多変量解析
第 2 回: 回帰分析 (1) 単回帰と最小 2 乗法

藤本 衡

2018 年 4 月 18 日

本日の内容

- 回帰分析とは？
- 単回帰分析の概要
- 最小 2 乗法の考え方
- 練習問題

回帰モデル (regression model)

- 目的変数 Y は説明変数 X の関数として表現できる「はず」だ、という考え
- X の観測値が得られれば、モデルによって Y を推定できる
- 説明変数は複数あってもよい

回帰モデル (回帰式)

$$Y = f(X_1, X_2, \ldots, X_N)$$

線形回帰モデル (linear regression model)

- **関数 f が一次 (線形) 関数の場合**
- **係数 $\alpha, \beta_1, \beta_2, \ldots, \beta_N$ は未知**

線形回帰モデル (線形回帰式)

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_N X_N$$

線形単回帰モデル (simple linear regression model)

- まずは話を簡単にしよう; 説明変数は 1 個だけ
  - 単回帰と区別するため、説明変数が 2 個以上の場合は重回帰と呼ぶ

線形単回帰モデル (線形単回帰式)

$$Y = \alpha + \beta X$$

未知パラメータ α, β の推定 (1)

- (X, Y) の観測値 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ を使おう
- でも観測値には誤差がつきもの $\longrightarrow$ 残差 ε の導入
  $\varepsilon_i = y_i - Y$

$$
\begin{aligned}
y_1 &= \alpha + \beta x_1 + \varepsilon_1 \\
y_2 &= \alpha + \beta x_2 + \varepsilon_2 \\
&\vdots \\
y_n &= \alpha + \beta x_n + \varepsilon_n
\end{aligned}
$$

残差に関する仮定

- **残差の期待値は** 0: $\mathbb{E}[\varepsilon_i] = 0$
- **任意の** i **について残差の分散は定数**: $\forall i, \mathbb{V}[\varepsilon_i] = \sigma^2$
- **相異なる個体の残差は無相関**: $i \neq j \Rightarrow \mathrm{Cov}[\varepsilon_i, \varepsilon_j] = 0$

未知パラメータ α, β の推定 (2)

- 残差 ε が小さいほど、当初のモデルに近いと言える

残差の評価尺度

$$\text{残差平方和:RSS} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \{y_i - (\alpha + \beta x_i)\}^2$$

$$\text{絶対残差和:SAR} = \sum_{i=1}^{n} |\varepsilon_i| = \sum_{i=1}^{n} |y_i - (\alpha + \beta x_i)|$$

$$\text{最大絶対残差:MAR} = \max_i \{|\varepsilon_i|\} = \max_i \{|y_i - (\alpha + \beta x_i)|\}$$

未知パラメータ α, β の推定 (3)

- RSS は α, β に関する凸 2 次関数である
  $\longrightarrow$ RSS が極小となる (α, β) が一意に定まる
  - SAR や MAR でも最小値は求められるが線形計画問題なので面倒

最小 2 乗法 (least square method)

$$\begin{cases} \dfrac{\partial RSS}{\partial \alpha} = \boxed{\text{計算してみて！}} = 0 \\ \dfrac{\partial RSS}{\partial \beta} = \boxed{\text{計算してみて！}} = 0 \end{cases} \tag{1}$$

最小 2 乗法

- (1) 式を満たす解を α^*, β^* とおく
- (1) 式は α, β に関する連立 1 次式
- 普通に答え出るんじゃね？

最小 2 乗法 (変形)

$$\begin{cases} \boxed{?\ 1}\alpha^* + \boxed{?\ 2}\beta^* - \boxed{?\ 3} = 0 \\ \boxed{?\ 4}\alpha^* + \boxed{?\ 5}\beta^* - \boxed{?\ 6} = 0 \end{cases} \qquad (2)$$

最小 2 乗法の解

$$\beta^* = \frac{s_{xy}}{s_x}, \quad \alpha^* = \bar{y} - \frac{s_{xy}}{s_x}\bar{x} \tag{3}$$

ただし

$$\begin{cases} s_x = \dfrac{1}{n}\sum(x_i - \bar{x}) = \dfrac{1}{n}\sum x_i - \bar{x} \\ \quad (x \text{ の分散}) \\ s_{xy} = \dfrac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y}) = \dfrac{1}{n}\sum x_i y_i - \bar{x}\bar{y} \\ \quad (x \text{ と } y \text{ の共分散}) \end{cases}$$

おまけ: 「回帰」分析の語源

- 回帰 (regression)
  - 並外れて高身長の親だけを集めて平均身長を求める
  - その子供の平均身長を求める $\longrightarrow$ 親世代より全体平均に近づく
    = "平均への回帰" (regression toward the mean)
  - 背景：「外れ値はたまたま出るもので、相関はそこまで大きくない」
- 回帰分析の対象は相関が極めて強いデータ
  $\Longrightarrow$ 本来とは逆の意味で使われている