

## R と回帰分析

### 1. 統計モデルと回帰分析

人間のみならず多くの動物は、学習機能を持っており、過去の経験から得られた知識・規則を今後の行動に生かしている。データ解析では、既知のデータから規則を導き出し、その規則を用いて未知の部分の説明したり、予測・推測したりする。例えば、体重(kg)と身長(m)のデータを用いた肥満度の指標 BMI(Body Mass Index)を考えよう。

$$\text{BMI} = \frac{\text{体重}}{\text{身長}^2}$$

この式は既知のデータと医学の知識から導いた関係式である。通常 BMI の値が 22 前後であれば標準であり、25 前後であると肥満傾向で、35 を超えると極度肥満であるといわれている。このようなデータ間の関係の規則を関数式で表すことができると、その数式を用いて未知なことを予測・推測することが可能である。このようなデータから導いた規則をここでは統計モデルと呼ぶ。

図 1 に平成 14 年度に総務省が発行した「情報通信白書」の中の、わが国のインターネットに関する報告を示す。横軸が時間軸で、縦軸がその年度のインターネットの利用者数である。図の右側には平成 17 年の予測値が示されている。このような時間の前後の順に並べた時系列データでは、しばしば現在までのデータの統計モデルを用いて今後を予測・推測する。

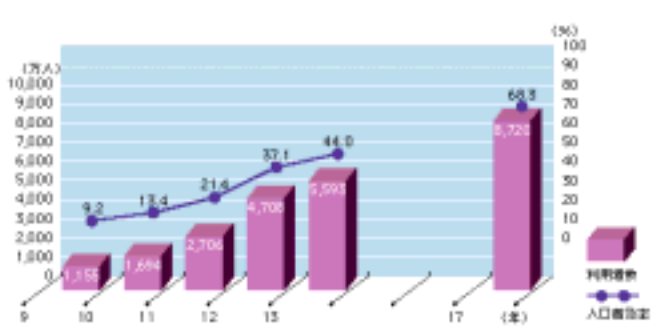


図 1 平成 14 年度の「情報通信白書」での予測の例 (平成 14 年度の「情報通信白書」より)

統計モデルの説明のため、身近な例として、身長と体重との関係を何らかの式で表すことを考えることにする。その具体的な関係がわからないので、抽象的な関数記号  $f$  で次のように表すことにする。

$$\text{体重} = f(\text{身長})$$

この式の中の「身長」を**説明変数**、「体重」を**被説明変数**、あるいは**目的変数**と呼ぶ。もし説明変数を  $x$ 、被説明変数を  $y$  で表すと上記の式は次のような式になる。

$$\begin{array}{c} \text{被説明変数} = f(\text{説明変数}) \\ \swarrow \quad \searrow \\ y = f(x) \end{array}$$

このような説明変数を用いて被説明変数を説明する関係を求める統計モデルを通常**回帰分析**と呼ぶ。説明変数が1つの場合は**単回帰分析**、説明変数が複数の場合は**重回帰分析**と呼ぶ。

図2に単回帰分析の例として2種類の説明変数と被説明変数の関係を示す。横軸が説明変数、縦軸が被説明変数である。

図2(a)では、説明変数と被説明変数との関係を直線で大まかな傾向を示している。このような直線関係でモデル化する回帰分析を**線形回帰分析**と呼ぶ。

図2(b)の説明変数と被説明変数のような非直線関係でモデル化する回帰分析を**非線形回帰分析**と呼ぶ。

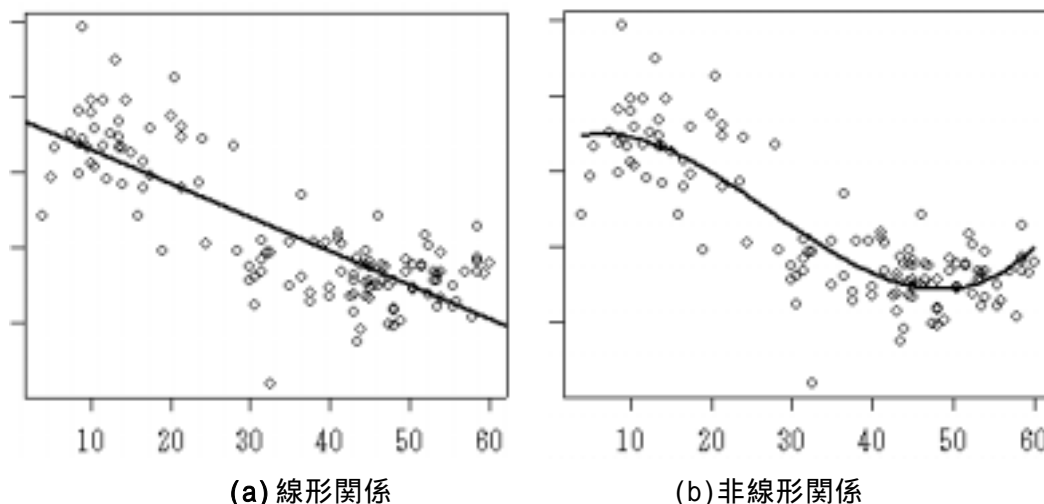


図2 説明変数と被説明変数との関係

## 2. 単回帰分析

単回帰分析は説明変数が 1 つである回帰分析であり、略して単回帰と呼ぶ。単回帰でも線形と非線形に分けられるが、特別な説明がない限り線形回帰を指す。

ここで言う線形とは、説明変数  $x$  と被説明変数  $y$  の関係を式

$$y = a + bx$$

で表すことが可能であることである。式の中の  $a$  を切片、 $b$  は直線の傾きに関する値で、説明変数  $x$  の係数と呼ぶ。

例えば、表 1 のような体重、身長データがあるとする。

表 1 体重、身長のデータ

身長	体重
165	50
170	60
172	65
175	65
170	70
172	75
183	80
187	85
180	90
185	95

まずデータセットを次のように作成する。

```
>taikei=matrix(0,10,2)
>taikei[,1]<-c(165,170,172,175,170,172,183,187,180,185)
>taikei[,2]<-c(50,60,65,65,70,75,80,85,90,95)
> colnames(taikei)<-c("身長","体重")
> taikei
      身長  体重
[1,]  165   50
<後略>
```

コマンド `plot(taikei)` で図 3 のような直線がない散布図を作成することができる。図 3 に身長を横軸、体重を縦軸にした散布図を示す。図 3 の直線は、散布図の点の傾向(トレンド)を表すもので、回帰直線と呼ぶ。

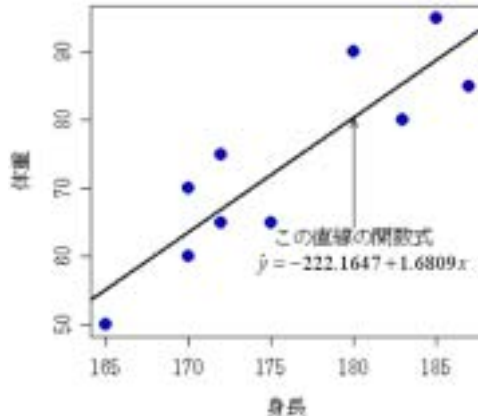


図 3 散布図と回帰直線

図 3 の中の直線の関数

$$\text{体重} = -222.1647 + 1.6809 \times \text{身長}$$

を線形回帰式、あるいは回帰直線と呼び、式の中の切片  $a$  は  $-222.647$  で、説明変数  $x$  の係数  $b$  は  $1.6809$  である。回帰直線は、適当に描いたものではなく、既知のデータに最もフィット (fit) が良いと考えられる直線である。回帰分析は最適と思われる回帰式の係数を求めるのが一つの主な目的である。R で線形回帰分析を行う関数の中で最も広く使用されているのは `lm` である。

`lm(formula, data, weights, subset, na.action)`

各引数の意味を簡潔に次に示す。

`formula` は回帰式に用いる被説明変数と説明変数、定数項を用いるか用いないかなどのモデルの形式である。例えば、回帰式を  $y = a_0 + a_1x$  にしたいときには

`formula` の部分は `y~x` と指定し、定数項を用いない回帰式  $y = a_1x$  にしたいときには `formula` の部分は `y~-1+x` (あるいは `y~x-1`) と指定する。

`data` は回帰分析に用いたデータセットの名前である。例えば、`data=taikei`、あるいは `data=` を省略して `taikei` のみでもよい。

`wights` は用いる説明変数に重みをつける引数である。初心者は無視してもよい。特に設定しない場合は重みを付けない。

`subset` はデータセットの中の一部を用いる際に、用いる部分を明示するための添字ベクトルの引数である。指定しない場合は、全てのデータを用いる。

`na.action` は欠損値扱いを指定する引数である。指定がない場合は、欠損値のデータを除いたデータを用いて計算を行う。よって、`na.action` を指定しない場合は、予測値や残差の数はデータセットから欠損値を除いた数と等しい。

関数 `lm` に用いるデータ形式はデータフレームである。マトリックス形式からデータフレーム形式には次のように関数 `data.frame` を用いて返還する。

```
>taikei.F<-data.frame(taikei)
```

`taikei.F` を用いた線形回帰関数 `lm` の使用例を示す。

```
>taikei.lm<-lm(体重~身長,data=taikei.F)
```

関数 `lm` 計算された結果の表示及び回帰分析に関する関連操作行うコマンドを表 2 に示す。

表 2 回帰分析の関連操作を行うコマンド

コマンド	内 容	使 用 例
<code>print</code>	要約より簡単な結果	<code>print(taikei.lm)</code>
<code>summary</code>	回帰分析結果の要約	<code>summary(taikei.lm)</code>
<code>coef</code>	回帰係数	<code>cofe(taikei.lm)</code> 、 <code>taikei.lm\$coef</code>
<code>fitted</code>	用いたデータの予測値	<code>fitted(taikei.lm)</code> 、 <code>taikei.lm\$fitted</code>
<code>deviance</code>	残差の平方和	<code>deviance(taikei.lm)</code>
<code>anova</code>	回帰係数の分散分析	<code>anova(taikei.lm)</code>
<code>predict</code>	新たなデータに対する予測値	<code>predict(taikei.lm)</code>
<code>plot</code>	回帰診断プロット	<code>plot(taikei.lm)</code>

次に関数 `summary` による結果について説明する。

```
>summary(taikei.lm)
```

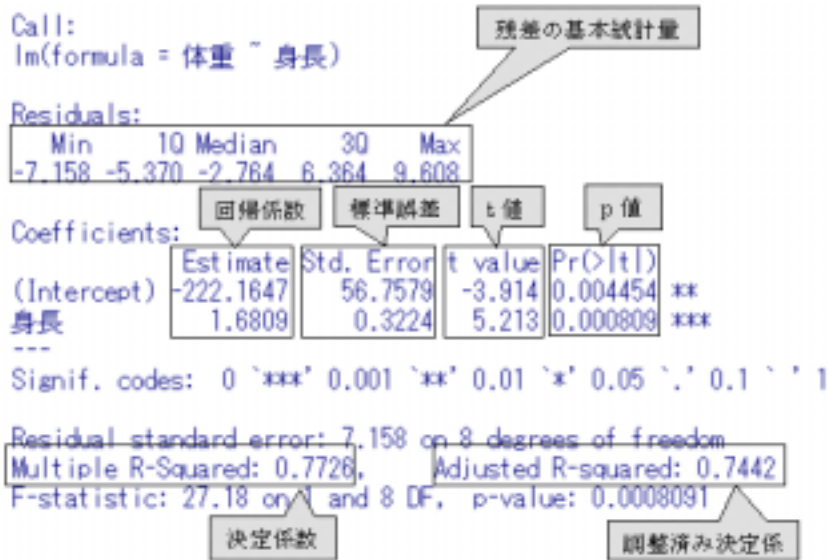


図4 回帰分析の主な結果

まず関数 `summary` が返した結果について順を追って説明を行い、次に予測値の求め方やグラフによる視覚的な考察などについて説明する。

## (1) 残差

データ(真のモデル)を

$$y = a + bx + e$$

回帰直線(推測モデル)を

$$\hat{y} = a + bx$$

としたとき、説明変数  $x$  が  $x_i$  を取った場合の両値の差  $e_i = y_i - \hat{y}_i$  を残差(residuals)と呼ぶ。図4にデータ、回帰直線、残差を示す。

$$\begin{cases} y_i = a + bx_i + e_i \\ \hat{y}_i = a + bx_i \end{cases}$$

$$i = 1, 2, 3, \dots, n$$

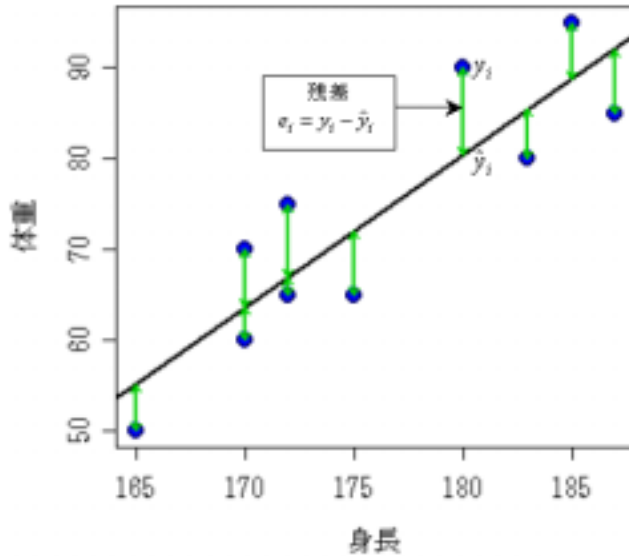


図5 回帰直線と残差

`summary(takei.lm)`の **Residuals** では残差の最小値、第1四分位数、中央値、第2四分位数、最大値が返される。次のコマンドで全ての残差を返すことができる。

```
>residuals(takei.lm)
#あるいは
>takei.lm$residuals
```

## (2) 回帰式と回帰係数

回帰直線  $\hat{y} = a + bx$  の係数(Coefficients)である  $a$ 、 $b$  は、残差の2乗の和

$$S_e = \sum e_i^2 = \sum (y_i - a_0 - a_1 x)^2$$

を最小化する連立方程式を解くことで次の解が得られる。

$$\begin{cases} b = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \\ a = \bar{y} - b\bar{x} \end{cases}$$

上記の `summary(taikei.lm)` で返されている結果の `Coefficients` の部分が回帰係数及びそれに関連する情報である。Intercept の行が切片  $a$  の推測値、その標準誤差、 $t$  値、 $P$  値である。その次の行が説明変数「身長」の係数と関連の情報である。回帰式は返されている回帰係数を用いて次のように構築する。

$$\text{体重} = -222.1647 + 1.6809 \times \text{身長}$$

回帰係数の標準誤差、 $t$  値は次の式で定義されている。

$$\text{残差の平方和} : S_e = \sum (y_i - \hat{y}_i)^2$$

$$\text{残差の不偏分散} : s_e^2 = \frac{S_e}{n - k - 1}$$

式中の  $n$  はデータの標本数、 $k$  は説明変数の数である。

$$\text{係数 } a \text{ の標準誤差} : SE(a) = \sqrt{s_e^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right]}$$

$$\text{係数 } b \text{ の標準誤差} : SE(b) = \sqrt{\frac{s_e^2}{\sum (x_i - \bar{x})^2}}$$

$$\text{係数 } a \text{ の } t \text{ 値} : t_a = \frac{a}{SE(a)}$$

$$\text{係数 } b \text{ の } t \text{ 値} : t_b = \frac{b}{SE(b)}$$

$t$  値に対応する  $P$  値は関数 `pt` で求めることが可能である。

この  $t$  値は「回帰係数がゼロである」という仮説検定の統計量である。 $P$  値が通常よく使われている有意水準 0.05(5%)、0.1(10%)、0.05(0.5%)より小さいときには、出力結果の  $P$  値の右にそれぞれ 1 つの星 "\*", 2 つの星 "\*\*", 3 つの星 "\*\*\*" で印をつける。

回帰係数は次のコマンドで返される。

```
> coefficients(taikei.lm)#あるいは
> taikei.lm$coefficients
```

### (3) 決定係数

回帰直線がどの程度データにフィットしているかを評価する指標として決定係数(coefficient of determination) がある。決定係数を通常  $R^2$  で示す。関数 `lm` では決定係数(Multiple R-Squared)と調整済み決定係数(Adjusted R-squared)が返される。決定係数と調整済み決定係数が 1 に近づくほど回帰直線がデータによくフィットしていると判断する。

決定係数と調整済み決定係数は次の式で定義されている。式の中の  $n$  は標本の数で、この  $k$  は説明変数の数である。単回帰では説明変数が 1 つであるので  $k=1$  である。

$$\text{決定係数: } R^2 = \frac{S_R}{S_T}$$

$$\text{調整済み決定係数: } \tilde{R}^2 = 1 - \frac{S_e / (n - k - 1)}{S_T / (n - 1)}$$

式の中の  $S_T$ 、 $S_R$ 、 $S_e$  は次のように定義されている。

$$y \text{ 偏差の平方和: } S_T = \sum (y_i - \bar{y})^2$$

$$\hat{y}_i \text{ 偏差の平方和: } S_R = \sum (\hat{y}_i - \bar{y})^2$$

$$\text{残差の平方和: } S_e = \sum (y_i - \hat{y}_i)^2$$

$$\text{平方和の関係: } S_T = S_R + S_e$$

関数 `lm` が返す結果の中の決定係数、調整済み決定係数の下部に  $F$  値とその  $P$  値が返される。 $F$  値と  $P$  値は「全ての回帰係数がゼロである」という帰無仮説の検定等計量である。 $F$  値は決定係数から求めることができ、この  $F$  値は自由度  $k$ 、 $n-k-1$  の  $F$  分布に従う。

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

### (4) 予測値

回帰式で計算される値を予測値、あるいは推測値と言う。関数 `summary` では予測値は返さない。予測値を返すためには `R` では関数 `predict` を用いる。考察の便利

のため回帰分析に用いたデータ、予測値、残差を次のように一覧表にして示す。

```
>予測値<-predict(taikei.lm)
>残差<-residuals(taikei.lm)
> data.frame(taikei.F,予測値,残差)
  身長 体重  予測値    残差
1  165  50  55.17854 -5.178535
2  170  60  63.58288 -3.582877
< 中略 >
9  180  90  80.39156  9.608440
10 185  95  88.79590  6.204098
```

## (5) グラフによる分析

単回帰の場合は、説明変数と非説明変数の散布図に求めた回帰直線を加えることでデータの傾向を概観することができる。次にそのコマンドと結果を示す。

```
>plot(taikei.F)
>abline(taikei.lm)
```

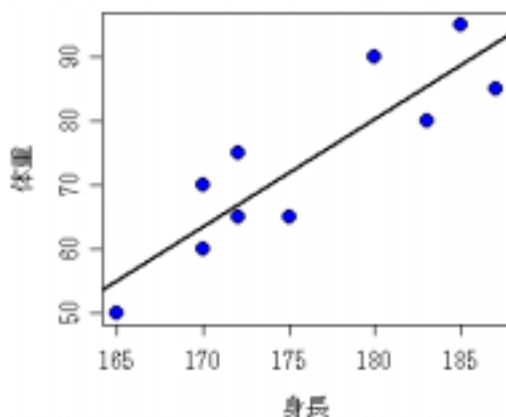


図6 データの散布図と回帰直線

また回帰分析ではしばしば残差を視覚的に分析を行う。R では回帰分析の残差などを視覚的に考察するための環境が用意されている。回帰分析の結果を次のコマンドのように関数 `plot` に代入するだけで図7に示す4つの異なるグラフが返される。

関数 `par` は作図の環境設定関数である。用いた引数 `mfrow=c(2,2)` は 1 画面に 2 行 2 列の図を作成する指定である。回帰分析の結果を関数 `plot` で作成した図を回帰診断図と呼ぶ。

```
>par(mfrow=c(2,2)) #par(mfrow=c(2,2),oma=c(2,2,2,1),mar=c(5,4,3,2))
> plot(taikei.lm) # plot(taikei.lm,cex=1.5,pch=21,bg="blue",col="blue")
```

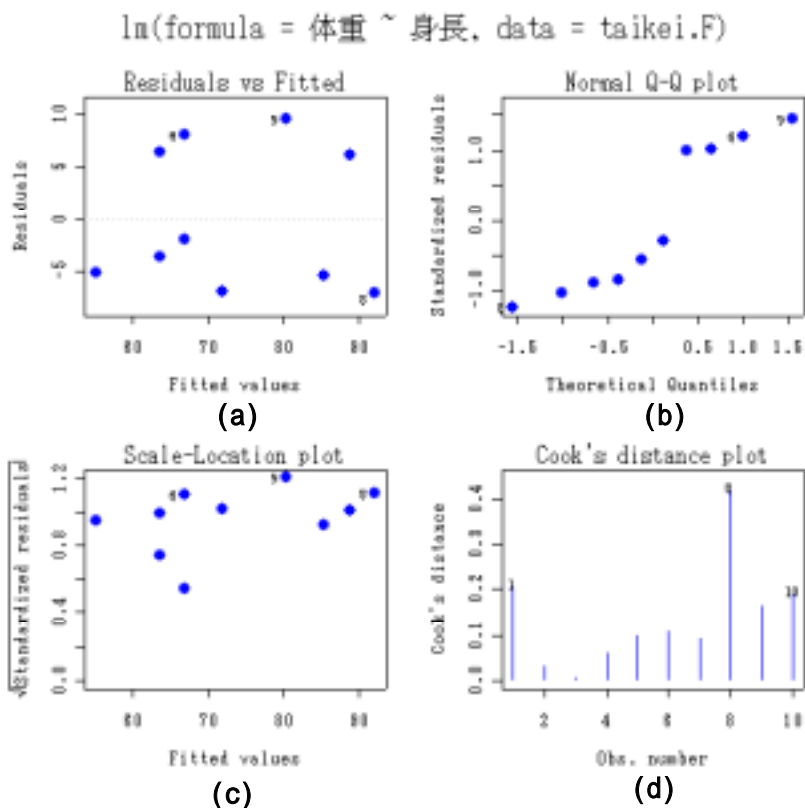


図 7 回帰診断図

### 残差とフィット値のプロット

図 7 (a) は残差とフィット値の散布図である。ここで言うフィット値は予測値である。図の横軸が予測値、縦軸が残差となっている。図から残差の全体像を概観することができる。この図では個体 6、8、9 の残差が相対的に大きいことがわかる。

### 残差の正規 Q-Q プロット

正規 Q-Q プロットはデータの正規性を考察するためにデータを視覚化する方法である。R では関数 `qqplot` を用いて Q-Q プロットを作成することができる。デー

タが正規分布に従うと、点が直線上に並べられる。通常の回帰分析では、残差が標準正規分布に従うという仮定の下で行っている。

図 7(b)は標準化した残差の Q-Q プロットである。ここで用いた例では標本データの数が少ないのでその正規性に関する議論には大きな意味がない。

### 残差の平方根プロット

図 7(c)は標準化した残差の絶対値の平方根を縦軸にし、予測値を横軸とした散布図である。この図の目的も残差の変動状況を考察することである。図から個体 6、8、9 の変動が相対的に大きいことが読みとられる。

### Cook の距離のプロット

Cook の距離は一種の距離の測度であり、R には関数 `cooks.distance` が用意されている。Cook の距離は回帰分析における影響度が大きいデータの検出などに多く用いられている。Cook の距離は全てデータを用いた場合と 1 つのデータを除いた後求めた回帰式による予測値を用いた場合との食い違いに関する距離の測度である。Cook の距離が大きいとそのデータが回帰式による予測値に大きく影響していることを意味する。よって、Cook の距離が大きいデータは異常値である可能性がある。Cook の距離が 0.5 以上であれば大きいと言われている。図 7(d)では個体 8 の Cook の距離が比較的大きいのが 0.5 以下であるので個体 8 が異常値であるとはいえない。

診断図を作成する `plot` では `which=n` の引数を用いて、4 種類の図の中の 1 つのみを作成することができる。この `n` は 1、2、3、4 でそれぞれ上記の 、 、 、

に対応する。例えば `plot(taikei.lm,which=1)` であれば残差対予測値の散布図が作成される。

回帰診断図をさらに加工したいときには、次のように個別のデータを求めて散布図を作成したほうが便利である。

残差対予測値の散布図

```
plot(resid(taikei.lm)) #plot(taikei.lm$resid)
abline(h=0)
```

Q-Q プロット

```
qqnorm(resid(taikei.lm)) #qqnorm(taikei.lm$resid)
qqline(resid(taikei.lm)) #qqline(taikei.lm$resid)
```

Cook の距離

```
plot(cooks.distance(taikei.lm))
```