4 決定係数

例 2: A 小学校の 5 人の小学生の学習時間 (X) と成績 (Y) のデータ。

	A	B	C	D	E
X	0.5	1	2	0.8	1.5
Y	3	3	4	1	4

1. 学習時間 2.5 の子の成績はいくらになるだろうか
2. ところで、この予測はどの程度信頼できるものだろうか

学習時間 2.5 の子の成績は?

1. $Y = \alpha + \beta X + u$ という関係式を考える
2. α, β の OLS 推定値を求める

$$
\begin{aligned}
\hat{\beta} &= \frac{S_{xy}}{S_x^2} & (\hat{\beta} = 1.3456) \\
\hat{\alpha} &= \bar{Y} - \beta \bar{X} & (\hat{\alpha} = 1.4391)
\end{aligned}
$$

3. Y の予測値を求める

$$
\begin{aligned}
\hat{Y} &= \hat{\alpha} + \hat{\beta} X \\
&= 1.4391 + 1.3456 * 2.5
\end{aligned}
$$

上の予測の信頼度は?

予測の信頼度を測る基本的なアイデアは、
　推定された回帰式の、既に与えられているデータに対する予測精度
を算定してみること。

$\hat{Y} = \hat{\alpha} + \hat{\beta} X$ 　　を予測値
$\hat{u} = Y - \hat{Y}$ 　　を残差　と呼ぶ。
ここまでの作業は、$Y = \alpha + \beta X + u$ を最小二乗法で推定して、

$$
\begin{aligned}
Y &= \hat{\alpha} + \hat{\beta} X + \hat{u} \\
&= \hat{Y} + \hat{u}
\end{aligned}
$$

という回帰式を得たということになる。

1. RSS (残差二乗和, Sum of the Squared Residuals)

$$RSS = \sum_{i=1}^{n} \hat{u_i}^2$$

たとえば、A 君の場合 $X = 0.5$ だから回帰式による予測値は $\hat{Y} = 1.4391 + 1.3456 * 0.5 = 2.1119$ 、しかし実測値は $Y = 3$ だから $\hat{u} = 3 - 2.1119 = 0.8881$ となる。

この残差を、与えられたデータすべてに関して計算して二乗和をとる。

$$\begin{aligned} RSS &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \\ &= \sum_{i=1}^{5}(Y_i - \hat{\alpha} - \hat{\beta}X_i)^2 \\ &= (3.0 - 1.4391 - 1.3456 * 0.5)^2 \\ &\quad +(3.0 - 1.4391 - 1.3456 * 1.0)^2 \\ &\quad + \cdots + (4.0 - 1.4391 - 1.3456 * 1.5)^2 \\ &= 3.44334 \end{aligned}$$

2. $RSS = TSS - ESS$ と分解できる (末尾の証明を参照のこと)。

$TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$: Y (実測値) の総変動 (Total Sum)

$ESS = \sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})^2$: $\hat{Y}$ (予測値) の総変動 (Explained Sum)

3. 決定係数 R^2 を次のように定めて、これを、回帰式の説明力 (既に与えられているデータに対する予測精度) の尺度とする。

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

$\bar{Y} = 3$ であるから、いまの場合には、$TSS = (3-3)^2 + (3-3)^2 + (4-3)^2 + (1-3)^2 + (4-3^2) = 6$ 。したがって

$$R^2 = 1 - \frac{3.44334}{6} = 0.57389$$

$ESS \geq 0, RSS \geq 0, TSS \geq 0$ より、

$$\boxed{0 \leq R^2 \leq 1}$$

- $R^2 = 0$ のとき $\rightarrow$ $RSS = TSS$ つまり $ESS = 0$, 説明力ゼロ
- $R^2 = 1$ のとき $\rightarrow$ $RSS = 0$ つまり $ESS = TSS$, 説明力 100 %

例題

disc5.xls のデータを用いて、次のことを調べてみよう。

- 日本経済の限界消費性向 (MPC, $\Delta C/\Delta Y$) は？
- 投資関数の説明変数として R と OS はどちらが better か？
- $\Delta I/\Delta R =$ ？
- $\Delta I/\Delta OS =$ ？

5 重回帰と要因分解

参考: $RSS = TSS - ESS$ の証明

まず、$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$ より残差の和はゼロとなる。

$$\sum_{i=1}^{n} \hat{u}_i = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta}X_i) = n\bar{Y} - n\hat{\alpha} - n\hat{\beta}\bar{X} = 0$$

また $\hat{\beta} = S_{XY}/S_X^2$ であるから、残差と説明変数は「直交する」。つまり、

$$\begin{aligned}
\sum_{i=1}^{n} \hat{u}_i X_i &= \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta}X_i)X_i \\
&= \sum_{i=1}^{n} \{Y_i - (\bar{Y} - \hat{\beta}\bar{X}) - \hat{\beta}X_i\}X_i \\
&= \sum_{i=1}^{n} (Y_i - \bar{Y})X_i - \hat{\beta}\sum_{i=1}^{n} (X_i - \bar{X})X_i \\
&= \sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X}) - \hat{\beta}\sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X}) \\
&= (nS_{XY}) - \hat{\beta}(nS_X^2) = 0
\end{aligned}$$

これらより、残差と予測値も直交する。つまり、

$$\sum_{i=1}^{n} \hat{Y}_i \hat{u}_i = \sum_{i=1}^{n} (\hat{\alpha} + \hat{\beta}X_i)\hat{u}_i = \hat{\alpha}\sum_{i=1}^{n} \hat{u}_i + \hat{\beta}\sum_{i=1}^{n} X_i \hat{u}_i = 0$$

これよりさらに、

$$\sum_{i=1}^{n} \hat{Y}_i^{\,2} = \sum_{i=1}^{n} \hat{Y}_i (Y_i - \hat{u}_i) = \sum_{i=1}^{n} \hat{Y}_i Y_i - \sum_{i=1}^{n} \hat{Y}_i \hat{u}_i = \sum_{i=1}^{n} \hat{Y}_i Y_i$$

また、予測値の平均 ($\bar{\hat{Y}}$) と実測値の平均 ($\bar{Y}$) は等しい。つまり、

$$\bar{\hat{Y}} = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \hat{u}_i) = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$$

以上の結果を利用すると

$$\begin{aligned}
RSS &= \sum_{i=1}^{n} \hat{u}_i^{\,2} = \sum_{i=1}^{n} \hat{u}_i (Y_i - \hat{Y}_i) = \sum_{i=1}^{n} \hat{u}_i (Y_i - \hat{\alpha} - \hat{\beta}X_i) \\
&= \sum_{i=1}^{n} \hat{u}_i Y_i = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)Y_i = \sum_{i=1}^{n} (Y_i^2 - \hat{Y}_i^{\,2}) \\
&= \sum_{i=1}^{n} (Y_i^2 - \hat{Y}_i^{\,2}) - n\bar{Y}^2 + n\bar{\hat{Y}}^{\,2} \\
&= \sum_{i=1}^{n} (Y_i^2 - 2Y_i\bar{Y} + \bar{Y}^2) - \sum_{i=1}^{n} (\hat{Y}_i - 2\hat{Y}_i\bar{\hat{Y}} + \bar{\hat{Y}}^{\,2}) \\
&= \sum_{i=1}^{n} (Y_i - \bar{Y})^2 - \sum_{i=1}^{n} (\hat{Y}_i - \bar{\hat{Y}})^2 \\
&= TSS - ESS
\end{aligned}$$