5　2 変量のデータ

5.1　散布図

各データ を xy 平面の座標上にプロットした**散布図**を描いて変量間に関係（傾向）があるかを調べてみる.

```
> x
    V1 V2
1  148 41
2  160 49
3  159 45
4  153 43
5  151 42
6  140 29
7  156 49
8  137 31
9  149 47
10 169 47
11 151 42
12 157 39
13 157 48
14 144 36
> plot(x$V2,x$V1,xlab="height",ylab="weight")
> plot(x$V2,x$V1,xlab="height",ylab="weight",pch="x")
> colnames(x)<-c("height","weight")
> x
   height weight
1     148     41
2     160     49
3     159     45
4     153     43
5     151     42
6     140     29
7     156     49
8     137     31
9     149     47
10    169     47
11    151     42
12    157     39
13    157     48
14    144     36
> plot(x$weight,x$height)
```

5.2　最小 2 乗法と回帰直線

散布図全体をみて、データ全体として右上がり，左下がりの傾向が見受けられたら，その傾向を示す直線を
散布図に描いてみよう．直線の方程式は

$$y = ax + b$$

と与えれるが, a, b の値をさだめる一つの方法が最小 2 乗法である．n 個のデータを $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ とする．このとき,

$$Q(a, b) = \sum_{i=1}^{n} \{y_i - (ax_i + b)\}^2 \tag{1}$$

を最小とする a, b を用いる．Q を a, b について偏微分することにより

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

$$b = \frac{(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{(x_i - \bar{x}_n)\ 2} \tag{3}$$

である．

```
> x<-read.table("height-weight")
> x
    V1 V2
1  148 41
2  160 49
3  159 45
4  153 43
5  151 42
6  140 29
7  156 49
8  137 31
9  149 47
10 169 47
11 151 42
12 157 39
13 157 48
14 144 36
> plot(x$V2 ~ x$V1,xlab="height",ylab="weight",pch="x") # x 軸に身長, y 軸に体重
> abline(lm(x$V2~x$V1),lty=1)   # x 軸に身長, y 軸に体重の回帰直線を書き込む
> summary(lm(x$V2~x$V1)) # 回帰直線の傾き (x$V1) と切片 (Intercept) を得る

Call:
lm(formula = x$V2 ~ x$V1)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9074 -1.5983  0.6302  2.0926  6.9528

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -50.4740    19.7680  -2.553 0.025311 *
x$V1          0.6075     0.1297   4.685 0.000528 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.951 on 12 degrees of freedom
Multiple R-Squared: 0.6465,     Adjusted R-squared: 0.617
F-statistic: 21.95 on 1 and 12 DF,  p-value: 0.000528

> 160* 0.6075-50.4740
[1] 46.726
```

演習 6　(1 を a, b で偏微分することにより，Q が最小となるときの a, b の値が (2 と (3 であたえらることを示せ．

演習 7　データ ”baseball” について x 軸を ”year ” とし，y 軸を ”Central” と ” Pacific” とした散布図と回帰直線をそれぞれ求めよ．

```
> y<-read.table("baseball")
> y
   V1 V2 V3
1  84 28 13
2  85 29 12
3  86 29 16
4  87 31 18
5  88 31 21
6  89 31 23
7  90 31 22
8  91 32 24
9  92 35 24
10 93 34 24
>
> colnames(y)<-c("year","Central","Pacific")
> y
   year Central Pacific
1    84      28      13
2    85      29      12
3    86      29      16
4    87      31      18
5    88      31      21
6    89      31      23
7    90      31      22
8    91      32      24
9    92      35      24
10   93      34      24
>
```