医学統計勉強会

東北大学病院循環器内科・東北大学臨床研究推進センター　共催

東北大学大学院医学系研究科EBM開発学寄附講座

宮田　敏

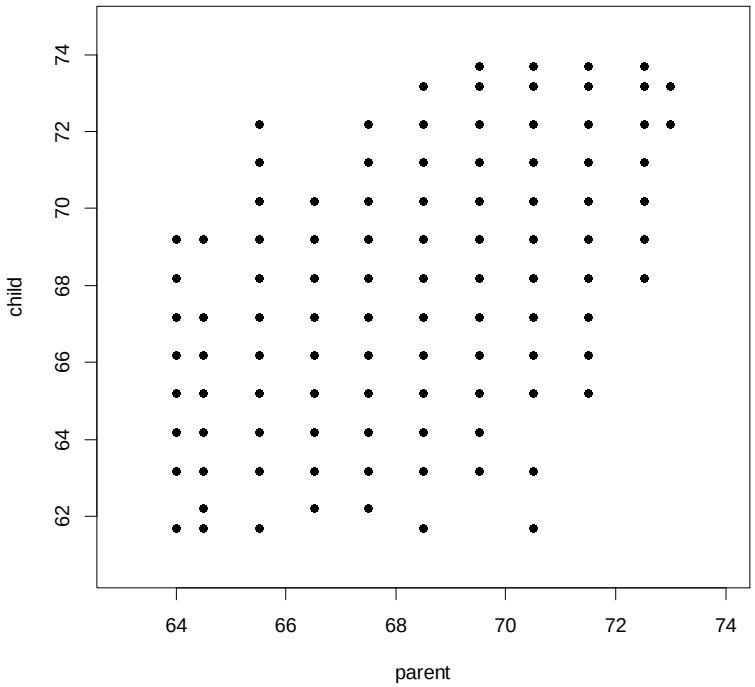
回帰分析

- 回帰分析 (regression analysis) は，一つの連続数（実数）の値を複数の変数によって説明，予測する多変量解析 (multivariate analysis) の一つ．

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

Y: response variable, 従属変数，被説明変数
$x_1, \ldots, x_k$: independent variable, 独立変数，説明変数
$\beta_0, \beta_1, \ldots, \beta_k$: regression coefficient, 回帰係数
ε: error term, 撹乱項、誤差項

Example 1: 親子の身長



205組の夫婦から生まれた，928人の成人した子供の身長（インチ）．

child: 子の身長
parent: 両親の身長の平均

データは0.1インチ刻みに丸められている．そのためデータ点が重なっている．

Galton, F. (1886). Regression Towards Mediocrity in Hereditary Stature
Journal of the Anthropological Institute, 15, 246-263

データの要約

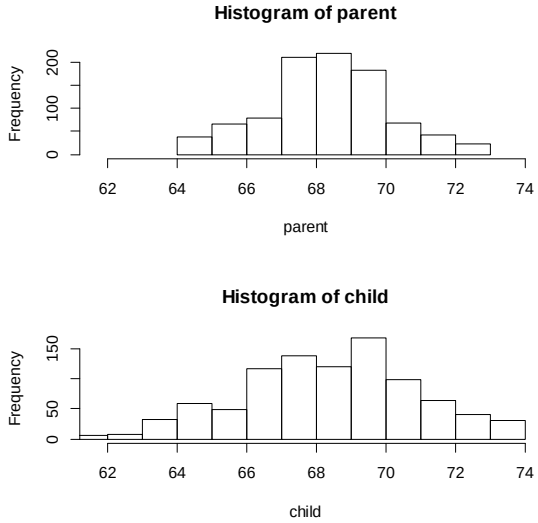
データが手に入ったら，まずデータを要約して，その傾向，特徴を把握する．

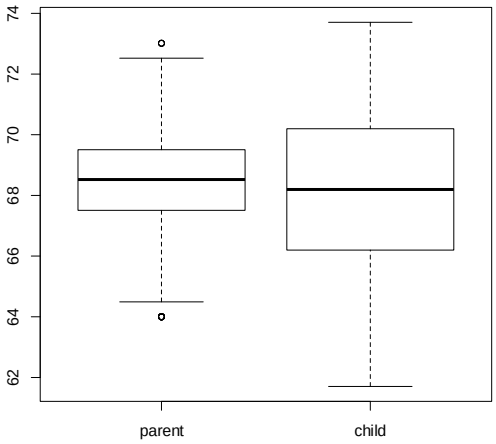
個々の変数の数量的要約：Galtonの親子の身長

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.		SD	IQR
parent	64	67.50	68.5	68.31	69.5	73		1.79	2
child	61.7	66.20	68.2	68.09	70.2	73.7		2.52	4

- 平均，中央値はほぼ等しく，分布の位置は同じ．
- 標準偏差（SD），四分位点間距離（IQR）は child の方が大きく，子の分布の散らばりの方が大きい．

個々の変数の視覚的要約：Galtonの親子の身長





グラフを比較するときは，軸を揃えることがコツ.

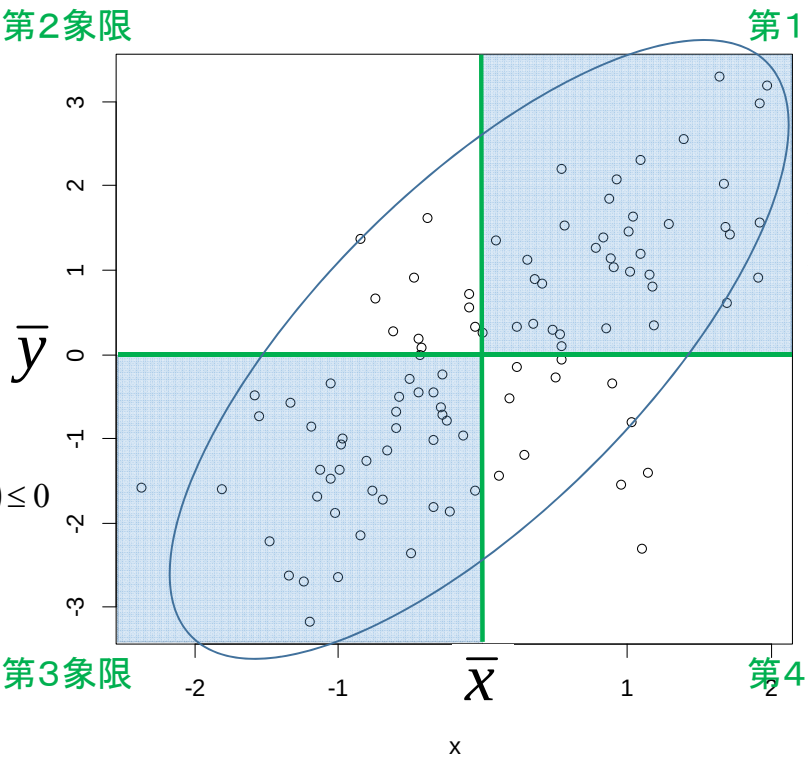
標本共分散

二つの変数 x と y が互いに影響し合っているとき，
x と y が如何に強く関係しあっているか知りたい.
このとき x と yの共分散(covariance)を以下で定義する.

$$\mathrm{Cov}(X,Y)=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})$$

x と yの間に正の(負の)相関があるとき、$\mathrm{Cov}(X, Y)$はそれぞれ正(負)になる.

正の相関の場合



$x_i \le \bar{x} \,\&\, y_i \le \bar{y}$
$\Leftrightarrow (x_i - \bar{x}) \le 0 \,\&\, (y_i - \bar{y}) \le 0$
$\Leftrightarrow (x_i - \bar{x})(y_i - \bar{y}) \ge 0$

$x_i \ge \bar{x} \,\&\, y_i \ge \bar{y}$
$\Leftrightarrow (x_i - \bar{x}) \ge 0 \,\&\, (y_i - \bar{y}) \ge 0$
$\Leftrightarrow (x_i - \bar{x})(y_i - \bar{y}) \ge 0$

正の相関の場合，
データの大部分は
第1，3象限にある。

$(x_i - \bar{x})(y_i - \bar{y}) \ge 0$

共分散

$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

は正になる。

標本相関係数

二つの変数xとyの間の相関係数 (correlation coefficient)を，以下で定義する．相関係数はxとyの線形関係の強さを測る量である．

$$\mathrm{Corr}(X,Y) = r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

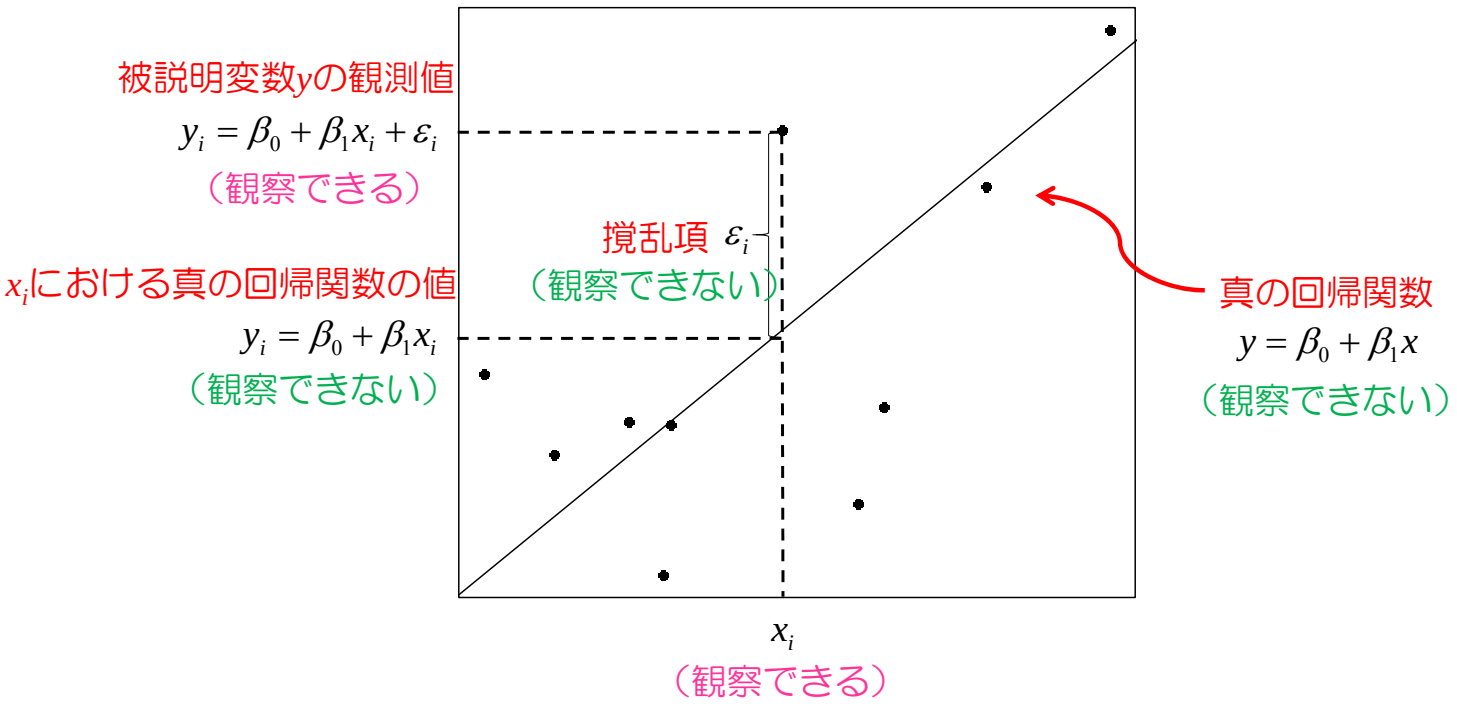
- 相関係数rはxとyの線形関係の強さを測る．
- $-1 \le r \le 1$
- $r = +(-)\,1$：正(負)の完全な相関，線形関係

回帰分析

二つの変数xとyの関係が線形（直線）で近似できるとする．このときxとyの関係を以下の回帰式 (regression equation) でモデル化する．

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Y: response variable, 従属変数， 被説明変数
x: independent variable, 独立変数， 説明変数
β_0, β_1: regression coefficient, 回帰係数
ε: error term, 撹乱項、 誤差項



回帰分析のモデルの仮定

- 線形性 (Linearity)：$Y = \beta_0 + \beta_1 x + \varepsilon$
  被説明変数yと説明変数xの関係は直線で近似できる.
- 独立性 (Independence) $\{(x_i, Y_i)\}_{i=1}^n$ は互いに独立である.
  あるサンプルの値が他のサンプルの値に影響しない.
- 正規性 (Normality)：$\varepsilon_i \sim N(0, \sigma^2)$, iid
  撹乱項 ε_iは正規分布に従う.
- 等分散性 (homoskedasticity)：σ^2.　分散一定

正規分布：Normal distribution

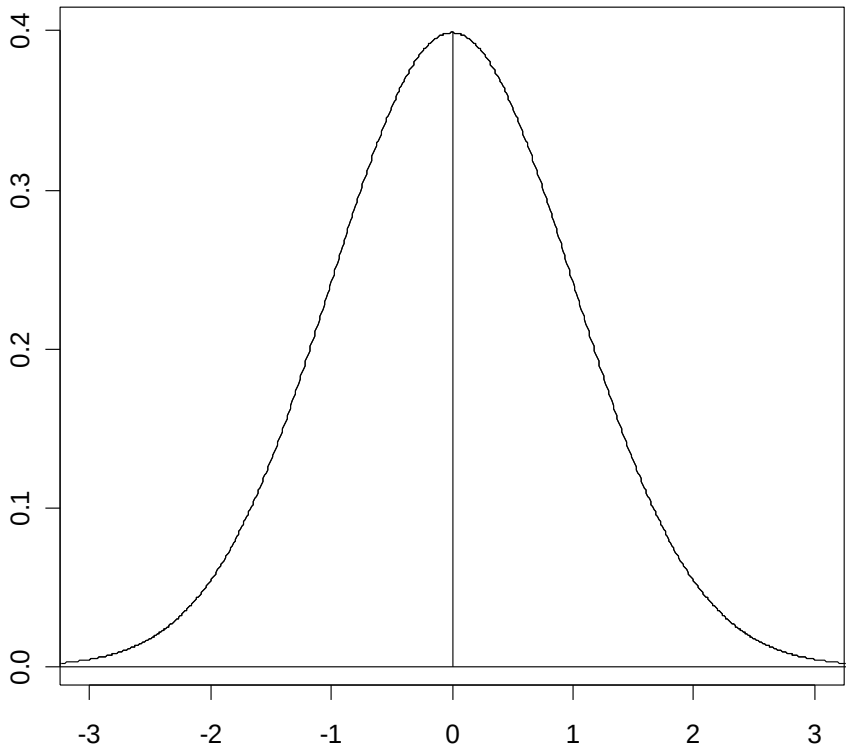
$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\{-(x-\mu)^2/(2\sigma^2)\}, -\infty < x < \infty$$

- $E(X) = \mu, V(X) = \sigma^2$.
- 分布の形状は μ とσ^2によって特徴づけ (parameterized) られる。
- 標準正規分布 (standard normal distribution)

$$f(z;0,1) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}, -\infty < z < \infty,$$

$$\Phi(z) = P(Z < z) = \int_{-\infty}^{z} f(y;0,1)dy.$$

- 釣り鐘型 (bell-shaped) で、左右対称な分布.



回帰係数の推定

最小二乗法 (ordinary least squares estimation, OLSE)

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} \{y_i - (\beta_0 + \beta_1 x_i)\}^2$$

$$\Rightarrow \begin{cases} \dfrac{\partial}{\partial \beta_0} \sum_{i=1}^{n} \{y_i - (\beta_0 + \beta_1 x_i)\}^2 = 0 \\ \dfrac{\partial}{\partial \beta_1} \sum_{i=1}^{n} \{y_i - (\beta_0 + \beta_1 x_i)\}^2 = 0 \end{cases} \Rightarrow \begin{cases} n\beta_0 + \beta_1 \left(\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} y_i \\ \beta_0 \left(\sum_{i=1}^{n} x_i\right) + \beta_1 \left(\sum_{i=1}^{n} x_i^2\right) = \sum_{i=1}^{n} x_i y_i \end{cases}$$

$$\Rightarrow \begin{cases} \hat{\beta}_1 = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \Big/ \sum_{i=1}^{n} (x_i - \bar{x})^2 \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}$$

推定された回帰直線：$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

最小二乗推定量の性質

- $E\left(\hat{\beta}_0\right)=\beta_0, E\left(\hat{\beta}_1\right)=\beta_1,$ 不偏性
- $\sigma^2_{\hat{\beta}_0}=\mathrm{var}\left(\hat{\beta}_0\right)=\dfrac{\sigma^2\Sigma x_i^2}{n\Sigma(x_i-\bar{x})^2},$
- $\sigma^2_{\hat{\beta}_1}=\mathrm{var}\left(\hat{\beta}_1\right)=\dfrac{\sigma^2}{\Sigma(x_i-\bar{x})^2},$
- $\hat{\beta}_0 \sim N\left(\beta_0,\sigma^2_{\hat{\beta}_0}\right), \hat{\beta}_1 \sim N\left(\beta_1,\sigma^2_{\hat{\beta}_1}\right),$ 正規性
- $s^2=\hat{\sigma}^2=\dfrac{\Sigma(y_i-\hat{y}_i)^2}{n-2}=\dfrac{\mathrm{SSE}}{n-2}, E\left(\hat{\sigma}^2\right)=\sigma^2,$

- 信頼区間 (Confidence Interval):

$$T=\frac{\hat{\beta}_1-\beta_1}{s_{\hat{\beta}_1}} \sim t_{n-2}, s_{\hat{\beta}_1}=\frac{s}{\sqrt{\Sigma(x_i-\bar{x})^2}}$$

$$\Rightarrow \text{ C.I. of } \beta_1 : \left(\hat{\beta}_1-s_{\hat{\beta}_1}t_{\alpha/2,n-2,}\hat{\beta}_1+s_{\hat{\beta}_1}t_{\alpha/2,n-2,}\right),$$

同様に, C.I. of $\beta_0 : \left(\hat{\beta}_0-s_{\hat{\beta}_0}t_{\alpha/2,n-2,}\hat{\beta}_0+s_{\hat{\beta}_0}t_{\alpha/2,n-2,}\right)$

- 仮説検定 (Hypothesis Testing):

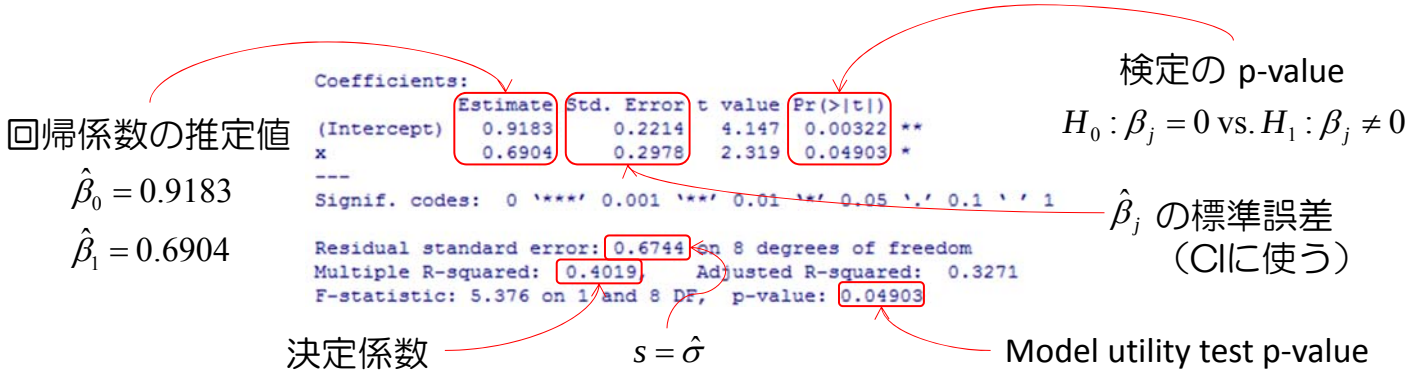
$H_0 : \beta_1=\beta_{10}$ vs. $H_a : \beta_1 \neq (\text{or} < \text{or} >)\ \beta_{10}$

検定統計量 : $T=\dfrac{\hat{\beta}_1-\beta_{10}}{s_{\hat{\beta}_1}} \sim t_{n-2}$ under H_0

$H_a: \beta_1 > \beta_{10}$	$t > t_{\alpha,n-2}$
$H_a: \beta_1 < \beta_{10}$	$t < -t_{\alpha,,n-2}$
$H_a: \beta_1 \neq \beta_{10}$	$\lvert t\rvert > t_{\alpha/2,n-2}$

回帰分析の結果

- 回帰係数の推定値　$\hat{\beta}_0 = 0.9183, \hat{\beta}_1 = 0.6904$
- 回帰係数の有意性検定のp値
- 決定係数（被説明変数の変動のうち回帰によって説明された変動の割合）　yの変動の40.19%が説明された
- Model utility test（回帰モデル全体の有意性検定。後でもう一度触れます）のp値　p=0.049
- 撹乱項の標準誤差　s=0.6744



多変量回帰分析

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon, \varepsilon \sim N\left(0, \sigma^2\right)$$

Y: response variable, 従属変数，被説明変数
$x_1, \ldots, x_k$: independent variable, 独立変数，説明変数
$\beta_0, \beta_1, \ldots, \beta_k$: regression coefficient, 回帰係数
ε: error term, 撹乱項、誤差項

モデルの仮定

1. パラメターに関する線形性 (Linearity)
2. 撹乱項の独立性 (Independence)
3. 撹乱項の正規性 (Normality)
4. 撹乱項の等分散性 (homoscedasticity)

最小二乗法によるパラメターの推定

- $\beta_0, \beta_1, \ldots, \beta_k$ は最小二乗法で推定される:

$$\min_{\beta_0,\ldots,\beta_k} \sum_{j=1}^{n} \left(y_j - \left(\beta_0 + \beta_1 x_{1j} + \cdots + \beta_k x_{kj}\right)\right)^2$$

$$\hat{\beta}_{k\times 1} = \left(X'X\right)^{-1} X'Y,$$

where $X = [x_{ij}]_{n\times(k+1)}, Y = (Y_1, \ldots, Y_n)'$

- $E(\hat{\beta}_j) = \beta_j,\ j = 1, \ldots, k$: unbiased.
- $V(\hat{\beta}) = \sigma^2 \left(X'X\right)^{-1}.$
- $\hat{\beta} \sim N\left(\beta, \sigma^2 \left(X'X\right)^{-1}\right)$

パラメターの推測、信頼区間、検定

- $s^2 = \hat{\sigma}^2 = \dfrac{\sum (y_i - \hat{y}_i)^2}{n-(k+1)} = \dfrac{\text{SSE}}{n-(k+1)}. E(s^2) = \sigma^2.$
- $T = \dfrac{\hat{\beta}_j - \beta_j}{s_{\hat{\beta}_j}} \sim t_{n-(k+1)}, s_{\hat{\beta}_j} = \left([s^2 (X'X)^{-1}]_{jj}\right)^{1/2},\ j = 1, \ldots, k.$
- C.I. of $\beta_j : (\hat{\beta}_j - s_{\hat{\beta}_j} t_{\alpha/2, n-(k+1)}, \hat{\beta}_j + s_{\hat{\beta}_j} t_{\alpha/2, n-(k+1)})$
- Hypothesis testing : $H_0 : \beta_j = \beta_{j0}$ vs. $H_a : \beta_j \neq \beta_{j0}$,

$$\text{Test statistic} : T = \frac{\hat{\beta}_j - \beta_{j0}}{s_{\hat{\beta}_j}} \sim t_{n-(k+1)} \text{ under } H_0.$$

決定係数 (coefficient of determination)

回帰係数の有意性検定は個々の係数の検定．回帰モデル全体のパフォーマンスを測る方法が必要．

SST (Total Sum of Squares): $\sum_{i=1}^{n}(y_i-\bar{y})^2$

yの全変動

SSE (Error Sum of Squares): $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$

回帰で説明されなかった変動

SSR (Regression sum of Squares): $\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$

回帰で説明された変動

R^2=決定係数= $\frac{SSR}{SST}$ =回帰によって説明された変動の割合

Model utility test

回帰モデル全体の有意性を検定するために，以下の Model Utility Test が知られている．

$$H_0:\beta_1=\cdots=\beta_k=0 \text{ vs. } H_1:\text{not } H_0$$

検定統計量：

$$F=\frac{R^2/k}{(1-R^2)/[n-(k+1)]}=\frac{SSR/k}{SSE/[n-(k+1)]}\sim F_{k,n-(k+1)} \text{ under } H_0$$

棄却域：Reject H_0 if $F>F_{\alpha,k,n-(k+1)}$

H_0は、$Y=\beta_0+\varepsilon$ と同義．回帰が全く無効．

回帰分析の結果のチェックポイント

- 回帰係数の推定値
- 回帰係数の有意性検定のp値
- 決定係数（被説明変数の変動のうち回帰によって説明された変動の割合）
- Model utility test（回帰モデル全体の有意性検定）のp値
- 撹乱項の標準誤差

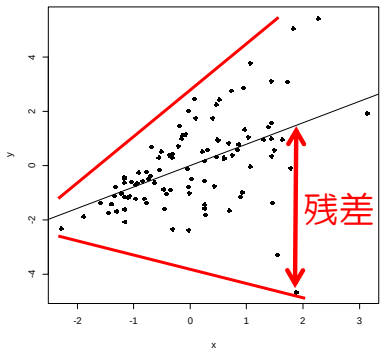
回帰診断

回帰モデルの仮定の確認：

- 線形性: Y と x の間の線形関係．　Y と xの間には非線形な関係がない．　x同士の間には線形関係がない．
  - –Multiple scatter plots.
- 独立性: $\{(x_i, Y_i)\}_{i=1}^{n}$ は互いに独立
  - –residual vs. fitted value plot
- 正規性: $\varepsilon \sim N(0, \sigma^2)$
  - –残差のQQ-norm plot　（後述する）
- 等分散性: $\mathrm{Var}(\varepsilon) = \sigma^2$
  - –residual vs. fitted value plot

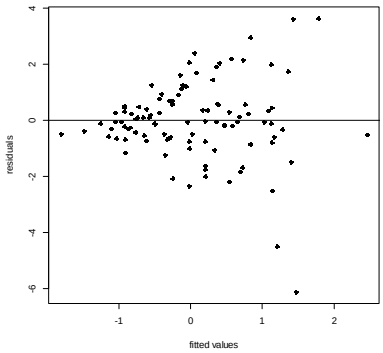
回帰診断（続き）

独立性，正規性，等分散性の仮定は，いずれも撹乱項についての仮定．撹乱項そのものは観察できないため，残差 (residuals)：$e_i = y_i - \hat{y}_i$ をレプリカとして使う．



分散増大傾向がある

Residuals vs. fitted value plot



残差も増大傾向がある

- 残差が均一ならば，等分散性の仮定は満たされる．
- 独立性の仮定が満たされる場合，残差プロットには特異なパターンがない．

分布の正規性の確認

標本分布の正規性の確認は，適切なモデルを選択する上で重要．

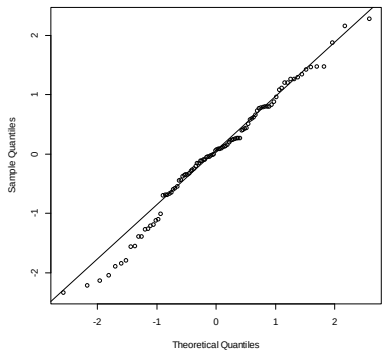
Definition: n個の標本を大きさ順に並べたとき，i番目に小さな標本は $[100(i-.5)/n]$ 標本パーセント点 (sample percentile) であるという．

例えば標本が，正規分布など特定の確率分布から抽出されたとする．このとき，その特定の分布の理論上の $[100(i-.5)/n]$パーセント点は，データの$[100(i-.5)/n]$標本パーセント点の近くにあるはずである．

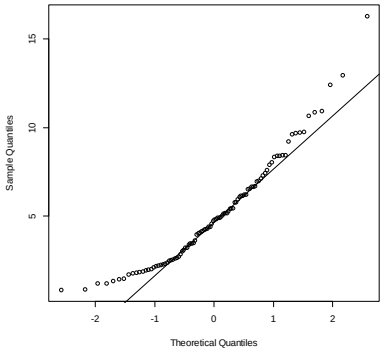
QQ-norm plot（Normal probability plot)

Definition: *n* 個の標本が得られたとき，標準正規分布の [*100(i-.5)/n*] パーセント点と，*i*番目に小さな観測値= [*100(i-.5)/n*] 標本パーセント点のプロットを，QQ-norm plotという。

正規分布

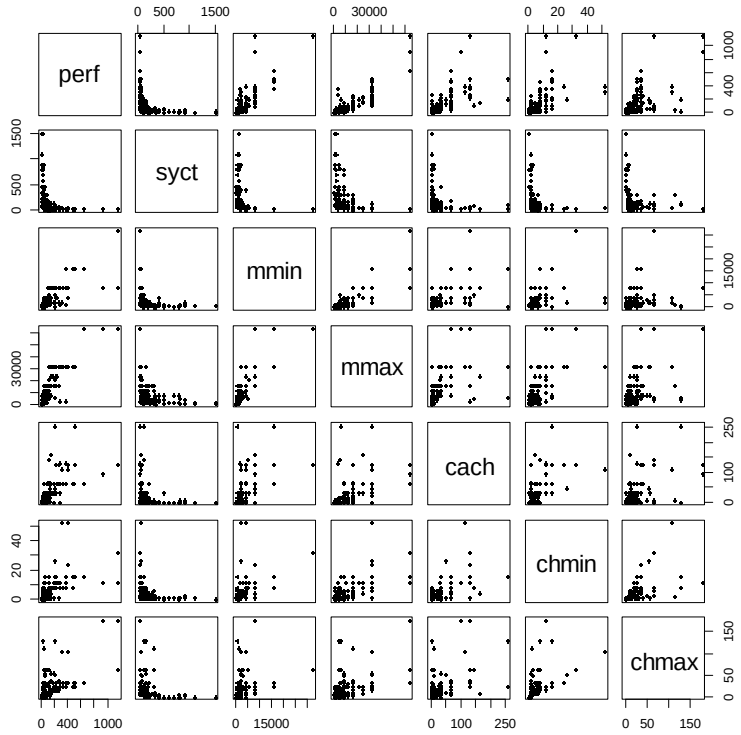


非正規分布



- 元のデータが正規分布から得られた場合，QQ-norm plotは直線上にプロットされる.
- 元データが正規分布に従わない場合、直線から外れる.（右図）

CPUデータ

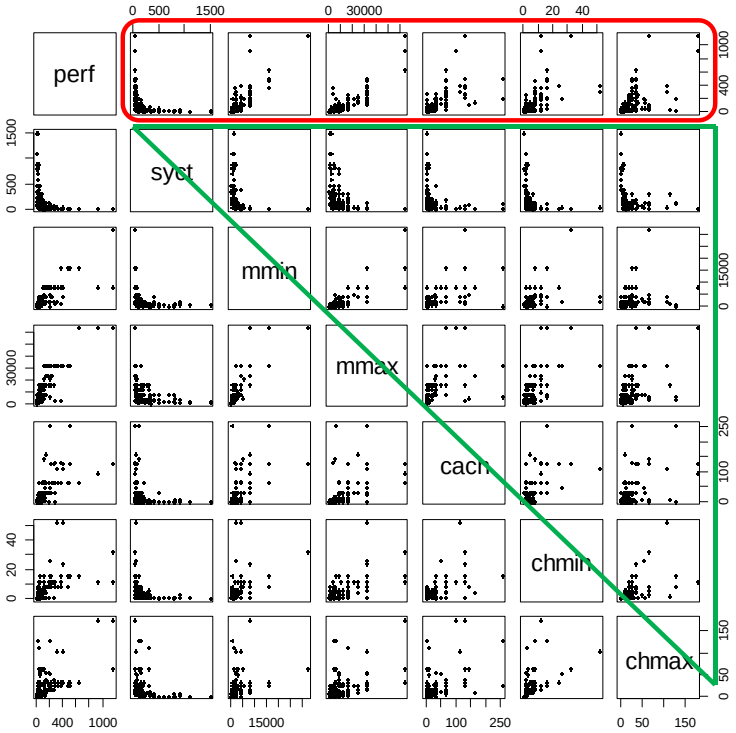


209のコンピュータのCPUの持つ，性能と各種特性値.

'name' Manufacturer and model
'syct' cycle time in nanoseconds
'mmin' minimum main memory in kilobytes
'mmax' maximum main memory in kilobytes
'cach' cache size in kilobytes
'chmin' minimum number of channels
'chmax' maximum number of channels
'perf' published performance on a benchmark mix relative to an IBM 370/158-3

P. Ein-Dor and J. Feldmesser (1987) Attributes of the performance of central processing units: a relative performance prediction model. *Comm. ACM.* **30**, 308–317.

CPUデータ（続き）



P. Ein-Dor and J. Feldmesser (1987) Attributes of the performance of central processing units: a relative performance prediction model. *Comm. ACM.* **30**, 308–317.

- *y*と*x*の間に，非線形な関係が存在しないか？
  - perf と syct の間に，明らかな非線形関係がある．
- 誤差項の分散は一定か？
  - mmin 等に，明らかな分差増大傾向がある．

- *x*同士の間に，線形な関係が存在しないか？
  - 例えば，mmin と mmax の間に明らかに線形関係がある．

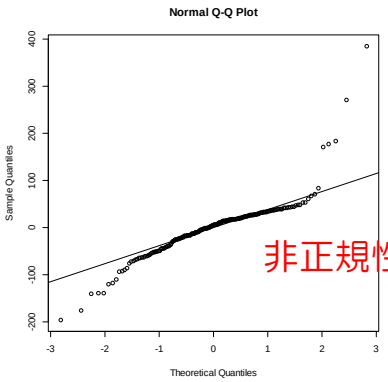
予備的な視覚的要約の段階でも，線形回帰モデルを当てはめるのは不適切であることがわかる．

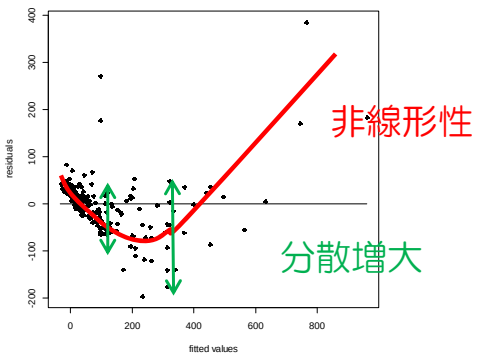
でも，ともかく回帰モデルを当てはめてみる．

CPUデータ（元データの回帰分析）

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.590e+01  8.045e+00  -6.948 4.99e-11 ***
syct         4.886e-02  1.752e-02   2.789  0.00579 **
mmin         1.529e-02  1.827e-03   8.371 9.42e-15 ***
mmax         5.571e-03  6.418e-04   8.680 1.33e-15 ***
cach         6.412e-01  1.396e-01   4.594 7.64e-06 ***
chmin       -2.701e-01  8.557e-01  -0.316  0.75263
chmax        1.483e+00  2.201e-01   6.738 1.64e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 59.99 on 202 degrees of freedom
Multiple R-squared:  0.8649,    Adjusted R-squared:  0.8609
F-statistic: 215.5 on 6 and 202 DF,  p-value: < 2.2e-16
```





- 説明変数の有意性検定は，chmin を除き，ほとんどが強く有意．

- 決定係数：R^2=0.8649　被説明変数の変動の86.49%が説明できた．

- Model utility test: p-value < 2.2×10^{-16}

回帰分析は，成功しているとしかいいようがない．
回帰診断の結果，モデルの仮定は破綻している。

変数変換

線形回帰モデルの仮定（線形性, 正規性, 等分散性）が満たされないとき，変数に何らかの変換を施すことで，モデルを改善できる場合がある.

例えば，撹乱項の分散が説明変数の値とともに大きくなる場合，logarithmic/power 変換が有効であることが多い.

被説明変数の予測値を得るには，まず変換された変数に対して線形回帰モデルを当てはめ，次にもとのモデルに逆変換する．最もよい変換を選ぶため，いくつかの変換を試してみる必要がある.

Box-Cox変換

対数変換，冪変換を組み合わせたBox-Cox変換により，分散の安定化と正規性の改善を同時に達成できる場合がある.

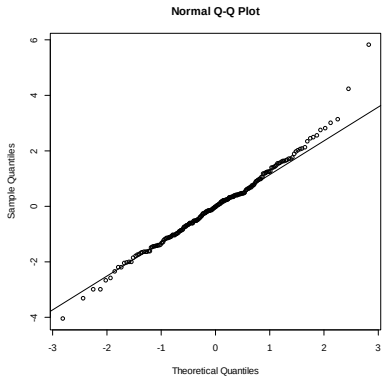
Box-Cox変換： $$y^{(\lambda)} = \begin{cases} (y^{\lambda} - 1)/\lambda : & \lambda \neq 0 \\ \log(y) : & \lambda = 0 \end{cases}$$

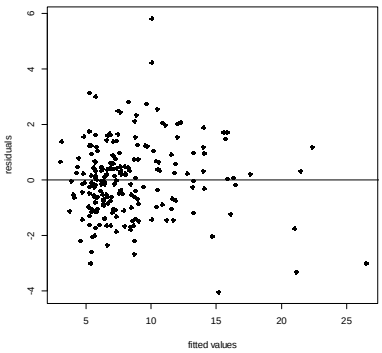
Box-Cox変換は，パラメーターλによって特徴付けられる.
パラメーターλは，モデルの適合度を最適化するように，ソフトウエアにより自動的に選択される.
（統計解析ソフトRなどが， Box-Cox変換を実装している）

CPUデータ（Box-Cox変換後）

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.214e+00  1.843e-01  28.284  < 2e-16 ***
syct        -1.681e-03  4.014e-04  -4.187 4.21e-05 ***
mmin         1.868e-04  4.186e-05   4.463 1.34e-05 ***
mmax         1.607e-04  1.471e-05  10.924  < 2e-16 ***
cach         2.792e-02  3.198e-03   8.731 9.56e-16 ***
chmin        2.774e-02  1.961e-02   1.415    0.159
chmax        8.330e-03  5.042e-03   1.652    0.100
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.375 on 202 degrees of freedom
Multiple R-squared:  0.8821,    Adjusted R-squared:  0.8786
F-statistic: 251.8 on 6 and 202 DF,  p-value: < 2.2e-16
```





- chmin，chmax は有意ではない．
- 決定係数：R^2=0.8821　被説明変数の変動の88.21（＞86.49）％が説明できた．
- Model utility test: p-value＜2.2×10^{-16}

回帰分析は，成功している．

Take Home Message

1. 回帰分析
2. 共分散と相関係数
3. 線形回帰モデル
   - 回帰係数の推定．最小二乗推定量の性質
4. 回帰診断：回帰モデルの仮定の確認
   - 散布図：線形性の確認
   - QQ-normプロット：残差の正規性の確認
   - 残差プロット：等分散性，独立性の確認
5. Box-Cox変換：分散の安定化と正規性の向上