総合科目　数理・自然 6「統計学入門」講義レジュメ　(今野)

直線のあてはめ

x 変量を「年齢」，y 変量を「血圧」とする．この場合，x は y をある程度左右する．このとき

1. x：独立変数（説明変数）
2. y：従属変数（被説明変数）

という．これを y の x 上への回帰という．年齢・血圧のデータを観れば，おおよそ 1 次式

$$y = bx + a \tag{1}$$

という関係式がみてとれる．

式 (1) における a と b をデータからみつけたい！

最小二乗法

$$L = \sum_{i=1}^{n} \{y_i - (bx_i + a)\}^2$$

を最小にする a と b を用いる．これは正規方程式

$$\begin{aligned} na + (\sum_{i=1}^{n} x_i)b &= \sum_{i=1}^{n} y_i \\ (\sum_{i=1}^{n} x_i)a + (\sum_{i=1}^{n} x_i^2)b &= \sum_{i=1}^{n} x_i y_i \end{aligned}$$

を解けばよいことがわかる．実際,

$$\hat{b} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}_n \bar{y}_n}{\sum_{i=1}^{n} x_i^2 - n(\bar{x}_n)^2}, \qquad \bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x, \tag{2}$$

$$\hat{a} = \bar{y}_n - b\bar{x}_n, \qquad \bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i, \tag{3}$$

であたられる．得られた a と b に方程式

$$y = \hat{b}x + \hat{a}$$

を y の x 上への回帰方程式という．

例

x	2	3	3	6	4	2	1	5
y	0.27	1.41	2.19	2.83	2.19	1.81	0.85	3.05

について回帰方程式を求めてみよう.

$$\begin{aligned}\hat{b} &= \frac{56.80 - 14.60 \times 26/8}{104 - 26 \times 26/8} = \frac{9.35}{1.95} = 0.48\\ \hat{a} &= \frac{14.60}{8} - 0.48 \times \frac{26}{8} = 0.27\end{aligned}$$

となり,

$$y = 0.48x + 0.27$$

となる.

つぎに, x の y 上への回帰方程式

$$x = dy + c$$

を求めてみよう.

$$\begin{aligned}\hat{d} &= \frac{56.80 - 14.60 \times 26/8}{32.96 - 14.60 \times 14.60/8} = \frac{9.35}{6.35} = 1.48\\ \hat{c} &= \frac{26}{8} - 1.48 \times \frac{14.60}{8} = 0.55\end{aligned}$$

となる. したがって, x の y 上への回帰方程式は

$$x = 1.48y + 0.55$$

となる.

例 つぎのデータについて回帰方程式

$$y = bx + a$$

を求めよう.

x	2	3	3	6	4	2	1	5
y	3.27	1.41	2.19	2.83	2.19	1.81	0.85	3.05

$$\begin{aligned}\sum_{i=1}^{8} x_i y_i &= 62.80 \quad \sum_{i=1}^{8} x_i = 26.00, \quad \sum_{i=1}^{8} y_i = 17.60, \quad \sum_{i=1}^{8} x_i^2 = 104.00,\\ \sum_{i=1}^{8} y_i^2 &= 43.58\end{aligned}$$

となる．

$$\hat{b} = \frac{62.80 - 26 \times 17.6/8}{104 - 30 \times 26/2} = \frac{5.6}{1.95} = 2.87$$

と

$$\hat{a} = \frac{17.60}{8} - 2.87\frac{26.00}{8} = -7.13$$

となる．よって，回帰方程式は

$$y = 2.87x - 7.13$$

となる．

回帰直線を作成するために R プログラム

```
> plot(height,weight)
> abline(ls.result)
> ls.result<-lsfit(height,weight)
> ls.result\$coef
  Intercept           X
-74.8740979   0.7622423
> plot(height,weight)
> abline(ls.result)
```

決定係数

表す変量 x と y の関係の強さを表す相関係数

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y}_n)^2}}$$

と直線の当てはまりの良さの間には関係がある．まず，回帰直線の傾きは

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2} = r\frac{S_y}{S_x}$$

と書き直すことができる．ただし，

$$S_x = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}, \qquad S_y = \sqrt{\sum_{i=1}^{n}(y_i - \bar{y}_n)^2}$$

である．また，回帰方程式を用いて x_i により予想された y の値 $\hat{y}_i = \hat{b}x_i + \hat{a}$ と y_i との差の二乗和 $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ を書きかえる：(3) と (2) に注意すれば

$$\begin{aligned} SS_E &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n \{(y_i - \bar{y}_n) - \hat{b}(x_i - x_i)\}^2 \\ &= \sum_{i=1}^n (y_i - \bar{y}_n)^2 - 2\sum_{i=1}^n \hat{b}(y_i - \bar{y}_n)(x_i - x_i) + \sum_{i=1}^n \hat{b}^2 (x_i - x_n)^2 \\ &= \sum_{i=1}^n (y_i - \bar{y}_n)^2 - \sum_{i=1}^n \frac{\{(y_i - \bar{y}_n)(x_i - x_i)\}^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \\ &= (1 - r^2) \sum_{i=1}^n (y_i - \bar{y}_n)^2 \end{aligned} \tag{4}$$

となる．よって，$r = \pm 1$ のとき，$SS_E = 0$ となり，y_i と $\hat{y}_i$ は一致し，y は x によって完全に決定される．r^2 は独立変数 x が従属変数 y を決定する強弱の度合いを表している．そこで，r^2 を決定係数という．$SS_0 = \sum_{i=1}^n (y_i - \bar{y}_n)^2$ は変量 y の散らばりである．

$$\begin{aligned} \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}_n) &= \sum_{i=1}^n (y_i - \hat{b}x_i - \bar{y}_n + \hat{b}\bar{x}_n)(\hat{b}x_i + \bar{y}_n - \hat{b}\bar{x}_n - \bar{y}_n) \\ &= \hat{b}\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n) - \hat{b}^2 \sum_{i=1}^n (x_i - \bar{x}_n)^2 = 0 \end{aligned}$$

となることに注意すれば

$$\begin{aligned} \sum_{i=1}^n (y_i - \bar{y}_n)^2 &= \sum_{i=1}^n (y_i - \hat{y}_i + \hat{y}_i - \bar{y}_n)^2 \\ &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + 2\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}_n) + \sum_{i=1}^n (\hat{y}_i - \bar{y}_n)^2 \\ &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y}_n)^2 \\ &\equiv SS_E + SS_R \end{aligned}$$

となる．この式と (4) から

$$SS_R = r^2 SS_0$$

を得る．SS_R は回帰により減少した y の散らばりと考えることができる．

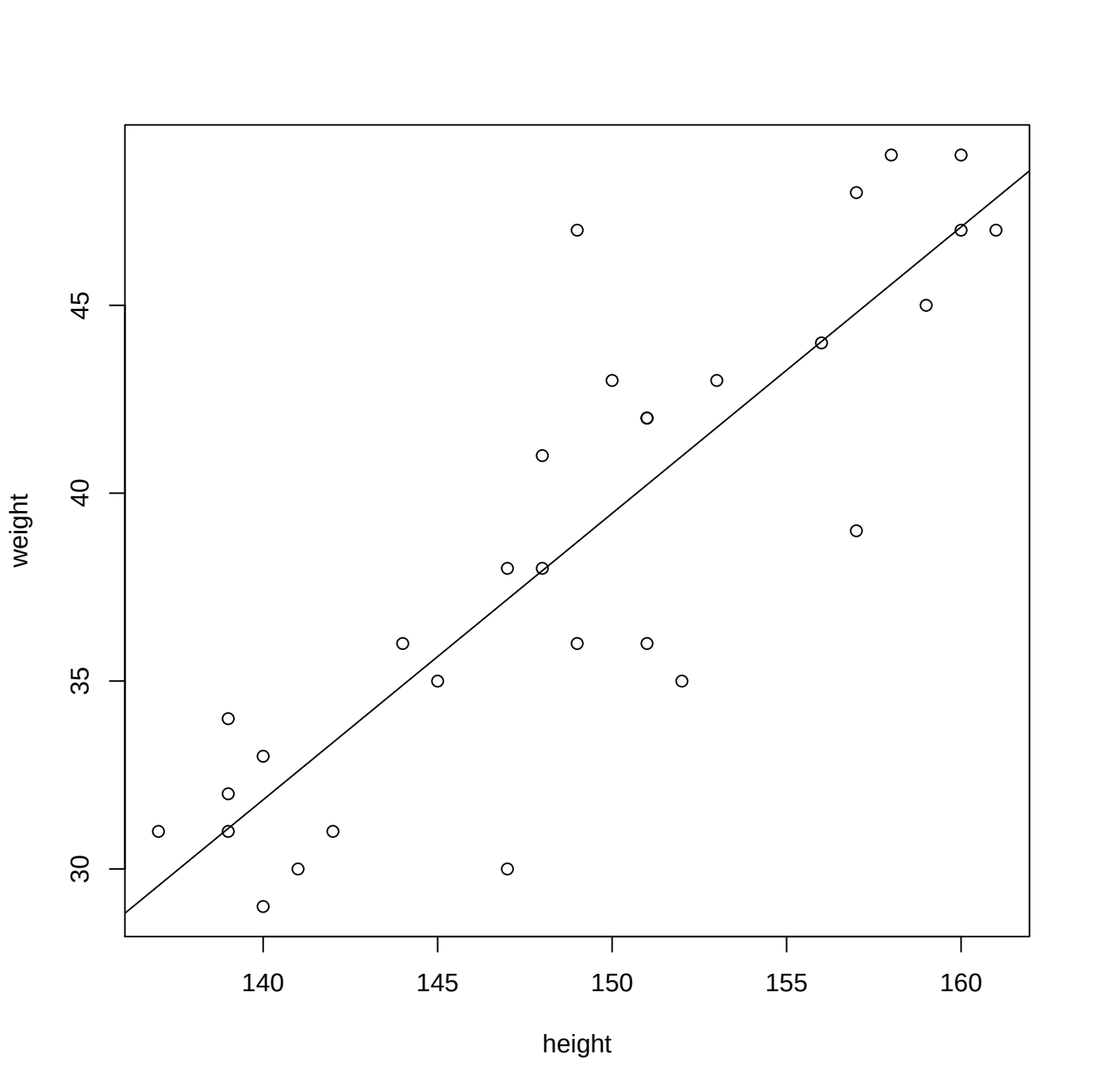
問題 4

下のデータは 1991 年から 1 年ごとのある地区の年間のゴミの排出量を調べた結果である,

x	1	2	3	4	5	6
y	49	44	42	33	27	12

以下の数表から，回帰方程式を求めよ. さらに，正しい回帰直線を用いて，1999 年のゴミの排出量の予測値を求め，その値を予測値として利用することの妥当性について議論せよ.

変量							合計
x	1	2	3	4	5	6	21
y	49	44	42	33	27	12	207
x^2	1	4	9	16	25	36	91
$x \cdot y$	49	88	126	132	135	72	602
y^2	2401	1936	1764	1089	729	144	8063


weight
30
35
40
45
140
145
150
155
160
height