3.4 決定係数 (回帰直線の有効性)

これまでは 2 変量の関係を回帰直線 $y = \hat{a} + \hat{b}x$ で表したが, いつもこのような統計手法を使えるかというと, そうとはいえない. では, データ $(x_1, y_1), ..., (x_n, y_n)$ に対して回帰直線を使うことが有効であるかないかを判断するにはどうすればよいのか?

ここで 1 つの判断材料を紹介する. 回帰直線 $y = \hat{a} + \hat{b}x$ の x の部分に $x_1, ..., x_n$ を代入したときの $\hat{y}_1, ..., \hat{y}_n$ を y_i の**予測値**もしくは**理論値**という.

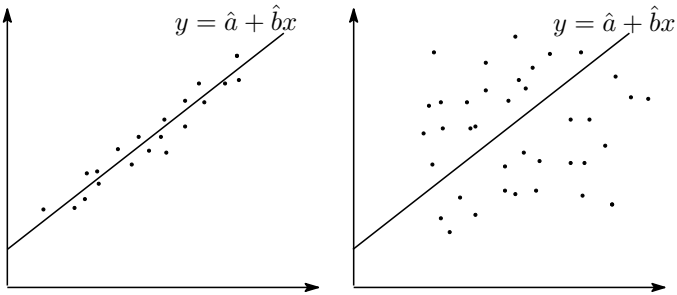
つまり y_i の**予測値**は $\hat{y}_i = \hat{a} + \hat{b}x_i$ $(i = 1, ..., n)$ となる. このとき, 以下の R^2 を**決定係数**とよぶ.

$$R^2 = \frac{\text{予測値}\hat{y}_1, ..., \hat{y}_n\text{の分散}}{y_1, ..., y_n\text{の分散}} = \frac{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{3.3}$$

ここで $\bar{y} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i$, $\bar{\hat{y}} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \hat{y}_i$ とする.

なんだか複雑な式であると思われるかもしれないが, Excel で自動的に計算してくれるので心配しなくてもよい.

ここで R^2 の数値の意味を説明する. R^2 はデータの何%を回帰直線でうまく表せているかを示している. このため R^2 は 1 に近いほど, 回帰直線でうまくデータを表せていることを意味する. これは**回帰直線の当てはまりが良い**という表現がよく使われる. 逆に R^2 が 0 に近い値をとるときは, 下図のように, 回帰直線 $y = \hat{a} + \hat{b}x$ は 効果がない ということを意味する.



当てはまりが良い
$R^2 \fallingdotseq 1$

当てはまりが悪い
$R^2 \fallingdotseq 0$