

# 回帰分析

## 予測と発展的なモデル

村田 昇

2020.10.23

## 講義の予定

- 第1日: 回帰モデルの考え方と推定
- 第2日: モデルの評価
- 第3日: モデルによる予測と発展的なモデル

## 回帰分析の復習

### 線形回帰モデル

- 目的変数  $y$  を説明変数  $x_1, \dots, x_p$  で説明する関係式を構成:
  - 説明変数:  $x_1, \dots, x_p$  (p 次元)
  - 目的変数:  $y$  (1 次元)
- 回帰係数  $\beta_0, \beta_1, \dots, \beta_p$  を用いた一次式:

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

- 誤差項を含む確率モデルで観測データを表現:

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \epsilon_i \quad (i = 1, \dots, n)$$

### 問題設定

- 確率モデル:

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

- 式の評価: 残差平方和 の最小による推定

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

### 解

- 解の条件: 正規方程式

$$X^\top X \boldsymbol{\beta} = X^\top \mathbf{y}$$

- 解の一意性: Gram 行列  $X^\top X$  が正則

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

## 寄与率

- 決定係数 (R-squared):

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

- 自由度調整済み決定係数 (adjusted R-squared):

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

## t-統計量による検定

- 回帰係数  $\beta_j$  が回帰式に寄与するか否かを検定する
  - 帰無仮説:  $\beta_j = 0$
  - 対立仮説:  $\beta_j \neq 0$  ( $\beta_j$  は役に立つ)
- t-統計量: 各係数ごと,  $\xi$  は  $(X^T X)^{-1}$  の対角成分

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

- p-値: 自由度  $n-p-1$  の  $t$  分布を用いて計算

## F-統計量による検定

- 説明変数のうち 1 つでも役に立つか否かを検定する
  - 帰無仮説:  $\beta_1 = \dots = \beta_p = 0$
  - 対立仮説:  $\exists j \beta_j \neq 0$  (少なくとも 1 つは役に立つ)
- F-統計量: 決定係数 (または残差) を用いて計算

$$F = \frac{n-p-1}{p} \frac{R^2}{1-R^2}$$

- p-値: 自由度  $p, n-p-1$  の  $F$  分布で計算

## 回帰モデルによる予測

### 予測

- 新しいデータ (説明変数)  $\mathbf{x}$  に対する予測値

$$\hat{y} = (1, \mathbf{x}^T) \hat{\beta}, \quad \hat{\beta} = (X^T X)^{-1} X^T \mathbf{y}$$

- 予測値は元データの目的変数の重み付け線形和

$$\hat{y} = \mathbf{w}(\mathbf{x})^T \mathbf{y}$$

- 重みは元データと新規データの説明変数で決定

$$\mathbf{w}(\mathbf{x})^T = (1, \mathbf{x}^T) (X^T X)^{-1} X^T$$

## 予測値の分布

- 推定量は以下の性質をもつ多変量正規分布

$$\begin{aligned}\mathbb{E}[\hat{\beta}] &= \beta \\ \text{Cov}(\hat{\beta}) &= \sigma^2 (X^T X)^{-1}\end{aligned}$$

- この性質を利用して以下の3つの値の違いを評価

$$\begin{aligned}\hat{y} &= (1, \mathbf{x}^T) \hat{\beta} && \text{(回帰式による予測値)} \\ \tilde{y} &= (1, \mathbf{x}^T) \beta && \text{(最適な予測値)} \\ y &= (1, \mathbf{x}^T) \beta + \epsilon && \text{(観測値)}\end{aligned}$$

( $\hat{y}$  と  $y$  は独立な正規分布に従うことに注意)

## 最適な予測値との差

- 差の分布は以下の平均・分散の正規分布

$$\begin{aligned}\tilde{\mathbf{x}}^T &= (1, \mathbf{x}^T) \\ \mathbb{E}[\tilde{y} - \hat{y}] &= \tilde{\mathbf{x}}^T \beta - \tilde{\mathbf{x}}^T \mathbb{E}[\hat{\beta}] = 0 \\ \text{Var}(\tilde{y} - \hat{y}) &= \underbrace{\sigma^2 \tilde{\mathbf{x}}^T (X^T X)^{-1} \tilde{\mathbf{x}}}_{\hat{\beta} \text{ の推定誤差による分散}} = \sigma^2 \gamma_c(\mathbf{x})^2\end{aligned}$$

- 正規化による表現

$$\frac{\tilde{y} - \hat{y}}{\sigma \gamma_c(\mathbf{x})} \sim \mathcal{N}(0, 1)$$

## 信頼区間

- 未知の分散を不偏分散で推定

$$Z = \frac{\tilde{y} - \hat{y}}{\hat{\sigma} \gamma_c(\mathbf{x})} \sim \mathcal{T}(n-p-1) \quad (t\text{-分布})$$

- 確率  $\alpha$  の信頼区間 (最適な予測値  $\tilde{y}$  が入ることが期待される区間)

$$(\hat{y} - C_\alpha \hat{\sigma} \gamma_c(\mathbf{x}), \hat{y} + C_\alpha \hat{\sigma} \gamma_c(\mathbf{x}))$$

ただし  $C_\alpha$  は以下を満たす定数

$$P(|Z| < C_\alpha | Z \sim \mathcal{T}(n-p-1)) = \alpha$$

## 観測値との差

- 差の分布は以下の平均・分散の正規分布

$$\begin{aligned} \mathbb{E}[y - \hat{y}] &= \tilde{\mathbf{x}}^\top \boldsymbol{\beta} - \tilde{\mathbf{x}}^\top \mathbb{E}[\hat{\boldsymbol{\beta}}] = 0 \\ \text{Var}(y - \hat{y}) &= \underbrace{\sigma^2 \tilde{\mathbf{x}}^\top (X^\top X)^{-1} \tilde{\mathbf{x}}}_{\hat{\boldsymbol{\beta}} \text{ の推定誤差による分散}} + \underbrace{\sigma^2}_{\text{誤差の分散}} = \sigma^2 \gamma_p(\mathbf{x})^2 \end{aligned}$$

- 正規化による表現

$$\frac{y - \hat{y}}{\sigma \gamma_p(\mathbf{x})} \sim \mathcal{N}(0, 1)$$

## 予測区間

- 未知の分散を不偏分散で推定

$$Z = \frac{y - \hat{y}}{\hat{\sigma} \gamma_p(\mathbf{x})} \sim \mathcal{T}(n-p-1) \quad (t\text{-分布})$$

- 確率  $\alpha$  の予測区間 (観測値  $y$  が入ることが期待される区間)

$$(\hat{y} - C_\alpha \hat{\sigma} \gamma_p(\mathbf{x}), \hat{y} + C_\alpha \hat{\sigma} \gamma_p(\mathbf{x}))$$

ただし  $C_\alpha$  は以下を満たす定数

$$P(|Z| < C_\alpha | Z \sim \mathcal{T}(n-p-1)) = \alpha$$

- $\gamma_p > \gamma_c$  なので信頼区間より広がる

## R: モデルからの予測

- 関数 `predict()` を用いた予測:

```
## モデルの作成
train <- data.frame(x1=..., x2=..., y=...)
est <- lm(y ~ x1 + x2, data=train)
fit <- predict(est) # あてはめ値の計算
## 新しいデータの予測
test <- data.frame(x1=..., x2=...) # 予測したいデータの説明変数
pred <- predict(est, # 予測値の計算
               newdata=test) # 説明変数のデータフレーム
cint <- predict(est, newdata=test,
               interval="confidence", level=0.95) # 信頼区間
pint <- predict(est, newdata=test,
               interval="prediction", level=0.95) # 予測区間
## 信頼区間, 予測区間の水準の既定値は 0.95
```

## R: 予測の例

- 東京の気候データによる例:

```

### 9,10月のデータでモデルを構築し, 8,11月のデータを予測
TW.data <- transform(read.csv("data/tokyo_weather_reg.csv"),
                      month=as.numeric(months(as.Date(date), # 月 (数値) を付加
                      abbreviate=TRUE)))

TW.model <- temp ~ solar + press # モデルの定義
TW.train <- subset(TW.data, # モデル推定用データ
                  subset= month %in% c(9,10))
TW.test  <- subset(TW.data, # 予測用データ
                  subset= month %in% c(8,11))
TW.est <- lm(TW.model, data=TW.train) # モデルの推定
summary(TW.est) # モデルの評価
## 予測値の計算
TW.fit <- predict(TW.est) # データのあてはめ値
TW.pred <- predict(TW.est, # 新規データの予測値
                  newdata=TW.test)

```

## 練習問題

- 東京の気候データを用いて以下の実験を試みなさい
  - 8月のデータで回帰式を推定する
  - 上記のモデルで9月のデータを予測する

```

## 8月と9月のデータを取り出すには, 例えば以下のようにすればよい
TW.data <- transform(read.csv("data/tokyo_weather_reg.csv"),
                      month=as.numeric(months(as.Date(date),
                      abbreviate=TRUE)))

TW.model <- temp ~ solar + press + cloud # モデルの定義
TW.train <- subset(TW.data, subset= month==8) # 推定用データ
TW.test  <- subset(TW.data, subset= month %in% 9) # 予測用データ
## %in% は集合に含まれるかをチェックする
## 例えば 7月, 9月を抽出するには month %in% c(7,9)

```

- 人工データを作成してモデルの予測について検討しなさい (宿題)

## 非線形の関係

### 非線形な関係のモデル化

- 目的変数  $Y$
- 説明変数  $X_1, \dots, X_p$
- 説明変数の追加で対応可能
  - 交互作用 (交差項):  $X_i X_j$  のような説明変数の積
  - 非線形変換:  $\log(X_k)$  のような関数による変換

### R: 線形でないモデル式の書き方

- 交互作用を記述するためには特殊な記法がある
- 非線形変換はそのまま関数を記述すればよい
- 1つの変数の多項式は関数  $I()$  を用いる

```

## 目的変数 Y, 説明変数 X1, X2, X3
## 交互作用を含む式 (formula) の書き方
Y ~ X1 + X1:X2      # X1 + X1*X2
Y ~ X1 * X2          # X1 + X2 + X1*X2
Y ~ (X1 + X2 + X3)^2 # X1 + X2 + X3 + X1*X2 + X2*X3 + X3*X1
## 非線形変換を含む式 (formula) の書き方

```

```
Y ~ f(X1)          # f(X1) (fは任意の関数)
Y ~ X1 + I(X2^2)   # X1 + X2^2
```

## 練習問題

- 東京の気候データ (9-11 月) を用いて気温を回帰する以下のモデルを検討しなさい
  - 日射量, 気圧, 湿度の線形回帰モデル
  - 湿度の対数を考えた線形回帰モデル
  - 最初のモデルにそれぞれの交互作用を加えたモデル
  - 更に 3 つの変数の積を加えたモデル
  - 自由にモデルを設定してみよ

## カテゴリカル変数

### カテゴリカル変数

- 悪性良性や血液型などの数値ではないデータ
- 適切な方法で数値に変換して対応:
  - 2 値の場合は 0,1 を割り当てる
    - 悪性:1
    - 良性:0
  - 3 値以上の場合は **ダミー変数** を利用する (カテゴリ数-1 個)
    - A 型: (1,0,0)
    - B 型: (0,1,0)
    - O 型: (0,0,1)
    - AB 型: (0,0,0)

### R: カテゴリカル変数の取り扱い

- 何も宣言しなくても通常は適切に対応してくれる
- 陽にカテゴリカル変数として扱いたい場合は関数 `factor()` を利用:

```
## factor 属性の与え方
X <- c("A", "S", "A", "B", "D")
Y <- c(85, 100, 80, 70, 30)
dat1 <- data.frame(X, Y)
dat2 <- transform(dat1,
                  X2=factor(X))
str(dat2) # 作成したデータフレームの素性を見る
dat3 <- transform(dat2,
                  X3=factor(X, levels=c("S", "A", "B", "C", "D")))
str(dat3) # dat2 とは factor の順序が異なる
dat4 <- transform(dat2,
                  Y2=factor(Y > 60))
str(dat4) # 条件の真偽で 2 値に類別される
```

## 練習問題

- 東京の気候データ (1 年分) を用いて気温を回帰する以下のモデルを検討しなさい
  - 降水の有無を表すカテゴリカル変数を用いたモデル  
(雨が降ると気温が変化することを検証するモデル)
  - 月をカテゴリカル変数として加えたモデル  
(月毎の気温の差を考慮して検証するモデル)