確率統計S　前回の解答　2017.12.22 分

次のデータが与えられたときの、回帰直線と相関係数を求めよ。

番号	1	2	3	4	5	6
x	7	9	11	13	15	17
y	62.1	66.6	50.1	46.4	41.3	35.3

x と y のそれぞれの標本平均、標本分散を求めると

$\bar{x} = 12.0, \quad s_x^2 = 14.0 \qquad \bar{y} = 50.3, \quad s_y^2 = 145.236$

相関係数と回帰直線を求めるために $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$ の値を求めると。

《$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ の場合》

番号	1	2	3	4	5	6	計
x	7	9	11	13	15	17	72
y	62.1	66.6	50.1	46.4	41.3	35.3	301.8
$x - \bar{x}$	-5	-3	-1	1	3	5	0
$y - \bar{y}$	11.8	16.3	-0.2	-3.9	-9	-15	0
積	-59	-48.9	0.2	-3.9	-27	-75	-213.6

相関係数：$r = \dfrac{-213.6}{(6-1)\sqrt{14.0}\sqrt{145.2}} = -0.94751\cdots = -0.948$

回帰直線:$\hat{\beta} = \dfrac{-213.6}{(6-1)\times 14.0} = -3.0514\cdots = -3.05, \quad \alpha = 50.3 - (-3.05)\times 12.00 = 86.9$

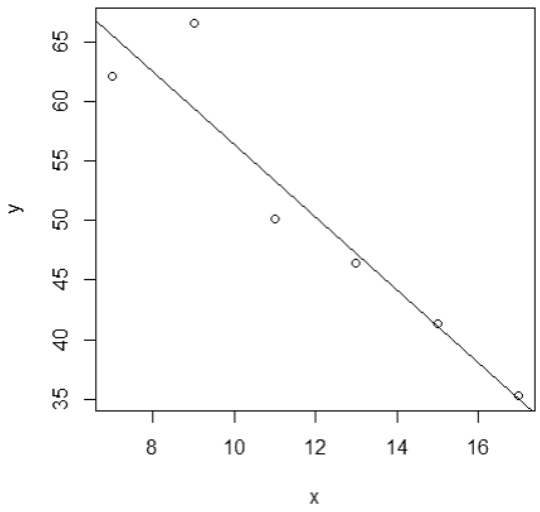
よって回帰直線は $y = -3.05x + 86.9$

《$\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$ の場合》

番号	1	2	3	4	5	6	計
x	7	9	11	13	15	17	72
y	62.1	66.6	50.1	46.4	41.3	35.3	301.8
積	434.7	599.4	551.1	603.2	619.5	600.1	3408

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \\ &= 3408 - 6 \times 12.0 \times 50.3 \\ &= -213.6\end{aligned}$$

なので、上の計算と一致している。



回帰直線における係数の区間推定と検定について

回帰直線のモデル式は誤差を考慮して

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

と表す。このモデル式で x_i と y_i はデータとして与えられた既知の値、α と β は真の値が存在するが未知なので推定する値であり、誤差については母平均 0，母分散 σ^2 の正規分布に従うと仮定する。最尤推定量を求めたいので、尤度関数を計算すると、

$$\begin{aligned} L(\alpha,\beta,\sigma^2) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum\varepsilon_i^2} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n \varepsilon_i^2\right\} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n \Big\{y_i - (\alpha+\beta x_i)\Big\}^2\right\} \end{aligned}$$

である。この尤度関数を最大にする α,β は、$\displaystyle\sum_{i=1}^n \Big\{y_i - (\alpha+\beta x_i)\Big\}^2$ を最小にすればよいので、最小 2 乗法の結果と一致する。

上記の方法で求めた値を $\hat{\alpha},\hat{\beta}$ とすると、$\hat{\alpha},\hat{\beta}$ の分布はそれぞれ母平均 α,β，母分散

$$\sigma_{\hat{\alpha}}^2 = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right), \quad \sigma_{\hat{\beta}}^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i-\bar{x})}$$

の正規分布に従う。実際には σ^2 が未知なので、推定した値（不偏推定量）

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n \Big\{y_i - (\hat{\alpha}+\hat{\beta}x_i)\Big\}^2$$

で置き換える。すると

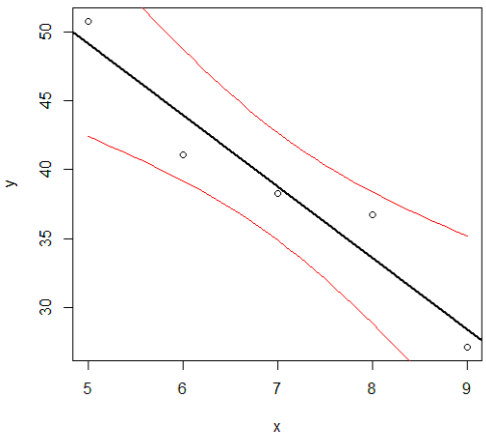
$$T_\alpha = \frac{\hat{\alpha}-\alpha}{\hat{\sigma}_{\hat{\alpha}}^2}, \quad T_\alpha = \frac{\hat{\beta}-\beta}{\hat{\sigma}_{\hat{\beta}}^2}$$

はそれぞれ自由度 $n-2$ の t 分布に従う。この結果を使えば区間推定や検定ができる。

さらに、任意の x_0 に対応する y_0 の区間推定については

$$\hat{\alpha}+\hat{\beta}x_0 \pm t_{n-2}(\alpha)\sqrt{\left\{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right\}\hat{\sigma}^2}$$

となるので、$x_0 = \bar{x}$ のときに、一番狭くなる双曲線のような形になっている。



資料置場

https://www.gen.kanagawa-it.ac.jp/takeda/class/