データ解析
第三回「回帰分析」

鈴木　大慈
理学部情報科学科
西八号館 W707 号室
s-taiji@is.titech.ac.jp

今日の講義内容

- 回帰分析 (lm)
- lm 関数返り値の解釈
- 回帰係数の有意性検定
- AIC によるモデル選択

線形回帰モデル

ガウス・マルコフモデル:

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_d x_d + \beta_{d+1} + \epsilon \quad (\epsilon \sim N(0, \sigma^2))$$

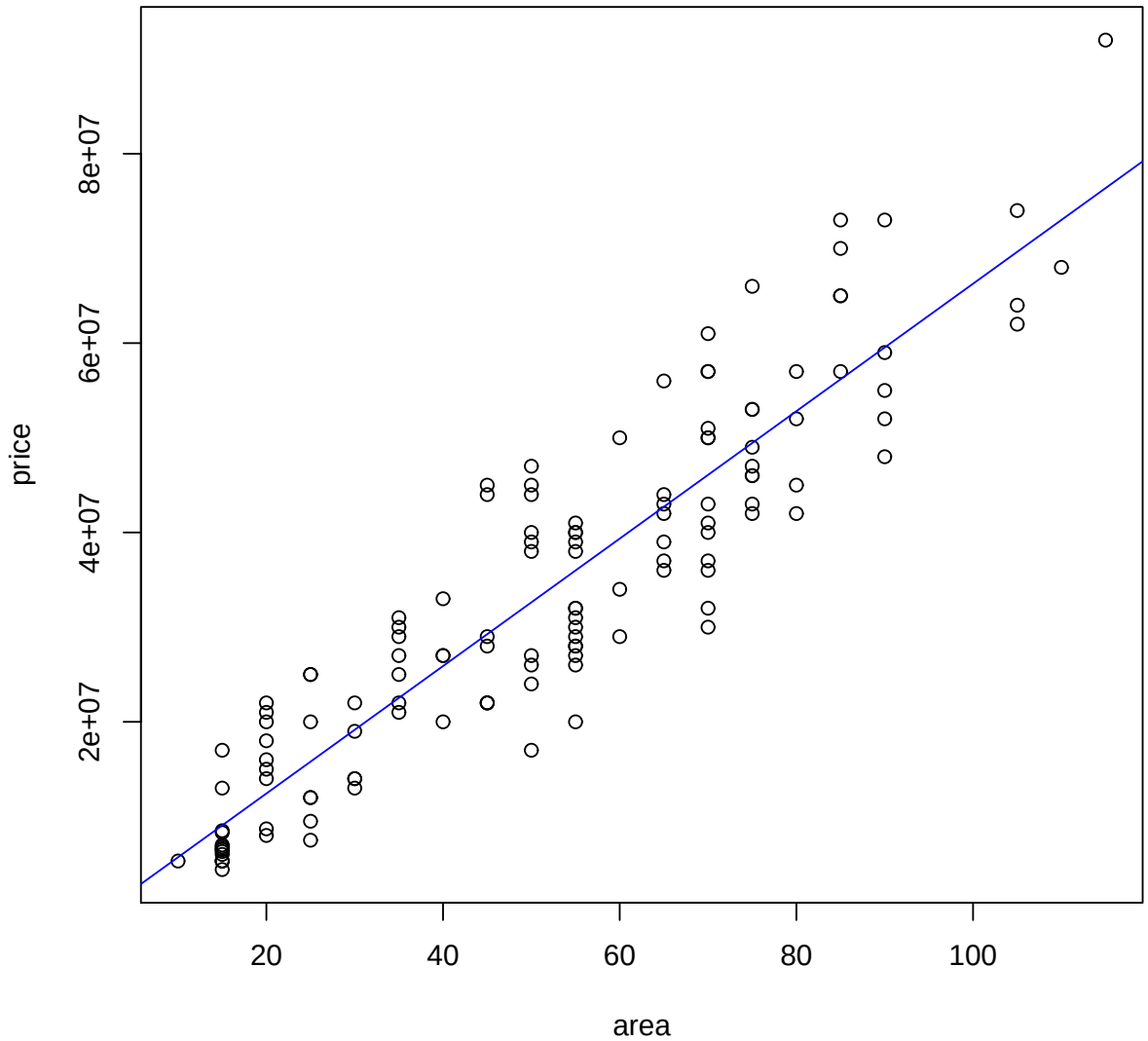
$y \in \mathbb{R}$: 従属変数
$\mathbf{x} \in \mathbb{R}^d$: 説明変数
β_i $(i = 1, \ldots, d)$:（偏）回帰係数, β_{d+1}: 切片 (切片項は省略することもある)

単回帰 $(d = 1)$

$$y = \beta_1 x_1 + \beta_2 + \epsilon$$

重回帰 $(d \geq 1)$

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_d x_d + \beta_{d+1} + \epsilon$$


price
2e+07
4e+07
6e+07
8e+07
20
40
60
80
100
area

最小二乗法

n サンプル (観測値)：$(y_i, \mathbf{x}_i) \in \mathbb{R} \times \mathbb{R}^d$ $(i = 1, \ldots, n)$

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n, \; X = \begin{bmatrix} — \; \mathbf{x}_1^\top \; — & 1 \\ \vdots & \vdots \\ — \; \mathbf{x}_n^\top \; — & 1 \end{bmatrix} \in \mathbb{R}^{n \times (d+1)}, \; \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix} \in \mathbb{R}^n,$$

とおき，$\boldsymbol{\beta}^*$ を真の回帰係数 (これを推定したい) とすると，

$$Y = X\boldsymbol{\beta}^* + \boldsymbol{\epsilon}.$$

最小二乗推定量 (最尤推定量)：

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

- 不偏推定量
- Cramer-Rao の下限を達成 (一様最小分散不偏推定量)

最尤推定量の従う分布

X 固定のもと，$\hat{\boldsymbol{\beta}}$ の分布は次式で与えられる：

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \sim N\left(\mathbf{0}, \sigma^2 \left(\frac{1}{n} X^\top X\right)^{-1}\right).$$

(導出)

$$\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top Y = (X^\top X)^{-1} X^\top (X\boldsymbol{\beta}^* + \boldsymbol{\epsilon})$$

$$\Rightarrow \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = \sqrt{n}(X^\top X)^{-1} X^\top \boldsymbol{\epsilon}. \tag{1}$$

$\epsilon_i \sim N(0, \sigma^2)$ より，右辺は平均 $\mathbf{0}$ の多変量正規分布．その分散共分散行列は $\mathrm{E}[n(X^\top X)^{-1} X^\top \boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top X (X^\top X)^{-1}] = n(X^\top X)^{-1} X^\top \mathrm{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top] X (X^\top X)^{-1} = \sigma^2 \left(\frac{1}{n} X^\top X\right)^{-1}$. □

これより回帰係数の信頼区間が求まるが，σ^2 を知らない場合はそれも推定する必要がある (次ページ).

最尤推定量の信頼区間

$\hat{\epsilon}$ を ϵ の推定量として

$$\hat{\epsilon} = Y - X\hat{\beta} = X(\beta^* - \hat{\beta}) + \epsilon = (I - X(X^\top X)^{-1}X^\top)\epsilon$$

とおく．すると，$S = (\frac{1}{n}X^\top X)^{-1}$ としたとき，

$$\frac{\sqrt{n}(\hat{\beta}_j - \beta_j^*)/\sqrt{S_{jj}}}{\sqrt{\|\hat{\epsilon}\|^2/(n-(d+1))}} \sim t(n-(d+1)) \quad (\text{自由度 } n-(d+1) \text{ の } t \text{ 分布}).$$

これより，β_j の信頼区間や検定が可能になる．

例: $\beta_j^* = 0$ の検定．もし

$$\frac{\sqrt{n}|\hat{\beta}_j|/\sqrt{S_{jj}}}{\sqrt{\|\hat{\epsilon}\|^2/(n-(d+1))}} \geq t_\alpha(n-(d+1))$$

ならば，$\beta_j^* = 0$ なる帰無仮説を棄却する．ただし，t_α は t 分布の上側 α 分位点．
($\beta_j^* = 0$ にしては $\hat{\beta}_j$ が大きすぎ)

導出

1. まず,
$$\hat{\epsilon} = Y - X\hat{\beta} = X(\beta^* - \hat{\beta}) + \epsilon = (I - X(X^\top X)^{-1}X^\top)\epsilon$$
に気をつける．これは平均 $\mathbf{0}$ の多変量正規分布で，かつ $\hat{\beta} - \beta^* = (X^\top X)^{-1}X^\top\epsilon$ (Eq. (1)) と独立である．独立性は，$\mathrm{E}[\hat{\epsilon}(\hat{\beta} - \beta^*)^\top] = O$ であることと，$\hat{\epsilon}$ と $\hat{\beta} - \beta^*$ の同時分布が正規分布であることより，それらは独立であることがわかる (チェックせよ).
2. 次に，$(I - X(X^\top X)^{-1}X^\top)$ は $n - (d+1)$ 次元部分空間への射影なので，$\|\hat{\epsilon}\|^2/\sigma^2$ は自由度 $n - (d+1)$ の χ^2 分布に従う.
3. また，最尤推定量の分布より $\sqrt{n}(\hat{\beta}_j - \beta_j^*)/\sqrt{S_{jj}} \sim N(0, \sigma^2)$ である.
4. よって，$\dfrac{\sqrt{n}(\hat{\beta}_j - \beta_j^*)/\sqrt{S_{jj}}}{\sqrt{\|\hat{\epsilon}\|^2/(n-(d+1))}}$ は互いに独立な正規分布と χ^2 分布の比なので t 分布となる．σ^2 は分母と分子で打ち消す. $\square$

回帰の有意性: F 検定

そもそも回帰に意味がないかもしれない．回帰式に説明力があることを検定.

$$H_0 : \beta_1^* = \beta_2^* = \cdots = \beta_d^* = 0 \quad \text{vs} \quad H_1 : \text{それ以外}$$

(H_0 でも切片の非ゼロは許す)

帰無仮説のもとで,

$$F = \frac{\|X\hat{\beta} - \mathbf{1}\bar{Y}\|^2/d}{\|\hat{\epsilon}\|^2/(n-(d+1))} \sim F(d, n-d-1).$$

ただし，$\bar{Y} = \frac{\sum_{i=1}^n y_i}{n}$.

導出は確率統計第二の講義を参照．また標準的な数理統計の教科書には大体書いてある.

決定係数 (R^2) と自由度調整済み決定係数 (R_{A}^2) :

$$R^2 = 1 - \frac{\|Y - X\hat{\beta}\|^2}{\|Y - \mathbf{1}\bar{Y}\|^2}, \quad R_{\mathrm{A}}^2 = 1 - \frac{\|Y - X\hat{\beta}\|^2/(n-d-1)}{\|Y - \mathbf{1}\bar{Y}\|^2/(n-1)}$$

決定係数が大きければ帰無仮説は棄却され，回帰式には説明力があることになる.

AIC による変数選択

重回帰分析では説明変数を追加するごとに残差は小さくなってゆく.
それでいいのか？
答え：**必ずしも良くはない**.

- 観測済みデータによく当てはまっても，未観測のデータへの当てはまりが悪くなることがある.
- **過適合** (overfit) という.
- 過適合を避けるにはサンプル数に比して複雑なモデルは使うべきではない.

赤池情報量規準，AIC (Akaike Information Criterion)

予測精度が一番良いモデルを選択するための規準
d 次元パラメトリックモデル $\mathcal{M}_d$ の AIC は次式で定義される:

$$\mathrm{AIC} = -2\log(p(\{X_i\}_{i=1}^n|\hat{\theta}_{\mathcal{M}_d})) + 2m.$$

ただし，$\hat{\theta}_{\mathcal{M}_d}$ は $\mathcal{M}_d$ 上の最尤推定量.

言い換えれば

$$\mathrm{AIC} = -2 \times \text{最尤推定量の対数尤度} + 2 \times \text{モデルの次元}$$

である．AIC が最小になるように説明変数の数を選ぶ.

AIC の解釈

- 次元が高い (複雑な) モデルであるほど負の対数尤度が小さくなるが，それにモデルの次元 (複雑さ) を罰則として加えることで過適合を防ぐ.
- モデルの次元とデータへの当てはまりのトレードオフ.
- AIC は真の分布との KL-divergence を漸近展開した 1 次項である.
- 予測誤差 (真との KL-divergence) を最小化．変数を選択するための客観的規準 (予測誤差) を一つ持ってきたところがミソ.

(導出は省略)

中古マンション価格データによる実演

国土交通省が公開している不動産取引価格情報から世田谷区の中古マンション取引価格データ (平成２５年度第３四半期分) を取得．ここから一部を抜粋したデータで回帰分析をやってみる．
http://www.land.mlit.go.jp/webland/download.html

従属変数：価格
説明変数：最寄り駅からの距離（徒歩），延床面積，建物の構造，建ぺい率，容積率，建築年，最寄り駅に急行が止まるか

ダミー変数

説明変数「建物の構造」は「RC」と「SRC」という文字列 (カテゴリカル変数).
これを $\{0, 1\}$ に置き換える．**ダミー変数**という.
RC $\to$ 1, SRC $\to$ 0 とした.

線形回帰分析のための関数

書式:

lm(formula, data, subset, weights, na.action,
method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
singular.ok = TRUE, contrasts = NULL, offset, ...)

引数，大事な部分だけ抜粋

formula	当てはめられるモデルのシンボリックな記述式
data	モデル中の変数を含むオプションのデータフレーム
subset	オプションのベクトルで，当てはめ過程で使われる観測値の部分集合を指定
weights	当てはめ過程で使われるオプションの重みベクトル
na.action	データが NA を含むときそれを処理する関数.

formula の書式

y ~ x	$y = \beta_1 x + b + \epsilon$
y ~ x1 + x2	$y = \beta_1 x1 + \beta_2 x2 + b + \epsilon$
y ~ x1 * x2	交互作用項を含んだモデル $y = \beta_1 x1 + \beta_2 x2 + \beta_3 x1 x2 + b + \epsilon$
y ~ x - 1	切片 (定数) 項を除外する．$y = \beta_1 x + \epsilon$
y ~ 1 + x + I(x^2)	多項式回帰: $y = b + \beta_1 x + \beta_2 x^2 + \epsilon$
y ~ x — z	z で条件付けしたときの y の x への単回帰.
y ~ . , data = データ名	あるデータに目的変数 y と説明変数 $x_1, \ldots, x_d$ が入っている場合，$y = b + \sum_{j=1}^{d} \beta_j x_j + \epsilon$ とする.

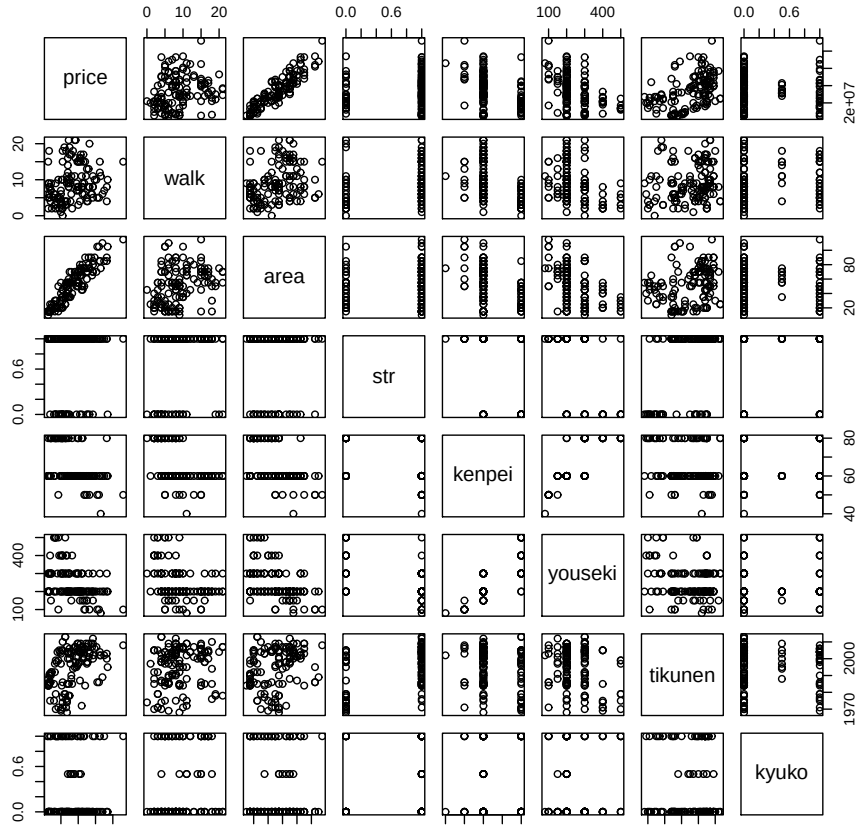
例：

```
x <- seq(-10,10); y<- 3*x + rnorm(21);
lm(y ~ x)  # y を x に回帰

x <- seq(-10,10); z <- seq(-10,10)^2; y<- 3*x + 4*z + rnorm(21);
datain <- data.frame(x=x,z=z,y=y)
lm(y ~ x, data=datain)  # y を x に回帰
lm(y ~ ., data=datain)  # y を y 以外の変数に回帰
```

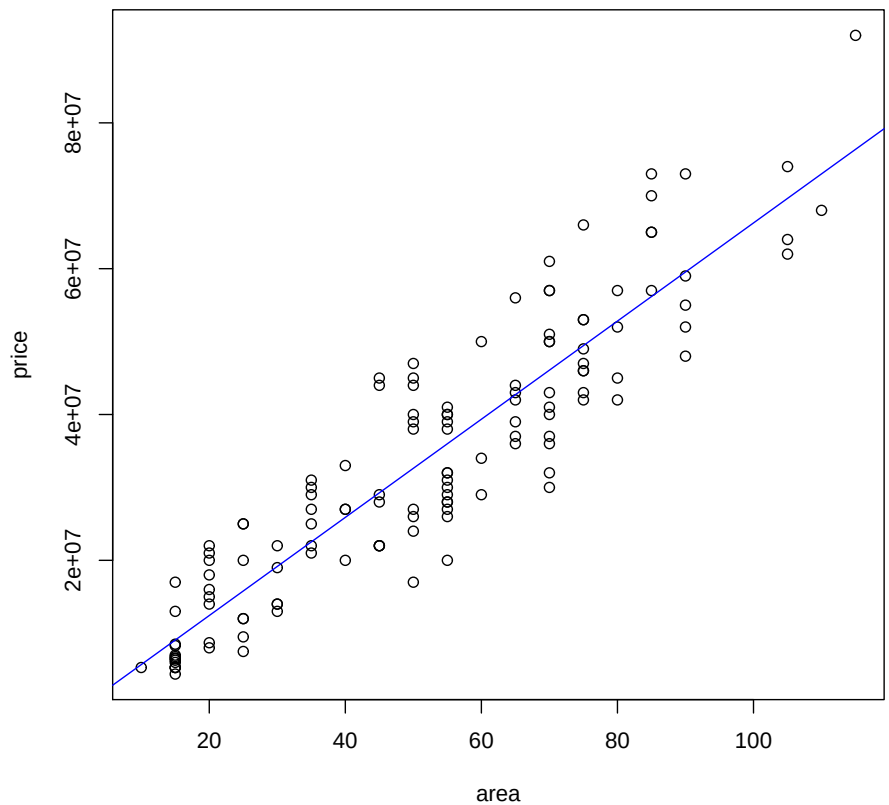
データの取得とプロット

```
x <- read.csv("setagaya_manshion.csv")
sman = data.frame(price=x[[1]],walk=x[[3]],area=x[[5]],
        str=ifelse(x[[6]]=="ＲＣ",1,0),kenpei=x[[7]],youseki=x[[8]],
        tikunen=x[[9]],kyuko=x[[10]])
        #構造はダミー変数 0-1 に置き換える.
plot(sman)    #散布図のプロット
```



回帰分析関数 (lm)

```
sman.lm <- lm(price ~ area,data=sman)  #回帰分析はこの一行で OK！
plot(sman$area,sman$price, xlab="area",ylab="price")  #結果をプロット
abline(sman.lm , lwd=1 , col="blue")
```



分析の要約 (summary)

```
> summary(sman.lm)

Call:
lm(formula = price ~ area, data = sman)

Residuals:
      Min        1Q    Median        3Q       Max
-16082568  -5634345   -789511   4806399  16822060

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1029167    1519818  -0.677      0.5
area          673025      26408  25.485   <2e-16 ***
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’ 1

Residual standard error: 7594000 on 128 degrees of freedom
Multiple R-squared: 0.8354,     Adjusted R-squared: 0.8341
F-statistic: 649.5 on 1 and 128 DF,  p-value: < 2.2e-16
```

要約の意味 (1)

```
Call:
lm(formula = price ~ area, data = sman)
```

呼び出し式

```
Residuals:
      Min         1Q     Median         3Q        Max
-16082568  -5634345    -789511    4806399   16822060
```

残差 $y_i - \hat{y}_i = y_i - \hat{\beta}^\top \mathbf{x}$ の要約統計量

要約の意味 (2)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1029167    1519818  -0.677      0.5
area          673025      26408  25.485   <2e-16 ***
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’
1
```

回帰係数の推定量，標準偏差，t-統計量，p-値
(Intercept) は切片．t-統計量の意味は前のスライドを参照．

ここの p-値がある有意水準より小さければ，その回帰係数が 0 であるという帰無仮説は棄却される．つまり，その変数は y を説明するための情報を持っているということ．

この例だと，切片項は有意でなく，area は有意である (たとえば有意水準 $\alpha = 0.05$ のとき)．

要約の意味 (3)

```
Residual standard error: 7594000 on 128 degrees of freedom
```

残差の標準偏差 (σ の推定量)．この場合，サンプル数 130 で変数は 1 変数+切片なので自由度は 130-2=128.

```
F-statistic: 649.5 on 1 and 128 DF,  p-value: < 2.2e-16
```

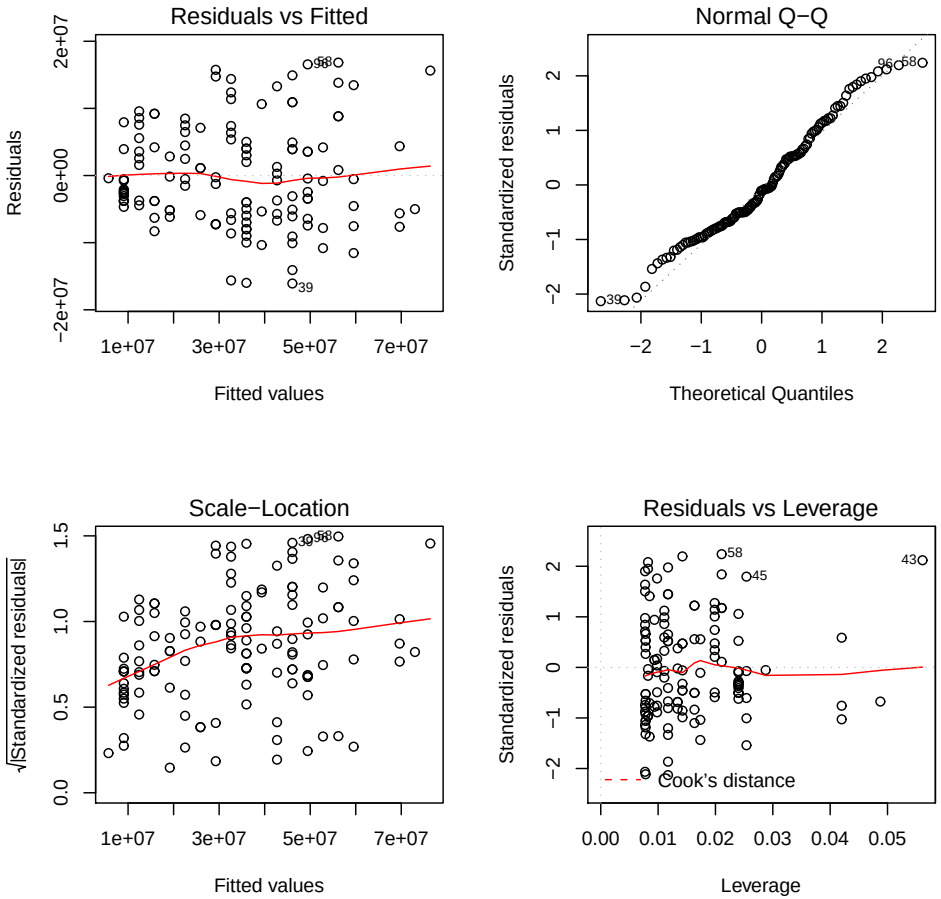
$\boldsymbol{\beta} = \mathbf{0}$ の検定における F 統計量．自由度は $(d, n - d - 1)$ (前のスライドを参照)．この場合は回帰は有意であると言える ($\alpha = 0.05$).

```
Multiple R-squared: 0.8354,     Adjusted R-squared: 0.8341
```

決定係数と自由度調整済み決定係数 (Adjusted R-squared)．Adjusted のほうが重要．大きいほど相関が強い.

回帰分析の診断

```
par(mfrow=c(2,2)) #一つの Figure に 2x2 の図を出力
plot(sman.lm)
```



(上左) 残差プロット:
予測値 $\hat{y}_i$ vs 残差 $y_i - \hat{y}_i$

(上右) 残差の Q-Q プロット:
正規分布との Q-Q プロット．対角線に近ければより正規分布に近い．残差の正規性はガウスマルコフモデルの仮定．

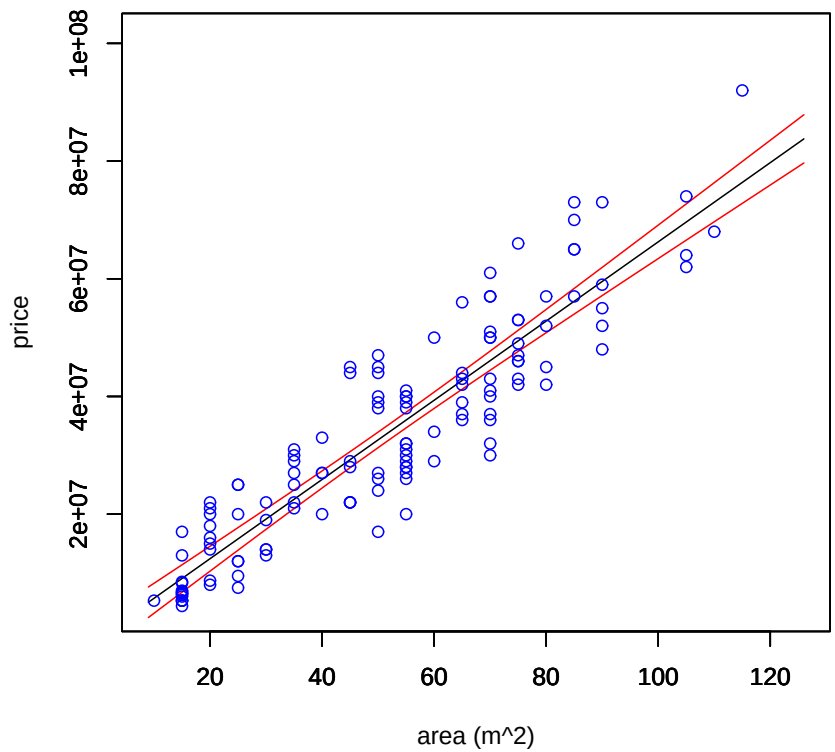
(下左) Scale-Location プロット:
規準化した残差の絶対値の平方根．この値は予測値の値に対して大体一定であるべき．そうでない場合，場所によってノイズの分散が異なることになる．

(下右) Cook 距離のプロット:
個々のデータが推定に与える影響を表した距離 (そのデータがない場合とある場合の予測値の変化量)．大きいと外れ値の可能性がある．0.5 を超えると「大きい」とされる．

予測値の信頼区間

predict 関数に interval="confidence" なるオプションを入れることで予測値 $(\hat{\beta}^\top X)$ の信頼区間を得ることができる.

```
area.plot <- seq(min(sman$area)*0.9,max(sman$area)*1.1,by=1) #信頼区間を計算する範囲の設定
sman.con <- predict(sman.lm, data.frame(area=area.plot),interval="confidence") #信頼区間
```



重回帰分析

築年数も含めて分析してみる.

```
> sman.lm3 <- lm(price ~ area + tikunen, data=sman)
> (AIC(sman.lm3))
[1] 4443.881
```

AIC(·) で AIC が計算できる.

すべての説明変数を用いて回帰分析

```
sman.lmall <- lm(price ~., data=sman)
sman.lmAIC <- step(sman.lmall)
summary(sman.lmAIC)
```

step(·) で AIC 最小のモデル (説明変数の組) を探索. walk + area + tikunen の三変数モデルが採用された.

正規性の検定

最後に残差の正規性を検定．正規性は棄却されない．

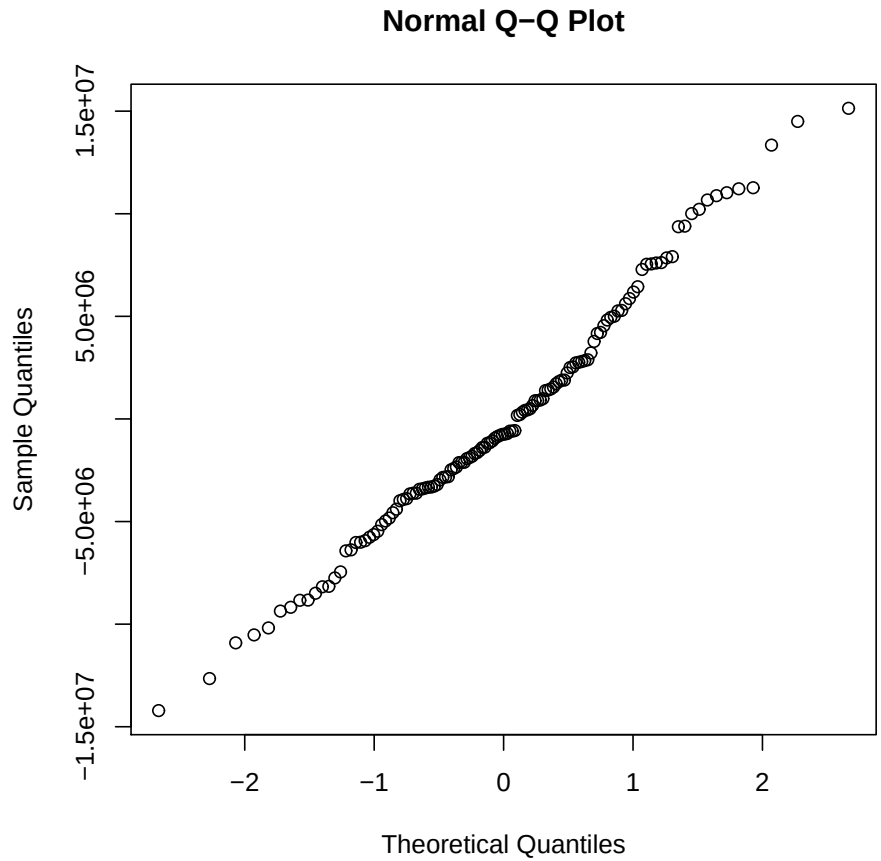
```
> shapiro.test(sman.lmAIC$residuals)

        Shapiro-Wilk normality test

data:  sman.lmAIC$residuals
W = 0.9874, p-value = 0.2806
```

残差の QQ プロット

```
qqnorm(sman.lmAIC$residual)
```



QQ プロットからも残差の分布が正規分布に近いことが確認できる.