = 0, 1$ とする．one-hot vector 機械学習

ここまで来たよ

5. 線形重回帰モデルの変数選択
   - 残差分析
   - 線形重回帰で使えるいろんな説明変数
   - 多重共線性
   - 重回帰の変数選択
   - 混合ガウス分布

多重共線性 multi-colinearity 久保川 統計学入門 p.249

線形重回帰で $\boldsymbol{\beta}$ は次のように推定された．$X^\top X$ は $p+1$ 次正方行列．

$$\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \boldsymbol{Y}$$

例$(p=2)$ $X^\top X = \begin{pmatrix} n & 0 & 0 \\ 0 & \sum_i (x_{i1} - \overline{x}_{\bullet 1})^2 & \sum_i (x_{i1} - \overline{x}_{\bullet 1})(x_{i2} - \overline{x}_{\bullet 2}) \\ 0 & \sum_i (x_{i2} - \overline{x}_{\bullet 2})(x_{i1} - \overline{x}_{\bullet 1}) & \sum_i (x_{i2} - \overline{x}_{\bullet 2})^2 \end{pmatrix}$

- $X^\top X$ が非正則で，逆行列が存在しなかったら？ 石村畑 §4.7
- 説明変数：$(x_{i1}, x_{i2})_{i=1,\dots,n}$ が一直線上にあったら？
- 説明変数：身長と足のサイズが実質的に比例していたら？

$\rightsquigarrow \hat{\boldsymbol{\beta}}$ が不定，回帰平面が定まらない，または解が存在しない線形代数 I
$2x_1 = x_2$ なら，$y = 2 + 6x_1 + 0x_2$，$y = 2 + 0x_1 + 3x_2$，
$y = 2 + 4x_1 + x_2, \dots$ を区別する理由がない．
数値計算では「$X^\top X$ の行列式が zero **に近い**」，ほぼ一直線上，だけで問題

説明変数の個数 p に対してデータ n が足りないケース

- 高い目標に練習が追いついてない
- $\operatorname{rank} X < p+1$ だと $X^\top X$ が非正則
- $X^\top X$ の対角成分は x_i の分散だから非負．各データ $(x_{i1}, \ldots, x_{in})$ に相関が強いと（極端なケース直線上にあると）rank が落ちる危険．

L05-Q1

Quiz(多重共線性)

1. 3 次元空間内の平面 P_1 上に，点 $(x_1, x_2, y) = (0,0,6), (2,0,0), (0,3,0)$ がある．P_1 を求めよう．
2. 3 次元空間内の平面 P_3 上に，点 $(x_1, x_2, y) = (0,0,6), (1,2,3), (2,4,0)$ がある．P_3 を求めよう．

略解

1. $y = 6 - 3x_1 - 2x_2$
2. $y = 6 - x_1 + x_2 + a(2x_1 - x_2)$, 任意定数 $a \in \mathbb{R}$.

ここまで来たよ

5. 線形重回帰モデルの変数選択
   - 残差分析
   - 線形重回帰で使えるいろんな説明変数
   - 多重共線性
   - 重回帰の変数選択
   - 混合ガウス分布

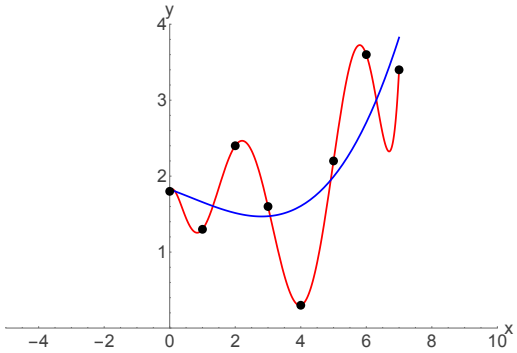
説明変数の個数 p はいくつに取ったらいい?

p が大きいといいこと

- 説明変数を追加すると, 実は常に決定係数 R^2 が増加する.
- p を増やして $p = n$ にすると，すべてのデータ点を平面上に載せられるちゃう.

p が大きいとまずいこと

- 単純な説明になってない（少ない説明変数できれいに説明したい）
- 多重共線性に陥る危険が増す
- $n = p$ では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機械学習



$\rightsquigarrow$ 適度に小さい p で，いい結果 (小さい e^2, 大きい R^2) を出したい

自由度調整済み決定係数 R^{*2} 久保川 統計学入門 p.248

バランスをとった改善方向

自由度調整済決定係数 R^{*2} を大きくする．

定義 (自由度調整済み決定係数 R^{*2} 久保川 統計学入門 p.248)

$$R^{*2} = 1 - \frac{S_{\mathrm{e}}/(n-p-1)}{S_{\mathrm{t}}/(n-1)}$$

n が大きいと大きいけど，p を増やすとペナルティがある．説明変数 p の個数が少ないほど，よいモデルである，ことを反映した指標．$n-p-1$, $n-1$ は自由度

変数選択：具体的にどの説明変数をいれる?残す?

説明変数のセット（モデル）を様々試し，R^{*2} が大きくなるモデルを選択する

一般に，

- 多くある説明変数候補のうち，どのセットを選ぶか（変数選択），
- そもそもどのモデルを選ぶか（線形重回帰でいいのか？）

は重要な問題．
指標が大きい（小さい）ものを選ぶ，という形で定式化される．指標は，予測の正確さと，変数の少なさのジレンマのバランスをとるもの．次のようなものが使われる．

- 自由度調整済み決定係数の最大化 R^{*2}
- 赤池情報量基準 AIC の最小化（線形重回帰の場合）
  $-2\log L + 2p = n\log(1-R^2) + 2(p+2) + C$
- ベイズ情報量基準 BIC の最小化 (線形重回帰の場合)
  $-2\log L + p\log n = n\log(1-R^2) + (p+2)\log n$

L は尤度．

ここまで来たよ

5. 線形重回帰モデルの変数選択
   - 残差分析
   - 線形重回帰で使えるいろんな説明変数
   - 多重共線性
   - 重回帰の変数選択
   - 混合ガウス分布

復習：同時分布と周辺分布

久保川 統計学入門 §6.1 | 確率統計 I(2025)L06

同時分布 X, Y 確率変数

$p(x,y) = P(X = x \text{かつ} Y = y) =$

$y \backslash x$	150	160	165	計
45	3/8	0	1/12	
50	1/8	1/3	1/12	
計				

周辺分布 X, Y の周辺分布 $p_{\mathrm{X}}(x)$,$p_{\mathrm{Y}}(y)$ は,

離散型$p_{\mathrm{X}}(x) = p(x,\bullet) = \sum_y p(x,y), \quad p_{\mathrm{Y}}(y) = p(\bullet,y) = \sum_x p(x,y)$

条件付き分布

命題 (条件付き確率を同時確率分布と周辺分布で表す式

久保川 統計学入門 定義 6.9**)**

$$P(X = x|Y = y) = p_{\mathrm{X|Y}}(x|y) = \frac{p(x,y)}{p_{\mathrm{Y}}(y)},$$

離散-連続型確率変数

確率変数の組 (X, Y) を考える.
X: 連続型確率変数,
Y: 離散型確率変数, $Y = y_1, \cdots, y_n$.
次を満たす確率密度関数 $f(x, y)$ で指定される.

定義 (離散-連続混合型分布)

$$P(a < X \leq b, Y = y_i) = \int_a^b f(x', y_i)\, \mathrm{d}x'.$$

$$P(a < X \leq b, y_i \leq Y \leq y_j) = \int_a^b \sum_{y_i \leq y' \leq y_j} f(x', y')\, \mathrm{d}x'$$

気分を出すために, 表や場合分けで $f(x,y)$. ただし, $Y=0,1$. x についての確率密度関数 $h_0(x), h_1(x)$.

$$f(x,y)=\begin{cases} h_0(x) & (y=0) \\ h_1(x) & (y=1) \\ 0 & (y\neq 0,1) \end{cases}$$

$y\backslash x$	$-\infty<x<+\infty$	
0	$h_0(x)$	
1	$h_1(x)$	

全 $=\int_{-\infty}^{+\infty} h_0(x)\,\mathrm{d}x+\int_{-\infty}^{+\infty} h_1(x)\,\mathrm{d}x=1.$

周辺分布は？ 条件付き分布は？

混合ガウス分布 GMM=Gaussian mixture model

正規分布 normal distribution = ガウス分布 Gaussian distribution
混合する = mix
混合ガウス分布を定義する準備として, 次の離散-連続混合型分布 $f(x,y)$ を考える.
$\pi_0+\pi_1=1, \pi_y\geq 0, \pi_{他}=0$ とする.

$$f(x,y)=\pi_y \frac{1}{(2\pi\sigma_y^2)^{1/2}}\mathrm{e}^{-\frac{(x-\mu_y)^2}{2\sigma_y^2}} = \begin{cases} \pi_0\cdot\frac{1}{(2\pi\sigma_0^2)^{1/2}}\mathrm{e}^{-\frac{(x-\mu_0)^2}{2\sigma_0^2}} & (y=0) \\ \pi_1\cdot\frac{1}{(2\pi\sigma_1^2)^{1/2}}\mathrm{e}^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} & (y=1) \\ 0 & (y\neq 0,1) \end{cases}$$

$$=\delta_{y0}\pi_0\frac{1}{(2\pi\sigma_0^2)^{1/2}}\mathrm{e}^{-\frac{(x-\mu_0)^2}{2\sigma_0^2}}+\delta_{y1}\pi_1\frac{1}{(2\pi\sigma_1^2)^{1/2}}\mathrm{e}^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}.$$

Y の周辺分布は離散型でベルヌイ分布 $B(1, \pi_1)$

$$f_{\mathrm{Y}}(y) = \int f(x, y)\mathrm{d}x = \int_{-\infty}^{+\infty} \pi_y \cdot \frac{1}{(2\pi\sigma_y^2)^{1/2}} \mathrm{e}^{-\frac{(x-\mu_y)^2}{2\sigma_y^2}} \mathrm{d}x = \pi_y.$$

X の周辺分布は連続型で GMM

定義 (GMM=混合ガウス分布)

連続型確率変数 X がパラメタ $(\pi_0, \pi_1, \mu_0, \mu_1, \sigma_0^2, \sigma_1^2)$ の GMM に従うとは, 次の確率密度関数にしたがうことをいう.

$$f_{\mathrm{X}}(x) = \sum_y f(x, y) = \pi_0 \cdot \frac{1}{(2\pi\sigma_0^2)^{1/2}} \mathrm{e}^{-\frac{(x-\mu_0)^2}{2\sigma_0^2}} + \pi_1 \cdot \frac{1}{(2\pi\sigma_1^2)^{1/2}} \mathrm{e}^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}.$$

復習：正規分布の確率

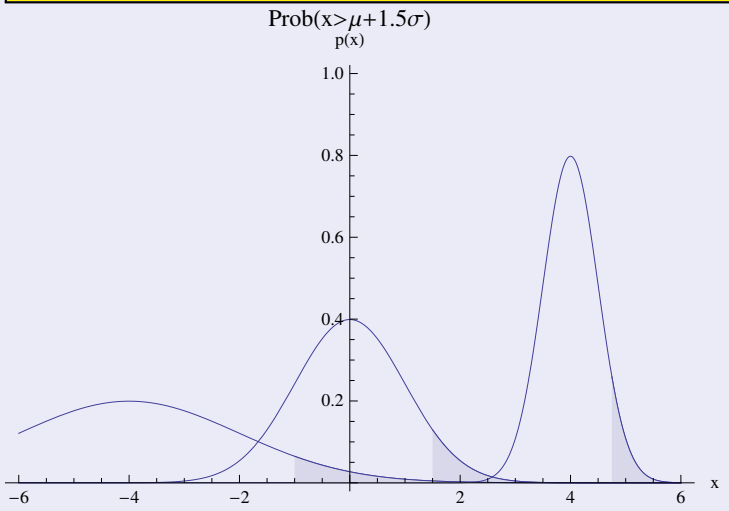
$Z \sim N(0, 1^2)$ とする．累積分布関数 $\Phi(z) = P(Z \leq z)$.
$X \sim N(\mu, \sigma)$ とする．

命題 (変数変換前後の確率は同じ 久保川 統計学入門 公式 5.10)

確率は変数変換，特に標準化 $Z = \frac{X-\mu}{\sigma} \sim N(0, 1^2)$ の前後で変わらない．

斜線部の面積はどれも同じ

$$
\begin{aligned}
P(c \leq X \leq d) &= P\left(\frac{c-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{d-\mu}{\sigma}\right) \\
&= P\left(\frac{c-\mu}{\sigma} \leq Z \leq \frac{d-\mu}{\sigma}\right) \\
&= \Phi\left(\frac{d-\mu}{\sigma}\right) - \Phi\left(\frac{c-\mu}{\sigma}\right)
\end{aligned}
$$



L05-Q2 石村畑 §6.7

Quiz(混合ガウス分布)

次の同時確率（密度）関数で指定される連続型確率変数 X, 離散型確率変数 Y を考える.

$$f(x,y) = \pi_y \cdot f(x|\mu_y, \sigma_y^2)$$

ただし，$f(x|\mu,\sigma^2)$ は $N(\mu,\sigma^2)$ の確率密度関数.
$\pi_0 = \frac{2}{3}, \pi_1 = 1/3, \pi_{他} = 0, \mu_0 = -4, \sigma_0^2 = 2^2, \mu_1 = 3, \sigma_1 = 3^2$ とする.

1. 周辺分布の確率密度関数のグラフ $f_{\mathrm{x}}(x)$ を描こう.
2. 確率 $P(X \leq -1)$ を，標準正規分布 $N(0,1^2)$ の累積分布関数 $\Phi(x)$ で表そう,

略解

1. 周辺分布は混合ガウス分布．二つ山のグラフ.

2. $\frac{2}{3}\Phi\left(\frac{-1-(-4)}{2}\right)+\frac{1}{3}\Phi\left(\frac{-1-3}{3}\right)$