

学籍番号:

公衆衛生学

疫学演習 2019-6-5 & 2019-6-12

氏名:

王 超辰 (<https://wangcc.me>)

# 1 問題1: 両群間計量データの平均値を比較する

200名の認知症患者を募集し, 認識能力テスト(cognitive test, COG), 及び脳萎縮の進行度 (brain atrophy, 脳体積の平均年間減少率, 単位は%) の検査を全員に行った. COG, 及び脳萎縮のデータは大きいほど認知症の進行度がより進んでいる. また, この200名の参加者から採取した血液検体を利用して, ある遺伝子の変異の有無を検査した. このデータは以下の表でまとめた:

変数	遺伝変異あり (n = 50)		遺伝変異なし (n = 150)	
	平均値 (mean)	標準偏差 (standard deviation)	平均値 (mean)	標準偏差 (standard deviation)
認識能力テスト, COG	69.2	9.0	60.2	9.0
脳萎縮度, atrophy, %/year	0.67	0.21	0.23	0.10

1. 帰無仮説を「遺伝子変異ありと変異なし両群の間に, COGの平均値は等しい」とする. 上記のデータ及び適宜な方法を使って検定せよ. 検定の結果を分かりやすく説明せよ. なお, 分散が等しいと仮定できる場合, 以下の式で両群の共通標準偏差が計算できる:

$$S = \sqrt{\frac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{n_A + n_B - 2}} \quad (1)$$

- $S_A$ : A群の標準偏差;
- $n_A$ : A群の人数;
- $S_B$ : B群の標準偏差;
- $n_B$ : B群の人数;
- $S$ : A群及びB群の共通標準偏差;
- $n_A + n_B - 2$ : 分散が等しい時の自由度.

また, EZR で t 値, 自由度 (degree of freedom) を使って P 値を計算する時, 以下のコマンドを利用してください:

```
2*pt(t value, degree of freedom, lower=FALSE)
```

## 1.1 答え

両群の標準偏差は 9.0 と推定され、分散が等しいと仮定できるから、Student の t 検定を行う:

$$T = \frac{\bar{X}_A - \bar{X}_B}{S\sqrt{1/n_A + 1/n_B}}$$

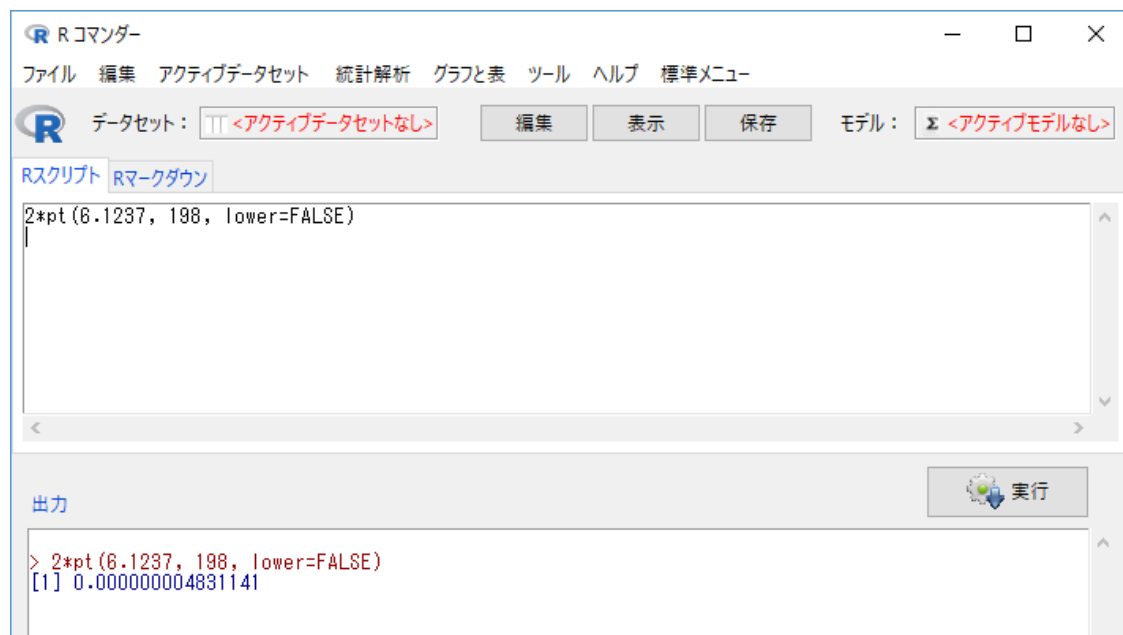
公式(1)により、共通分散/標準偏差を推定する:

$$\begin{aligned} \therefore S &= \sqrt{\frac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{n_A + n_B - 2}} \\ \therefore S &= \sqrt{\frac{(50 - 1)9^2 + (150 - 1)9^2}{50 + 150 - 2}} = 9 \\ \Rightarrow T &= \frac{\bar{X}_A - \bar{X}_B}{S\sqrt{1/n_A + 1/n_B}} \\ &= \frac{69.2 - 60.2}{9 \times \sqrt{1/50 + 1/150}} \\ &= \frac{9}{9 \times 0.1633} = 6.1237 \end{aligned}$$

自由度 (degree of freedom)は  $n_A + n_B - 2 = 198$ , P値の計算は EZR を利用する:

```
2*pt(6.1237, 198, lower=FALSE)
```

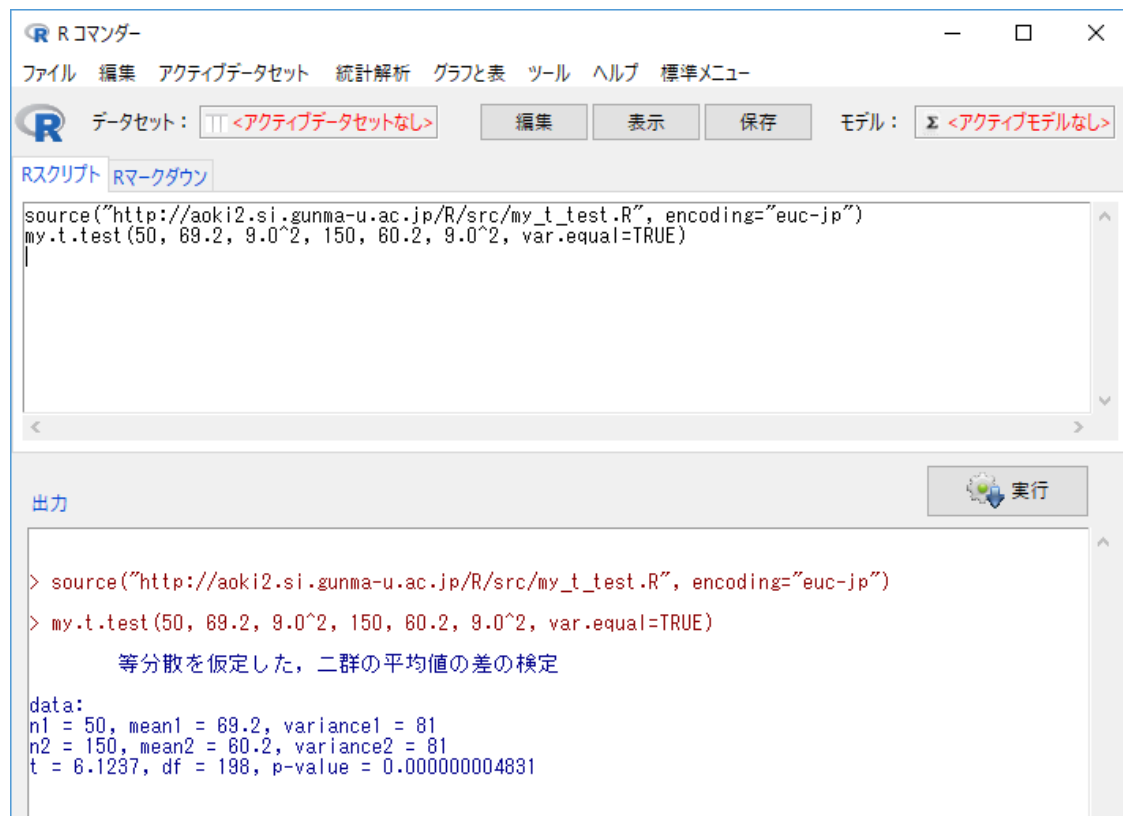
```
## [1] 4.831141e-09
```



以下のコードをRスクリプトに入力して,実行をクリックしてください.自分の検定結果とは一致するかを確認してください.

```
source("http://aoki2.si.gunma-u.ac.jp/R/src/my_t_test.R", encoding="euc-jp")
my.t.test(50, 69.2, 9.0^2, 150, 60.2, 9.0^2, var.equal=TRUE)
```

```
##
## 等分散を仮定した, 二群の平均値の差の検定
##
## data:
## n1 = 50, mean1 = 69.2, variance1 = 81
## n2 = 150, mean2 = 60.2, variance2 = 81
## t = 6.1237, df = 198, p-value = 4.831e-09
```



手計算の結果とは一致していると確認できる.

この検定結果は「両群のCOG平均値が等しい」という帰無仮説を棄却するために非常に強い証拠を提供したと言える.

There is strong evidence against the null hypothesis that the means of COG are the same in the two groups.

2. この患者データから、遺伝子変異ありとなしの群の間に脳萎縮度 (atrophy) の比較を 1. と同じ方法で検定してもよいか? どの検定方法を使えば 1. と同じ検定方法を使えるかどうかを判断できるを説明せよ. 実際にこの検定方法を行ってください.

なお, EZR で F 値, 両群の分散, 両群それぞれの自由度 (df) を使って P 値を計算する時に, 以下のコマンドを利用してください:

```
2*pf(F value, df in group 1, df in group 2, lower=FALSE)
```

## 1.2 答え

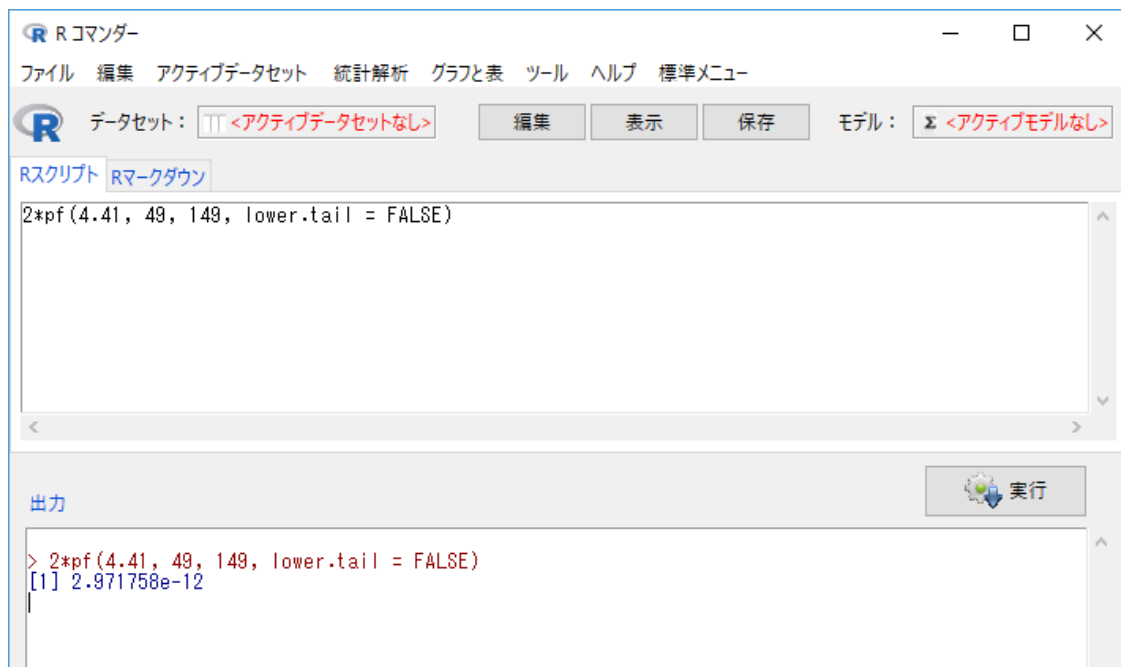
テーブルから両群の標準偏差はそれぞれ 0.21, 0.10 だと推定され, 分散 (variance) が等しいという前提が満たされていない. 1. の検定方法を使う時には, 両群の分散が等しいという前提条件が必須だから, 同じ Student t 検定を行うことができない. 「両群の分散」が等しいという帰無仮説を検定するには F 検定を利用する:

$$\begin{aligned} F &= \frac{S_A^2}{S_B^2} \\ &= \frac{0.21^2}{0.10^2} \\ &= 4.41 \end{aligned}$$

自由度 (degree of freedom) はそれぞれ  $n_A - 1 = 49$ ;  $n_B - 2 = 149$ , P 値の計算は EZR を利用する:

```
2*pf(4.41, 49, 149, lower.tail = FALSE)
```

```
## [1] 2.971758e-12
```



以下のコードをRスクリプトに入力して,実行をクリックしてください.自分の検定結果とは一致するかを確認してください.

```
source("http://aoki2.si.gunma-u.ac.jp/R/src/my_var_test.R", encoding="euc-jp")
my.var.test(50, 0.21^2, 150, 0.1^2)
```

```
##
## 二次データから, 二群の等分散性の検定
##
## data:  n1 = 50, variance1 = 0.0441, n2 = 150, variance2 = 0.01
## F = 4.41, num df = 49, denom df = 149, p-value = 2.972e-12
```



手計算の結果とは一致していると確認できる.

この検定結果は「両群の脳萎縮度の分散が等しい」という帰無仮説を棄却するために非常に強い証拠を提供したと言える.

There is strong evidence against the null hypothesis that the variances of atrophy are the same in the two groups.

3. 2.の結果を踏まえて、帰無仮設「両群の脳萎縮度の平均値が等しい」を検定せよ。なお、両群の分散が等しいという前提が満たされていない時に、自由度(df)の計算式は以下となる:

$$df = \frac{(S_A^2/n_A + S_B^2/n_B)^2}{(S_A^2/n_A)^2/(n_A - 1) + (S_B^2/n_B)^2/(n_B - 1)} \quad (2)$$

また, EZR で t 値, 自由度 (df) を使って P 値を計算する時, 以下のコマンドを利用してください:

```
2*pt(t value, df, lower=FALSE)
```

### 1.3 答え

2.の検定結果から、「両群の脳萎縮度の分散が等しい」という帰無仮説を棄却されたため, Welch の t 検定を採用する。

$$\begin{aligned} \Rightarrow T &= \frac{\bar{X}_A - \bar{X}_B}{\sqrt{S_A^2/n_A + S_B^2/n_B}} \\ &= \frac{0.67 - 0.23}{\sqrt{0.21^2/50 + 0.10^2/150}} \\ &= \frac{0.44}{\sqrt{0.0009486667}} = 14.28551 \end{aligned}$$

自由度は公式(2)により計算できる:

$$\begin{aligned} df &= \frac{(S_A^2/n_A + S_B^2/n_B)^2}{(S_A^2/n_A)^2/(n_A - 1) + (S_B^2/n_B)^2/(n_B - 1)} \\ &= \frac{(0.21^2/50 + 0.10^2/150)^2}{(0.21^2/50)^2/(50 - 1) + (0.10^2/150)^2/(150 - 1)} \\ &= 58.58105 \end{aligned}$$

P値の計算は EZR を利用する:

```
2*pt(14.28551, 58.58105, lower=FALSE)
```

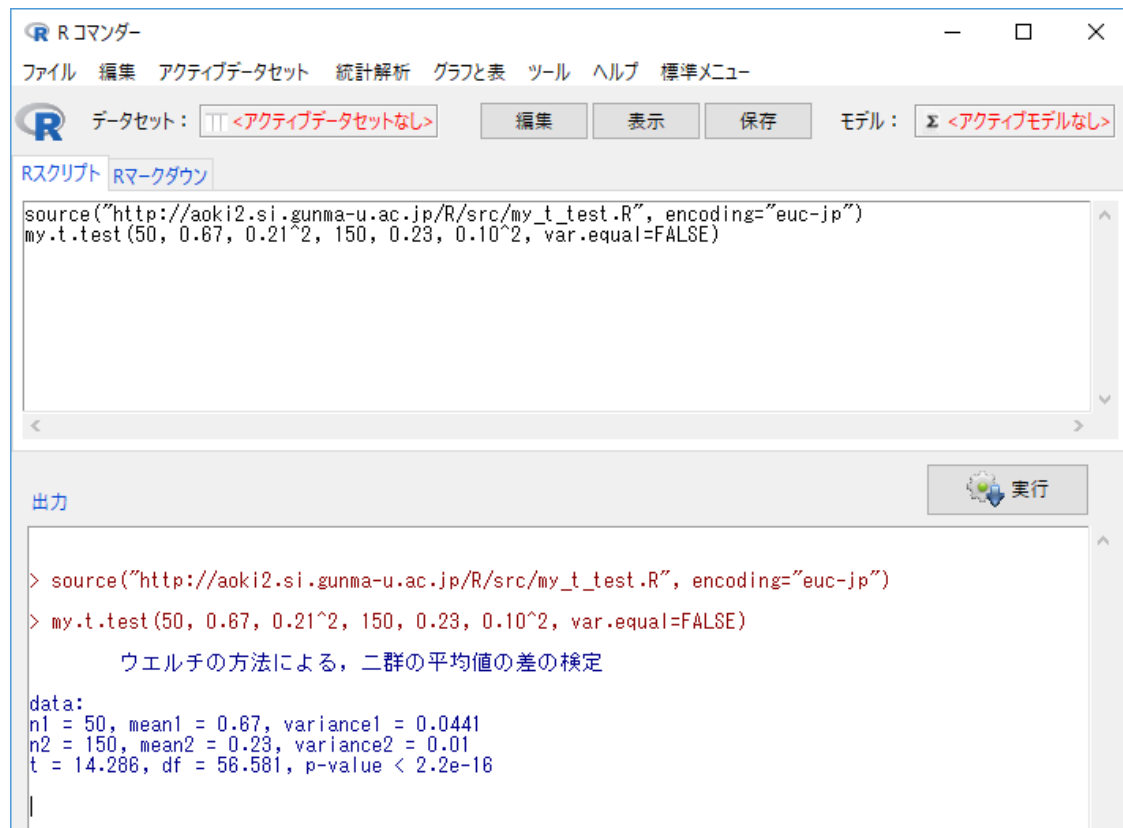
```
## [1] 9.601543e-21
```



以下のコードをRスクリプトに入力して,実行をクリックしてください.自分の検定結果とは一致するかを確認してください.

```
source(" http://aoki2.si.gunma-u.ac.jp/R/src/my_t_test.R" , encoding=" euc-jp" )  
my.t.test(50, 0.67, 0.21^2, 150, 0.23, 0.10^2, var.equal=FALSE)
```

```
##  
## ウェルチの方法による, 二群の平均値の差の検定  
##  
## data:  
## n1 = 50, mean1 = 0.67, variance1 = 0.0441  
## n2 = 150, mean2 = 0.23, variance2 = 0.01  
## t = 14.286, df = 56.581, p-value < 2.2e-16
```



The screenshot shows the R Commander window. The menu bar includes 'ファイル', '編集', 'アクティブデータセット', '統計解析', 'グラフと表', 'ツール', 'ヘルプ', and '標準メニュー'. The 'データセット' field shows '<アクティブデータセットなし>' and the 'モデル' field shows '<アクティブモデルなし>'. The 'Rスクリプト' tab is active, displaying the following code:

```
source("http://aoki2.si.gunma-u.ac.jp/R/src/my_t_test.R", encoding="euc-jp")
my.t.test(50, 0.67, 0.21^2, 150, 0.23, 0.10^2, var.equal=FALSE)
```

The '出力' (Output) window shows the execution results:

```
> source("http://aoki2.si.gunma-u.ac.jp/R/src/my_t_test.R", encoding="euc-jp")
> my.t.test(50, 0.67, 0.21^2, 150, 0.23, 0.10^2, var.equal=FALSE)

ウェルチの方法による、二群の平均値の差の検定

data:
n1 = 50, mean1 = 0.67, variance1 = 0.0441
n2 = 150, mean2 = 0.23, variance2 = 0.01
t = 14.286, df = 56.581, p-value < 2.2e-16
```

手計算の結果とは一致していると確認できる。

この検定結果は「両群の脳萎縮度の平均値が等しい」という帰無仮説を棄却するために非常に強い証拠を提供したと言える。

There is strong evidence against the null hypothesis that the means of atrophy are the same in the two groups.



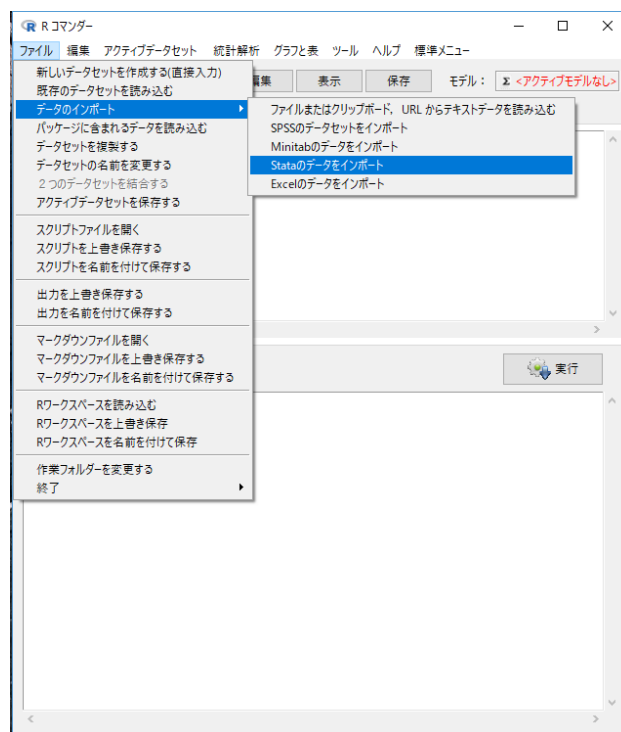
## 2 問題2:線形回帰モデル

190名の乳幼児の性別(1 = 男,2 = 女),年齢(月, months),体重(kg)のデータを収集した.このデータを用いて,以下の問題を解答したい:

- ・ 子供の年齢が一ヶ月の増加によって,体重はどれぐらい増えているか?
- ・ 男の子は女の子と比べて,平均的に体重はどれぐらい大きい/小さい?

### 2.1 データのインポート

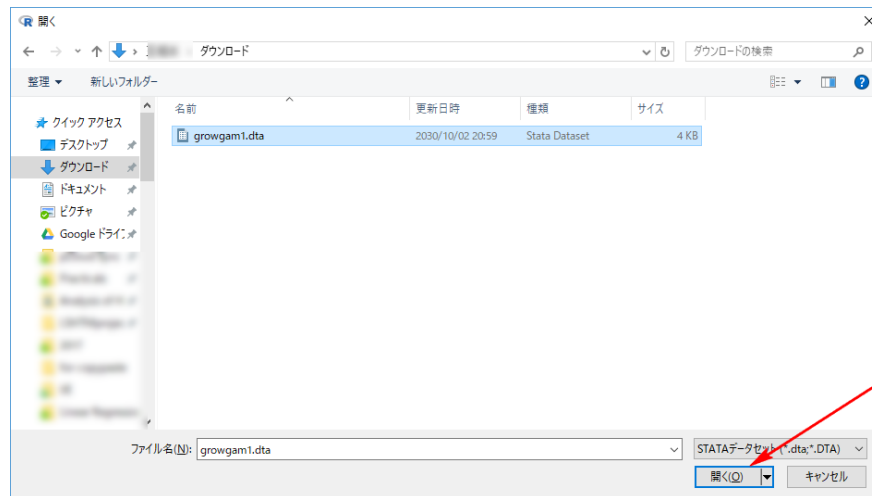
#### 2.1.1 ステップ 1



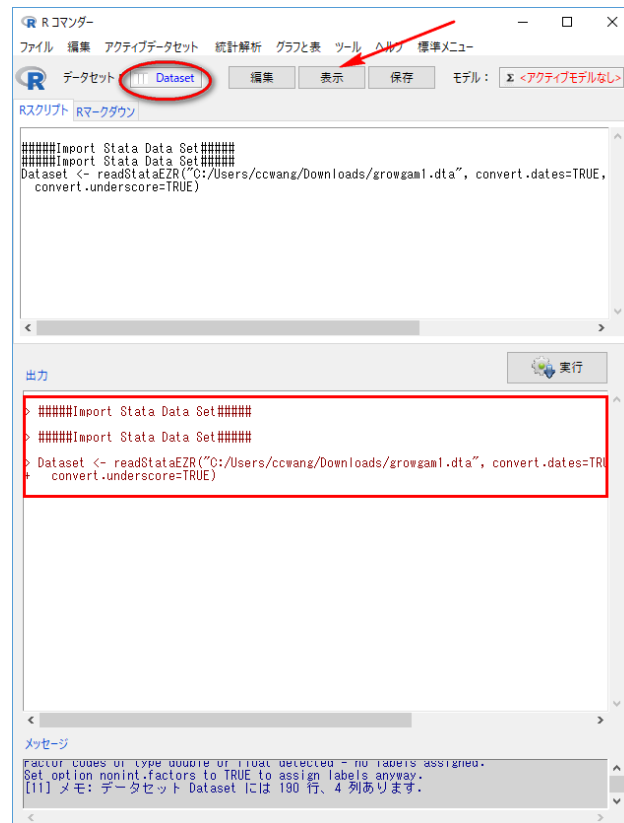
## 2.1.2 ステップ2



## 2.1.3 ステップ3



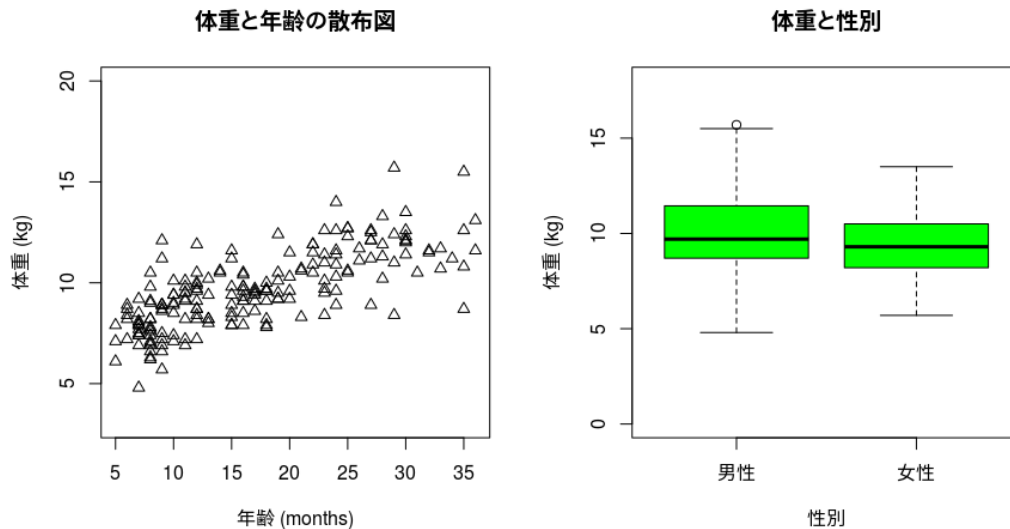
## 2.1.4 ステップ4



## 2.1.5 ステップ5

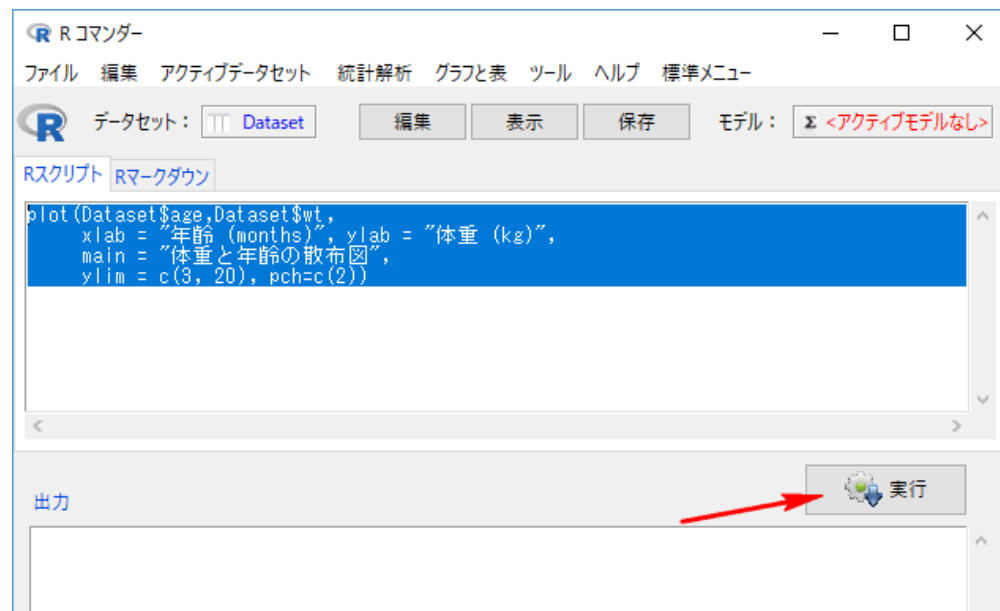
	sex	age	wt	len
1	2	23	8.4	73.2
2	2	22	10.9	84.4
3	2	6	7.2	68.7
4	1	24	10.3	83.7
5	1	14	10.5	79.2
6	2	18	9.6	75.8
7	2	30	11.4	84.4
8	1	24	11.4	84.8
9	1	17	9.4	74.3
10	2	27	12.5	82.6
11	2	18	9.1	74.1
12	1	30	12.0	86.4
13	1	18	7.8	71.9
14	2	15	8.5	77.6
15	2	13	8.0	72.2
16	1	11	9.1	72.4
17	2	8	10.5	71.2
18	2	9	7.2	67.4
19	1	8	6.9	62.7
20	1	16	9.6	79.4
21	2	25	10.5	81.5
22	1	18	9.7	80.7
23	1	29	8.4	81.2
24	2	10	9.4	72.7
25	1	8	7.0	68.3
26	1	9	6.9	68.8
27	1	6	8.7	68.6
28	1	25	12.3	85.4
29	2	16	9.3	78.5
30	1	29	15.7	95.5

## 2.2 体重と年齢の散布図,性別により体重の箱ひげ図



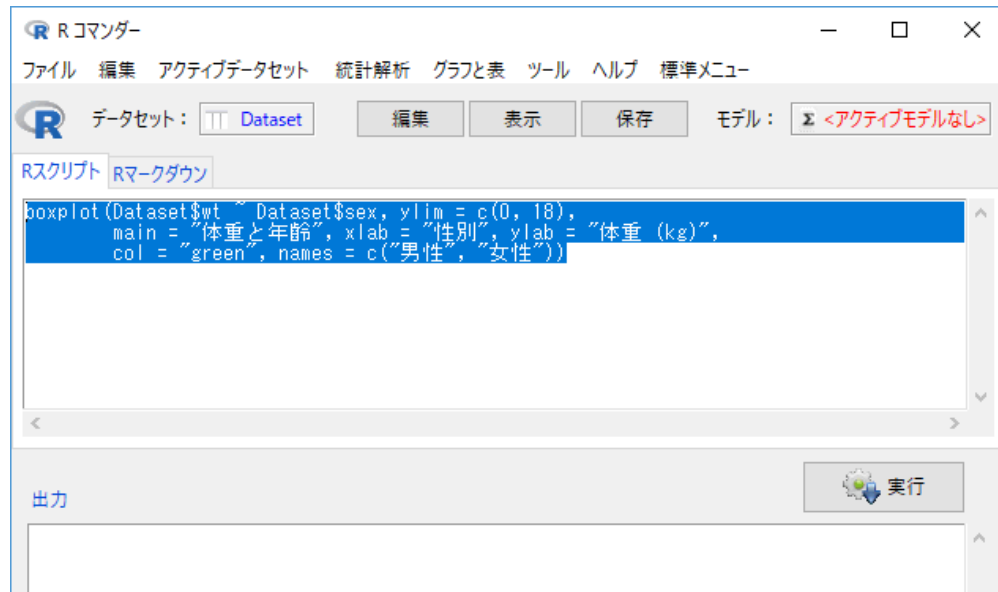
上記左のグラフを描くため,以下のコードをRスクリプトに入力して,実行をクリックしてください。

```
plot(Dataset$age, Dataset$wt,
     xlab = "年齢 (months)", ylab = "体重 (kg)",
     main = "体重と年齢の散布図",
     ylim = c(3, 20), pch=c(2))
```



性別により体重の箱ひげ図を描くため,以下のコードをRスクリプトに入力して,実行をクリックしてください.

```
boxplot(Dataset$wt ~ Dataset$sex, ylim = c(0, 18),  
        main = " 体重と年齢", xlab = " 性別", ylab = " 体重 (kg)",  
        col = "green", names = c(" 男性", " 女性" ))
```



2.3 年齢,体重それぞれの平均値,分散を求めよ;また,年齢と体重の相関係数を算出せよ.なお,EZRで計量データの平均値を計算するには,コマンド `mean(変数名)` を使う;共分散を計算したい時に,コマンド `cor(変数1, 変数2)` を利用する.

以下のコードをRスクリプトに入力して,実行をクリックしてください.(結果を下の余白に記入すること)

```
# 年齢の平均値  
mean(Dataset$age)
```

```
## [1] 16.97895
```

```
# 年齢の分散  
var(Dataset$age)
```

```
## [1] 69.5022
```

```
# 体重の平均値  
mean(Dataset$wt)
```

```
## [1] 9.644737
```

```
# 体重の分散  
var(Dataset$wt)
```

```
## [1] 3.513068
```

```
# 体重と年齢の共分散 covariance
cov(Dataset$wt, Dataset$age)
```

```
## [1] 11.49089
```

2.4 年齢を説明変数, 体重を目的変数とする場合, 年齢の傾き(回帰係数), と切片を求めよ. なお, 分散と共分散の定義を以下とする,  $\bar{X}$  は  $X$  の平均値を示す:

- ・ 分散 variance:

$$\begin{aligned}\text{Var}(X) &= \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n - 1} \\ &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}\end{aligned}$$

- ・ 共分散 covariance:

$$\begin{aligned}\text{Cov}(X, Y) &= \frac{(X_1 - \bar{X})(Y_1 - \bar{Y}) + (X_2 - \bar{X})(Y_2 - \bar{Y}) + \cdots + (X_n - \bar{X})(Y_n - \bar{Y})}{n - 1} \\ &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}\end{aligned}$$

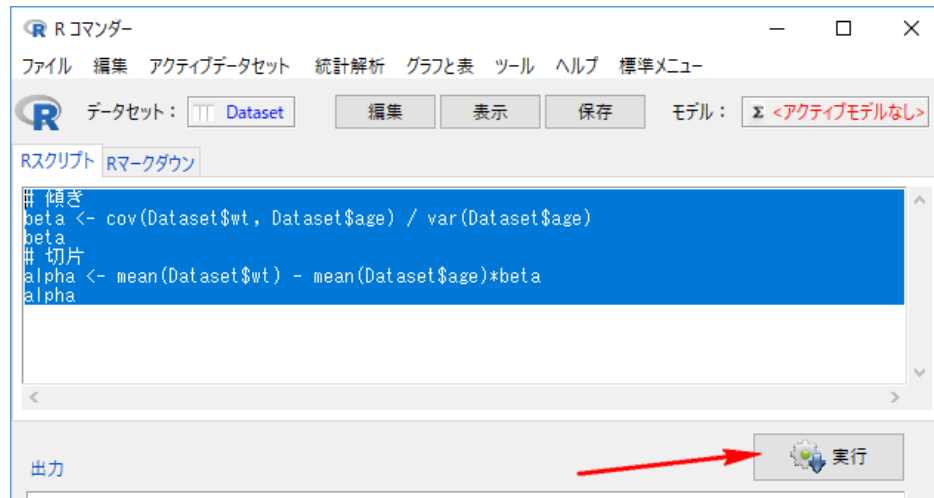
以下のコードをRスクリプトに入力して, 実行をクリックしてください. (結果を下の余白に記入すること)

```
# 傾き (slope)
beta <- cov(Dataset$wt, Dataset$age) / var(Dataset$age)
beta
```

```
## [1] 0.1653314
```

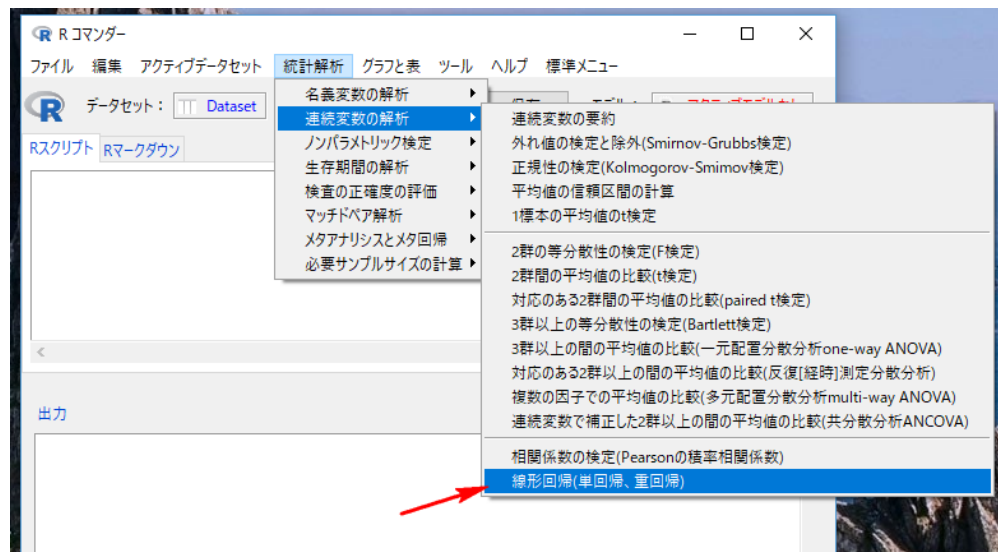
```
# 切片 (intercept)
alpha <- mean(Dataset$wt) - mean(Dataset$age)*beta
alpha
```

```
## [1] 6.837584
```

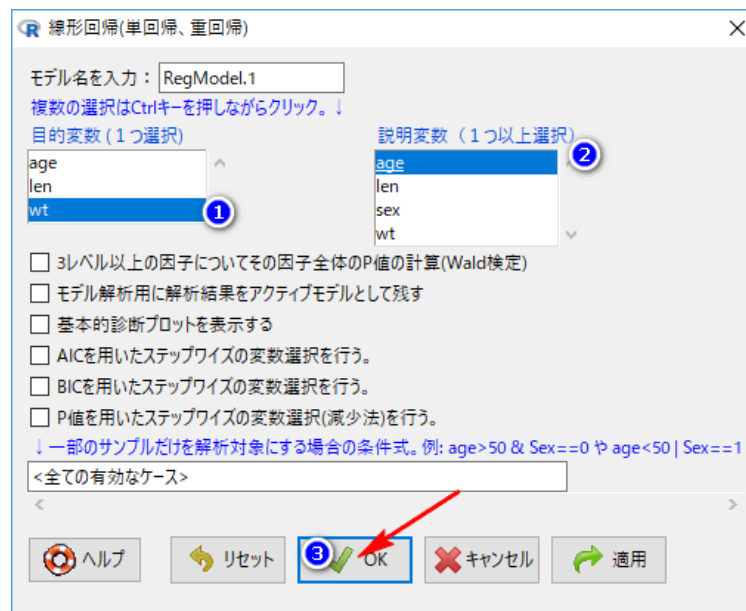


## 2.5 実際にEZRで線形モデルを作ってみよう:

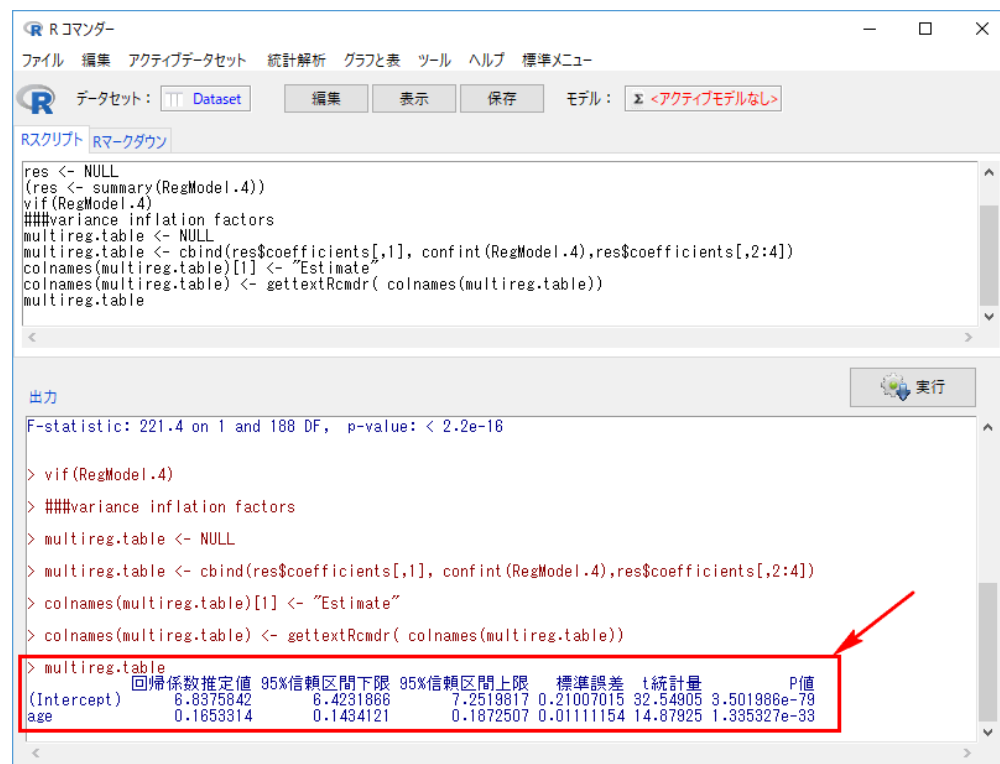
### 2.5.1 ステップ1



## 2.5.2 ステップ2



## 2.5.3 ステップ3



自分の計算結果とは一致するかを確認してください。



2.6 今まで計算した傾きと切片の数字を用いて,年齢と体重の関係を線形と考える場合の計算式を記入せよ.傾きと切片の計算結果の意味をそれぞれ記述せよ.

2.6.1 答え

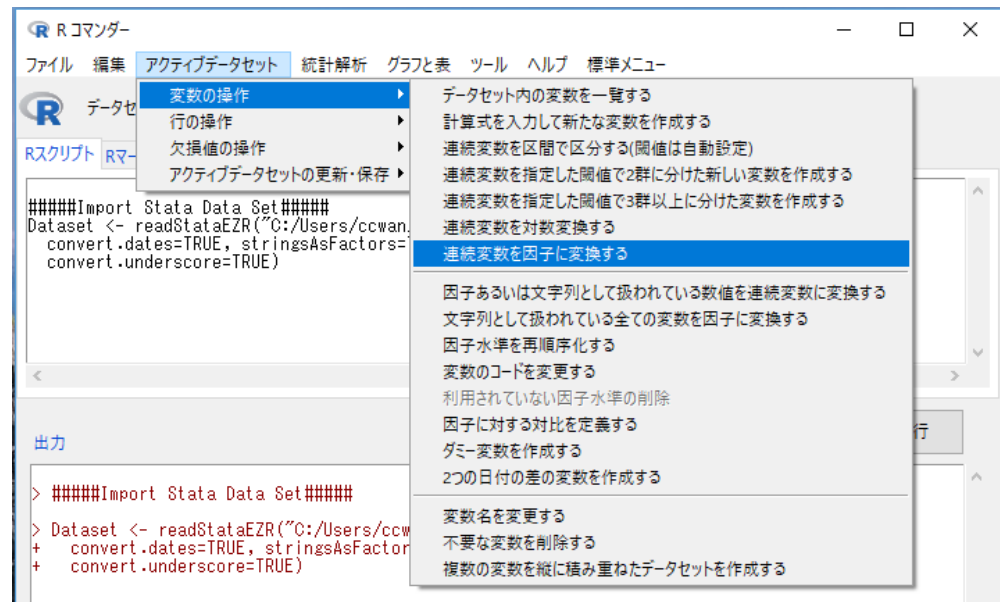
$$Y = 6.838 + 0.165X$$

- ・  $Y$  : 体重 (kg);
- ・  $X$  : 年齢 (months);
- ・ 0.165 : 子供の年齢が1ヶ月伸びると,体重が平均的 0.165 kg (165 g) 高くなる;
- ・ 6.838 : 子供の年齢が0ヶ月の時に,体重の平均値は 6.838 kg.

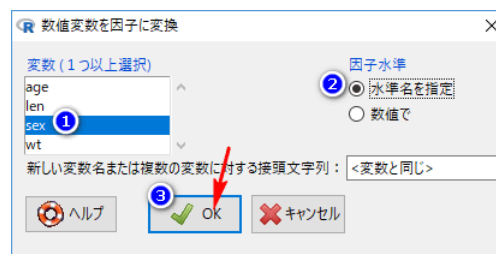
## 2.7 性別を説明変数に入れたモデルを作る

### 2.7.1 性別変数を因子 (factor) に変換する

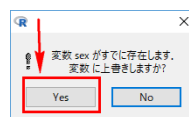
#### 2.7.1.1 ステップ1



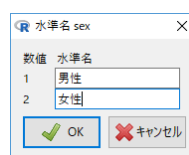
#### 2.7.1.2 ステップ2



#### 2.7.1.3 ステップ3

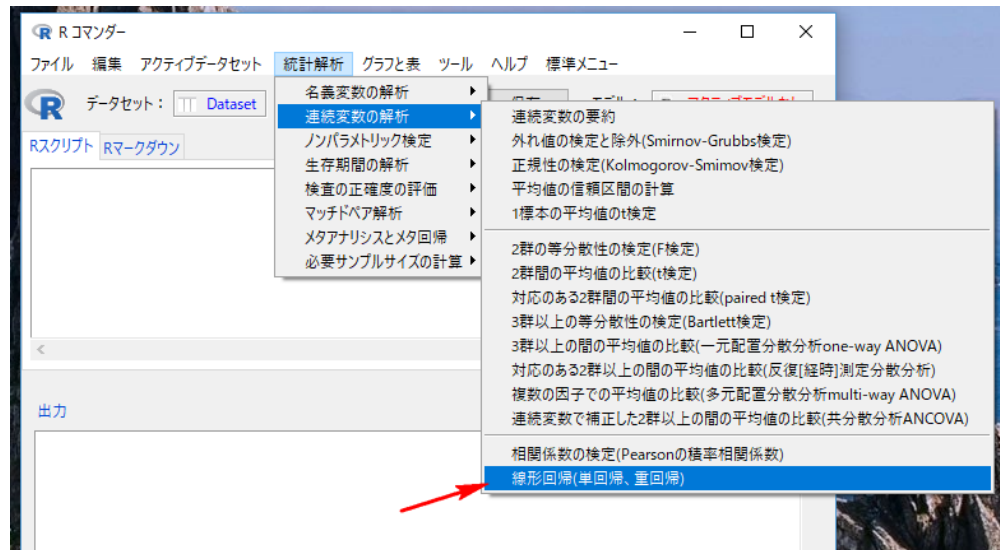


#### 2.7.1.4 ステップ4-水準名に男性,女性を入力する

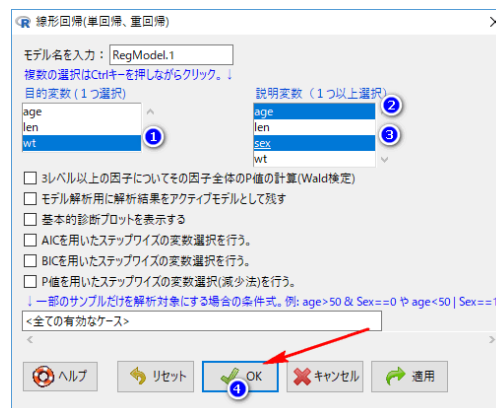


## 2.7.2 重回帰線形モデルを作る

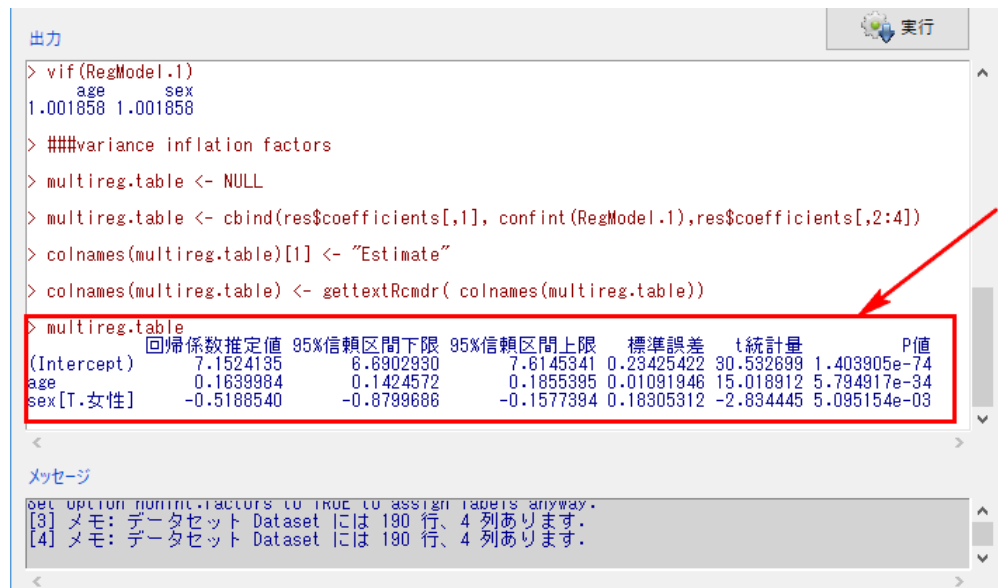
## 2.7.2.1 ステップ1



## 2.7.2.2 ステップ2—複数の説明変数を選択する時に control キーを押しながらマウスで変数名をクリックする



## 2.7.3 重回帰線形モデルの結果を確認する



```

> vif(RegModel.1)
      age      sex 
1.001858 1.001858 

> ###variance inflation factors

> multireg.table <- NULL

> multireg.table <- cbind(res$coefficients[,1], confint(RegModel.1), res$coefficients[,2:4])

> colnames(multireg.table)[1] <- "Estimate"

> colnames(multireg.table) <- gettextRcmdr( colnames(multireg.table))

> multireg.table
      回帰係数推定値 95%信頼区間下限 95%信頼区間上限 標準誤差  t統計量  P値
(Intercept)      7.1524135      6.8902930      7.6145341  0.23425422  30.532699  1.403905e-74
age              0.1639984      0.1424572      0.1855395  0.01091946  15.018912  5.794917e-34
sex[T.女性]     -0.5188540     -0.8799686     -0.1577394  0.18305312  -2.834445  5.095154e-03
  
```

メッセージ

```

get OPTION HONINT.FACTORS TO TRUE TO ASSIGN labels anyway.
[3] メモ: データセット Dataset には 190 行、4 列あります。
[4] メモ: データセット Dataset には 190 行、4 列あります。
  
```

2.8 重回帰線形モデルの計算結果を用いて、体重の平均値を年齢と性別の線形モデルで表示せよ。各回帰係数の意味を説明せよ。

2.9 答え

$$Y = 7.152 + 0.164X_1 - 0.519X_2$$

- $Y$  : 体重(kg);
- $X_1$  : 年齢 (months);
- $X_2 = 1$  : 女性;
- $X_2 = 0$  : 男性;
- 7.152 : 男の子が年齢 0 ヶ月の時の平均体重;
- 0.164 : 同じ性別の子供の年齢が1ヶ月高くなることによって、体重が平均的に 0.164 kg増える;
- -0.519 : 子供年齢が同じ時に、女の子は男の子と比べ、体重が平均的に 0.519 kg低い。

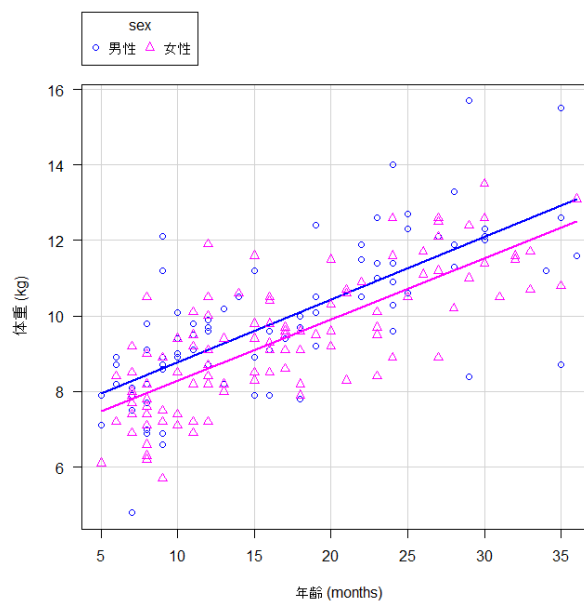
2.10 上記の重回帰線形モデルを用いて、年齢が34ヶ月の女の子の体重の予測値を計算せよ。

2.11 答え

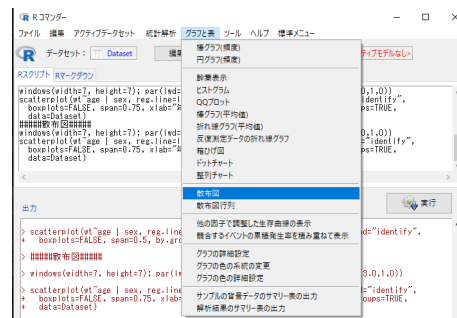
- ・  $X_1 = 34$ ;
- ・  $X_2 = 1$ ;

$$Y = 7.152 + 0.164 \times 34 - 0.519 \times 1 \\ = 12.209 \text{ (kg)}$$

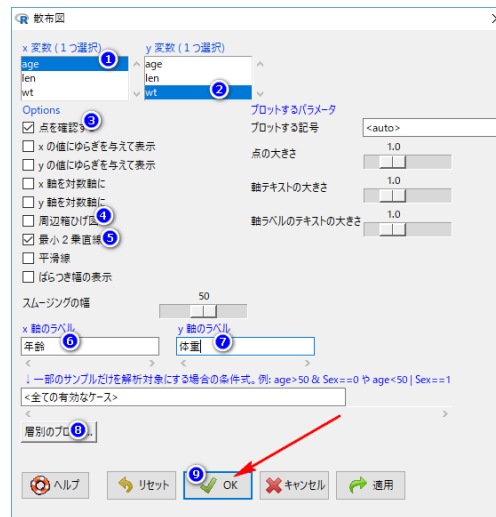
2.12 男女別の年齢と体重の散布図を描く



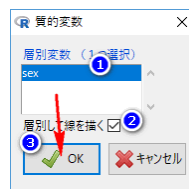
2.12.1 ステップ1



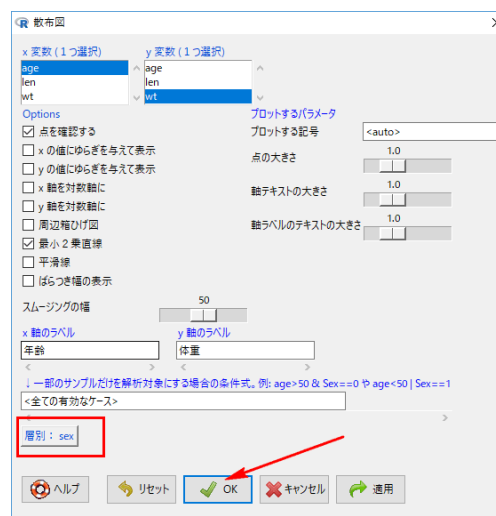
## 2.12.1.1 ステップ2



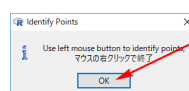
## 2.12.1.2 ステップ3



## 2.12.1.3 ステップ4



## 2.12.1.4 ステップ5



3 問題3:  $\chi^2$  検定, オッズ比, ロジスティクス回帰モデル

1990年代, アフリカナイジェリア北部でオンコセルカ症 (回旋糸状虫症, onchocerciasis; river blindness と呼ばれる) が流行していた.

#### 4 問題4:生存分析



5 参考図書:

- 1.「Rによる保健医療データ解析演習」,中澤 港, (<http://minato.sip21c.org/msb/medstatbookx.pdf>)
- 2.「みんなの医療統計 12日間で基礎理論とEZRを完全マスター!」,新谷 歩.