クレジット:
UTokyo Online Education 統計データ解析 Ⅱ 2018 小池祐太



東京大学 THE UNIVERSITY OF TOKYO

CC BY NC ND

# 統計データ解析 II (平成３０年度)

東京大学大学院数理科学研究科
統計データ解析教育研究グループ

村田　昇 (早稲田大学, 東京大学)
吉田朋広 (東京大学)
小池祐太 (東京大学)

# 第 5 章 重回帰分析

## 5.1. 目的

**回帰分析** *(regression analysis)* とは, ある 1 種類の変数/データを別の変数/データ (1 種類もしくは複数) によって説明もしくは予測するための関係式 (**回帰 (方程) 式** *(regression equation)*) を構成することを目的とする分析法である.

- 説明される側のデータは, 目的変数, 被説明変数, 従属変数, 応答変数などと呼ばれる.
- 説明する側のデータは, 説明変数, 独立変数, 共変量などと呼ばれる.

説明変数が 1 種類の場合を**単回帰** *(simple regression)*, 複数の場合を**重回帰** *(multiple regression)* と呼ぶ.

以下, 目的変数を $Y$, 説明変数を $X_1, \ldots, X_p$ で表すことにし, 組 $(Y, X_1, \ldots, X_p)$ に対する $n$ 個の観測データ

$$\{(y_i, x_{i1}, \ldots, x_{ip})\}_{i=1}^n \tag{5.1}$$

が得られている状況を考える. $Y$ を $X_1, \ldots, X_p$ で説明するための関係式は, 一般にはある $p$ 変数関数 $f$ を使って,

$$Y = f(X_1, \ldots, X_p) \tag{5.2}$$

と書ける. この形では一般的すぎて分析に不向きのため, 多くの場合 $f$ として 1 次関数のみを考える. すなわち, ある定数 $\beta_0, \beta_1, \ldots, \beta_p$ によって

$$f(x_1, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

と書ける場合を考察する. この場合を**線形回帰** *(linear regression)* と呼ぶ. 一般の場合は**非線形回帰** *(nonlinear regression)* と呼ばれる. なお, 例えば $X_j^2$ や $X_j X_k$ といった 2 次式や, $\log X_j$ などの適切な非線形関数で変換した説明変数を新たな説明変数として加えることにすれば, 説明変数と目的変数の間のある程度非線形な関係を線形回帰モデルでも表すことができることに注意する.

線形回帰の場合, 回帰式 (5.2) は

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{5.3}$$

という形になる. パラメーター $\beta_0, \beta_1, \ldots, \beta_p$ は**回帰係数** *(regression coefficient)* と呼ばれ, 特に $\beta_0$ を**定数項** *(constant term)* と呼ぶ. 回帰係数は未知なので, 回帰式の構成にはそれらをデータから決定することが必要となる. 次節でこの点について議論する.

## 5.2. 回帰係数の推定

一般にあるモデルを考えたとき, そのモデルに含まれる未知パラメーターを観測データから (何らかの意味で) 決定する作業を**推定** *(estimation)* と呼ぶ. 我々の目標は, 回帰式 (5.3) に含まれるパラメーター $\beta_0, \beta_1, \ldots, \beta_p$ を観測データ (5.1) から推定することである.

**5.2.1. 確率モデル.** 実際のデータは観測誤差などのランダムな変動を含むと考えられるため, 回帰式 (5.3) が観測データに対してそのまま成立するとは期待しづらい. そのため, 統計学では, データのばらつきを表す項を $\epsilon_i$ として, 以下の形の確率モデルを分析することを考える:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \qquad i = 1, \ldots, n. \tag{5.4}$$

このモデルにおいて, $\epsilon_1, \ldots, \epsilon_n$ は**誤差項** *(error term)* もしくは**撹乱項** *(disturbance term)* と呼ばれる. 誤差項は確率変数であり, しばしば平均 0 の正規乱数列と仮定される.

**5.2.2. 最小二乗法.** 回帰係数の推定には通常**最小二乗法** *(least squares)* が用いられる. 最小二乗法の考え方は以下の通りである. 回帰係数 $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ を 1 つ決めたとき, 回帰式では説明できない目的変数の変動は,

$$e_i(\boldsymbol{\beta}) = y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}), \qquad i = 1, \ldots, n$$

で与えられる. これらの変動 $e_1(\boldsymbol{\beta}), \ldots, e_n(\boldsymbol{\beta})$ はいずれも絶対値が小さいほど当てはまりがよいと考えられる. そこで, 最小二乗法では, $e_1(\boldsymbol{\beta}), \ldots, e_n(\boldsymbol{\beta})$ の平方和

$$S(\boldsymbol{\beta}) := \sum_{i=1}^{n} e_i(\boldsymbol{\beta})^2$$

を最小にするように回帰係数 $\boldsymbol{\beta}$ を決定する. $S(\boldsymbol{\beta})$ は**残差平方和** *(residual sum of squares)* と呼ばれ, $S(\boldsymbol{\beta})$ を最小にする $\boldsymbol{\beta}$ は**最小二乗推定量** *(least squares estimator)* と呼ばれる. 最小二乗推定量はしばしば記号 $\hat{\boldsymbol{\beta}}$ で表される.

**5.2.3. 線形回帰式の行列による表現.** 最小二乗推定量の計算には, モデル (5.4) を行列を用いて表現しておくと便利である.

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix},$$

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \qquad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

とおくと, モデル (5.4) は

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

と表すことができる. なお, 行列 $\boldsymbol{X}$ は**デザイン行列** *(design matrix)* と呼ばれることがある. 更に, 一般に列ベクトル $\boldsymbol{a} = (a_1, \ldots, a_n)^\top$ に対して, $\boldsymbol{a}^\top \boldsymbol{a}$ は $\boldsymbol{a}$ の各成分の二乗の総和 $\sum_{i=1}^n a_i^2$ に一致することに注意すれば, 残差平方和は

$$S(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

と表せる.

**5.2.4. 正規方程式.** もし $\boldsymbol{\beta}$ が最小二乗推定量ならば, 点 $\boldsymbol{\beta}$ における残差二乗和の勾配は零ベクトルとなる必要がある[1]:

$$\nabla S(\beta) := \left( \frac{\partial S}{\partial \beta_0}(\boldsymbol{\beta}), \frac{\partial S}{\partial \beta_1}(\boldsymbol{\beta}), \ldots, \frac{\partial S}{\partial \beta_p}(\boldsymbol{\beta}) \right)^\top = \boldsymbol{0}. \tag{5.5}$$

[1] 例えば参考文献 4. の第 II 章定理 8.1 参照

各 $j = 0, 1, \ldots, p$ について偏微分を計算すると,

$$\frac{\partial S}{\partial \beta_j}(\boldsymbol{\beta}) = -2 \sum_{i=1}^{n} \left( y_i - \sum_{k=0}^{p} \beta_k x_{ik} \right) x_{ij}$$

となる. 但し, $x_{i0} = 1$ $(i = 1, \ldots, n)$ とおいた. 従って方程式 (5.5) は,

$$\sum_{i=1}^{n} x_{ij} \left( \sum_{k=0}^{p} x_{ik} \beta_k \right) = \sum_{i=1}^{n} x_{ij} y_i \qquad (j = 0, 1, \ldots, p)$$

と書き直せる. $x_{ij}$ が行列 $\boldsymbol{X}$ の $(i, j)$ 成分に対応していることに注意すれば, 上の方程式は

$$\boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{\beta} = \boldsymbol{X}^\top \boldsymbol{y} \tag{5.6}$$

と書ける. 方程式 (5.6) を**正規方程式** *(normal equations)* と呼ぶ.

一般に, $\boldsymbol{\beta}$ が方程式 (5.5) の解であることは, $\boldsymbol{\beta}$ が $S(\boldsymbol{\beta})$ を最小にするための必要条件であって十分条件であるとは限らない. しかし, いまの状況では実は十分条件となっていることを示すことができる (5.7 節定理 5.3 参照). 従って正規方程式の解はすべて最小二乗推定量である. さらに, 正規方程式は常に解を持つことが証明できる (節定理 5.2 参照). 以上の結果から, どのような状況下であっても最小二乗推定量は常に存在する. しかし, 実際のデータ解析をする上では, 最小二乗推定量がただ一つだけ存在する状況が好ましい (例えば推定量の選択によって分析結果が変化してしまうことが避けられる). 次の定理はそのような状況が起きるための必要十分条件を与える:

**定理 5.1.** 正規方程式がただ一つの解をもつための必要十分条件は, $(p+1)$ 次正方行列 $\boldsymbol{X}^\top \boldsymbol{X}$ が正則であることであり, このとき正規方程式の解は

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \tag{5.7}$$

で与えられる.

**証明.** $\boldsymbol{X}^\top \boldsymbol{X}$ が正則ならば, 正規方程式の両辺に $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ をかけることで, 正規方程式が (5.7) で与えられるただ一つの解をもつことがわかる. 逆に正規方程式がただ一つの解をもつ場合に $\boldsymbol{X}^\top \boldsymbol{X}$ が正則となることを背理法で示す. もし $\boldsymbol{X}^\top \boldsymbol{X}$ が非正則ならば, 方程式 $\boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{b} = \boldsymbol{0}$ は非自明な解 $\boldsymbol{b}^*$ をもつ.[2] いま $\hat{\boldsymbol{\beta}}$ を解とすると, $\hat{\boldsymbol{\beta}} + \boldsymbol{b}^*$ も明らかに正規方程式の解であるが, $\hat{\boldsymbol{\beta}} + \boldsymbol{b}^* \neq \hat{\boldsymbol{\beta}}$ であるため, これは正規方程式の解の一意性に矛盾する. □

定理 5.1 より正規方程式の解の一意性は行列 $\boldsymbol{X}^\top \boldsymbol{X}$ の正則性に帰着される. $\boldsymbol{X}^\top \boldsymbol{X}$ は $\boldsymbol{X}$ の **Gram 行列** *(Gram matrix)* と呼ばれることがある. $\boldsymbol{X}^\top \boldsymbol{X}$ が正則であることは, $\boldsymbol{X}$ の列ベクトルが 1 次独立であることと同値である.[3] $\boldsymbol{X}$ の各列は各説明変数の観測データからなるベクトルで与えられているから, これは回帰式の構築に利用する説明変数たちが互いに異なる情報をもつという意味だと解釈できる. 説明変数の間の関係の 1 次従属関係への近さの度合いは**多重共線性** *(multicollinearity)* と呼ばれる. データ解析をする上では説明変数は多重共線性があまり強くならないように選択すべきである. 例えば, 似たような動きをする説明変数を重複して利用することは避けるべきである.

以下では特に断らない限り $\boldsymbol{X}^\top \boldsymbol{X}$ は正則であると仮定して議論を進める.

---

[2] 例えば, 参考文献 1. の系 2.3.6 参照.

[3] $\boldsymbol{X}^\top \boldsymbol{X}$ が正則であることは, $\boldsymbol{X}^\top \boldsymbol{X}$ の階数が $(p+1)$ であることと同値であり, 補題 5.1 より $\boldsymbol{X}^\top \boldsymbol{X}$ の階数は $\boldsymbol{X}^\top$ の階数と一致するため.

**5.2.5. 最小二乗法による線形回帰式の推定の幾何学的解釈.** 一般に $\hat{\boldsymbol{\beta}}$ を回帰係数の推定値とすると, 目的変数の観測データ $\boldsymbol{y}$ から観測誤差の影響を除いた目的変数の真の値が,

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

によって推定できる. $\hat{\boldsymbol{y}}$ を**あてはめ値** *(fitted values)* または**予測値** *(predicted values)* と呼ぶ. $\hat{\boldsymbol{\beta}}$ が最小二乗推定量であれば, あてはめ値 $\hat{\boldsymbol{y}}$ は, 幾何学的には, ベクトル $\boldsymbol{y}$ からデザイン行列 $\boldsymbol{X}$ の列ベクトルによって張られる (超) 平面 $L[X]$ に垂線を下ろした際の垂線の足となる.[4] 図 1 に $n = 3, p+1 = 2$ の場合に最小二乗法による推定の様子を図示したものを示す.

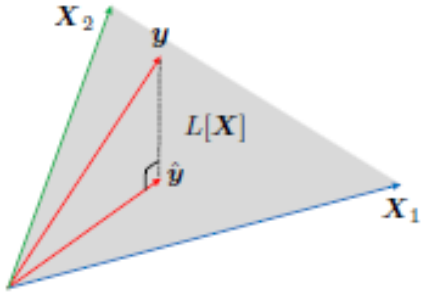



図 1. $n = 3, p+1 = 2$ の場合の最小二乗法による推定の図示. $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ はそれぞれデザイン行列 $\boldsymbol{X}$ の第 1 列, 第 2 列からなるベクトルに対応している. グレーの平面が $L[\boldsymbol{X}]$ に対応.

特に, ベクトル $\hat{\boldsymbol{\epsilon}} := \boldsymbol{y} - \hat{\boldsymbol{y}}$ はあてはめ値 $\hat{\boldsymbol{y}}$ に直交する: $\hat{\boldsymbol{\epsilon}} \cdot \hat{\boldsymbol{y}} = 0$. $\hat{\boldsymbol{\epsilon}}$ は**残差** *(residuals)* と呼ばれ, 回帰式による目的変数のあてはめ値と実際の観測値とのずれを表す.

**5.2.6. 線形回帰式と標本平均.** 最小二乗法によって構成した線形回帰式から定まる超平面は, 目的変数および説明変数それぞれの標本平均からなる点を必ず通ることが知られている. このことを正確に述べるためにいくつか記号を定義する. 各 $i = 1, \ldots, n$ について, 説明変数の $i$ 番目の観測データに対応するベクトル $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ を導入する. そして,

$$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

とおく. $\bar{\boldsymbol{x}}$ および $\bar{y}$ はそれぞれ説明変数および目的変数の標本平均となっている. このとき, 等式

$$\bar{y} = (1\ \bar{\boldsymbol{x}}^\top)\hat{\boldsymbol{\beta}} \tag{5.8}$$

が成り立つことが知られている (導出は 5.8 節を参照). この等式を幾何学的に解釈すると, 線形回帰式 $y = (1\ \boldsymbol{x}^\top)\hat{\boldsymbol{\beta}}$ によって定まる超平面は常に点 $(\bar{\boldsymbol{x}}^\top, \bar{y})$ を通るということになる.

[4] 定理 5.3 より $\hat{\boldsymbol{y}}$ は目的変数の観測データ $\boldsymbol{y}$ の $L[X]$ への直交射影となるため.

**5.2.7. R での実行.** R では線形回帰分析を実行するための関数 `lm()` が用意されている. モデル (5.4) において, 目的変数 $Y$ および説明変数 $X_1, \ldots, X_p$ の観測データに対応するベクトルがそれぞれ `y` および `x1`,...,`xp` で与えられているとする. このとき, モデル (5.4) の回帰係数の推定は, コマンド

`lm(y ~ x1 + ··· + xp)`

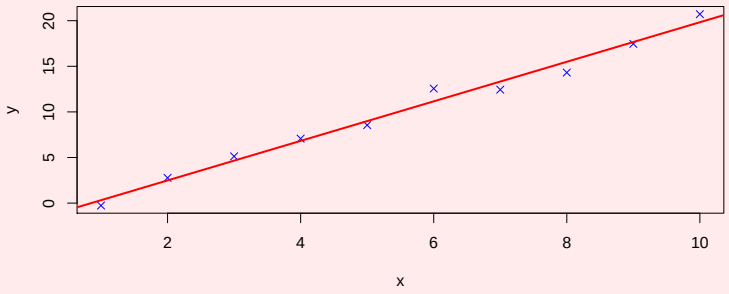
で実行できる. また, 実際のデータを使って解析する際は, データセットの一部の変数を目的変数および説明変数として回帰分析をすることが多い. そのような場合, データセットに対応するデータフレームを `dat` とすれば, 以下のコマンドで回帰係数の推定を実行できる:

`lm(`$Y$ の変数名 `~` $X_1$ の変数名 `+ ··· +` $X_p$ の変数名`, data = dat)`

ここで, `dat` は列が各変数に対応するような形式になっている必要がある.

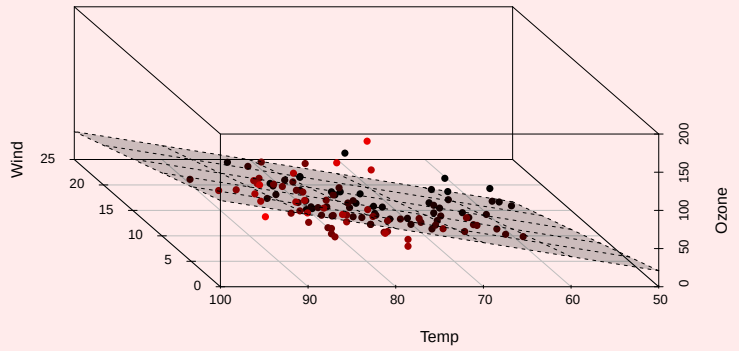
```
> # 人工データに対する単回帰分析の例
> # モデル: y = -1 + 2x
> set.seed(123) # 乱数の初期値の固定
> x <- c(7, 2, 6, 4, 3, 10, 9, 1, 8, 5) # 説明変数の観測データ
> epsilon <- rnorm(length(x)) # 誤差項
> y <- -1 + 2 * x + epsilon # 目的変数の観測データ
> (out <- lm(y ~ x)) # 回帰係数の推定

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     -1.848        2.168

> (b <- coef(out)) # 推定された回帰係数の出力
(Intercept)           x
  -1.848358    2.167815
> ## 最小二乗推定量の計算公式との確認
> X <- cbind(rep(1, length(x)), x) # デザイン行列
> G <- crossprod(X) # Gram 行列 (t(X) %*% X と同じ)
> solve(G) %*% crossprod(X, y)
       [,1]
  -1.848358
x  2.167815
> ## あてはめ値と残差が直交することの確認
> (yhat <- fitted(out)) # あてはめ値
         1          2          3          4          5          6          7
13.3263484  2.4872725 11.1585332  6.8229029  4.6550877 19.8297940 17.6619788
         8          9         10
 0.3194573 15.4941636  8.9907180
> (epshat <- resid(out)) # 残差
         1          2          3          4          5          6          7
-0.8868241  0.2825500  1.4001751  0.2476055  0.4742001  0.8852710 -0.2010626
         8          9         10
-0.5845185 -1.1810165 -0.4363800
> yhat %*% epshat
             [,1]
[1,] 1.820766e-14
> ## 回帰直線が標本平均を通ることの確認
> c(1, mean(x)) %*% b
         [,1]
[1,] 10.07463
> mean(y)
[1] 10.07463
```

```
> ## 散布図と回帰直線の描画
> plot(x, y, pch = 4, col = "blue") # 散布図
> abline(b, col = "red", lwd = 2) # 直線 y = b[1] + b[2] * x の描画
```




```
> # R の組込データセット airquality による重回帰分析の例
> # Ozone を目的変数 Wind, Temp を説明変数とする
> model <- Ozone ~ Wind + Temp # モデルの定義．formula クラスが返る
> # class(model)
> (est <- lm(model, data = airquality)) # 回帰係数の推定

Call:
lm(formula = model, data = airquality)

Coefficients:
(Intercept)         Wind         Temp
    -71.033       -3.055        1.840

> # 他の書き方
> # my.data <- na.omit(subset(airquality,select = c(Ozone,Wind,Temp)))
> # (est <- lm(formula = Ozone ~ Wind+Temp, data = my.data))
> (b <- coef(est)) # 推定された回帰係数の出力

(Intercept)        Wind        Temp
 -71.033218   -3.055491    1.840179

> ## 最小二乗推定量の計算公式との一致を確認
> dat <- model.frame(model, data = airquality) # model に必要な変数の抽出
> # または dat <- subset(airquality,select = c(Ozone,Wind,Temp))
> dat <- na.omit(dat) # 欠測の削除
> X <- model.matrix(model, data = dat) # デザイン行列
> # または X <- cbind(rep(1,nrow(dat)),dat$Wind,dat$Temp)
> G <- crossprod(X) # Gram 行列 (t(X) %*% X と同じ)
> solve(G) %*% crossprod(X, dat$Ozone)
                  [,1]
(Intercept) -71.033218
Wind         -3.055491
Temp          1.840179

> ## あてはめ値と残差が直交することの確認
> yhat <- fitted(est) # あてはめ値
> epshat <- resid(est) # 残差
> yhat %*% epshat
             [,1]
[1,] -7.48912e-12

> ## 回帰式が標本平均を通ることの確認
> colMeans(X) %*% b
         [,1]
[1,] 42.12931
```

```
> # ここで, X は切片項を含むことに注意. head(X)
> mean(dat$Ozone)
[1] 42.12931
> ## 散布図と回帰式の定める平面の描画
> library(scatterplot3d) # パッケージのロード
> obj <- scatterplot3d(dat[ ,c("Wind", "Temp", "Ozone")],
+                      pch = 16, angle = -60,
+                      highlight.3d = TRUE)
> # いろいろ試す
> # obj <- scatterplot3d(dat[ ,c("Wind", "Temp", "Ozone")],
> #           pch = 16, angle = -120, type = "h", highlight.3d = T)
> obj$plane3d(b, draw_polygon = TRUE) # 回帰式の定める平面の追加
```




(lse.r)

## 5.3. 分析の評価

ここでは関数 lm( ) のアウトプットに関数 summary( ) を適用した際に表示される, 分析結果の評価をするための各種指標について解説する.

**5.3.1. 残差.** 関数 summary( ) のアウトプットの “Residuals” の欄には残差 $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$ に対する五数要約 (最小値, 第 1 分位点, メディアン, 第 3 分位点, 最大値) が表示される. 残差は絶対値が小さいほど回帰式の観測データへのあてはまりがよいこととなるので, 残差のばらつきは小さいほどよい.

```
> # 人工データに対する単回帰分析の例
> # モデル: y = -1 + 2x
> set.seed(123) # 乱数の初期値の固定
> x <- c(7, 2, 6, 4, 3, 10, 9, 1, 8, 5) # 説明変数の観測データ
> epsilon <- rnorm(length(x)) # 誤差項
> y <- -1 + 2 * x + epsilon # 目的変数の観測データ
> out <- lm(y ~ x) # 回帰分析の実行
> summary(out)
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.18102 -0.54748  0.02327  0.42629  1.40018

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.84836    0.58486   -3.16   0.0134 *
x            2.16782    0.09426   23.00 1.36e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8562 on 8 degrees of freedom
Multiple R-squared:  0.9851,        Adjusted R-squared:  0.9832
F-statistic: 528.9 on 1 and 8 DF,  p-value: 1.356e-08
> quantile(resid(out)) # 自力で残差の五数要約を計算する場合
         0%         25%         50%         75%        100%
-1.18101648 -0.54748388  0.02327146  0.42628757  1.40017507
```

(resid.r)

**5.3.2. 標準誤差.** モデル (5.4) において, 誤差項 $\epsilon_1, \ldots, \epsilon_n$ は平均 0, 分散 $\sigma^2$ の正規乱数であると仮定する. 最小二乗推定量 $\hat{\boldsymbol{\beta}}$ は誤差項に依存して変化するため確率変数であるが, いまの仮定の下では平均 $\boldsymbol{\beta}$, 共分散行列 $\sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ の $(p+1)$ 変量正規分布に従うことが知られている. 特に, 行列 $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ の対角成分を $\xi_0, \xi_1, \ldots, \xi_p$ とすれば, 各 $j = 0, 1, \ldots, p$ について, $\hat{\beta}_j$ は平均 $\beta_j$, 分散 $\sigma^2\xi_j$ の正規分布に従う. 正規分布の分散, もしくはその平方根である標準偏差は, 確率変数の値の平均からの離れやすさを表す指標だと解釈できる (大きいほど離れやすい). $\hat{\beta}_j$ の値は “真の” 回帰係数値 $\beta_j$ に近ければ近いほどよい推定であるといえるから, その標準偏差 $\sigma\sqrt{\xi_j}$ は $\hat{\beta}_j$ の推定精度を評価するのに利用できる. ただし, $\sigma$ は未知パラメーターだから, データから推定する必要がある. $\sigma$ の推定量としては, 通常

$$\hat{\sigma} = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n} \hat{\epsilon}_i^2}$$

が利用される (この推定量が利用される理由の直感的説明については注意 5.1 を参照). こうして推定値 $\hat{\beta}_j$ の精度を評価するための指標 $\hat{\sigma}\sqrt{\xi}_j$ が得られるが, これを $\hat{\beta}_j$ の**標準誤差** *(standard error)* と呼ぶ.

標準誤差は, 関数 `summary()` のアウトプットの “Coefficients” の欄の 2 列目 “Std. Error” で確認できる. また, 関数 `summary()` のアウトプットの “Residual standard error” の欄で $\hat{\sigma}$ の値を確認できる ($\hat{\sigma}$ は残差のばらつき具合を表す指標として利用できる).

**注意 5.1.** $\sigma^2$ は「データ」$\epsilon_1, \ldots, \epsilon_n$ の分散の「理論値」と考えられるから,「データ」から計算される分散 $\frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2$ によってよく近似できるはずである (平均 0 であることがわかっているため, 標本平均の調整と $\frac{1}{n}$ を $\frac{1}{n-1}$ に変更する操作は不要). しかし, 誤差項 $\epsilon_1, \ldots, \epsilon_n$ ももちろん観測できないため, これらを残差 $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$ で代替したもの $\frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^2$ を考えるのが自然である. しかし, 残差 $\hat{\epsilon}_i$ の構成には $p+1$ 個の未知パラメーター $\beta_0, \beta_1, \ldots, \beta_p$ の推定値が含まれているため, この分実質的なサンプル数が $p+1$ だけ減少する (このことを厳密に定式化するためには, 少し進んだ確率論の概念が必要). そのため, $\frac{1}{n}$ を $\frac{1}{n-p-1}$ に置き換える必要がある.

```
> ## 人工データに対する単回帰分析の例
> ## モデル: y = -1 + 2x
> set.seed(123) # 乱数の初期値の固定
> x <- c(7, 2, 6, 4, 3, 10, 9, 1, 8, 5) # 説明変数の観測データ
> epsilon <- rnorm(length(x)) # 誤差項
> y <- -1 + 2 * x + epsilon # 目的変数の観測データ
```

```
> out <- lm(y ~ x) # 回帰分析の実行
> summary(out)
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.18102 -0.54748  0.02327  0.42629  1.40018

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.84836    0.58486   -3.16   0.0134 *
x            2.16782    0.09426   23.00 1.36e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8562 on 8 degrees of freedom
Multiple R-squared:  0.9851,        Adjusted R-squared:  0.9832
F-statistic: 528.9 on 1 and 8 DF,  p-value: 1.356e-08
> coef(summary(out))[ ,2] # 標準誤差のみ抽出
(Intercept)           x
 0.58486347  0.09425928
> summary(out)$sigma # 誤差項の標準偏差の推定値
[1] 0.8561524
> ## 誤差項の標準偏差が標準誤差に与える影響を確認する
> epsilon <- rnorm(length(x), sd = 2) # 誤差項 (標準偏差 2 倍)
> y <- -1 + 2 * x + epsilon # 目的変数の観測データ
> out <- lm(y ~ x) # 回帰分析の実行
> summary(out)
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.88242 -0.91287 -0.02386  0.67309  2.22199

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.5993     0.9030  -3.986  0.00403 **
x             2.5485     0.1455  17.512 1.15e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.322 on 8 degrees of freedom
Multiple R-squared:  0.9746,        Adjusted R-squared:  0.9714
F-statistic: 306.7 on 1 and 8 DF,  p-value: 1.154e-07
> coef(summary(out))[ ,2] # 標準誤差のみ抽出
(Intercept)           x
  0.9029526   0.1455240
> summary(out)$sigma # 誤差項の標準偏差の推定値
[1] 1.321787
```

(se.r)

**5.3.3. $t$ 値と $p$ 値.** $(n-p-1)\hat{\sigma}^2/\sigma^2$ は自由度 $n-p-1$ のカイ二乗分布に従い, かつ $\hat{\boldsymbol{\beta}}$ と独立であることが知られている. $(\hat{\beta}_j - \beta_j)/(\sigma\sqrt{\xi_j})$ が標準正規分布に従う

ことに注意すれば, $\hat{\beta}_j - \beta_j$ を $\hat{\beta}_j$ の標準誤差で割った量

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{\xi_j}}$$

は自由度 $n-p-1$ の $t$ 分布に従うことがわかる. よって, もし $\beta_j = 0$ であったならば, 統計量

$$t = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{\xi_j}}$$

は自由度 $n-p-1$ の $t$ 分布に従う. この $t$ を $\hat{\beta}_j$ の $t$ **値** *(t-value)* と呼ぶ. また, 自由度 $n-p-1$ の $t$ 分布に従う確率変数の絶対値が $|t|$ を超える理論上の確率

$$2\int_{|t|}^{\infty} f(x)dx, \quad \text{但し, } f(x) \text{ は自由度 } n-p-1 \text{ の } t \text{ 分布の確率密度関数} \tag{5.9}$$

を $\hat{\beta}_j$ の $p$ **値** *(p-value)* と呼ぶ.

$\beta_j = 0$ が成り立つということは, $j$ 番目の説明変数 $X_j$ は回帰式に寄与していないこととなるから, 回帰式から除外しても問題無いことが示唆される. 上で定義した $t$ 値および $p$ 値は, $\beta_j = 0$ であるか否かという仮説を検証するのに利用できる. 実際, もし $\beta_j = 0$ という仮説が正しければ, 確率 (5.9) はそれほど小さくはならないはずなので, もし $p$ 値が想定よりも小さい場合, はじめの仮説である $\beta_j = 0$ が誤っているという結論した方が自然である. 統計の言葉で言うと, $t$ 値及び $p$ 値は, 仮説検定

$$H_0 : \beta_j = 0 \qquad \text{vs} \qquad H_1 : \beta_j \neq 0$$

に対する検定統計量の $t$ 値と $p$ 値となっている. また, ここでいうはじめに想定する $p$ 値の下限を**有意水準** *(significance level)* といい, 通常 0.01 もしくは 0.05 とすることが多い. $p$ 値が有意水準より小さいような回帰係数の推定値をもつ説明変数は**有意** *(significant)* であるといわれる. 有意な説明変数は目的変数の変動を説明するのに有用であるといえる.

$t$ 値および $p$ 値は, 関数 `summary()` のアウトプットの "Coefficients" の欄の 3–4 列目 "`t value`" および "`Pr(>|t|)`" で確認できる. また, $p$ 値の横についているアスタリスク等の記号は, $p$ 値がどの程度小さいかを示している. 例えば, '*' は $p$ 値が 0.05 以下であることを示し, '**' は $p$ 値が 0.05 以下であることを示す (記号の意味は "Coefficients" の欄の下部に書いてある).

```
> ## 人工データに対する単回帰分析の例
> ## モデル: y = -1 + 2x
> set.seed(123) # 乱数の初期値の固定
> x <- c(7, 2, 6, 4, 3, 10, 9, 1, 8, 5) # 説明変数の観測データ
> epsilon <- rnorm(length(x)) # 誤差項
> y <- -1 + 2 * x + epsilon # 目的変数の観測データ
> z <- runif(length(x)) # モデルに不要な説明変数
> out <- lm(y ~ x + z) # 回帰分析の実行
> summary(out)
Call:
lm(formula = y ~ x + z)

Residuals:
     Min       1Q   Median       3Q      Max
-1.21071 -0.42234 -0.03279  0.40204  1.36660

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.60022    0.96901  -2.683   0.0314 *
x            2.17912    0.09526  22.875 7.73e-08 ***
z            1.12036    1.14921   0.975   0.3621
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8588 on 7 degrees of freedom
Multiple R-squared:  0.9869,        Adjusted R-squared:  0.9831
F-statistic: 263.3 on 2 and 7 DF,  p-value: 2.586e-07
> coef(summary(out))[ ,3:4] # t 値と p 値のみ抽出
               t value     Pr(>|t|)
(Intercept) -2.6833759 3.138539e-02
x           22.8751932 7.731516e-08
z            0.9748969 3.620894e-01
```

(tpvalues.r)

**5.3.4. 決定係数. 決定係数** *(coefficient of determination, R squared)* は線形回帰分析のあてはまり具合を評価するためのもっとも代表的な指標である. 決定係数は記号 $R^2$ で表され, 回帰モデルによる目的変数のあてはめ値 $\hat{y}_1, \dots, \hat{y}_n$ と実際の観測データ $y_1, \dots, y_n$ の相関の 2 乗として定義される:

$$R^2 = \frac{\left(\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})\right)^2}{\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2 \sum_{i=1}^n (y_i - \bar{y})^2}. \tag{5.10}$$

ここに, $\bar{\hat{y}}$ と $\bar{y}$ はそれぞれ $\hat{y}_1, \dots, \hat{y}_n$ と $y_1, \dots, y_n$ の平均を表す:

$$\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^n \hat{y}_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^n y_i.$$

あてはめ値と実際の観測データの変動が近いほどあてはまりが良いと考えられるので, 決定係数は高ければ高いほどよい.

決定係数は以下のようにも書くことができる:

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}. \tag{5.11}$$

この式を示すために, まず $\bar{\hat{y}} = \bar{y}$ であることに注意する. 実際, 5.8 節の記号を使えば, $\hat{y}_i = (1, \boldsymbol{x}_i^\top)\hat{\boldsymbol{\beta}}$ と書けるので,

$$\bar{\hat{y}} = (1, \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i^\top)\hat{\boldsymbol{\beta}} = (1, \bar{\boldsymbol{x}}^\top)\hat{\boldsymbol{\beta}} = \bar{y}$$

となる. 従って,

$$\begin{aligned}
\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y}) &= \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})(y_i - \hat{y}_i) + \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2 \\
&= \sum_{i=1}^n \hat{y}_i \underbrace{(y_i - \hat{y}_i)}_{\hat{\epsilon}_i} - \bar{\hat{y}} \sum_{i=1}^n (y_i - \hat{y}_i) + \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2 \\
&= \hat{\boldsymbol{y}} \cdot \hat{\boldsymbol{\epsilon}} - n\bar{\hat{y}}(\bar{y} - \bar{\hat{y}}) + \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2 \\
&= \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2
\end{aligned}$$

が成り立つ. これを (5.10) に代入して (5.11) を得る. (5.11) の分子と分母をそれぞれ $n-1$ で割ることで, 決定係数はあてはめ値の分散を目的変数の観測データの分散で割ったものだとも解釈できる. すなわち, 目的変数の観測データの分散のうち何パーセントを回帰モデルが説明できているかを表す指標とも解釈できる.

決定係数はさらに以下のようにも書き直せる:

$$R^2 = 1 - \frac{\sum_{i=1}^n \hat{\epsilon}_i^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2}. \tag{5.12}$$

実際,

$$\begin{aligned}\sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n \hat{\epsilon}_i^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^n \hat{\epsilon}_i(\hat{y}_i - \bar{y}) \\ &= \sum_{i=1}^n \hat{\epsilon}_i^2 + \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2 + 2\hat{\boldsymbol{\epsilon}} \cdot \hat{\boldsymbol{y}} - 2\bar{y} \underbrace{\sum_{i=1}^n \hat{\epsilon}_i}_{n(\bar{y} - \bar{\hat{y}})} \\ &= \sum_{i=1}^n \hat{\epsilon}_i^2 + \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2\end{aligned}$$

より

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

であるから, これを (5.11) に代入して (5.12) を得る. (5.12) より, 決定係数は説明変数が多いほど高くなることがわかるため, 決定係数は本来回帰式に不要である説明変数の効果を過剰に見積もっているおそれがある. この問題を解消するために, 推定されて得られた未知パラメーターの影響を考慮して以下のように決定係数を修正したものが**自由度調整済み決定係数** *(adjusted R squared)* である:

$$\bar{R}^2 = 1 - \frac{\frac{1}{n-p-1}\sum_{i=1}^n \hat{\epsilon}_i^2}{\frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y}_i)^2}.$$

なお, 決定係数は**寄与率**とも呼ばれる.

決定係数および自由度調整済み決定係数は, それぞれ関数 `summary()` のアウトプットの “Multiple R-squared” および “Adjusted R-squared” の欄で確認できる.

```
> ## 人工データに対する例
> ## モデル: y = 1 - 2 * x1 + 3 * x2
> set.seed(123) # 乱数の初期値の固定
> x1 <- runif(50) # 説明変数 x1 の観測データ
> x2 <- runif(50) # 説明変数 x2 の観測データ
> x3 <- runif(50) # (不要な) 説明変数 x3 の観測データ
> epsilon <- rnorm(50, sd = 0.3) # 誤差項
> y <- 1 - 2 * x1 + 3 * x2 + epsilon # 目的変数の観測データ
> out1 <- lm(y ~ x1) # x1 による回帰分析の実行
> summary(out1)
Call:
lm(formula = y ~ x1)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6499 -0.7082 -0.1138  0.6251  1.6303

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.2684     0.2477   9.158 4.16e-12 ***
x1           -1.6855     0.4155  -4.056 0.000182 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.856 on 48 degrees of freedom
Multiple R-squared:  0.2553,        Adjusted R-squared:  0.2398
F-statistic: 16.45 on 1 and 48 DF,  p-value: 0.0001823
> summary(out1)$r.squared # 決定係数のみ抽出
[1] 0.2552839
> summary(out1)$adj.r.squared # 自由度調整済み決定係数のみ抽出
[1] 0.239769
> out2 <- lm(y ~ x1 + x3) # x1 と x3 による回帰分析の実行
> summary(out2)
Call:
lm(formula = y ~ x1 + x3)

Residuals:
     Min       1Q   Median       3Q      Max
-1.53192 -0.74658  0.02437  0.63994  1.51550

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.0764     0.3244   6.400 6.65e-08 ***
x1           -1.6957     0.4164  -4.073 0.000177 ***
x3            0.3828     0.4167   0.918 0.363057
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8574 on 47 degrees of freedom
Multiple R-squared:  0.2684,        Adjusted R-squared:  0.2373
F-statistic: 8.622 on 2 and 47 DF,  p-value: 0.000646
> summary(out2)$r.squared # 決定係数のみ抽出 (out1 より上昇)
[1] 0.2684152
> summary(out2)$adj.r.squared # 自由度調整済み決定係数のみ抽出 (out1 より下降)
[1] 0.2372839
> out3 <- lm(y ~ x1 + x2) # x1 と x2 による回帰分析の実行
> summary(out3)
Call:
lm(formula = y ~ x1 + x2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.52199 -0.17122 -0.06054  0.15732  0.64859

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.98181    0.09455   10.38 9.36e-14 ***
x1          -1.90178    0.12513  -15.20  < 2e-16 ***
x2           2.93287    0.13310   22.04  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.257 on 47 degrees of freedom
Multiple R-squared:  0.9343,        Adjusted R-squared:  0.9315
F-statistic: 334.1 on 2 and 47 DF,  p-value: < 2.2e-16
> summary(out3)$r.squared # 決定係数のみ抽出 (out1 より上昇)
[1] 0.9342752
> summary(out3)$adj.r.squared # 自由度調整済み決定係数のみ抽出 (out1 より上昇)
[1] 0.9314784
```

(rsquared.r)

**5.3.5. $F$ 値.** $t$ 値は個々の説明変数の要・不要を判断するための指標であったが, 説明変数のうち 1 つでも目的変数の説明の役に立つものがあるか否かを判定するための指標に回帰モデルの $F$ **値** *($F$-value)* がある. これは, 現在の説明変数を用いて回帰分析を実行することに意味があるかどうかを検証するための指標ともいえる. 回帰モデルの $F$ 値は次式で定義される:

$$F = \frac{\frac{1}{p}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \frac{n-p-1}{p}\frac{R^2}{1-R^2}.$$

もしすべての説明変数が不要, すなわち $\beta_1 = \cdots = \beta_p = 0$ であったならば, $F$ は自由度 $p, n-p-1$ の $F$ 分布に従うことが知られている. 従って, 自由度 $p, n-p-1$ の $F$ 分布に従う確率変数が $F$ を超える理論上の確率

$$\int_F^\infty f(x)dx, \quad \text{但し, } f(x) \text{ は自由度 } p, n-p-1 \text{ の } F \text{ 分布の確率密度関数} \tag{5.13}$$

はそれほど小さくはならないはずなので, この確率が想定より小さければ回帰分析に意味があると結論付けられる. 統計の言葉で言うと, $F$ 値及び確率 (5.13) は, 仮説検定

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \qquad \text{vs} \qquad H_1 : \text{ある } j = 1, \ldots, p \text{ に対して, } \beta_j \neq 0$$

に対する検定統計量の $F$ 値と $p$ 値となっている.

回帰モデルの $F$ 値および確率 (5.13) は, それぞれ関数 `summary()` のアウトプットの “F-statistic” およびその隣の “p-value” の欄で確認できる.

```
> ## 人工データに対する例
> ## モデル: y = 1 - 2 * x1
> set.seed(123) # 乱数の初期値の固定
> x1 <- runif(50) # 説明変数 x1 の観測データ
> x2 <- runif(50) # (不要な) 説明変数 x2 の観測データ
> x3 <- runif(50) # (不要な) 説明変数 x3 の観測データ
> epsilon <- rnorm(50, sd = 0.3) # 誤差項
> y <- 1 - 2 * x1 + epsilon # 目的変数の観測データ
> out1 <- lm(y ~ x2 + x3) # x2 と x3 による回帰分析の実行
> summary(out1)
Call:
lm(formula = y ~ x2 + x3)

Residuals:
     Min       1Q   Median       3Q      Max
-1.14898 -0.56041 -0.01379  0.39085  1.26912

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1122     0.2213   0.507    0.614
x2           -0.2077     0.3269  -0.635    0.528
x3           -0.1018     0.3076  -0.331    0.742

Residual standard error: 0.6242 on 47 degrees of freedom
Multiple R-squared:  0.01261,        Adjusted R-squared:  -0.02941
F-statistic: 0.3001 on 2 and 47 DF,  p-value: 0.7421
> ## F 値に対応する$p$値の計算
> res <- summary(out1)$fstatistic # F 値と自由度
> 1 - pf(res[1], df1 = res[2], df2 = res[3])
   value
0.742135
> out2 <- lm(y ~ x1 + x2) # x1 と x2 による回帰分析の実行
> summary(out2)
```

```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.52199 -0.17122 -0.06054  0.15732  0.64859

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.98181    0.09455  10.384 9.36e-14 ***
x1          -1.90178    0.12513 -15.198  < 2e-16 ***
x2          -0.06713    0.13310  -0.504    0.616
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.257 on 47 degrees of freedom
Multiple R-squared:  0.8327,        Adjusted R-squared:  0.8255
F-statistic: 116.9 on 2 and 47 DF,  p-value: < 2.2e-16
> ## F 値に対応する$p$値の計算
> res <- summary(out2)$fstatistic # F 値と自由度
> 1 - pf(res[1], df1 = res[2], df2 = res[3])
value
    0
```

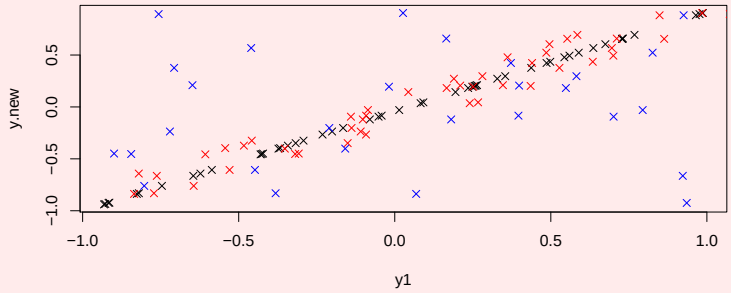
(fstatistic.r)

## 5.4. 予測

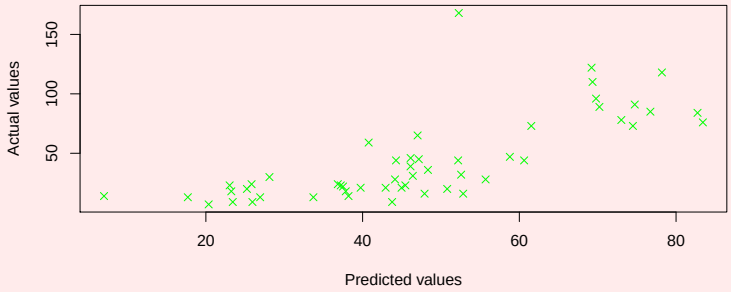
回帰分析の目的の 1 つは, 説明変数の新規データが与えられたときに, そのデータに対応する目的変数の値を予測することであるが, これは関数 `predict()` で実行できる.

```
> ### 人工データに対する例
> ### モデル: y = 1 - 2 * x1
> set.seed(123) # 乱数の初期値の固定
> x1 <- runif(50) # 説明変数 x1 の観測データ
> x2 <- runif(50) # (不要な) 説明変数 x2 の観測データ
> epsilon <- rnorm(50, sd = 0.3) # 誤差項
> y <- 1 - 2 * x1 + epsilon # 目的変数の観測データ
> dat <- data.frame(x1 = x1, x2 = x2, y = y) # 観測データからなるデータフレーム
> mod1 <- lm(y ~ x1, data = dat) # x1 による回帰分析の実行 (正しいモデル)
> summary(mod1)
Call:
lm(formula = y ~ x1, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-0.72257 -0.15628 -0.00511  0.16938  0.60658

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.08640    0.07911   13.73   <2e-16 ***
x1          -2.08166    0.13270  -15.69   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2734 on 48 degrees of freedom
Multiple R-squared:  0.8368,        Adjusted R-squared:  0.8334
F-statistic: 246.1 on 1 and 48 DF,  p-value: < 2.2e-16
```

```
> mod2 <- lm(y ~ x1 + x2, data = dat) # x1 と x2 による回帰分析の実行 (冗長なモデル)
> summary(mod2)

Call:
lm(formula = y ~ x1 + x2, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-0.72530 -0.16045 -0.00355  0.16389  0.60793

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.07781    0.10162  10.606 4.63e-14 ***
x1          -2.08311    0.13450 -15.488  < 2e-16 ***
x2           0.01956    0.14306   0.137    0.892
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2762 on 47 degrees of freedom
Multiple R-squared:  0.8368,        Adjusted R-squared:  0.8299
F-statistic: 120.5 on 2 and 47 DF,  p-value: < 2.2e-16

> mod3 <- lm(y ~ x2, data = dat) # x2 による回帰分析の実行 (誤ったモデル)
> summary(mod3)

Call:
lm(formula = y ~ x2, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-1.27851 -0.53343  0.02397  0.48671  1.31837

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.07731    0.19179   0.403    0.689
x2          -0.15422    0.34866  -0.442    0.660

Residual standard error: 0.6753 on 48 degrees of freedom
Multiple R-squared:  0.004059,        Adjusted R-squared:  -0.01669
F-statistic: 0.1956 on 1 and 48 DF,  p-value: 0.6602

> ## 新規データに対する予測
> new.dat <- data.frame(x1 = runif(50), x2 = runif(50,-10,10)) # 説明変数の新規データ
> y.new <- 1 - 2 * new.dat$x1 # 新規データに対する目的変数の真値
> y1 <- predict(mod1, newdata = new.dat) # mod1 による予測値
> y2 <- predict(mod2, newdata = new.dat) # mod2 による予測値
> y3 <- predict(mod3, newdata = new.dat) # mod3 による予測値
> ## 決定係数による評価
> cor(y.new, y1)^2

[1] 1

> cor(y.new, y2)^2

[1] 0.9578643

> cor(y.new, y3)^2

[1] 0.01301964

> ## 散布図による可視化
> plot(y.new ~ y1, pch = 4)
> lines(y.new ~ y2, type = "p", pch = 4, col = "red")
> lines(y.new ~ y3, type = "p", pch = 4, col = "blue")
```




```
> ### データセット airquality による例
> ### Ozone を目的変数とする回帰分析
> ### 5-7 月のデータを用いてモデルを構築し，8-9 月のデータを予測
> dat1 <- subset(airquality, Month %in% 5:7) # モデル推定用データ
> dat2 <- subset(airquality, Month %in% 8:9) # 予測用データ
> mod <- lm(Ozone ~ Solar.R + Temp, data = dat1) # モデルの推定
> summary(mod)



Call:
lm(formula = Ozone ~ Solar.R + Temp, data = dat1)

Residuals:
    Min      1Q  Median      3Q     Max
-35.689 -16.611   0.847   9.239  75.396

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -121.92537   22.42757  -5.436 1.23e-06 ***
Solar.R        0.03973    0.03176   1.251    0.216
Temp           2.03384    0.31523   6.452 2.77e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.17 on 56 degrees of freedom
  (33 observations deleted due to missingness)
Multiple R-squared:  0.5109,        Adjusted R-squared:  0.4934
F-statistic: 29.24 on 2 and 56 DF,  p-value: 2.014e-09



> myozone <- predict(mod, newdata = dat2) # 予測
> cor(myozone, dat2$Ozone, use = "c")^2 # 予測値に対する決定係数 (NA は除く)



[1] 0.544632



> plot(dat2$Ozone ~ myozone, pch = 4, col = "green",
+      xlab = "Predicted values", ylab = "Actual values") # 散布図
```




(predict.r)

## 5.5. 発展的なモデル

**5.5.1. 変数が多い場合のモデルの記述法.** データフレーム dat において, 1 つの変数 A を目的変数としそれ以外を説明変数とするようなモデルを推定したい場合は,

```
lm(A ~ ., data = dat)
```

を実行する. また, 変数 A を目的変数とし (変数 A および) 変数 B 以外を説明変数とするようなモデルを推定したい場合は,

```
lm(A ~ .  - B, data = dat)
```

を実行する. 特に, 定数項をモデルから除外したい場合は,

```
lm(A ~ .  - 1, data = dat)
```

とする.

```
> ### データセット airquality による例
> ## Ozone を目的変数, それ以外を説明変数とする
> mod1 <- lm(Ozone ~., data = airquality)
> summary(mod1)
Call:
lm(formula = Ozone ~ ., data = airquality)

Residuals:
    Min      1Q  Median      3Q     Max
-37.014 -12.284  -3.302   8.454  95.348

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -64.11632   23.48249  -2.730  0.00742 **
Solar.R       0.05027    0.02342   2.147  0.03411 *
Wind         -3.31844    0.64451  -5.149 1.23e-06 ***
Temp          1.89579    0.27389   6.922 3.66e-10 ***
Month        -3.03996    1.51346  -2.009  0.04714 *
Day           0.27388    0.22967   1.192  0.23576
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.86 on 105 degrees of freedom
  (42 observations deleted due to missingness)
Multiple R-squared:  0.6249,        Adjusted R-squared:  0.6071
F-statistic: 34.99 on 5 and 105 DF,  p-value: < 2.2e-16
```

```
> ## Ozone を目的変数, Ozone, Month, Day 以外を説明変数とする
> mod2 <- lm(Ozone ~. - Month - Day, data = airquality)
> summary(mod2)
Call:
lm(formula = Ozone ~ . - Month - Day, data = airquality)

Residuals:
    Min      1Q  Median      3Q     Max
-40.485 -14.219  -3.551  10.097  95.619

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -64.34208   23.05472  -2.791  0.00623 **
Solar.R       0.05982    0.02319   2.580  0.01124 *
Wind         -3.33359    0.65441  -5.094 1.52e-06 ***
Temp          1.65209    0.25353   6.516 2.42e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.18 on 107 degrees of freedom
  (42 observations deleted due to missingness)
Multiple R-squared:  0.6059,        Adjusted R-squared:  0.5948
F-statistic: 54.83 on 3 and 107 DF,  p-value: < 2.2e-16
```

(lm-model.r)

**5.5.2. 質的データの利用.** 身長や体重など, 数値として扱えるデータを**量的データ** *(quantitative data)* と呼ぶ. 他方, 性別や血液型など, 数値として扱えないデータ (分類を表すようなデータ) を**質的データ** *(qualitative data)* と呼ぶ. 質的データはそのままでは回帰分析の説明変数として利用できないが, 以下で説明する**数量化** *(quantification)* と呼ばれる操作を施すことで量的データと同じように扱うことができる.

まず, 例として, (性別, 身長) の観測データ $(x_1, y_1), \dots, (x_n, y_n)$ が与えられたときに, 性別による身長の違いを検証するために, 性別を説明変数, 身長を目的変数とする線形回帰分析を実行したいとする. 性別のデータ $x_1, \dots, x_n$ は数値でないためこのままでは説明変数として扱えないので, 次の**ダミー変数** *(dummy variable)* と呼ばれる変数を導入する:

$$z_i = \begin{cases} 1 & x_i = \text{男の場合}, \\ 0 & x_i = \text{女の場合}. \end{cases}$$

このとき, 数値データ $z_1, \dots, z_n$ を説明変数とする回帰モデルは以下のようになる:

$$y_i = \beta_0 + \beta_1 z_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & x_i = \text{男の場合}, \\ \beta_0 + \epsilon_i & x_i = \text{女の場合}. \end{cases}$$

従って, 係数 $\beta_0$ は女性の平均身長, $\beta_0 + \beta_1$ は男性の平均身長, $\beta_1$ は女性と男性の平均身長の違いを表すと解釈できる. このように, ダミー変数の導入によって質的データも回帰式の説明変数として取り扱うことができる.

上の例では 2 種類の分類 (男か女) をとる質的データの数量化を説明したが, 一般に $k$ 種類 ($k \geq 2$) の分類 $C_1, \dots, C_k$ をもつ質的データ $x_1, \dots, x_n$ を数量化するには, 以下のように定義される $k-1$ 個のダミー変数 $\boldsymbol{z}_j = (z_{1j}, \dots, z_{nj})^\top$ $(j = 1, \dots, k-1)$ を導入する必要がある:

$$z_{ij} = \begin{cases} 1 & x_i = C_j \text{ の場合}, \\ 0 & x_i \neq C_j \text{ の場合}. \end{cases}$$

これらのダミー変数を説明変数として回帰式を推定した場合, 定数項は $C_k$ に分類されるデータに対する目的変数の平均値と解釈でき, $\boldsymbol{z}_j$ の回帰係数は $C_j$ に分類されるデータと $C_k$ に分類されるデータの間の目的変数の平均値の差と解釈できる.

R には質的データを表すためのクラス factor が用意されている. たいていの場合, 数値データとして扱えないデータは必要に応じて factor クラスに変換されるため, ユーザー側で明示的にクラスを変換する必要はない. また, factor クラスの説明変数を関数 lm( ) のモデル式に加えると, 自動的にダミー変数へと変換されるため, ユーザー側で明示的にダミー変数へと変換する必要はない.

```
> ## データセット ToothGrowth による例
> ## モルモットにビタミン C/オレンジジュースを与えた場合の
> ## 歯の成長度を記録したデータ
> out <- lm(len ~ supp, data = ToothGrowth)
> # ビタミン C とオレンジジュースで成長度に違いは出るか?
> summary(out)
Call:
lm(formula = len ~ supp, data = ToothGrowth)

Residuals:
     Min       1Q   Median       3Q      Max
-12.7633  -5.7633   0.4367   5.5867  16.9367

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   20.663      1.366  15.127   <2e-16 ***
suppVC        -3.700      1.932  -1.915   0.0604 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.482 on 58 degrees of freedom
Multiple R-squared:  0.05948,        Adjusted R-squared:  0.04327
F-statistic: 3.668 on 1 and 58 DF,  p-value: 0.06039
> ### ビタミン C=1 のダミー変数の係数が負のため,
> ### オレンジジュースの方が効果が高いと予想される
> ### しかし, 差は 5%水準で有意でない
> out <- lm(len ~ supp + dose, data = ToothGrowth)
> # 投与量も説明変数として加える
> summary(out)
Call:
lm(formula = len ~ supp + dose, data = ToothGrowth)

Residuals:
   Min     1Q Median     3Q    Max
-6.600 -3.700  0.373  2.116  8.800

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.2725     1.2824   7.231 1.31e-09 ***
suppVC       -3.7000     1.0936  -3.383   0.0013 **
dose          9.7636     0.8768  11.135 6.31e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.236 on 57 degrees of freedom
Multiple R-squared:  0.7038,        Adjusted R-squared:  0.6934
F-statistic: 67.72 on 2 and 57 DF,  p-value: 8.716e-16
> ### ビタミン C/オレンジジュースの差が 1%水準で有意となる
> ### これは,「投与量が等しい」という条件下で効果を比較
> ### した場合, オレンジジュースの方が効果が高いことを
> ### 統計的にサポートしている
```

(dummy.r)

見かけ上量的データであるような変数を質的データとして扱いたい場合は, 以下の例のようにデータを明示的に factor クラスへと変換しておく必要がある.

```
> # 2016年の東京における降水の有無と気温の関係を調べる
> ## データの読み込み
> kikou <- read.csv("kikou2016.csv", fileEncoding = "sjis")
> ## 降水の有無を表す変数をデータフレームに追加
> kikou <- transform(kikou, 降水の有無 = as.factor(降水量 > 0))
> m1 <- lm(気温 ~ 降水の有無, data = kikou) # モデルの推定
> summary(m1) # 雨の日に気温が高いという結果
Call:
lm(formula = 気温 ~ 降水の有無, data = kikou)

Residuals:
     Min       1Q   Median       3Q      Max
-14.8609  -7.1126   0.6546   6.4218  16.0702

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     15.8298     0.4952  31.969   <2e-16 ***
降水の有無TRUE   1.8311     0.8373   2.187   0.0294 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.639 on 364 degrees of freedom
Multiple R-squared:  0.01297,        Adjusted R-squared:  0.01026
F-statistic: 4.783 on 1 and 364 DF,  p-value: 0.02939
> ## 東京では冬より夏の方が降水が多いことを考慮して, 月を表す
> ## ダミー変数を追加する
> kikou <- transform(kikou, 月 = as.factor(月))
> m2 <- lm(気温 ~ 降水の有無 + 月, data = kikou)
> summary(m2) # 雨の日の方が気温が低いという結果. ただし結果は5%水準で有意でない
Call:
lm(formula = 気温 ~ 降水の有無 + 月, data = kikou)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3101 -1.6307 -0.0326  1.7465 11.6514

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.1482     0.4663  13.184  < 2e-16 ***
降水の有無TRUE  -0.5234     0.2921  -1.792   0.0740 .
月2              1.2238     0.6699   1.827   0.0686 .
月3              4.1457     0.6590   6.291 9.35e-10 ***
月4              9.5253     0.6687  14.244  < 2e-16 ***
月5             14.1482     0.6584  21.489  < 2e-16 ***
月6             16.4843     0.6732  24.486  < 2e-16 ***
月7             19.3949     0.6597  29.397  < 2e-16 ***
月8             21.2212     0.6654  31.890  < 2e-16 ***
月9             18.4961     0.6701  27.603  < 2e-16 ***
月10            12.7432     0.6597  19.315  < 2e-16 ***
月11             5.4853     0.6687   8.203 4.43e-15 ***
月12             2.8514     0.6584   4.331 1.94e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.588 on 353 degrees of freedom
Multiple R-squared:  0.8901,        Adjusted R-squared:  0.8864
F-statistic: 238.4 on 12 and 353 DF,  p-value: < 2.2e-16
```

(dummy-kikou.r)

**5.5.3. 交互作用モデル・変数の非線形変換.** 冒頭で述べたように, モデル (5.3) において説明変数として $X_jX_k$(交差項と呼ばれる) や $\log X_j$ といったものを新たに加えることで, 目的変数と説明変数 $X_1, \ldots, X_p$ の非線形な関係をモデル化することができる. R においてはこのような回帰式の柔軟なモデル化を可能にするためのモデル式の記述法が実装されている. 以下の実行例を参考にしてほしい.

```
> ## MASS パッケージのデータセット Boston による例
> ## ボストン近郊の家の価格のデータ
> ## 変数 medv がボストン近郊の 506 の地域での家の価格の
> ## メディアンを表す
> library(MASS) # MASS パッケージのロード
> ### 様々な交互作用モデル
> # medv を rm(平均部屋数) で回帰
> mod1 <- lm(medv ~ rm, data = Boston)
> summary(mod1) # 部屋が多いほど価格は上昇
Call:
lm(formula = medv ~ rm, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-23.346  -2.547   0.090   2.986  39.433

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -34.671      2.650  -13.08   <2e-16 ***
rm             9.102      0.419   21.72   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.616 on 504 degrees of freedom
Multiple R-squared:  0.4835,        Adjusted R-squared:  0.4825
F-statistic: 471.8 on 1 and 504 DF,  p-value: < 2.2e-16
> # rm と dis(ボストンのオフィス街への距離) の交差項を追加
> mod2 <- lm(medv ~ rm + rm:dis, data = Boston)
> summary(mod2) # 距離が遠いほど部屋数が価格に与える影響は上昇
Call:
lm(formula = medv ~ rm + rm:dis, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-21.178  -2.896  -0.118   2.594  40.150

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -32.92685    2.65037 -12.423   <2e-16 ***
rm            8.49150    0.44154  19.231   <2e-16 ***
rm:dis        0.08668    0.02211   3.921    1e-04 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.524 on 503 degrees of freedom
Multiple R-squared:  0.4988,        Adjusted R-squared:  0.4969
F-statistic: 250.3 on 2 and 503 DF,  p-value: < 2.2e-16
> # rm,dis および rm と dis の交差項を説明変数とする
> mod3 <- lm(medv ~ rm * dis, data = Boston)
> summary(mod3) # 上述の効果に加え, 距離が遠いほど価格は下落
```

```
Call:
lm(formula = medv ~ rm * dis, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-18.423  -3.276   0.104   2.831  38.061

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.2533     4.8953  -3.116  0.00194 **
rm            5.7020     0.7851   7.263 1.45e-12 ***
dis          -5.7579     1.3500  -4.265 2.39e-05 ***
rm:dis        0.9855     0.2119   4.652 4.22e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.415 on 502 degrees of freedom
Multiple R-squared:  0.5164,        Adjusted R-squared:  0.5135
F-statistic: 178.7 on 3 and 502 DF,  p-value: < 2.2e-16
> # rm,dis,crim(犯罪率) とそれらの交差項をすべて説明変数とする
> mod4 <- lm(medv ~ (rm + dis + crim)^2, data = Boston)
> summary(mod4)
Call:
lm(formula = medv ~ (rm + dis + crim)^2, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-19.557  -2.965  -0.659   2.512  36.447

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -22.05731    5.31340  -4.151 3.89e-05 ***
rm            7.30399    0.84247   8.670  < 2e-16 ***
dis          -4.11598    1.32447  -3.108  0.00199 **
crim          1.92106    0.30224   6.356 4.67e-10 ***
rm:dis        0.65476    0.20725   3.159  0.00168 **
rm:crim      -0.22725    0.04331  -5.248 2.28e-07 ***
dis:crim     -0.52385    0.09094  -5.760 1.47e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.714 on 499 degrees of freedom
Multiple R-squared:  0.6186,        Adjusted R-squared:  0.614
F-statistic: 134.9 on 6 and 499 DF,  p-value: < 2.2e-16
> ## crim の係数が正のため，一見犯罪率が高い地域ほど家賃が高く見えるが，
> ## crim と他変数の交差項が負のため他の変数の大きさ次第
> ### 説明変数の非線形変換
> summary(lm(medv ~ dis, data = Boston))
Call:
lm(formula = medv ~ dis, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-15.016  -5.556  -1.865   2.288  30.377

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.3901     0.8174  22.499  < 2e-16 ***
dis           1.0916     0.1884   5.795 1.21e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 8.914 on 504 degrees of freedom
Multiple R-squared:  0.06246,        Adjusted R-squared:  0.0606
F-statistic: 33.58 on 1 and 504 DF,  p-value: 1.207e-08
> mod5 <- lm(medv ~ log(dis), data = Boston) # dis の対数で回帰
> summary(mod5) # 決定係数が (多少) 増加
Call:
lm(formula = medv ~ log(dis), data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-13.599  -5.485  -2.114   2.168  32.780

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  16.6131     0.9473  17.537   <2e-16 ***
log(dis)      4.9828     0.7261   6.862    2e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.804 on 504 degrees of freedom
Multiple R-squared:  0.08545,        Adjusted R-squared:  0.08363
F-statistic: 47.09 on 1 and 504 DF,  p-value: 1.997e-11
> mod6 <- lm(medv ~ dis + I(dis^2), data = Boston) # dis の 2 次式で回帰
> summary(mod6) # 距離が遠くなるにつれて, 距離の価格への影響は弱まる
Call:
lm(formula = medv ~ dis + I(dis^2), data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-13.205  -5.210  -2.114   2.344  32.987

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.74348    1.54254   8.261 1.28e-15 ***
dis          4.13317    0.73300   5.639 2.86e-08 ***
I(dis^2)    -0.31317    0.07302  -4.289 2.16e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.764 on 503 degrees of freedom
Multiple R-squared:  0.09554,        Adjusted R-squared:  0.09194
F-statistic: 26.57 on 2 and 503 DF,  p-value: 1.077e-11
```

(cross.r)

## 5.6. 参考文献


1. 二木昭人著「基礎講義 線形代数学」, 培風館 (1999 年).
2. G. James, D. Witten, T. Hastie, R. Tibshirani 著「An Introduction to Statistical Learning」, Springer (2013 年).
3. U. リゲス著, 石田基広訳「R の基礎とプログラミング技法」, 丸善出版 (2012 年).
4. 杉浦光夫著「解析入門 I」, 東京大学出版会 (1980 年).
5. 吉田朋広著「数理統計学」, 朝倉書店 (2006 年).

## 5.7. 参考: 正規方程式の性質

正規方程式の性質を調べるには, 直交射影の概念を導入しておくのが便利である. 一般に, $\mathbb{R}^n$ の部分線形空間 $U$ が与えられたとき, $\mathbb{R}^n$ は $U$ と $U$ の直交補空間 $U^\perp := \{\boldsymbol{a} \in \mathbb{R}^n : \text{すべての } \boldsymbol{b} \in U \text{ に対して } \boldsymbol{a}^\top \boldsymbol{b} = 0\}$ の直和に分解できる.[5] 特に, 各 $\boldsymbol{a} \in \mathbb{R}^n$ に対して, $\boldsymbol{a} - \boldsymbol{b} \in U^\perp$ となるような $\boldsymbol{b} \in U$ がただ一つ存在する. 従って, $P\boldsymbol{a} = \boldsymbol{b}$ として $\mathbb{R}^n$ から $U$ への写像 $P$ を定めることができるが, この写像 $P$ を $U$ への**直交射影** *(orthogonal projection)* と呼ぶ. 特に, $n \times m$ 行列 $A$ が与えられたとき, $L[A] := \{A\boldsymbol{b} : \boldsymbol{b} \in \mathbb{R}^m\}$ と定義し, $L[A]$ への直交射影を記号 $P_A$ で表すことにする.

**補題 5.1.** 任意の $n \times m$ 行列 $A$ に対して, $L[A^\top A] = L[A^\top]$ が成り立つ.

**証明**. $L[A^\top A] \subset L[A^\top]$ は明らかだから, $L[A^\top A] \supset L[A^\top]$ を示す. $\boldsymbol{a} \in L[A^\top]$ とすると, ある $\boldsymbol{b} \in \mathbb{R}^n$ が存在して $\boldsymbol{a} = A^\top \boldsymbol{b}$ と書ける. このとき $A^\top(\boldsymbol{b} - P_A\boldsymbol{b}) = 0$ が成り立つ. 実際, $\boldsymbol{v} := \boldsymbol{b} - P_A\boldsymbol{b} \in L[A]^\perp$ であるから,

$$0 = (A(A^\top \boldsymbol{v}))^\top \boldsymbol{v} = \boldsymbol{v}^\top AA^\top \boldsymbol{v} = (A^\top \boldsymbol{v})^\top A^\top \boldsymbol{v}$$

が成り立つ. 上式右辺は $A^\top \boldsymbol{v}$ の各成分の二乗和に等しいから, これは $A^\top \boldsymbol{v} = \boldsymbol{0}$ を意味する. 従って, $\boldsymbol{a} = A^\top(P_A\boldsymbol{b})$ が成り立つ. $P_A\boldsymbol{b} \in L[A]$ だから, これは $\boldsymbol{a} \in L[A^\top A]$ を意味する. □

補題 5.1 から次の結果が直ちに従う.

**定理 5.2.** 正規方程式は常に解をもつ.

正規方程式の解はすべて最小二乗推定量となることが証明できる:

**定理 5.3.** $\hat{\boldsymbol{\beta}}$ を正規方程式の解とすると, $P_{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ が成り立つ. 更に, $\hat{\boldsymbol{\beta}}$ は最小二乗推定量である. すなわち, $S(\hat{\boldsymbol{\beta}})$ は関数 $S$ の最小値である.

**証明**. 任意の $\boldsymbol{b} \in \mathbb{R}^{p+1}$ に対して,

$$(\boldsymbol{X}\boldsymbol{b})^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{b}^\top \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{b}^\top(\boldsymbol{X}^\top \boldsymbol{y} - \boldsymbol{X}^\top \boldsymbol{X}\hat{\boldsymbol{\beta}}) = 0$$

が成り立つから, $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \in L[\boldsymbol{X}]^\perp$. これは $P_{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ を意味する. 次に, $\boldsymbol{\beta}$ を $\mathbb{R}^{p+1}$ の任意の元とすると,

$$\begin{aligned} S(\boldsymbol{\beta}) &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + 2(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}) \\ &\quad + (\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}) \end{aligned}$$

が成り立つ. ここで, $\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta} \in L[X]$ だから, 前半の結果より

$$(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}) = 0$$

である. また $(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}) \geq 0$ である. 以上より,

$$S(\boldsymbol{\beta}) \geq (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = S(\hat{\boldsymbol{\beta}})$$

が成り立つ. 従って $S(\hat{\boldsymbol{\beta}})$ は関数 $S$ の最小値である. □

## 5.8. 参考: (5.8) 式の導出

まず記号を導入する.

$$\overline{\boldsymbol{x}^2} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\top, \qquad \overline{\boldsymbol{x}y} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i y_i$$

[5] 例えば, 参考文献 1. の定理 5.3.7 参照.

とおく. このとき,

$$\boldsymbol{X}^\top \boldsymbol{X} = \begin{pmatrix} 1 & \cdots & 1 \\ \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_n \end{pmatrix} \begin{pmatrix} 1 & \boldsymbol{x}_1^\top \\ \vdots & \vdots \\ 1 & \boldsymbol{x}_n^\top \end{pmatrix} = n \begin{pmatrix} 1 & \bar{\boldsymbol{x}}^\top \\ \bar{\boldsymbol{x}} & \overline{\boldsymbol{x}^2} \end{pmatrix},$$

$$\boldsymbol{X}^\top \boldsymbol{y} = \begin{pmatrix} 1 & \cdots & 1 \\ \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_n \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = n \begin{pmatrix} \bar{y} \\ \overline{\boldsymbol{x}y} \end{pmatrix}$$

が成り立つ. ここで, Gram 行列 $\boldsymbol{X}^\top \boldsymbol{X}$ の逆行列を計算するために, 次のブロック行列に対する逆行列の計算公式を用いる:

**定理 5.4.** $A$ を $q$ 次正方行列, $B$ を $q \times r$ 行列, $C$ を $r \times q$ 行列, $D$ を $q$ 次正方行列とし, $(q+r)$ 次正方行列

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

を考える. $A$ は正則であると仮定し, $G = D - CA^{-1}B$ とおく. このとき $\det M = \det A \det G$ が成り立つ. さらに, $M$ が正則ならば, $G$ も正則であり,

$$M^{-1} = \begin{pmatrix} A^{-1} + A^{-1}BG^{-1}CA^{-1} & -A^{-1}BG^{-1} \\ -G^{-1}CA^{-1} & G^{-1} \end{pmatrix} \tag{5.14}$$

が成り立つ.

**証明**. 等式

$$\begin{pmatrix} E_p & O \\ -CA^{-1} & E_q \end{pmatrix} \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} E_p & -A^{-1}B \\ O & E_q \end{pmatrix} = \begin{pmatrix} A & O \\ O & G \end{pmatrix}$$

が成り立つので, 両辺の行列式をとって $\det M = \det A \det G$ を得る. さらに $M$ が正則ならば, $\det G = \det M / \det A \neq 0$ となるので $G$ も正則であり, 上の等式から (5.14) を得る. □

定理 5.4 より, $V_x = \overline{\boldsymbol{x}^2} - \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\top$ とおくと, $V_x$ は正則であり,

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \frac{1}{n} \begin{pmatrix} 1 + \bar{\boldsymbol{x}}^\top V_x^{-1} \bar{\boldsymbol{x}} & -\bar{\boldsymbol{x}}^\top V_x^{-1} \\ -V_x^{-1}\bar{\boldsymbol{x}} & V_x^{-1} \end{pmatrix}$$

が成り立つ. 従って,

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \frac{1}{n} \begin{pmatrix} 1 + \bar{\boldsymbol{x}}^\top V_x^{-1} \bar{\boldsymbol{x}} & -\bar{\boldsymbol{x}}^\top V_x^{-1} \\ -V_x^{-1}\bar{\boldsymbol{x}} & V_x^{-1} \end{pmatrix} n \begin{pmatrix} \bar{y} \\ \overline{\boldsymbol{x}y} \end{pmatrix} \\ &= \begin{pmatrix} \bar{y} + \bar{\boldsymbol{x}}^\top V_x^{-1} \bar{\boldsymbol{x}} \bar{y} - \bar{\boldsymbol{x}}^\top V_x^{-1} \overline{\boldsymbol{x}y} \\ -V_x^{-1}\bar{\boldsymbol{x}}\bar{y} + V_x^{-1}\overline{\boldsymbol{x}y} \end{pmatrix} \\ &= \begin{pmatrix} \bar{y} - \bar{\boldsymbol{x}}^\top V_x^{-1} V_{xy} \\ V_x^{-1} V_{xy} \end{pmatrix} \end{aligned}$$

が成り立つ. ただし, $V_{xy} = \overline{\boldsymbol{x}y} - \bar{\boldsymbol{x}}\bar{y}$ とおいた. 特に,

$$(1 \ \bar{\boldsymbol{x}}^\top)\hat{\boldsymbol{\beta}} = \bar{y} - \bar{\boldsymbol{x}}^\top V_x^{-1} V_{xy} + \bar{\boldsymbol{x}}^\top V_x^{-1} V_{xy} = \bar{y}$$

が成り立つ.

$\bar{\boldsymbol{x}} = (\bar{x}_a)_{a=1,...,p}$ とするとき，$V_x = (s_{ab})_{a,b=1,...,p}$, $V_{xy} = (s_{ya})_{a=1,...,p}$ は次のように表される：

$$s_{ab} = \frac{1}{n}\sum_{i=1}^{n}(x_{ia} - \bar{x}_a)(x_{ib} - \bar{x}_b) \qquad (a, b = 1, ..., p)$$

$$s_{ya} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(x_{ia} - \bar{x}_a) \qquad (a = 1, ..., p)$$

$\hat{\beta}$ を $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)^\top$ と表し，$\tilde{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_p)^\top$ とする．このとき，

$$\hat{\beta}_a = \sum_{b=1}^{p} s^{ab} s_{yb} \qquad (a = 1, ..., p)$$

$$\hat{\beta}_0 = \bar{y} - \sum_{a=1}^{p} \bar{x}_a \hat{\beta}_a$$

である．ここで，$(s^{ab})_{a,b=1,...,p} = \big((s_{ab})_{a,b=1,...,p}\big)^{-1}$.