

回帰 (regression)

データをもとに、ある変数（従属変数）を別の変数（独立変数）で予測する式を作るための統計的手法を、「**回帰分析**」(regression analysis) という。

- とくに、独立変数が1つならば**単回帰分析**とよぶ。

回帰直線 (regression line)

回帰直線

- 散布図の各点 (x_i, y_i) が近くに分布するような直線を回帰直線という。

$$y = ax + b$$

- 回帰係数 (回帰直線の傾き) : a
 - 回帰直線のy切片 ($x=0$ のときのyの値) : b
 - 独立変数 (説明変数: 予測に使う変数) : x
 - 従属変数 (目的変数、基準変数: 予測したい変数) : y
- なお、回帰式は必ず $(\bar{x}, \bar{y})$ を通る

最小二乗法

- 観測値 (または実測値) y_i と推定値 (または予測値) $\hat{y}_i$ との差 (残差 e) の二乗が最小になるような a と b を求める。
- 次の値が最小となるような、 a と b を求める。

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- e^2 を足したものを、残差平方和 S_e という

$$S_e = \sum e^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

回帰式の計算

- x を独立変数 (横軸)、 y を従属変数 (縦軸) としたときの回帰式 (y への x からの回帰式) は次のようになる。

$$y = r \cdot \frac{s_y}{s_x} x + \left(\bar{y} - \frac{s_{xy}}{s_x^2} \cdot \bar{x} \right)$$

- なお、

$$a = r \cdot \frac{s_y}{s_x}$$
$$b = \bar{y} - \frac{s_{xy}}{s_x^2} \cdot \bar{x} = \bar{y} - \bar{x}a$$

- 相関係数: r
- 2変数の標準偏差: s_x, s_y
- 2変数の共分散 (偏差積和の平均)

$$s_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

- 回帰式を変形すると、次のようになる。

$$\begin{aligned} y - \bar{y} &= \frac{s_{xy}}{s_x^2} (x - \bar{x}) \\ &= \frac{S_{xy}}{S_{xx}} (x - \bar{x}) \end{aligned}$$

- 2変数の平均値: $\bar{x}, \bar{y}$
- 2変数の偏差積和: S_{xy}

$$S_{xy} = \sum_{i=1}^n d_x \cdot d_y = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

- 偏差平方和: S_{xx}

$$S_{xx} = \sum_{i=1}^n d_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2$$

標準誤差

- 予測値と実測値のずれ (予測値の誤差; 残差 e) の標準偏差を、標準誤差 s_e という

$$\begin{aligned} s_e &= \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y})^2}{n-2}} \\ &= \sqrt{\frac{S_e}{n-2}} \end{aligned}$$

- $n-2$ で割っているのは、2つの係数を推定したことによる自由度 (後日説明) の修正のため

決定係数

- 相関係数の二乗を決定係数 (または寄与率) という。

- 決定係数: R^2

$$\begin{aligned} R^2 &= \left(\frac{1}{n} \frac{S_{xy}}{s_x s_y} \right)^2 \\ &= \frac{S_{xy}^2}{n^2 \cdot s_x^2 s_y^2} \\ &= \frac{S_{xy}^2}{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2} \\ &= \frac{S_{xy}^2}{S_{xx} S_{yy}} \end{aligned}$$

- 偏差平方和と残差平方和を使うと、次のように書くこともできる

$$\begin{aligned}
 R^2 &= \frac{S_{\hat{y}y}}{S_{yy}} \\
 &= \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} \\
 &= 1 - \frac{S_e}{S_{yy}}
 \end{aligned}$$

- 決定係数は、0から1の値をとる。

$$0 \leq R^2 \leq 1$$

- 推定 (回帰式) の精度を表す指標である
 - 従属変数 y の分散の何%を予測値 $\hat{y}$ の分散が説明しているか、を示す
 - だいたい、0.5以上であれば精度が高いといえる

回帰の概念

- 予測値 $\hat{y}$ と従属変数の平均 $\bar{y}$ との差は、一般に独立変数 x とその平均 $\bar{x}$ との差より小さくなる
 - 予測値と平均の差 $(\hat{y} - \bar{y})$ との差は、独立変数と平均の差 $(x - \bar{x})$ との差より小さくなる
 - 予測値は独立変数に比べて平均に近づく
- 統計学的な現象で、「回帰効果」や「平均への回帰」ともいう
 - 例) 1回目の試験の結果が偏っていた (とくに良い、悪いなど) 人について、2回目の試験結果を調べると、その平均値は1回目の結果よりも1回目の全体の平均値に近くなる (時間的には逆で考えてもよい)
 - 例) 父親と子どもの身長を比較して、とくに身長の高い父親でも、とくに身長の低い父親からでも、子どもたちの身長は父親たちの身長より平均に近くなる
 - 例) とくに身長の高い人たちの父親の身長は、子どもたちの身長よりも平均に近い (全体の身長の分布は、父親たちも子どもたちも同じ)