回帰分析

予測と発展的なモデル

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

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

- **誤差項** を含む確率モデルで観測データを表現

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

問題設定

- 確率モデル

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

- 式の評価 : **残差平方和** の最小化による推定

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

解

- 解の条件 : **正規方程式**

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

- 解の一意性 : **Gram 行列** $X^{\top}X$ が正則

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

実データによる例

- 東京の 8 月の気候 (気温, 降雨, 日射, 降雪, 風速, 気圧, 湿度, 雲量) に関するデータ (の一部)

	month	day	day_of_week	temp	rain	solar	snow	wdir	wind	press	humid	cloud
213	8	1	Sun	28.7	0.0	26.58	0	SSE	3.2	1000.2	76	2.3
214	8	2	Mon	28.6	0.5	19.95	0	SE	3.4	1006.1	80	7.0
215	8	3	Tue	29.0	3.0	19.89	0	S	4.0	1009.9	80	6.3
216	8	4	Wed	29.5	0.0	26.52	0	S	3.0	1008.2	76	2.8
217	8	5	Thu	29.1	0.0	26.17	0	SSE	2.8	1005.1	74	5.8
218	8	6	Fri	29.1	0.0	24.82	0	SSE	2.9	1004.2	75	4.0
219	8	7	Sat	27.9	2.0	11.43	0	NE	2.5	1003.1	85	9.0
220	8	8	Sun	25.9	90.5	3.43	0	N	3.0	998.0	97	10.0
221	8	9	Mon	28.1	2.0	13.34	0	S	6.1	995.4	84	6.0
222	8	10	Tue	31.0	0.0	22.45	0	SSW	4.7	996.3	58	4.8
223	8	11	Wed	29.2	0.0	21.12	0	SE	2.9	1008.0	61	9.3
224	8	12	Thu	26.0	0.5	8.34	0	SSE	2.4	1008.8	84	9.5
225	8	13	Fri	22.5	20.5	4.36	0	NE	2.7	1008.0	97	10.0
226	8	14	Sat	22.3	77.0	2.76	0	N	2.7	1003.6	100	10.0

- 作成した線形回帰モデルを検討する
  - 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)
- 説明変数と目的変数の関係

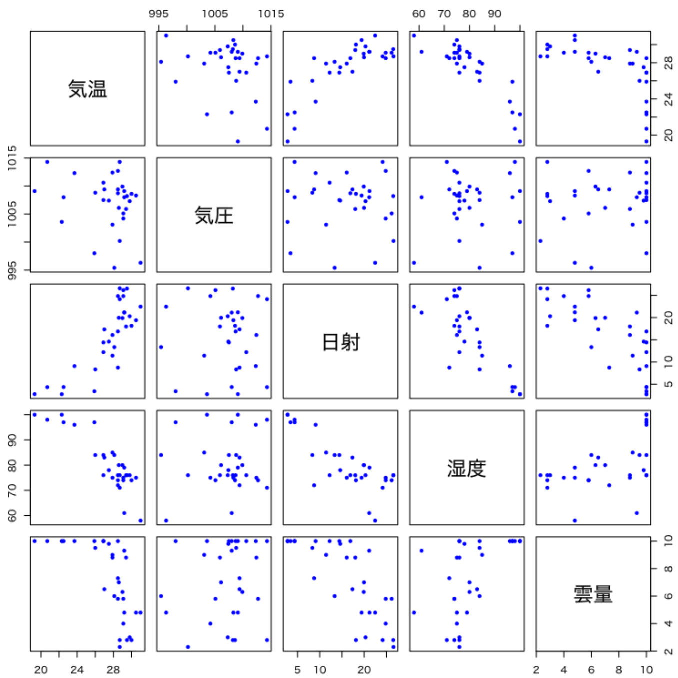

図 1: 説明変数と目的変数の散布図

- 観測値とあてはめ値の比較

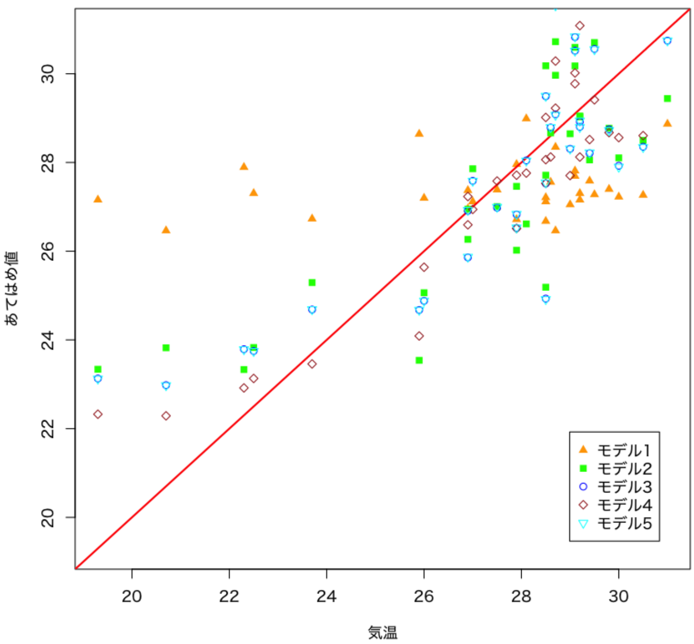

図 2: モデルの比較

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

モデルの評価

- 決定係数
  – モデル 1 : 気温 = F(気圧)
    ```
    [1] "R2: 0.0483 ; adj. R2: 0.0155"
    ```
  – モデル 2 : 気温 = F(日射)
    ```
    [1] "R2: 0.663 ; adj. R2: 0.651"
    ```
  – モデル 3 : 気温 = F(気圧, 日射)
    ```
    [1] "R2: 0.703 ; adj. R2: 0.681"
    ```
  – モデル 4 : 気温 = F(気圧, 日射, 湿度) (3 より改善している)
    ```
    [1] "R2: 0.83 ; adj. R2: 0.811"
    ```
  – モデル 5 : 気温 = F(気圧, 日射, 雲量) (3 より改善していない)
    ```
    [1] "R2: 0.703 ; adj. R2: 0.67"
    ```

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 H_0: $\beta_1 = \cdots = \beta_p = 0$
  - 対立仮説 H_1: $\exists j\ \beta_j \neq 0$ (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

モデルの評価

- 決定係数と F 統計量
  - モデル 1 : 気温 = F(気圧)

    ```
    [1] "R2: 0.0483 ; adj. R2: 0.0155 ; F-stat: 1.47 ; p-val: 0.235"
    ```
  - モデル 2 : 気温 = F(日射)

    ```
    [1] "R2: 0.663 ; adj. R2: 0.651 ; F-stat: 57 ; p-val: 2.52e-08"
    ```
  - モデル 3 : 気温 = F(気圧, 日射)

    ```
    [1] "R2: 0.703 ; adj. R2: 0.681 ; F-stat: 33.1 ; p-val: 4.23e-08"
    ```
  - モデル 4 : 気温 = F(気圧, 日射, 湿度)

    ```
    [1] "R2: 0.83 ; adj. R2: 0.811 ; F-stat: 43.8 ; p-val: 1.65e-10"
    ```
  - モデル 5 : 気温 = F(気圧, 日射, 雲量)

    ```
    [1] "R2: 0.703 ; adj. R2: 0.67 ; F-stat: 21.3 ; p-val: 2.81e-07"
    ```

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する
  - 帰無仮説 H_0: $\beta_j = 0$
  - 対立仮説 H_1: $\beta_j \neq 0$ (β_j は役に立つ)
- t 統計量: 各係数ごと，ζ は $(X^\top X)^{-1}$ の対角成分

$$t = \frac{\hat{\beta}_j}{\hat{\sigma}\zeta_j}$$

- p 値: 自由度 $n-p-1$ の t 分布を用いて計算

モデルの評価

- 回帰係数の推定値と t 統計量
  - モデル 1 : 気温 = F(気圧)

    ```
                 Estimate Std. Error t value Pr(>|t|)
    (Intercept)  162.000      111.00    1.46    0.155
    press         -0.134        0.11   -1.21    0.235
    ```
    - 気圧単体では回帰係数は有意ではない
  - モデル 2 : 気温 = F(日射)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	22.50	0.7210	31.20	7.59e-24
solar	0.31	0.0411	7.55	2.52e-08

    * 日射単体の回帰係数は有意となる

- モデル 3 : 気温 = F(気圧, 日射)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	145.000	63.1000	2.29	2.98e-02
press	-0.121	0.0627	-1.93	6.34e-02
solar	0.308	0.0393	7.85	1.50e-08

    * 気圧は日射と組み合わせることで有意となる

• 回帰係数の推定値と t 統計量 (つづき)

- モデル 4: 気温 = F(気圧, 日射, 湿度)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	147.000	48.7000	3.02	0.005470
press	-0.108	0.0484	-2.24	0.033800
solar	0.134	0.0492	2.73	0.011100
humid	-0.158	0.0353	-4.49	0.000121

- モデル 5: 気温 = F(気圧, 日射, 雲量)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	145.00000	65.0000	2.23	3.42e-02
press	-0.12200	0.0648	-1.88	7.15e-02
solar	0.31000	0.0624	4.97	3.31e-05
cloud	0.00686	0.1710	0.04	9.68e-01

    * このモデルでは雲量は有用でないことが示唆される

モデルの診断 (参考)

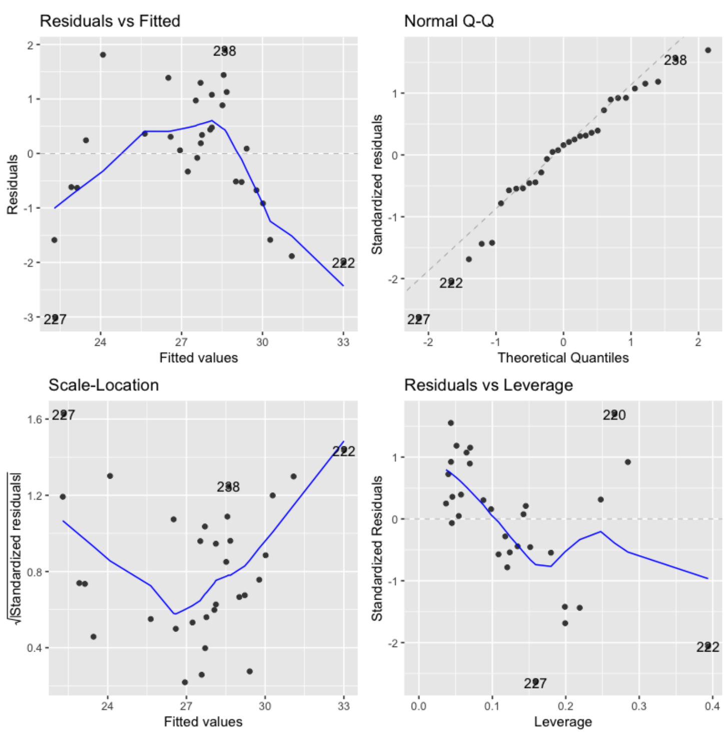

図 3: モデル 4 の診断プロット

回帰モデルによる予測

予測

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

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

- 予測値は元データの目的変数の重み付け線形和

$$\hat{y} = \boldsymbol{w}(\boldsymbol{x})^\top \boldsymbol{y}, \qquad \boldsymbol{w}(\boldsymbol{x})^\top = (1, \boldsymbol{x}^\top)(X^\top X)^{-1} X^\top$$

  - 重みは元データと新規データの説明変数で決定

予測値の性質

- 推定量は以下の性質をもつ多変量正規分布

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

- この性質を利用して以下の 3 つの値の違いを評価

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

  - $\hat{y}$ と y は独立な正規分布に従うことに注意

信頼区間

最適な予測値との差

- 差の分布は以下の平均・分散をもつ正規分布に従う

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

- 正規化による表現

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

信頼区間

- 未知の分散を不偏分散で推定

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

- 確率 α の信頼区間

$$\mathcal{I}_\alpha^c = (\hat{y} - C_\alpha \hat{\sigma}\gamma_c(\boldsymbol{x}),\ \hat{y} + C_\alpha \hat{\sigma}\gamma_c(\boldsymbol{x}))$$

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

  - 最適な予測値 $\tilde{y}$ が入ることが期待される区間

予測区間

観測値との差

- 差の分布は以下の平均・分散をもつ正規分布に従う

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

- 正規化による表現

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

予測区間

- 未知の分散を不偏分散で推定

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

- 確率 α の予測区間

$$\mathcal{I}_\alpha^p = \left(\hat{y} - C_\alpha \hat{\sigma}\gamma_p(\boldsymbol{x}),\ \hat{y} + C_\alpha \hat{\sigma}\gamma_p(\boldsymbol{x})\right)$$

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

  - 観測値 y が入ることが期待される区間
  - $\gamma_p > \gamma_c$ なので信頼区間より広くなる

実習

R：予測値と区間推定

- 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：モデルからの予測

- 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 <- read.csv("data/tokyo_weather.csv")
tw_train <- subset(tw_data, # モデル推定用データ
                   subset= month %in% c(9,10)) # %in% は集合に含むか
tw_test  <- subset(tw_data, # 予測用データ
                   subset= month %in% c(8,11))

tw_model <- temp ~ solar + press # モデルの定義
tw_est <- lm(tw_model, data=tw_train) # モデルの推定
summary(tw_est) # モデルの評価
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)
```

- グラフ表示の例

```
## 予測結果を図示
myColor <- rep("black",12)
myColor[8:11] <- c("red","orange","violet","blue") # 色の定義
with(tw_train,
     plot(temp ~ tw_fit, pch=1, col=myColor[month],
          xlab="fitted", ylab="observed"))
with(tw_test,
     points(temp ~ tw_pred, pch=4, col=myColor[month]))
abline(0,1,col="gray") # 予測が完全に正しい場合のガイド線
legend("bottomright",inset=.05, pch=15, # 凡例の作成
       legend=c("Aug","Sep","Oct","Nov"), col=myColor[8:11])
```

R：区間表示のための関数

- 関数 plotrix::plotCI()

```
library(plotrix) # 標準の plot を拡張した関数が用意されている
plotCI(x,y=NULL,uiw,liw=uiw,ui=NULL,li=NULL,err="y",...)
## x: データの x 座標
## y: データの y 座標
## uiw,liw: 区間の上下の幅
```

```
## ui,li: 区間の上下の値（幅か値のどちらかを指定する）
## err: 区間を付加する座標
```

- 関数 `predict()` の返り値 (lwr/upr) を用いる場合は ui/li を利用すればよい
- 標準誤差の定数倍を計算して用いる場合は uiw/liw を利用すればよい

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 <- read.csv("data/tokyo_weather.csv")
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
```

発展的なモデル

非線形性を含むモデル

- 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

カテゴリカル変数を含むモデル

- 数値ではないデータ
  - 悪性良性
  - 血液型
- 適切な方法で数値に変換して対応:
  - 2 値の場合は 1,0 (真，偽) 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)

実習

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
```

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
- 陽に扱う場合は関数 factor() を利用する

```
## 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 (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
- 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)
  - 上記に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する)

補足

R：モデルの探索

- 変数が増えるとモデルの比較が困難
- 関数 step() を用いて自動化することができる

```
## モデルの探索
adv_data <- read.csv('https://www.statlearning.com/s/Advertising.csv',
                     row.names=1)
summary(lm(sales ~ radio, data=adv_data))
summary(lm(sales ~ TV + radio, data=adv_data))
summary(lm(sales ~ TV + radio + newspaper, data=adv_data))
summary(init <- lm(sales ~ TV * radio * newspaper, data=adv_data))
```

```
opt <- step(init) # step 関数による探索（最大のモデルから削減増加を行う）
summary(opt)
```

  - 最適とは限らないので注意は必要

R : car package

- 回帰モデルの評価
  - 与えられたデータの再現
  - 新しいデータの予測
- モデルの再構築のための視覚化
  - **residual plots**: 説明変数・予測値と残差の関係
  - **marginal-model plots**: 説明変数と目的変数・モデルの関係
  - **added-variable plots**: 説明変数・目的変数をその他の変数で回帰したときの残差の関係
  - **component+residual plots**: 説明変数とそれ以外の説明変数による残差の関係

  などが用意されている

例題

- これまでに用いたデータでモデルを更新して評価してみよう
  - 変数間の線形回帰の関係について仮説を立てる
  - モデルのあてはめを行い評価する
    * 説明力があるのか? (F 統計量，t 統計量，決定係数)
    * 残差に偏りはないか? (様々な診断プロット)
    * 変数間の線形関係は妥当か? (様々な診断プロット)
  - 検討結果を踏まえてモデルを更新する (評価の繰り返し)

次回の予定

- **第 1 回: 主成分分析の考え方**
- 第 2 回: 分析の評価と視覚化