重回帰分析

残差分析・変数選択

内容

- 重回帰分析
  - 残差分析
  - 歯の咬耗度データの分析
  - 「R」で変数選択 ~ step 関数 ~

重回帰分析と単回帰分析

- 体重を予測する問題
  - 分析1…「身長」のみから体重を予測
  - 分析2…「身長」と「ウエスト」の両方を用いて体重を予測

    分析1と比べて大きな改善

    ⇒「体重」に関する推測では「身長」だけでは不十分
- 重回帰分析における問題 ~モデルの構築~
  - 適切なモデルで分析しているか？
  - 適切な変数をモデルに組み込んでいるか？

モデル選択・変数選択の問題

残差分析

残差における仮定

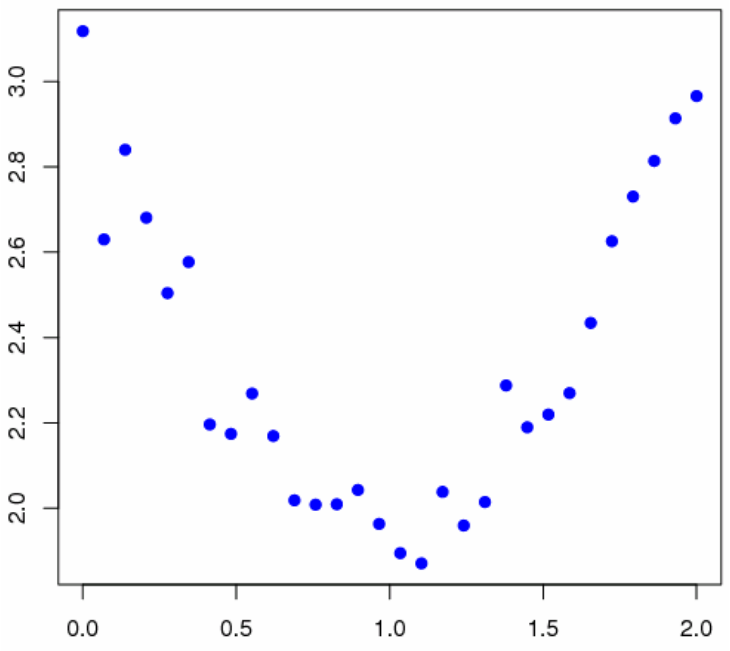
- 回帰分析における残差
  - モデルに組み込んだ変数では説明しれない「偶然誤差」

- 適切なモデルのもとでの残差に関する仮定
  - 残差に正規分布を仮定する
  - 残差の期待値は0
  - 残差の分散は等しい
  - それぞれの残差は互いに独立である

モデルチェック

モデルチェック：例

- 右の図のデータに対する適切なモデルとは？
  - １次式によるモデル
  - ２次式によるモデル
- データの構造
  - $y_i = (x_i - 1)^2 + 2 + \varepsilon_i$
  - $\varepsilon_i \sim N(0, 0.1^2)$
  - $i = 1, \ldots, 30$



１次式によるモデル

- １次式を仮定して分析を行うと、次の結果を得る

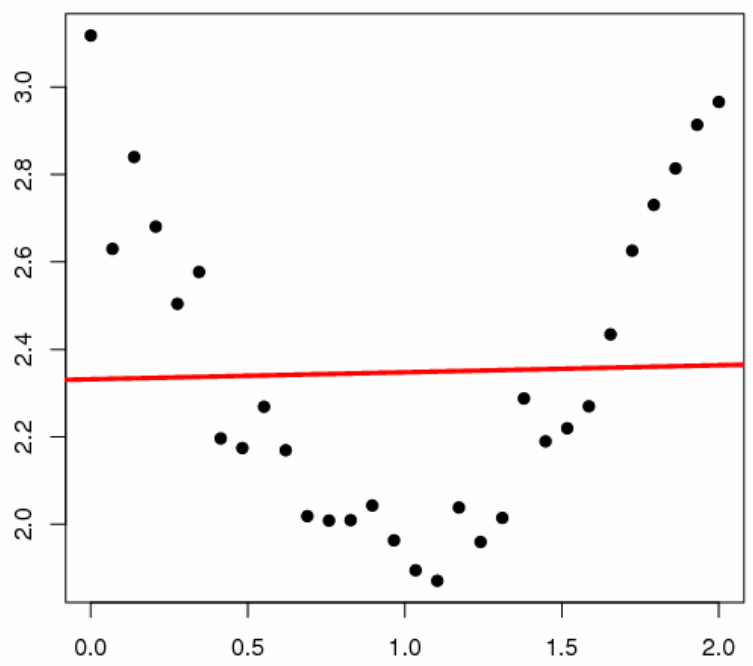
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- 仮定したモデルは適切か？

残差分析

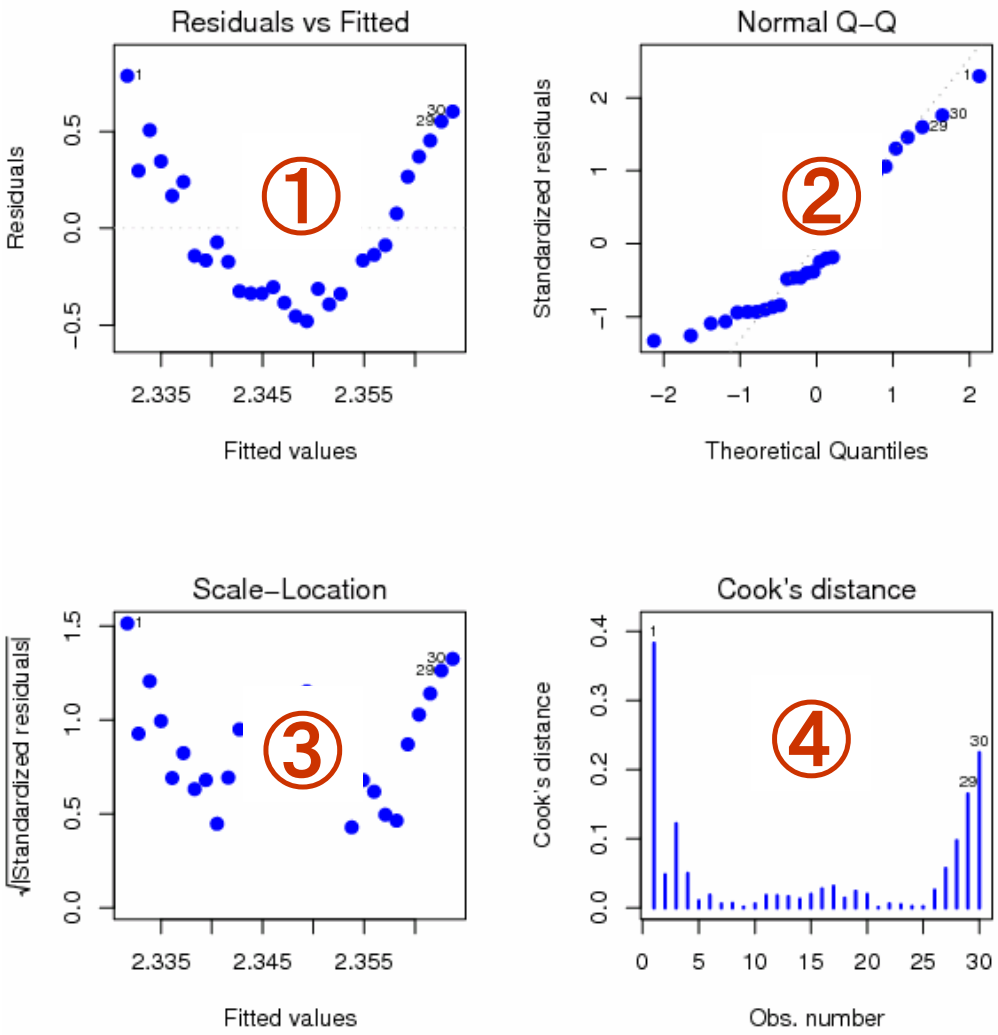
残差の仮定を満たしているか？

回帰診断プロット



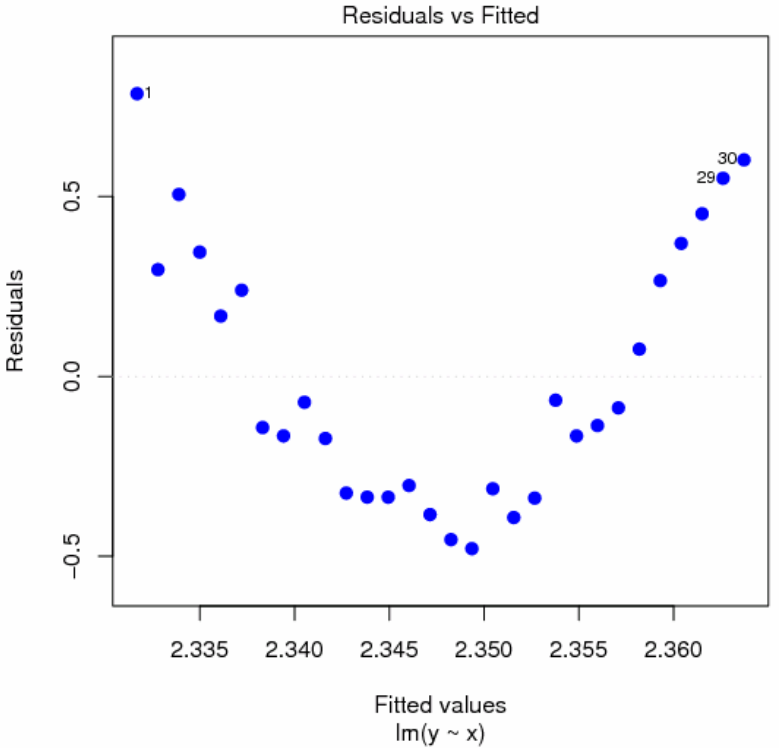
残差分析：1次式

- 回帰診断プロット
  - ① 残差(y)と予測値(x)
  - ② 正規Q-Qプロット
  - ③ 規準化残差と予測値
  - ④ Cookの距離
- 用途
  - ① 残差のふるまい
  - ② 正規性の検証
  - ③ 残差の大きさ
  - ④ 外れ値の探索



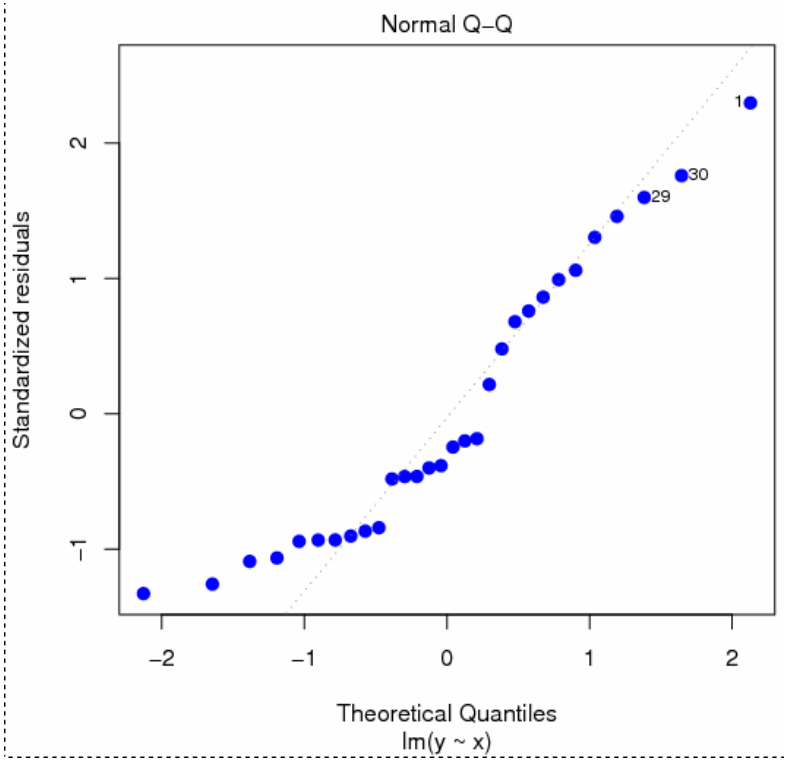
①　残差のふるまい

- 横軸：予測値、縦軸：残差
- 残差の全体像の把握
- 相対的に大きい残差には
  番号がふられる（1, 29, 30）
- 残差の独立性と系列相関の有無
- 系列相関に関する検定
  ダービン・ワトソン統計量
  杉山 髙一著
  「多変量データ解析入門」



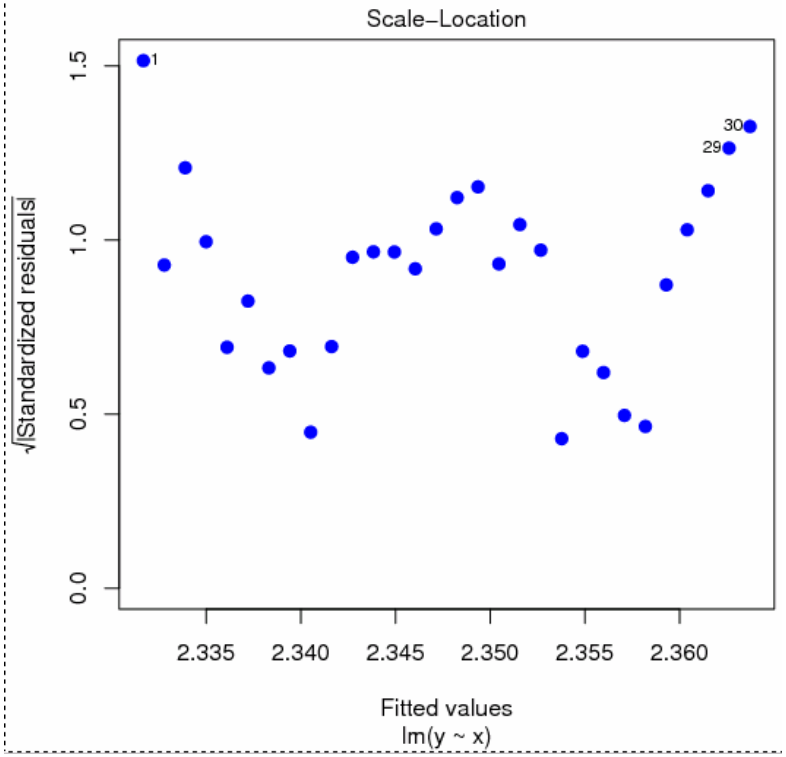
② 正規Q-Qプロット

- 名称：正規Q-Qプロット
- 横軸 ： 正規分布の
  縦軸 ： 規準化残差の
  経験分布関数による
- 残差が正規分布に従っている
  ⇒点が直線上に並べられる。
- 残差が正規分布からずれている
  ⇒点が直線からはずれる
- 残差の仮定 ： 標準正規分布
- 相対的に、直線から外れている
  データには番号がふられる（1, 29, 30）



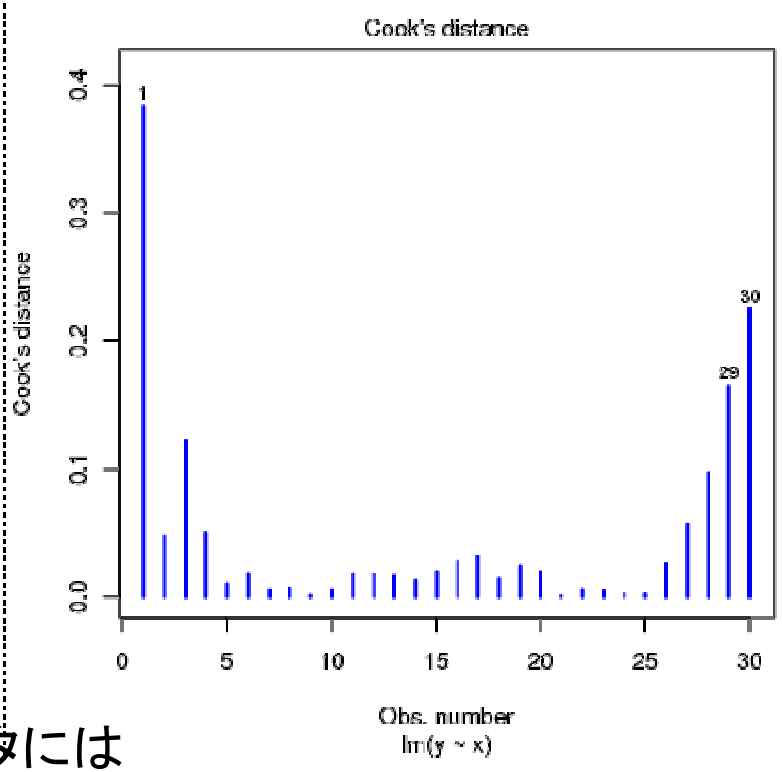
③　残差の大きさ

- 縦軸 ： 規準化した残差の絶対値の平方根
  横軸 ： 予測値
- 残差の変動の考察
- 相対的に大きい残差には
  番号がふられる（1, 29, 30）



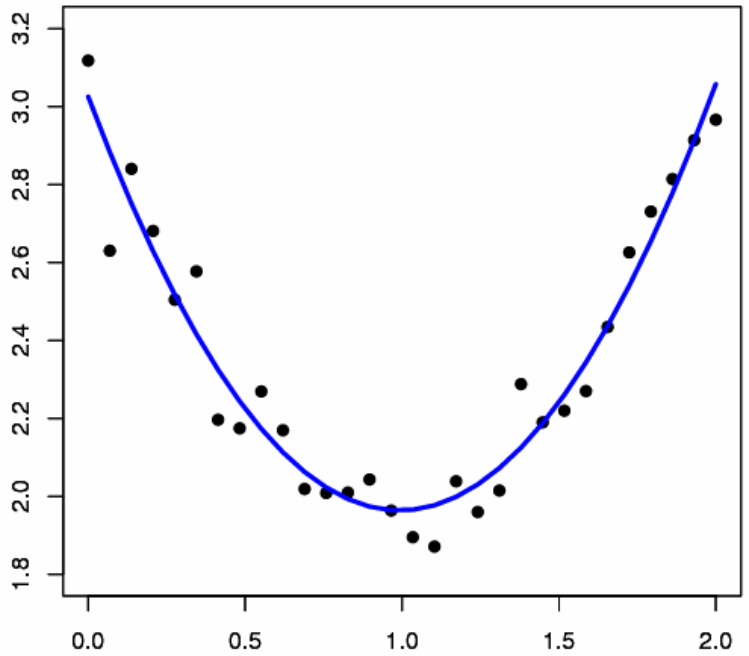
④ Cookの距離

- Cookの距離
  - 個々のデータが回帰式の推定に及ぼす影響を表した距離
- Cookの距離が大きいデータ
  ⇒回帰式の推定に大きく影響
  ⇒外れ値の可能性
- 「R」では、
  「Cookの距離≧0.5」
  ならば大きいとしている
  （絶対的なものではない）
- Cookの距離が相対的に大きいデータには番号がふられる（1, 29, 30）



２次式によるモデル

- モデルに2次式を仮定すると、次の結果を得る
- 回帰診断プロットによる
  残差の検討

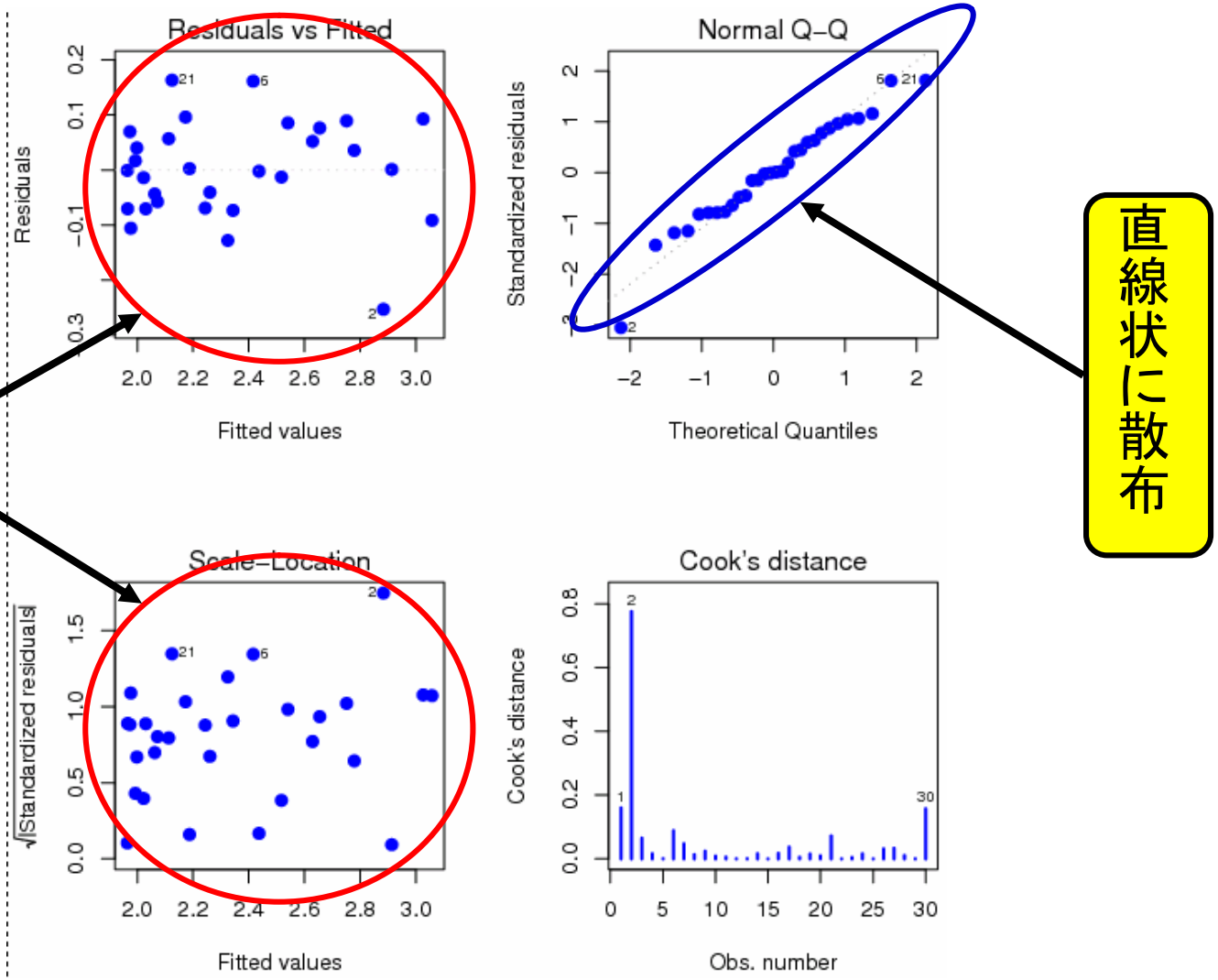


残差分析：2次式

期待値0

残差は適当にばらついている

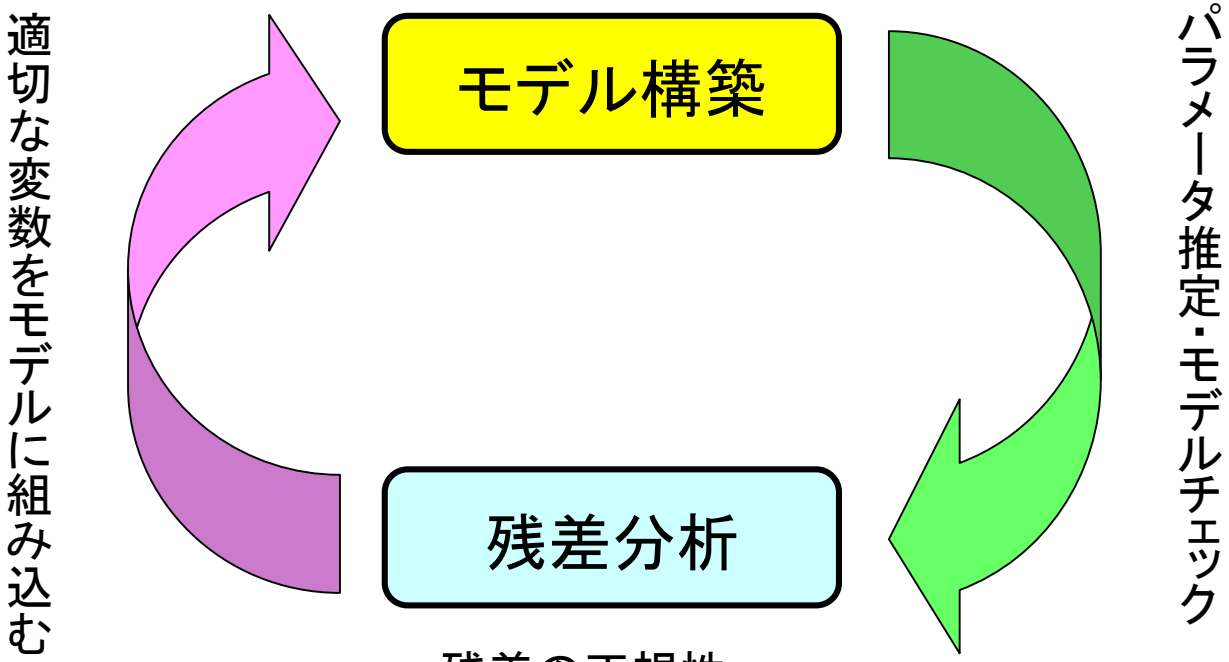
独立性



直線状に散布

実データにおけるモデル構築

- 主に変数を追加する場合



- 残差の正規性
- 残差の期待値は0
- 残差の分散は等しい
- それぞれの残差は互いに独立

プログラム：参考　①

- 1次式のあてはめで用いたプログラム

```
x <- seq(from=0, to=2, length.out=30)
e <- rnorm(30, 0, 0.1)
y <- (x-1)^2+2+e

result <- lm(y~x)
plot(x, y, pch=19, col="black")
abline(result, col="red", lwd=3)

par(mfrow=c(2,2))
for(i in 1:4){
    plot(result, which=i, add.smooth=F, pch=21,
    bg="blue", col="blue", lwd=2)
}
```

プログラム：参考　②

- 2次式のあてはめで用いたプログラム

```
x2 <- x^2
result <- lm(y ~ x+x2)

plot(x, y, xlim=c(0,2), ylim=c(1.8,3.2), pch=19)
par(new=T)
plot(x, fitted(result), type="l", xlim=c(0,2),
      ylim=c(1.8,3.2), ann=F, col="blue", lwd=3)

par(mfrow=c(2,2))
for(i in 1:4){
     plot(result, which=i, add.smooth=F, pch=21,
          bg="blue", col="blue", lwd=2)
}
```

プログラムの説明 （回帰診断）

```
par(mfrow=c(2,2))
for(i in 1:4){
      plot(result, which=i, add.smooth=F, pch=21,
            bg="blue", col="blue", lwd=2)
}
```

- 「R」では for 文も使うことができます。
- par(mfrow=c(2,2))
  次に描く図やグラフを描くスペースを2行2列に分割
- plot(lm.obj) ： 回帰診断プロットの出力
- その他の引数については R-Tips をご覧下さい

歯の咬耗度データの分析

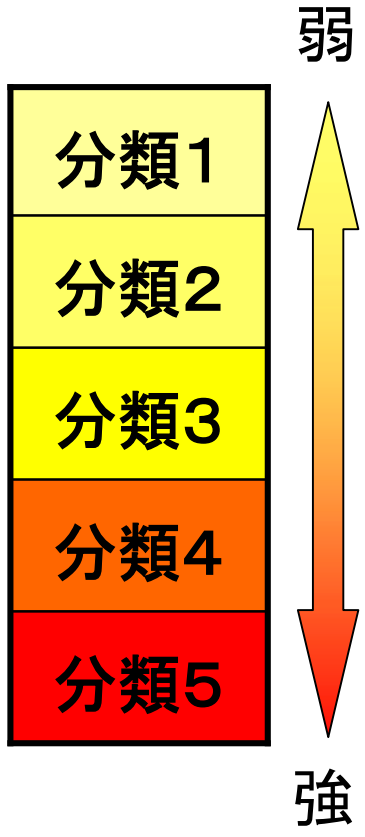
～ 変数選択 ～

日本大学名誉教授（松戸歯学部）尾崎公教授による「歯の咬耗度」のデータです。このデータを用いて分析の説明をいたします。

歯のデータの分析

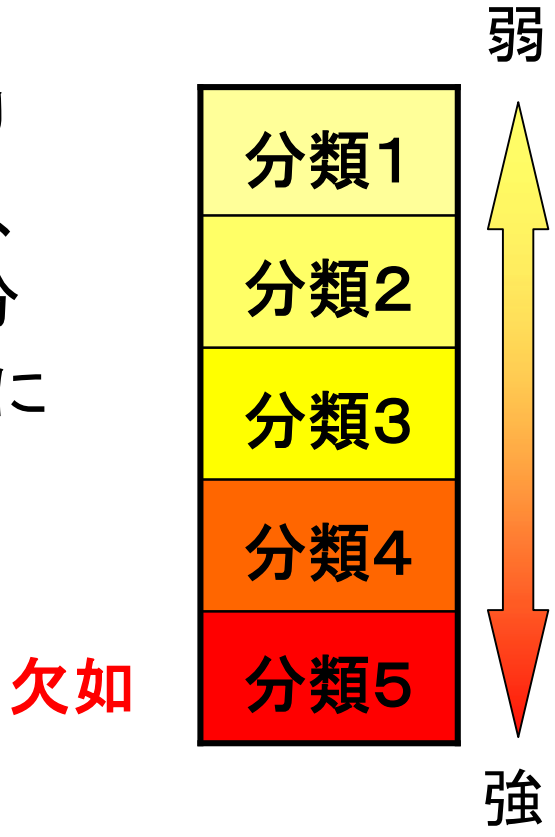
- 歯の咬耗度による年齢推定
- データ
  - 189人、28本の歯の咬耗度を測定
  - 歯の摩耗の度合いは5段階
  - 各分類に、どのような数値を割りふるか、すなわち数量化が重要な問題になる。

欠如



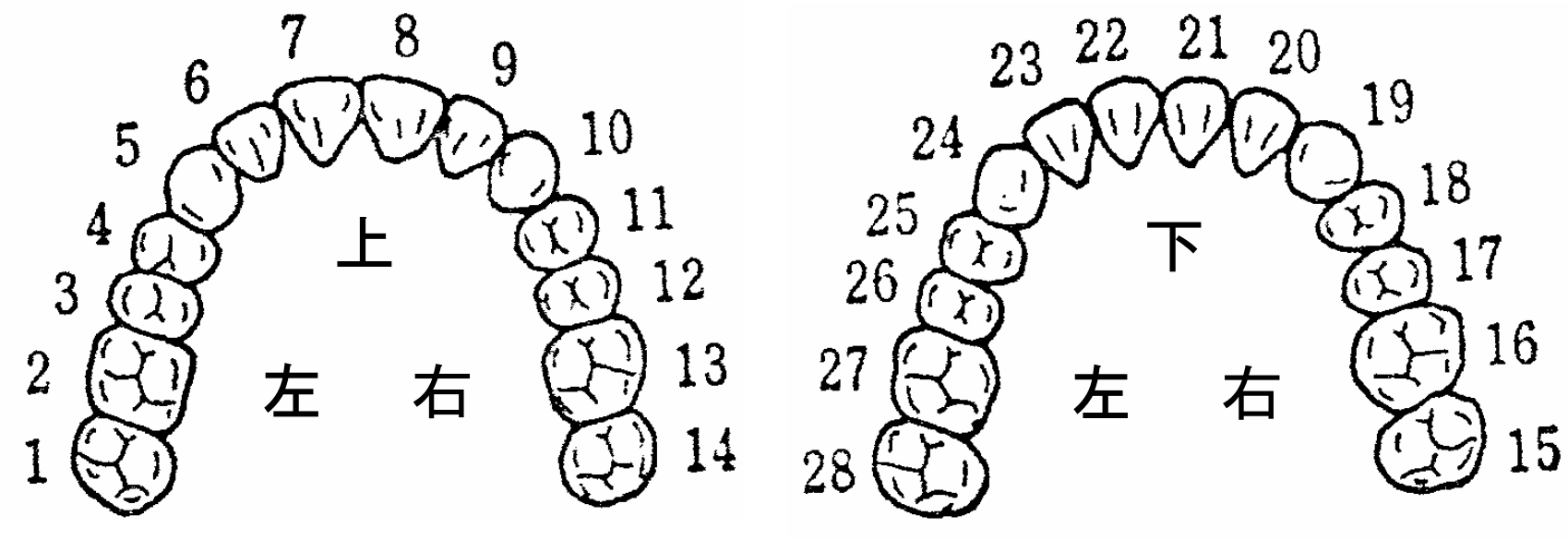
歯のデータの分析

- 数量化について
  - ここでは、数量化分析等を用いたりして検討した結果、分類1には1.0、分類2には2.0、分類3には3.0、分類4には4.0、欠如した歯の分類5には4.0を与えた。
  - 詳しい変数の分類・分析法
    杉山 髙一著
    「多変量データ解析入門」



変数名の定義

- 歯の変数名と対応関係



歯の咬耗度データ

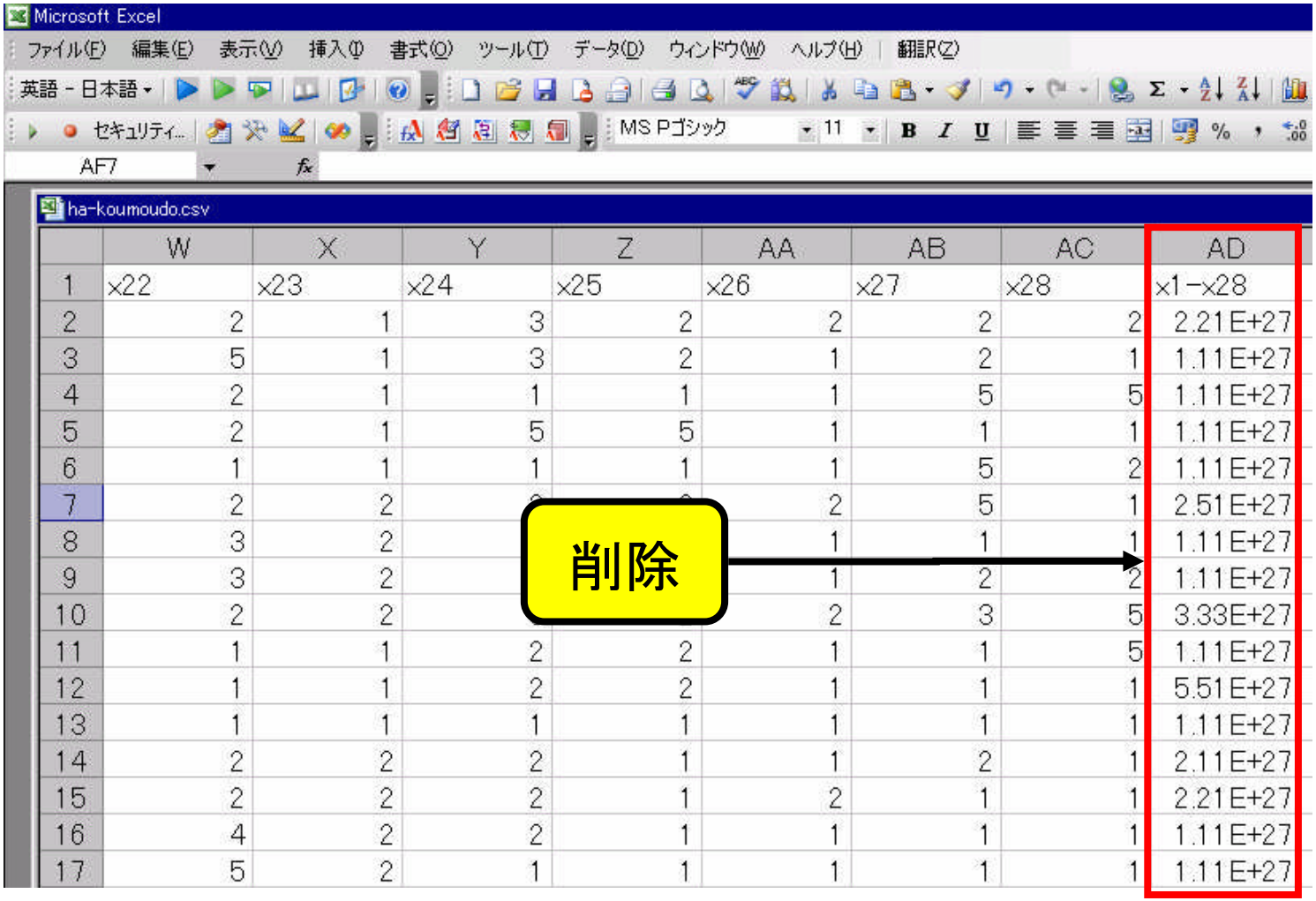
ha-koumoudo.csv

	A	B	C	D
1	age	x1	x2	x3
2	21	2	2	1
3	21	1	1	1
4	21	1	1	1
5	21	1	1	1
6	22	1	1	1
7	22	2	5	1
8	22	1	1	1
9	23	1	1	1
10	23	3	3	3
11	23	1	1	1
12	23	5	5	1
13	24	1	1	1
14	24	2	1	1
15	24	2	2	1
16	24	1	1	1

…

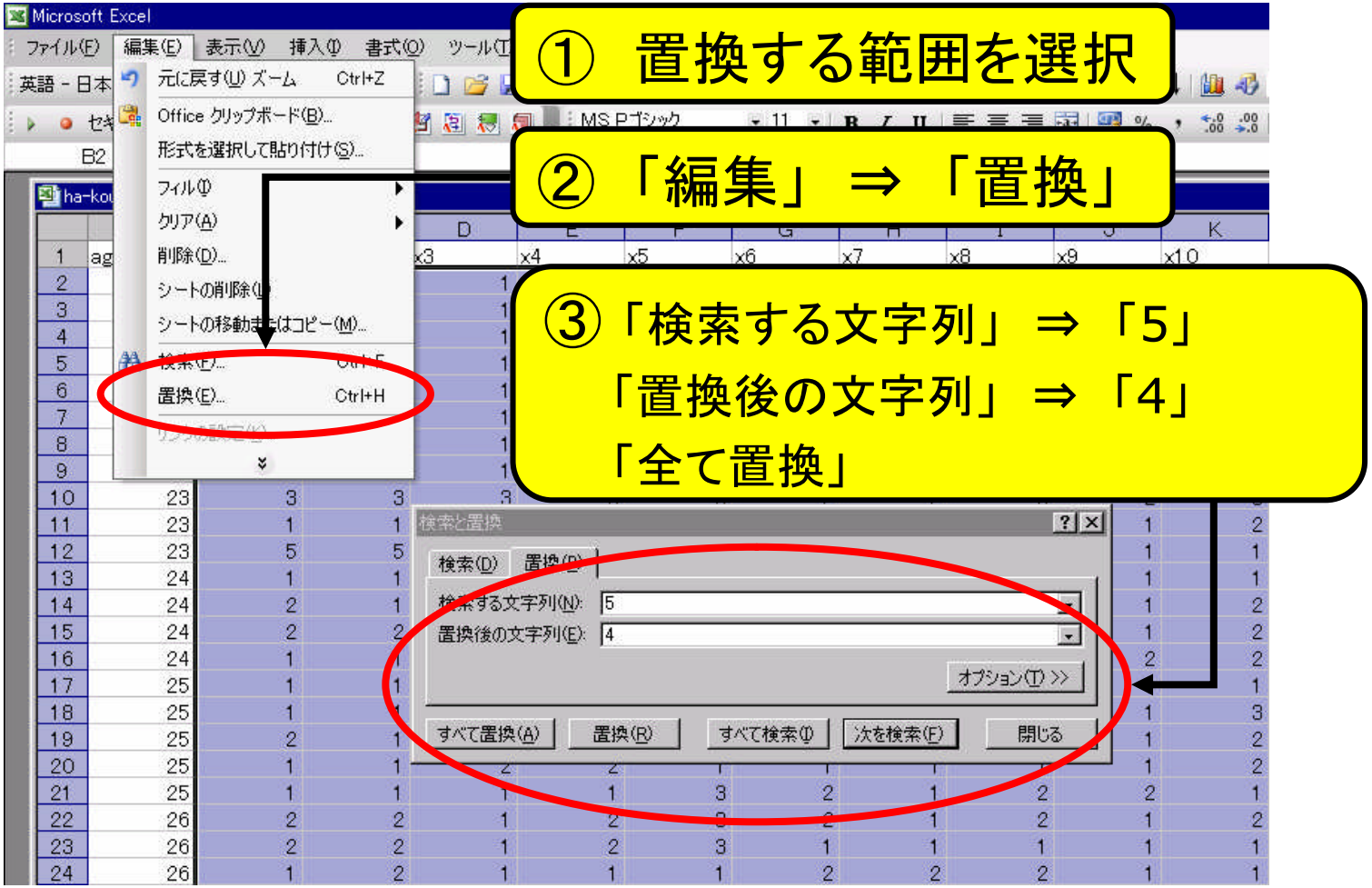
AA	AB	AC	AD
x26	x27	x28	x1−x28
2	2	2	2.21E+27
1	2	1	1.11E+27
1	5	5	1.11E+27
1	1	1	1.11E+27
1	5	2	1.11E+27
2	5	1	2.51E+27
1	1	1	1.11E+27
1	2	2	1.11E+27
2	3	5	3.33E+27
1	1	5	1.11E+27
1	1	1	5.51E+27
1	1	1	1.11E+27
1	2	1	2.11E+27
2	1	1	2.21E+27
1	1	1	1.11E+27

データ加工：不要なデータの削除



	W	X	Y	Z	AA	AB	AC	AD
1	x22	x23	x24	x25	x26	x27	x28	x1−x28
2	2	1	3	2	2	2	2	2.21E+27
3	5	1	3	2	1	2	1	1.11E+27
4	2	1	1	1	1	5	5	1.11E+27
5	2	1	5	5	1	1	1	1.11E+27
6	1	1	1	1	1	5	2	1.11E+27
7	2	2			2	5	1	2.51E+27
8	3	2			1	1	1	1.11E+27
9	3	2			1	2	2	1.11E+27
10	2	2			2	3	5	3.33E+27
11	1	1	2	2	1	1	5	1.11E+27
12	1	1	2	2	1	1	1	5.51E+27
13	1	1	1	1	1	1	1	1.11E+27
14	2	2	2	1	1	2	1	2.11E+27
15	2	2	2	1	2	1	1	2.21E+27
16	4	2	2	1	1	1	1	1.11E+27
17	5	2	1	1	1	1	1	1.11E+27

データ加工：特定のデータの置換



プログラム

- 「ディレクトリの変更」を忘れずにしましょう

```
koumoudo <- read.csv(“ha-koumoudo.csv", header=T)

result1 <- lm(age~., data=koumoudo)
result2 <- step(result1)

summary(result1)
summary(result2)
```

プログラムの説明

```
result1 <- lm(age~., data=koumoudo)
result2 <- step(result1)
```

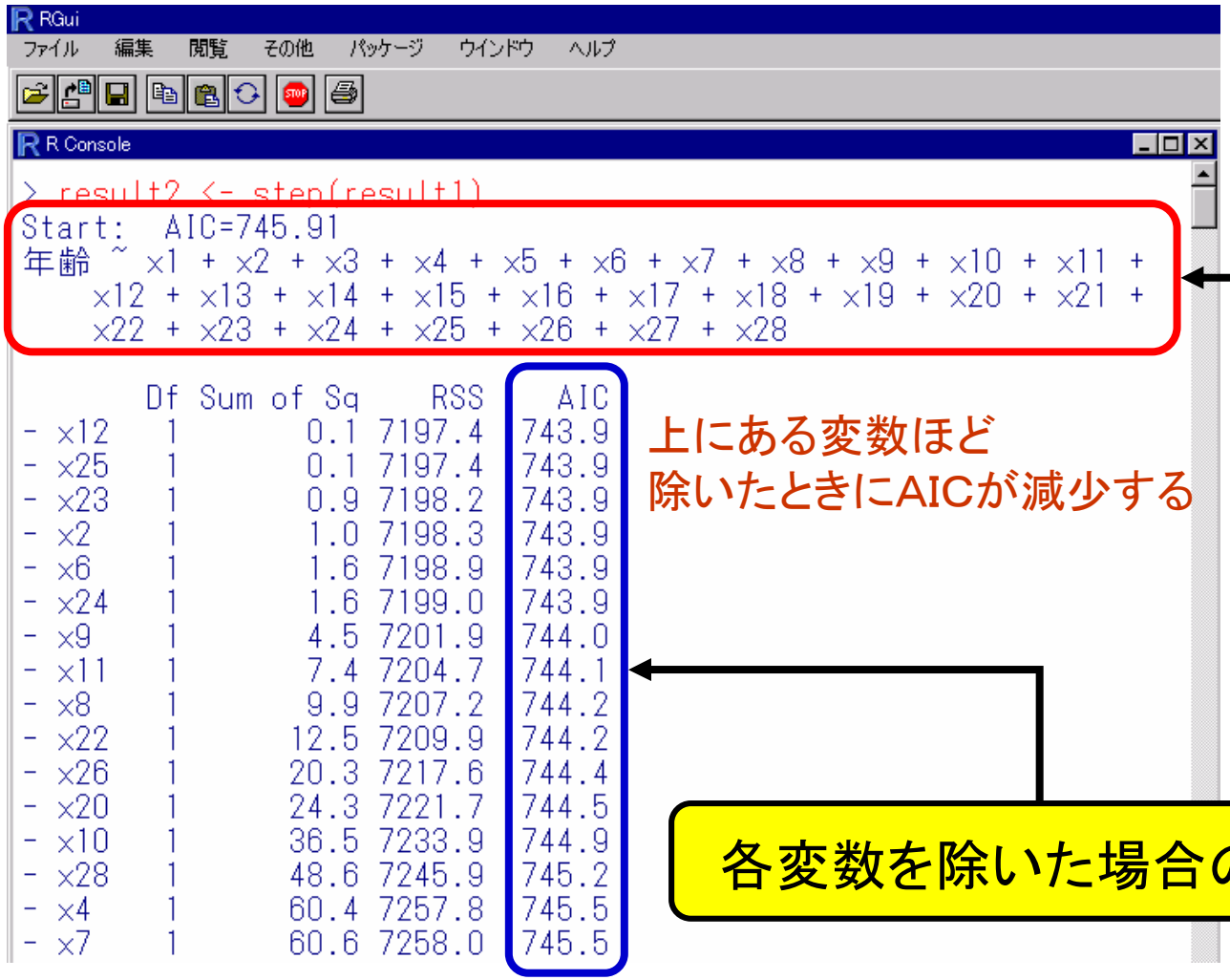
- lm(目的変数~., data=データ)
  - 「目的変数~.」とすると、目的変数以外の全ての変数を説明変数として分析を行う
- step(lm.obj)
  - lm関数により得たモデルに対して、AIC基準で変数選択を行う関数

変数選択基準

- 変数選択基準
  - 残差平方和
  - 決定係数・自由度調整済み決定係数
  - 各変数に対する有意性検定（ｔ検定・F検定）
  - AIC基準　AIC最小のモデルを最適なモデルとする
- 参考文献
  - 杉山 髙一著　：「多変量データ解析入門」
  - 小西 貞則著　：「情報量規準」
  - 早川 毅著　　：「回帰分析の基礎」

step関数

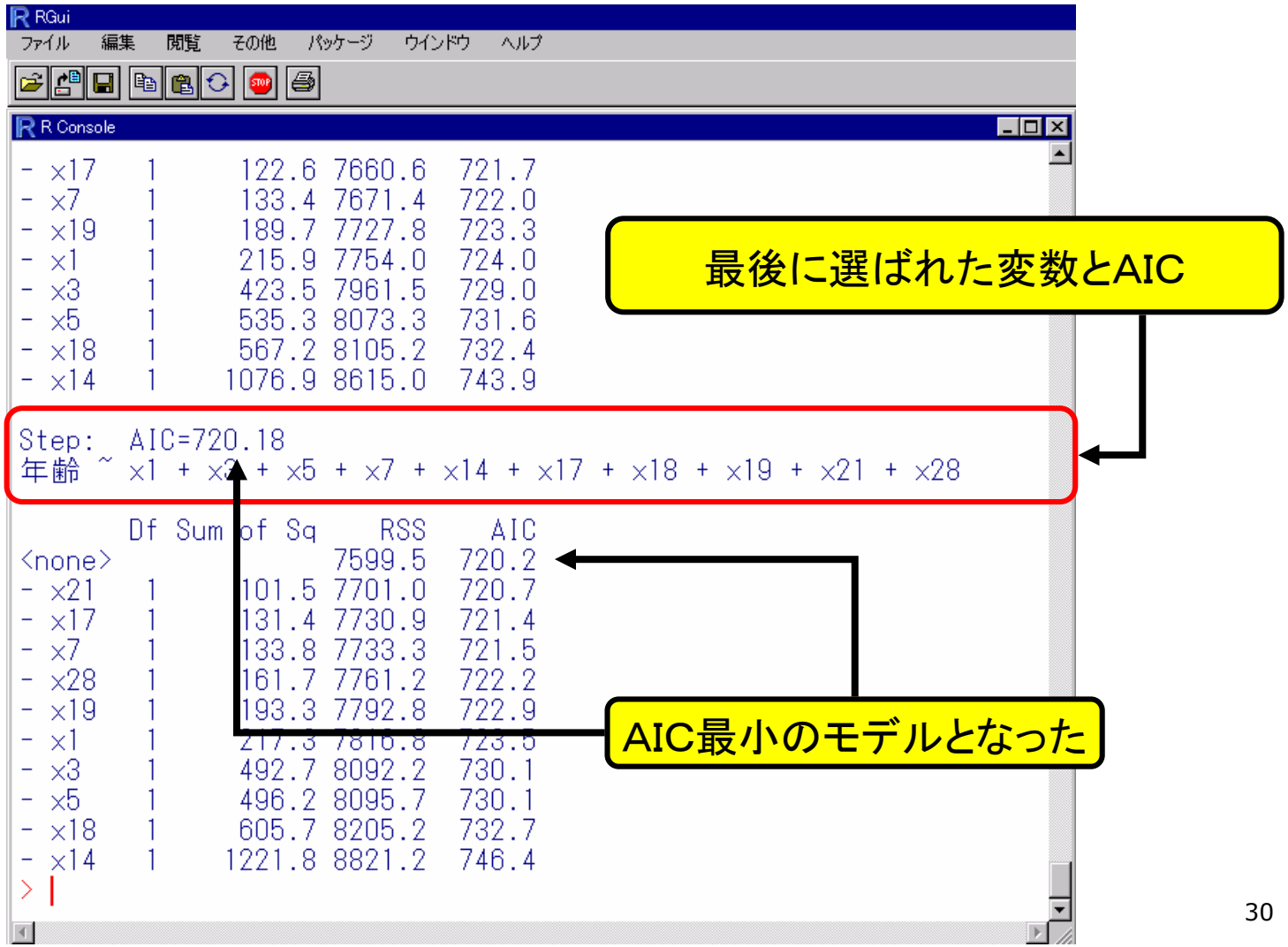
全ての変数を含んだモデルと、そのAIC



上にある変数ほど
除いたときにAICが減少する

各変数を除いた場合のAIC

step関数


RGui
ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ
R Console
- x17 1 122.6 7660.6 721.7
- x7 1 133.4 7671.4 722.0
- x19 1 189.7 7727.8 723.3
- x1 1 215.9 7754.0 724.0
- x3 1 423.5 7961.5 729.0
- x5 1 535.3 8073.3 731.6
- x18 1 567.2 8105.2 732.4
- x14 1 1076.9 8615.0 743.9
Step: AIC=720.18
年齢 ~ x1 + x3 + x5 + x7 + x14 + x17 + x18 + x19 + x21 + x28
Df Sum of Sq RSS AIC
<none> 7599.5 720.2
- x21 1 101.5 7701.0 720.7
- x17 1 131.4 7730.9 721.4
- x7 1 133.8 7733.3 721.5
- x28 1 161.7 7761.2 722.2
- x19 1 193.3 7792.8 722.9
- x1 1 217.3 7816.8 723.5
- x3 1 492.7 8092.2 730.1
- x5 1 496.2 8095.7 730.1
- x18 1 605.7 8205.2 732.7
- x14 1 1221.8 8821.2 746.4
>
最後に選ばれた変数とAIC
AIC最小のモデルとなった

分析結果：変数選択前1

R Console

```
> summary(result1)

Call:
lm(formula = 年齢 ~ ., data = koumoudo)

Residuals:
     Min       1Q   Median       3Q      Max
-25.1007  -4.1198  -0.3587   4.1069  17.5973

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.23351    2.84803   1.135 0.257928
x1           1.30429    0.77118   1.691 0.092730 .
x2          -0.09868    0.66918  -0.147 0.882946
x3           2.60590    0.85306   3.055 0.002639 **
x4          -0.90700    0.78242  -1.159 0.248092
x5           2.39122    1.02422   2.335 0.020802 *
x6          -0.17492    0.93010  -0.188 0.851062
x7           1.10322    0.95021   1.161 0.247362
x8           0.42742    0.91023   0.470 0.639299
x9           0.25651    0.80690   0.318 0.750973
x10          0.85730    0.95144   0.901 0.368913
x11          0.34070    0.84285   0.404 0.686587
```

分析結果：変数選択前2

R Console

```
x12          0.03559    0.83469   0.043 0.966047
x13          0.72367    0.59708   1.212 0.227293
x14          2.74277    0.69398   3.952 0.000116 ***
x15          0.87986    0.65966   1.334 0.184159
x16         -0.81182    0.68504  -1.185 0.237748
x17          1.13763    0.71870   1.583 0.115421
x18          2.82679    0.82123   3.442 0.000737 ***
x19         -1.61175    1.19751  -1.346 0.180234
x20         -0.68798    0.93543  -0.735 0.463126
x21          1.28287    0.97422   1.317 0.189783
x22         -0.46201    0.87524  -0.528 0.598321
x23         -0.11959    0.85463  -0.140 0.888889
x24          0.19930    1.04952   0.190 0.849629
x25          0.03549    0.83081   0.043 0.96
x26         -0.49507    0.73751  -0.671 0.50
x27          1.00290    0.68716   1.459 0.14
x28          0.70950    0.68281   1.039 0.300335
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.707 on 160 degrees of freedom
Multiple R-squared: 0.7944,     Adjusted R-squared: 0.7584
F-statistic: 22.08 on 28 and 160 DF,  p-value: < 2.2e-16
```

自由度調整済み決定係数

分析結果：変数選択後

R Console

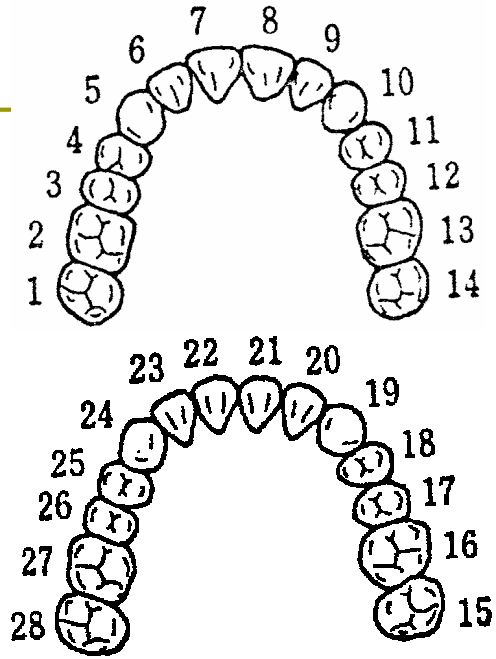
```
> summary(result2)

Call:
lm(formula = 年齢 ~ x1 + x3 + x5 + x7 + x14 + x17 + x18 + x19 +
    x21 + x28, data = koumoudo)

Residuals:
     Min       1Q   Median       3Q      Max
-24.3694  -5.0871  -0.1068   4.0291  17.6537

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.7750     2.0810   2.295 0.022925 *
x1            1.4693     0.6513   2.256 0.025292 *
x3            2.3844     0.7019   3.397 0.000840 ***
x5            2.8666     0.8409   3.409 0.000806 ***
x7            1.2176     0.6877   1.771 0.078345 .
x14           3.2680     0.6109   5.349 2.69e-07 ***
x17           1.1497     0.6554   1.754 0.081116 .
x18           2.7463     0.7291   3.767 0.000225 ***
x19          -2.1144     0.9938  -2.128 0.034745 *
x21           1.0810     0.7012   1.542 0.124959
x28           1.0872     0.5587   1.946 0.053232 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.534 on 178 degrees of freedom
Multiple R-squared: 0.7829,     Adjusted R-squared: 0.7707
F-statistic:  64.2 on 10 and 178 DF,  p-value: < 2.2e-16
```

変数選択により改善

解析結果の比較

変数選択	変数の数	選ばれた変数
前	28	
後	10	1, 3, 5, 7, 14, 17, 18, 19, 21, 28

変数選択	自由度調整済み決定係数	AIC
前	0.7584	745.91
後	0.7707	720.18



変数選択

- 適切に変数を選択することにより、モデルが改善された
- 変数が減ることにより、意味づけや解釈が容易になる
- データを収集する側にもメリット（コストや時間）

参考URL

- 統計科学研究所のウェブサイト
  http://www.statistics.co.jp/index.htm
- R-Tips
  http://cse.naro.affrc.go.jp/takezawa/r-tips/r2.html
- JIN'S PAGE
  http://www1.doshisha.ac.jp/~mjin/R/