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) 第 13 回

小池祐太

2018 年 1 月 10 日

1. 分散分析
   - 一元配置
   - 二元配置

2. 回帰分析
   - 回帰モデル
   - 回帰係数の点推定
     - 最小二乗法
     - R での実行
   - 回帰係数の区間推定
   - 回帰係数の有意性の検定
   - 決定係数

分散分析

- 前回の講義で述べた平均の差の検定は, 2 つのグループ間で平均の差があるか否かを検定する方法であった
- **分散分析**とは, 大雑把にいうと, 2 つ以上のグループ間で平均の差があるか否かを検定する方法
  - 例えば, ある小売店について,「売上高は月によって差があるか」という仮説を検定したり, また, ある銘柄の株価について「収益率は曜日によって差があるか」という仮説を検定するのに分散分析は有用

分散分析

- 分散分析の基本的な考え方は, データの変動からグループ間での変動と観測誤差のみに起因する変動を抽出し, 両者を比較すること
  - もしグループ間で平均に差がなければ, グループ間での変動は観測誤差のみに起因する変動と自由度を除いて本質的な差がないはず
  - 逆にグループ間で平均に差があれば, 前者はその分だけ変動が増して後者より大きくなるはずなので, 両者の比較によって目的の検定が実行できる
- 従って, 分散分析は「分散の分析」というよりむしろ「データの変動の分析」であるといえる

一元配置

- まず, グループ分けが 1 種類の場合を考え, p 個のグループ $A_1, A_2, \ldots, A_p$ があるとする
- 統計学では, グループ分けのことを**因子**と呼び, 因子内の各グループのことを**水準**と呼ぶことが多いため, 以下これらの用語を用いる
- 各 $i = 1, 2, \ldots, p$ について n_i 個の観測データ $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$ が与えられている状況を考える (例えば, $A_1, A_2, \ldots, A_p$ が月に対応し, $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$ が i 月の各日における売上高に対応していると考えれば良い)

一元配置

- 観測データは以下のモデルに従うと仮定する:

$$Y_{ij} = \mu_i + \varepsilon_{ij} \quad (i = 1, \ldots, p;\ j = 1, \ldots, n_i). \tag{1}$$

  - μ_i は定数であり, 水準 A_i における観測データの平均値を表す
  - ε_{ij} は確率変数であり, $\varepsilon_{11}, \ldots, \varepsilon_{1n_1}, \ldots, \varepsilon_{p1}, \ldots, \varepsilon_{pn_p}$ は独立同分布で平均 0, 分散 σ^2 の正規分布に従うと仮定する
- 水準 $A_1, A_2, \ldots, A_p$ の間の平均値に差があるか否かを検定する問題は, 以下のように定式化できる:

$$H_0 : \mu_1 = \cdots = \mu_p \qquad \text{vs} \qquad H_1 : \text{ある } i, j \text{ に対して} \mu_i \neq \mu_j.$$

一元配置

- 冒頭で述べたように, 分散分析ではデータの変動から因子間での変動と観測誤差のみに起因する変動を抽出し, 両者を比較することで検定を構成する
- まず, データ全体の標本平均 $\bar{Y}_{\cdot\cdot}$ および水準 A_i における標本平均 $\bar{Y}_{i\cdot}$ を以下で定義する:

$$\bar{Y}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{n_i} Y_{ij}, \qquad \bar{Y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \quad (i = 1, \ldots, p).$$

ただし, $n := \sum_{i=1}^{p} n_i$ は全サンプル数を表す

一元配置

- 次に, 各水準内でのデータの変動 (の合計)SS_W, 水準間でのデータの変動 SS_B を以下で定義する:

$$SS_W = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2,$$

$$SS_B = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 = \sum_{i=1}^{p} n_i (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2.$$

  - SS_W を**級内変動**, SS_B を**級間変動**と呼ぶ
  - いまの設定では, 級内変動 SS_W は観測誤差にみに起因して生じる

一元配置

- 仮に帰無仮説 H_0 が正しければ, 水準内でのデータの変動・水準間でのデータの変動ともに観測誤差のみが原因で生じるはずなので, 自由度を除けば本質的な違いはないはずである
- 逆に, 対立仮説 H_1 が正しければ, 水準間でのデータの変動は観測誤差のみならず, 水準間での平均値 $\mu_1, \ldots, \mu_p$ の異質性にも影響されるはずなので, SS_B は SS_W より本質的に大きくなるはずである
- 数学的には, 帰無仮説の下で, $SS_W/(n-p)$, $SS_B/(p-1)$ はともに σ^2 の不偏推定量となることが示せる

一元配置

- 従って, 検定統計量として

$$F = \frac{SS_B/(p-1)}{SS_W/(n-p)}$$

  を考えるのが自然である
- 対立仮説の下では F は大きな値をとるはずなので, この検定は右片側検定となる
- 帰無仮説の下で次の事実が成り立つことが知られている: SS_B, SS_W は独立であり, SS_B は自由度 $p-1$ の χ^2 分布に従い, SS_W は自由度 $n-p$ の χ^2 分布に従う
- 従って, 帰無仮説の下で F は自由度 $p-1, n-p$ の F 分布に従う (6.3.6 節参照)

一元配置

- よって, $\alpha \in (0,1)$ に対して, 自由度 $p-1, n-p$ の F 分布の $100(1-\alpha)$%分位点を $F_{1-\alpha}(p-1, n-p)$ とすれば, H_0 の下では

$$P(F > F_{1-\alpha}(p-1, n-p)) = \alpha$$

  が成り立つ
- 以上より, 有意水準を α とする場合, 棄却域を

$$(F_{1-\alpha}(p-1, n-p), \infty)$$

  と設定すれば, 第一種過誤の上限が α となる

一元配置

- 具体的な検定の手順としては, データから検定統計量 F の値を計算し,

$$F > F_{1-\alpha}(p-1, n-p)$$

であった場合には帰無仮説を棄却する

- もしくは, $f(x)$ を自由度 $p-1, n-p$ の F 分布の確率密度関数として, p 値

$$\int_F^\infty f(x)dx$$

が α 未満であった場合に帰無仮説を棄却するという手順をとっても同等

分散分析表 (一元配置の場合)

	自由度	平方和	平均平方和	F 値	p 値
級間	$p-1$	SS_B	$SS_B/(p-1)$	F	$\int_F^\infty f(x)dx$
級内	$n-p$	SS_W	$SS_W/(n-p)$		

一元配置

- モデル (1) では各水準の効果をその水準における平均値で表していたが, 因子 A 全体の平均効果を μ で表して, 平均 μ を基準とした各水準 A_i の相対的な効果 α_i で表すことも可能
- すなわち,

$$\mu_i = \mu + \alpha_i, \qquad \mu = \frac{1}{n}\sum_{i=1}^{p} n_i \mu_i$$

とする

一元配置

- このとき,

$$\sum_{i=1}^{p} n_i \alpha_i = 0$$

であるから, 帰無仮説 H_0 は

$$\alpha_1 = \cdots = \alpha_p = 0$$

と同等

- R には分散分析を実行するための関数 `aov()` が用意されている
- 実行例 `anova-oneway.r`

二元配置

- 次に, 因子が 2 種類ある場合を考え, 一方の因子の水準間の平均値に差があるか否かを検定する問題を考える (もう一方の因子の水準間で平均値に差があるかは問わない)
- **例** いくつかの薬の効能を比較するために何人かの被験者にそれぞれの薬を投与して治験結果を集めた場合
  - 「薬の種類」と「被験者番号」の 2 種類の因子
  - 「薬の種類」という因子間での薬の効能の差を検証したい
  - 薬の効き目には個人差があると考えられるため, 「被験者番号」という因子間で効能に差があることは許容したい

二元配置

- 2 種類の因子 A, B があるとし, 因子 A には a 個の水準 $A_1, \ldots, A_a$ があり, 因子 B には b 個の水準 $B_1, \ldots, B_b$ があるとする
- 因子 A, B の水準がそれぞれ A_i, B_j であるようなデータの観測値が Y_{ij} で与えられているとし, 以下のモデルに従うとする:

$$Y_{ij} = \alpha_i + \beta_j + \varepsilon_{ij} \quad (i = 1, \ldots, a;\ j = 1, \ldots, b). \tag{2}$$

  - α_i, β_j はともに定数であり, それぞれ因子 A, B の水準 A_i, B_j における効果を表す
  - ε_{ij} は確率変数であり, $\varepsilon_{11}, \ldots, \varepsilon_{1b}, \ldots, \varepsilon_{a1}, \ldots, \varepsilon_{ab}$ は独立同分布で平均 0, 分散 σ^2 の正規分布に従うと仮定する

二元配置

- 上の例でいうと, 因子 A が「薬の種類」, 因子 B が「被験者番号」に対応し, α_i は薬 A_i の効能を, β_j は被験者 B_j 固有の薬の効きやすさに対応すると考えられる
  - 薬の効能に差があるか否かという検定は, 因子 A の水準間の効果に差があるか否かを検定する問題となる
- 因子 A の水準間の効果に差があるか否かの検定は以下のように定式化できる:

$$H_0 : \alpha_1 = \cdots = \alpha_a \qquad \text{vs} \qquad H_1 : \text{ある } i_1, i_2 \text{ に対して} \alpha_{i_1} \neq \alpha_{i_2}.$$

二元配置

- 一元配置の場合と同様に, データの変動から因子間での変動と観測誤差のみに起因する変動を抽出し, 両者を比較することで検定を構成する
- データ全体の標本平均 $\bar{Y}_{..}$, 因子 A の水準 A_i における標本平均 $\bar{Y}_{i.}$, および因子 B の水準 B_j における標本平均 $\bar{Y}_{.j}$ をそれぞれ以下で定義:

$$\bar{Y}_{..} = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b} Y_{ij}, \quad \bar{Y}_{i.} = \frac{1}{b}\sum_{j=1}^{b} Y_{ij} \quad (i = 1, \ldots, a),$$

$$\bar{Y}_{.j} = \frac{1}{a}\sum_{i=1}^{a} Y_{ij} \quad (j = 1, \ldots, b).$$

二元配置

- 因子 A 内でのデータの変動 SS_A および因子 B 内でのデータの変動 SS_B を以下で定義:

$$SS_A = b\sum_{i=1}^{a}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2, \quad SS_B = a\sum_{j=1}^{b}(\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot})^2.$$

  - SS_A を**行間変動**, SS_B を**列間変動**と呼ぶ
- 仮に帰無仮説 H_0 が正しければ, 因子 A 内でのデータの変動 SS_A は観測誤差のみが原因で生じるはずなので, SS_A を観測誤差による変動と比較するのが自然である

二元配置

- 観測誤差による変動は次の統計量で計算できる:

$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{b}(Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot})^2.$$

  - 実際, $\bar{Y}_{i\cdot}, \bar{Y}_{\cdot j}, \bar{Y}_{\cdot\cdot}$ はそれぞれ

$$\alpha_i + \frac{1}{b}\sum_{j=1}^{b}\beta_j, \quad \frac{1}{a}\sum_{i=1}^{a}\alpha_i + \beta_j,$$

$$\frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}(\alpha_i + \beta_j) = \frac{1}{a}\sum_{i=1}^{a}\alpha_i + \frac{1}{b}\sum_{j=1}^{b}\beta_j$$

    の推定量とみなせるため, $Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot}$ は観測誤差 ε_{ij} に対応するものと考えられる
  - SS_E は**誤差変動**と呼ばれる

二元配置

- 帰無仮説 H_0 が正しければ, 変動 SS_A, SS_E はともに観測誤差のみが原因で生じるはずなので, 自由度を除けば本質的な違いはないはずである
- 逆に, 対立仮説 H_1 が正しければ, 因子 A 内でのデータの変動は観測誤差のみならず, 因子 A 内の水準間での効果 $\alpha_1, \ldots, \alpha_a$ の異質性にも影響されるはずなので, SS_A は SS_E より本質的に大きくなるはずである
- 数学的には, 帰無仮説の下で, $SS_A/(a-1)$, $SS_E/\{(a-1)(b-1)\}$ はともに σ^2 の不偏推定量となることが示せる

二元配置

- 従って, 検定統計量として

$$F_A = \frac{SS_A/(a-1)}{SS_E/\{(a-1)(b-1)\}}$$

  を考えるのが自然である
- 対立仮説の下では F_A は大きな値をとるはずなので, この検定は右片側検定となる
- 帰無仮説の下で次の事実が成り立つことが知られている: SS_A, SS_E は独立であり, SS_A は自由度 $a-1$ の χ^2 分布に従い, SS_E は自由度 $(a-1)(b-1)$ の χ^2 分布に従う
- 従って, 帰無仮説の下で F_A は自由度 $a-1, (a-1)(b-1)$ の F 分布に従う (6.3.6 節参照)

二元配置

- よって, $\alpha \in (0,1)$ に対して, 自由度 $a-1,(a-1)(b-1)$ の F 分布の $100(1-\alpha)$%分位点を $F_{1-\alpha}(a-1,(a-1)(b-1))$ とすれば, H_0 の下では

$$P(F_A > F_{1-\alpha}(a-1,(a-1)(b-1))) = \alpha$$

  が成り立つ
- 以上より, 有意水準を α とする場合, 棄却域を

$$(F_{1-\alpha}(a-1,(a-1)(b-1)),\infty)$$

  と設定すれば, 第一種過誤の上限が α となる

二元配置

- 具体的な検定の手順としては, データから検定統計量 F_A の値を計算し,

$$F_A > F_{1-\alpha}(a-1, (a-1)(b-1))$$

であった場合には帰無仮説を棄却する

- もしくは, $f_{a-1,(a-1)(b-1)}(x)$ を自由度 $a-1, (a-1)(b-1)$ の F 分布の確率密度関数として, p 値

$$\int_{F_A}^{\infty} f_{a-1,(a-1)(b-1)}(x)dx$$

が α 未満であった場合に帰無仮説を棄却するという手順をとっても同等である

- 因子 A ではなく因子 B の水準間の平均の差に関心がある場合は, A と B の役割を入れ替えて同様の議論を行えばよい

分散分析表 (二配置の場合)

	自由度	平方和	平均平方和	F 値	p 値
因子 A	$a-1$	SS_A	$\frac{SS_A}{a-1}$	F_A	$\int_{F_A}^{\infty} f_{a-1,(a-1)(b-1)}(x)dx$
因子 B	$b-1$	SS_B	$\frac{SS_B}{b-1}$	F_B	$\int_{F_B}^{\infty} f_{b-1,(a-1)(b-1)}(x)dx$
誤差	$(a-1)(b-1)$	SS_E	$\frac{SS_E}{(a-1)(b-1)}$		

二元配置

- 一元配置の場合と同様に, モデル (2) において各水準の効果を全体の平均 μ^* に対する相対効果で表すことも可能である
- 実際, $\bar{\alpha} = \frac{1}{a}\sum_{i=1}^{a}\alpha_i, \bar{\alpha} = \frac{1}{b}\sum_{j=1}^{b}\beta_j$ とし,

$$\mu^* = \bar{\alpha} + \bar{\beta},\ \alpha_i^* = \alpha_i - \bar{\alpha},\ \beta_j^* = \beta_j - \bar{\beta}$$

とおけば, α_i^*, β_j^* はそれぞれ水準 A_i, B_j の相対効果に対応し, モデル (2) は

$$Y_{ij} = \mu^* + \alpha_i^* + \beta_j^* + \varepsilon_{ij} \quad (i = 1, \ldots, a;\ j = 1, \ldots, b)$$

と書き直せる

二元配置

- このとき

$$\sum_{i=1}^{a} \alpha_i^* = \sum_{j=1}^{b} \beta_j^* = 0$$

であるから, 帰無仮説 H_0 は,

$$\alpha_1^* = \cdots = \alpha_a^* = 0$$

と同等となる

- 実行例 `anova-twoway.r`

回帰分析

- **回帰分析**
  - 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
  - 説明される側のデータ $\cdots$ **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 $\cdots$ **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
  - 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)

回帰モデル

- 以下では, 説明変数を X, 目的変数を Y で表す
- Y を X で説明するための関係式としては様々なモデルが考えられるが, 本講義では Y を X の一次関数でモデル化する場合 (**線形回帰**) を考える:

$$Y = \alpha + \beta X \tag{3}$$

  - 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

回帰係数の点推定

- モデル (3) 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
- 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 (3) 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

回帰係数の点推定

- 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:

$$Y_i = \alpha + \beta X_i + \epsilon_i, \qquad i = 1, \ldots, n. \tag{4}$$

  - $\epsilon_1, \ldots, \epsilon_n$ 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.

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
  - 回帰係数 α, β 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 $e_i(\alpha, \beta) = Y_i - (\alpha + \beta X_i)$ $(i = 1, \ldots, n)$ で与えられる
  - $e_1(\alpha, \beta), \ldots, e_n(\alpha, \beta)$ の絶対値がいずれも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 α, β を決定する

最小二乗法

- $S(\alpha, \beta)$ は**残差平方和**と呼ばれる
- $S(\alpha, \beta)$ 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, しばしば記号 $(\hat{\alpha}, \hat{\beta})$ で表される
- 最小二乗推定量は仮定 (A) の下でただ一つだけ存在し,

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \quad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} \tag{5}$$

で与えられる. ただし,

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

R での実行

- R で線形回帰分析を実行するためには関数 `lm( )` を用いる
- 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き, モデル (4) の回帰係数の推定は,

  `lm(y ~ x)`

  で実行できる
- データフレームの一部の変数を指定して線形回帰を実行することも可能 (以下の実行例参照)
- 実行例 `slr.r`

回帰係数の区間推定

- パラメーター α, β の区間推定について議論する
- 誤差項に関する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

回帰係数の区間推定

- 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

回帰係数の区間推定

- $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$ は**残差**と呼ばれ, 以下を満たす:

$$\sum_i \hat{\epsilon}_i = 0, \qquad \sum_i \hat{\epsilon}_i X_i = 0. \tag{6}$$

- 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:

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

回帰係数の区間推定

- 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
- 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

回帰係数の区間推定

- $(n-2)s.e.(\hat{\beta})^2/\mathrm{Var}[\hat{\beta}]$ 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,

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

  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

回帰係数の区間推定

- α の信頼区間も同様の方法で構成できる (配布資料参照)
- 実行例 `slr-ci.r`

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
- 線形回帰モデル (4) においてこれを検証するには, 次の検定問題を考えればよい:

$$H_0 : \beta = 0 \qquad \text{vs} \qquad H_1 : \beta \neq 0 \tag{7}$$

- 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
- 以下, 引き続き条件 (C) を仮定した下で, 上の検定を実行する方法を説明する

回帰係数の有意性の検定

- 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

回帰係数の有意性の検定

- 以上より, 有意水準を $\gamma \in (0,1)$ とする場合, 検定 (7) 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

$$2\int_{|t|}^{\infty} f(x)dx \tag{8}$$

  が γ 未満の場合に帰無仮説を棄却するとしても同等 ($f(x)$ は自由度 $n-2$ の t 分布の確率密度関数を表す)

回帰係数の有意性の検定

- 検定統計量の値 t を $\hat{\beta}$ の t **値**と呼び, 検定の p 値 (8) を $\hat{\beta}$ の p **値**と呼ぶ
- 定数項 α についても類推の議論で検定を実行することが可能
- 実行例 `slr-test.r`

決定係数

- 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

決定係数

- $\hat{\epsilon}_i = Y_i - \hat{Y}_i$ $(i = 1, \ldots, n)$ が成り立つことに注意すると, (6) 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) 変動に対応している
- 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)

決定係数

- R^2 は以下のように書き直すことも可能である:

$$R^2 = \left\{ \frac{\sum_i (\hat{Y}_i - \bar{Y})(Y_i - \bar{Y})}{\sqrt{\sum_i (Y_i - \bar{Y})^2} \cdot \sqrt{\sum_i (\hat{Y}_i - \bar{Y})^2}} \right\}^2 . \tag{9}$$

- 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い

決定係数

- R^2 の以下のようにも書き直せる:

$$R^2 = 1 - \frac{\frac{1}{n}\sum_i \hat{\epsilon}_i^2}{\frac{1}{n}\sum_i (Y_i - \bar{Y})^2}. \tag{10}$$

- 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

決定係数

- 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)
- 実行例 `slr-rsquared.r`