クレジット:
UTokyo Online Education 統計データ解析 Ⅱ 2018 小池祐太

東京大学 THE UNIVERSITY OF TOKYO

CC BY NC ND

統計データ解析 (II) 第 8 回

小池祐太

2018 年 5 月 30 日

1. 重回帰分析: 復習
2. 最小二乗法による線形回帰式の推定の幾何学的解釈
3. 線形回帰式と標本平均
4. 分析の評価
   - 残差
   - 標準誤差
   - t 値と p 値
   - 決定係数
   - F 値
5. 予測

重回帰分析

- **回帰分析 (regression analysis)**
  - ある 1 種類の変数/データを別の変数/データ (1 種類もしくは複数) によって説明もしくは予測するための関係式 (**回帰 (方程) 式 (regression equation)**) を構成することを目的とする分析法
- 説明される側のデータは, 目的変数, 被説明変数, 従属変数, 応答変数などと呼ばれる
- 説明する側のデータは, 説明変数, 独立変数, 共変量などと呼ばれる
- 説明変数が 1 種類の場合を**単回帰 (simple regression)**, 複数の場合を**重回帰 (multiple regression)** と呼ぶ

重回帰分析

- 目的変数を Y, 説明変数を $X_1, \ldots, X_p$ で表すことにし, 組 $(Y, X_1, \ldots, X_p)$ に対する n 個の観測データ

$$\{(y_i, x_{i1}, \ldots, x_{ip})\}_{i=1}^{n} \tag{1}$$

  が得られている状況を考える
- 観測データは次のモデルに従うとする:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \qquad i = 1, \ldots, n. \tag{2}$$

  - $\beta_0, \beta_1, \ldots, \beta_p$ は未知パラメーター (**回帰係数 (regression coefficients)**)
  - $\epsilon_1, \ldots, \epsilon_n$ (**誤差項 (error term)**): 独立な確率変数 (多くの場合それぞれ平均 0, 分散 σ^2 の正規分布に従うと仮定)

最小二乗法

- 回帰係数 $\boldsymbol{\beta} := (\beta_0, \beta_1, \ldots, \beta_p)^\top$ は**最小二乗法 (least squares)** によって決定 (推定) する
- すなわち, **残差平方和 (residual sum of squares)**

$$S(\boldsymbol{\beta}) := \sum_{i=1}^{n} |y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})|^2$$

を最小化するベクトル $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)^\top$ を $\boldsymbol{\beta}$ の推定量とする (**最小二乗推定量 (least squares estimator)**)

- R での実行: 関数 `lm()`

線形回帰式の行列による表現

- 最小二乗推定量の計算のために, モデル (2) を行列を用いて表現する

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix},$$

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \qquad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

とおくと, モデル (2) は

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

と表すことができる

線形回帰式の行列による表現

- 行列 $\boldsymbol{X}$ は**デザイン行列 (design matrix)** と呼ばれることがある
- 最小二乗推定量がただ一つだけ存在するための必要十分条件は, **グラム行列 (Gram matrix)** $\boldsymbol{X}^\top \boldsymbol{X}$ が正則であることである (配布資料の定理 5.1)
- 以下では $\boldsymbol{X}^\top \boldsymbol{X}$ の正則性を常に仮定する

最小二乗法による線形回帰式の推定の幾何学的解釈

- 一般に $\hat{\boldsymbol{\beta}}$ を回帰係数の推定値とすると, 目的変数の観測データ $\boldsymbol{y}$ から観測誤差の影響を除いた目的変数の真の値が,

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

  によって推定できる
  - $\hat{\boldsymbol{y}}$ を**あてはめ値 (fitted values)** または**予測値 (predicted values)** と呼ぶ
- 特に $\hat{\boldsymbol{\beta}}$ が最小二乗推定量のとき, デザイン行列 $\boldsymbol{X}$ の列ベクトルたちで張られる $\mathbb{R}^n$ の部分線形空間 (超平面) を $L[\boldsymbol{X}]$ と書くことにすれば, $\hat{\boldsymbol{y}}$ はベクトル $\boldsymbol{y}$ の $L[\boldsymbol{X}]$ への直交射影となることがわかる (資料の定理 5.3)
  - すなわち, $\hat{\boldsymbol{y}}$ は, $L[\boldsymbol{X}]$ に属するベクトル $\boldsymbol{z}$ で $\boldsymbol{y} - \boldsymbol{z}$ が $L[\boldsymbol{X}]$ に属するすべてのベクトルに直交するような唯一のもの

最小二乗法による線形回帰式の推定の幾何学的解釈

- 幾何学的には, ベクトル $\boldsymbol{y}$ からデザイン行列 $\boldsymbol{X}$ の列ベクトルによって張られる (超) 平面に垂線を下ろした際の垂線の足が $\hat{\boldsymbol{y}}$ となる
- 次頁に $n = 3$, $p + 1 = 2$ の場合に最小二乗法による推定の様子を図示したものを示す

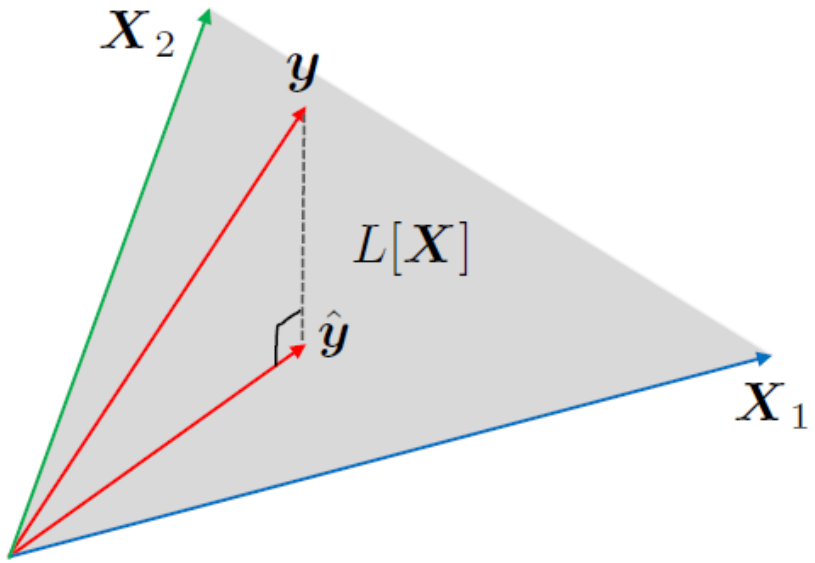

図 1: $n = 3$, $p + 1 = 2$ の場合の最小二乗法による推定の図示. $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ はそれぞれデザイン行列 $\boldsymbol{X}$ の第 1 列, 第 2 列からなるベクトルに対応している. グレーの平面が $L[\boldsymbol{X}]$ に対応.

最小二乗法による線形回帰式の推定の幾何学的解釈

- 特に, ベクトル $\hat{\epsilon} := \boldsymbol{y} - \hat{\boldsymbol{y}}$ はあてはめ値 $\hat{\boldsymbol{y}}$ に直交する:

$$\hat{\epsilon} \cdot \hat{\boldsymbol{y}} = 0.$$

- $\hat{\epsilon}$ は**残差 (residuals)** と呼ばれ, 回帰式による目的変数のあてはめ値と実際の観測値とのずれを表す
- 実行例 `lse.r`

線形回帰式と標本平均

- 各 $i = 1, \ldots, n$ について, 説明変数の i 番目の観測データに対応するベクトル $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ を導入する
- 説明変数および目的変数の標本平均を考える

$$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

- このとき, $\hat{\boldsymbol{\beta}}$ を最小二乗推定量とすれば, $\bar{y} = (1, \bar{\boldsymbol{x}}^\top)\hat{\boldsymbol{\beta}}$ が成り立つことが確認できる (資料参照)
- 幾何学的には, 方程式 $y = (1, \boldsymbol{x}^\top)\hat{\boldsymbol{\beta}}$ によって定まる超平面は常に点 $(\bar{\boldsymbol{x}}^\top, \bar{y})$ を通るということである
- 実行例 `lse.r`

分析の評価

- 関数 `lm( )` のアウトプットに関数 `summary( )` を適用した際に表示される, 分析結果の評価をするための各種指標について解説する

残差

- 関数 `summary( )` のアウトプットの “Residuals” の欄には残差 $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$ に対する五数要約 (最小値, 第 1 分位点, メディアン, 第 3 分位点, 最大値) が表示される
- 残差は絶対値が小さいほど回帰式の観測データへのあてはまりがよいこととなるので, 残差のばらつきは小さいほどよい
- 実行例 `resid.r`

標準誤差

- モデル (2) において, 誤差項 $\epsilon_1, \ldots, \epsilon_n$ は平均 0, 分散 σ^2 の正規乱数であると仮定する
- 最小二乗推定量 $\hat{\boldsymbol{\beta}}$ は誤差項に依存して変化するため確率変数であるが, いまの仮定の下では平均 $\boldsymbol{\beta}$, 共分散行列 $\sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ の $(p+1)$ 変量正規分布に従うことが知られている
- 特に, 行列 $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ の対角成分を $\xi_0, \xi_1, \ldots, \xi_p$ とすれば, 各 $j = 0, 1, \ldots, p$ について, $\hat{\beta}_j$ は平均 β_j, 分散 $\sigma^2 \xi_j$ の正規分布に従う
- 正規分布の分散, もしくはその平方根である標準偏差は, 確率変数の値の平均からの離れやすさを表す指標だと解釈できる (大きいほど離れやすい)

標準誤差

- $\hat{\beta}_j$ の値は “真の” 回帰係数値 β_j に近ければ近いほどよい推定であるといえるから, その標準偏差 $\sigma\sqrt{\xi_j}$ は $\hat{\beta}_j$ の推定精度を評価するのに利用できる
- ただし, σ は未知パラメーターだから, データから推定する必要がある
- σ の推定量としては, 通常

$$\hat{\sigma} = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}\hat{\epsilon}_i^2}$$

が利用される (この推定量が利用される理由の直感的説明については資料の注意 5.1 を参照)

標準誤差

- こうして推定値 $\hat{\beta}_j$ の精度を評価するための指標

$$\hat{\sigma}\sqrt{\xi_j}$$

  が得られるが, これを $\hat{\beta}_j$ の**標準誤差**と呼ぶ
- 標準誤差は, 関数 `summary( )` のアウトプットの “Coefficients” の欄の 2 列目 “Std. Error” で確認できる
- また, 関数 `summary( )` のアウトプットの “Residual standard error” の欄で $\hat{\sigma}$ の値を確認できる ($\hat{\sigma}$ は残差のばらつき具合を表す指標として利用できる)
- 実行例 `se.r`

t 値と p 値

- $(n-p-1)\hat{\sigma}^2/\sigma^2$ は自由度 $n-p-1$ のカイ二乗分布に従い, かつ $\hat{\boldsymbol{\beta}}$ と独立であることが知られている
- $(\hat{\beta}_j - \beta_j)/(\sigma\sqrt{\xi_j})$ が標準正規分布に従うことに注意すれば, $\hat{\beta}_j - \beta_j$ を $\hat{\beta}_j$ の標準誤差で割った量

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{\xi_j}}$$

は自由度 $n-p-1$ の t 分布に従うことがわかる (配布資料 4 章 4.4.4 節参照)

t 値と p 値

- よって, もし $\beta_j = 0$ であったならば, 統計量

$$t = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{\xi_j}}$$

は自由度 $n - p - 1$ の t 分布に従う
  - この t を $\hat{\beta}_j$ の t **値**と呼ぶ
- また, 自由度 $n - p - 1$ の t 分布に従う確率変数の絶対値が $|t|$ を超える理論上の確率

$$2\int_{|t|}^{\infty} f(x)dx, \quad (f(x) \text{ は自由度 } n - p - 1 \text{ の } t \text{ 分布の確率密度関数}) \tag{3}$$

を $\hat{\beta}_j$ の p **値**と呼ぶ

t 値と p 値

- $\beta_j = 0$ が成り立つということは, j 番目の説明変数 X_j は回帰式に寄与していないこととなるから, 回帰式から除外しても問題無いことが示唆される
- 上で定義した t 値および p 値は, $\beta_j = 0$ であるか否かという仮説を検証するのに利用できる
- 実際, もし $\beta_j = 0$ という仮説が正しければ, 確率 (3) はそれほど小さくはならないはずなので, もし p 値が想定よりも小さい場合, はじめの仮説である $\beta_j = 0$ が誤っているという結論した方が自然である

t 値と p 値

- 統計の言葉で言うと, t 値及び p 値は, 仮説検定

$$H_0 : \beta_j = 0 \qquad \text{vs} \qquad H_1 : \beta_j \neq 0$$

  に対する検定統計量の t 値と p 値となっている
- また, ここでいうはじめに想定する p 値の下限を**有意水準**といい, 通常 0.01 もしくは 0.05 とすることが多い
- p 値が有意水準より小さいような回帰係数の推定値をもつ説明変数は**有意**であるといわれる
  - 有意な説明変数は目的変数の変動を説明するのに有用であるといえる

t 値と p 値

- t 値および p 値は, 関数 `summary( )` のアウトプットの “Coefficients” の欄の 3–4 列目 “`t value`” および “`Pr(>|t|)`” で確認できる
- また, p 値の横についているアスタリスク等の記号は, p 値がどの程度小さいかを示している
- 例えば, ‘*’ は p 値が 0.05 以下であることを示し, ‘**’ は p 値が 0.01 以下であることを示す (記号の意味は “Coefficients” の欄の下部に書いてある)
- 実行例 `tpvalues.r`

決定係数

- **決定係数**は線形回帰分析のあてはまり具合を評価するためのもっとも代表的な指標である
- 決定係数は記号 R^2 で表され, 回帰モデルによる目的変数のあてはめ値 $\hat{y}_1, \ldots, \hat{y}_n$ と実際の観測データ $y_1, \ldots, y_n$ の相関の 2 乗として定義される:

$$R^2 = \frac{\left(\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})\right)^2}{\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2 \sum_{i=1}^n (y_i - \bar{y})^2}. \tag{4}$$

ここに, $\bar{\hat{y}}$ と $\bar{y}$ はそれぞれ $\hat{y}_1, \ldots, \hat{y}_n$ と $y_1, \ldots, y_n$ の平均を表す:

$$\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^n \hat{y}_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^n y_i.$$

決定係数

- あてはめ値と実際の観測データの変動が近いほどあてはまりが良いと考えられるので, 決定係数は高ければ高いほどよい.
- 決定係数は以下のようにも書くことができる (資料参照, もしくは自分で確認):

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{5}$$

- (5) の分子と分母をそれぞれ $n-1$ で割ることで, 決定係数はあてはめ値の分散を目的変数の観測データの分散で割ったものだとも解釈できる
- すなわち, 目的変数の観測データの分散のうち何パーセントを回帰モデルが説明できているかを表す指標とも解釈できる

決定係数

- 決定係数はさらに以下のようにも書き直せる (資料参照, もしくは自分で確認):

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}. \tag{6}$$

- (6) より, 決定係数は説明変数を付け加えるほど高くなることがわかるため, 決定係数は本来回帰式に不要である説明変数の効果を過剰に見積もっているおそれがある

決定係数

- この問題を解消するために, 推定されて得られた未知パラメーターの影響を考慮して以下のように決定係数を修正したものが**自由度調整済み決定係数**である:

$$\bar{R}^2 = 1 - \frac{\frac{1}{n-p-1}\sum_{i=1}^{n}\hat{\epsilon}_i^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}.$$

- なお, 決定係数は**寄与率**とも呼ばれる
- 決定係数および自由度調整済み決定係数は, それぞれ関数 `summary( )` のアウトプットの “Multiple R-squared” および “Adjusted R-squared” の欄で確認できる
- 実行例 `rsquared.r`

F 値

- t 値は個々の説明変数の要・不要を判断するための指標であったが, 説明変数のうち 1 つでも目的変数の説明の役に立つものがあるか否かを判定するための指標に回帰モデルの F **値**がある
- これは, 現在の説明変数を用いて回帰分析を実行することに意味があるかどうかを検証するための指標ともいえる
- 回帰モデルの F 値は次式で定義される:

$$F = \frac{\frac{1}{p}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \frac{n-p-1}{p}\frac{R^2}{1-R^2}.$$

F 値

- もしすべての説明変数が不要, すなわち $\beta_1 = \cdots = \beta_p = 0$ であったならば, F は自由度 $p, n-p-1$ の F 分布に従うことが知られている
- したがって, 自由度 $p, n-p-1$ の F 分布に従う確率変数が F を超える理論上の確率

$$\int_F^\infty f(x)dx, \quad f(x) \text{ は自由度 } p, n-p-1 \text{ の } F \text{ 分布の確率密度関数} \tag{7}$$

はそれほど小さくはならないはずなので, この確率が想定より小さければ回帰分析に意味があると結論付けられる

F 値

- 統計の言葉で言うと, F 値及び確率 (7) は, 仮説検定

$$H_0 : \beta_1 = \cdots = \beta_p = 0$$
$$\text{vs} \qquad H_1 : \text{ある } j = 1, \ldots, p \text{ に対して, } \beta_j \neq 0$$

  に対する検定統計量の F 値と p 値となっている
- 回帰モデルの F 値および確率 (7) は, それぞれ関数 `summary( )` のアウトプットの “F-statistic” およびその隣の “p-value” の欄で確認できる
- 実行例 `fstatistics.r`

予測

- 回帰分析の目的の 1 つは, 説明変数の新規データが与えられたときに, そのデータに対応する目的変数の値を予測することであるが, これは関数 `predict()` で実行できる
- 実行例 `predict.r`