多変量解析の基礎（回帰分析）
―理論とRによる演習―

本稿のWebページ

古橋武

多変量解析

- 回帰分析

について基礎理論を解説し，
Rによる演習を行います．

回帰分析とは

データ $P_1, P_2, \ldots, P_n$ が与えられたときに，このデータ分布を近似するモデルを同定する手法．

1入力1出力のとき

データ分布を直線で近似している例

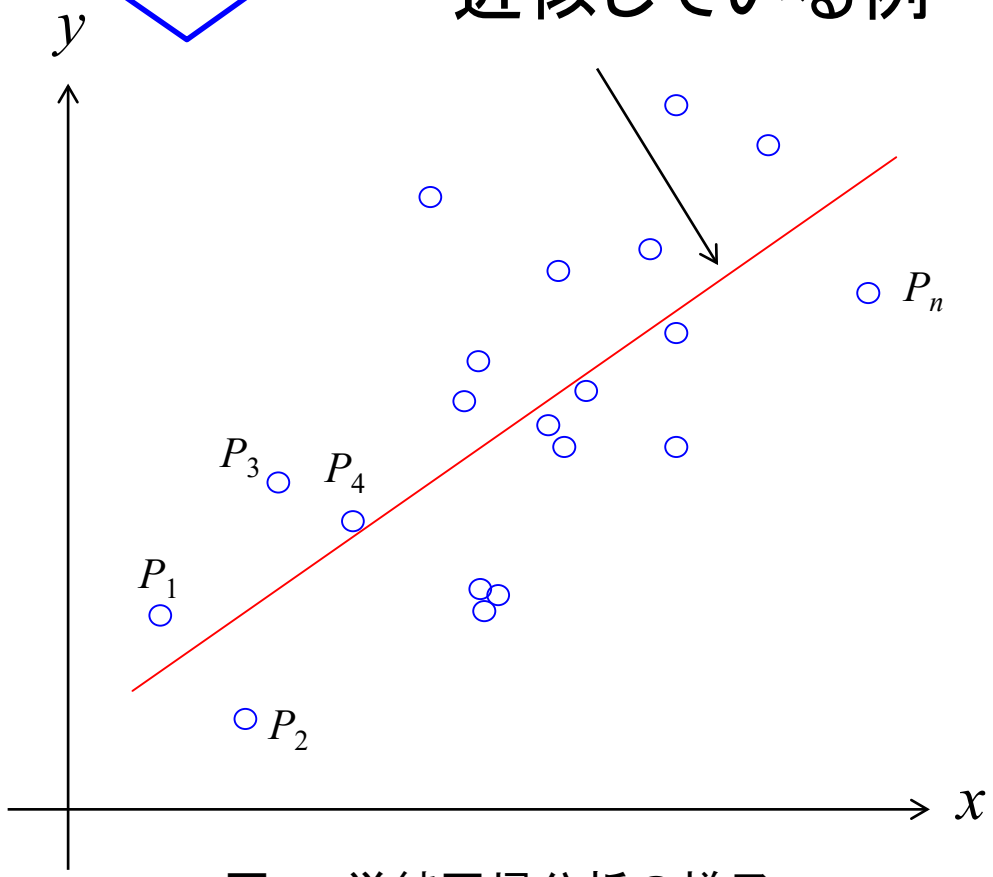
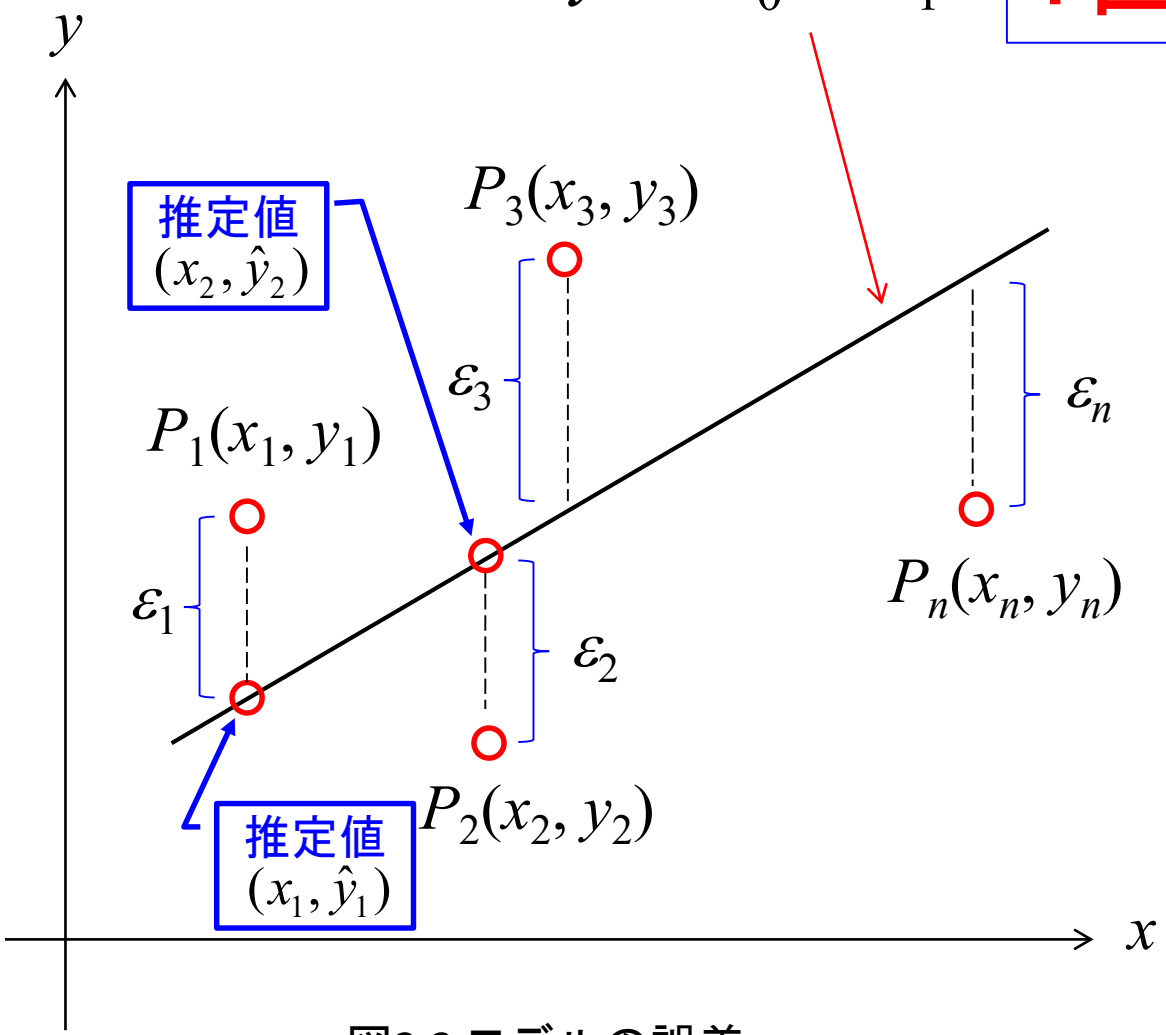


図2.1 単純回帰分析の様子

2. 単純回帰分析
2.1 基礎理論

$$\hat{y} = a_0 + a_1 x$$ **：回帰式**



$$E = \sum_{i=1}^{n} \varepsilon_i^2 \rightarrow \text{最小}$$

とする a_0, a_1 を求めます.

図2.2 モデルの誤差

$$E = \sum_{i=1}^{n} \varepsilon_i^2$$

$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (2.4)$$

$$= \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2$$

誤差Eを最小とするa_0, a_1は次式の通りです.

$$a_0 = \overline{y} - \overline{x} a_1$$

$$a_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \qquad (2.8)$$

２．２　Rによる計算

1. 「回帰分析」フォルダをパソコンの「マイドキュメント」にコピーしてください．
2. Rのインストールがまだの人は，http://cran.r-project.org/よりRの最新バージョンをインストールしてください．本稿ではR 3.0.2 for Windowsを用いた場合について解説します．無事インストールできるとデスクトップに「R　i386 3.0.2」のアイコンが現れます．64ビットパソコンでは「R x64 3.0.2」のアイコンも現れます．
3. R　i386 3.0.2（もしくはR x64 3.0.2）のアイコンをダブルクリックすることでRを立ち上げることができます．
4. 「ファイル」→「スクリプトを開く」とクリックしていくと「マイドキュメント」のフォルダが開かれます．
5. 「回帰分析」フォルダをダブルクリックして「**回帰分析_身体測定_基礎式.R**」のアイコンをダブルクリックすると(2.8)式の計算をするスクリプトが「**Rエディタ**」のウィンドウに開かれます．

表 2.1 身体測定結果

身長	座高
170.6	88.1
164.5	88.5
161	87.2
170.5	88.5
171	88.4
170	91.8
165	89.7
173	91
166.8	86.2
173.6	92
176.3	93
172.5	91.4
182	93.7
179	96.3
176.3	97
175.5	94.5
169	95.4
170.4	92.5
176.3	91
172.8	94.1

「Rエディタ」内の(2.8)式の計算をするスクリプト

```
x_身体測定 <- read.csv("C:/Users/Furuhashi/Documents/回帰分析/身体測定_身長_座高.csv")
x_身体測定

plot(x_身体測定$身長,x_身体測定$座高,col="red", pch=1)
XX <- as.matrix(x_身体測定)                                    #データフレームを行列へ変換

y <- XX[,2]                                                    # 座高データを抽出

x <- XX[,1]                                                    # 身長データを抽出

n <- length(x)                                                 # ベクトルxの要素数を得る

mean_y <- mean(y)                                              # 座高データの平均値を計算

mean_x <- mean(x)                                              # 身長データの平均値を計算

sum_xy <- t(x) %*% y                                           # Σxi*yi の計算

sum_xx <- t(x) %*% x                                           # Σxi^2 の計算

a1 <- (sum_xy - n * mean_x * mean_y)/(sum_xx - n * mean_x^2)   #a1の計算
a0 <- mean_y - mean_x * a1                                     #a0の計算


abline(a0, a1)                                                 # 回帰式の描画
```

スクリプトの1行目の実行

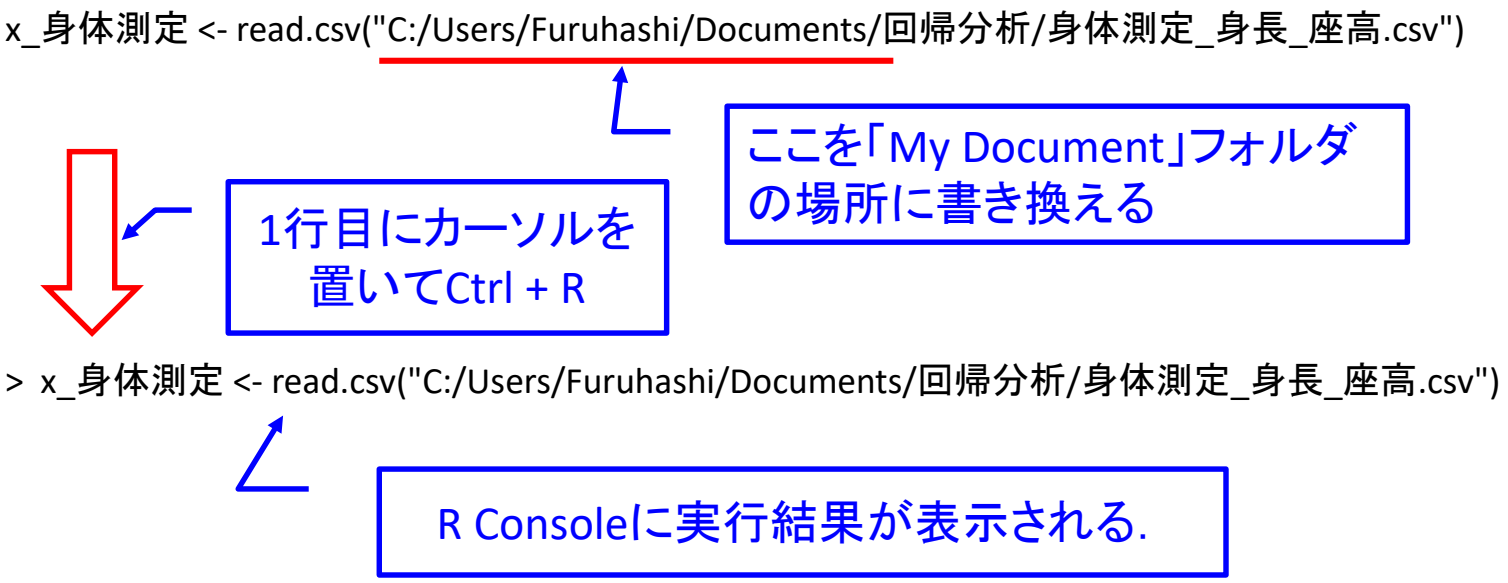


図2.3 スクリプトの1行目の実行

スクリプトの2行目の実行

x_身体測定

を実行すると以下が表示されます.

身体測定_身長_座高.csvを読み込んだ結果が「R Console」に表示されます.

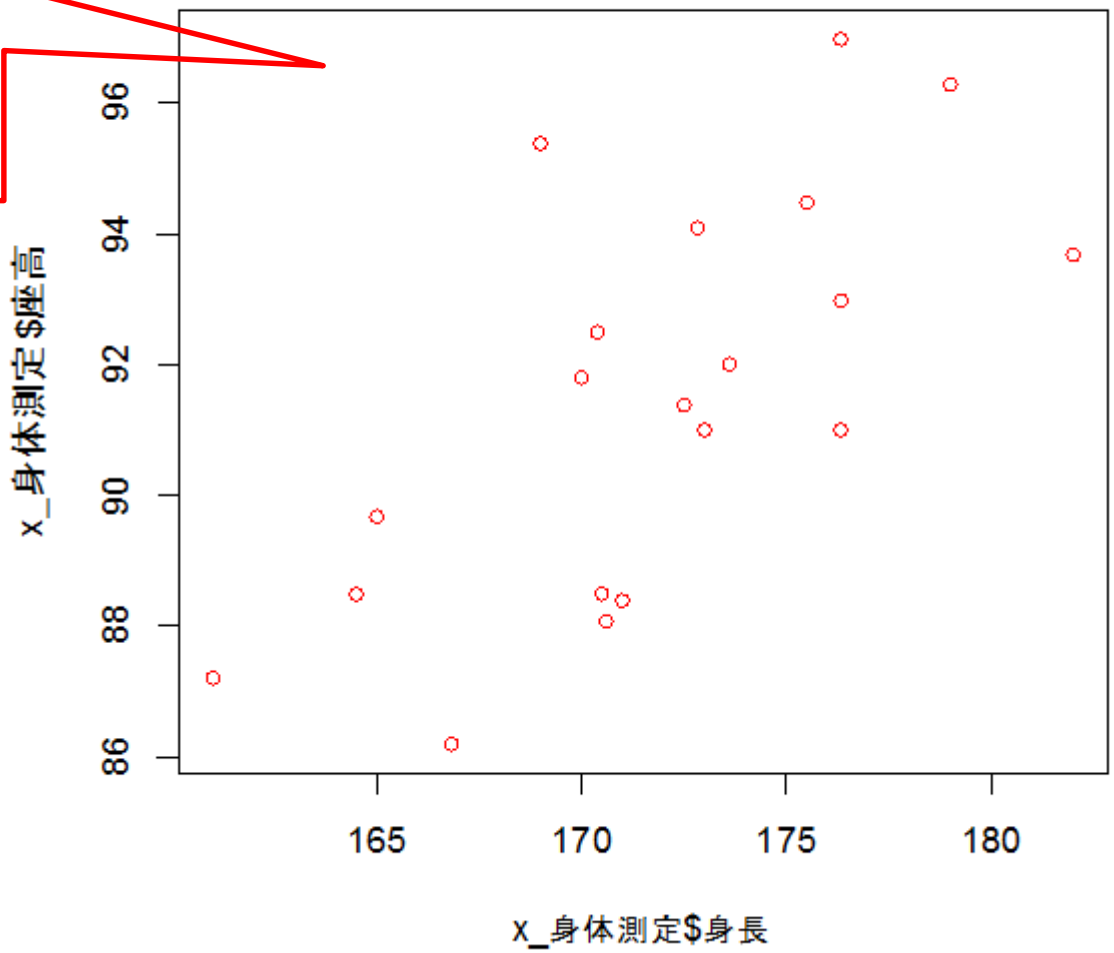
	身長	座高
1	170.6	88.1
2	164.5	88.5
3	161.0	87.2
4	170.5	88.5
5	171.0	88.4
.	.	.

図2.4 スクリプトの2行目の実行結果

スクリプトの3行目の実行

```
plot(x_身体測定$身長,x_身体測定$座高,col="red“, pch=1)
```

20人の人の身長と座高がプロットされます.



colはマーカの色
pchはマーカの種類を指定するパラメータ

0: □
1: ○
2: △
3: ＋
4: ×
5: ◇

図2.5 スクリプトの3行目の実行結果

「Rエディタ」内の(2.8)式の計算をするスクリプト

```
x_身体測定 <- read.csv("C:/Users/Furuhashi/Documents/回帰分析/身体測定_身長_座高.csv")
x_身体測定

plot(x_身体測定$身長,x_身体測定$座高,col="red", pch=1)

XX <- as.matrix(x_身体測定)          #データフレームを行列へ変換

y <- XX[,2]                          # 座高データを抽出

x <- XX[,1]                          # 身長データを抽出

n <- length(x)                       # ベクトルxの要素数を得る

mean_y <- mean(y)                    # 座高データの平均値を計算

mean_x <- mean(x)                    # 身長データの平均値を計算

sum_xy <- t(x) %*% y                 # Σxi*yi の計算

sum_xx <- t(x) %*% x                 # Σxi^2 の計算

a1 <- (sum_xy - n * mean_x * mean_y)/(sum_xx - n * mean_x^2)   #a1の計算
a0 <- mean_y - mean_x * a1                                     #a0の計算

abline(a0, a1)                       # 回帰式の描画
```

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

$\bar{y}$

$\bar{x}$

$\sum_{i=1}^{n} x_i y_i$ の計算

$\sum_{i=1}^{n} x_i^2$ の計算

$$a_0 = \bar{y} - \bar{x} a_1$$

$$a_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

最後の行

```
abline(a0, a1)
```

により，回帰式が描画されます．

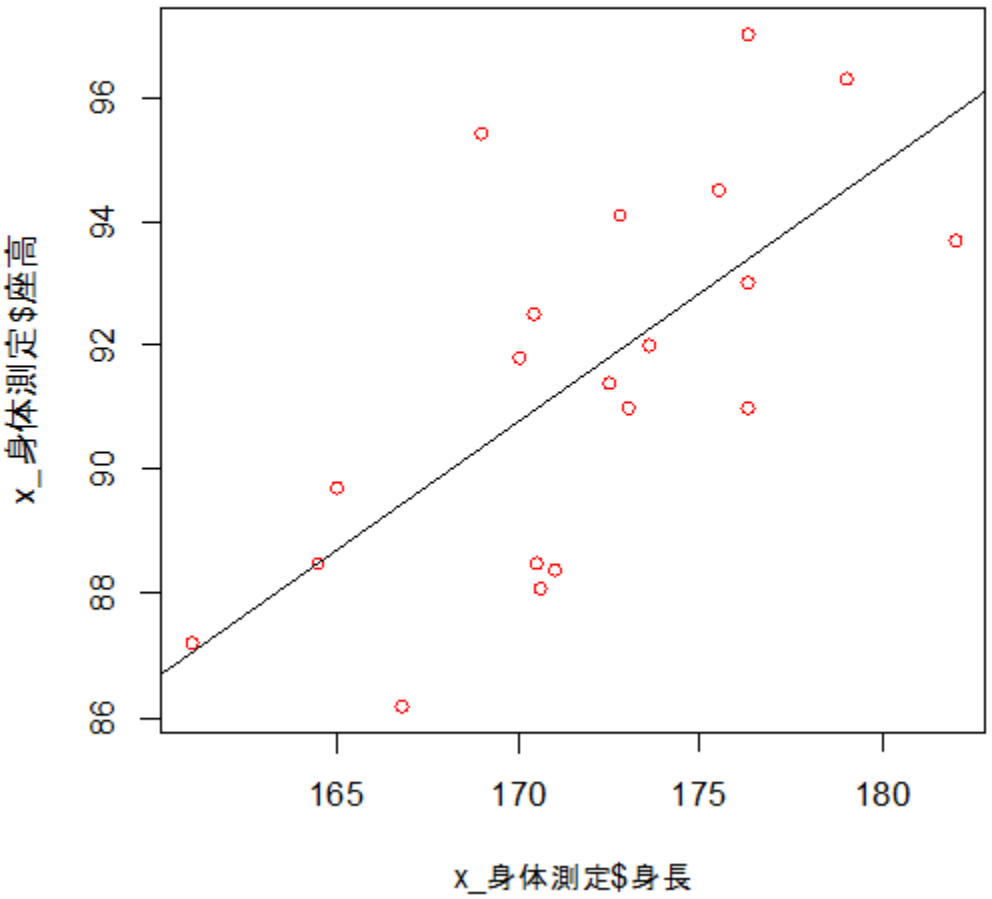


図2.6 全スクリプトの実行結果

2.3 行列による表現

単純回帰モデルを単純回帰モデルを用いるとテータ点$P_i(x_i, y_i)$は

$$y_i = a_0 + a_1 x_i + \varepsilon_i$$

と表すことができます

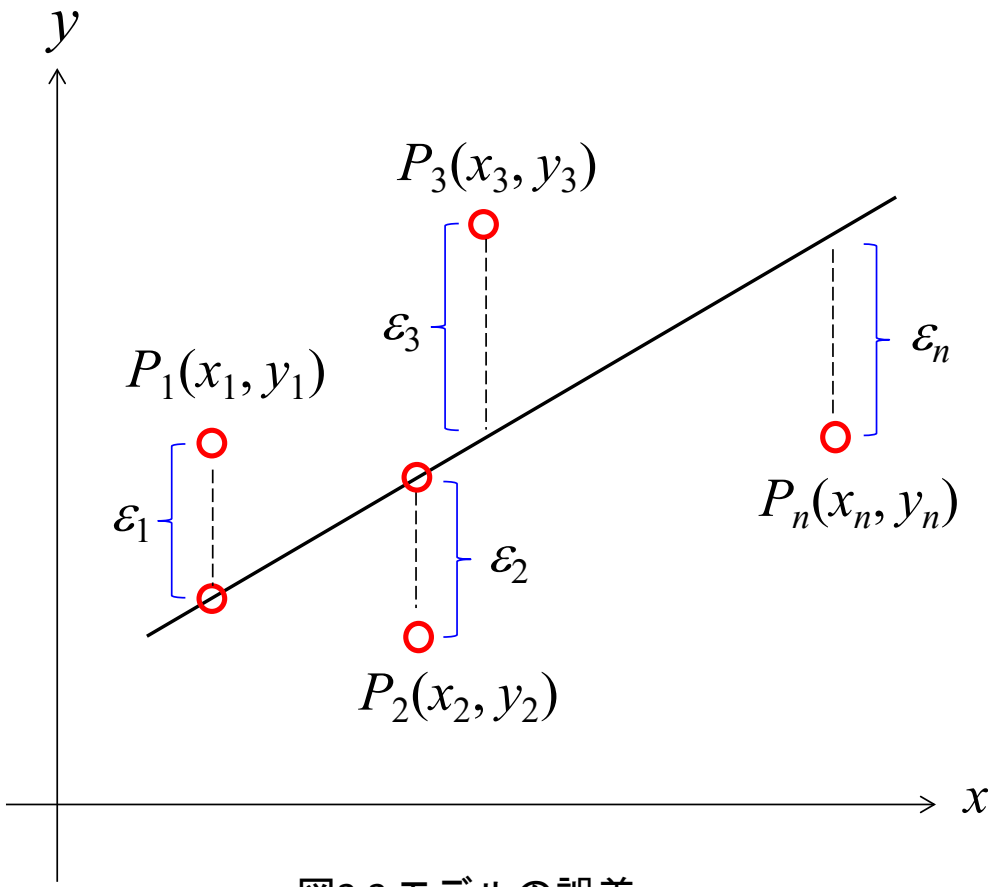


図2.2 モデルの誤差

並べて表記すると

$$\begin{aligned} y_1 &= a_0 + a_1 x_1 + \varepsilon_1 \\ y_2 &= a_0 + a_1 x_2 + \varepsilon_2 \\ &\cdot \\ y_n &= a_0 + a_1 x_n + \varepsilon_n \end{aligned} \tag{2.11}$$

となります．ここで

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \tag{2.12}$$

$$A = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

とおくと，(2.11)式は

$$Y = XA + \varepsilon \qquad (2.13)$$

と簡潔に表すことができます．

(2.4)式の誤差Eを最小とするAは

$$A = (X^t X)^{-1} X^t Y \qquad (2.20)$$

と求められます．

2.4 Rによる行列計算

2.2項と同じ身長と座高のデータを用いて，ベクトル，行列を用いた回帰分析の計算をRにより実行します．

1．Rを立ち上げ，「回帰分析」フォルダにある「**回帰分析_身体測定_行列・ベクトル.R**」ファイルを開いてください．
2．スクリプトを1行ずつ逐次実行させるには，Ctrl+Rを押し続けることでできます．
3．全スクリプトを一括実行させるには，Ctrl+Aを押して全スクリプトを選択した後に，Ctrl+Rを押すことでできます．

回帰分析_身体測定_行列・ベクトル.R

Rエディタ内の表示

```
x_身体測定 <- read.csv("C:/Users/Furuhashi/Documents/回帰分析/身体測定_身長_座高.csv")

plot(x_身体測定$身長,x_身体測定$座高,col="red", pch=1)

XX <- as.matrix(x_身体測定)          #データフレームを行列へ変換

Y <- XX[,2]                          # 座高データを抽出

X <- cbind(c(1:1), XX[,1])           # 要素が1のベクトルと身長データを結合

XtX <- t(X) %*% X                    # X^t Xを計算

inv_XtX = solve(XtX)                 # X^t Xの逆行列を計算

A <- inv_XtX %*% t(X) %*% Y          # 係数ベクトルを計算

abline(A[1], A[2])                   # 回帰式の描画
```

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

$$X^t X$$

$$\left(X^t X\right)^{-1}$$

$$A = \left(X^t X\right)^{-1} XY$$

2.2項のスクリプトと比べると，本項のスクリプトは簡単になっています．

2.2項の回帰分析_身体測定_基礎式.R

```
x_身体測定 <- read.csv("C:/Users/Furuhashi/Documents/回帰分析/身体測定_身長_座高.csv")
plot(x_身体測定$身長,x_身体測定$座高,col="red", pch=1)
XX <- as.matrix(x_身体測定)          #データフレームを行列へ変換
y <- XX[,2]                          # 座高データを抽出
x <- XX[,1]                          # 身長データを抽出
n <- length(x)                       # ベクトルxの要素数を得る
mean_y <- mean(y)                    # 座高データの平均値を計算
mean_x <- mean(x)                    # 身長データの平均値を計算
sum_xy <- t(x) %*% y                 # Σxi*yi の計算
sum_xx <- t(x) %*% x                 # Σxi^2 の計算
a1 <- (sum_xy - n * mean_x * mean_y)/(sum_xx - n * mean_x^2)          #a1の計算
a0 <- mean_y - mean_x * a1                                            #a0の計算
abline(a0, a1)                       # 回帰式の描画
```

2.4項の回帰分析_身体測定_行列・ベクトル.R

```
x_身体測定 <- read.csv("C:/Users/Furuhashi/Documents/回帰分析/身体測定_身長_座高.csv")
plot(x_身体測定$身長,x_身体測定$座高,col="red", pch=1)
XX <- as.matrix(x_身体測定)          #データフレームを行列へ変換
Y <- XX[,2]                          # 座高データを抽出
X <- cbind(c(1:1), XX[,1])           # 要素が1のベクトルと身長データを結合
XtX <- t(X) %*% X                    # X^t Xを計算
inv_XtX = solve(XtX)                 # X^t Xの逆行列を計算
A <- inv_XtX %*% t(X) %*% Y          # 係数ベクトルを計算
abline(A[1], A[2])                   # 回帰式の描画
```

2.5 Rの組み込み関数(lm())による計算

回帰分析_身体測定_組込関数.R

```
x_身体測定 <- read.csv("C:/Users/Furuhashi/Documents/回帰分析/身体測定_身長_座高.csv")

plot(x_身体測定$身長,x_身体測定$座高,col="red", pch=1)

x_回帰分析 <- lm(座高~身長 , data=x_身体測定)

abline(x_回帰分析, lwd = 1, col = "blue")
```

1行で回帰分析を実行できます.

図2.6 と全く同じ結果が得られます.

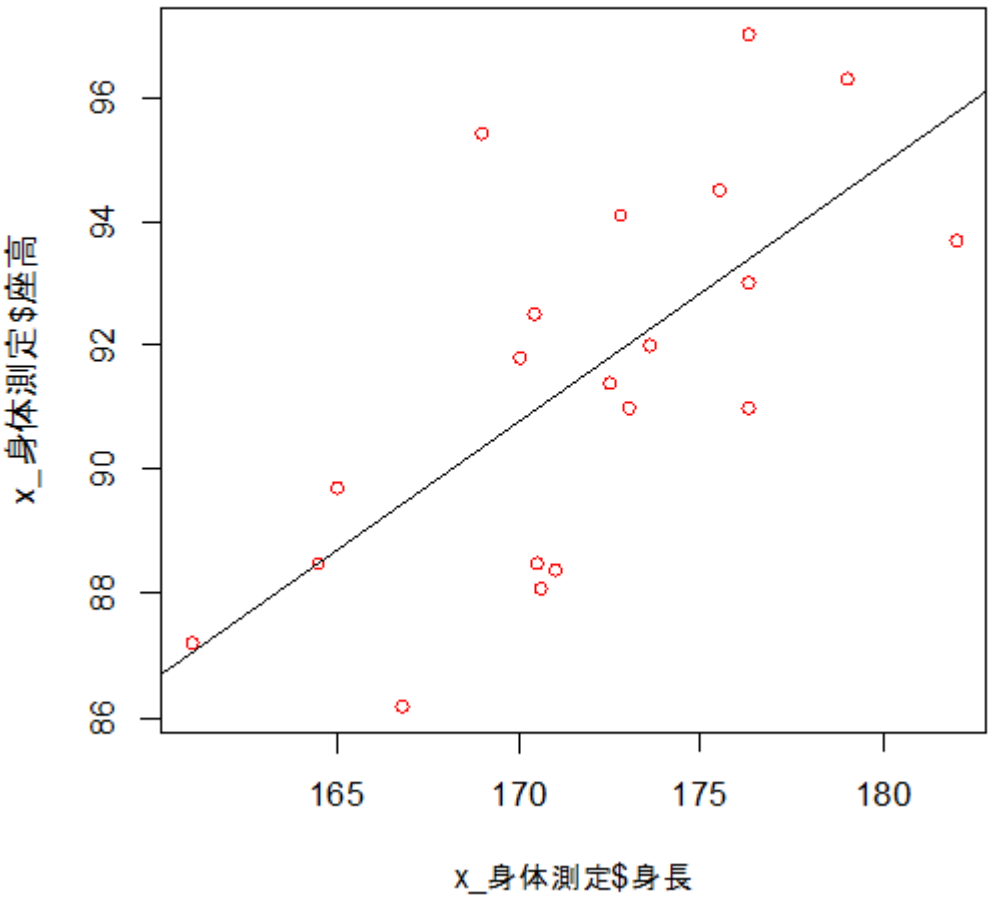


図2.6 全スクリプトの実行結果

2.4項の回帰分析_身体測定_行列・ベクトル.R

```
x_身体測定 <- read.csv("C:/Users/Furuhashi/Documents/回帰分析/身体測定_身長_座高.csv")
plot(x_身体測定$身長,x_身体測定$座高,col="red", pch=1)
XX <- as.matrix(x_身体測定)          #データフレームを行列へ変換
Y <- XX[,2]                          # 座高データを抽出
X <- cbind(c(1:1), XX[,1])           # 要素が1のベクトルと身長データを結合
XtX <- t(X) %*% X                    # X^t Xを計算
inv_XtX = solve(XtX)                 # X^t Xの逆行列を計算
A <- inv_XtX %*% t(X) %*% Y          # 係数ベクトルを計算
abline(A[1], A[2])                   # 回帰式の描画
```

2.5項の回帰分析_身体測定_組込関数.R

```
x_身体測定 <- read.csv("C:/Users/Furuhashi/Documents/回帰分析/身体測定_身長_座高.csv")

plot(x_身体測定$身長,x_身体測定$座高,col="red", pch=1)

x_回帰分析 <- lm(座高~身長 , data=x_身体測定)

abline(x_回帰分析, lwd = 1, col = "blue")
```

おわりに

回帰分析について解説しました．Rのノウハウ書としないために基礎理論を述べ，その理論展開に沿ったRの計算例を紹介しました．lm()関数を利用する方が実践的ではありますが，**理論を理解してこそ**，これらの関数を使いこなせることと思います．

なお，本スライドの内容の詳細は

「多変量解析の基礎I（回帰分析）[kindle版]」

にまとめて，Amazonより出版しています．