

線形重回帰モデルの精度評価

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(2026-10-08 Thu)

最終更新: Time-stamp: "2026-10-15 Thu 07:26 JST hig"

今日の目標

- 多変量正規分布のアフィン変換による母平均値，母共分散の変換を説明できる
- 線形重回帰モデルの残差，決定係数を説明できる



復習：線形重回帰モデル

多変量解析☆演習 (2026)L03

$$Y = \alpha_0 + \beta_1(x_1 - \bar{x}_{\bullet 1}) + \beta_2(x_2 - \bar{x}_{\bullet 2}) + \epsilon$$

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

ダイヤルの平均 $\bar{x}_{\bullet j} = \frac{1}{n} \sum_i x_{ij}$. 行列 X とベクトル $\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\epsilon}$ で, また ϵ_i は独立より, 次のようにまとめられる.

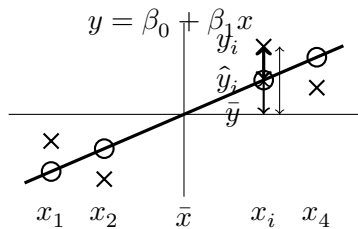
$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \epsilon_i \sim N(0, \sigma^2) \text{ (独立同一分布)}$$

$$\begin{bmatrix} Y \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X \\ \vdots \\ X_n \end{bmatrix} \begin{bmatrix} \beta \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} - \bar{x}_{\bullet 1} & x_{12} - \bar{x}_{\bullet 2} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} - \bar{x}_{\bullet 1} & x_{n2} - \bar{x}_{\bullet 2} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \alpha_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

最尤推定 $\rightsquigarrow$ 残差二乗和 S_e 最小化 $\rightsquigarrow \hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \mathbf{Y} \rightsquigarrow \hat{\mathbf{Y}} = X\hat{\boldsymbol{\beta}} = X(X^\top X)^{-1} X^\top \mathbf{Y}$

$\mathbf{Y}, \mathbf{y}$: データ中の目的変数, $\hat{\mathbf{Y}}, \hat{\mathbf{y}}$: 推定値, 回帰平面上の値.



ここまで来たよ

4. 線形重回帰モデルの精度評価

- 多変量正規分布の重回帰への適用
- 重回帰係数の推定量のしたがう分布
- 残差と平方和の分解と決定係数
- 線形重回帰で使えるいろんな説明変数

アフィン変換のもとでの母平均値・母共分散の変換

命題 (アフィン変換のもとでの母平均値・母共分散の変換)

一般の確率変数 $\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$ と, $q \times n$ 行列 $A = (a_{ij})$, $\mathbf{b} \in \mathbb{R}^q$ に対して,

$$\mathbf{Y}' = A\mathbf{Y} + \mathbf{b}$$

$$\boxed{\mathbf{y}'} = \boxed{A} \boxed{\mathbf{y}} + \boxed{\mathbf{b}}$$

を考えると,

$$E[A\mathbf{Y} + \mathbf{b}] = AE[\mathbf{Y}] + \mathbf{b}, \quad \text{Cov}(A\mathbf{Y} + \mathbf{b}) = ACov(\mathbf{Y})A^\top$$

母平均値は簡単. 一方母共分散は, $\text{Cov}(A\mathbf{Y} + \mathbf{b})_{ij} = E[(\sum_k a_{ik}Y_k + b_i)(\sum_\ell a_{j\ell}Y_\ell + b_j)] - E[(\sum_k a_{ik}Y_k + b_i)]E[(\sum_\ell a_{j\ell}Y_\ell + b_j)] = \sum_{k,\ell} a_{ik} \text{Cov}(Y_k, Y_\ell) a_{j\ell} = \sum_{k,\ell} a_{ik} \text{Cov}(Y_k, Y_\ell) (A^\top)_{\ell j} = ACov(\mathbf{Y})A^\top$.

多変量正規分布のアフィン変換

命題

$q \times n$ 行列 A の $\text{rank } A = q$ のとき,

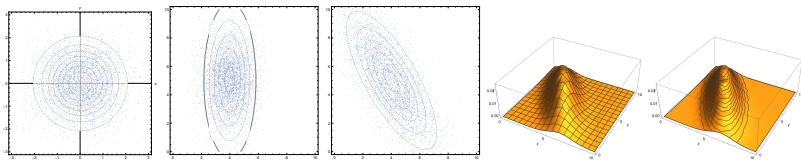
$\mathbf{Y} \sim N(\boldsymbol{\mu}, \Sigma)$ ならば,

$\mathbf{Y}' = A\mathbf{Y} + \mathbf{b}$ は $N(A\boldsymbol{\mu} + \mathbf{b}, A\Sigma A^\top)$ にしたがう.

- ここでは証明できない.
- 正確な証明 確率統計 II
- 部分的には正規分布の再生性からわかる

久保川 統計学入門 公式 6.21

等高線が傾いた楕円の 2 次元正規分布



等高線が円 $D = x^2 + y^2$

等高線が傾いた楕円 $D = x^2 + cxy + 2y^2$

$$D = \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} x \\ y \end{pmatrix}$$

から出発して、回転と拡大縮小する.

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \sigma_x & 0 \\ 0 & \sigma_y \end{pmatrix} \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \end{pmatrix} = A \begin{pmatrix} x \\ y \end{pmatrix},$$

$$D = \begin{pmatrix} x' & y' \end{pmatrix} (A^{-1})^\top \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^{-1} A^{-1} \begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} x' & y' \end{pmatrix} (A \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} A^\top)^{-1} \begin{pmatrix} x' \\ y' \end{pmatrix}$$

L04-Q1

石村畑 §6.7

Quiz(2次元正規分布のアフィン変換)

多変量確率分布 $\mathbf{Y} \sim N(\mathbf{0}, \Sigma)$ を考える. ただし, $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}$.
 $\mathbf{Y}' = A\mathbf{Y} + \mathbf{b}$ を考える.

$$A = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}, \mathbf{b} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$$

とする.

- ① $\mathbf{Y}$ の分布の等高線を描こう.
- ② $\mathbf{Y}'$ はどのような分布にしたがうか.
- ③ $\mathbf{Y}'$ の分布の等高線を描こう.

略解

① $\frac{y_1^2}{1^2} + \frac{y_2^2}{2^2} = c^2$. 中心原点, 長軸は y_2 軸にそい長径 $2c$, 単軸は y_1 軸にそい短径 c の楕円.

② $\mathbf{Y}' \sim N(\mathbf{b}, \Sigma')$, $\Sigma' = A \Sigma A^\top = \begin{pmatrix} 5 & -3 \\ -3 & 5 \end{pmatrix}$.

③ 等高線は,

$$(\mathbf{y}' - \mathbf{b})^\top \begin{pmatrix} 5 & -3 \\ -3 & 5 \end{pmatrix}^{-1} (\mathbf{y}' - \mathbf{b}) = c^2.$$

$$(\mathbf{y}' - \begin{pmatrix} 2 \\ 3 \end{pmatrix})^\top \frac{1}{16} \begin{pmatrix} 5 & 3 \\ 3 & 5 \end{pmatrix} (\mathbf{y}' - \begin{pmatrix} 2 \\ 3 \end{pmatrix}) = c^2.$$

長軸が $y_1 = -y_2$, 単軸が $y_1 = y_2$ になるように $\pi/4$ 回転し, 中心を $\mathbf{b}$ に移すように平行移動した楕円.

この長軸短軸方向, 長径短径の比は, Σ'^{-1} の固有値固有ベクトルを求めることによって決められる.

Σ'^{-1} の固有値は, $1/2, 1/8$, 対応する固有ベクトルの一つは $\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}$,

ここまで来たよ

4. 線形重回帰モデルの精度評価

- 多変量正規分布の重回帰への適用
- 重回帰係数の推定量のしたがう分布
- 残差と平方和の分解と決定係数
- 線形重回帰で使えるいろんな説明変数

重回帰係数の推定量のしたがう分布

β : 仏のみ知る真の母数

$\hat{\beta}$: x_{ij} のもとでの $\mathbf{Y}$ の繰り返し測定から毎回求まる推定量．確率変数．だから，分布が考えられる．

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 E)$$

なので， $\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2 E)$ ．一方，回帰係数の推定量 $\hat{\beta}$ は

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

というアフィン変換で得られるので，やはり多変量正規分布にしたがう．

$$E[\hat{\beta}] = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top E[\mathbf{Y}] = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X}\beta = \beta$$

$$\text{Cov}(\hat{\beta}) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \sigma^2 E((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top)^\top = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1},$$

$$\hat{\beta} \sim N(\beta, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}) \quad \text{久保川 統計学入門 公式 12.1, 12.6}$$

もっともな結果．母平均値が β ．不偏．ダイヤルの設定値 $\mathbf{X}$ の共分散 $\times n = \mathbf{X}^\top \mathbf{X}$ が大きいほど誤差＝分散は小さくなる．

重回帰係数に対してできること

久保川 統計学入門 公式 12.2,12.8

- ① 区間推定 (β_0 や $\beta_1, \dots$ の信頼区間は?)
- ② 検定 帰無仮説 0: $\beta_0 = 0 \Leftrightarrow$ 切片はいらない
- ③ 検定 帰無仮説 1: $\beta_1 = 0 \Leftrightarrow Y$ は x_1 に依存しない
 - ▶ 対立仮説 1: Y は x_1 に依存する

ある Python ライブラリで偏回帰係数・仮説検定

import statsmodels.formula.api as smf E03_03 任意
 重回帰なら formula='y ~ x1 + x2'
 青い表の行で, β_0, β_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 に蓄えられた結果を取り出す
```

目的変数は height OLS Regression Results (Adjusted R-squared)

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

t検定のp値 (有意確率) 95%信頼区間

	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
 説明変数weightの
 係数 β_1

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.

ここまで来たよ

4. 線形重回帰モデルの精度評価

- 多変量正規分布の重回帰への適用
- 重回帰係数の推定量のしたがう分布
- 残差と平方和の分解と決定係数
- 線形重回帰で使えるいろんな説明変数

残差

残差 (縦) ベクトル $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - X\hat{\boldsymbol{\beta}}$.

$$\begin{aligned}\text{残差二乗和 } S_e &= \sum_i e_i e_i = \mathbf{e}^\top \mathbf{e} = (\mathbf{y} - X\hat{\boldsymbol{\beta}})^\top (\mathbf{y} - X\hat{\boldsymbol{\beta}}) \\ &= \mathbf{y}^\top \mathbf{y} - 2\hat{\boldsymbol{\beta}}^\top X^\top \mathbf{y} + \hat{\boldsymbol{\beta}}^\top X^\top X \hat{\boldsymbol{\beta}}\end{aligned}$$

各項はスカラー (1 次正方行列) なので、転置しても同じことを使った.

$$\mathbf{y}^\top X \hat{\boldsymbol{\beta}} = (\mathbf{y}^\top X \hat{\boldsymbol{\beta}})^\top = \hat{\boldsymbol{\beta}}^\top X^\top \mathbf{y}.$$

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

最終的に、残差は $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - X(X^\top X)^{-1} X^\top \mathbf{y}$.

残差分析

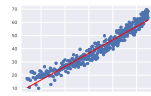
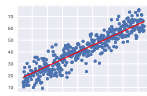
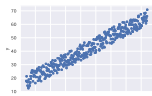
久保川 統計学入門 §12.2

単回帰 $p = 1$ で.

$$\epsilon_i = y_i - (\beta_0 + \beta_1 x_i) \sim N(0, \sigma^2).$$

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

- ① あとで出てくる決定係数 $R^2 = 1 - \frac{\sum \text{残差}^2}{\text{水平な直線からのずれ}^2}$ が 1 に近いのか?

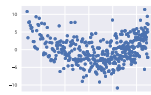
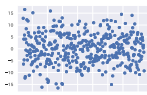
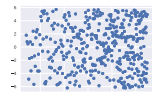


$$R^2 = 0.944$$

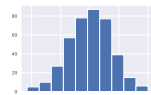
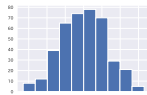
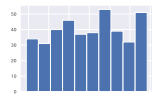
$$0.828$$

$$0.932$$

- ② (x, e) の散布図を描く. 残差 $e_i = y_i - \hat{y}_i$ は測定値の予測値からのずれ.



- ③ 残差 e のヒストグラムを描く (横軸のスケールの違いに注意).



全平方和の分解

$$\begin{aligned} S_t &= |\mathbf{y} - \bar{\mathbf{y}}|^2 \\ &= |(\mathbf{y} - \hat{\mathbf{y}}) + (\hat{\mathbf{y}} - \bar{\mathbf{y}})|^2 \\ &\stackrel{!}{=} |\mathbf{y} - \hat{\mathbf{y}}|^2 + |\hat{\mathbf{y}} - \bar{\mathbf{y}}|^2 \\ &= S_e + S_r \end{aligned}$$

- 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

$\mathbb{R}^n$ で, 2つのベクトルは直交する!

ここでは証明できないが, 線形重回帰と, 一定範囲のモデルで成立する.

L04-Q2

石村畑 §6.7

Quiz(平方和分解と決定係数)

次のデータから求めた回帰直線は $y = 3 + 2x$ である.

x	y
1	6
2	6
4	10
5	14

- ① 総平方和 S_t を求めよう
- ② 回帰平方和 S_r を求めよう
- ③ 残差平方和 S_e を求めよう
- ④ 決定係数 R^2 を求めよう.
- ⑤ 自由度調整済み決定係数 R^{*2} を求めよう.

決定係数と重相関係数

久保川 統計学入門 p.243

定義 (決定係数 R^2 (R^2 -score))

$$R^2 = 1 - \frac{S_e}{S_t} = \frac{S_r}{S_t}$$

R^2 が大きいほどよいモデル+推定結果. 残差が小さく, 回帰による説明がよい. 変動のうち, 回帰で説明できる部分の比率. $0 \leq R^2 \leq 1$.

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

$$R^{*2} = 1 - \frac{S_e/(n-p-1)}{S_t/(n-1)}$$

大きいほどよいモデル+推定結果.

説明変数 p の個数が少ないのに目的変数をよく説明するほど, よいモデルである, ことを反映した指標.

説明変数の個数の違うモデルの間での比較に役立つ (次回).

実は, $n-p-1$, $n-1$ は, S_e , S_t のしたがうカイ二乗分布の「自由度」 確率統計 I

定義 (重相関係数)

重相関係数 = データ $(\hat{y}_i, y_i)_{i=1, \dots, n}$ の相関係数

ここでは証明は手に余る.

命題 (決定係数と重相関係数)

線形重回帰において, 決定係数 $R^2 = (\text{重相関係数})^2$

線形重回帰以外にも, 一定範囲のモデルで成立する.

命題 (重相関係数と相関関数の関係)

線形~~単~~重回帰において,

重相関係数 = データ $(\hat{y}_i, y_i)_{i=1, \dots, n}$ の相関係数 =
データ $(x_i, y_i)_{i=1, \dots, n}$ の相関係数

ここまで来たよ

4. 線形重回帰モデルの精度評価

- 多変量正規分布の重回帰への適用
- 重回帰係数の推定量のしたがう分布
- 残差と平方和の分解と決定係数
- 線形重回帰で使えるいろんな説明変数

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.

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

「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 の線型結合 石村畑 §5.3.

- β_j : 元々の意味 x_j を 1 増やしたとき y が増える量.
- $x_2 = x_1^2$ とか $x_3 = x_1 x_2$ でもよい (多項式回帰).
- $x_4 = \frac{1}{x_1}$ 双曲回帰, $x_5 = e^{x_1}$.
- $x_2 = 3x_1$ とかはだめ. 関数の意味で線形独立である必要. $\leadsto$ 多重共線性

名義変数のダミー変数

(男, 女), (A,B,AB,O) のようなデータの区別は, 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.

- $x_2 = \text{if(性別==女) return 1; else return 0}$ でもよい (真偽値に対するダミー変数). 2群で定数項 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 $\dots \rightarrow$ 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 $x_A = 0, 1, x_B = 0, 1, x_{AB} = 0, 1, x_O = 0, 1$ とする. one-hot vector 機械学習