§3 データ分析で注意すべき点　演習問題　解答

問題の難易度の目安【易】★☆☆　【基礎】★★☆　【標準】★★★

1 (★☆☆)(はずれ値)

同じ実験を 6 回繰り返して，次の測定データが得られたとする：

実験データ	2.3	2.5	1.8	3.0	2.4	0.6

(1) 平均と分散を求めよ．

(2) 0.6 だけが異常に低い値と思われる．このデータだけ取り除いた，残り 5 個のデータの平均と分散を求めよ．

(3) データ 0.6 の扱いについて考察せよ．

解 (1)　平均は，$x=\dfrac{2.3+2.5+1.8+3.0+2.4+0.6}{6}=\mathbf{2.1}$ であり，分散 σ^2 は

$$\sigma^2=\frac{2.3^2+2.5^2+1.8^2+3.0^2+2.4^2+0.6^2}{6}=\mathbf{0.57}.$$

(2)　0.6 を除いた 5 つのデータに関する平均は，$\dfrac{2.3+2.5+1.8+3.0+2.4}{5}=\mathbf{2.4}$ であり，分散 $\tilde{\sigma}^2$ は

$$\tilde{\sigma}^2=\frac{2.3^2+2.5^2+1.8^2+3.0^2+2.4^2}{5}=\mathbf{0.15}.$$

すなわち，データ 0.6 を取り除くと，分散が大幅に小さくなり，同じ実験を繰り返した実験データとしての妥当性が高まったと考えられる．

(3)　0.6 は，はずれ値として除外するのが適当と考えられる．∎

Remark

データの中には信頼できないものが混ざっている可能性が常にある．明らかに他のデータとかけ離れた値のデータをはずれ値といい，それを取り除くことでデータの精度が高くなると考えられが，はずれ値が本当に間違ったデータなのか，実は正しいデータなのかは分からない．データの性格を考えたり，はずれ値を入れた場合と取り除いた場合で結果を比較してみたりして，推測する必要がある．はずれ値の存在は常にデータ解析の障害となる．

2 (★★☆)(1 次式によるデータの変換①)

$\alpha, \beta, \gamma, \delta$ を定数とし，$\beta \neq 0, \delta \neq 0$ とする．2 次元データ $(x_1, y_1), \ldots, (x_N, y_N)$ を

$$u_k := \frac{x_k - \alpha}{\beta}, \quad v_k := \frac{y_k - \gamma}{\delta}$$

により，データ $(u_1, v_1), \ldots, (u_N, v_N)$ に変換する．変換前後の共分散 s_{xy} と s_{uv} に対して，等式

$$s_{xy} = \beta\delta s_{uv}$$

が成り立つことを示せ．また，$\beta, \delta > 0$ ならば，変換前後の相関係数 r_{xy} と r_{uv} に関して，等式

$$r_{xy} = r_{uv}$$

が成り立つことを示せ．

解 $u_k = \frac{x_k-\alpha}{\beta}, \quad v_k = \frac{y_k-\gamma}{\delta} \iff x_k = \beta u_k + \alpha, \quad y_k = \delta v_k + \gamma$ であるから，平均をとると $\bar{x} = \beta\bar{u} + \alpha, \quad \bar{y} = \delta\bar{v} + \gamma$ を得る．これより，

$$x_k - \bar{x} = \beta(u_k - \bar{u}), \quad y_k - \bar{y} = \delta(v_k - \bar{v}).$$

したがって，

$$s_{xy} = \frac{1}{N}\sum_{k=1}^{N}(x_k - \bar{x})(y_k - \bar{y}) = \beta\delta\frac{1}{N}\sum_{k=1}^{N}(u_k - \bar{x})(v_k - \bar{y}) = \beta\delta s_{uv}.$$

また，$\beta, \delta > 0$ のとき $s_x = \beta s_u, \quad s_y = \delta s_v$ であるから，

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\beta\delta s_{uv}}{\beta s_u \delta s_v} = r_{uv}.$$

■

3 (★★☆)(1 次式によるデータの変換②)

ある駅の不動産屋で 8 件の賃貸物件 (1LDK) の駅からの徒歩時間 (分) と 1 ヶ月の賃貸料 (万) を調べたところ次の表のようになった：

徒歩時間	1	3	3	4	6	7	7	9
賃貸料	8	6	5	7	6	5	6	5

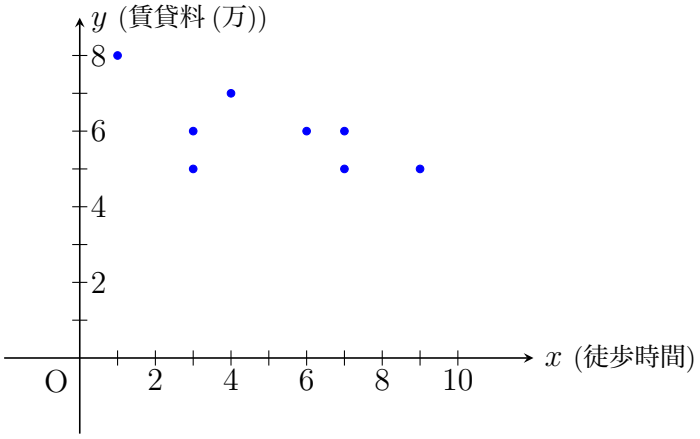
徒歩時間を変数 x，賃貸料を変数 y とし，次の問いに答えよ．

(1) 散布図を描け．

(2) 相関係数 r_{xy} を求めよ．また，徒歩時間と賃貸料にはどの程度の相関があると言えるか．

(3) 回帰直線を求め，散布図に描け．

解 (1)　散布図は以下のようになる：



(2)　$\bar{x} = \frac{1+3+3+4+6+7+7+9}{8} = 5$ かつ $\bar{y} = \frac{8+6+5+7+6+5+6+5}{8} = 6$ であるから，$u_k := x_k - 5$, $v_k := y_k - 6$ $(k = 1, \ldots, 8)$ とおくと，

$$\bar{u} = \bar{v} = 0.$$

u, v に関するデータ一覧は以下の通り：

徒歩時間 u	-4	-2	-2	-1	1	2	2	4
賃貸料 v	2	0	-1	1	0	-1	0	-1

u の分散は $\sigma_u^2 = \frac{1}{8}\sum_{k=1}^{8}(u_k - \bar{u})^2 = \frac{1}{8}\sum_{k=1}^{8} u_k^2 = 6.25$，$v$ の分散は $\sigma_v^2 = \frac{1}{8}\sum_{k=1}^{8}(v_k - \bar{u})^2 = \frac{1}{8}\sum_{k=1}^{8} v_k^2 = 1.$
よって，

$$\sigma_u = \sqrt{6.25} = 2.5, \quad \sigma_v = 1.$$

一方，uv の共分散 σ_{uv} は

$$\sigma_{uv} = \frac{1}{8}\sum_{k=1}^{8}(u_k - \bar{u})(v_k - \bar{u}) = \frac{1}{8}\sum_{k=1}^{8} u_k v_k = -1.625$$

ゆえに，**2** の結果を用いて，

$$r_{xy} = r_{uv} = \frac{\sigma_{uv}}{\sigma_u \sigma_v} = -\frac{1.625}{2.5} = \mathbf{-0.65}.$$

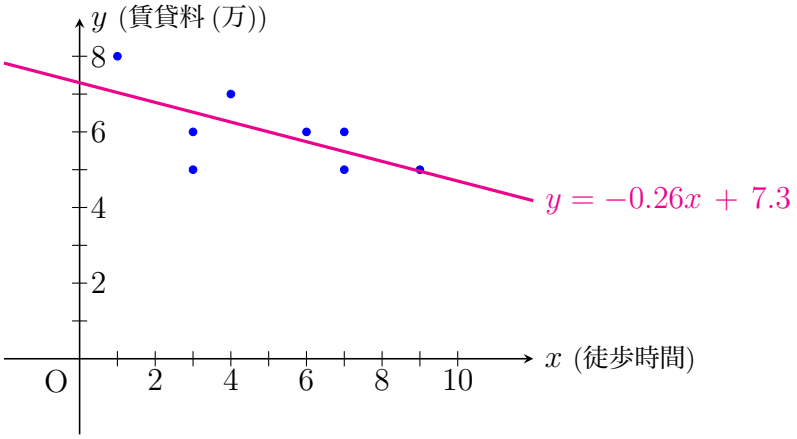
これより $0.4 \leqq |r_{xy}| \leqq 0.7$ であるから，徒歩時間 x と賃貸料 y には**おおむね相関関係がある.**

(3)　再び **2** の結果を用いて，

$$\hat{a} := \frac{r_{xy}}{\sigma_x} = \frac{r_{uv}}{\sigma_u} = \frac{-0.65}{2.5} = -0.26$$

$$\hat{b} := \bar{y} - \hat{a}\bar{x} = 6 - (-0.26) \times 5 = 7.3$$

とおくとき，求める回帰直線は，$y = \hat{a}x + \hat{b} = \mathbf{-0.26}\boldsymbol{x} + \mathbf{7.3}$ であり，散布図に書き込むと，下の桃色の直線となる：



Remark

2 つの種類のデータ間に相関関係があるからといって，必ずしも因果関係があるわけではない．すなわち，説明変数となるデータが原因で，目的変数となるデータの値が変化すると結論することは証拠不十分である．

(例 1)　因果関係が逆：説明変数と目的変数の設定が逆．

(例 2)　疑似相関：説明変数 x と目的変数 y には直接的な因果関係はなく，なんらかの見えない要因 z があって，z と x，z と y の間のそれぞれに因果関係がある可能性がある．

(例 3)　偶然の一致：単なる偶然でデータ間に相関関係があっただけで，原因となる要素がないか，あるにしても非常に複雑な要因のため，因果関係を見出す意味を持たない．

4 (★★☆)(第 3 ファクターの影響を無視した場合の相関係数)

3 つ組データ $(x_1, y_1, z_1), \ldots, (x_N, y_N, z_N)$ を考える．以下 $(y_1, z_1), \ldots, (y_N, z_N)$ の関係を調べる際に，第 3 ファクター x_i の影響がなくなるようにするために，

$$\begin{cases} \hat{y}_i := \dfrac{s_{xy}}{s_x^2} x_i + \bar{y} - \dfrac{s_{xy}}{s_x^2} \bar{x} \\ \hat{z}_i := \dfrac{s_{xz}}{s_x^2} x_i + \bar{y} - \dfrac{s_{xz}}{s_x^2} \bar{x} \end{cases}$$

とおいて，$(y_1 - \hat{y}_1, z_1 - \hat{z}_1), \ldots, (y_N - \hat{y}_N, z_N - \hat{z}_N)$ の相関係数を調べよう．以下簡単のため

$$y_i' := y_i - \hat{y}_i, \qquad z_i' := z_i - \hat{z}_i, \quad i = 1, \ldots, N$$

とおく．

(1) $\bar{y'} = 0$，$\bar{z'} = 0$ を確かめよ．また，共分散 $s_{y'z'}$ が

$$s_{y'z'} = s_{yz} - \frac{s_{xy}s_{xz}}{s_x^2}$$

で与えられることを示せ．

(2) $s_{y'}^2$，$s_{z'}^2$ について

$$s_{y'}^2 = s_y^2\left\{1 - \left(\frac{s_{xy}}{s_x s_y}\right)^2\right\}, \qquad s_{z'}^2 = s_z^2\left\{1 - \left(\frac{s_{xz}}{s_x s_z}\right)^2\right\}$$

で与えられることを示せ．

(3) (1), (2) を用いて，第 3 ファクター x を除いた新しいデータ (y', z') に関する相関係数 $r_{y'z'\backslash x} := \dfrac{s_{y'z'}}{s_{y'}s_{z'}}$ は

$$r_{y'z'\backslash x} = \frac{r_{yz} - r_{xy}r_{xz}}{\sqrt{1 - r_{xy}^2}\sqrt{1 - r_{xz}^2}}$$

で与えられることを示せ．$r_{y'z'\backslash x}$ を**偏相関係数**という．ここに，r_{yz} はデータ (y, z) に関する相関係数であり，r_{xy}，r_{xz} についても同様である．

解 (1)　$\overline{y'} = 0$ のみを示す ($\overline{z'} = 0$ は同様に示される)．$y_i' := y_i - \hat{y}_i$ $(i = 1, \ldots, N)$ に対して，平均は

$$\begin{aligned}
\overline{y'} = \frac{1}{N}\sum_{i=1}^{N} y_i' &= \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i) \\
&= \bar{y} - \frac{1}{N}\sum_{i=1}^{N}\left(\frac{s_{xy}}{s_x^2}x_i + \bar{y} - \frac{s_{xy}}{s_x^2}\bar{x}\right) \\
&= \bar{y} - \left[\frac{s_{xy}}{s_x^2}\bar{x} + \bar{y} - \frac{s_{xy}}{s_x^2}\bar{x}\right] = 0.
\end{aligned}$$

次に，$\overline{y'} = \overline{z'} = 0$ であるから，

$$\begin{aligned}
s_{y'z'} &= \frac{1}{N}\sum_{i=1}^{N}\left(y_i' - \overline{y'}\right)\left(z_i' - \overline{z'}\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)\,(z_i - \hat{z}_i) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left((y_i - \bar{y}) - \frac{s_{xy}}{s_x^2}(x_i - \bar{x})\right)\left((z_i - \bar{z}) - \frac{s_{xz}}{s_x^2}(x_i - \bar{x})\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})(z_i - \bar{z}) - \frac{1}{N}\sum_{i=1}^{N}\frac{s_{xz}}{s_x^2}(y_i - \bar{y})(x_i - \bar{x})
\end{aligned}$$

$$
\begin{aligned}
&\quad - \frac{1}{N}\sum_{i=1}^{N}\frac{s_{xy}}{s_x^2}(x_i-\bar{x})(z_i-\bar{z}) + \frac{1}{N}\sum_{i=1}^{N}\frac{s_{xy}s_{xz}}{s_x^4}(x_i-\bar{x})^2 \\
&= s_{yz} - \frac{s_{xz}}{s_x^2}s_{xy} - \cancel{\frac{s_{xy}}{s_x^2}s_{xz}} + \cancel{\frac{s_{xy}s_{xz}}{s_x^4}s_x^2} \\
&= s_{yz} - \frac{s_{xy}s_{xz}}{s_x^2}.
\end{aligned}
$$

(2)　y' の分散 $s_{y'}^2$ について，

$$
\begin{aligned}
s_{y'}^2 = \frac{1}{N}\sum_{i=1}^{N}\left(y_i'-\overline{y'}\right)^2 &= \frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\left((y_i-\bar{y}) - \frac{s_{xy}}{s_x^2}(x_i-\bar{x})\right)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}(y_i-\bar{y})^2 - \frac{2}{N}\sum_{i=1}^{N}\frac{s_{xy}}{s_x^2}(x_i-\bar{x})(y_i-\bar{y}) + \frac{1}{N}\sum_{i=1}^{N}\frac{s_{xy}^2}{s_x^4}(x_i-\bar{x})^2 \\
&= s_y^2 - 2\frac{s_{xy}^2}{s_x^2} + \frac{s_{xy}^2}{s_x^4}s_x^2 \\
&= s_y^2\left\{1-\left(\frac{s_{xy}}{s_x s_y}\right)^2\right\}.
\end{aligned}
$$

同様に，$s_{z'}^2 = s_z^2\left\{1-\left(\frac{s_{xz}}{s_x s_z}\right)^2\right\}$ も示される．

(3)　(1),(2) より，x を除いた新しいデータ (y', z') に関する相関係数 $r_{y'z'\backslash x} \equiv \dfrac{s_{y'z'}}{s_{y'}s_{z'}}$ は，

$$
r_{y'z'\backslash x} = \frac{s_{y'z'}}{s_{y'}s_{z'}} = \frac{r_{yz}-r_{xy}r_{xz}}{\sqrt{1-r_{xy}^2}\sqrt{1-r_{xz}^2}}
$$

で与えられる．∎

5 (★★☆)(偏相関係数)

ある会社の社員の体重と年収について，データを取ってみたところ，それらの間には相関関係があった．3 次元データ (x, y, z) を

$$(x, y, z) = (\text{年齢},\ \text{体重},\ \text{年収})$$

とする．ここで，体重 y と年収 z の相関係数 r_{yz} は $r_{yz} = 0.90$ で高い相関を持っていた．ところが A さんはこの相関に疑問を感じたため，年収・体重の両方に影響を及ぼ

している第3のファクターとして年齢xがあるのではないかと思い，実際に調べてみたところ，

・年齢x　体重yの相関係数$r_{xy}=0.75$

・年齢x　年収zの相関係数$r_{xz}=0.80$

であった．このとき年齢xの影響を取り除いた年収y，体重zの偏相関係数を求めよ．

解 4 より，年齢xの影響を取り除いた年収y，体重zの偏相関係数$r_{yz\backslash x}$は

$$r_{yz\backslash x}=\frac{r_{yz}-r_{xy}r_{xz}}{\sqrt{1-r_{xy}^2}\sqrt{1-r_{xz}^2}}=\frac{0.90-0.80\times 0.75}{\sqrt{1-0.75^2}\sqrt{1-0.8^2}}\approx\frac{0.3}{0.66\times 0.6}\approx\mathbf{0.66}.$$

■