

医学統計勉強会

第4回 回帰分析

帝京大学臨床研究センター（TARC）・帝京大学大学院公衆衛生学研究科 共催

帝京大学大学院公衆衛生学研究科

宮田 敏

回帰分析

回帰分析 (regression analysis) は、一つの連続変数（実数値）の値を複数の変数によって説明、予測する**多変量解析 (multivariable analysis)** の一つ。

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \epsilon_i \sim N(0, \sigma^2)$$

Y : response variable, 従属変数, 被説明変数, 応答変数

x_1, \dots, x_k : independent variable, 独立変数, 説明変数, 共変量

$\beta_0, \beta_1, \dots, \beta_k$: regression coefficient, 回帰係数

ϵ : error term, 誤差項

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。

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

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

共分散の値そのものは、解釈が難しい。正負の符号が、興味の対象。

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。相関係数は単位に依存しない。

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

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

- 相関係数は、**単位に依存しない**。

$$\text{Corr}(X, Y) = \text{Corr}(aX + c, bY + d), \text{ for } ab > 0$$

例えば、温度の摂氏 (° C) と華氏 (° F) は

$$^{\circ}\text{C} = (5 \div 9) \times (^{\circ}\text{F} - 32)$$

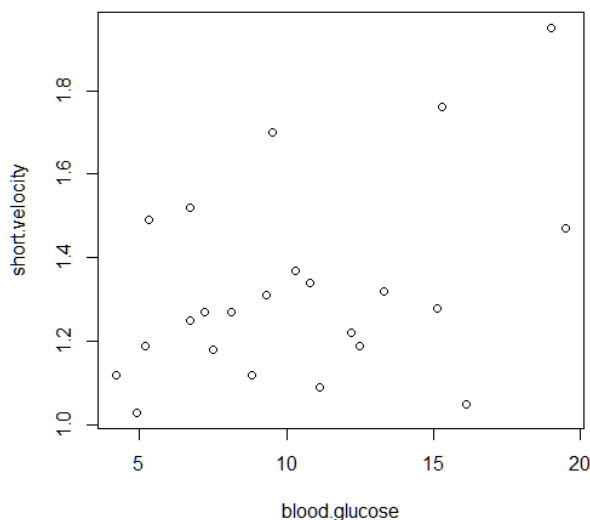
という関係があるが、摂氏で測っても華氏で測っても相関係数は同じ。

- 相関係数は、線形関係の強さを測るだけ。X が 1 単位増加したときの Y の変化量のような、**量的な評価はできない**。

⇒ 量的な評価は、線形回帰分析による。

例：I型糖尿病患者の空腹時血糖値と心室内周短縮速度

blood.glucose: 空腹時血糖値
chort.velocity: 心室内周短縮速度



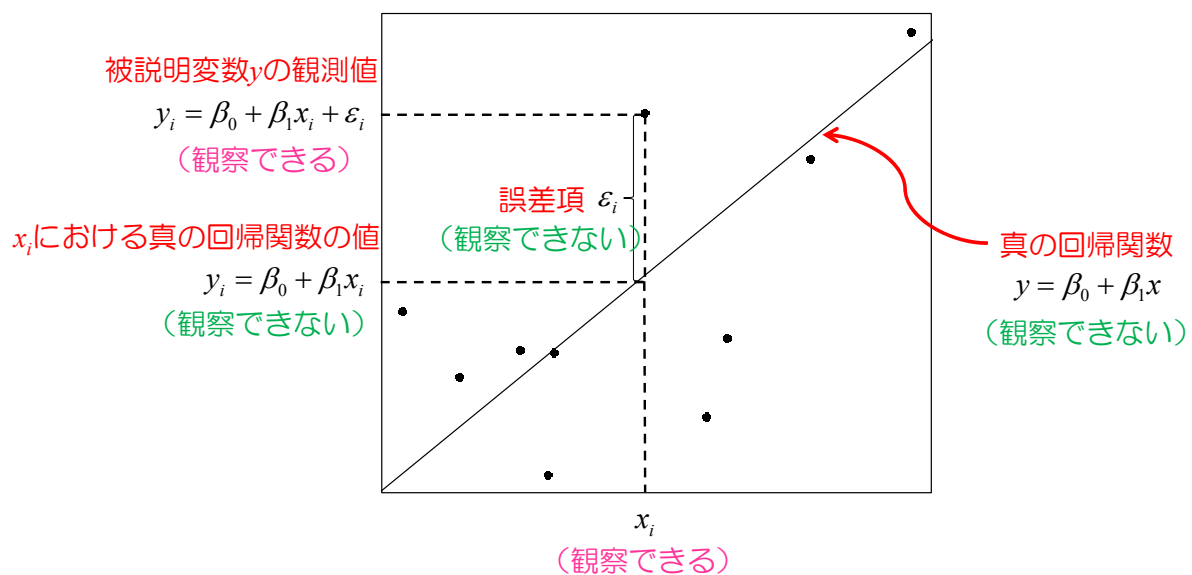
```
R Console
> library(ISwR)
警告メッセージ:
パッケージ 'ISwR' はバージョン 3.4.4 の R の下で壊れました
>
> # 散布図 #
> thuesen <- na.omit(thuesen) # 欠測値を除く #
> attach(thuesen)
>
> plot(blood.glucose, short.velocity)
> plot(thuesen) # 上と同じ
>
> # 数値的データの要約(二変量) #
> cov(blood.glucose, short.velocity) # 共分散
[1] 0.4289723
> cov(thuesen) # 分散共分散行列
      blood.glucose short.velocity
blood.glucose 19.5320158  0.42897233
short.velocity 0.4289723  0.05424387
> cor(blood.glucose, short.velocity) # 相関係数
[1] 0.4167546
> cor(thuesen) # 相関行列
      blood.glucose short.velocity
blood.glucose 1.0000000  0.4167546
short.velocity 0.4167546  1.0000000
```

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, 応答変数
 x : independent variable, 独立変数, 説明変数, 共変量
 β_0, β_1 : regression coefficient, 回帰係数
 ε : error term, 誤差項



観察可能な (x_i, y_i) から, $y = \beta_0 + \beta_1 x$ を推定したい。

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
誤差項 ε_i は正規分布に従う.
- **等分散性 (homoskedasticity)** : σ^2 . 分散一定

回帰モデルの推定（最小二乗法）

回帰モデルの仮定の下で、回帰係数 β_0, β_1 を推定したい。誤差項 ε は期待値 0 の正規分布に従うから、50% の確率で負の値を取り、50% の確率で正の値を取る。

回帰直線は、データの真ん中を通る

⇔ 回帰直線とデータとの距離は最小であるはず。

個々のデータ (x_i, y_i) と回帰直線 $y_i = \beta_0 + \beta_1 x_i$ の間の距離を**二乗距離**で測る。 $(y_i - (\beta_0 + \beta_1 x_i))^2 = \epsilon_i^2$

データ全体と回帰直線の距離は、二乗距離の総和 = **残差二乗和** $\sum_{i=1}^n \epsilon_i^2$ に等しい。

データ全体と回帰直線の距離は**残差二乗和** (residual sum of squares)

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

に等しいから、これを回帰係数 β_0, β_1 について最小化する。

$$\begin{aligned} & \min_{\beta_0, \beta_1} \sum_{i=1}^n \{y_i - (\beta_0 + \beta_1 x_i)\}^2 \\ \Rightarrow & \begin{cases} \frac{\partial}{\partial \beta_0} \sum_{i=1}^n \{y_i - (\beta_0 + \beta_1 x_i)\}^2 = 0 \\ \frac{\partial}{\partial \beta_1} \sum_{i=1}^n \{y_i - (\beta_0 + \beta_1 x_i)\}^2 = 0 \end{cases} \Rightarrow \begin{cases} n\beta_0 + \beta_1 \left(\sum_{i=1}^n x_i\right) = \sum_{i=1}^n y_i \\ \beta_0 \left(\sum_{i=1}^n x_i\right) + \beta_1 \left(\sum_{i=1}^n x_i^2\right) = \sum_{i=1}^n x_i y_i \end{cases} \\ \Rightarrow & \begin{cases} \hat{\beta}_1 = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) / \sum_{i=1}^n (x_i - \bar{x})^2 \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases} \end{aligned}$$

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

- $E(\hat{\beta}_0) = \beta_0, E(\hat{\beta}_1) = \beta_1$, **不偏性**
- $\sigma_{\hat{\beta}_0}^2 = \text{var}(\hat{\beta}_0) = \frac{\sigma^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}$,
- $\sigma_{\hat{\beta}_1}^2 = \text{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$,
- $\hat{\beta}_0 \sim N(\beta_0, \sigma_{\hat{\beta}_0}^2), \hat{\beta}_1 \sim N(\beta_1, \sigma_{\hat{\beta}_1}^2)$, **正規性**
- $s^2 = \hat{\sigma}^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{\text{SSE}}{n-2}, E(\hat{\sigma}^2) = \sigma^2$,

- 信頼区間 (Confidence Interval):

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

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

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

- 仮説検定 (Hypothesis Testing):

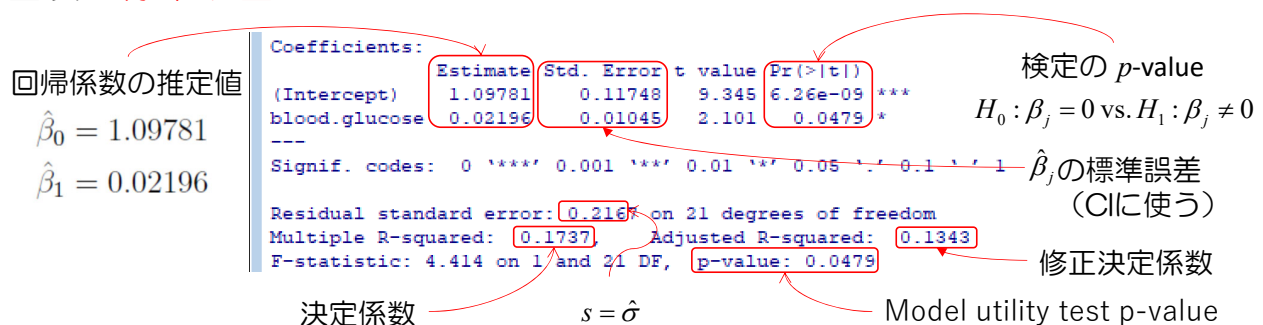
$$H_0 : \beta_1 = \beta_{10} \text{ vs. } H_a : \beta_1 \neq (\text{or } < \text{ or } >) \beta_{10}$$

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

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

回帰分析の結果

- 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) y の変動の17.37%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 $p = 0.0479$
- 誤差項の標準誤差 $s = 0.216$



回帰診断

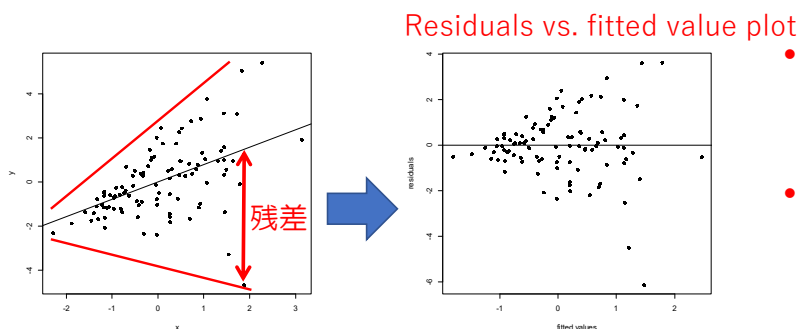
回帰モデルの仮定の確認：

- **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**: $\text{Var}(\varepsilon) = \sigma^2$
–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, いずれも誤差項についての仮定。
誤差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

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。

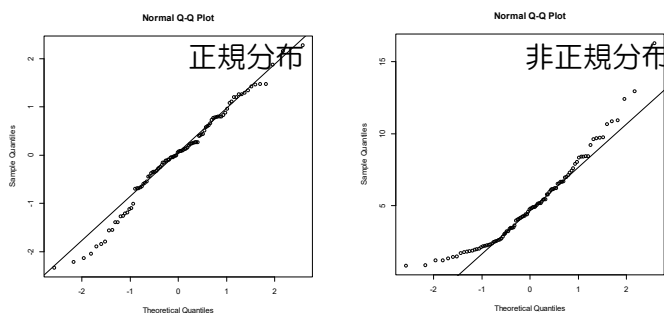
Definition: n 個の標本を大きさ順に並べたとき、 i 番目に小さな標本は $[100(i - 0.5)/n]$ **標本パーセント点 (sample percentile)** であるという。

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 $[100(i - 0.5)/n]$ パーセント点は、データの $[100(i - 0.5)/n]$ 標本パーセント点の近くにあるはずである。

⇒ 標本が正規分布から抽出されたのであれば、理論上のパーセント点と標本パーセント点のプロットは、45度線付近にプロットされる。

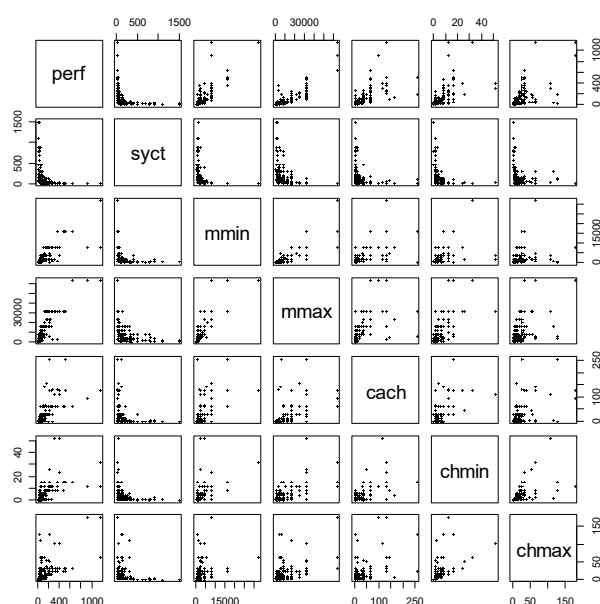
QQ-norm plot (Normal probability plot)

Definition: n 個の標本が得られたとき、標準正規分布の $[100(i - 0.5)/n]$ パーセント点と、 i 番目に小さな観測値 = $[100(i - 0.5)/n]$ 標本パーセント点のプロットを、**QQ-norm plot**という。



- 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

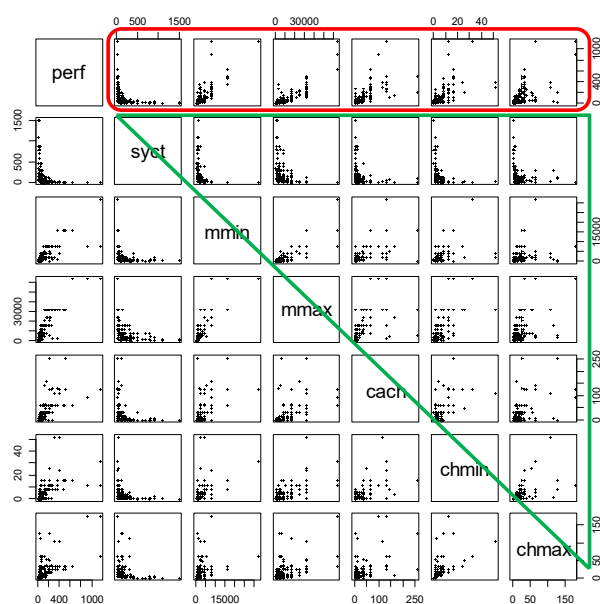


P. Ein-Dor and J. Feldmesser (1987) Attributes of the performance of central processing units: a relative performance prediction model. *Comm. ACM*. **30**, 308–317.

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

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、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、とにかく回帰モデルを当てはめてみる。

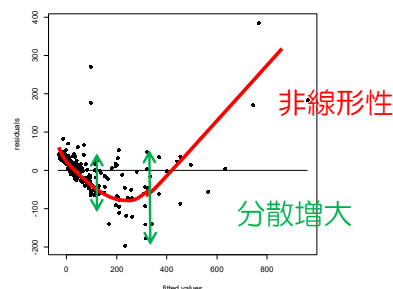
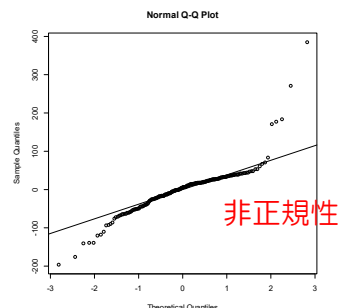
CPUデータ（元データの回帰分析）

```
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: $p\text{-value} < 2.2 \times 10^{-16}$

回帰分析は、成功しているとしかいいようがない。しかし、**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、誤算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，パラメター λ によって特徴付けられる。
パラメター λ は，モデルの適合度を最適化するように，ソフトウェ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)

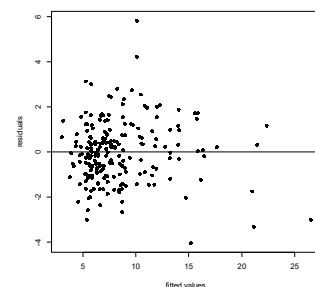
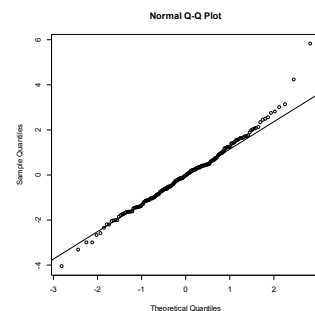
CPUデータ (Box-Cox変換後)

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.214e+00  1.843e-01  28.284 < 2e-16 ***
syst         -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。



回帰分析における変数選択

多くの説明変数の候補の中から、最良のセットを選びたい。そのためにはモデルの良さ (fitness) を測る尺度が必要。 (**赤池の情報量基準, Akaike's Information Criterion, AIC**)

$$AIC = \frac{RSS}{\sigma^2} + 2k + \text{const.} \quad \sigma^2 \text{ known}$$

$$AIC = n \log(RSS/n) + 2k + \text{const.} \quad \sigma^2 \text{ unknown}$$

$$RSS = \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}))^2$$

k : パラメーターの数。第一項はモデルが如何にデータに近く当てはめられたかを測る。第二項はモデルの複雑さを測る。全体として、AICはモデルの**フィットのよさ**と**複雑さ**をバランスする尺度になっている。

2024/10/23

医学統計勉強会 第3回 連続変数の比較

25

最適な変数の組み合わせを探索するための手順:

1. 説明変数の候補を準備する
2. すべての候補変数を用いたfull modelを推定する
3. **stepAIC** (MASS library)コマンドで最適な説明変数の組を探す

Example 13.11: ボールの接着剤の剪断強度

```
library(Devore7)
data(xmp13.11)
xmp13.11.lm <- lm(Strength~Force+Power+Temperature+Time, data=xmp13.11)

library(MASS)
xmp13.11.step <- stepAIC(xmp13.11.lm)
summary(xmp13.11.step)
```

2024/10/23

医学統計勉強会 第3回 連続変数の比較

26

```
> library(Devore6)
> data(xmp13.11)
> xmp13.11.lm <- lm(Strength~Force+Power+Temperature+Time)
> library(MASS)
> xmp13.11.step <- stepAIC(xmp13.11.lm)
Start: AIC=102.86
Strength ~ Force + Power + Temperature + Time

      Df Sum of Sq  RSS   AIC
- Force    1    26.88 692.00 102.15
- Time     1    40.04 705.16 102.72
<none>                 665.12 102.96
- Temperature 1    252.20 917.32 110.61
- Power      1   1341.02 2006.13 134.08

Step: AIC=102.15
Strength ~ Power + Temperature + Time

      Df Sum of Sq  RSS   AIC
- Time     1    40.04 732.04 101.84
<none>                 692.00 102.15
- Temperature 1    252.20 944.20 109.47
- Power      1   1341.02 2033.02 132.48

Step: AIC=101.84
Strength ~ Power + Temperature

      Df Sum of Sq  RSS   AIC
<none>                 732.04 101.84
- Temperature 1    252.20 984.24 108.72
- Power      1   1341.02 2073.06 131.07
> summary(xmp13.11.step)

Call:
lm(formula = Strength ~ Power + Temperature, data = xmp13.11)

Residuals:
    Min       1Q   Median       3Q      Max
-11.3233  -2.8087  -0.8483   3.1892   9.4600

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -24.90167    10.07207  -2.472  0.02001 *
Power         0.49833     0.07086   7.033 1.47e-07 ***
Temperature  0.12967     0.04251   3.050 0.00508 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.207 on 27 degrees of freedom
Multiple R-squared:  0.6852,    Adjusted R-squared:  0.6619
F-statistic: 28.98 on 2 and 27 DF,  p-value: 1.674e-07

> |
```

変数選択の実際（私の場合）

医学統計の性質上，説明変数は取りこぼしなく広めに選択したい。しかし，候補の数がサンプル数より大きければ，最初に全ての変数を用いたモデルは推定できない。また，欠測が多いとサンプル数が減る。

1. まず**単変量解析**で，変数ごとの p 値を出す。
2. ある程度 p 値が小さい変数に，**候補を絞る**。
（ $p < 0.2$ 程度で，あまりきつく絞らない）
3. 絞った候補から，**変数減少法**で選択する。（割合多めの変数が選択される）Rでは `stepAIC()` コマンドを用いる（MASSパッケージが必要）。

変数選択の問題

- 変数選択の結果選ばれたモデルに、有意でない変数が含まれる。
変数選択は、被説明変数の変動をもっともよく説明する変数の組み合わせを探索する。個々の変数が有意でなくても、組み合わせ全体として最適であると解釈する。
- 変数選択の結果、興味のある変数がモデルから除かれてしまった。
上記に矛盾するようだが、回帰の目的は被説明変数を説明することだけではない。変数間の関係を推測するため、興味ある変数を強制投入してもよい。

参考文献：

Peter Dalgaard (著), 岡田 昌史 (監修, 翻訳) 「Rによる医療統計学 原書2版」 ISBN-13: 978-4621087756