クレジット:
UTokyo Online Education 統計データ解析 I 2017 小池祐太

東京大学 THE UNIVERSITY OF TOKYO

CC BY NC ND

統計データ解析Ⅰ (平成２９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 (首都大学東京, 東京大学)

第 11 章 回帰分析

回帰分析とは, ある変量やデータを別の変量・データを用いて説明・予測するためのモデル (**回帰モデル**) を構築することを目的とする分析法である. 回帰分析においては, 説明される側の変量・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, 説明する側の変量・データは**説明変数・独立変数・共変量**などと呼ばれる. 目的変数・説明変数ともに複数個あってもよいが, 目的変数については変数ごとにそれぞれ回帰モデルを構築すればよいので, 通常は 1 つの場合を考える. 説明変数については, 1 つの場合を**単回帰**, 2 つの場合を**重回帰**として区別することが多い. この講義では単回帰のみ扱う (重回帰は「統計データ解析 II」の講義で取り扱う).

11.1. 回帰モデル

以下では, 説明変数を X, 目的変数を Y で表すことにする. Y を X で説明するための関係式は, 一般にはある関数 $f(x)$ を使って,

$$Y = f(X) \tag{11.1}$$

と書ける. しかし, このモデルでは一般的すぎて分析に不向きのため, 通常は f の関数形に何らかの制約を課す. 最も広く利用されているのは, $f(x)$ として一次関数のみ考えるというものである. すなわち, ある定数 α, β が存在して,

$$f(x) = \alpha + \beta x$$

と書ける場合のみを分析対象とする. この場合 (11.1) 式は

$$Y = \alpha + \beta X \tag{11.2}$$

となる. モデル (11.2) を分析対象とする回帰分析を**線形回帰**と呼び, f としてより一般的な関数形を許す回帰分析を**非線形回帰**と呼ぶ. この講義では線形回帰分析を取り扱う. モデル (11.2) において, α は**定数項**, β は X の**回帰係数**と呼ばれる.

なお, 非線形な関係であっても, データに適切な変数変換 (二乗する, 対数をとるなど) を施すことで, 線形な関係に変換可能な場合や, 線形な関係で近似できる場合がよくあることに注意しておく.

11.2. 回帰係数の点推定

モデル (11.2) は未知のパラメーター α, β を含むから, これらを観測データから推定してやる必要がある. この節ではこの問題について議論する.

11.2.1. 観測データ. n 個の個体について説明変数と目的変数の組 (X, Y) を観測して得られたデータ $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ が与えられているとする. 実際のデータには観測誤差のようなランダムな変動が含まれていると考えられるから, モデル (11.2) が観測データに対してもそのまま成立するとは考えづらい. そのため, データのランダムな変動を表す項を $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ として, 以下の形の確率モデルを分析することを考える:

$$Y_i = \alpha + \beta X_i + \epsilon_i, \qquad i = 1, \ldots, n. \tag{11.3}$$

$\epsilon_1, \ldots, \epsilon_n$ は**誤差項**もしくは**撹乱項**と呼ばれる. 以下の分析では次の仮定をおく:

(A) データ $X_1, \ldots, X_n$ は確率変数ではなく確定値であり, 一定値ではない. すなわち, $X_1 = \cdots = X_n$ ではない.

(B) 誤差項 $\epsilon_1, \ldots, \epsilon_n$ は独立同分布な確率変数列であり, 平均 0, 分散 σ^2 である.

11.2.2. 最小二乗法. 回帰モデルの推定には通常**最小二乗法**が用いられる. 最小二乗法の考え方は以下の通りである. パラメーターの組 (α, β) を 1 つ決めたとき, 回帰モデルでは説明できない目的変数の変動は,

$$e_i(\alpha, \beta) = y_i - (\alpha + \beta X_i), \qquad i = 1, \ldots, n$$

で与えられる. これらの変動 $e_1(\alpha, \beta), \ldots, e_n(\alpha, \beta)$ 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, $e_1(\alpha, \beta), \ldots, e_n(\alpha, \beta)$ の平方和

$$S(\alpha, \beta) := \sum_{i=1}^{n} e_i(\alpha, \beta)^2 = \sum_{i=1}^{n} \{Y_i - (\alpha + \beta X_i)\}^2$$

を最小にするようにパラメーター (α, β) を決定する. $S(\alpha, \beta)$ は**残差平方和**と呼ばれ, $S(\alpha, \beta)$ を最小にするパラメーターの組 (α, β) は**最小二乗推定量**と呼ばれる. 最小二乗推定量はしばしば記号 $(\hat{\alpha}, \hat{\beta})$ で表される.

最小二乗推定量は具体的に求めることができる. 実際, 最小二乗推定量はもし存在すれば次の連立方程式の解とならなければならない:

$$\begin{cases} \frac{\partial S}{\partial \alpha} = -2 \sum_{i=1}^{n} \{Y_i - (\alpha + \beta X_i)\} = 0, \\ \frac{\partial S}{\partial \beta} = -2 \sum_{i=1}^{n} \{Y_i - (\alpha + \beta X_i)\} X_i = 0. \end{cases} \tag{11.4}$$

この式を整理して, α, β に関する連立一次方程式

$$\begin{cases} n\alpha + (\sum_i X_i)\beta = \sum_i Y_i, \\ (\sum_i X_i)\alpha + (\sum X_i^2)\beta = \sum_i X_i Y_i \end{cases}$$

を得る (**正規方程式**と呼ばれる). この連立一次方程式を解くと以下の解を得る:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}, \quad \hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X} \tag{11.5}$$

ただし,

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i, \qquad \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

(11.5) 式で与えられる $(\hat{\alpha}, \hat{\beta})$ が実際に $S(\alpha, \beta)$ を最小化していることは, 具体的な計算によって確認することができる (演習問題).

11.2.3. 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 (11.3) において, 説明変数 X および目的変数 Y の観測データに対応するベクトルがそれぞれ `x` および `y` で与えられているとする. このとき, モデル (11.3) の回帰係数の推定は,

`lm(y ~ x)`

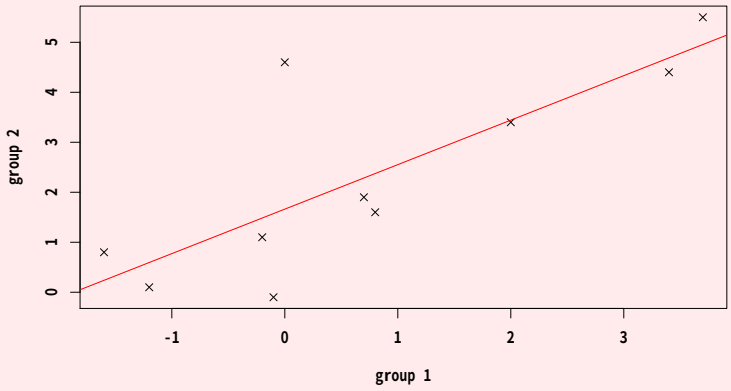
で実行できる. また, 実際のデータを使って解析する際は, データセットの一部の変数を目的変数および説明変数として回帰分析をすることが多い. そのような場合, データセットに対応するデータフレームを `dat` とすれば, 以下のコマンドで回帰係数の推定を実行できる:

`lm(`Y の変数名 ~ X の変数名`, data = dat)`

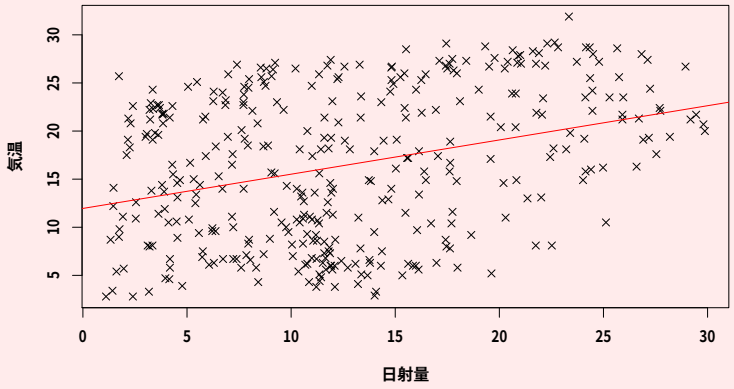
ここで, `dat` は列が各変数に対応するような形式になっている必要がある.

```
> ## データセット sleep による例
> x <- subset(sleep, group == 1, extra, drop = TRUE)
> y <- subset(sleep, group == 2, extra, drop = TRUE)
> (out <- lm(y ~ x)) # 線形回帰分析の実行
```

```
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     1.6625       0.8899

> coef(out) # 推定されたパラメーター値

(Intercept)           x
  1.6625378   0.8899497

> # 最少二乗推定量の計算公式との確認
> (beta.hat <- cov(x, y)/var(x))

[1] 0.8899497

> (alpha.hat <- mean(y) - beta.hat * mean(x))

[1] 1.662538

> # データの散布図と回帰直線の図示
> plot(x, y, xlab = "group 1", ylab = "group 2", pch = 4)
> abline(reg = out, col = "red")
```




```
> ## 気候データによる例
> kikou <- read.csv("kikou2016.csv", fileEncoding = "sjis")
> (out <- lm(気温 ~ 日射量, data = kikou)) # 気温を日射量で説明

Call:
lm(formula = 気温 ~ 日射量, data = kikou)

Coefficients:
(Intercept)       日射量
    11.9571       0.3559

> # データの散布図と回帰直線の図示
> plot(気温 ~ 日射量, data = kikou, pch = 4)
> abline(reg = out, col = "red")
> confint(out)

                2.5 %     97.5 %
(Intercept) 10.4372617 13.4769302
日射量       0.2514435  0.4602951
```



(slr.r)

11.2.4. Gauss-Markov の定理. 最小二乗推定量は以下の性質をもつことが確認できる:

(1) $\hat{\alpha}, \hat{\beta}$ は不偏推定量である:

$$E(\hat{\alpha}) = \alpha, \qquad E(\hat{\beta}) = \beta.$$

(2) $\hat{\alpha}, \hat{\beta}$ は $Y_1, \ldots, Y_n$ の線形和で表される. すなわち, $(X_1, \ldots, X_n$ に依存するかもしれない) 定数 $a_1, \ldots, a_n, b_1, \ldots, b_n$ が存在して,

$$\hat{\alpha} = \sum_{i=1}^{n} a_i Y_i, \qquad \hat{\beta} = \sum_{i=1}^{n} b_i Y_i$$

が成り立つ.

(3) $\hat{\alpha}, \hat{\beta}$ の分散は次式で与えられる:

$$\mathrm{Var}(\hat{\alpha}) = \frac{\sigma^2 \sum_i X_i^2}{n \sum_i (X_i - \bar{X})^2}, \qquad \mathrm{Var}(\hat{\beta}) = \frac{\sigma^2}{n \sum_i (X_i - \bar{X})^2}.$$

実は, 最小二乗推定量は上の性質 (1) および (2) を満たすもののうち分散が最小のものであるということが知られている. すなわち, 次の定理が成り立つ:

定理 11.1 (Gauss-Markov の定理). α, β の推定量 $\tilde{\alpha}, \tilde{\beta}$ が以下の 2 条件を満たすとする:

(1) $\tilde{\alpha}, \tilde{\beta}$ は不偏推定量である:

(2) $\tilde{\alpha}, \tilde{\beta}$ は $Y_1, \ldots, Y_n$ の線形和で表される. すなわち, $(X_1, \ldots, X_n$ に依存するかもしれない) 定数 $a_1, \ldots, a_n, b_1, \ldots, b_n$ が存在して,

$$\tilde{\alpha} = \sum_{i=1}^{n} a_i Y_i, \qquad \tilde{\beta} = \sum_{i=1}^{n} b_i Y_i \tag{11.6}$$

が成り立つ.

このとき, 次の不等式が成り立つ:

$$\mathrm{Var}(\tilde{\alpha}) \geq \frac{\sigma^2 \sum_i X_i^2}{n \sum_i (X_i - \bar{X})^2}, \qquad \mathrm{Var}(\tilde{\beta}) \geq \frac{\sigma^2}{n \sum_i (X_i - \bar{X})^2}.$$

演習 11.1. 最小二乗推定量について調べてみよう.

(1) 正規方程式の解が (11.5) 式で与えられることを実際に確認してみよ.

(2) (11.5) 式で与えられる $(\hat{\alpha}, \hat{\beta})$ が実際に $S(\alpha, \beta)$ を最小化していることを確認してみよ.

11.3. 回帰係数の区間推定

この節ではパラメーター α, β の区間推定について議論する. そのために, 誤差項に関して以下の仮定を追加する:

(C) ϵ_i たちは正規分布に従う.

上の仮定と命題 8.1 より, $\hat{\alpha}, \hat{\beta}$ もそれぞれ正規分布に従うことがわかり, 平均と分散は

$$E(\hat{\alpha}) = \alpha, \qquad E(\hat{\beta}) = \beta,$$
$$\mathrm{Var}(\hat{\alpha}) = \frac{\sigma^2 \sum_i X_i^2}{n \sum_i (X_i - \bar{X})^2}, \qquad \mathrm{Var}(\hat{\beta}) = \frac{\sigma^2}{n \sum_i (X_i - \bar{X})^2}$$

で与えられる. 従って, もし σ^2 が既知であれば, 8.4.1 節と同様の議論によって α, β の信頼区間をそれぞれ構成できる. 一般には σ^2 は既知でないため, データから推定する必要がある. σ^2 が ϵ_i たちに共通の分散であったことと, ϵ_i たちの平均は 0 であること, および

$$\epsilon_i = Y_i - (\alpha + \beta X_i) \qquad (i = 1, \ldots, n)$$

と書き直せることに注意すれば,

$$\hat{\epsilon}_i = Y_i - (\hat{\alpha} + \hat{\beta} X_i), \qquad i = 1, \ldots, n$$

と定義して, $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$ の二乗の平均 $\frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2$ を σ^2 の推定量として考えるのが自然なように思える. $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$ は**残差**と呼ばれ, 以下を満たす ((11.4) 式より従う):

$$\sum_i \hat{\epsilon}_i = 0, \qquad \sum_i \hat{\epsilon}_i X_i = 0. \tag{11.7}$$

実際には

$$E[\hat{\epsilon}_i^2] = \frac{n-2}{n}\sigma^2 \qquad (i = 1, \ldots, n)$$

となることがわかるため, $\frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2$ を σ^2 の不偏推定量となるように補正した以下の推定量が利用される:

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

従って, $\hat{\alpha}, \hat{\beta}$ の分散の推定量として

$$s.e.(\hat{\alpha})^2 := \frac{\hat{\sigma}^2 \sum_i X_i^2}{n \sum_i (X_i - \bar{X})^2}, \qquad s.e.(\hat{\beta})^2 := \frac{\hat{\sigma}^2}{\sum_i (X_i - \bar{X})^2}$$

を考えるのが自然である. これらの推定量の平方根をとって得られる $\hat{\alpha}, \hat{\beta}$ の標準偏差の推定量 $s.e.(\hat{\alpha}), s.e.(\hat{\beta})$ をそれぞれ $\hat{\alpha}, \hat{\beta}$ の**標準誤差**と呼ぶ.

以上の準備の下, α, β の信頼区間を構成する方法を説明する. まず, $(n-2)s.e.(\hat{\alpha})^2 / \mathrm{Var}[\hat{\alpha}]$ は $\hat{\alpha}$ と独立で, かつ自由度 $n-2$ の χ^2 分布に従うことが知られている. 従って,

$$\frac{\hat{\alpha} - \alpha}{s.e.(\hat{\alpha})} \left(= \frac{(\hat{\alpha} - \alpha)/\sqrt{\mathrm{Var}[\hat{\alpha}]}}{\sqrt{\frac{(n-2)s.e.(\hat{\alpha})^2/\mathrm{Var}[\hat{\alpha}]}{n-2}}} \right)$$

は自由度 $n-2$ の t 分布に従うことがわかる (6.3.5 節参照). 以上より, $\gamma \in (0,1)$ に対して,

$$[\hat{\alpha} - t_{1-\gamma/2}(n-2) \cdot s.e.(\hat{\alpha}), \hat{\alpha} + t_{1-\gamma/2}(n-2) \cdot s.e.(\hat{\alpha})]$$

は α の $100(1-\gamma)\%$信頼区間を与えることがわかる ($t_{1-\gamma/2}(n-2)$ は自由度 $n-2$ の t 分布の $100(1-\gamma/2)\%$分位点を表す).

β の信頼区間の構成も同様の議論でできる. $(n-2)s.e.(\hat{\beta})^2/\operatorname{Var}[\hat{\beta}]$ は $\hat{\beta}$ と独立で, かつ自由度 $n-2$ の χ^2 分布に従うことが知られているので,

$$\frac{\hat{\beta}-\beta}{s.e.(\hat{\beta})}\left(=\frac{(\hat{\beta}-\beta)/\sqrt{\operatorname{Var}[\hat{\beta}]}}{\sqrt{\frac{(n-2)s.e.(\hat{\beta})^2/\operatorname{Var}[\hat{\beta}]}{n-2}}}\right)$$

は自由度 $n-2$ の t 分布に従うことがわかる (6.3.5 節参照). 以上より, $\gamma \in (0,1)$ に対して,

$$[\hat{\beta}-t_{1-\gamma/2}(n-2)\cdot s.e.(\hat{\beta}), \hat{\beta}+t_{1-\gamma/2}(n-2)\cdot s.e.(\hat{\beta})]$$

は β の $100(1-\gamma)\%$信頼区間を与えることがわかる.

```
> ## データセット sleep による例
> x <- subset(sleep, group == 1, extra, drop = TRUE)
> y <- subset(sleep, group == 2, extra, drop = TRUE)
> (out <- lm(y ~ x)) # 線形回帰分析の実行
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     1.6625       0.8899
> confint(out) # 95%信頼区間
                2.5 %   97.5 %
(Intercept) 0.6358501 2.689225
x           0.3366380 1.443261
> confint(out, level = 0.99) # 99%信頼区間
                 0.5 %   99.5 %
(Intercept) 0.16863991 3.156436
x           0.08484491 1.695054
```

(slr-ci.r)

演習 11.2. 最小二乗推定量の性質について調べてみよう.

(1) (11.7) 式が成立することを実際に確認してみよ.
(2) $(\hat{\alpha}-\alpha)/s.e.(\hat{\alpha})$, $(\hat{\beta}-\beta)/s.e.(\hat{\beta})$ がそれぞれ自由度 $n-2$ の t 分布に従うことをシミュレーションで確認してみよ.

11.4. 回帰係数の有意性の検定

回帰分析において, 説明変数 X が目的変数 Y を説明・予測するのに本当に役立っているか検証することは重要である. 線形回帰モデル (11.3) においてこれを検証するには, 検定問題

$$H_0 : \beta = 0 \qquad \text{vs} \qquad H_1 : \beta \neq 0 \tag{11.8}$$

を考えればよい. この検定は β の**有意性の検定**と呼ばれ, 帰無仮説 H_0 が有意水準 γ で棄却されるとき, β は有意水準 γ で**有意である**といわれる. この節では, 前節に引き続き (C) を仮定した下で, 上の検定を実行する方法を説明する.

前節で述べたことから, 帰無仮説 H_0 の下で, 統計量

$$t = \frac{\hat{\beta}}{s.e.(\hat{\beta})}$$

は自由度 $n-1$ の t 分布に従う. 一方, 対立仮説 H_1 が正しければ, $\hat{\beta}$ は 0 でない値 β に近い値を取ることが期待されるから, $|t|$ は 0 から離れた値を取ることが予想され

る. 以上より, 有意水準を $\gamma \in (0,1)$ とする場合, 検定 (11.8) は次の手順で実行できる: データから検定統計量 t の値を計算し,

$$|t| > t_{1-\gamma/2}(n-2)$$

であった場合には帰無仮説を棄却する. もしくは, 検定の p 値

$$2\int_{|t|}^{\infty} f(x)dx \tag{11.9}$$

が γ 未満の場合に帰無仮説を棄却するとしても同等である. ここに, $f(x)$ は自由度 $n-2$ の t 分布の確率密度関数を表す. なお, 検定統計量の値 t を $\hat{\beta}$ の t **値**と呼び, 検定の p 値 (11.9) を $\hat{\beta}$ の p **値**と呼ぶ.

定数項 α についても同様の方法で検定を実行することが可能であるが, 詳細は省略する.

```
> ## 気候データによる例
> kikou <- read.csv("kikou2016.csv", fileEncoding = "sjis")
> (mod1 <- lm(気温 ~ 日射量, data = kikou)) # 気温を日射量で説明
Call:
lm(formula = 気温 ~ 日射量, data = kikou)

Coefficients:
(Intercept)       日射量
    11.9571       0.3559
> summary(mod1) # パラメーター推定値・標準誤差・t 値・p 値などを表示
Call:
lm(formula = 気温 ~ 日射量, data = kikou)

Residuals:
     Min       1Q   Median       3Q      Max
-14.0428  -6.4502  -0.2706   7.2320  13.1273

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.9571     0.7729  15.471  < 2e-16 ***
日射量        0.3559     0.0531   6.702 7.86e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.254 on 364 degrees of freedom
Multiple R-squared:  0.1098,        Adjusted R-squared:  0.1074
F-statistic: 44.91 on 1 and 364 DF,  p-value: 7.863e-11
> ### 日射量の回帰係数の p 値は非常に小さいので, 日射量は気温の説明に
> ### 有用であると結論できそう
> (mod2 <- lm(気温 ~ 降水量, data = kikou)) # 気温を降水量で説明
Call:
lm(formula = 気温 ~ 降水量, data = kikou)

Coefficients:
(Intercept)       降水量
   16.23425      0.04855
> summary(mod2) # パラメーター推定値・標準誤差・t 値・p 値などを表示
Call:
lm(formula = 気温 ~ 降水量, data = kikou)

Residuals:
     Min       1Q   Median       3Q      Max
-16.6869  -6.9685   0.1832   6.6494  15.6658
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.23425    0.42543  38.159   <2e-16 ***
降水量       0.04855    0.02956   1.642    0.101
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.661 on 364 degrees of freedom
Multiple R-squared:  0.007354,        Adjusted R-squared:  0.004626
F-statistic: 2.697 on 1 and 364 DF,  p-value: 0.1014
> ### 降水量の回帰係数の p 値は 0.101 なので，有意であるとはいえない
```

(slr-test.r)

演習 11.3. 回帰係数の有意性の検定について, そのサイズおよび検出力をシミュレーションによって計算してみよ.

11.5. 決定係数

前節で議論した回帰係数の有意性の検定では, 説明変数 X が目的変数 Y の説明・予測に役立つかどうかを検証することはできたが, 実際に X が Y の変動をどの程度説明できているかということについては何も述べていない. このことを評価する指標として**決定係数**がある (**寄与率**と呼ばれることもある). 決定係数は次式で定義される:

$$R^2 := \frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2}.$$

ただし,

$$\hat{Y}_i := \hat{\alpha} + \hat{\beta} X_i \qquad (i = 1, \ldots, n)$$

であり, $\hat{Y}_1, \ldots, \hat{Y}_n$ は**あてはめ値**または**予測値**と呼ばれる. $\hat{\epsilon}_i = Y_i - \hat{Y}_i$ $(i = 1, \ldots, n)$ が成り立つことに注意すると, (11.7) 式より

$$\frac{1}{n} \sum_{i=1}^{n} \hat{Y}_i = \bar{Y}$$

が成り立つ. この式より, R^2 の分子・分母はそれぞれあてはめ値・目的変数の (標本平均まわりでの) 変動に対応しており, 従って回帰モデルが目的変数の変動を何割程度説明できているかを測る評価指標であると解釈できる (従って大きいほど説明力が高いと解釈される).

R^2 は以下のように書き直すことも可能である:

$$R^2 = \left\{ \frac{\sum_i (\hat{Y}_i - \bar{Y})(Y_i - \bar{Y})}{\sqrt{\sum_i (Y_i - \bar{Y})^2} \cdot \sqrt{\sum_i (\hat{Y}_i - \bar{Y})^2}} \right\}^2. \tag{11.10}$$

すなわち, R^2 は目的変数の観測データとあてはめ値の相関の二乗に等しく, 回帰モデルによるあてはめが目的変数にどの程度連動しているかを測る指標であるとも解釈できる. さらに, 等式 $\hat{Y}_i - \bar{Y} = \hat{\beta}(X_i - \bar{X})$ を使うことで, 上の式は

$$R^2 = \left\{ \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (Y_i - \bar{Y})^2} \cdot \sqrt{\sum_i (X_i - \bar{X})^2}} \right\}^2$$

とも書ける. すなわち, R^2 は説明変数と目的変数の観測データの間の相関の二乗にも等しくなっている.

(11.7) 式を使うことで, R^2 のさらなる別表示として次式を得る:

$$R^2 = 1 - \frac{\frac{1}{n}\sum_i \hat{\epsilon}_i^2}{\frac{1}{n}\sum_i (Y_i - \bar{Y})^2}. \tag{11.11}$$

この式において, 右辺第 2 項の分子, 分母はそれぞれ確率変数 ϵ_i, Y_i の分散の標本分散による推定値ともみなせる. この観点から考えると, 推定量としては不偏なものを使った方がよいと考えられる. そこで, 標本分散を対応する不偏推定量で置き換えた以下のような評価指標が考えられる:

$$\bar{R}^2 = 1 - \frac{\frac{1}{n-2}\sum_i \hat{\epsilon}_i^2}{\frac{1}{n-1}\sum_i (Y_i - \bar{Y})^2}.$$

これを**自由度調整済み決定係数**と呼ぶ (**自由度調整済み寄与率**と呼ばれることもある).

```
> ## データセット sleep による例
> x <- subset(sleep, group == 1, extra, drop = TRUE)
> y <- subset(sleep, group == 2, extra, drop = TRUE)
> mod <- lm(y ~ x) # 線形回帰分析の実行
> (out <- summary(mod)) # 下から二行目に決定係数と自由度調整済み決定係数が表示される
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6735 -0.4673 -0.3365  0.3979  2.9375

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.6625     0.4452   3.734  0.00575 **
x             0.8899     0.2399   3.709  0.00596 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.288 on 8 degrees of freedom
Multiple R-squared:  0.6323,        Adjusted R-squared:  0.5863
F-statistic: 13.76 on 1 and 8 DF,  p-value: 0.005965
> coef(out) # パラメーター推定値・標準誤差・t 値・p 値
             Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 1.6625378  0.4452237 3.734163 0.005753296
x           0.8899497  0.2399439 3.708990 0.005964996
> out$r.squared # 決定係数
[1] 0.6322957
> out$adj.r.squared # 自由度調整済み決定係数
[1] 0.5863326
> # 計算式の確認
> ybar <- mean(y) # 目的変数の標本平均
> yhat <- fitted(mod) # あてはめ値
> sum((yhat - ybar)^2)/sum((y - ybar)^2) # 定義式
[1] 0.6322957
> cor(yhat, y)^2 # あてはめ値と目的変数の相関の二乗
[1] 0.6322957
> ehat <- resid(mod) # 残差
> 1 - mean(ehat^2)/mean((y - ybar)^2) # 残差を使った表示
[1] 0.6322957
> n <- length(y)
> 1 - (sum(ehat^2)/(n - 2))/(sum((y - ybar)^2)/(n - 1)) # 自由度調整済み決定係数の定義式
[1] 0.5863326
```

```
(slr-rsquared.r)
```

演習 11.4. 決定係数について調べてみよう.

(1) (11.10) 式が成り立つことを確認してみよ.
(2) (11.11) 式が成り立つことを確認してみよ.

11.6. 参考文献

1. 東京大学教養学部統計学教室編「統計学入門」, 東京大学出版会 (1991 年).
2. 吉田朋広著「数理統計学」, 朝倉書店 (2006 年).