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) 第 9 回

小池祐太

2018 年 6 月 7 日

1. 重回帰分析: 復習
2. 分析の評価
   - 決定係数
   - F 値
3. 予測
4. 発展的なモデル
   - 変数が多い場合のモデルの記述法
   - 質的データの利用
   - 交互作用モデル・変数の非線形変換

重回帰分析

- **回帰分析 (regression analysis)**
  - 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
- 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

重回帰分析

- 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

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

  が得られている状況を考える
- 観測データは次のモデルに従うとする:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \qquad i = 1, \ldots, n. \tag{2}$$

  - $\beta_0, \beta_1, \ldots, \beta_p$ は未知パラメーター (**回帰係数 (regression coefficients)**)
  - $\epsilon_1, \ldots, \epsilon_n$ (**誤差項 (error term)**): 独立な確率変数 (多くの場合それぞれ平均 0, 分散 σ^2 の正規分布に従うと仮定)

最小二乗法

- 回帰係数 $\boldsymbol{\beta} := (\beta_0, \beta_1, \ldots, \beta_p)^\top$ は**最小二乗法 (least squares)** によって決定 (推定) する
- すなわち, **残差平方和 (residual sum of squares)**

$$S(\boldsymbol{\beta}) := \sum_{i=1}^{n} |y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})|^2$$

を最小化するベクトル $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)^\top$ を $\boldsymbol{\beta}$ の推定量とする (**最小二乗推定量 (least squares estimator)**)

- R での実行: 関数 `lm()`

あてはめ値と残差

- 最小二乗推定量と説明変数の観測データを使って, 目的変数の観測データ y_i から観測誤差の影響を除いた値の予測値を

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$$

  で計算できる
  - **あてはめ値 (fitted values)** または**予測値 (predicted values)** と呼ぶ (R での実行: 関数 `fitted()`)
- 目的変数の観測値 y_i からのあてはめ値 $\hat{y}_i$ のずれ $\hat{\epsilon}_i := y_i - \hat{y}_i$ は**残差 (residual)** と呼ばれる (R での実行: 関数 `resid()`)
- あてはめ値と残差は常に直交する ($\sum_{i=1}^{n} \hat{y}_i \hat{\epsilon}_i = 0$)

分析の評価

- 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

決定係数

- **決定係数**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 $\hat{y}_1, \ldots, \hat{y}_n$ と実際の観測データ $y_1, \ldots, y_n$ の相関の 2 乗として定義される:

$$R^2 = \frac{\left(\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})\right)^2}{\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2 \sum_{i=1}^n (y_i - \bar{y})^2}. \tag{3}$$

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

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

決定係数

- 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 (資料参照, もしくは自分で確認):

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{4}$$

- (4) 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

- 決定係数はさらに以下のようにも書き直せる (資料参照, もしくは自分で確認):

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}. \tag{5}$$

- (5) より, 決定係数は説明変数を付け加える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

決定係数

- 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**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
- 実行例 `rsquared.r`

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**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:

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

F 値

- 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
- したがって, 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

$$\int_F^\infty f(x)dx, \quad f(x) \text{ は自由度 } p, n-p-1 \text{ の } F \text{ 分布の確率密度関数} \tag{6}$$

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

F 値

- 統計の言葉で言うと, F 値及び確率 (6) は, 仮説検定

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

  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 (6) 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
- 実行例 `fstatistics.r`

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
- 実行例 `predict.r`

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
- 他にもより複雑な回帰モデルを記述するための記法がいくつかあるので, 必要に応じて自分で調べて欲しい
- 実行例 `lm-model.r`

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

質的データの利用

- まず, 例として, (性別, 身長) の観測データ $(x_1, y_1), \ldots, (x_n, y_n)$ 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 $x_1, \ldots, x_n$ 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:

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

質的データの利用

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

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

- したがって, 係数 β_0 は女性の平均身長, $\beta_0 + \beta_1$ は男性の平均身長, β_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

質的データの利用

- 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 $C_1, \ldots, C_k$ をもつ質的データ $x_1, \ldots, x_n$ 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 $\boldsymbol{z}_j = (z_{1j}, \ldots, z_{nj})^\top$ $(j = 1, \ldots, k-1)$ 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

質的データの利用

- 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
  - ただし, 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, 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 (以下の実行例参照)
- 実行例 `dummy.r`, `dummy-kikou.r`

交互作用モデル・変数の非線形変換

- 前述のように, モデル

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

  において, 説明変数として $X_j X_k$(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
- 実行例 `cross.r`