

回帰分析

モデルの評価

村田 昇

講義概要

- 第1回：回帰モデルの考え方と推定
- 第2回：モデルの評価
- 第3回：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 を 説明変数 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

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

簡潔な表現のための行列

- デザイン行列 (説明変数)

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

簡潔な表現のためのベクトル

- ベクトル (目的変数・誤差・回帰係数)

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

問題の記述

- 確率モデル

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \text{確率分布}$$

- 回帰式の推定: **残差平方和** の最小化

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

解の表現

- 解の条件: **正規方程式**

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

- 解の一意性: **Gram 行列** $\mathbf{X}^\top \mathbf{X}$ が正則

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

最小二乗推定量の性質

- **あてはめ値** $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ は \mathbf{X} の列ベクトルの線形結合
- **残差** $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \hat{\mathbf{y}}$ はあてはめ値 $\hat{\mathbf{y}}$ と直交

$$\hat{\boldsymbol{\epsilon}}^\top \hat{\mathbf{y}} = 0$$

- 回帰式は説明変数と目的変数の **標本平均** を通過

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

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)

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

- 不偏分散で補正

解析の事例

実データによる例

- 気象庁より取得した東京の気候データ
 - 気象庁 <https://www.data.jma.go.jp/gmd/risk/obsdl/index.php>
 - データ https://noboru-murata.github.io/multivariate-analysis/data/tokyo_weather.csv

東京の8月の気候の分析

- データの一部

日付	気温	降雨	日射	降雪	風向	風速	気圧	湿度	雲量
2022-08-01	30.60	0.00	24.53	0.00	SSE	2.80	1010.10	72.00	8.80
2022-08-02	31.60	0.00	24.78	0.00	SSE	2.50	1008.80	71.00	9.80
2022-08-03	31.50	0.00	21.24	0.00	SSE	2.30	1005.10	75.00	7.30
2022-08-04	24.60	18.00	3.46	0.00	NE	2.70	1006.00	89.00	10.00
2022-08-05	23.80	0.00	7.65	0.00	NE	2.90	1006.10	83.00	9.80
2022-08-06	25.20	0.00	17.06	0.00	SSE	2.40	1008.10	73.00	10.00
2022-08-07	27.60	0.00	14.45	0.00	SSE	2.20	1009.30	80.00	8.30
2022-08-08	29.80	0.00	22.52	0.00	S	4.50	1008.50	75.00	4.80
2022-08-09	30.90	0.00	25.50	0.00	S	5.50	1006.90	69.00	6.80
2022-08-10	30.50	0.00	25.99	0.00	S	5.30	1007.20	70.00	6.00
2022-08-11	29.50	0.00	22.90	0.00	S	5.40	1007.50	75.00	6.00
2022-08-12	28.30	2.00	15.36	0.00	S	5.80	1007.50	81.00	9.80

- 気温を説明する5種類の線形回帰モデルを検討
 - モデル1: 気温 = F(気圧)
 - モデル2: 気温 = F(日射)
 - モデル3: 気温 = F(気圧, 日射)
 - モデル4: 気温 = F(気圧, 日射, 湿度)
 - モデル5: 気温 = F(気圧, 日射, 雲量)

分析の視覚化

- 関連するデータの散布図
- モデル1の推定結果
- モデル2の推定結果
- モデル3の推定結果
- 観測値とあてはめ値の比較

モデルの比較

- 寄与率による比較

モデル	決定係数	自由度調整済み決定係数
1 気温 = F(気圧)	0.064	0.031
2 気温 = F(日射)	0.641	0.628
3 気温 = F(気圧, 日射)	0.741	0.722
4 気温 = F(気圧, 日射, 湿度)	0.758	0.731
5 気温 = F(気圧, 日射, 雲量)	0.760	0.733

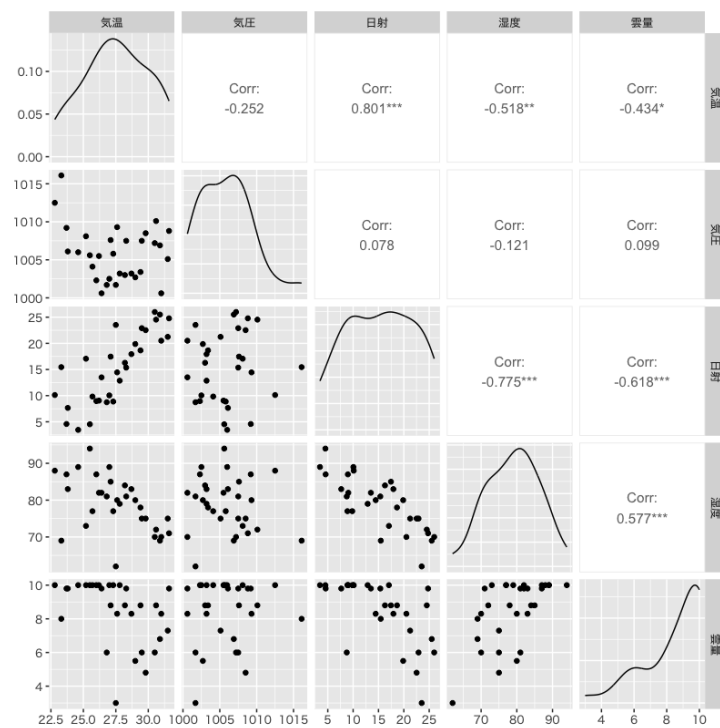


図 1: 散布図

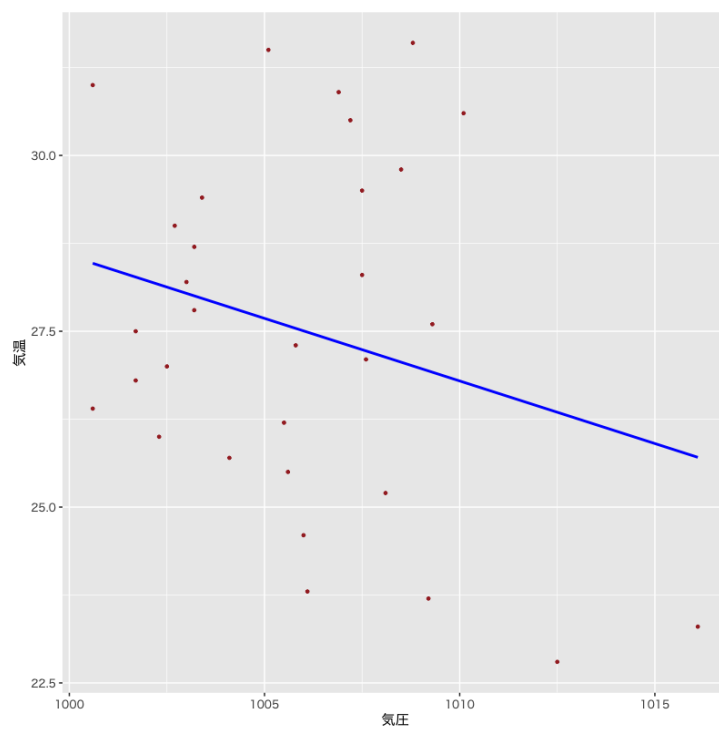


図 2: モデル 1

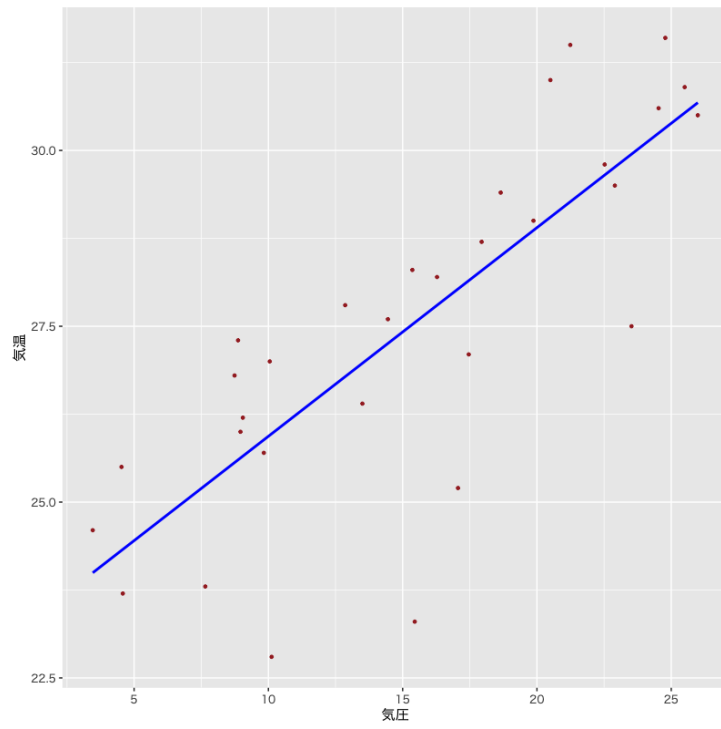


図 3: モデル 2

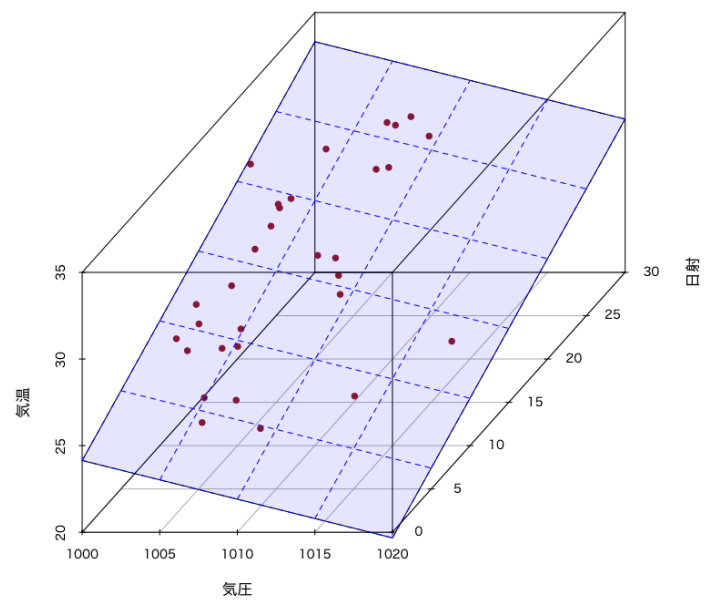


図 4: モデル 3

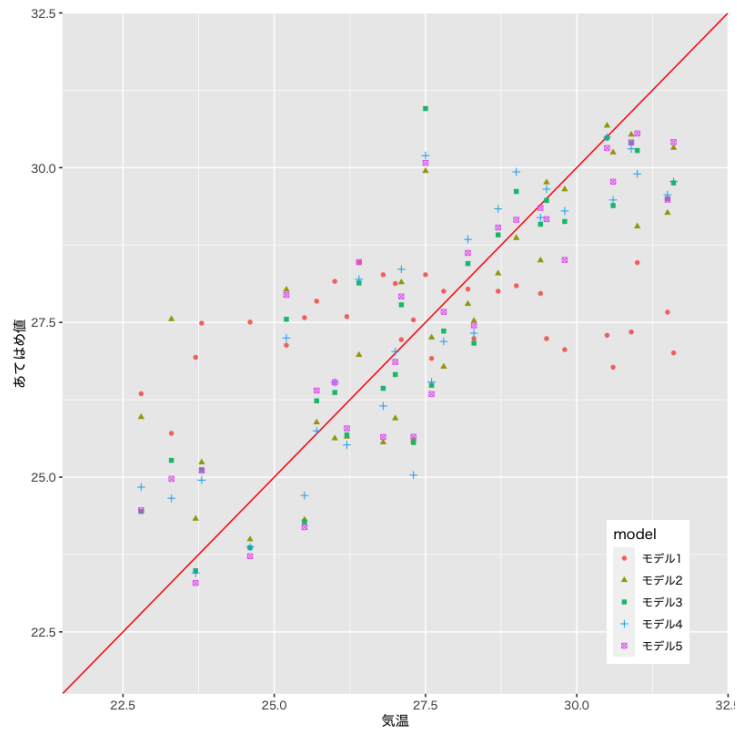


図 5: モデルの比較

あてはめ値の性質

あてはめ値

- さまざまな表現

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} \\ (\hat{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \text{ を代入}) \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\end{aligned}\tag{A}$$

$$\begin{aligned}(\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \text{ を代入}) \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\epsilon} \\ &= \mathbf{X}\boldsymbol{\beta} + \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\epsilon}\end{aligned}\tag{B}$$

- (A) あてはめ値は **観測値の重み付けの和** で表される
- (B) あてはめ値と観測値は **誤差項** の寄与のみ異なる

あてはめ値と誤差

- 残差と誤差の関係

$$\begin{aligned}\hat{\boldsymbol{\epsilon}} &= \mathbf{y} - \hat{\mathbf{y}} \\ &= \boldsymbol{\epsilon} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\epsilon} \\ &= (\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\boldsymbol{\epsilon}\end{aligned}\tag{C}$$

- (C) 残差は **誤差の重み付けの和** で表される

ハット行列

- 定義

$$H = X(X^T X)^{-1} X^T$$

- ハット行列 H による表現

$$\hat{y} = Hy$$

$$\hat{\epsilon} = (I - H)\epsilon$$

- あてはめ値や残差は H を用いて簡潔に表現される

ハット行列の性質

- 観測データ (デザイン行列) のみで計算される
- 観測データと説明変数の関係を表す
- 対角成分 (**テコ比**; leverage) は観測データが自身の予測に及ぼす影響の度合を表す

$$\hat{y}_j = (H)_{jj}y_j + (\text{それ以外のデータの寄与})$$

- $(A)_{ij}$ は行列 A の (i, j) 成分
- テコ比が小さい: 他のデータでも予測が可能
- テコ比が大きい: 他のデータでは予測が困難

推定量の統計的性質

最小二乗推定量の性質

- 推定量と誤差の関係

$$\begin{aligned}\hat{\beta} &= (X^T X)^{-1} X^T y \\ &= (X^T X)^{-1} X^T (X\beta + \epsilon) \\ &= (X^T X)^{-1} X^T X\beta + (X^T X)^{-1} X^T \epsilon \\ &= \beta + (X^T X)^{-1} X^T \epsilon\end{aligned}$$

- 正規分布の重要な性質 (**再生性**)
正規分布に従う独立な確率変数の和は正規分布に従う

推定量の分布

- **誤差の仮定**: 独立, 平均 0 分散 σ^2 の正規分布
- 推定量は以下の多変量正規分布に従う

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

- 通常 σ^2 は未知, 必要な場合には不偏分散で代用

$$\hat{\sigma}^2 = \frac{S}{n-p-1} = \frac{1}{n-p-1} \hat{\epsilon}^T \hat{\epsilon} = \frac{1}{n-p-1} \sum_{i=1}^n \epsilon_i^2$$

- これらの性質を利用してモデルの評価を行う

実習

R : 乱数を用いた人工データの生成

- 正規乱数を用いた線形単回帰モデル

```
set.seed(987) # 乱数のシード値を設定
x_obs <- c(1,3,5,7) # 説明変数の観測値
epsilon <- rnorm(length(x_obs), sd = 0.5) # 誤差項の生成
y_obs <- 2 - 3 * x_obs + epsilon # 目的変数の観測値
my_data <- tibble(x = x_obs, y = y_obs) # データフレームの作成
beta_lm <- lm(y ~ x, data = my_data) # 回帰係数の推定
coef(beta_lm) # 回帰係数の取得
summary(beta_lm) # 分析結果の概要の表示
```

R : 数値実験 (Monte-Carlo 法)

- 実験のためのコードは以下のようになる

```
mc <- 5000 # 実験回数を指定
mc_trial <- function() { # 1回の試行を行うプログラム
  ## 乱数生成と推定の処理
  return(返回值)}
mc_data <-
  replicate(mc, my_trial()) |> # Monte-Carlo 実験
  t() |> as_tibble() # 転置 (関数 t()) してデータフレームに変換
#' 適切な統計・視覚化処理 (下記は例)
mc_data |>
  summarise(across(everything(), var)) # 各列の分散の計算
ggpairs(mc_data) # 散布図行列の描画
tibble(x = mc_data[[k]]) |> # k 列目のベクトルで新しいデータフレームを作成
ggplot(aes(x = x)) + geom_histogram() # k 列目のデータのヒストグラム
```

練習問題

- 最小二乗推定量の性質を数値実験 (Monte-Carlo 法) により確認しなさい

- 以下のモデルに従う人工データを生成する

説明変数の観測データ :

$\{1, 20, 13, 9, 5, 15, 19, 8, 3, 4\}$

確率モデル :

$$y = -1 + 2 \times x + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 2)$$

- 観測データから回帰係数を推定する
- 実験を複数回繰り返し推定値 $(\hat{\beta}_0, \hat{\beta}_1)$ の分布を調べる

誤差の評価

各係数の推定量の分布

- 推定された回帰係数の精度を評価
 - 誤差 ϵ の分布は平均 0 分散 σ^2 の正規分布
 - $\hat{\beta}$ の分布: $p+1$ 変量正規分布

$$\hat{\beta} \sim N(\beta, \sigma^2(X^T X)^{-1})$$

- $\hat{\beta}_j$ の分布: 1 変量正規分布

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2((X^T X)^{-1})_{jj}) = N(\beta_j, \sigma^2 \zeta_j^2)$$

* $(A)_{jj}$ は行列 A の (j, j) (対角) 成分

標準誤差

- 標準誤差 (standard error): $\hat{\beta}_j$ の標準偏差の推定量

$$\text{s.e.}(\hat{\beta}_j) = \hat{\sigma} \zeta_j = \sqrt{\frac{1}{n-p-1} \sum_{i=1}^n \hat{\epsilon}_i^2} \cdot \sqrt{((X^T X)^{-1})_{jj}}$$

- 未知母数 σ^2 は不偏分散 $\hat{\sigma}^2$ で推定
- $\hat{\beta}_j$ の精度の評価指標

実習

練習問題

- 数値実験により標準誤差の性質を確認しなさい
 - 人工データを用いて標準誤差と真の誤差を比較する

```
#' 標準誤差は以下のようにして取り出せる
data_lm <- lm(formula, data)
summary(data_lm)$coefficients # 係数に関する情報はリストの要素として保管されている
summary(data_lm)$coefficients[,2] # 列番号での指定
summary(data_lm)$coef[, "Std. Error"] # 列名での指定, coef と省略してもよい
```

- 広告費と売上データを用いて係数の精度を議論する
- 東京の気候データを用いて係数の精度を議論する

係数の評価

t 統計量

- 回帰係数の分布 に関する定理
 - t 統計量 (t -statistic)

$$t = \frac{\hat{\beta}_j - \beta_j}{\text{s.e.}(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma} \zeta_j}$$

は自由度 $n-p-1$ の t 分布に従う

- 証明には以下の性質を用いる
 - $\hat{\sigma}^2$ と $\hat{\beta}$ は独立となる
 - $(\hat{\beta}_j - \beta_j)/(\sigma\zeta_j)$ は標準正規分布に従う
 - $(n-p-1)\hat{\sigma}^2/\sigma^2 = S(\hat{\beta})/\sigma^2$ は自由度 $n-p-1$ の χ^2 分布に従う

t 統計量による検定

- 回帰係数 β_j が回帰式に寄与するか否かを検定
 - 帰無仮説 $H_0: \beta_j = 0$ (t 統計量が計算できる)
 - 対立仮説 $H_1: \beta_j \neq 0$
- p 値: 確率変数の絶対値が $|t|$ を超える確率

$$(p \text{ 値}) = 2 \int_{|t|}^{\infty} f(x) dx \quad (\text{両側検定})$$

- $f(x)$ は自由度 $n-p-1$ の t 分布の確率密度関数
- 帰無仮説 H_0 が正しければ p 値は小さくならない

実習

練習問題

- 数値実験により t 統計量の性質を確認しなさい
 - 人工データを用いて t 統計量の分布を確認する

```
#' t 統計量とその p 値は以下のようにして取り出せる
data_lm <- lm(formula, data)
summary(data_lm)$coef[,c("t value", "Pr(>|t|)")] # 列名での指定
summary(data_lm)$coef[,3:4] # 列番号での指定
```

- 広告費と売上データを用いて係数の有意性を議論する
- 東京の気候データを用いて係数の有意性を議論する

モデルの評価

F 統計量

- ばらつきの比に関する定理:

$$\beta_1 = \dots = \beta_p = 0 \text{ ならば } F \text{ 統計量 (F-statistic)}$$

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

は自由度 $p, n-p-1$ の F 分布に従う

- 証明には以下の性質を用いる
 - S_r と S は独立となる
 - S_r/σ^2 は自由度 p の χ^2 分布に従う
 - S/σ^2 は自由度 $n-p-1$ の χ^2 分布に従う

F 統計量を用いた検定

- 説明変数のうち 1 つでも役に立つか否かを検定
 - 帰無仮説 $H_0: \beta_1 = \dots = \beta_p = 0$ (S_r が χ^2 分布になる)
 - 対立仮説 $H_1: \exists j \beta_j \neq 0$
- p 値: 確率変数の値が F を超える確率

$$(p \text{ 値}) = \int_F^{\infty} f(x)dx \quad (\text{片側検定})$$

- $f(x)$ は自由度 $p, n-p-1$ の F 分布の確率密度関数
- 帰無仮説 H_0 が正しければ p 値は小さくならない

実習

練習問題

- 数値実験により F 統計量の性質を確認しなさい
 - 人工データを用いて F 統計量の分布を確認しなさい

```
#' F統計量とその自由度は以下のようにして取り出せる
data_lm <- lm(formula, data)
summary(data_lm)$fstat
summary(data_lm)$fstatistic # 省略しない場合
```

- 広告費と売上データのモデルの有効性を議論しなさい
- 東京の気候データのモデルの有効性を議論しなさい

補足

R: 診断プロット

- 回帰モデルのあてはまりを視覚的に評価
 - Residuals vs Fitted:** あてはめ値 (予測値) と残差の関係
 - Normal Q-Q:** 残差の正規性の確認
 - Scale-Location:** あてはめ値と正規化した残差の関係
 - Residuals vs Leverage:** 正規化した残差とレバレッジの関係

などが用意されている

```
#' 関数 stats::lm() による推定結果の診断プロット
tw_lm6 <- lm(temp ~ press + solar + rain, data = tw_subset)
#' 関数 ggfortify::autoplot() を利用する
#' 必要であれば 'install.packages("ggfortify")' を実行
library(ggfortify)
autoplot(tw_lm6)
#' 診断プロットは 1 から 6 まで用意されており 1, 2, 3, 5 がまとめて表示される
#' 個別に表示する場合は 'autoplot(tw_lm6, which = 1)' のように指定する
#' 詳細は '?ggfortify::autoplot.lm' を参照
```

次回の予定

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