データ解析

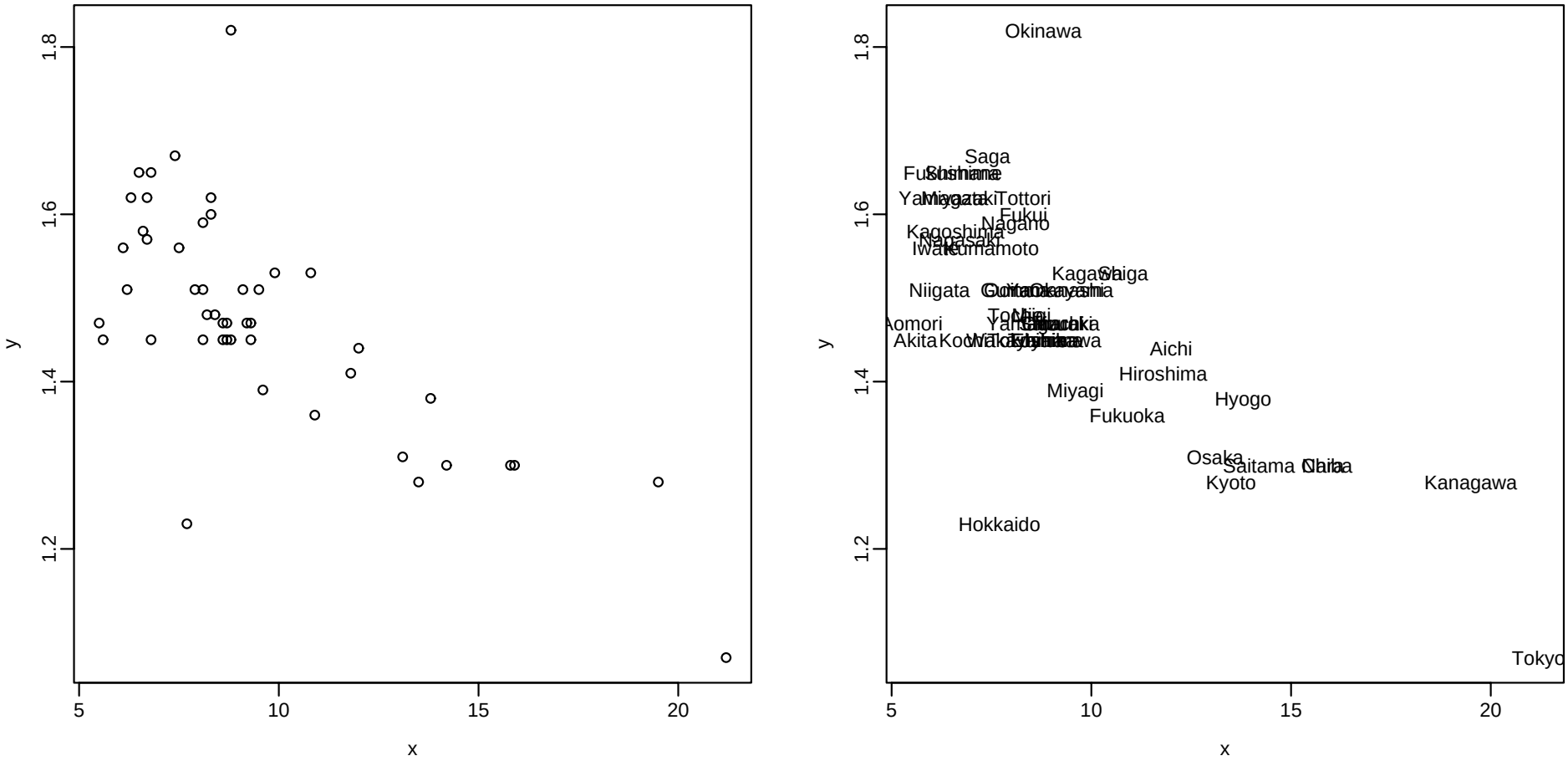
Rによる多変量解析入門

(4) 回帰分析(I)

lec20021031 下平英寿 shimo@is.titech.ac.jp

直線のあてはめ

散布図 (scatter plot)



データ

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

```
> ### 散布図
> ax <- "E09504"
> X2000$jitem[ax]
                                    E09504
"最終学歴が大学・大学院卒の者の割合 "
> x <- X2000$x[,ax]
> ay <- "A05203"
> X2000$jitem[ay]
                   A05203
"合 計 特 殊 出 生 率  "
> y <- X2000$x[,ay]
> rbind(x,y)
  Hokkaido Aomori Iwate Miyagi Akita Yamagata Fukushima Ibaraki Tochigi Gumma
x     7.70   5.50  6.10   9.60  5.60     6.30      6.50    9.30    8.20  8.10
y     1.23   1.47  1.56   1.39  1.45     1.62      1.65    1.47    1.48  1.51
  Saitama Chiba Tokyo Kanagawa Niigata Toyama Ishikawa Fukui Yamanashi Nagano
x    14.2  15.9 21.20    19.50    6.20   8.80     9.30   8.3      9.10   8.10
y     1.3   1.3  1.07     1.28    1.51   1.45     1.45   1.6      1.51   1.59
  Gifu Shizuoka Aichi  Mie Shiga Kyoto Osaka Hyogo Nara Wakayama Tottori
```

```
x 8.70      9.20 12.00 8.40 10.80 13.50 13.10 13.80 15.8      8.10    8.30
y 1.47      1.47  1.44 1.48  1.53  1.28  1.31  1.38  1.3      1.45    1.62
  Shimane Okayama Hiroshima Yamaguchi Tokushima Kagawa Ehime Kochi Fukuoka Sa
x    6.80    9.50     11.80       8.60       8.60   9.90  8.70  6.80   10.90 7.
y    1.65    1.51      1.41       1.47       1.45   1.53  1.45  1.45    1.36 1.
  Nagasaki Kumamoto Ooita Miyazaki Kagoshima Okinawa
x     6.70     7.50  7.90     6.70      6.60    8.80
y     1.57     1.56  1.51     1.62      1.58    1.82
> ## 散布図を点で描く
> plot(x,y)
> ## 散布図を県名で描く
> myplot(x,y)
```

平均，分散，標準偏差，共分散，相関

```
> mymean1
function(x) sum(x)/length(x)
> mymean1(x)
[1] 9.540426
> mymean1(y)
[1] 1.472979
> myvar1
function(x,y=x) sum((x-mymean1(x))*(y-mymean1(y)))/(length(x)-1)
> sqrt(myvar1(x))
[1] 3.438950
> sqrt(myvar1(y))
[1] 0.1331380
> myvar1(x,y)/sqrt(myvar1(x)*myvar1(y))
[1] -0.7296628
> mycor1
function(x,y) myvar1(x,y)/sqrt(myvar1(x)*myvar1(y))
> mycor1(x,y)
[1] -0.7296628
```

単回帰モデル (simple regression model)

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i; \quad i = 1, 2, \ldots, n$$

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_2 + \epsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_n + \epsilon_n \end{aligned}$$

y_i	目的変数，従属変数，応答変数
x_i	説明変数，独立変数，予測変数
ϵ_i	誤差
β_k	回帰係数，偏回帰係数

最小二乗法 (least squares method)

誤差　$\epsilon_i = y_i - (\beta_0 + \beta_1 x_i)$

誤差の二乗和　$S = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \{y_i - (\beta_0 + \beta_1 x_i)\}^2$

$= A_{00}\beta_0^2 + 2A_{01}\beta_0\beta_1 + A_{11}\beta_1^2$

Sが最小になるようにβ_0とβ_1を調節する

$$\bar{x} = \frac{1}{n}\sum x_i, \quad \bar{y} = \frac{1}{n}\sum y_i,$$

$$S_{xx} = \sum (x_i - \bar{x})^2, \quad S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$$

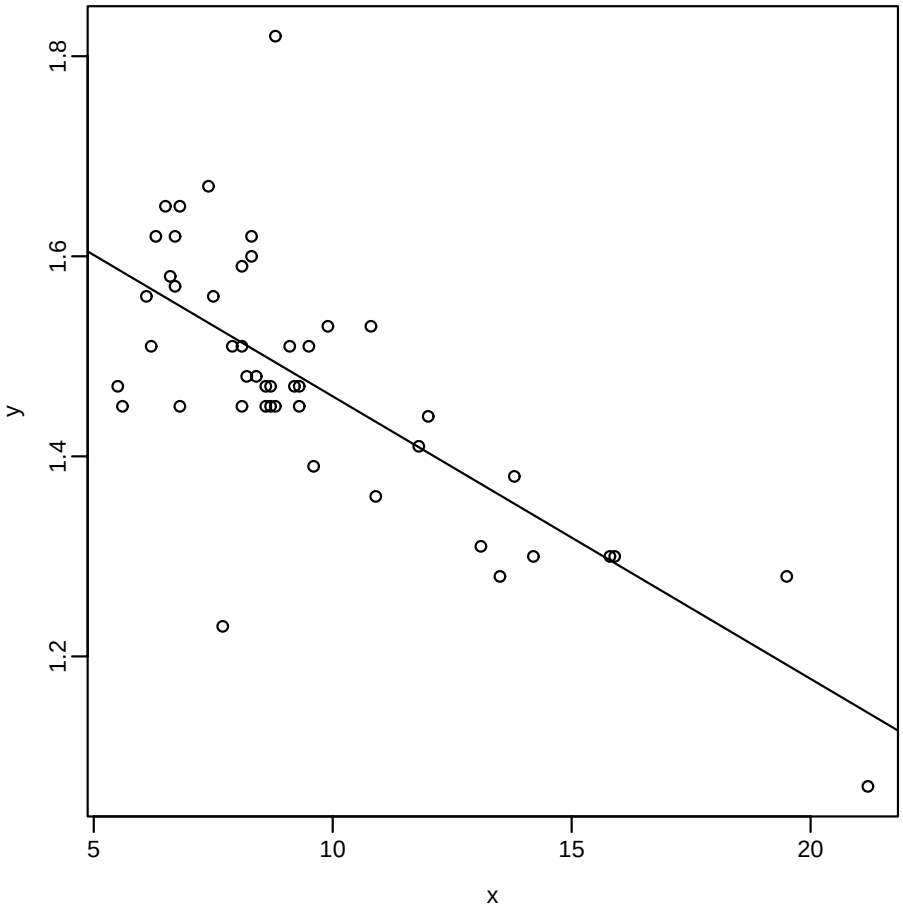
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

$$\boxed{\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i}$$

数値例１

```
> b1 <- myvar1(x,y)/myvar1(x,x)
> b0 <- mymean1(y) - b1 * mymean1(x)
> coef <- c(b0,b1)
> coef
[1]  1.74248324 -0.02824869
> ## 散布図に回帰直線を描く
> plot(x,y)
> abline(b0,b1)
> ## 予測
> pred <- b0 + b1 * x # この計算では，スカラ（長さ１のベクトル）と長さ４７のベクトルを足し算している．長さの違うベクトルを足し算すると，短いほうが繰り返し用いられる．
> rbind(pred,y)
     Hokkaido   Aomori    Iwate   Miyagi    Akita Yamagata Fukushima  Ibaraki
pred 1.524968 1.587115 1.570166 1.471296 1.584291 1.564516  1.558867 1.479770
y    1.230000 1.470000 1.560000 1.390000 1.450000 1.620000  1.650000 1.470000
      Tochigi    Gumma  Saitama    Chiba    Tokyo Kanagawa  Niigata   Toyama
pred 1.510844 1.513669 1.341352 1.293329 1.143611 1.191634 1.567341 1.493895
```

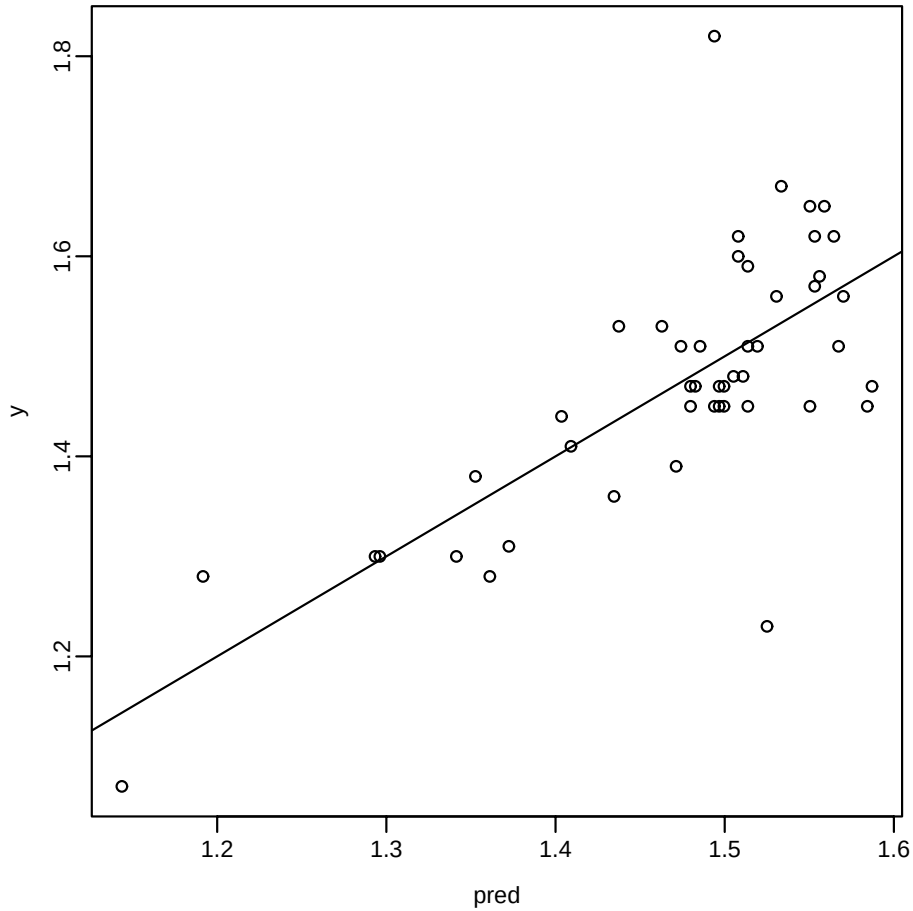
```
y    1.480000 1.510000 1.300000 1.300000 1.070000 1.280000 1.510000 1.450000
     Ishikawa    Fukui Yamanashi   Nagano     Gifu Shizuoka    Aichi      Mie
pred 1.479770 1.508019  1.485420 1.513669 1.496720 1.482595 1.403499 1.505194
y    1.450000 1.600000  1.510000 1.590000 1.470000 1.470000 1.440000 1.480000
        Shiga    Kyoto    Osaka    Hyogo     Nara Wakayama  Tottori  Shimane
pred 1.437397 1.361126 1.372425 1.352651 1.296154 1.513669 1.508019 1.550392
y    1.530000 1.280000 1.310000 1.380000 1.300000 1.450000 1.620000 1.650000
      Okayama Hiroshima Yamaguchi Tokushima   Kagawa    Ehime    Kochi  Fukuo
pred 1.474121  1.409149  1.499545  1.499545 1.462821 1.496720 1.550392 1.4345
y    1.510000  1.410000  1.470000  1.450000 1.530000 1.450000 1.450000 1.3600
         Saga Nagasaki Kumamoto    Ooita Miyazaki Kagoshima  Okinawa
pred 1.533443 1.553217 1.530618 1.519319 1.553217  1.556042 1.493895
y    1.670000 1.570000 1.560000 1.510000 1.620000  1.580000 1.820000

> ## 予測と実測値の散布図
> plot(pred,y)
> abline(0,1)
```



(x_i, y_i)**の散布図**

$$y = \beta_0 + \beta_1 x$$

$$\widehat{\beta}_0 = 1.74,\ \widehat{\beta}_1 = -0.028$$



$(\widehat{y}_i, y_i)$**の散布図**

$$\widehat{y} = y$$

最小二乗推定量の導出

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0$$

$$\sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i^2$$

$$\begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ n & \sum x_i^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ n & \sum x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

ベクトル表現

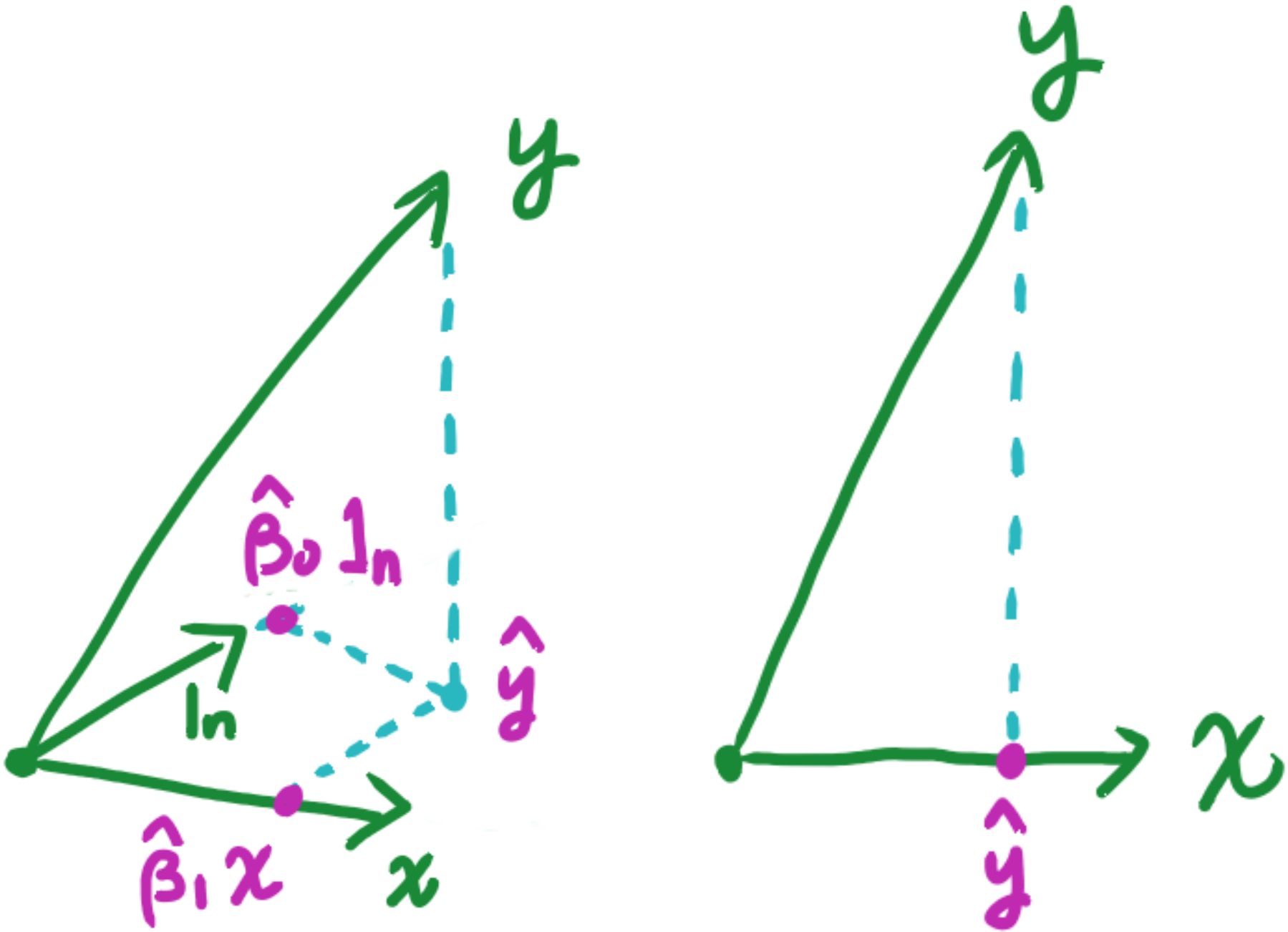
$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\begin{aligned} \boldsymbol{y} &= \beta_0 \mathbf{1}_n + \beta_1 \boldsymbol{x} + \boldsymbol{\epsilon} \\ &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \end{aligned}$$

$$\boldsymbol{X} = \left[\mathbf{1}_n, \boldsymbol{x}\right], \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$S = \|\boldsymbol{\epsilon}\|^2 = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$$

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$


y
$\hat{\beta}_0 1_n$
1_n
$\hat{y}$
$\hat{\beta}_1 x$
x
y
x
$\hat{y}$

中心化（centering）

$$x_i \leftarrow x_i - \bar{x}, \qquad y_i \leftarrow y_i - \bar{y}$$

$$y_i = \beta_1 x_i + \epsilon_i$$

$$\boldsymbol{y} = \beta_1 \boldsymbol{x} + \boldsymbol{\epsilon}$$

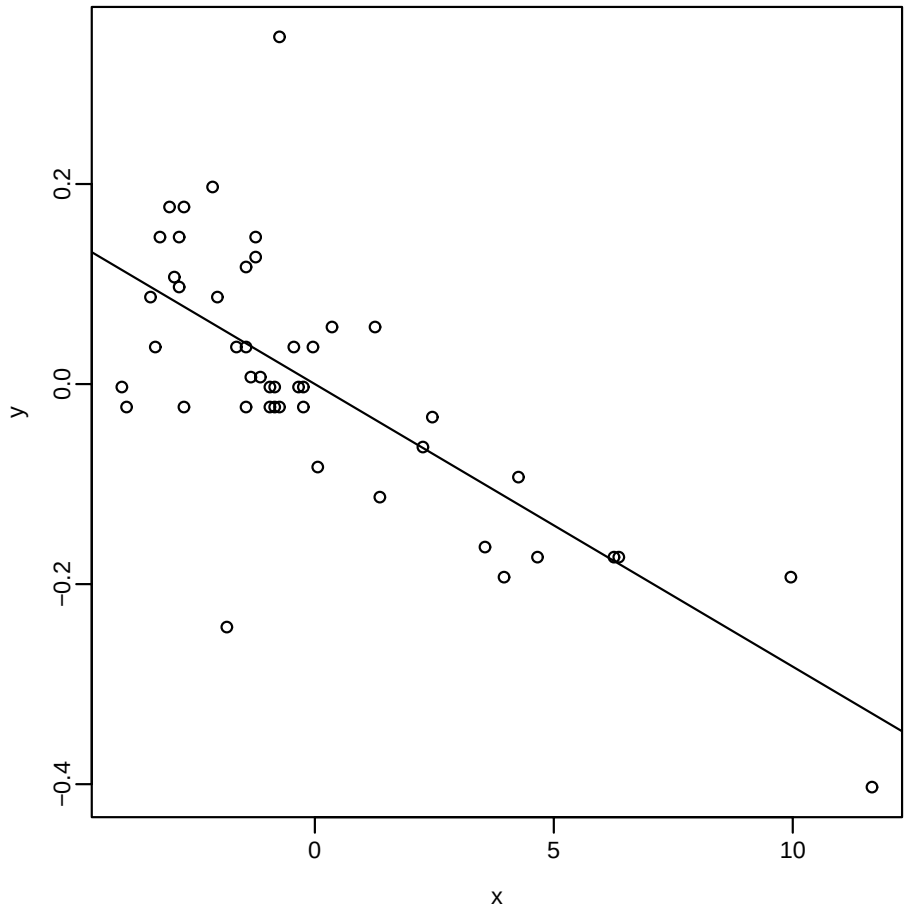
$$S_{xx} = \sum x_i^2 = \boldsymbol{x}'\boldsymbol{x} = \|\boldsymbol{x}\|^2, \qquad S_{xy} = \sum x_i y_i = \boldsymbol{x}'\boldsymbol{y}$$

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\boldsymbol{x}'\boldsymbol{y}}{\|\boldsymbol{x}\|^2}$$

yのxへの射影 $\quad \hat{y} = \left(\dfrac{\boldsymbol{x}'\boldsymbol{y}}{\|\boldsymbol{x}\|^2}\right)\boldsymbol{x}$

```
> n <- length(y)
> n
[1] 47
> i <- rep(1,n)
> i
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[39] 1 1 1 1 1 1 1 1 1
> i %*% x / n
         [,1]
[1,] 9.540426
> i %*% y / n
         [,1]
[1,] 1.472979
> x <- x - i * (i %*% x / n)
> y <- y - i * (i %*% y / n)
> mymean1(x)
[1] -1.530690e-15
> mymean1(y)
```

```
[1] -1.842498e-16
> (x %*% y)/(x %*% x)  # = b1
             [,1]
[1,] -0.02824869
> plot(x,y)
> abline(0,(x %*% y)/(x %*% x) )
```



回帰分析

重回帰モデル (multiple regression model)

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i; \quad i = 1, \ldots, n$$

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{x}_k = \begin{bmatrix} x_{1k} \\ \vdots \\ x_{nk} \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

$$\boldsymbol{y} = \beta_0 \mathbf{1}_n + \beta_1 \boldsymbol{x}_1 + \cdots + \beta_p \boldsymbol{x}_p + \boldsymbol{\epsilon}$$

$$\boldsymbol{X} = [\mathbf{1}_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p] = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}$$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

射影

$$\|\boldsymbol{\epsilon}\|^2 = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$$

を最小にするには

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

ハット行列（射影行列）

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

$1_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$の張る空間への$\boldsymbol{y}$の射影

$$\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$$

もし$\boldsymbol{X}'\boldsymbol{X}$が退化している場合には

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{X}^{+}\boldsymbol{y}$$

導出

$$\begin{aligned}\|\boldsymbol{\epsilon}\|^2 &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{\beta}\end{aligned}$$

これをβで微分して

$$\frac{\partial\|\boldsymbol{\epsilon}\|^2}{\partial\boldsymbol{\beta}} = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = 0$$

すなわち正規方程式(normal equation)

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{y}$$

これを解いて

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

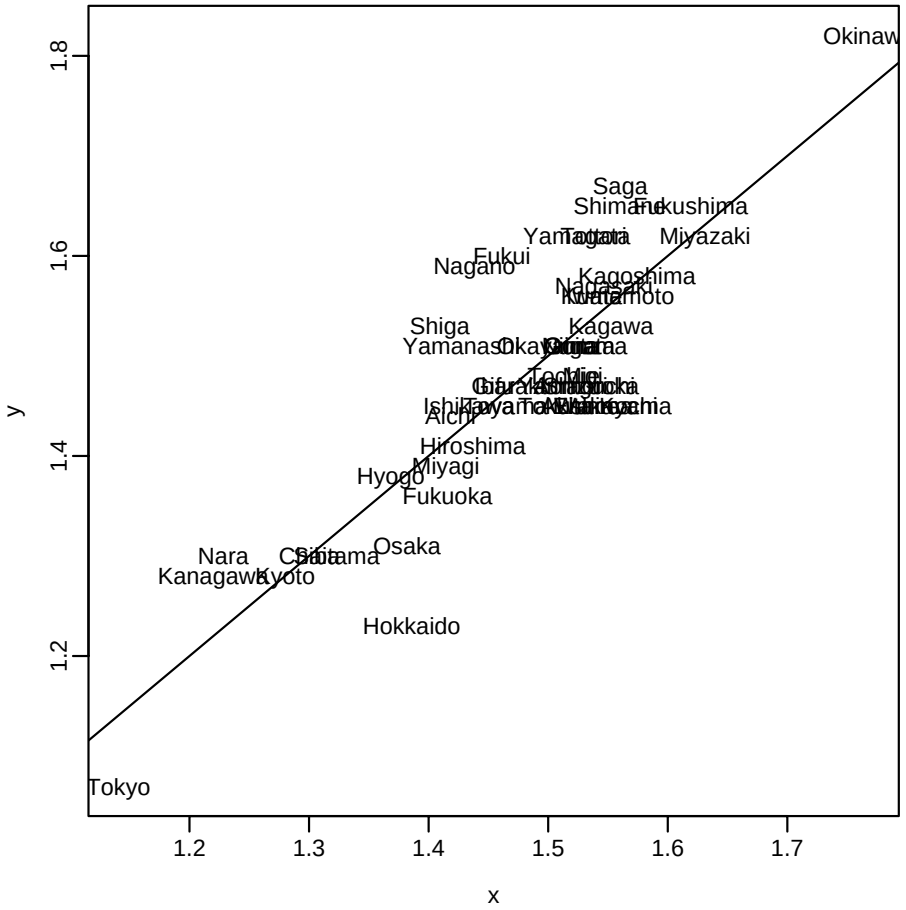
数値例２

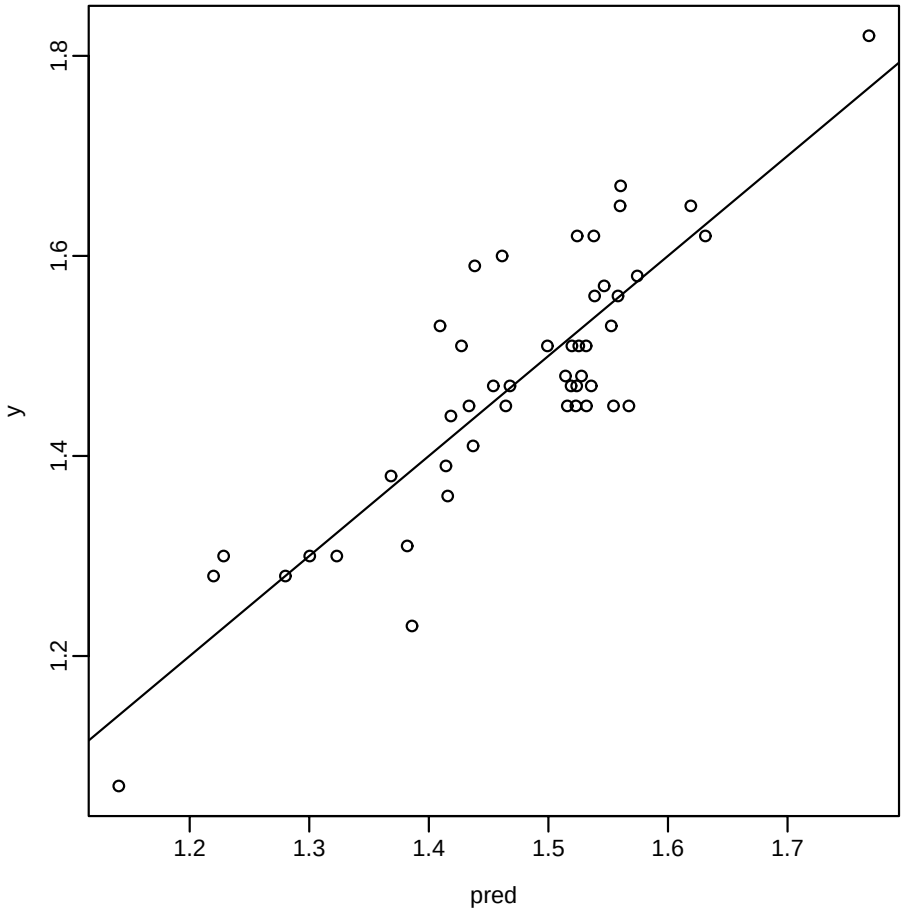
```
> ## 多変量で線形
> ax <- c("E09504","A0410302","C01301","B02101"); x <- X2000$x[,ax]
> ay <- "A05203"; y <- X2000$x[,ay]
> X2000$jitem[c(ax,ay)]
                                    E09504                                    A0410302
"最終学歴が大学・大学院卒の者の割合 " "未　婚　者　割　合［20～24歳・女］ "
                                    C01301                                     B02101
          "県民１人当たり県民所得　 "                  "年　平　均　気　温　 "
                                    A05203
            "合 計 特 殊 出 生 率　 "
> f <- mylsfit(x,y)
> f$tsummary
, ,  = Y


                 Estimate      Std.Err    t-value     Pr(>|t|)
Intercept  3.636118e+00 6.108140e-01  5.9529052 4.642519e-07
E09504    -1.655659e-02 6.570354e-03 -2.5198932 1.562857e-02
```

```
  A0410302  -2.727500e-02 7.438516e-03 -3.6667262 6.850351e-04
  C01301     1.228960e-05 4.635665e-05  0.2651098 7.922217e-01
  B02101     2.012511e-02 5.098161e-03  3.9475245 2.951848e-04

> f$fsummary
  Mean Sum Sq R Squared  F-value Df 1 Df 2       Pr(>F)
Y  0.07061925 0.7431179 30.37478    4   42 6.678249e-12
> x1 <- cbind(rep(1,length(y)),x)
> coef <- solve(t(x1) %*% x1) %*% (t(x1) %*% y)
> coef

                    [,1]
           3.636118e+00
  E09504   -1.655659e-02
  A0410302 -2.727500e-02
  C01301    1.228960e-05
  B02101    2.012511e-02
> pred <- x1 %*% coef
> t(cbind(pred,y))
```

```
  Hokkaido   Aomori    Iwate   Miyagi    Akita Yamagata Fukushima  Ibaraki
  1.385854 1.518977 1.538630 1.414345 1.515823 1.524002  1.619096 1.467897
y 1.230000 1.470000 1.560000 1.390000 1.450000 1.620000  1.650000 1.470000
   Tochigi    Gumma  Saitama    Chiba    Tokyo Kanagawa  Niigata   Toyama
  1.514316 1.531548 1.323061 1.300365 1.140643 1.219962 1.525388 1.464334
y 1.480000 1.510000 1.300000 1.300000 1.070000 1.280000 1.510000 1.450000
  Ishikawa    Fukui Yamanashi   Nagano     Gifu Shizuoka    Aichi      Mie
  1.433561 1.461327  1.427325 1.438374 1.453923 1.535835 1.418394 1.527675
y 1.450000 1.600000  1.510000 1.590000 1.470000 1.470000 1.440000 1.480000
     Shiga    Kyoto    Osaka    Hyogo     Nara Wakayama  Tottori  Shimane
  1.409378 1.280022 1.381912 1.368422 1.228403 1.554402 1.538035 1.559977
y 1.530000 1.280000 1.310000 1.380000 1.300000 1.450000 1.620000 1.650000
   Okayama Hiroshima Yamaguchi Tokushima   Kagawa    Ehime    Kochi  Fukuoka
  1.499213  1.436992   1.52368  1.523004 1.552604 1.531791 1.567354 1.415729
y 1.510000  1.410000   1.47000  1.450000 1.530000 1.450000 1.450000 1.360000
      Saga Nagasaki Kumamoto    Ooita Miyazaki Kagoshima  Okinawa
  1.560404 1.546608 1.558161 1.519602 1.631219  1.574310 1.768123
y 1.670000 1.570000 1.560000 1.510000 1.620000  1.580000 1.820000
> myplot(pred,y)
```

```
> abline(0,1)
> plot(pred,y)
> abline(0,1)
```





数値例３

```
> ###  まず単回帰，それからダミー変数（１）
> ## 単回帰
> ax <- "C01301"; ay <- "C04602"
> X2000$jitem[c(ax,ay)]
                                    C01301
                "県民１人当たり県民所得　"
                                    C04602
"個 人 預 貯 金 残 高（人口１人当たり） "
> x <- X2000$x[,ax,drop=F]
> y <- X2000$x[,ay,drop=F]
> f <- mylsfit(x,y)
> f$tsummary
, ,  = C04602


              Estimate      Std.Err   t-value     Pr(>|t|)
Intercept 26.9369511 102.88756993 0.2618096 7.946630e-01
```

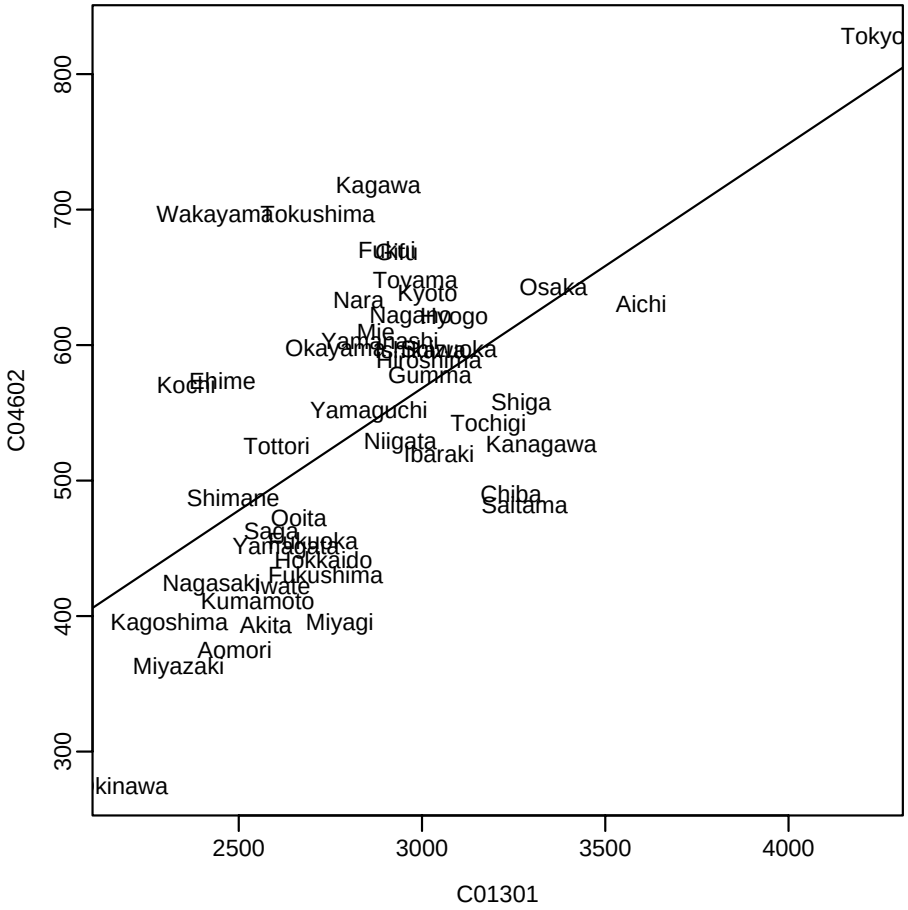
```
  C01301     0.1804065   0.03580613 5.0384246 8.093436e-06

> f$fsummary
       Mean Sum Sq R Squared  F-value Df 1 Df 2       Pr(>F)
C04602    90.60172 0.3606658 25.38572    1   45 8.093436e-06
> myplot(x,y)
> abline(f)
> ## 中心化してダミー変数を使う
> x <- x - mean(x)
> y <- y - mean(y)
> x <- cbind(x,nihonregion)
> x
                C01301 tohoku kanto shinetsu tokai kinki chugoku shikoku kyush
Hokkaido  -118.659574      1     0        0     0     0       0       0
Aomori    -360.659574      1     0        0     0     0       0       0
Iwate     -230.659574      1     0        0     0     0       0       0
Miyagi     -73.659574      1     0        0     0     0       0       0
Akita     -275.659574      1     0        0     0     0       0       0
Yamagata  -220.659574      1     0        0     0     0       0       0
```

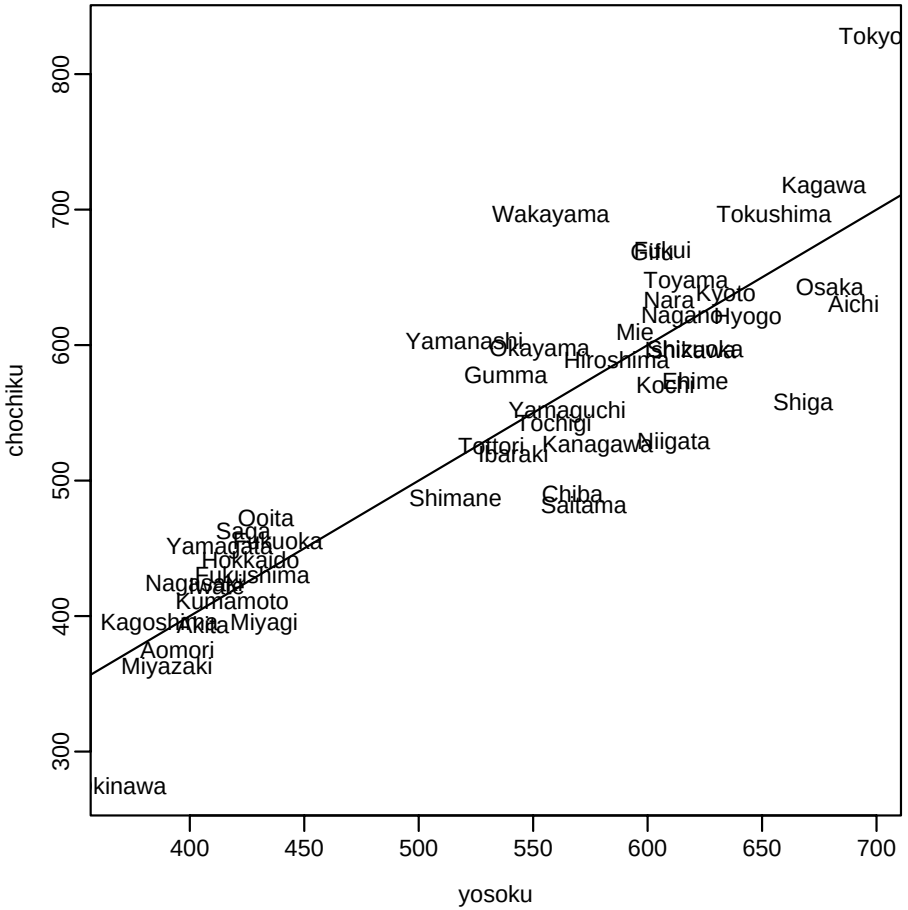
Fukushima	-112.659574	1	0	0	0	0	0
Ibaraki	197.340426	0	1	0	0	0	0
Tochigi	331.340426	0	1	0	0	0	0
Gumma	172.340426	0	1	0	0	0	0
Saitama	430.340426	0	1	0	0	0	0
Chiba	393.340426	0	1	0	0	0	0
Tokyo	1380.340426	0	1	0	0	0	0
Kanagawa	476.340426	0	1	0	0	0	0
Niigata	91.340426	0	0	1	0	0	0
Toyama	132.340426	0	0	1	0	0	0
Ishikawa	146.340426	0	0	1	0	0	0
Fukui	54.340426	0	0	1	0	0	0
Yamanashi	35.340426	0	1	0	0	0	0
Nagano	119.340426	0	0	1	0	0	0
Gifu	81.340426	0	0	0	1	0	0
Shizuoka	223.340426	0	0	0	1	0	0
Aichi	748.340426	0	0	0	1	0	0
Mie	24.340426	0	0	0	1	0	0
Shiga	421.340426	0	0	0	0	1	0

Kyoto	165.340426	0	0	0	0	1	0	0
Osaka	509.340426	0	0	0	0	1	0	0
Hyogo	238.340426	0	0	0	0	1	0	0
Nara	-22.659574	0	0	0	0	1	0	0
Wakayama	-413.659574	0	0	0	0	1	0	0
Tottori	-245.659574	0	0	0	0	0	1	0
Shimane	-364.659574	0	0	0	0	0	1	0
Okayama	-85.659574	0	0	0	0	0	1	0
Hiroshima	169.340426	0	0	0	0	0	1	0
Yamaguchi	5.340426	0	0	0	0	0	1	0
Tokushima	-133.659574	0	0	0	0	0	0	1
Kagawa	31.340426	0	0	0	0	0	0	1
Ehime	-393.659574	0	0	0	0	0	0	1
Kochi	-492.659574	0	0	0	0	0	0	1
Fukuoka	-146.659574	0	0	0	0	0	0	0
Saga	-260.659574	0	0	0	0	0	0	0
Nagasaki	-423.659574	0	0	0	0	0	0	0
Kumamoto	-297.659574	0	0	0	0	0	0	0
Ooita	-185.659574	0	0	0	0	0	0	0

```
Miyazaki  -513.659574       0     0          0     0     0        0        0
Kagoshima -538.659574       0     0          0     0     0        0        0
Okinawa   -666.659574       0     0          0     0     0        0        0
> f <- mylsfit(x,y,intercept=F)  # 定数項は含めないで回帰分析
> f$tsummary
, ,  = C04602


              Estimate     Std.Err    t-value     Pr(>|t|)
  C01301     0.1320462  0.03475661  3.7991678 5.099313e-04
  tohoku   -98.8497878 22.79470611 -4.3365239 1.027886e-04
  kanto    -25.7297094 25.16274109 -1.0225321 3.129950e-01
  shinetsu  58.3071977 25.97661830  2.2446031 3.069535e-02
  tokai     50.2255783 30.22020273  1.6619868 1.047444e-01
  kinki     71.3354400 24.03080205  2.9685002 5.157827e-03
  chugoku   23.1530379 25.95441651  0.8920654 3.779721e-01
  shikoku  131.8024396 29.99034941  4.3948284 8.608699e-05
  kyushu   -83.2049623 24.21735814 -3.4357572 1.444296e-03
```

```
> f$fsummary
       Mean Sum Sq R Squared  F-value Df 1 Df 2       Pr(>F)
C04602     57.4674 0.7827953 15.21668    9   38 4.247037e-10
> my <- mean(X2000$x[,ay])
> myplot(f$pred+my,y+my,xlab="yosoku",ylab="chochiku")
> abline(0,1)
```





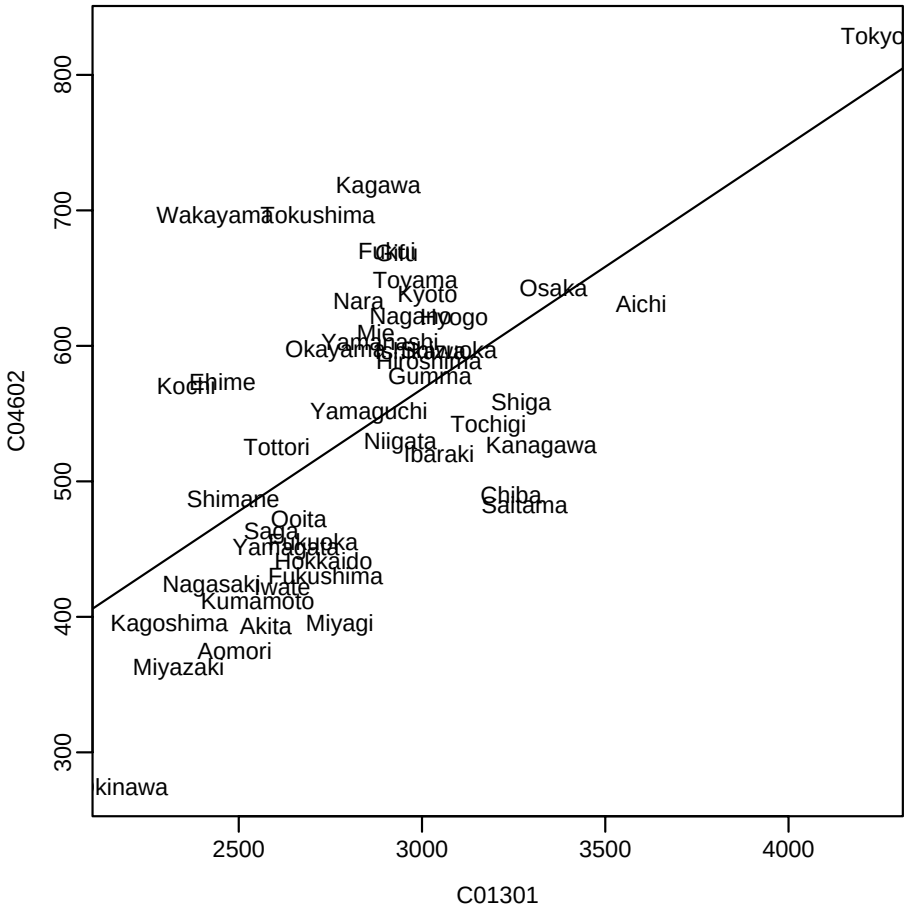
```
> ###  まず単回帰，それからダミー変数 （２）
> ## 単回帰
> ax <- "C01301"; ay <- "C04602"
> X2000$jitem[c(ax,ay)]
                                      C01301
                  "県民１人当たり県民所得　"
                                      C04602
"個 人 預 貯 金 残 高（人口１人当たり） "
> X2000$junit[c(ax,ay)]
                  C01301                        C04602
   "(千円:thousand yen)" "(万円:10 thousand yen)"
> x <- X2000$x[,ax,drop=F]
> y <- X2000$x[,ay,drop=F]
> f <- mylsfit(x,y)
> f$tsummary
, ,  = C04602
```

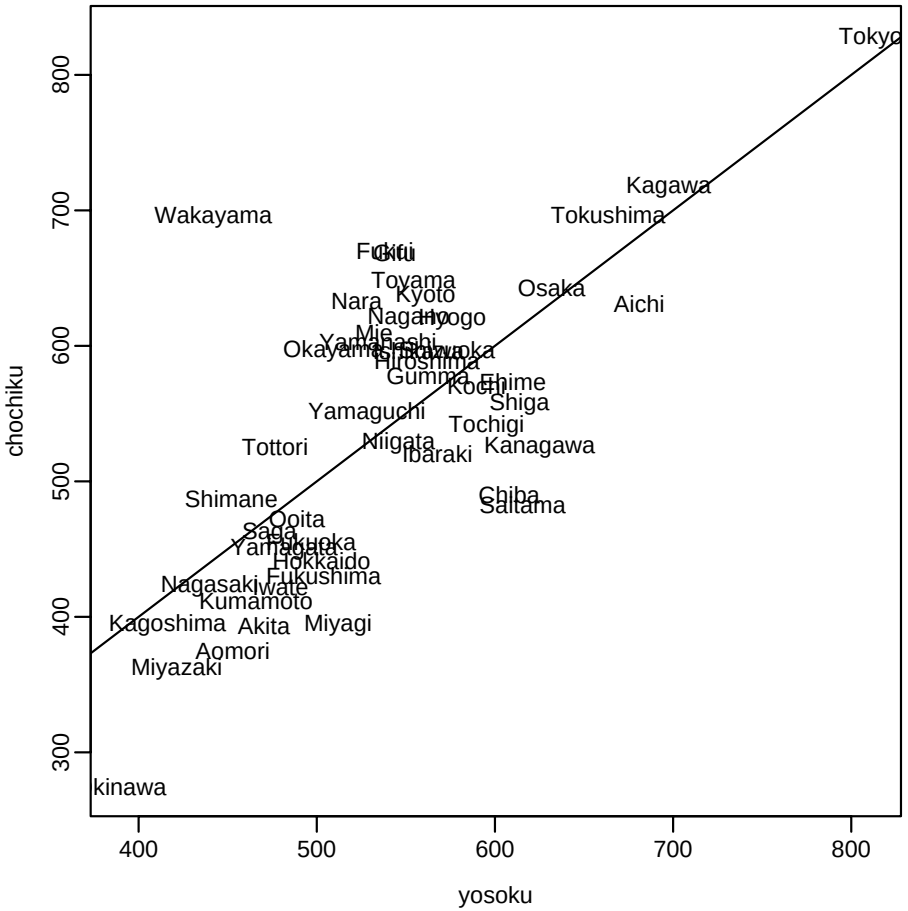
```
              Estimate      Std.Err   t-value     Pr(>|t|)
  Intercept 26.9369511 102.88756993 0.2618096 7.946630e-01
  C01301     0.1804065   0.03580613 5.0384246 8.093436e-06

> f$fsummary
       Mean Sum Sq R Squared  F-value Df 1 Df 2       Pr(>F)
C04602    90.60172 0.3606658 25.38572    1   45 8.093436e-06
> myplot(x,y)
> abline(f)
> ## ダミー変数
> nihonregion[,"shikoku"]
 Hokkaido    Aomori     Iwate    Miyagi     Akita  Yamagata Fukushima   Ibara
        0         0         0         0         0         0         0
  Tochigi     Gumma   Saitama     Chiba     Tokyo  Kanagawa   Niigata    Toya
        0         0         0         0         0         0         0
 Ishikawa     Fukui Yamanashi     Nagano      Gifu  Shizuoka     Aichi       N
        0         0         0         0         0         0         0
    Shiga     Kyoto     Osaka     Hyogo      Nara  Wakayama   Tottori   Shima
        0         0         0         0         0         0         0
```

```
  Okayama Hiroshima Yamaguchi Tokushima    Kagawa     Ehime     Kochi   Fukuo
       0         0         0         1         1         1         1
    Saga  Nagasaki  Kumamoto     Ooita  Miyazaki Kagoshima   Okinawa
       0         0         0         0         0         0         0
> x <- cbind(X2000$x[,ax,drop=F],nihonregion[,"shikoku",drop=F])
> t(x)
        Hokkaido Aomori Iwate Miyagi Akita Yamagata Fukushima Ibaraki Tochigi
C01301      2731   2489  2619   2776  2574     2629      2737    3047    3181
shikoku        0      0     0      0     0        0         0       0       0
        Gumma Saitama Chiba Tokyo Kanagawa Niigata Toyama Ishikawa Fukui
C01301   3022    3280  3243  4230     3326    2941   2982     2996  2904
shikoku     0       0     0     0        0       0      0        0     0
        Yamanashi Nagano Gifu Shizuoka Aichi  Mie Shiga Kyoto Osaka Hyogo Nar
C01301       2885   2969 2931     3073  3598 2874  3271  3015  3359  3088 282
shikoku         0      0    0        0     0    0     0     0     0     0
        Wakayama Tottori Shimane Okayama Hiroshima Yamaguchi Tokushima Kagawa
C01301      2436    2604    2485    2764      3019      2855      2716   2881
shikoku        0       0       0       0         0         0         1      1
        Ehime Kochi Fukuoka Saga Nagasaki Kumamoto Ooita Miyazaki Kagoshima
```

```
C01301   2456  2357    2703 2589     2426     2552  2664     2336      2311
shikoku     1     1       0    0        0        0     0        0         0
        Okinawa
C01301     2183
shikoku       0
> f <- mylsfit(x,y)
> f$tsummary
, ,  = C04602


               Estimate     Std.Err    t-value     Pr(>|t|)
  Intercept -59.1667194 92.49159483 -0.6396983 5.256857e-01
  C01301      0.2057250  0.03192249  6.4445142 7.486358e-08
  shikoku   163.9675326 42.22411062  3.8832679 3.422263e-04

> f$fsummary
      Mean Sum Sq R Squared  F-value Df 1 Df 2     Pr(>F)
C04602     79.0721 0.5238522 24.20414    2   44 8.1343e-08
> myplot(f$pred,y,xlab="yosoku",ylab="chochiku")
```

```
> abline(0,1)
```


Tokyo
Kagawa
Wakayama
Tokushima
Toyama
Osaka
Aichi
Nara
Kyoto
Hyogo
Mie
Okayama
Hiroshima
Gumma
Kochi
Ehime
Shiga
Yamaguchi
Tochigi
Tottori
Niigata
Ibaraki
Kanagawa
Chiba
Saitama
Shimane
Ooita
Saga
Hokkaido
Fukushima
Nagasaki
Iwate
Kumamoto
Kagoshima
Akita
Miyagi
Aomori
Miyazaki
kinawa
C04602
C01301
300
400
500
600
700
800
2500
3000
3500
4000


Tokyo
Kagawa
Wakayama
Tokushima
Toyama
Osaka
Aichi
Nara
Kyoto
Hyogo
Mie
Okayama
Hiroshima
Gumma
Kochi
Ehime
Shiga
Yamaguchi
Tochigi
Tottori
Niigata
Ibaraki
Kanagawa
Chiba
Saitama
Shimane
Ooita
Saga
Hokkaido
Fukushima
Nagasaki
Iwate
Kumamoto
Kagoshima
Akita
Miyagi
Aomori
Miyazaki
kinawa
chochiku
yosoku
300
400
500
600
700
800
400
500
600
700
800

数値例４

```
> ### 多項式回帰
> psinit("20021020k04sp1.eps",pty="s")
> ax <- "B02304"; ay <- "B02102"
> X2000$jitem[c(ax,ay)]
                                                  B02304
                        "雪　　日　　数（年　間）　"
                                                  B02102
"最　高　気　温（日最高気温の月平均の最高値）　"
> X2000$junit[c(ax,ay)]
 B02304  B02102
 "(日)" "(°C)"
> x <- X2000$x[,ax]
> y <- X2000$x[,ay]
> myplot(x,y)
> ## 線形回帰（１次式）
> f1 <- mylsfit(x,y)
> f1$coefficients
```

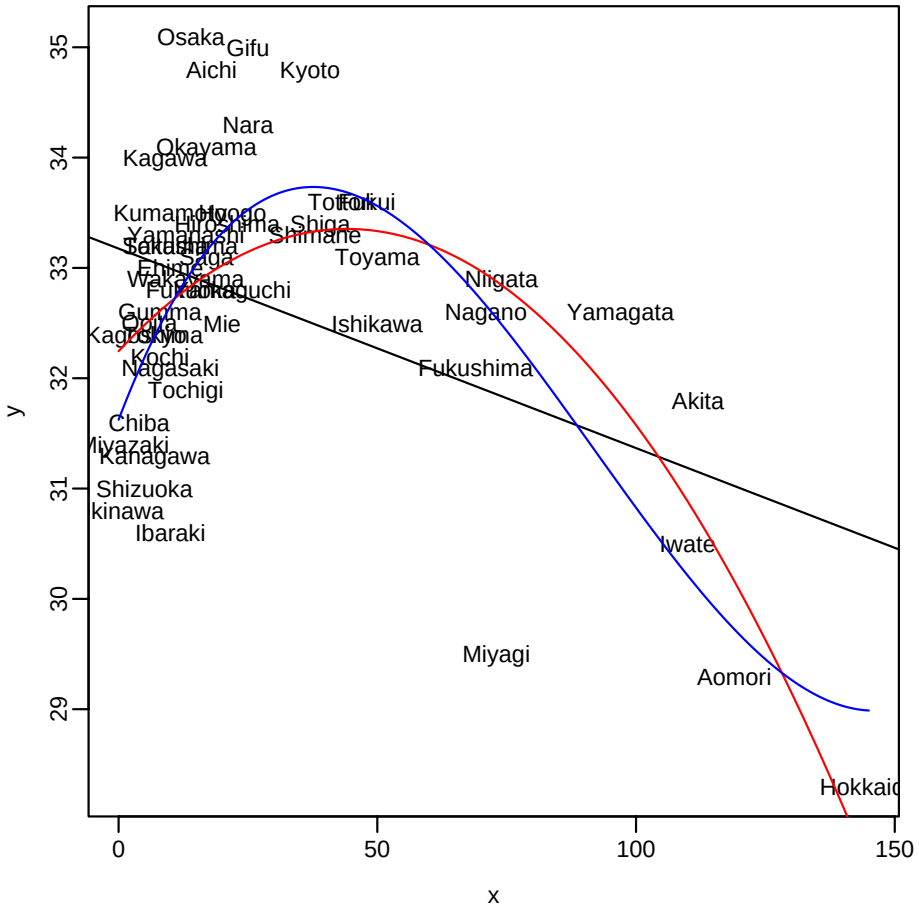
```
                        Y
  Intercept 33.17502116
  X         -0.01808129
> f1$tsummary
, ,  = Y


               Estimate     Std.Err    t-value     Pr(>|t|)
  Intercept 33.17502116 0.260778107 127.215515 3.462132e-59
  X         -0.01808129 0.005409449  -3.342537 1.678815e-03

> f1$fsummary
  Mean Sum Sq R Squared  F-value Df 1 Df 2      Pr(>F)
Y    1.304014 0.1988971 11.17256    1   45 0.001678815
> abline(f1)
> ## 多項式回帰（２次式）
> xx <- cbind(x,x^2)
> dimnames(xx)[[2]] <- c("x","x^2")
> t(xx)
```

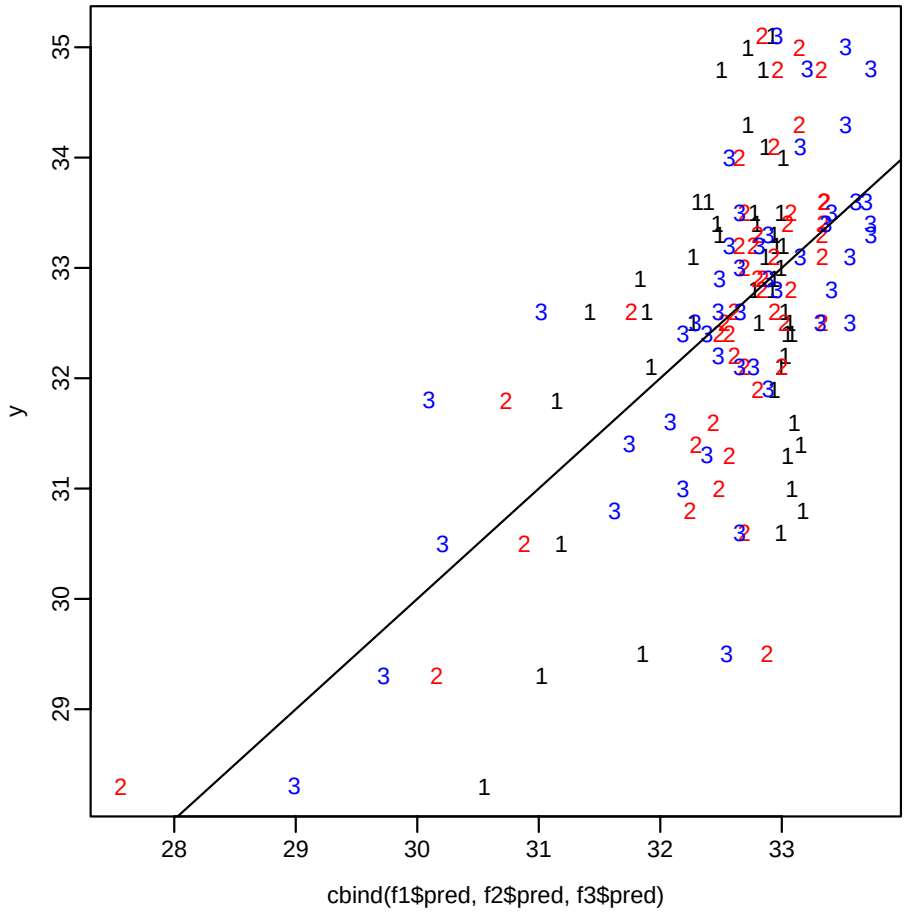
```
     Hokkaido Aomori Iwate Miyagi Akita Yamagata Fukushima Ibaraki Tochigi Gun
x         145    119   110     73   112        97         69       10       13
x^2     21025  14161 12100   5329 12544      9409       4761      100      169
     Saitama Chiba Tokyo Kanagawa Niigata Toyama Ishikawa Fukui Yamanashi Naga
x          9     4     7         7      74     50        50    48         13
x^2       81    16    49        49    5476   2500      2500  2304        169    50
     Gifu Shizuoka Aichi Mie Shiga Kyoto Osaka Hyogo Nara Wakayama Tottori
x      25        5    18  20    39    37    14    22   25        13       43
x^2   625       25   324 400  1521  1369   196   484  625       169     1849
     Shimane Okayama Hiroshima Yamaguchi Tokushima Kagawa Ehime Kochi Fukuoka
x         38      17         21        22        12      9    10     8       14
x^2     1444     289        441       484       144     81   100    64      196
     Saga Nagasaki Kumamoto Ooita Miyazaki Kagoshima Okinawa
x      17       10        10     6        1         5       0
x^2   289      100       100    36        1        25       0
> f2 <- mylsfit(xx,y)
> f2$coefficients

                          Y
```

```
  Intercept 32.2452328361
  x          0.0502327170
  x^2       -0.0005693289
> f2$tsummary
, ,  = Y


                 Estimate      Std.Err    t-value     Pr(>|t|)
  Intercept 32.2452328361 0.2895669764 111.356734 1.396604e-55
  x          0.0502327170 0.0149957168   3.349804 1.667099e-03
  x^2       -0.0005693289 0.0001193666  -4.769583 2.055703e-05

> f2$fsummary
  Mean Sum Sq R Squared  F-value Df 1 Df 2       Pr(>F)
Y    1.070696 0.4719236 19.66064    2   44 7.931007e-07
> xs <- seq(min(x),max(x),length=300)
> ys <- cbind(1,xs,xs^2) %*% f2$coefficients
> lines(xs,ys,col=2)
> ## 多項式回帰（３次式）
```

```
> xx <- cbind(x,x^2,x^3)
> dimnames(xx)[[2]] <- c("x","x^2","x^3")
> f3 <- mylsfit(xx,y)
> f3$coefficients
                          Y
  Intercept  3.162375e+01
  x          1.227429e-01
  x^2       -2.052218e-03
  x^3        7.450801e-06
> f3$tsummary
, ,  = Y

                 Estimate      Std.Err   t-value     Pr(>|t|)
  Intercept  3.162375e+01 3.621251e-01 87.328250 4.800261e-50
  x          1.227429e-01 3.122303e-02  3.931166 3.026254e-04
  x^2       -2.052218e-03 5.806757e-04 -3.534189 9.925295e-04
  x^3        7.450801e-06 2.862603e-06  2.602806 1.263671e-02
```

```
> f3$fsummary
  Mean Sum Sq R Squared  F-value Df 1 Df 2       Pr(>F)
Y    1.006674 0.5437978 17.08548    3   43 1.877689e-07
> xs <- seq(min(x),max(x),length=300)
> ys <- cbind(1,xs,xs^2,xs^3) %*% f3$coefficients
> lines(xs,ys,col=4)
> dev.off()
null device
          1
> ## 予測と観測値の散布図
> psinit("20021020k04sp2.eps",pty="s")
> matplot(cbind(f1$pred,f2$pred,f3$pred),y,col=c(1,2,4))
> abline(0,1)
> dev.off()
null device
          1
```


Osaka
Gifu
Aichi
Kyoto
Nara
Okayama
Kagawa
Shiga
Shimane
Toyama
Niigata
Nagano
Yamagata
Ishikawa
Mie
Fukushima
Akita
Chiba
Kanagawa
Shizuoka
Ibaraki
Iwate
Miyagi
Aomori
Tochigi
Nagasaki
Kochi
y
x
0
50
100
150
29
30
31
32
33
34
35


y
cbind(f1$pred, f2$pred, f3$pred)
28
29
30
31
32
33
29
30
31
32
33
34
35

あてはまりのよさ

重相関係数

y_i と $\hat{y}_i$ の相関を重相関係数 R と呼ぶ．

$$R = \frac{S_{y\hat{y}}}{\sqrt{S_{yy}S_{\hat{y}\hat{y}}}}$$

$$S_{y\hat{y}} = \sum(y_i - \bar{y})(\hat{y}_i - \bar{y}), S_{yy} = \sum(y_i - \bar{y})^2, S_{\hat{y}\hat{y}} = \sum(\hat{y}_i - \bar{y})^2$$

まず中心化 $\boldsymbol{y} \leftarrow \boldsymbol{y} - \bar{y}\mathbf{1}_n$，$\hat{\boldsymbol{y}} \leftarrow \hat{\boldsymbol{y}} - \bar{y}\mathbf{1}_n$ をおこなうと

$$R = \frac{\boldsymbol{y}'\hat{\boldsymbol{y}}}{\|\boldsymbol{y}\|\|\hat{\boldsymbol{y}}\|}$$

あてはまりの良さを R または R^2 で判断する

```
> ### 重相関係数
> ## 数値例４（１次式）
> t(round(cbind(f1$pred,y),2))
  Hokkaido Aomori Iwate Miyagi Akita Yamagata Fukushima Ibaraki Tochigi Gumma
Y    30.55  31.02 31.19  31.86 31.15    31.42     31.93   32.99   32.94 33.03
```

```
y    28.30  29.30 30.50  29.50 31.80    32.60     32.10   30.60   31.90 32.60
  Saitama Chiba Tokyo Kanagawa Niigata Toyama Ishikawa Fukui Yamanashi Nagano
Y   33.01  33.1 33.05    33.05   31.84  32.27    32.27 32.31     32.94  31.89
y   33.20  31.6 32.40    31.30   32.90  33.10    32.50 33.60     33.30  32.60
   Gifu Shizuoka Aichi   Mie Shiga Kyoto Osaka Hyogo  Nara Wakayama Tottori
Y 32.72    33.08 32.85 32.81 32.47 32.51 32.92 32.78 32.72    32.94    32.4
y 35.00    31.00 34.80 32.50 33.40 34.80 35.10 33.50 34.30    32.90    33.6
  Shimane Okayama Hiroshima Yamaguchi Tokushima Kagawa Ehime Kochi Fukuoka
Y   32.49   32.87      32.8     32.78     32.96  33.01 32.99 33.03   32.92
y   33.30   34.10      33.4     32.80     33.20  34.00 33.00 32.20   32.80
   Saga Nagasaki Kumamoto Ooita Miyazaki Kagoshima Okinawa
Y 32.87    32.99    32.99 33.07    33.16     33.08   33.18
y 33.10    32.10    33.50 32.50    31.40     32.40   30.80
> mycor1(f1$pred,y)
[1] 0.4459788
> mycor1(f1$pred,y)^2
[1] 0.1988971
> f1$fsummary
  Mean Sum Sq R Squared  F-value Df 1 Df 2      Pr(>F)
```

```
Y    1.304014 0.1988971 11.17256    1   45 0.001678815
> ## 数値例４（２次式）
> mycor1(f2$pred,y)
[1] 0.686967
> mycor1(f2$pred,y)^2
[1] 0.4719236
> f2$fsummary
  Mean Sum Sq R Squared  F-value Df 1 Df 2        Pr(>F)
Y    1.070696 0.4719236 19.66064    2   44 7.931007e-07
> ## 数値例４（３次式）
> mycor1(f3$pred,y)
[1] 0.7374264
> mycor1(f3$pred,y)^2
[1] 0.5437978
> f3$fsummary
  Mean Sum Sq R Squared  F-value Df 1 Df 2        Pr(>F)
Y    1.006674 0.5437978 17.08548    3   43 1.877689e-07
```

残差 (residual)

$$e_i = y_i - \widehat{y}_i, \quad i = 1, \ldots, n$$

$$\boldsymbol{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \hat{\boldsymbol{y}} = \begin{bmatrix} \widehat{y}_1 \\ \vdots \\ \widehat{y}_n \end{bmatrix}$$

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$$

ハット行列を用いると

$$\boldsymbol{e} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{y}$$

$\boldsymbol{HX} = \boldsymbol{X}$ **なので** $\boldsymbol{X}'\boldsymbol{e} = 0$**. とくに**

$$1_n'\boldsymbol{e} = 0, \quad \hat{\boldsymbol{y}}'\boldsymbol{e} = 0$$

ピタゴラスの定理

$$\|\boldsymbol{y}\|^2 = \|\hat{\boldsymbol{y}} + \boldsymbol{e}\|^2 = \|\hat{\boldsymbol{y}}\|^2 + 2\hat{\boldsymbol{y}}'\boldsymbol{e} + \|\boldsymbol{e}\|^2 = \|\hat{\boldsymbol{y}}\|^2 + \|\boldsymbol{e}\|^2$$

```
> ### 残差
> e <- y-f1$pred
> t(round(e,2))

  Hokkaido Aomori Iwate Miyagi Akita Yamagata Fukushima Ibaraki Tochigi Gun
Y    -2.25  -1.72 -0.69  -2.36  0.65     1.18      0.17   -2.39   -1.04 -0.

  Saitama Chiba Tokyo Kanagawa Niigata Toyama Ishikawa Fukui Yamanashi Naga
Y    0.19  -1.5 -0.65    -1.75    1.06   0.83     0.23  1.29      0.36   0.

  Gifu Shizuoka Aichi   Mie Shiga Kyoto Osaka Hyogo Nara Wakayama Tottori
Y 2.28    -2.08  1.95 -0.31  0.93  2.29  2.18  0.72 1.58    -0.04     1.2

  Shimane Okayama Hiroshima Yamaguchi Tokushima Kagawa Ehime Kochi Fukuoka
Y    0.81    1.23       0.6      0.02      0.24   0.99  0.01 -0.83   -0.12

  Saga Nagasaki Kumamoto Ooita Miyazaki Kagoshima Okinawa
```

```
  Y 0.23    -0.89     0.51 -0.57    -1.76     -0.68   -2.38
> sum(e)
[1] 9.947598e-14
> sum(f1$pred * e)
[1] 3.312767e-12
> mymean1(e)
[1] 2.116510e-15
> mycor1(f1$pred,e)
[1] 1.675907e-15
> ## ピタゴラスの定理
> sum(y^2)
[1] 49980.06
> sum(f1$pred^2)
[1] 49903.54
> sum(e^2)
[1] 76.52033
> sum(f1$pred^2) + sum(e^2) - sum(y^2)
[1] -7.275958e-12
```

重相関係数と残差の関係

まず中心化$\boldsymbol{y} \leftarrow \boldsymbol{y} - \bar{y}1_n$, $\hat{\boldsymbol{y}} \leftarrow \hat{\boldsymbol{y}} - \bar{y}1_n$と中心化しておくと，

$$R = \frac{\boldsymbol{y}'\hat{\boldsymbol{y}}}{\|\boldsymbol{y}\|\|\hat{\boldsymbol{y}}\|}$$

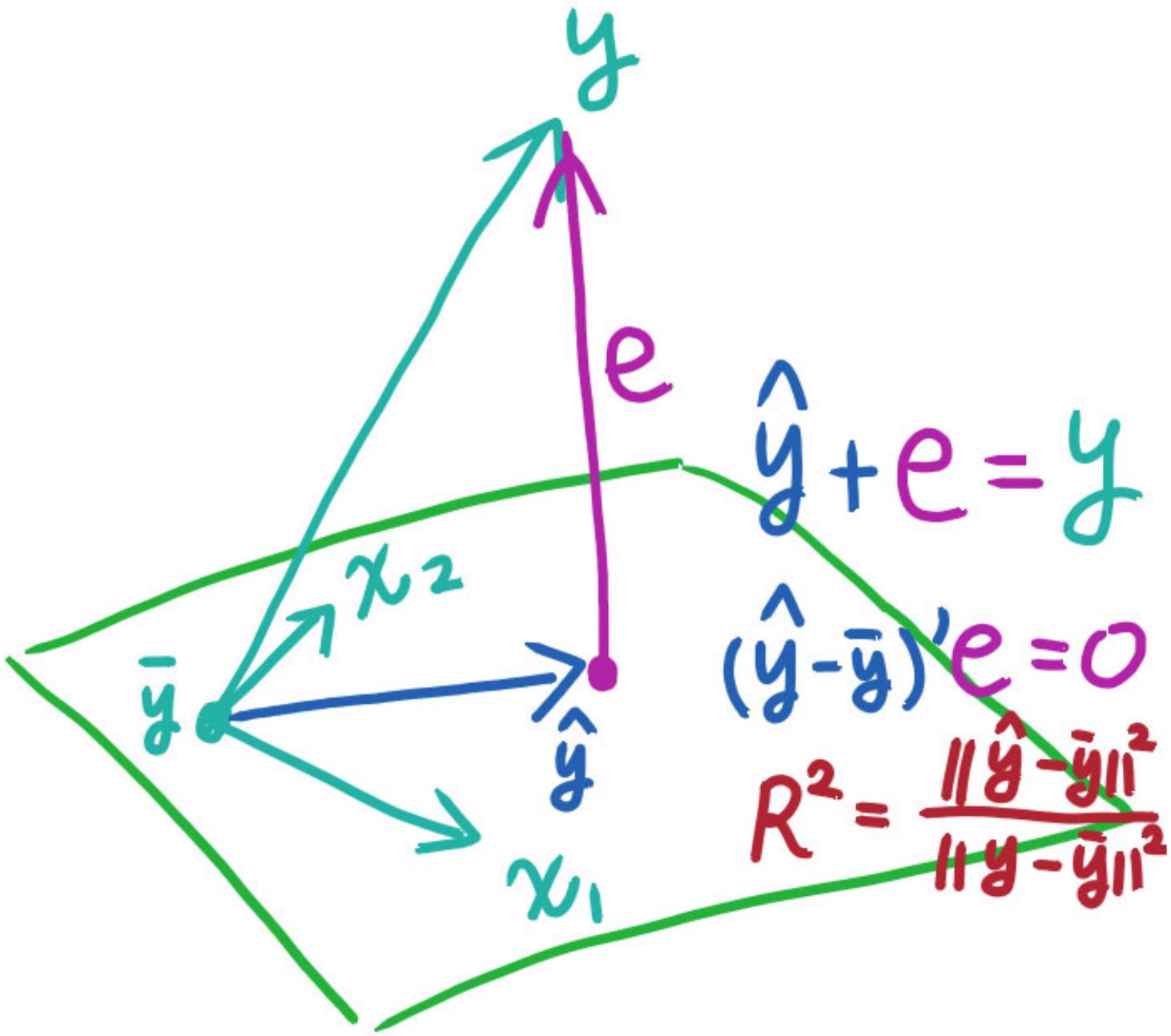
$e'\hat{\boldsymbol{y}} = 0$なので$\boldsymbol{y}'\hat{\boldsymbol{y}} = (\hat{\boldsymbol{y}} + e)'\hat{\boldsymbol{y}} = \|\hat{\boldsymbol{y}}\|^2$，すなわち

$$R^2 = \frac{\|\hat{\boldsymbol{y}}\|^2}{\|\boldsymbol{y}\|^2} = \frac{S_{\hat{y}\hat{y}}}{S_{yy}}$$

ピタゴラスの定理$\|\boldsymbol{y}\|^2 = \|\hat{\boldsymbol{y}}\|^2 + \|\boldsymbol{e}\|^2$もしくは$S_{yy} = S_{\hat{y}\hat{y}} + S_{ee}$より

$$R^2 = 1 - \frac{\|\boldsymbol{e}\|^2}{\|\boldsymbol{y}\|^2} = 1 - \frac{S_{ee}}{S_{yy}}$$

```
> ### 重相関係数と残差の関係
> e <- y-f1$pred # 残差
> cor(f1$pred,y)^2 # R^2
        [,1]
Y 0.1988971
> sum(f1$pred^2)/sum(y^2) # 中心化しないとダメ
[1] 0.998469
> sum((f1$pred-mymean1(f1$pred))^2)/sum((y-mymean1(y))^2) # R^2
[1] 0.1988971
> myvar1(f1$pred)/myvar1(y) # これでもOK
[1] 0.1988971
> 1-sum(e^2)/sum(y^2) # 中心化しないとダメ
[1] 0.998469
> 1-sum(e^2)/sum((y-mymean1(y))^2) # R^2
[1] 0.1988971
> 1-myvar1(e)/myvar1(y) # これでもOK
[1] 0.1988971
```


y
e
ŷ + e = y
x2
ȳ
ŷ
(ŷ - ȳ)'e = 0
R² = ||ŷ - ȳ||² / ||y - ȳ||²
x1

部分回帰（Subset Regression）

モデル (k)

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i; \qquad i = 1, \ldots, n$$

すなわち

$$\beta_{k+1} = \cdots = \beta_p = 0$$

$$\boldsymbol{X}^{(k)} = \left[\mathbf{1}_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\right], \qquad \boldsymbol{\beta}^{(k)} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix}$$

$$\boldsymbol{y} = \boldsymbol{X}^{(k)} \boldsymbol{\beta}^{(k)} + \boldsymbol{\epsilon}$$

$$\hat{\boldsymbol{\beta}}^{(k)} = (\boldsymbol{X}^{(k)\prime} \boldsymbol{X}^{(k)})^{-1} \boldsymbol{X}^{(k)\prime} \boldsymbol{y}$$

一般に $\hat{\beta}^{(k)}$ **と** $\hat{\beta}$ **の対応する要素は一致しない:** $\qquad \hat{\beta}_i^{(k)} \neq \hat{\beta}_i$

```
> ### 部分回帰
> t(f1$coef)

    Intercept           X
  Y  33.17502 -0.01808129
> t(f2$coef)

    Intercept           x           x^2
  Y  32.24523 0.05023272 -0.0005693289
> t(f3$coef)

    Intercept          x           x^2           x^3
  Y  31.62375 0.1227429 -0.002052218 7.450801e-06
```

直交化とＱＲ分解

$$\boldsymbol{X} = \left[\boldsymbol{x}_0, \ldots, \boldsymbol{x}_p\right], \quad \boldsymbol{Q} = \left[\boldsymbol{q}_0, \ldots, \boldsymbol{q}_p\right], \quad \boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I}_{p+1}$$

$$\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}, \quad r_{ij} = 0, i > j; \quad \boldsymbol{\gamma} = \boldsymbol{R}\boldsymbol{\beta}$$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \boldsymbol{Q}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

$$\hat{\boldsymbol{\gamma}} = \boldsymbol{Q}'\boldsymbol{y}, \quad \hat{\boldsymbol{\beta}} = \boldsymbol{R}^{-1}\hat{\boldsymbol{\gamma}} = \boldsymbol{R}^{-1}\boldsymbol{Q}'\boldsymbol{y}$$

c.f. $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = (\boldsymbol{R}'\boldsymbol{R})^{-1}\boldsymbol{R}'\boldsymbol{Q}' = \boldsymbol{R}^{-1}\boldsymbol{Q}'$

$\gamma_k = r_{kk}\beta_k + \cdots + r_{kp}\beta_p$**に注意すると**

$$\gamma_{k+1} = \cdots = \gamma_p = 0 \Leftrightarrow \beta_{k+1} = \cdots = \beta_p = 0$$

モデル(k)

$$\hat{\boldsymbol{\gamma}}^{(k)} = \boldsymbol{Q}^{(k)\prime}\boldsymbol{y}, \quad \hat{\gamma}_i^{(k)} = \hat{\gamma}_i$$

```
> ### 直交化とQR分解
> xx <- cbind(1,x,x^2,x^3)
> dimnames(xx)[[2]] <- c(1,"x","x^2","x^3")
> Q <- qr.Q(qr(xx))
> R <- qr.R(qr(xx))
> cbind(xx,Q)
           1   x   x^2     x^3
Hokkaido   1 145 21025 3048625 -0.145865 -0.46469866  0.58635783  0.54555490
Aomori     1 119 14161 1685159 -0.145865 -0.35684268  0.16893012 -0.16646877
Iwate      1 110 12100 1331000 -0.145865 -0.31950791  0.05955370 -0.25620527
Miyagi     1  73  5329  389017 -0.145865 -0.16602055 -0.20035745 -0.12673425
Akita      1 112 12544 1404928 -0.145865 -0.32780453  0.08229878 -0.24191600
Yamagata   1  97  9409  912673 -0.145865 -0.26557992 -0.06654976 -0.28255814
Fukushima  1  69  4761  328509 -0.145865 -0.14942732 -0.21017241 -0.08803184
Ibaraki    1  10   100    1000 -0.145865  0.09532280  0.05944675 -0.01420506
Tochigi    1  13   169    2197 -0.145865  0.08287788  0.02700782  0.03316958
Gumma      1   8    64     512 -0.145865  0.10361942  0.08218756 -0.05056613
Saitama    1   9    81     729 -0.145865  0.09947111  0.07070567 -0.03189642
```

Chiba	1	4	16	64	-0.145865	0.12021265	0.13034481	-0.13536963
Tokyo	1	7	49	343	-0.145865	0.10776772	0.09389241	-0.07023124
Kanagawa	1	7	49	343	-0.145865	0.10776772	0.09389241	-0.07023124
Niigata	1	74	5476	405224	-0.145865	-0.17016885	-0.19734628	-0.13616848
Toyama	1	50	2500	125000	-0.145865	-0.07060948	-0.20807451	0.08707592
Ishikawa	1	50	2500	125000	-0.145865	-0.07060948	-0.20807451	0.08707592
Fukui	1	48	2304	110592	-0.145865	-0.06231287	-0.20317131	0.10168360
Yamanashi	1	13	169	2197	-0.145865	0.08287788	0.02700782	0.03316958
Nagano	1	71	5041	357911	-0.145865	-0.15772393	-0.20571087	-0.10754201
Gifu	1	25	625	15625	-0.145865	0.03309820	-0.08268060	0.14536604
Shizuoka	1	5	25	125	-0.145865	0.11606434	0.11797104	-0.11261593
Aichi	1	18	324	5832	-0.145865	0.06213635	-0.02259767	0.09415285
Mie	1	20	400	8000	-0.145865	0.05383973	-0.04087907	0.11265321
Shiga	1	39	1521	59319	-0.145865	-0.02497810	-0.17006988	0.14995720
Kyoto	1	37	1369	50653	-0.145865	-0.01668149	-0.16026134	0.15599117
Osaka	1	14	196	2744	-0.145865	0.07872957	0.01664078	0.04711818
Hyogo	1	22	484	10648	-0.145865	0.04554312	-0.05826859	0.12799090
Nara	1	25	625	15625	-0.145865	0.03309820	-0.08268060	0.14536604
Wakayama	1	13	169	2197	-0.145865	0.08287788	0.02700782	0.03316958

```
Tottori    1  43  1849    79507 -0.145865 -0.04157133 -0.18701133  0.13242784
Shimane    1  38  1444    54872 -0.145865 -0.02082980 -0.16527710  0.15321596
Okayama    1  17   289     4913 -0.145865  0.06628465 -0.01312251  0.08367402
Hiroshima  1  21   441     9261 -0.145865  0.04969142 -0.04968532  0.12070886
Yamaguchi  1  22   484    10648 -0.145865  0.04554312 -0.05826859  0.12799090
Tokushima  1  12   144     1728 -0.145865  0.08702619  0.03759783  0.01831088
Kagawa     1   9    81      729 -0.145865  0.09947111  0.07070567 -0.03189642
Ehime      1  10   100     1000 -0.145865  0.09532280  0.05944675 -0.01420506
Kochi      1   8    64      512 -0.145865  0.10361942  0.08218756 -0.05056613
Fukuoka    1  14   196     2744 -0.145865  0.07872957  0.01664078  0.04711818
Saga       1  17   289     4913 -0.145865  0.06628465 -0.01312251  0.08367402
Nagasaki   1  10   100     1000 -0.145865  0.09532280  0.05944675 -0.01420506
Kumamoto   1  10   100     1000 -0.145865  0.09532280  0.05944675 -0.01420506
Ooita      1   6    36      216 -0.145865  0.11191603  0.10582024 -0.09090882
Miyazaki   1   1     1        1 -0.145865  0.13265757  0.16880393 -0.21008090
Kagoshima  1   5    25      125 -0.145865  0.11606434  0.11797104 -0.11261593
Okinawa    1   0     0        0 -0.145865  0.13680587  0.18206958 -0.23719159
> R
                    1          x        x^2         x^3
```

```
Hokkaido -6.855655 -226.0907 -15932.541 -1542516.5
Aomori    0.000000 -241.0622 -28925.150 -3410817.3
Iwate     0.000000    0.0000   8969.815  1785209.7
Miyagi    0.000000    0.0000      0.000   351663.8
> ir <- solve(R)
> ir
          [,1]          [,2]           [,3]           [,4]
1   -0.145865  0.136805874  0.1820695765 -2.371916e-01
x    0.000000 -0.004148307 -0.0133771326  2.767379e-02
x^2  0.000000  0.000000000  0.0001114850 -5.659501e-04
x^3  0.000000  0.000000000  0.0000000000  2.843625e-06
> g3 <- t(Q) %*% y # gamma
> t(g3)
          [,1]     [,2]      [,3]     [,4]
[1,] -223.3485 4.358715 -5.106775 2.620177
> t(ir %*% g3)  # betaに等しいハズ

               1         x           x^2           x^3
  [1,] 31.62375 0.1227429 -0.002052218 7.450801e-06
```

```
> t(f3$coef) # beta

    Intercept         x          x^2          x^3
Y  31.62375 0.1227429 -0.002052218 7.450801e-06
> t(ir[1:2,1:2] %*% g3[1:2]) # gamma (1次)

             1           x
[1,] 33.17502 -0.01808129
> t(f1$coef)

    Intercept           X
Y  33.17502 -0.01808129
> t(ir[1:3,1:3] %*% g3[1:3]) # gamma (2次)

             1          x           x^2
[1,] 32.24523 0.05023272 -0.0005693289
> t(f2$coef)

    Intercept          x           x^2
Y  32.24523 0.05023272 -0.0005693289
```

平方和の分解

$$\hat{\boldsymbol{y}} = \boldsymbol{Q}\hat{\boldsymbol{\gamma}} = \hat{\gamma}_0\boldsymbol{q}_0 + \cdots + \hat{\gamma}_p\boldsymbol{q}_p$$

$$\|\hat{\boldsymbol{y}}\|^2 = \|\hat{\gamma}_0\boldsymbol{q}_0\|^2 + \cdots + \|\hat{\gamma}_p\boldsymbol{q}_p\|^2 = \hat{\gamma}_0^2 + \cdots + \hat{\gamma}_p^2$$

$$\boldsymbol{q}_0 = \frac{1}{\sqrt{n}}\mathbf{1}_n, \quad \hat{\gamma}_0 = \sqrt{n}\,\bar{y}$$

モデル (k)

$$\|\boldsymbol{Q}^{(k)}\hat{\boldsymbol{\gamma}}^{(k)} - \bar{y}\mathbf{1}_n\|^2 = \hat{\gamma}_1^2 + \cdots + \hat{\gamma}_k^2$$

$$\|\boldsymbol{e}\|^2 = \|\boldsymbol{y}\|^2 - \hat{\gamma}_0^2 - \cdots - \hat{\gamma}_k^2 = \|\boldsymbol{y} - \bar{y}\mathbf{1}_n\|^2 - \hat{\gamma}_1^2 - \cdots - \hat{\gamma}_k^2$$

$$R^2 = \frac{\hat{\gamma}_1^2 + \cdots + \hat{\gamma}_k^2}{\|\boldsymbol{y} - \bar{y}\mathbf{1}_n\|^2}$$

```
> ### 平方和の分解
> sum((f1$pred-mean(f1$pred))^2) # 予測値の平方和（１次）
[1] 18.99839
> sum(g3[2]^2) # g[2]^2
[1] 18.99839
> sum((f2$pred-mean(f2$pred))^2) # 予測値の平方和（２次）
[1] 45.07754
> sum(g3[2:3]^2) # g[2]^2+g[3]^2
[1] 45.07754
> sum((f3$pred-mean(f3$pred))^2) # 予測値の平方和（３次）
[1] 51.94287
> sum(g3[2:4]^2) # g[2]^2+g[3]^2+g[4]^2
[1] 51.94287
> cor(f1$pred,y)^2 # R^2 （１次）
        [,1]
Y 0.1988971
> sum(g3[2]^2)/sum((y-mean(y))^2)
[1] 0.1988971
```

```
> cor(f2$pred,y)^2 # R^2 (２次)
        [,1]
Y 0.4719236
> sum(g3[2:3]^2)/sum((y-mean(y))^2)
[1] 0.4719236
> cor(f3$pred,y)^2 # R^2 (３次)
        [,1]
Y 0.5437978
> sum(g3[2:4]^2)/sum((y-mean(y))^2)
[1] 0.5437978
```

第４回 課題

１．直線当てはめ（単回帰）の関数kaiki1を作れ．

```
kaiki1 <- function(x,y) {
# x,yは同じ長さの実数ベクトル
# y = coef[1] + coef[2]*x + residの形の単回帰分析を行う
# 以下のcoef, pred, residを計算する
#   coef(係数)は２次元ベクトル
#   resid(残差)はyと同じ長さのベクトル
#   pred(予測値) = coef[1] + coef[2]*xはyと同じ長さのベクトル
# 次の行は結果をリストとして返す.
list(coef,pred,resid)
}
```

２．重回帰分析の関数kaiki2を作れ．

```
kaiki2 <- function(x,y) {
# xはn * p次元の行列
# yは長さnのベクトル
```

```
# y = coef[1] + coef[2]*x[,1] + ... + coef[p+1]*x[,p] + resid
#  の形の重回帰分析を行う
# 以下の coef, pred, resid を計算する
#   coef(係数)は p+1 次元ベクトル
#   resid(残差)は y と同じ長さのベクトル
#   pred(予測値)は y と同じ長さのベクトル
# 次の行は結果をリストとして返す.
list(coef,pred,resid)
}
```

3. kaiki2の返す値から重相関係数の二乗 R^2 を計算する関数 jyusokansq を作れ.

```
jyusokansq <- function(kout) {
# kout$pred は予測値, kout$resid は残差
# これらから重相関係数の二乗を計算し rsq に代入
rsq
}
```

4. kaiki2, jyusoukansqを使い，X2000$xから適当な項目を選んで重回帰分析する．係数$\beta$と重相関係数$R$を計算する．myfunc20020919.Rにあるmylsfitをつかって同じ分析をして，結果が同じになるかどうか確認する．

5. 上で得られた結果について，predをx軸，yをy軸とするプロットをする．x軸＝y軸となる直線を描く(abline(0,1)をつかう)．さらに県名を使ったプロットをする．myfunc20020919.Rのmyplot関数を参考にせよ．プロットはmyfunc20020919.Rにあるpsinit関数などを使いepsファイルとして出力し，それをプリンタで印刷する．

```
psinit("ファイル名") # これ以後のプロットの結果をファイルにeps形式で書き出す
ここでプロットをおこなう...
dev.off() # ファイルをクローズする
```