応用統計学 2017 第2回 最小二乗法，重回帰分析

2017年10月4日 (水)

清 智也 sei@mist.i.u-tokyo.ac.jp
http://ur0.pw/yTzt

- 単回帰：目的変数，説明変数，最小二乗法，回帰式，回帰係数，予測値，残差[1]。
- 重回帰：計画行列，正規方程式，決定係数，重相関係数，ダミー変数[2]。
- 後日扱う内容：回帰モデル，変数選択，一般化線形モデル。

演習問題

前回までの演習問題（の一部）の略解は，講義用 web ページからダウンロードできる。以下では行列の転置を $'$ で表す。

問題 2-1. 目的変数 $\boldsymbol{y}$ を説明変数 $\boldsymbol{x}$ で説明する回帰式 $\hat{y}(x) = \hat{a} + \hat{b}x$ の回帰係数は

$$\hat{a} = \bar{y} - \hat{b}\bar{x}, \quad \hat{b} = r_{xy}(s_y/s_x)$$

で与えられることを示せ。ただし $\bar{x}, \bar{y}$ は平均，s_x, s_y は標準偏差，r_{xy} は相関係数を表す。

問題 2-2. 講義の web ページにある「LPGA データ」は 2017 年女子オープンゴルフ選手権競技の 3 日目までのスコア $\boldsymbol{x}$ と 4 日目のスコア $\boldsymbol{y}$ からなる CSV ファイルである[3]。これをダウンロードし，$\boldsymbol{y}$ を $\boldsymbol{x}$ で説明する回帰式を求めよ。また散布図を描き，回帰直線を重ね描きせよ。

問題 2-3. 2016 年度の東京の日最高気温（℃）の月平均値 y_t $(t = 1, \ldots, 12)$ は

$$10.6, 12.2, 14.9, 20.3, 25.2, 26.3, 29.7, 31.6, 27.7, 22.6, 15.5, 13.8$$

であった。目的変数を y_t，説明変数を $x_{t1} = \cos(2\pi t/12)$，$x_{t2} = \sin(2\pi t/12)$ として回帰式を求めよ。

なお，上のデータは気象庁のサイト http://www.jma.go.jp/から「各種データ・資料」に進み，地点を「東京都 東京」，年月日を「2016 年」，データの種類を「2016 年の月ごとの値を表示」とすれば得られる。

[1] simple regression: response variable, explanatory variable, least squares method, regression equation, regression coefficient, predicted value, residual.

[2] multiple regression: design matrix, normal equation, coefficient of determination, multiple correlation coefficient, dummy variable.

[3] 日本女子プロゴルフ協会ホームページ (https://www.lpga.or.jp) より作成。

問題 2-4. 目的変数 $\boldsymbol{y}$ を p 個の説明変数 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ で説明する回帰式

$$\hat{y}(x_1, \ldots, x_p) = \hat{a} + \hat{b}_1 x_1 + \cdots + \hat{b}_p x_p, \quad (x_1, \ldots, x_p) \in \mathbb{R}^p,$$

は点 $(\bar{x}_1, \ldots, \bar{x}_p, \bar{y})$ を通ることを示せ。また，$\mathbf{1}_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ が互いに直交しているとき，$\hat{a}, \hat{b}_1, \ldots, \hat{b}_p$ はどのように表されるか。

問題 2-5 (Simpson's paradox revisited)**.** $\boldsymbol{y}$, $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ がそれぞれ

$$\boldsymbol{y} = (4, 5, 6, 1, 2, 3)', \quad \boldsymbol{x}_1 = (1, 2, 3, 4, 5, 6)', \quad \boldsymbol{x}_2 = (1, 1, 1, -1, -1, -1)'$$

であるとする。$\boldsymbol{y}$ を $\boldsymbol{x}_1$ と $\boldsymbol{x}_2$ で説明する場合の回帰係数と，$\boldsymbol{x}_1$ だけで説明する場合の回帰係数を比較せよ。ただし回帰式には定数項も含めるものとする。

問題 2-6. ランク p の行列 $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ に対し，$\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ と定義される行列 $\boldsymbol{P}$ は直交射影行列になること，すなわち $\boldsymbol{P}^2 = \boldsymbol{P}$ かつ $\boldsymbol{P}' = \boldsymbol{P}$ が満たされることを示せ。また $\boldsymbol{P}$ はどのような部分空間への射影になっているか説明せよ。

問題 2-7. 計画行列 $\boldsymbol{X}$ の QR 分解 $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$ が得られているとき，回帰係数を数値的に求める上で有利な点を説明せよ。ただし $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$ が $\boldsymbol{X}$ の QR 分解であるとは，$\boldsymbol{Q}$ が列直交（$\boldsymbol{Q}'\boldsymbol{Q}$ が単位行列）で，$\boldsymbol{R}$ が正の対角成分を持つ上三角行列となることをいう。

宿題

問題 2-8. アンケート結果を講義の web ページからダウンロードし，睡眠時間を通学時間で説明する回帰式を求めよ。