回帰分析

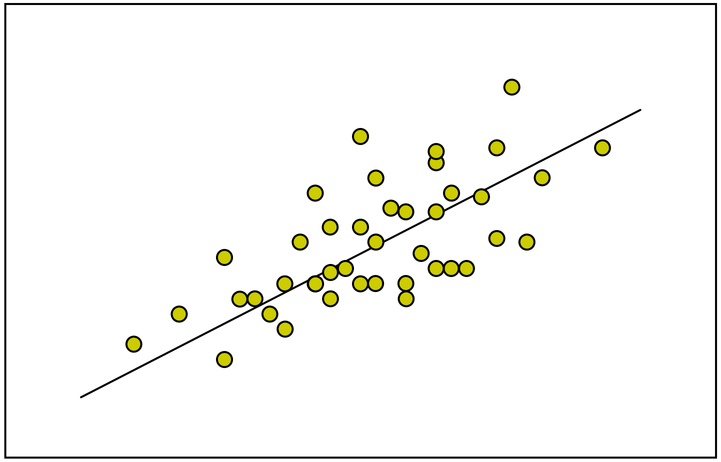
京都大学工学研究科
小山田研究室　TA
田中 哲平

回帰分析 Regression analysis

- 目的
  - XとYとの定量的な関係の構造（モデル）を求める
    - YとXとの関係を求める（相関分析）のではない
  - 従属変数の予測
  - 独立変数の影響

$$y = \beta_1 + \beta_2 x$$

従属変数　　独立変数



回帰方程式

$$y = \beta_1 + \beta_2 x$$

- 線形回帰
  - 非線形回帰であっても、関数変換等により線形モデルに変形または近似できるものも多い
- 例
  - 弾性モデル
    - $z=a\omega^b$
  - 指数回帰
    - $z=ab^x$

両辺の対数をとる

$$\log z = \log a + b \log \omega$$

$$\log z = \log a + x \log b$$

母回帰方程式

i番目の要素

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$$

- β_1, β_2　:(偏)回帰係数
- X_i　:すでに確定した値　(確率変数ではない)
- ε_i　:誤差項
  - $E(\varepsilon_i) = 0$,　$i = 1, 2, \cdots, n$
  - $V(\varepsilon_i) = \sigma^2$, $i = 1, 2, \cdots, n$
  - 異なった誤差項は無相関

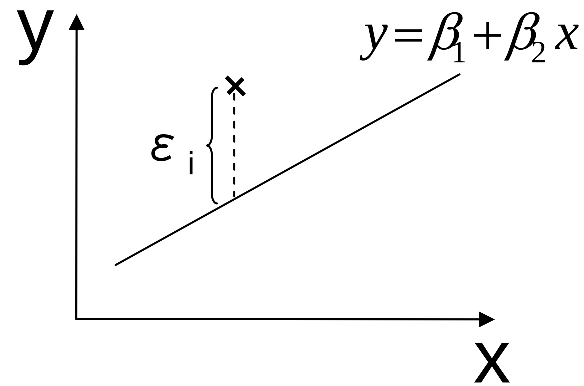
$$\Rightarrow \quad E(Y_i) = \beta_1 + \beta_2 X_i$$

回帰係数の推定

- 最小二乗法
  - 予測誤差の2乗和を最小にする

$$S=\sum(\varepsilon_i)^2$$

$$=\sum\left\{Y_i-(\beta_1+\beta_2X_i)\right\}^2$$



この式が最小となるβ_1，β_2を探す

最小二乗法の計算

- S を回帰係数で偏微分してゼロとおく

$$\begin{cases} \dfrac{\partial S}{\partial \beta_1} = -2\sum \{Y_i - (\beta_1 + \beta_2 X_i)\} = 0 \\ \dfrac{\partial S}{\partial \beta_2} = -2\sum \{Y_i - (\beta_1 + \beta_2 X_i)\} X_i = 0 \end{cases}$$

- 正規方程式
  - 偏微分により得られた連立方程式を解く

$$\begin{cases} n\beta_1 + (\sum X_i)\beta_2 = \sum Y_i \\ (\sum X_i)\beta_1 + (\sum X_i^2)\beta_2 = \sum X_i Y_i \end{cases}$$

標本(偏)回帰係数

- 正規方程式の解

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}, \qquad \hat{\beta}_2 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})}$$

$\hat{\beta}_1, \hat{\beta}_2$:標本(偏)回帰係数

$\overline{X}$:X_iを標本とする標本平均

$\overline{Y}$:Y_iを標本とする標本平均

回帰値、回帰残差

- 回帰値

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

- 回帰残差

$$\begin{aligned}\hat{e}_i &= Y_i - \hat{Y}_i \\ &= Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\end{aligned}$$

誤差項の平均ゼロ $\sum \hat{e}_i = 0$

分散は独立変数に依存せず一定 $\sum \hat{e}_i X_i = 0$

推定値の誤差標準

- 誤差項ε_iの分散σ^2は回帰方程式の当てはまりの良さを表す

$$s^2 = \frac{\sum(\hat{e}_i)^2}{n-2}$$

n：サンプリング数

s：推定値の誤差標準

→ 小さいほど、よく適合している

回帰残差の平方和を(n-2)で割るのは、先ほどの回帰残差の条件を満たすべきための制限が加わり、自由度が2失われているため

3つの平方和

- 偏差平方和 $S_{yy} = \sum (y_i - \bar{y})^2$
- 回帰平方和 $SS_R = \sum (\hat{y}_i - \bar{y})^2$
- 残差平方和 $SS_E = \sum (y_i - \hat{y}_i)^2$

3つの平方和の関係式

$$S_{yy} = S_R + S_e$$

決定係数(寄与率)

$$R^2 = SS_R / S_{yy}$$

$$= 1 - SS_E / S_{yy}$$

$$0 < R \leq 1$$

Rの値が大きいほど、回帰モデルが良いことを表す？

変数を増やせば、残差は減少し値が大きくなるので、必ずしも決定係数の値が大きいものが良い回帰モデルとは限らない

自由度修正済み決定係数

$$R_{ad}{}^{2} = 1 - \frac{SS_E / n - p - 1}{S_{yy} / n - 1}$$

n：サンプリング数

p：説明変数の数