R と RStudio の使い方

芳賀敏郎（2011）医薬品開発のための統計解析　第 1 部　基礎
4 相関・回帰
4.5 回帰分析適用上の諸問題

テキストと利用上の注意

●テキスト

芳賀敏郎（2011）医薬品開発のための統計解析
第１部　基礎　改訂版、サイエンティスト社、p.275
（サイトへアップすることに対して、サイエンティスト社の了解を得ています）

●R による解析事例を紹介

R スクリプトの出力結果を紹介します（tidyverse 系には次期バージョンで対応します）
R スクリプト（文字コードUTF-8に設定）を、このサイトからダウンロードできます
R スクリプトを [Compile Report] することにより、Word または HTML で見ることができます
R と RStudio の設定と基本的な使い方は「R と RStudio の使い方」を参照してください
R の出力結果の見方は、テキストとそれを解説した PDF ファイルを参照してください
グラフ表示は、解析手段として、必要最小限の表現に止めています

●自己責任で利用

上記のことを理解した上で、自己責任により利用してください

第１部　基礎

1. 統計の基礎・・・・・1.1 宝くじの期待値と分散、1.2 サイコロの目の数の期待値と分散
1.3 分散の加法性・中心極限定理・正規分布、1.4 統計的推測、1.5 モデル

2. １組のデータの解析　2.1 データの特徴の記述、2.2 データのグラフ表示と外れ値
2.3 対数変換と対数正規分布、2.4 平均に関する推測（母標準偏差 σ 既知）
2.5 分散に関する推測、2.6 平均に関する推測（母標準偏差 σ 未知）

3. ２組のデータの解析　3.1 データのグラフ化、3.2 平均値の差の t 検定、3.3 分散の違いの検定
3.4 分散が異なる場合の平均値の差の比較
3.5 対応のある場合の平均値の差の t 検定、3.6 検出力と n の決め方
3.7 ノンパラメトリック検定

4. 相関・回帰・・・・・4.1 散布図、4.2 相関係数、4.3 回帰モデルとモデルの推定
4.4 誤差を考慮した推定、**4.5 回帰分析適用上の諸問題**

Anscombe の例

- 表示4.5.1 Anscombe の例

  スクリプトファイル：Green1-4-5a.R

  利用した関数

  as.matrix、apply、rbind、mean、sd

  signif、readxl::reas_excel

  方法

  Excelファイルからデータを読込

  データを データフレーム df に付値

  as.matrix 関数で df から行列を作成

  mean 関数とsd 関数で計算

  signif 関数で数値を丸めて表示

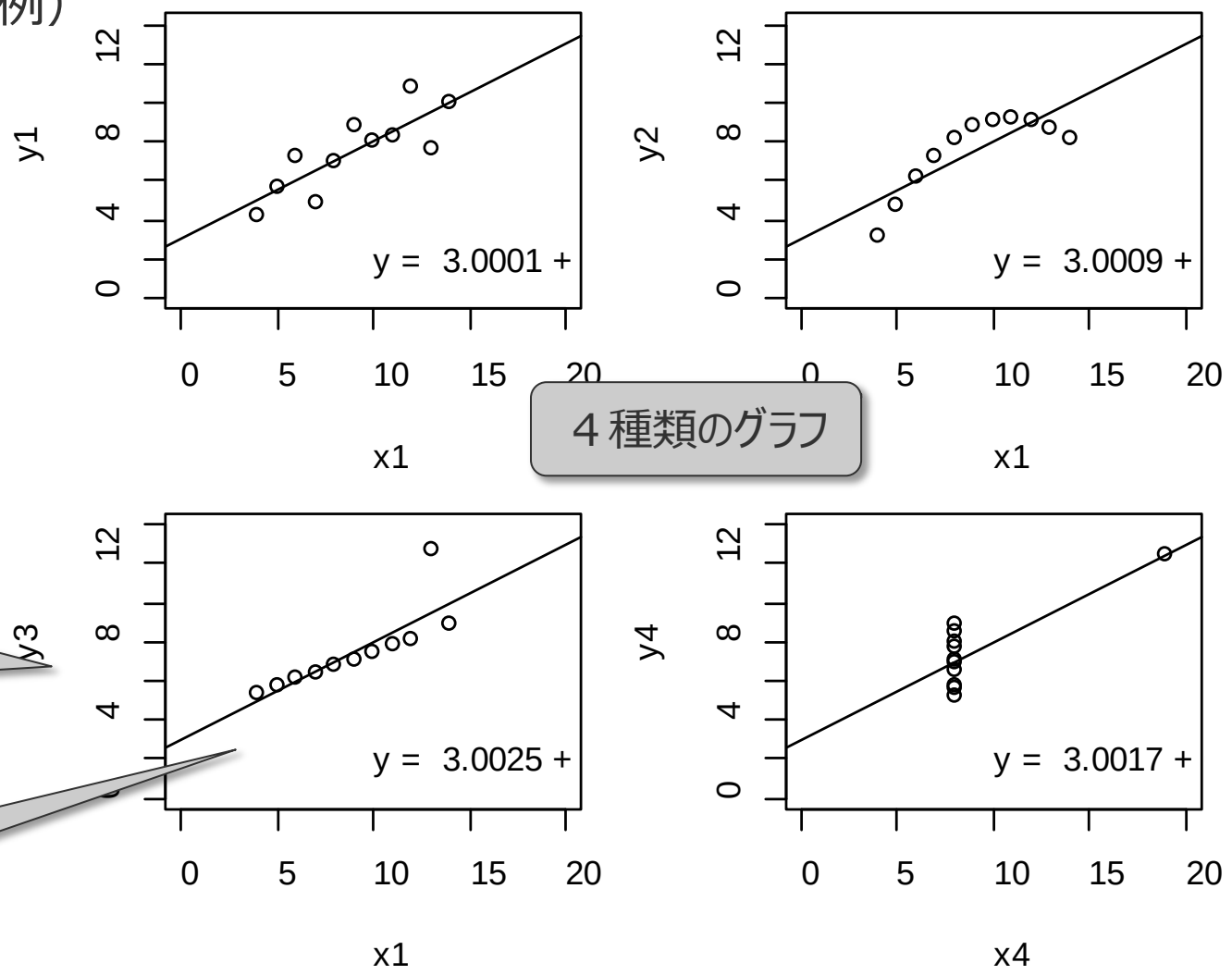

```
mn <- apply(mx, 2, mean)
```

```
df <- read_excel("Green1-4.xlsx",
                 sheet = "4-Anscombe")
df <- data.frame(df)

mx <-  as.matrix(df)
mn <- apply(mx, 2, mean)
std <- apply(mx, 2, sd)
signif(rbind(mx, "mean" = mn, "SD" = std),
       digits = 4)
```

```
##          x1     y1    y2    y3     x4     y4
##      10.000  8.040 9.140  7.46  8.000  6.580
##       8.000  6.950 8.140  6.77  8.000  5.760
##      13.000  7.580 8.740 12.74  8.000  7.710
##           . . . . . . . . . . . . . .
##       7.000  4.820 7.260  6.42  8.000  7.910
##       5.000  5.680 4.740  5.73  8.000  6.890
## mean  9.000  7.501 7.501  7.50  9.000  7.501
## SD    3.317  2.032 2.032  2.03  3.317  2.031
```

Anscombe の例

- 表示4.5.2 散布図と回帰直線（Anscombe の例）
  - スクリプトファイル：Green1-4-5a.R
  - 利用した関数
    - lm、plot、abline、for、list、rep
    - ifelse
  - 方法
    - lm 関数で回帰分析
    - ４つのグラフを for 文で繰り返し描画



４種類のグラフ

xとy
の組み合わせ

回帰式

Anscombe の例

- 表示4.5.2 散布図と回帰直線（Anscombe の例）
  スクリプトファイル：Green1-4-5a.R

```
mdl <- list(as.formula(y1 ~ x1),
            as.formula(y2 ~ x1),
            as.formula(y3 ~ x1),
            as.formula(y4 ~ x4))
```

リスト mdl の要素

```
> mdl
[[1]]
y1 ~ x1

[[2]]
y2 ~ x1

[[3]]
y3 ~ x1

[[4]]
y4 ~ x4
```

リストの要素は
[[番号]] で指定

```
##            x1     y1    y2    y3     x4     y4
##        10.000  8.040 9.140  7.46  8.000  6.580
##         8.000  6.950 8.140  6.77  8.000  5.760
##        13.000  7.580 8.740 12.74  8.000  7.710
##         9.000  8.810 8.770  7.11  8.000  8.840
##        11.000  8.330 9.260  7.81  8.000  8.470
##        14.000  9.960 8.100  8.84  8.000  7.040
##         6.000  7.240 6.130  6.08  8.000  5.250
##         4.000  4.260 3.100  5.39 19.000 12.500
##        12.000 10.840 9.130  8.15  8.000  5.560
##         7.000  4.820 7.260  6.42  8.000  7.910
##         5.000  5.680 4.740  5.73  8.000  6.890
## mean    9.000  7.501 7.501  7.50  9.000  7.501
## SD      3.317  2.032 2.032  2.03  3.317  2.031
```

Anscombe の例

- 表示4.5.2 散布図と回帰直線（Anscombe の例）
  スクリプトファイル：Green1-4-5a.R

mdl[[1]]
mdl[[2]]
mdi[[3]]
mdl[[4]]

```
mdl <- list(as.formula(y1 ~ x1),
            as.formula(y2 ~ x1),
            as.formula(y3 ~ x1),
            as.formula(y4 ~ x4))
```

空のリストを作成

４要素からなる
ベクトルを作成

(1) 回帰分析の結果を付値
リストを要素とするリスト

(2) 回帰係数と切片の取り出し
ベクトル

(3) 回帰式を文字列で付値

```
lm_out <- list()    # 回帰分析の結果を付値するリスト

lab <- rep(" ", 4)  # 回帰式を付値するベクトル

for (i in 1:4) {
  lm_out[[i]] <- lm(mdl[[i]], data = df)

  res <- round(lm_out[[i]][[1]], digits = 4)

  lab[i] <- paste("y = ", res[1],
                  ifelse(res[2] > 0, "+", "-"),
                  abs(res[2]),"x")
}
```

Anscombe の例

- 表示4.5.2 散布図と回帰直線（Anscombe の例）
  スクリプトファイル：Green1-4-5a.R

```
リスト lm_out の構造
  lm_out[[1]]
      [[1]] 切片と傾き
      [[2]] 残差
      . . . . .
  lm_out[[2]]
      [[1]] 切片と傾き
      [[2]] 残差
      . . . . .
  lm_out[[3]]
      [[1]] 切片と傾き
      . . . . .
  lm_out[[4]]
      [[1]] 切片と傾き
      . . . . .
```

mdl[[1]]
mdl[[2]]
mdi[[3]]
mdl[[4]]

```
mdl <- list(as.formula(y1 ~ x1),
            as.formula(y2 ~ x1),
            as.formula(y3 ~ x1),
            as.formula(y4 ~ x4))
```

```
lm_out <- list()    # 回帰分析の結果を付値するリスト

lab <- rep(" ", 4) # 回帰式を付値するベクトル

for (i in 1:4) {
  lm_out[[i]] <- lm(mdl[[i]], data = df)

  res <- round(lm_out[[i]][[1]], digits = 4)

  lab[i] <- p                es[1],
                             s[2] > 0,
                        abs(res[2]),"x")
}
```

res[1]：切片
res[2]：傾き

lm_out[[i]]の
切片と傾き

Anscombe の例

- 表示4.5.2 散布図と回帰直線（Anscombe の例）
  スクリプトファイル：Green1-4-5a.R

```
mdl <- list(as.formula(y1 ~ x1),
            as.formula(y2 ~ x1),
            as.formula(y3 ~ x1),
            as.formula(y4 ~ x4))
```

空のリストを作成

４要素からなる
ベクトルを作成

(1) 回帰分析の結果を付値
リストを要素とするリスト

(2) 回帰係数と切片の取り出し
ベクトル

(3) 回帰式を文字列で付値

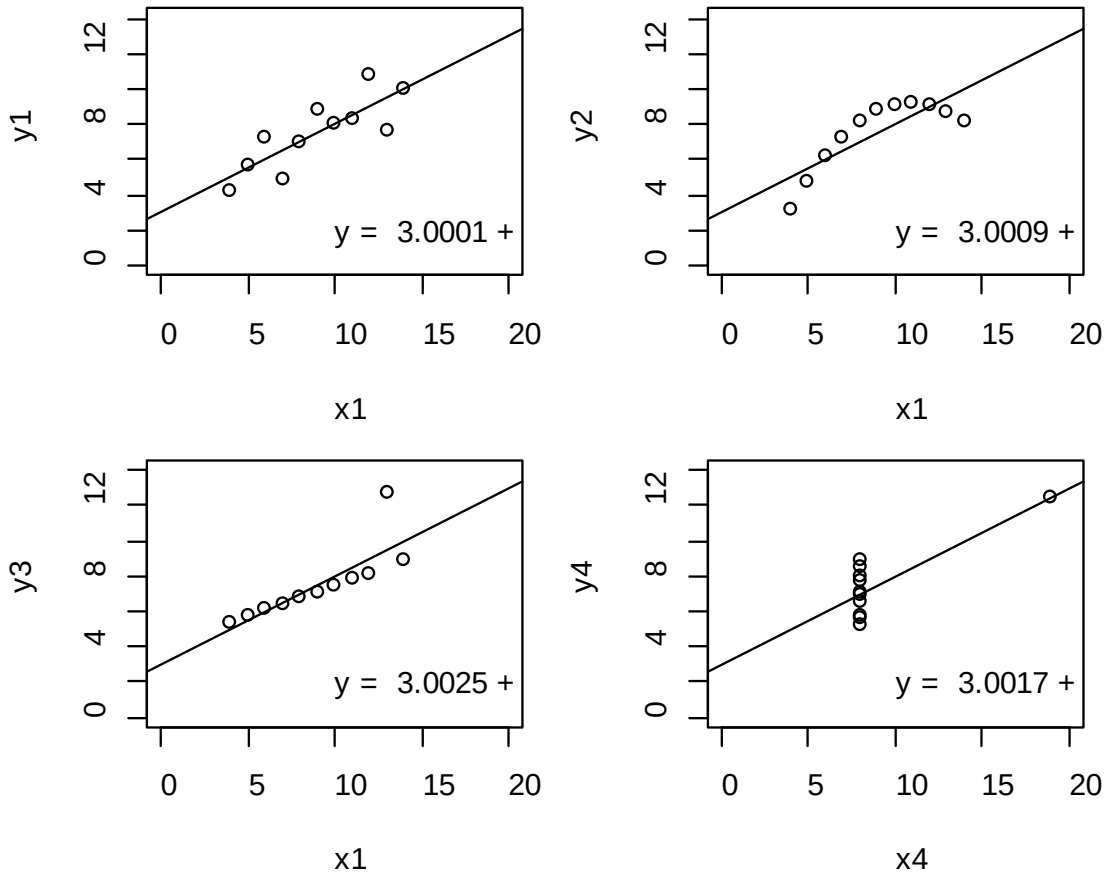
```
lm_out <- list()    # 回帰分析の結果を付値するリスト

lab <- rep(" ", 4)  # 回帰式を付値するベクトル

for (i in 1:4) {
  lm_out[[i]] <- lm(mdl[[i]], data = df)

  res <- round(lm_out[[i]][[1]], digits = 4)

  lab[i] <- paste("y = ", res[1],
                  ifelse(res[2] > 0, "+", "-"),
                  abs(res[2]),"x")
}
```

res[1]：切片

res[2]：傾き

Anscombe の例

- 表示4.5.2 散布図と回帰直線（Anscombe の例）
  スクリプトファイル：Green1-4-5a.R



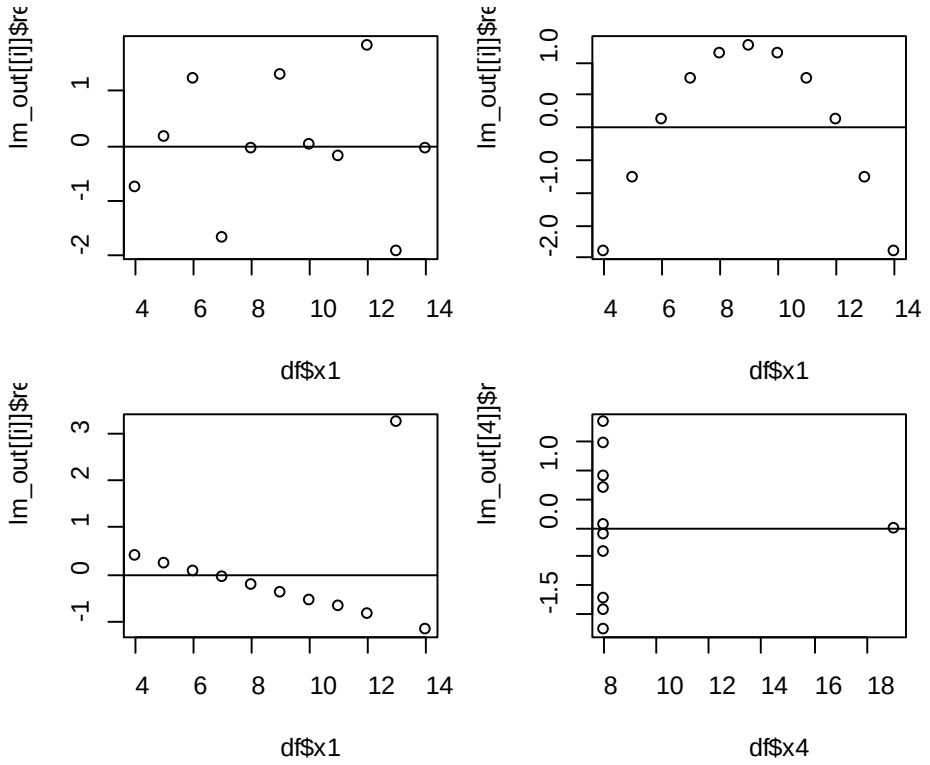

```
mdl <- list(as.formula(y1 ~ x1),
            as.formula(y2 ~ x1),
            as.formula(y3 ~ x1),
            as.formula(y4 ~ x4))
```

```
par(mfrow = c(2, 2)) # 画面を４分割
par(mar = c(4, 4, 1, 1)) # 余白
                         # §4.1 参照

for (i in 1:4) {
  plot(mdl[[i]], data = df,
       xlim = c(0, 20),
       ylim = c(0,14))
  abline(lm_out[[i]])
  text(x = 10, y = 2,
       label = lab[i])
}
```

Anscombe の例

- 表示4.5.2 散布図と回帰直線（Anscombe の例）
  - スクリプトファイル：Green1-4-5a.R
  - 利用した関数
    - lm、summary
    - plot、abline
  - 方法
    - lm 関数で回帰分析
    - rbind 関数で切片、傾き、決定係数を表示
    - plot 関数で残差をプロット

```
##          Intercept        x1 R-squared
## y1 ~ x1   3.000091 0.5000909 0.6665425
## y2 ~ x1   3.000909 0.5000000 0.6662420
## y3 ~ x1   3.002455 0.4997273 0.6663240
## y4 ~ x4   3.001727 0.4999091 0.6667073
```



残差の検討

- 回帰診断図

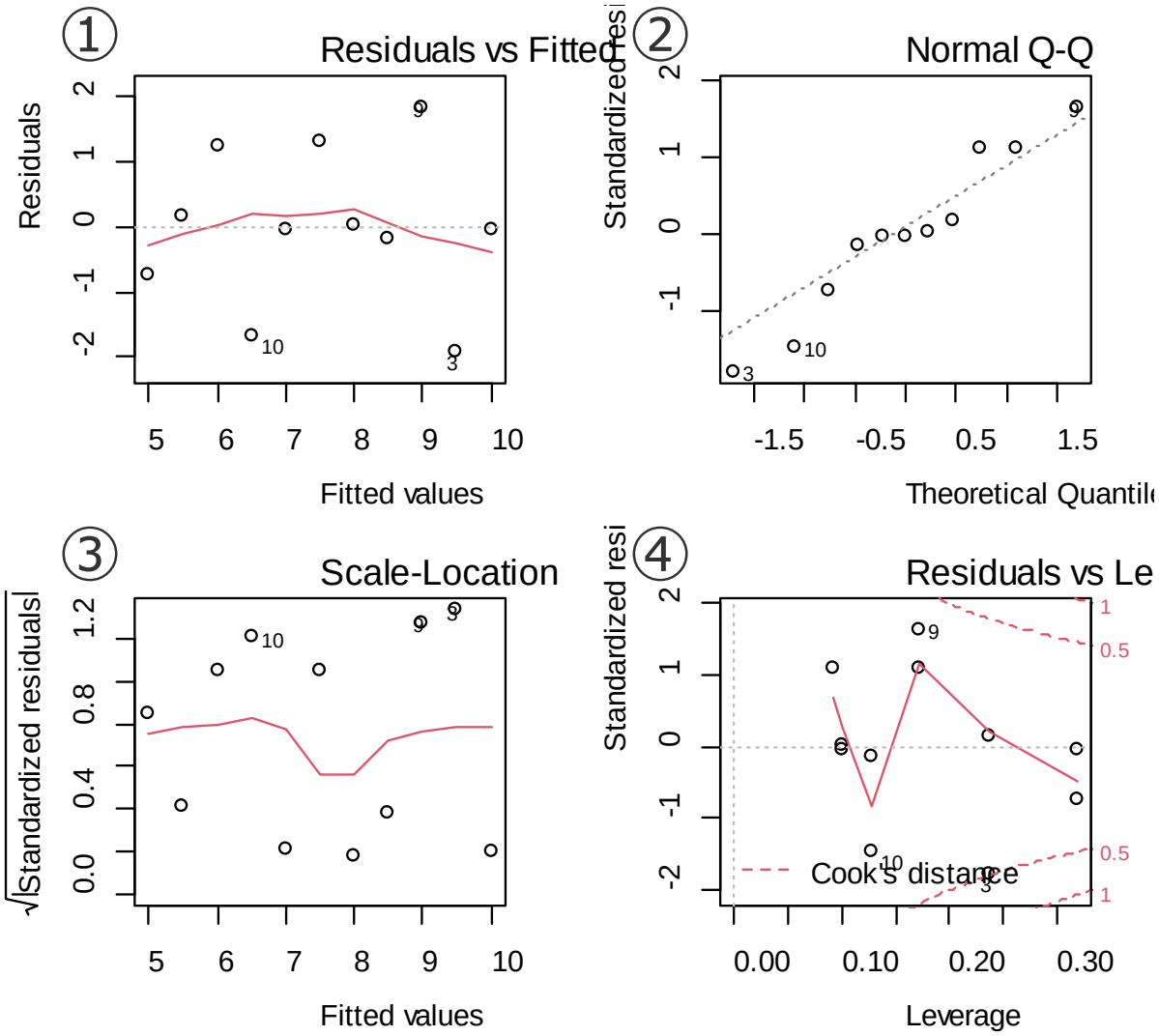
  スクリプトファイル：Green1-4-5a.R

  利用した関数：lm、plot

  方法

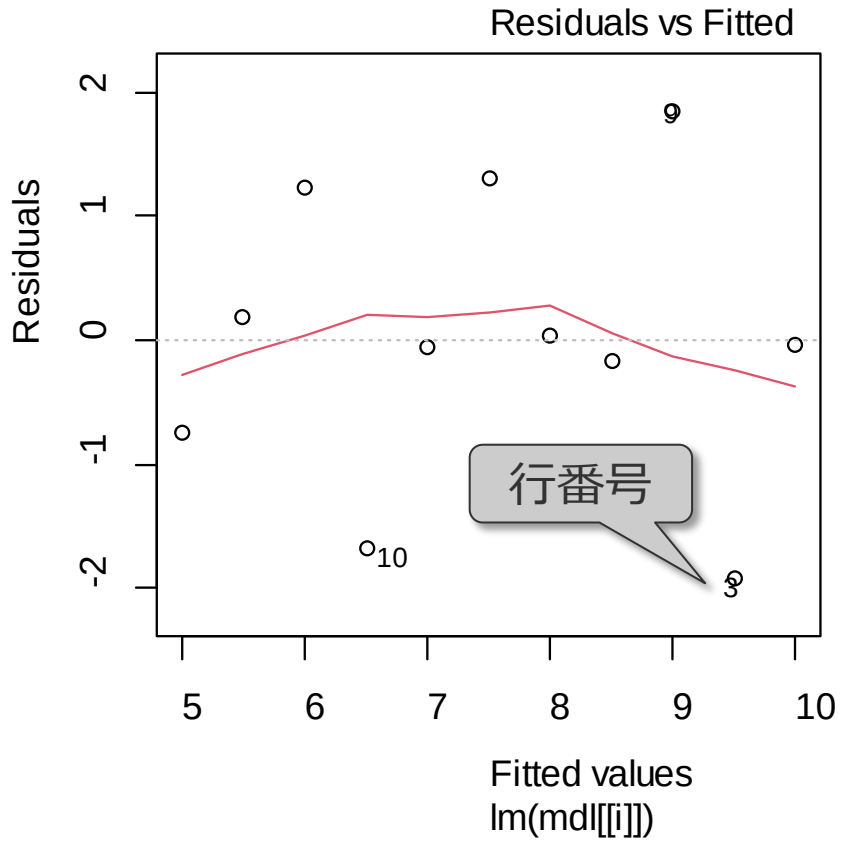
```
lm_out[[1]] <- lm(mdl[[1]], data = df)

par(mfrow = c(2, 2))
plot(lm_out[[1]])

par(mfrow = c(1, 1))
plot(lm_out[[1]], which = 1) #①
plot(lm_out[[1]], which = 2) #②
plot(lm_out[[1]], which = 3) #③
plot(lm_out[[1]], which = 4)
plot(lm_out[[1]], which = 5) #④
plot(lm_out[[1]], which = 6)
```

plot.lm のヘルプを参照

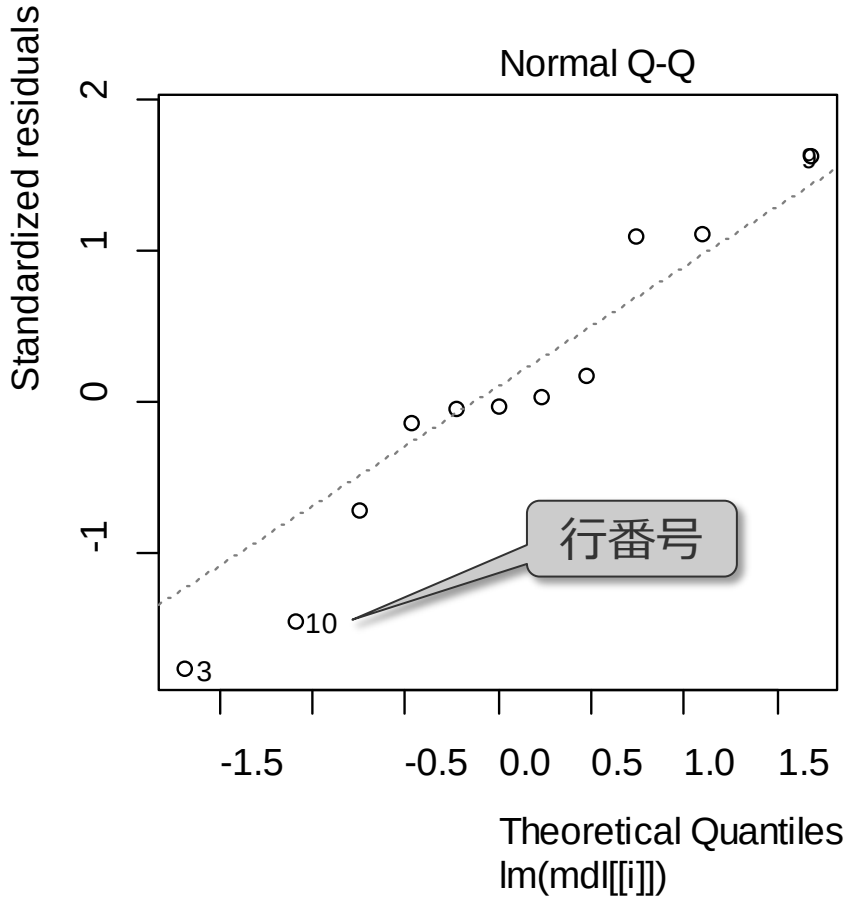


残差の検討


```
plot(lm_out[[1]], which = 1)
#残差と予測値
```

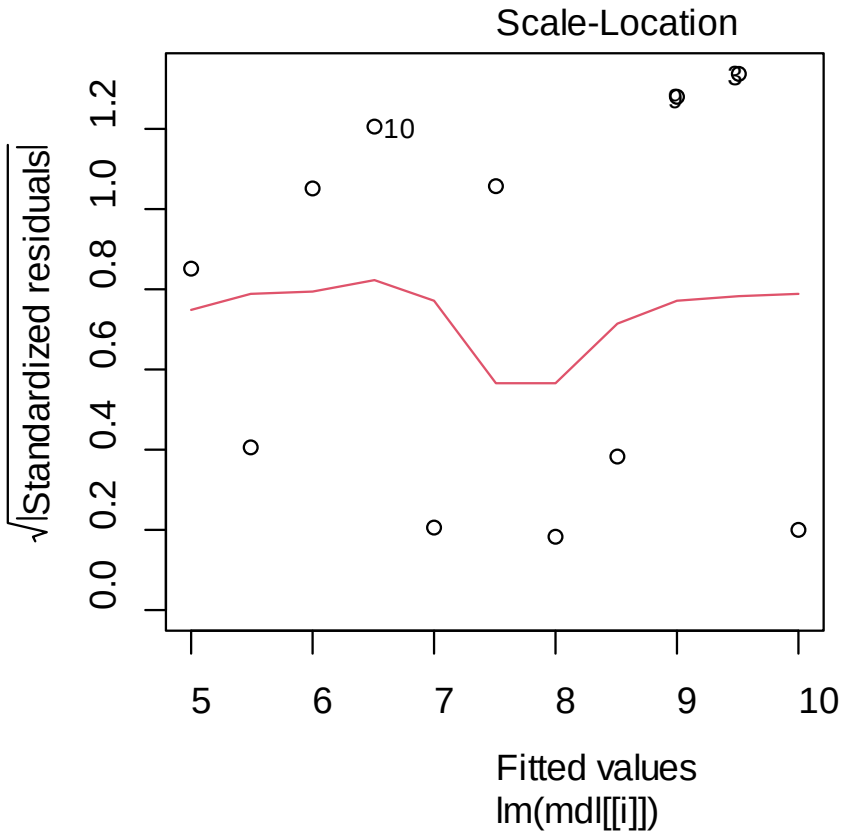



```
plot(lm_out[[1]], which = 2)
#基準化残差の正規Q-Qプロット
```

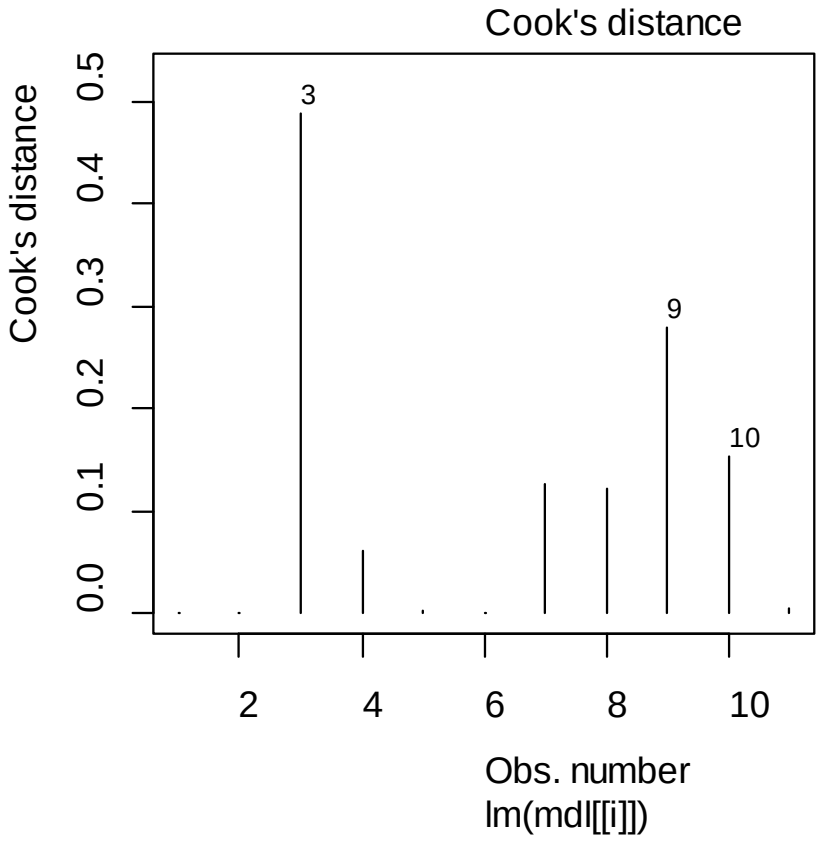


残差の検討


```
plot(lm_out[[1]], which = 3)
#基準化残差の絶対値の平方根と予測値
```

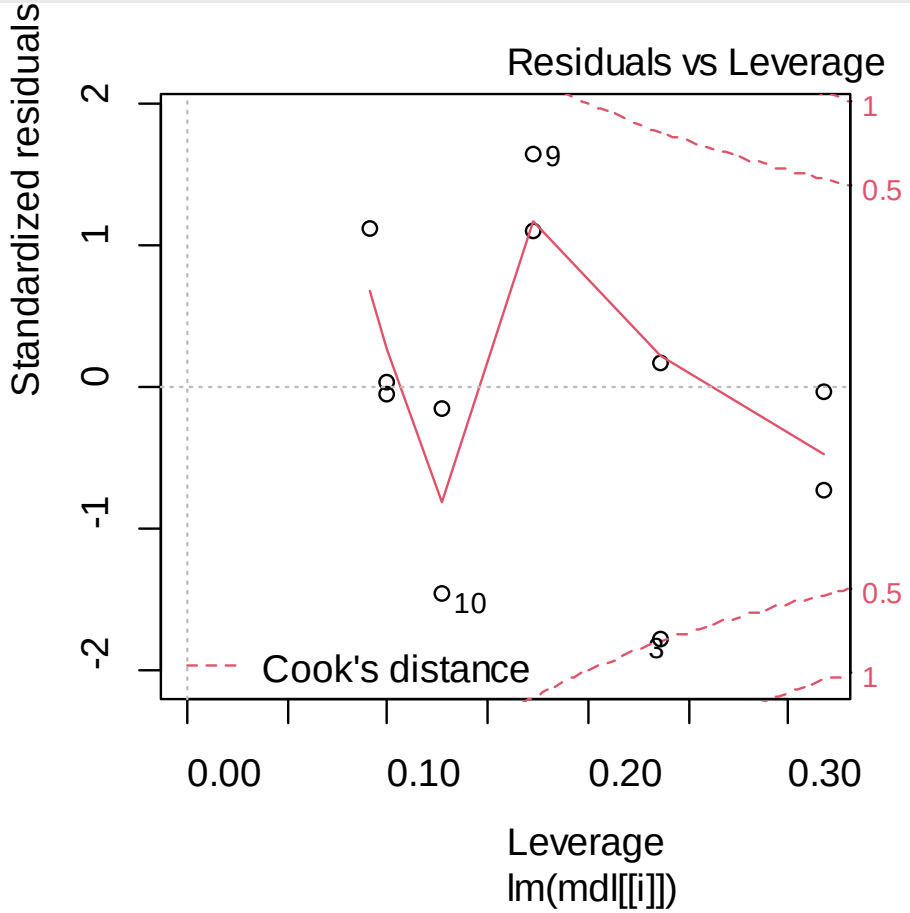



```
plot(lm_out[[1]], which = 4)
# クックの距離
```

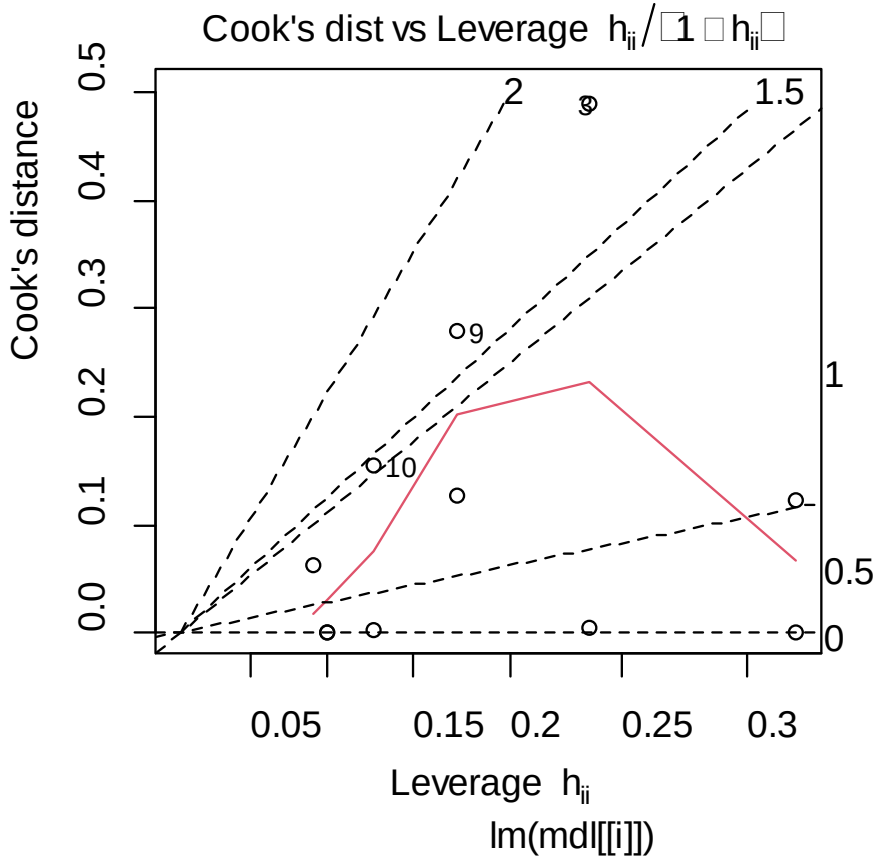


残差の検討


```
plot(lm_out[[1]], which = 5)
#基準化残差とてこ比
```




```
plot(lm_out[[1]], which = 6)
# クックの距離とてこ比/(1 −てこ比)
```



変数の選び方(1) 血糖と血圧

- 変数の選び方(1)　血糖と血圧
  - スクリプトファイル：Green1-4-5b.R
  - 利用した関数：cor
  - 方法：cor 関数で相関係数行列を出力

```
df <- read_excel("Green1-4.xlsx",
                 sheet = "4-ensyu")
df <- data.frame(df)

df1 <- df[c("SBP", "DBP", "BS")]
del <- df$SBP - df$DBP
df2 <- data.frame(df1, DELTA = del)
##    SBP DBP  BS DELTA
## 1  126  78  95    48
## 2  104  70  88    34
      . . . . . . .
## 39 102  68  80    34
## 40 138  70 112    68
```

```
cor(df1)

##            SBP        DBP          BS
## SBP 1.0000000 0.76542352 0.36089821
## DBP 0.7654235 1.00000000 0.05811234
## BS  0.3608982 0.05811234 1.00000000
```

```
cor(df2)

##              SBP        DBP          BS      DELTA
## SBP   1.0000000 0.76542352 0.36089821 0.65628864
## DBP   0.7654235 1.00000000 0.05811234 0.01679135
## BS    0.3608982 0.05811234 1.00000000 0.49259947
## DELTA 0.6562886 0.01679135 0.49259947 1.00000000
```

DELTA = SBP − DBP
（上の血圧－下の血圧）

変数の選び方(1) 血糖と血圧

- 変数の選び方(1)　血糖と血圧
  スクリプトファイル：Green1-4-5b.R
  利用した関数：lm、stargazer::stargazer
  方法：lm 関数で回帰分析

```
df <- read_excel("Green1-4.xlsx",
                 sheet = "4-ensyu")
df <- data.frame(df)

df1 <- df[c("SBP", "DBP", "BS")]
del <- df$SBP - df$DBP
df2 <- data.frame(df1, DELTA = del)
##    SBP DBP  BS DELTA
## 1  126  78  95    48
## 2  104  70  88    34
      . . . . . .
## 39 102  68  80    34
## 40 138  70 112    68
```

```
lm_out_s <- lm(BS ~ SBP, data = df2)

lm_out_d <- lm(BS ~ DELTA, data = df2)
stargazer(lm_out_s, lm_out_d, type = "text")
```

変数の選び方(1) 血糖と血圧

- 変数の選び方(1)　血糖と血圧
  スクリプトファイル：Green1-4-5b.R
  利用した関数：lm、stargazer::stargazer
  方法：lm 関数で回帰分析
  stargrazer 関数で
  ２種類の回帰分析の
  結果を出力

```
lm_out_s <- lm(BS ~ SBP, data = df2)

lm_out_d <- lm(BS ~ DELTA, data = df2)
stargazer(lm_out_s, lm_out_d, type = "text")
```

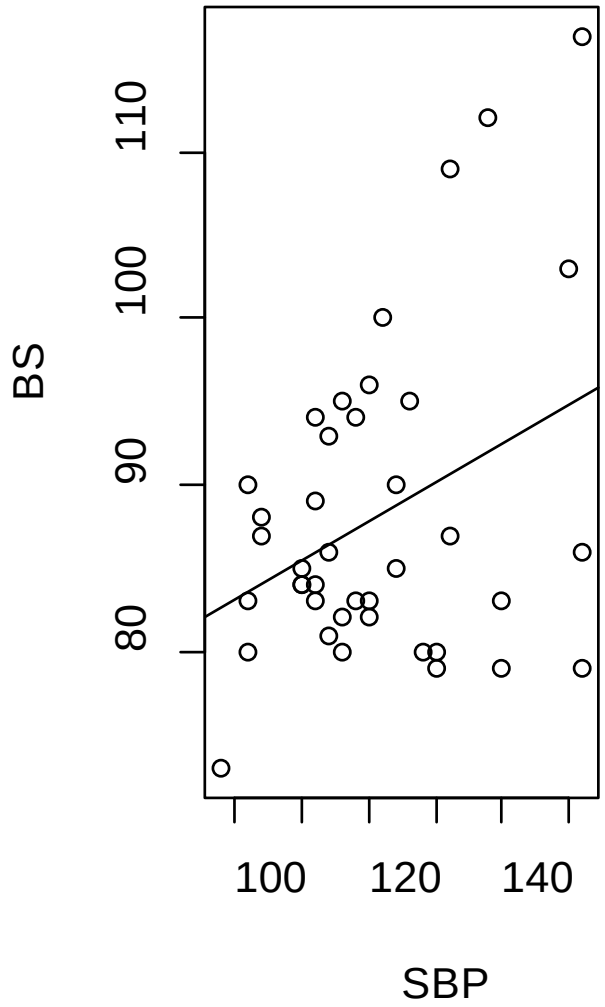
```
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                                BS
##                           (1)           (2)
## -----------------------------------------------
## SBP                     0.235**
##                         (0.099)
## DELTA                                 0.498***
##                                       (0.143)
## Constant                59.592***     66.470***
##                         (12.025)      (6.333)
## -----------------------------------------------
## Observations              40            40
## R2                       0.130         0.243
## Adjusted R2              0.107         0.223
## Residual Std. Error (df = 38)  9.002   8.400
## F Statistic (df = 1; 38)       5.691**  12.175***
```

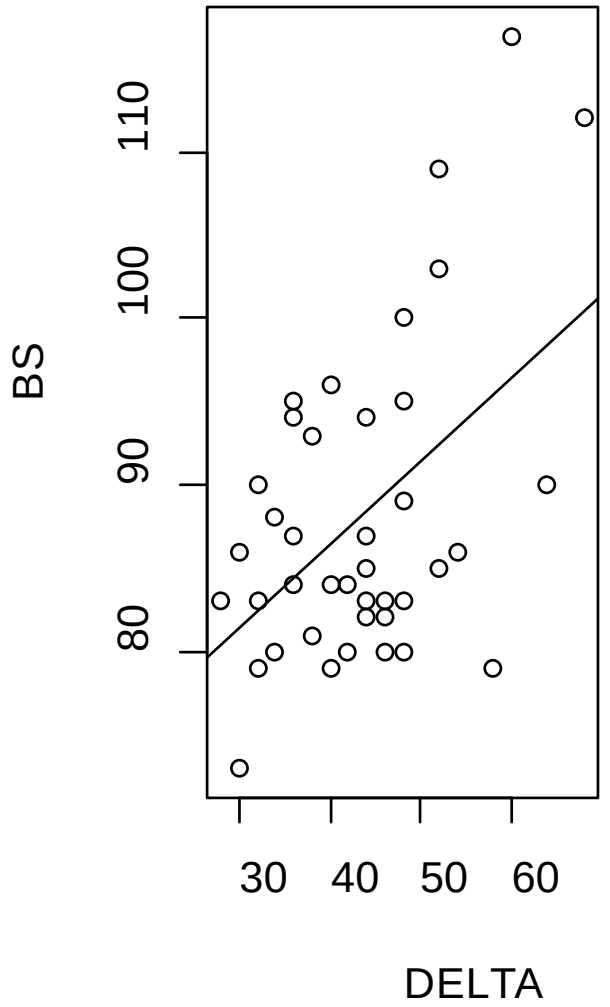
$2.385^2 = 5.69$
$3.489^2 = 12.17$

変数の選び方(1) 血糖と血圧

p.253

- 変数の選び方(1)　血糖と血圧
  - スクリプトファィル：Green1-4-5b.R
  - 利用した関数：lm、plot、abline
  - 方法
    - lm 関数で回帰分析
    - plot 関数と abline 関数で回帰直線を描画





変数の選び方(1) 血糖と血圧

- 変数の選び方(1)　血糖と血圧：重回帰分析
  スクリプトファイル：Green1-4-5b.R

```
lm_out_sb <- lm(BS ~ SBP + DBP, data = df2)
stargazer(lm_out_s, lm_out_d, lm_out_sb,
          type = "text")
```

```
## ================================================================================
##                                (1)                    (2)                    (3)
## --------------------------------------------------------------------------------
## SBP                         0.235**                                    0.498***
##                             (0.099)                                    (0.145)
## DELTA                                              0.498***
##                                                    (0.143)
## DBP                                                                   -0.455**
##                                                                       (0.192)
## Constant                   59.592***              66.470***              63.158***
##                             (12.025)               (6.333)               (11.452)
## --------------------------------------------------------------------------------
## Observations                  40                     40                     40
## R2                           0.130                  0.243                  0.245
## Adjusted R2                  0.107                  0.223                  0.204
## Residual Std. Error    9.002 (df = 38)        8.400 (df = 38)        8.499 (df = 37)
## F Statistic         5.691** (df = 1; 38) 12.175*** (df = 1; 38) 6.008*** (df = 2; 37)
```

変数の選び方(2) 血糖と体重・身長

- 変数の選び方(2)　血糖と血圧：単回帰分析と重回帰分析

  スクリプトファイル：Green1-4-5c.R

  利用した関数：lm、stargazer::stargazer

  方法：lm 関数で回帰分析、

  　stargazer 関数で結果を出力

BMI = weight/eright^2

```
df1

##    weight height  BS       BMI
## 1      75    169  95 26.25958
## 2      75    164  88 27.88519
## 3      68    161 117 26.23356
     . . . . . . . . .
## 39     58    168  80 20.54989
## 40     70    160 112 27.34375
```

```r
df1 <- df[c("weight", "height", "BS")]

# BMI を計算
bmi <- with(df1, 10000 * weight / height^2)
df1 <- cbind(df1, BMI = bmi)

# 対数変換
df2 <- log10(df1)

# 単回帰帰分析
lm_out1  <- lm(BS ~ BMI, data = df1)
summary(lm_out1)

# 重回帰分析
lm_out2 <- lm(BS ~ height + weight, data = df2)
summary(lm_out2)
```

BMI = weight/eright^2

変数の選び方(2) 血糖と体重・身長

- 変数の選び方(2)　血糖と血圧：重回帰分析

  スクリプトファイル：Green1-4-5c.R

  利用した関数

  lm、summary

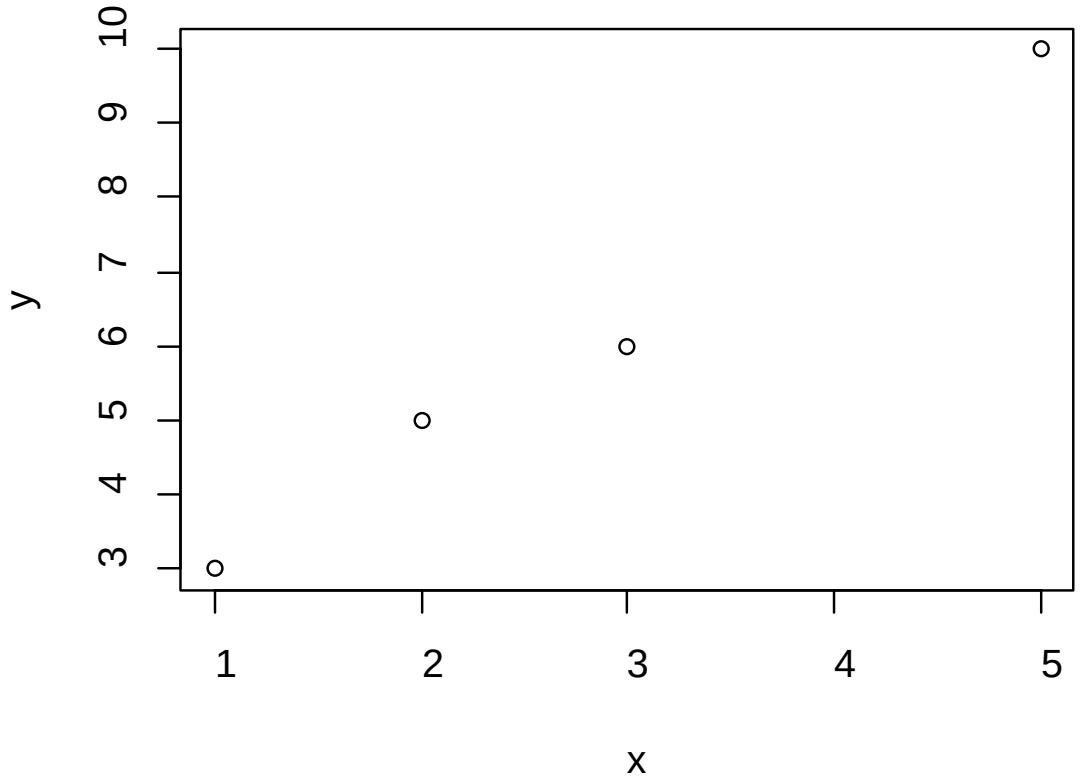
```
## Call:
## lm(formula = BS ~ height + weight, data = df2)

## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.06082 -0.02588 -0.01256  0.01860  0.09402
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.8009     1.0613   3.581 0.000978 ***
## height       -1.2698     0.5175  -2.454 0.018971 *
## weight        0.5323     0.1428   3.728 0.000644 ***

## Residual standard error: 0.03855 on 37 degrees of freedom
## Multiple R-squared:  0.2853, Adjusted R-squared:  0.2467
## F-statistic: 7.385 on 2 and 37 DF,  p-value: 0.002001
```

補遺(4)：原点を通る回帰式

p.258

- 表示 4.6.1 原点を通る回帰式
  - スクリプトファィル
    - Green1-4-5d.R
  - 利用した関数
    - lm、stargazer::stargazer
    - plot、abline
  - 方法
    - lm 関数で回帰分析

```
df = data.frame(y = c(3, 5, 6, 10),
                x = c(1, 2, 3,  5))

plot(y ~ x, data = df)
```



補遺(4)：原点を通る回帰式

- 表示 4.6.1 原点を通る回帰式
  - スクリプトファイル
    - Green1-4-5d.R
  - 利用した関数：lm、 stargazer::stargazer
  - 方法
    - 原点を通る回帰分析では、決定係数を指標として判断に用いない

```
lm_out  <- lm(y ~ x, data = df)       # 通常の回帰式
lm_out0 <- lm(y ~ x + 0, data = df)   # 原点を通る回帰式
stargazer(lm_out, lm_out0, type = "text")
```

```
## ===============================================================
##                                 (1)                    (2)
## ---------------------------------------------------------------
## x                            1.714***               2.077***
##                              (0.128)                (0.123)
##
## Constant                      1.286*
## ---------------------------------------------------------------

##                              (0.399## Observations
4                       4
## R2                            0.989                  0.990
## Adjusted R2                   0.984                  0.986
## Residual Std. Error      0.378 (df = 2)         0.768 (df = 3)
## F Statistic         180.000*** (df = 1; 2) 285.261*** (df = 1; 3)
```

切片がない

補遺(4)：原点を通る回帰式

- 表示 4.6.1 原点を通る回帰式

  スクリプトファィル

  　Green1-4-5d.R

  利用した関数

  　lm、summary

  方法

  　lm 関数で回帰分析

```
## Call:
## lm(formula = y ~ x + 0, data = df)
##
## Residuals:
##       1       2       3       4
##  0.9231  0.8462 -0.2308 -0.3846
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## x    2.077      0.123   16.89 0.000452 ***
## ---
##
## Residual standard error: 0.7679 on 3 degrees of freedom
## Multiple R-squared:  0.9896, Adjusted R-squared:  0.9861
## F-statistic: 285.3 on 1 and 3 DF,  p-value: 0.000452
```

切片がない

テキストでいう
「新しい寄与率」
この数値で判断しない

- 作成　　片瀬雅彦
- 作成時期　　2021年8月15日