

学籍番号:

公衆衛生学

疫学演習 2019-6-5 & 2019-6-12

氏名:

# 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:

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

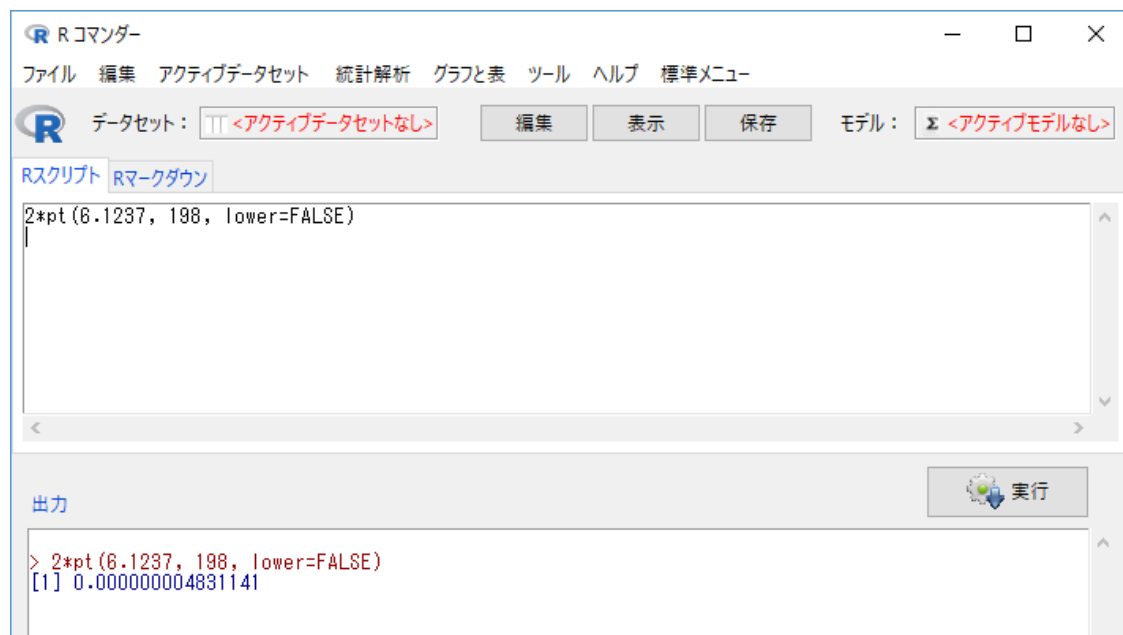
公式(1)により、共通分散/標準偏差を推定する:

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

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)
```

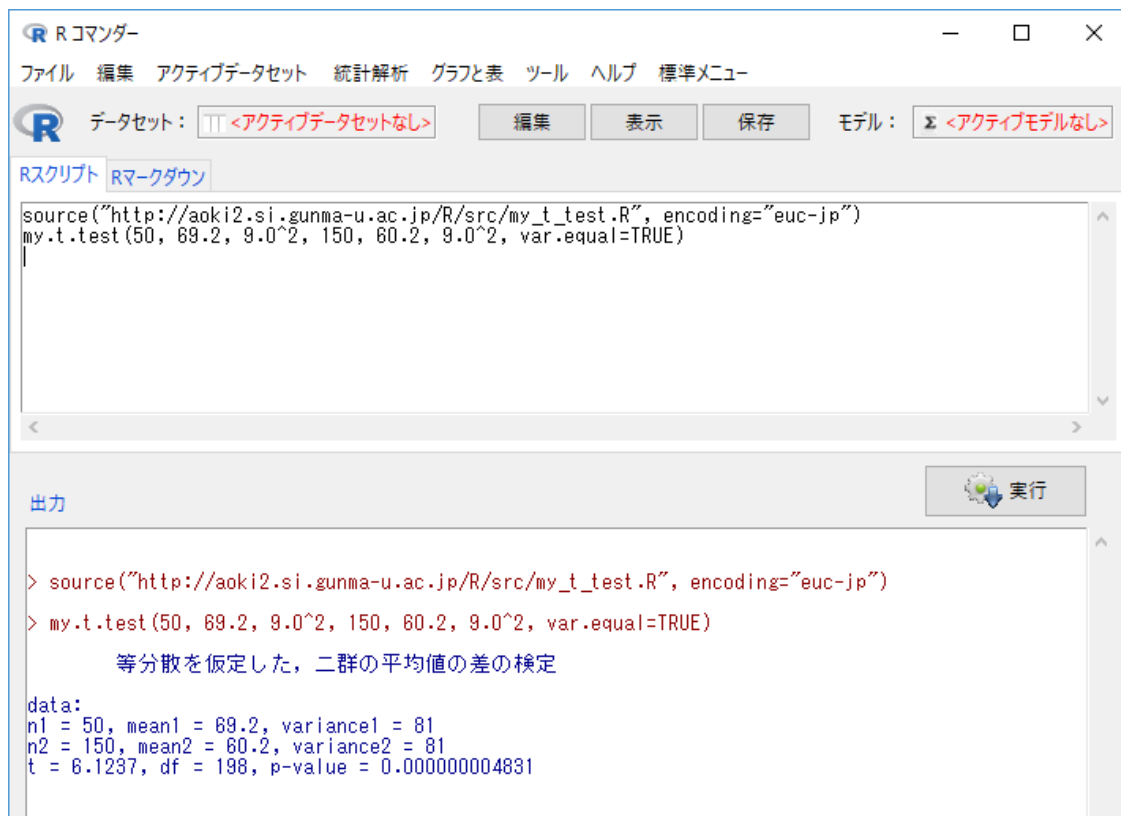
```
## [1] 4.831141e-09
```



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
```



手で計算した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

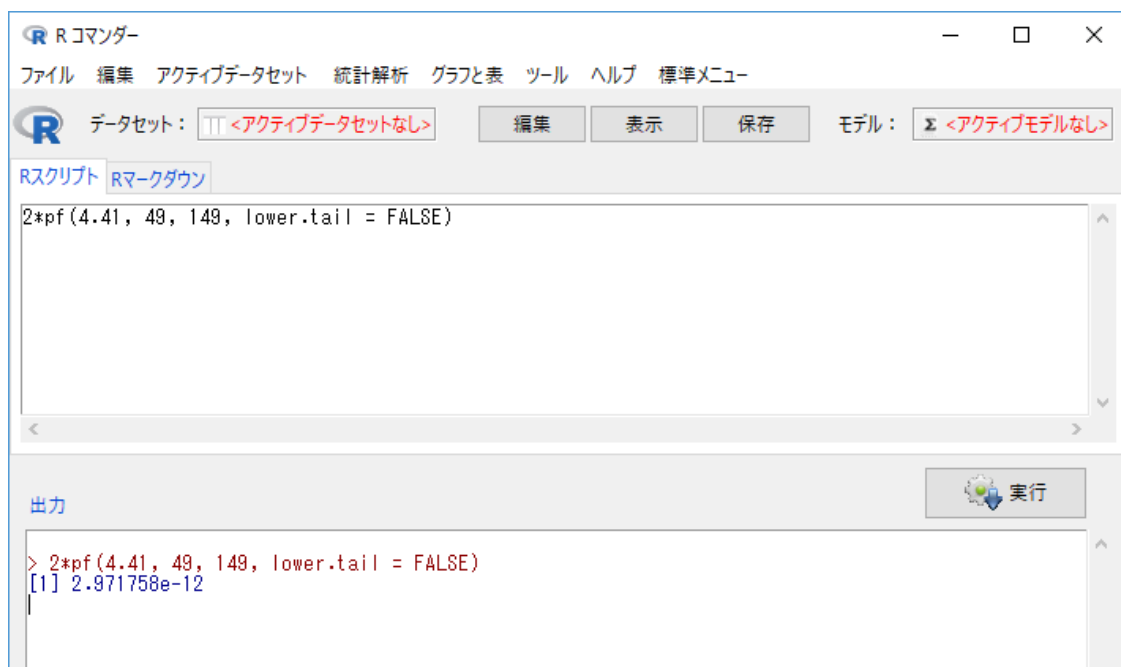
テーブルから両群の標準偏差はそれぞれ 0.21, 0.10 だと推定され, 分散 (variance) が等しいという前提が満たされていない. 1. の検定方法を使う時には, 両群の分散が等しいという前提条件が必須だから, 同じ Student t 検定を行うことができない. 「両群の分散」が等しいという帰無仮説を検定するには F 検定を利用する:

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

自由度 (degree of freedom) はそれぞれ  $n_A - 1 = 49$ ;  $n_B - 1 = 149$ , 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
```



手で計算した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、帰無仮説「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)
```

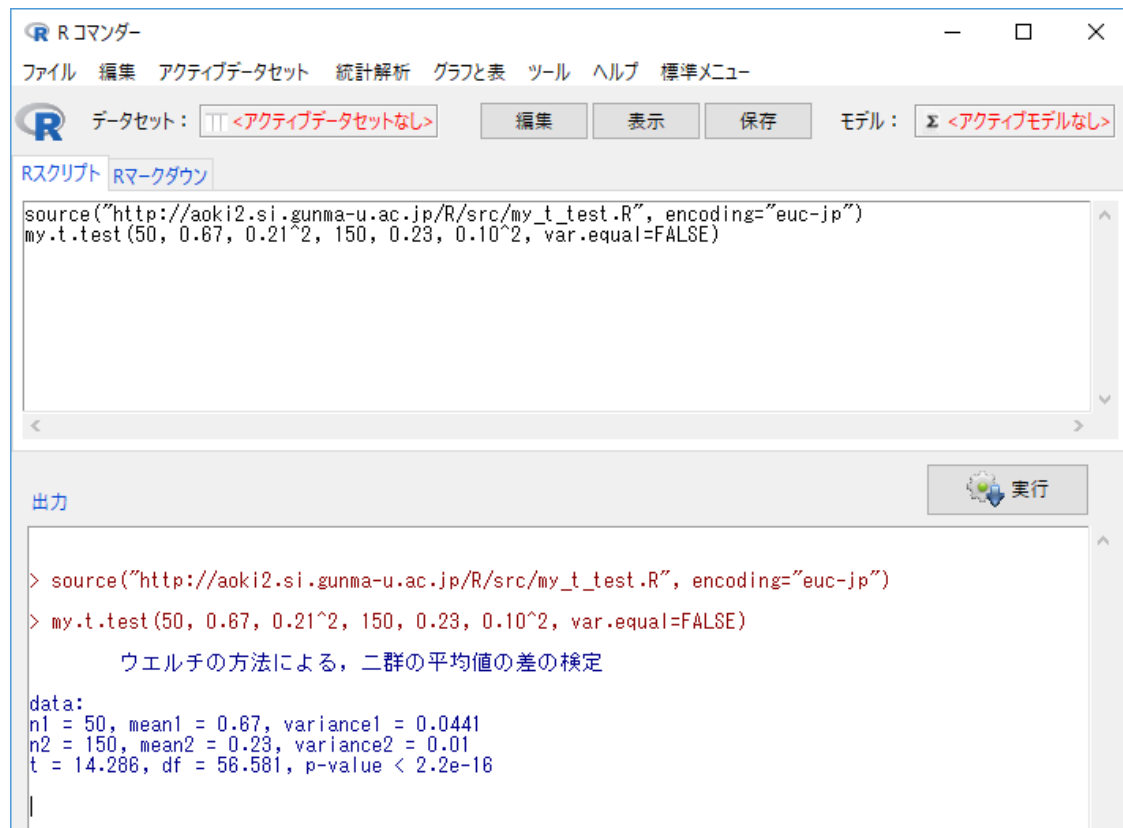
```
## [1] 9.601543e-21
```



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 R 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
```

手で計算した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.



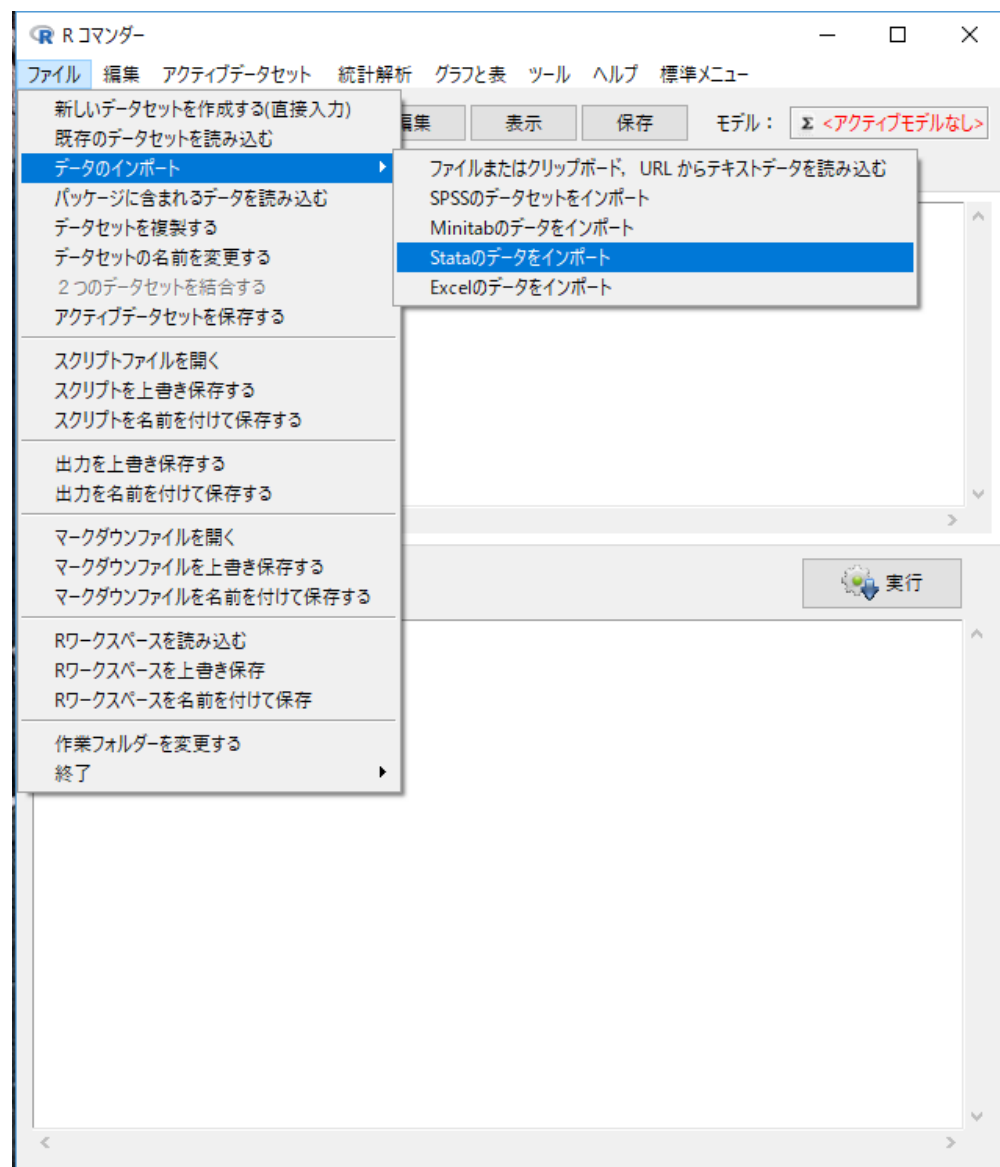
## 2 問題2:線形回帰モデル

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, 体重はどれくらい増えているか?
- ・ 男の子は女の子と比べて, 平均的に体重はどれくらい大きい/小さい?

### 2.1 データのインポート

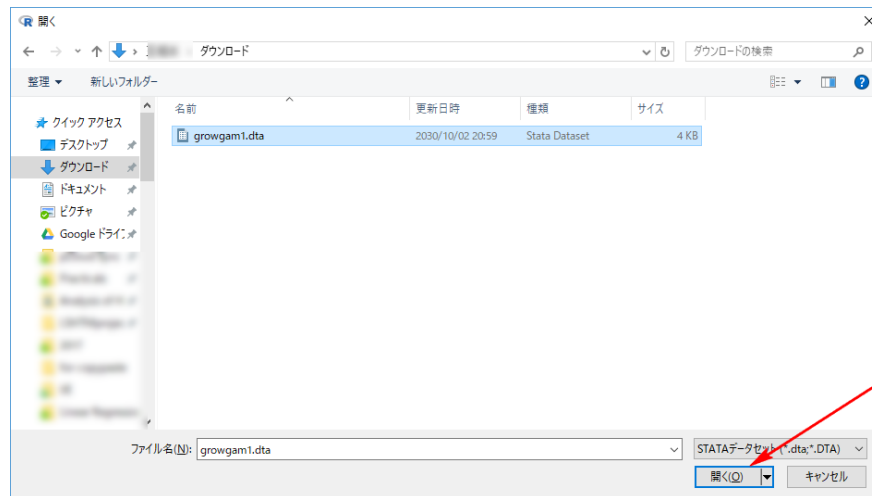
#### 2.1.1 ステップ 1



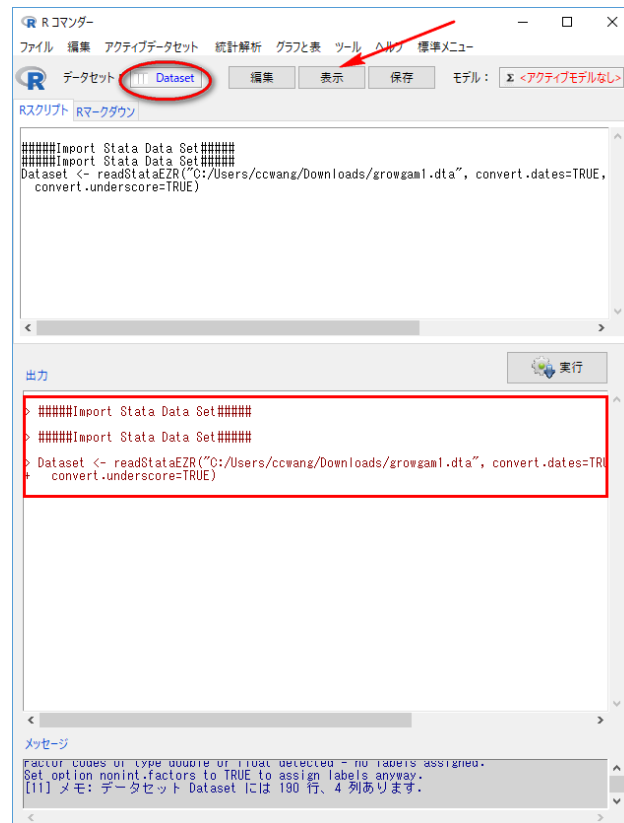
## 2.1.2 ステップ2



## 2.1.3 ステップ3



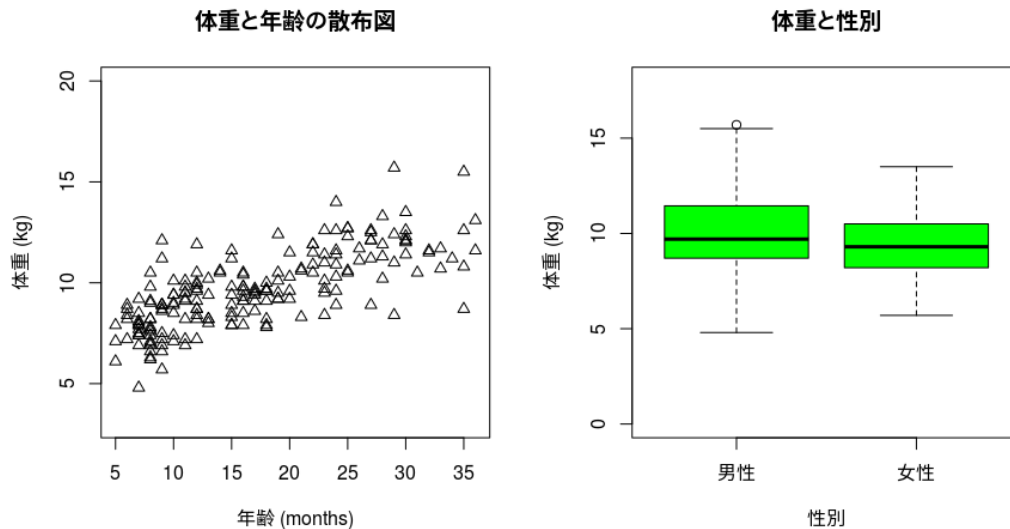
## 2.1.4 ステップ4



## 2.1.5 ステップ5

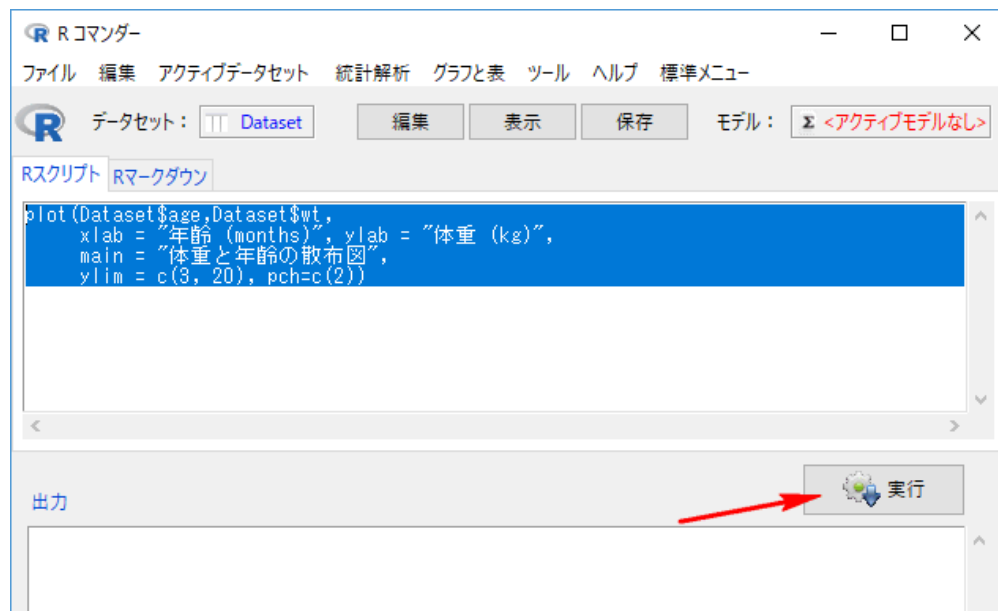
	sex	age	wt	len
1	2	23	8.4	73.2
2	2	22	10.9	84.4
3	2	6	7.2	68.7
4	1	24	10.3	83.7
5	1	14	10.5	79.2
6	2	18	9.6	75.8
7	2	30	11.4	84.4
8	1	24	11.4	84.8
9	1	17	9.4	74.3
10	2	27	12.5	82.6
11	2	16	9.1	74.1
12	1	30	12.0	86.4
13	1	18	7.8	71.9
14	2	15	8.5	77.6
15	2	13	8.0	72.2
16	1	11	9.1	72.4
17	2	8	10.5	71.2
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	61.2
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



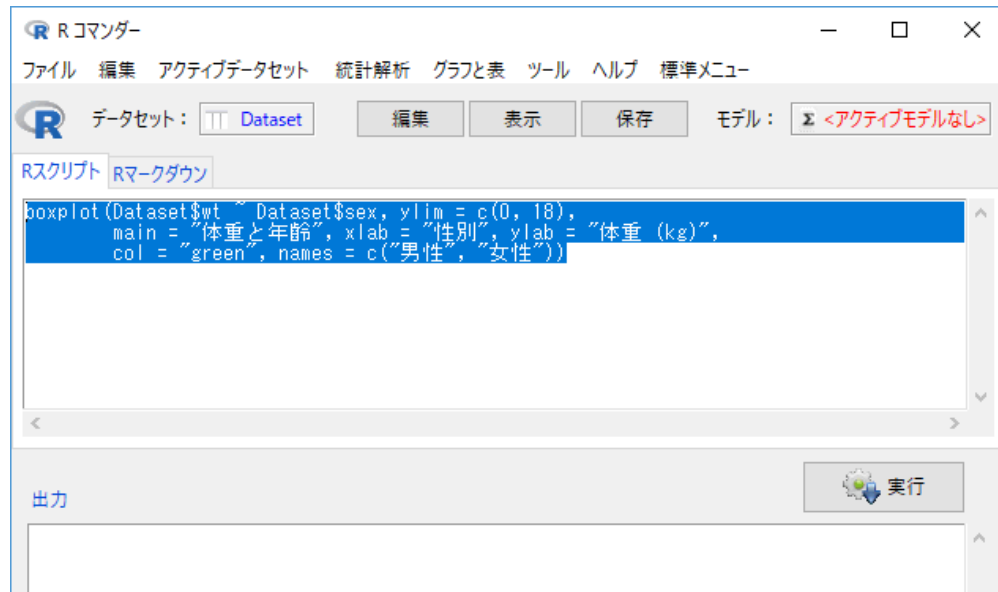
上記左のグラフを描くため,以下のコードをRスクリプトに入力して,実行をクリックしてください。

```
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:

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

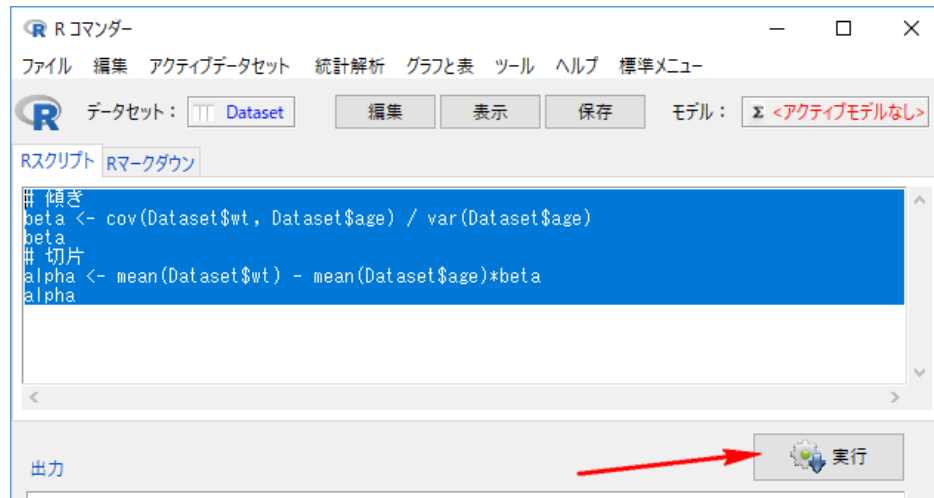
以下のコードをRスクリプトに入力して, 実行をクリックしてください. (結果を下の余白に記入すること)

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

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

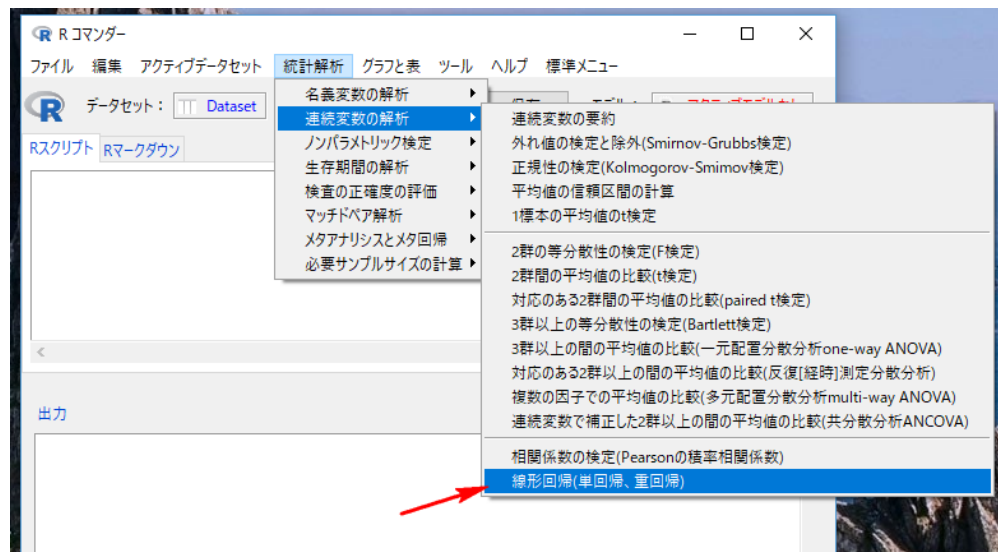
```
# 切片 (intercept)
alpha <- mean(Dataset$wt) - mean(Dataset$age)*beta
alpha
```

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

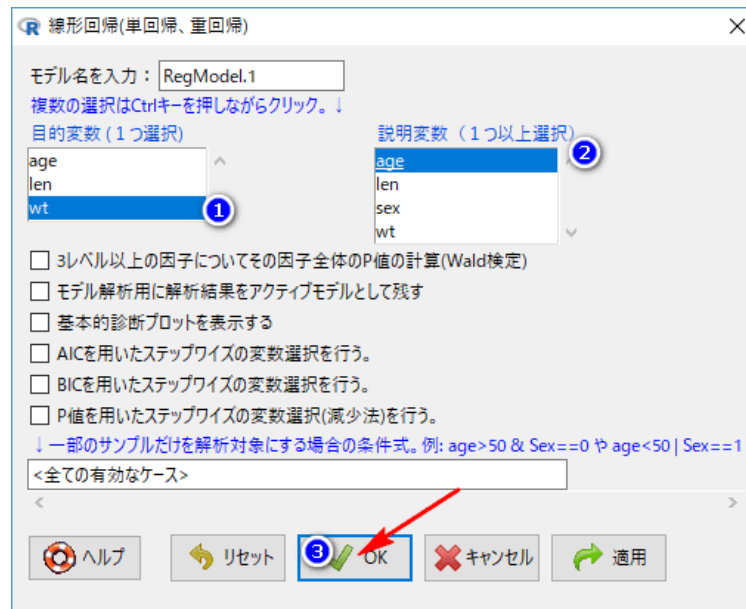


## 2.5 実際にEZRで線形モデルを作ってみよう:

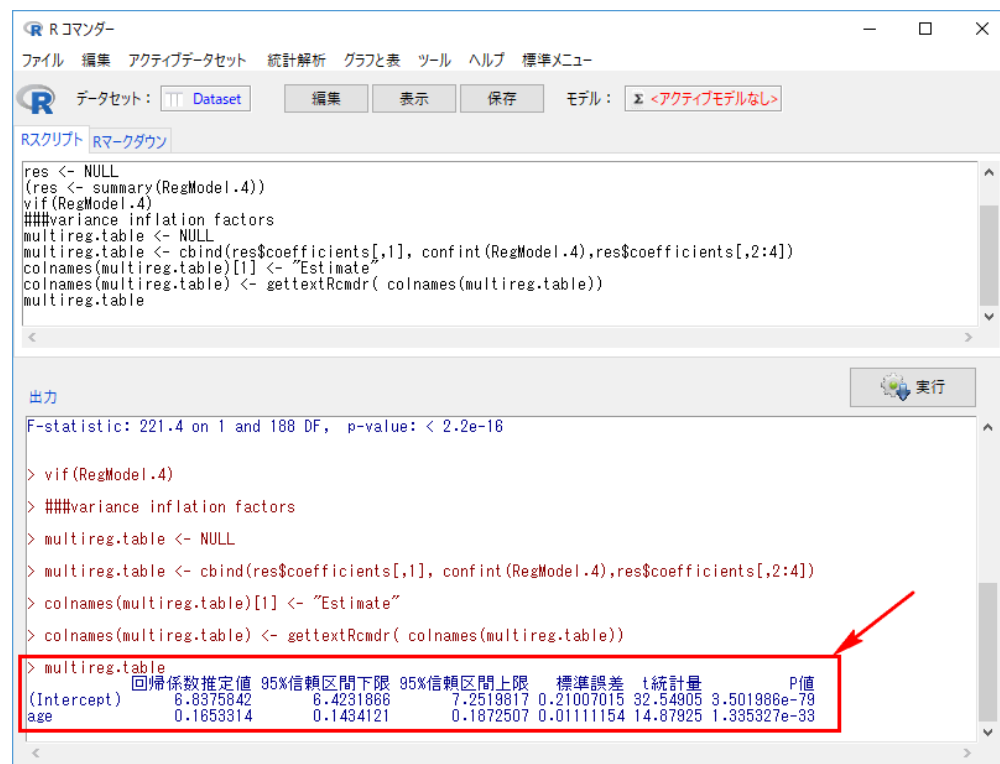
### 2.5.1 ステップ1



## 2.5.2 ステップ2



## 2.5.3 ステップ3



推定された回帰係数と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

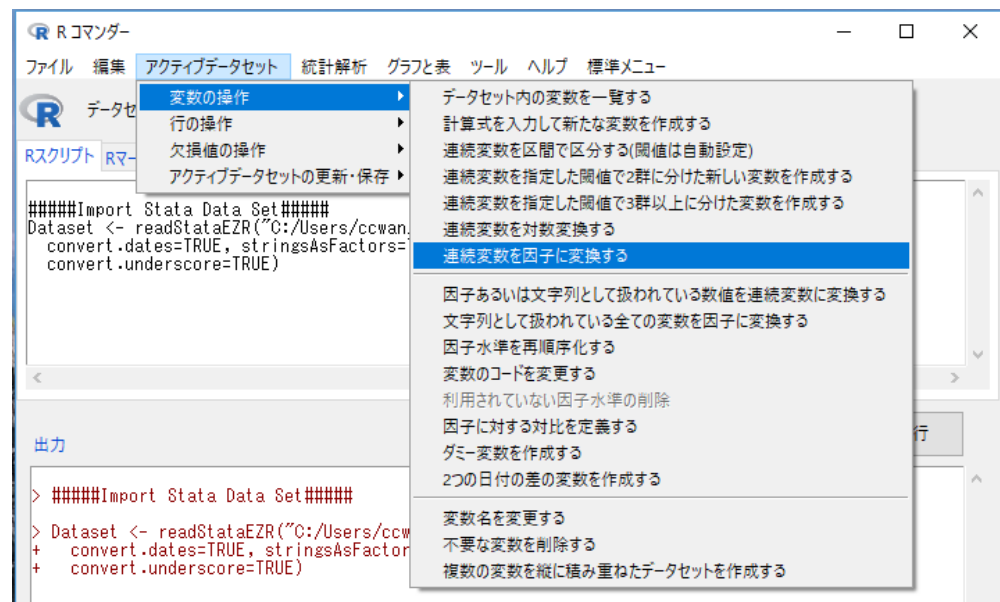
$$Y = 6.838 + 0.165X$$

- ・  $Y$  : 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.

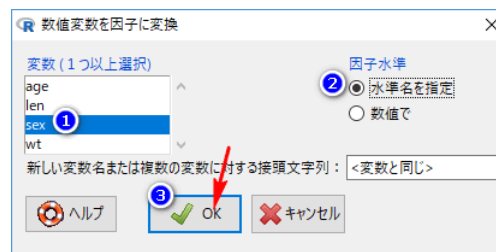
2.7 性別を説明変数に入れたモデルを作る

2.7.1 性別変数を因子 (factor) に変換する

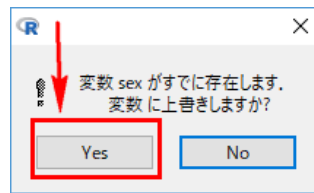
2.7.1.1 ステップ1



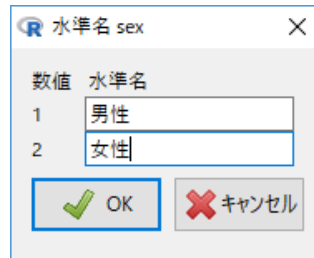
2.7.1.2 ステップ2



## 2.7.1.3 ステップ3

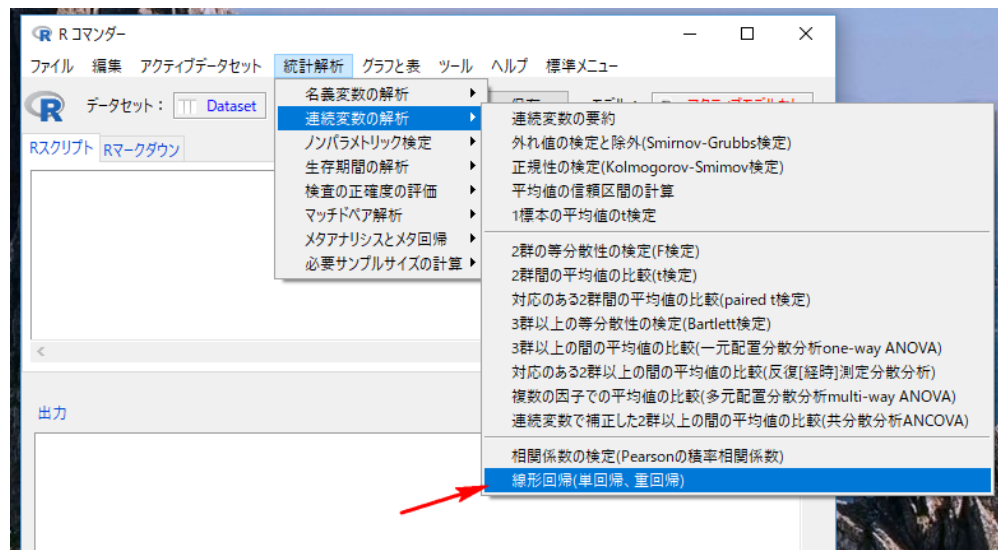


## 2.7.1.4 ステップ4-水準名に男性,女性を入力する

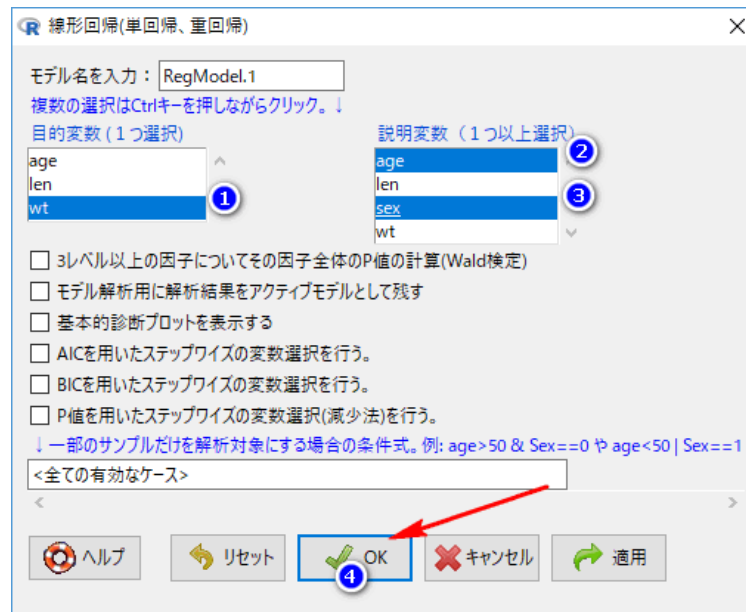


## 2.7.2 重回帰線形モデルを作る

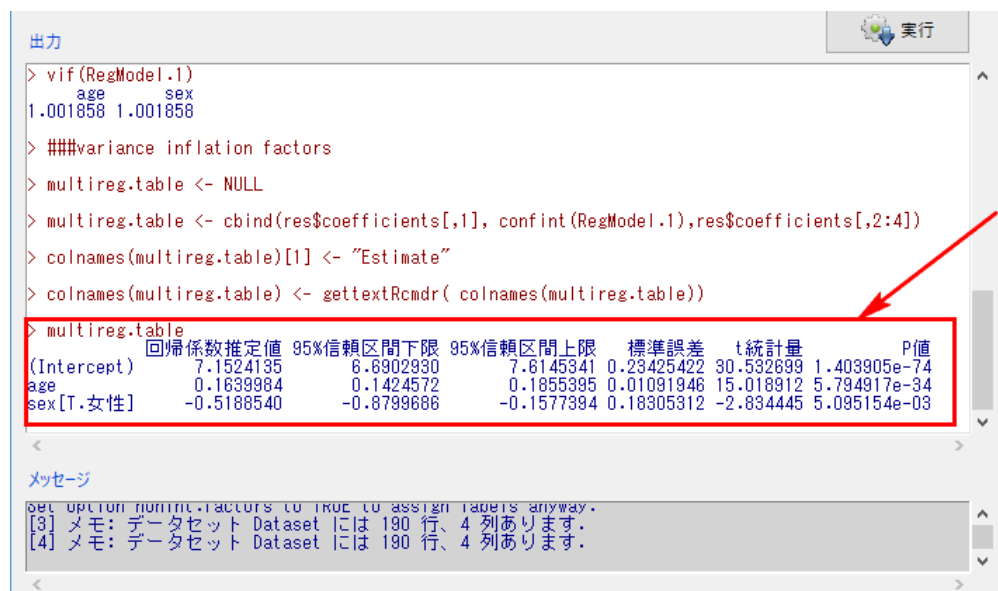
## 2.7.2.1 ステップ1



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



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.

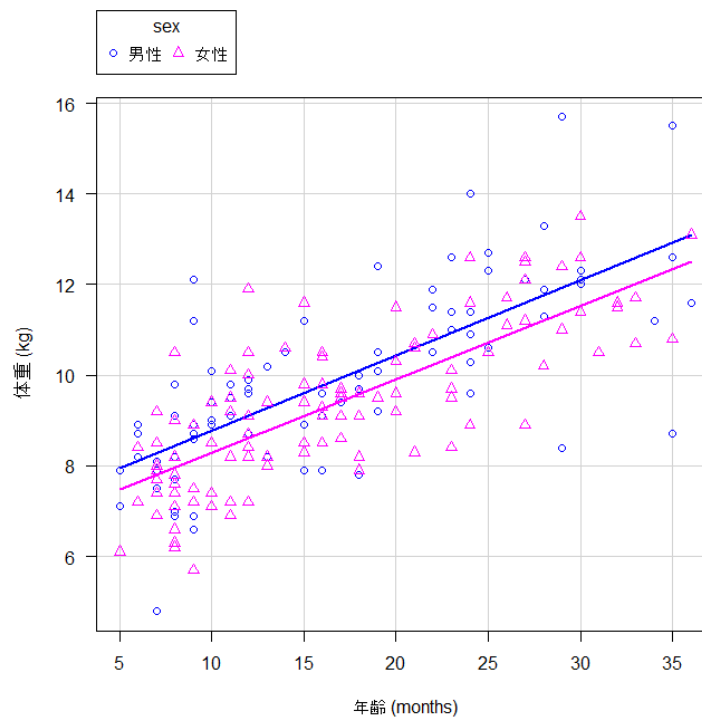
2.10 上記の重回帰線形モデルを用いて,年齢が34ヶ月の女の子の体重の予測値を計算せよ.

2.11 答え

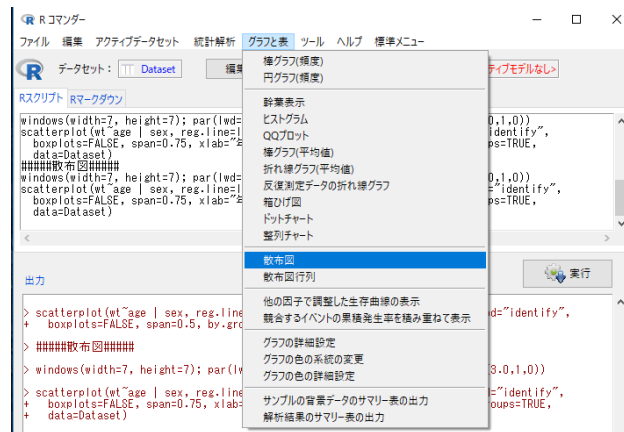
- ・  $X_1 = 34$ ;
- ・  $X_2 = 1$ ;

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

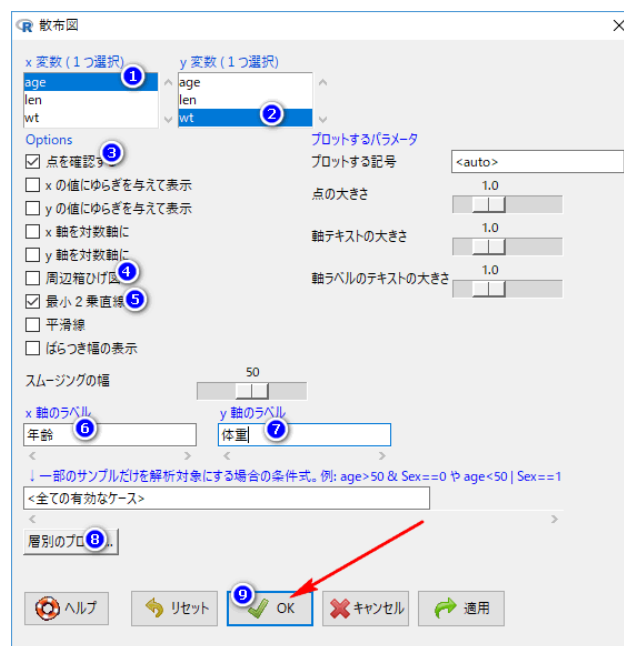
2.12 男女別の年齢と体重の散布図を描く



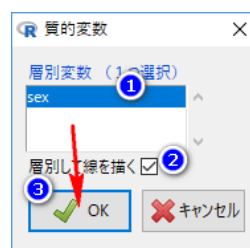
### 2.12.1 ステップ1



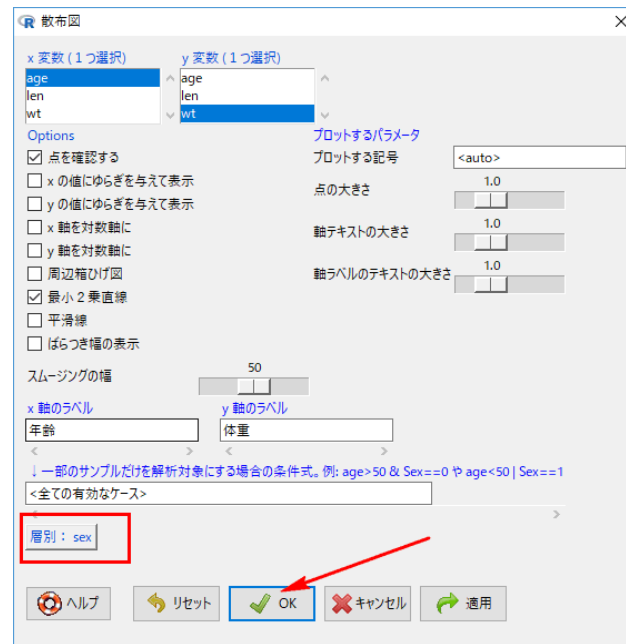
### 2.12.2 ステップ2



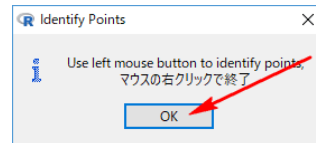
### 2.12.3 ステップ3



## 2.12.4 ステップ4



## 2.12.5 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. ある研究チームが流行していた地域の34個の村に住む15歳以上の全住民に目の検査を行った. 目の検査を受けた住民はWHOの診断基準を元に, 「視覚障害 (visually impaired)」と「視力正常 (normal vision)」に分類された. 対象者は年間観察され, その期間中に死亡者を登録された.

## 3.1 視覚障害と死亡の関係

視覚障害の有無と死亡リスクの関連を見るために, 以下の表をまとめた:

死亡	視力正常	視覚障害	合計
0	3874 (97.56%)	287 (87.77%)	4161 (96.81%)
1	97 (2.44%)	40 (12.23%)	137 (3.19%)
合計	3971 (100%)	327 (100%)	4298 (100%)

3.1.1 もし、視覚障害と対象者の死亡リスクと関連がない場合、下の表(各セルの期待者数)を入力せよ:

死亡	視力正常	視覚障害	合計
0	$3971 \times 4161 / 4298 = 3844.4232$	$327 \times 4161 / 4298 = 316.5768$	4161 (96.81%)
1	$3971 \times 137 / 4298 = 126.5768$	$327 \times 173 / 4298 = 10.4232$	137 (3.19%)
合計	3971 (100%)	327 (100%)	4298 (100%)

3.1.2 上記の2つの表の数字を使って  $\chi^2$  統計量を計算せよ

3.1.3 答え

$$\chi^2 = \frac{(3874 - 3844.4232)^2}{3844.4232} + \frac{(287 - 316.5768)^2}{316.5768} + \frac{(97 - 126.5768)^2}{126.5768} + \frac{(40 - 10.4232)^2}{10.4232}$$

$$= 93.829$$

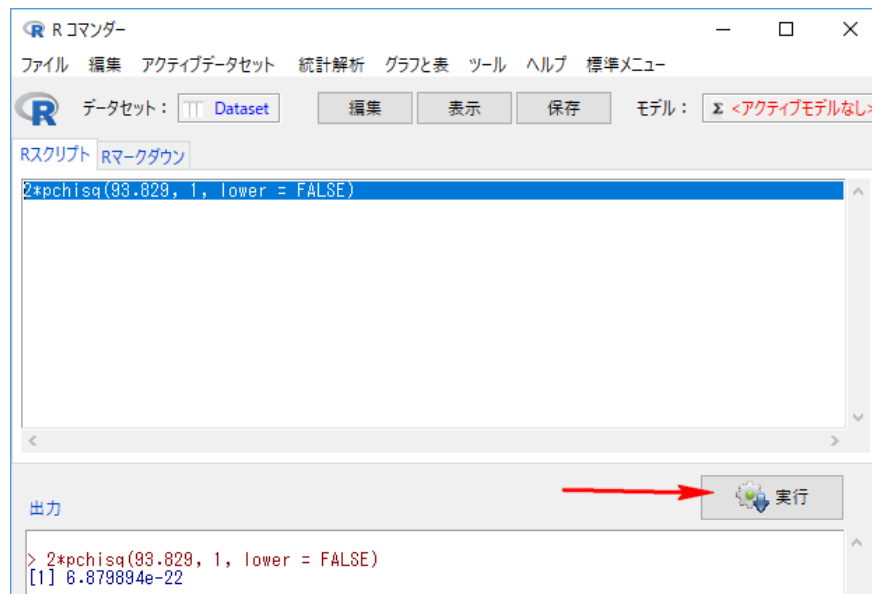
3.1.4  $2 \times 2$  の分割表では、自由度は  $(2 - 1) \times (2 - 1) = 1$  .

EZRで、 $\chi^2$ 統計量と自由度(df)を使って P 値を計算したい場合、以下のコマンドが利用できる:

`2*pchisq(chisquare統計量, df, lower = FALSE)`

`2*pchisq(93.829, 1, lower = FALSE)`

## [1] 6.879894e-22



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。

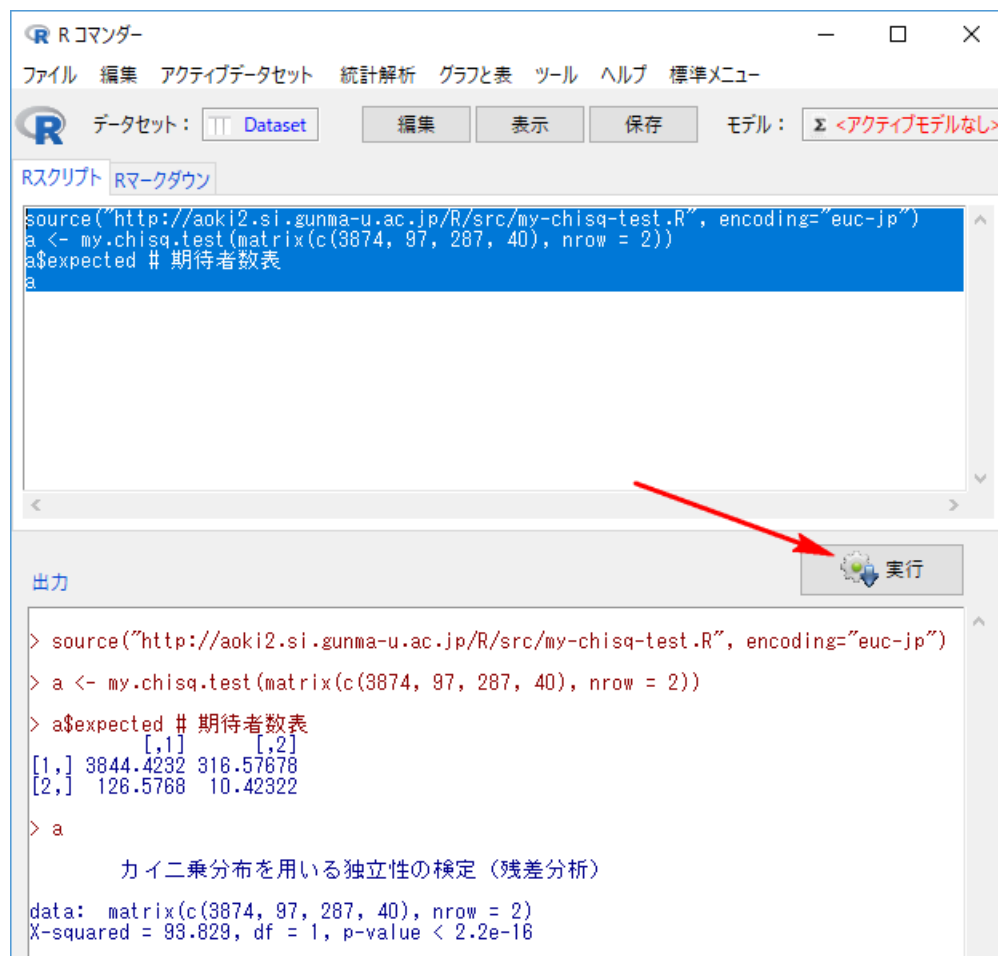


```
source("http://aoki2.si.gunma-u.ac.jp/R/src/my-chisq-test.R", encoding="euc-jp")
a <- my.chisq.test(matrix(c(3874, 97, 287, 40), nrow = 2))
a$expected # 期待者数表
```

```
##           [,1]      [,2]
## [1,] 3844.4232 316.57678
## [2,] 126.5768  10.42322
```

```
a
```

```
##
## カイ二乗分布を用いる独立性の検定（残差分析）
##
## data: matrix(c(3874, 97, 287, 40), nrow = 2)
## X-squared = 93.829, df = 1, p-value < 2.2e-16
```



手で計算した結果とは一致していると確認できる。

この検定結果は「視覚障害と対象者の死亡リスクと関連がない」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 there is no association between visual impairment and risk of death.)

3.1.5 視覚障害と死亡の関係を示すテーブルの数を元に、下表を完成せよ:

	視力正常	視覚障害	トータル
リスク (risk)	0.0244	0.1223	0.0319
オッズ (odds)	0.0250	0.1394	0.0329
対数オッズ (log-odds)	-3.689	-1.9704	-3.414

では、視覚障害と死亡の関連を示すオッズ比は:

$$OR = 0.1394 \div 0.025 = 5.576$$

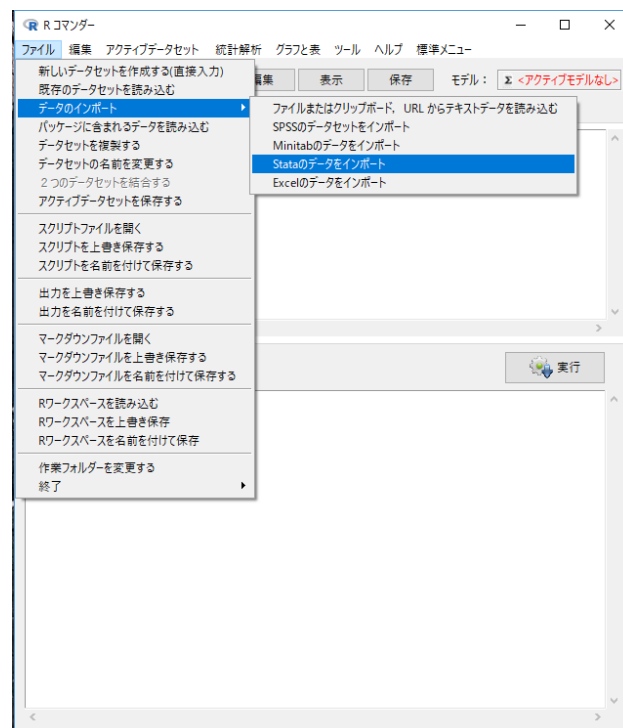
このオッズ比の対数を取った値  $\log(OR)$  は:

$$\log(OR) = 1.717$$

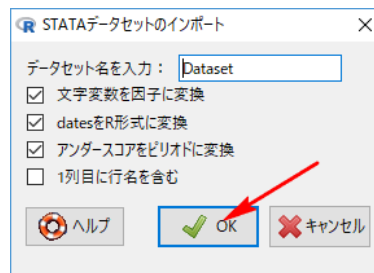
3.1.6 EZRでロジスティクス回帰モデルを作る

3.1.6.1 ステップ1ーデータのインポート

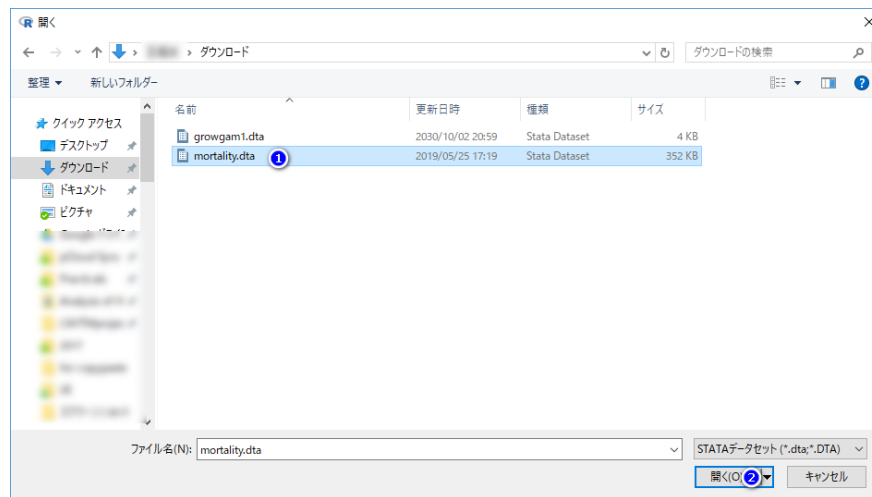
1.



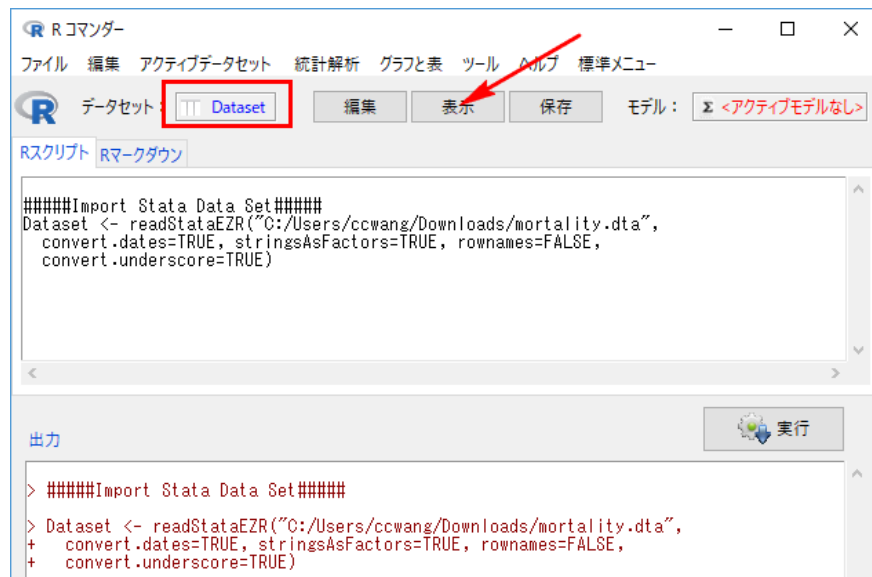
2.



3.



4.



5.

	id	area	district	vcode	compound.size	age	sex	ethnic	religion
1	1	0	Kahugu	1	16	59	Male	Gure	Christian
2	2	0	Kahugu	1	16	39	Female	Gure	Christian
3	3	0	Kahugu	1	16	38	Female	Gure	Christian
4	4	0	Kahugu	1	16	27	Female	Gure	Christian
5	5	0	Kahugu	1	16	19	Male	Gure	Christian
6	6	0	Kahugu	1	16	15	Male	Gure	Christian
7	14	0	Kahugu	1	20	74	Male	Gure	Christian
8	15	0	Kahugu	1	20	56	Female	Gure	Christian
9	16	0	Kahugu	1	20	45	Female	Gure	Christian
10	17	0	Kahugu	1	20	43	Female	Gure	Christian
11	18	0	Kahugu	1	20	22	Male	Gure	Christian
12	19	0	Kahugu	1	20	17	Male	Gure	Christian
13	22	0	Kahugu	1	20	19	Male	Gure	Christian
14	23	0	Kahugu	1	20	24	Male	Gure	Christian
15	24	0	Kahugu	1	20	21	Male	Gure	Christian
16	29	0	Kahugu	1	20	18	Female	Gure	Christian
17	30	0	Kahugu	1	20	30	Male	Gure	Christian
18	31	0	Kahugu	1	20	25	Female	Gure	Christian
19	33	0	Kahugu	1	6	71	Male	Gure	Traditional
20	34	0	Kahugu	1	6	28	Female	Gure	Christian
21	35	0	Kahugu	1	6	33	Male	Gure	Christian
22	38	0	Kahugu	1	7	30	Male	Gure	Christian
23	39	0	Kahugu	1	7	25	Female	Gure	Christian
24	42	0	Kahugu	1	7	20	Male	Gure	Christian
25	43	0	Kahugu	1	7	20	Female	Gure	Christian
26	44	0	Kahugu	1	22	75	Male	Gure	Christian
27	45	0	Kahugu	1	22	65	Female	Gure	Christian
28	46	0	Kahugu	1	22	52	Female	Gure	Christian
29	47	0	Kahugu	1	22	49	Female	Gure	Christian
30	48	0	Kahugu	1	22	17	Male	Gure	Christian

## 3.1.6.2 ステップ2ーロジスティクスモデルを作る

6.

R コマンドー

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 標準メニュー

データセット: Dataset

Rスクリプト Rマークダウン

```
#####Import Stata Data Set#####
Dataset <- readStataEZR("C:/Users/ccwang/Downloads/mortality.dta",
  convert.dates=TRUE, stringsAsFactors=TRUE, rownames=FALSE,
  convert.underscore=TRUE)
```

出力

```
> #####Import Stata Data Set#####
> Dataset <- readStataEZR("C:/Users/ccwang/Downloads/mortality.dta",
+   convert.dates=TRUE, stringsAsFactors=TRUE, rownames=FALSE,
+   convert.underscore=TRUE)
```

統計解析

- 名義変数の解析
  - 頻度分布
- 連続変数の解析
  - 比率の信頼区間の計算
- ノンパラメトリック検定
  - 1標本の比率の検定
- 生存期間の解析
  - 2群の比率の差の信頼区間の計算
- 検査の正確度の評価
  - 2群の比率の比の信頼区間の計算
- マッチドペア解析
  - 分割表の直接入力と解析
- メタアナリシスとメタ回帰
  - 分割表の作成と群間の比率の比較(Fisherの正確検定)
  - 対応のある比率の比較(二分割表の対称性の検定、McNemar検定)
  - 対応のある3群以上の比率の比較(Cochran Q検定)
  - 比率の傾向の検定(Cochran-Armitage検定)
- 必要サンプルサイズの計算

二値変数に対する多変量解析(ロジスティック回帰)

実行

7.

二値変数に対する多変量解析(ロジスティック回帰)

モデル名を入力: GLM.1

変数 (ダブルクリックして式に入れる)

age  
agebin  
agegrp  
area  
bmi  
bmigrp  
compound.size  
diastolic  
died  
district [因子]

モデル式: + \* : / %in% - ^ ( )

目的変数 ~ 説明変数

層別化因子は + strata(####)と入力する ↑ ↑ ↑

☐ 3レベル以上の因子についてその因子全体のP値の計算(Wald検定)

☐ モデル解析用に解析結果をアクティブモデルとして残す

☐ ROC曲線を表示する

☐ 基本的診断プロットを表示する

☐ 傾向スコア変数を自動作成する

☐ AICを用いたステップワイズの変数選択を行う。

☐ BICを用いたステップワイズの変数選択を行う。

☐ P値を用いたステップワイズの変数選択(減少法)を行う。

↓一部のサンプルだけを解析対象にする場合の条件式。例: age>50 & Sex==0 や age<50 | Sex==1

<全ての有効なケース>

ヘルプ リセット OK キャンセル 適用

8.

二値変数に対する多変量解析(ロジスティック回帰)

モデル名を入力: GLM.1

変数 (ダブルクリックして式に入れる)

mfperm  
mfpos  
occupation [因子]  
pulse  
religion [因子]  
sex [因子]  
systolic  
vcode  
vimp  
weight

モデル式: + \* : / %in% - ^ ( )

目的変数 died ~ 説明変数 vimp

層別化因子は + strata(####)と入力する ↑ ↑ ↑

☐ 3レベル以上の因子についてその因子全体のP値の計算(Wald検定)

☐ モデル解析用に解析結果をアクティブモデルとして残す

☐ ROC曲線を表示する

☐ 基本的診断プロットを表示する

☐ 傾向スコア変数を自動作成する

☐ AICを用いたステップワイズの変数選択を行う。

☐ BICを用いたステップワイズの変数選択を行う。

☐ P値を用いたステップワイズの変数選択(減少法)を行う。

↓一部のサンプルだけを解析対象にする場合の条件式。例: age>50 & Sex==0 や age<50 | Sex==1

<全ての有効なケース>

ヘルプ リセット OK キャンセル 適用

9.

```
##variance inflation factors
odds <- NULL
odds <- data.frame(exp( summary(GLM.1)$coef[,1:2] %*% rbind(c(1,1,1), 1.96*c(0,
-1,1))))
odds <- cbind(odds, summary(GLM.1)$coefficients[,4])
odds <- signif(odds, digits=3)
names(odds) <- gettextRcmdr(c("odds ratio", "Lower 95%CI", "Upper 95%CI",
"p.value"))
odds
```

出力

```
> ##variance inflation factors
> odds <- NULL
> odds <- data.frame(exp( summary(GLM.1)$coef[,1:2] %*% rbind(c(1,1,1), 1.96*c(0,
+ -1,1))))
> odds <- cbind(odds, summary(GLM.1)$coefficients[,4])
> odds <- signif(odds, digits=3)
> names(odds) <- gettextRcmdr(c("odds ratio", "Lower 95%CI", "Upper 95%CI",
+ "p.value"))
> odds
```

	オッズ比	95%信頼区間下限	95%信頼区間上限	P値
(Intercept)	0.025	0.0205	0.0306	8.91e-282
vimp	5.570	3.7800	8.2000	3.71e-18

メッセージ

```
get OPTION nonlin.factors to TRUE to assign labels anyway.
[3] メモ: データセット Dataset には 4298 行、28 列あります。
[4] エラー: model contains fewer than 2 terms
```

計算したオッズ比はこの結果とは一致しているかを確認してください。

10. 出力のところをスクロールアップすると  $\log(\text{OR})$  の結果が確認できる

```
##variance inflation factors
odds <- NULL
odds <- data.frame(exp( summary(GLM.1)$coef[,1:2] %*% rbind(c(1,1,1), 1.96*c(0,
-1,1))))
odds <- cbind(odds, summary(GLM.1)$coefficients[,4])
odds <- signif(odds, digits=3)
names(odds) <- gettextRcmdr(c("odds ratio", "Lower 95%CI", "Upper 95%CI",
"p.value"))
odds
```

出力

```
> GLM.1 <- glm(died ~ vimp, family=binomial(logit), data=Dataset)
> summary(GLM.1)
```

```
Call:
glm(formula = died ~ vimp, family = binomial(logit), data = Dataset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5108  -0.2224  -0.2224  -0.2224   2.7247

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.6873      0.1028  -35.870  <2e-16 ***
vimp          1.7167      0.1976   8.687  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

メッセージ

```
get OPTION nonlin.factors to TRUE to assign labels anyway.
[3] メモ: データセット Dataset には 4298 行、28 列あります。
[4] エラー: model contains fewer than 2 terms
```

- ・ (intercept/切片) の  $-3.689$  は”視力正常”群の対数オッズであることを確認できる;
- ・ **vimp** (視覚障害)の回帰係数 **1.7167** は”視覚障害”と”視力正常”群に比べた  $\log(\text{OR}) = \log(\text{odds in 視覚障害}) - \log(\text{odds in 視力正常}) = -1.9704 - (-3.689)$ である.

### 3.2 年齢の影響を考慮する

	視覚障害 (0 = no, 1 = yes)									
死亡	0	1	0	1	0	1	0	1	0	1
1 = yes	29	2	38	10	15	11	15	17	97	40
0 = no	2301	22	1271	124	212	69	90	72	3874	287
n	2330	24	1309	134	227	80	105	89	3971	327
年齢	15-34		35-54		55-64		65 +		Total	

上記のデータをよく見ると、視覚障害のオッズは年齢と共に上昇している (年齢が15-34歳群の  $(2 + 22)/(29 + 2301) = 0.010$  から年齢が65歳以上群の  $(17 + 72)/(15 + 90) = 0.848$  に上げた). しかし、年齢の上昇と共に、死亡のオッズも上がる. 年齢はここで、交絡因子 (confounder) と定義される.

3.2.1 以上のデータと解説をよく理解した上で、下表を完成せよ:

	オッズ		
年齢	視力正常	視覚障害	オッズ比
15-34	$29/2301 = 0.01260$	0.0909	7.214
35-54	0.02990	0.08065	2.6973
55-64	0.07075	0.15942	2.2533
65+	0.16667	0.23611	1.4166

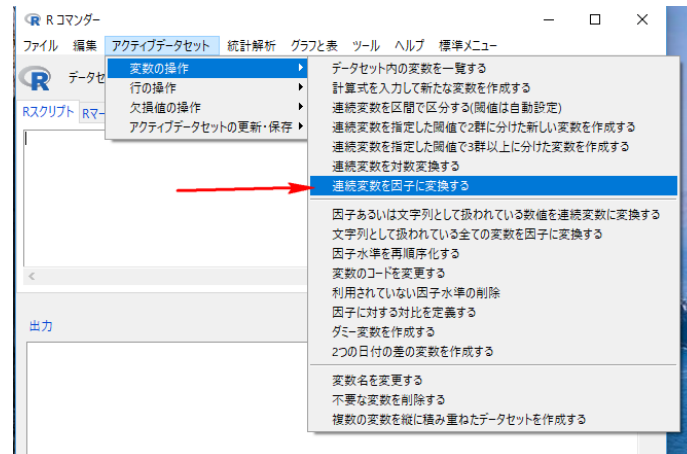
Look at the stratum-specific odds ratios, it appears that there may be a trend and that the odds ratio for visual impairment may decline with age.

各年齢層では、視覚障害と死亡の関係を評価するオッズ比は大きく異なることがわかり、年齢と共に減っていく傾向があるかもしれない. 年齢を考慮せずに視覚障害と死亡との関係を評価することは妥当ではないことが示唆される.

## 3.2.2 EZRで年齢グループを調整したロジスティクス回帰モデルを作る

## 3.2.2.1 年齢グループ agegrp 変数を因子 (factor) に変換する

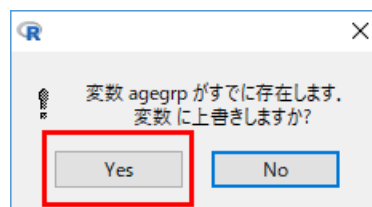
1.



2.

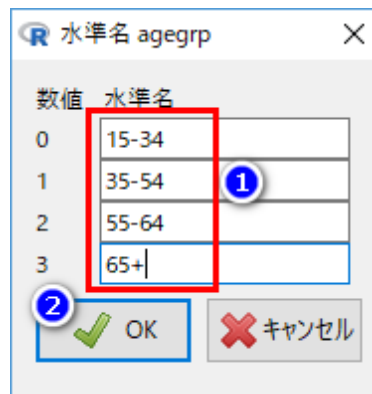


3.



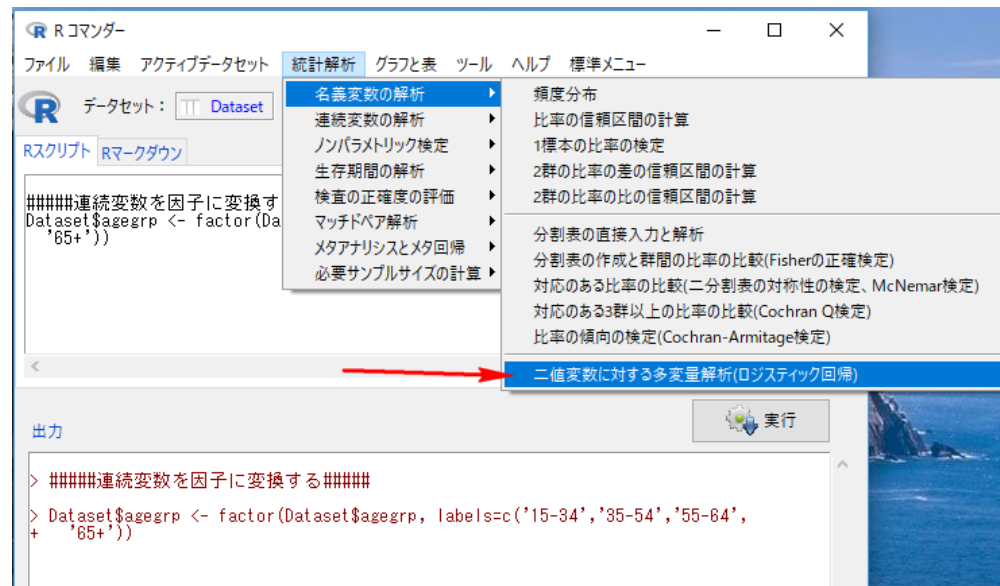


## 4. 水準名に各年齢グループの名前を入力する

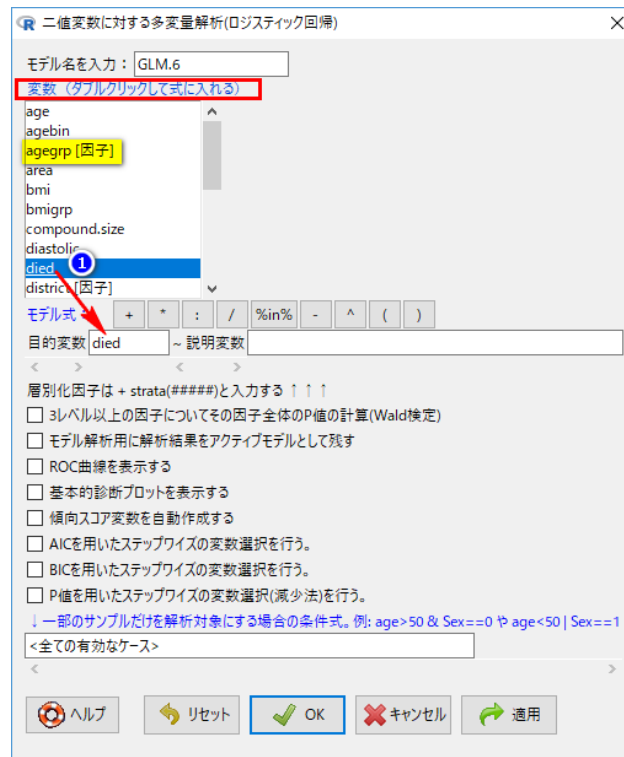


## 3.2.2.2 多変量ロジスティクス回帰モデルを作る

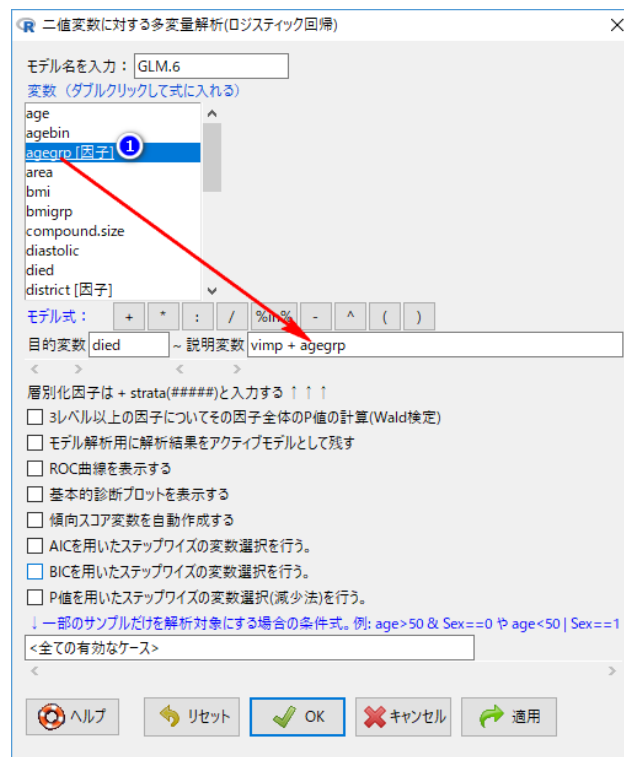
## 1.



2. agegrp が因子になったことが確認できる。  
died (死亡) を目的変数へ,vimp (視覚障害) を説明変数へ



3. agegrp も目的変数へ移動すると自動的に + が入れられる.OKをクリックする。



4. 視覚障害と死亡の関係を評価する年齢調整オッズ比が計算される。

```
## Variance inflation factors
odds <- NULL
odds <- data.frame(exp( summary(GLM.6)$coef[,1:2] %*% rbind(c(1,1,1), 1.98*c(0,
-1,1))))
odds <- cbind(odds, summary(GLM.6)$coefficients[,4])
odds <- signif(odds, digits=3)
names(odds) <- gettextRcmdr(c("odds ratio", "Lower 95%CI", "Upper 95%CI",
p.value"))
odds
```

	オッズ比	95%信頼区間下限	95%信頼区間上限	P値
(Intercept)	0.0132	0.00925	0.0188	1.47e-126
vimp	2.2000	1.41000	3.4400	5.33e-04
agegrp[T.35-54]	2.3500	1.48000	3.7400	2.77e-04
agegrp[T.55-64]	5.4200	3.06000	9.5100	4.20e-09
agegrp[T.65+]	5.9000	5.54000	17.7000	9.84e-15

3.2.2.3 単変量ロジスティクス回帰モデルで評価した粗オッズ比 (crude odds ratio) と比べ、年齢調整オッズ比はどう変わったかを説明せよ。

3.2.2.4 答え 年齢を考慮していない場合、視覚障害者は視力正常者と比べ、三年間の間に死亡するオッズが5.57倍であり、95%信頼区間が 3.78 - 8.20 と推定される。

多変量ロジスティクス回帰モデルを用いて、視覚障害と死亡の关系到年齢の交絡考慮した後、オッズ比が 2.20 になり、95%信頼区間が 1.41 - 3.44 と推定される。このオッズ比が大きく変化した(小さくなった)ことは、年齢がこの関連の強い交絡因子であることを示唆される。また、年齢調整したオッズ比の95%信頼区間は 1 を跨いでいない。以上の結果を踏まえて、「視覚障害者は視力正常者と比べ、観察期間中に死亡するオッズが有意に高いこと」を支持するために、非常に強い証拠を提供した。

(なお、この解析は、「各年齢層内の視覚障害と死亡の关系が等しい」という前提が仮定される。つまり、2.20は各年齢層の視覚障害と死亡の関係を評価する共通オッズ比 [common odds ratio] である。)

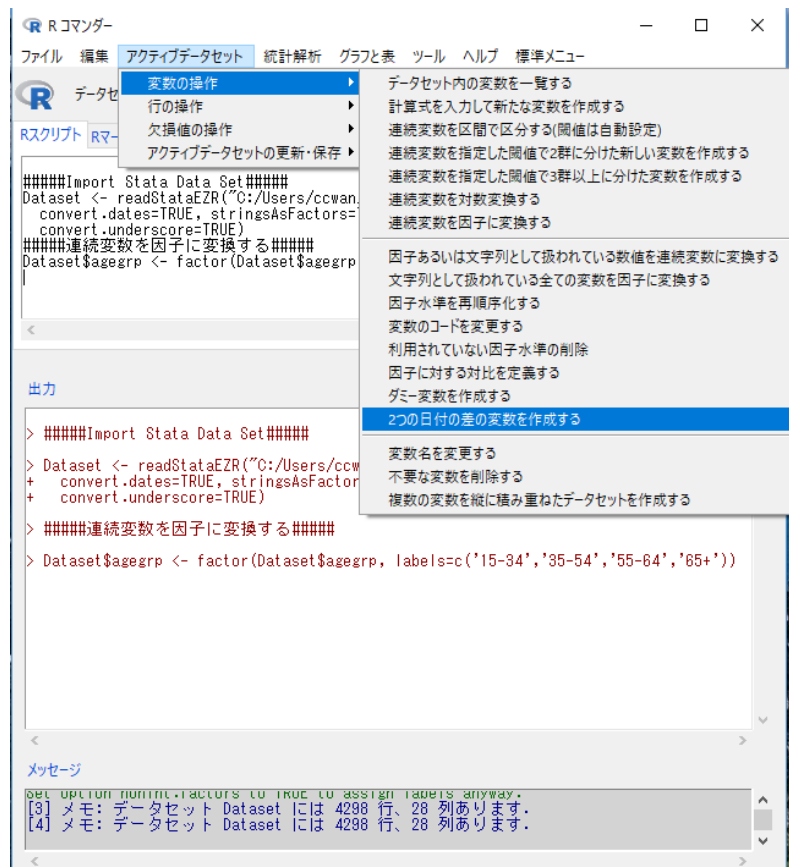
## 4 問題4:生存分析

問題3の研究で,実は対象者が研究に参加した時点と研究終了時点(死亡,打ち切り,または研究期間が終了した)の時間も記録されている:

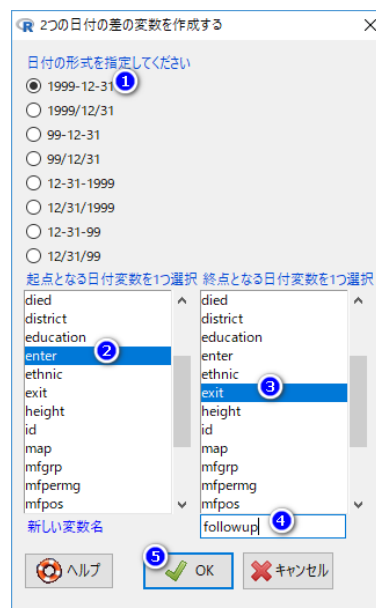
```
## # A tibble: 4,298 x 6
##       id agegrp enter      exit      vimp died
##   <dbl> <fct> <date>    <date>    <dbl+lbl> <dbl>
## 1     1  1 55-64 1989-05-09 1992-02-05 0 [Normal]      0
## 2     2  2 35-54 1989-05-09 1992-02-05 0 [Normal]      0
## 3     3  3 35-54 1989-05-09 1992-02-05 1 [Visually impaired] 0
## 4     4  4 15-34 1989-05-09 1992-02-05 0 [Normal]      0
## 5     5  5 15-34 1989-05-09 1992-02-05 0 [Normal]      0
## 6     6  6 15-34 1989-05-09 1989-08-12 0 [Normal]      0
## 7     7 14 65+   1989-05-09 1992-02-05 0 [Normal]      0
## 8     8 15 55-64 1989-05-11 1992-02-05 1 [Visually impaired] 0
## 9     9 16 35-54 1989-05-12 1992-02-05 1 [Visually impaired] 0
## 10    17 35-54 1989-05-09 1992-02-05 0 [Normal]      0
## 11    18 15-34 1989-05-09 1992-02-05 0 [Normal]      0
## 12    19 15-34 1989-05-12 1992-02-05 0 [Normal]      0
## 13    22 15-34 1989-05-09 1992-02-05 0 [Normal]      0
## 14    23 15-34 1989-05-09 1992-02-05 0 [Normal]      0
## 15    24 15-34 1989-05-09 1992-02-05 0 [Normal]      0
## 16    29 15-34 1989-05-10 1992-02-06 0 [Normal]      0
## 17    30 15-34 1989-05-11 1992-02-06 0 [Normal]      0
## 18    31 15-34 1989-05-09 1992-02-05 0 [Normal]      0
## 19    33 65+   1989-05-09 1992-02-05 0 [Normal]      0
## 20    34 15-34 1989-05-09 1992-02-06 0 [Normal]      0
## 21    35 15-34 1989-05-11 1992-02-05 0 [Normal]      0
## 22    38 15-34 1989-05-15 1991-09-05 0 [Normal]      1
## 23    39 15-34 1989-05-10 1992-02-06 0 [Normal]      0
## 24    42 15-34 1989-05-10 1992-02-05 0 [Normal]      0
## 25    43 15-34 1989-05-09 1992-02-06 0 [Normal]      0
## 26    44 65+   1989-05-14 1992-02-05 1 [Visually impaired] 0
## 27    45 65+   1989-05-14 1992-02-05 0 [Normal]      0
## 28    46 35-54 1989-05-14 1992-02-08 0 [Normal]      0
## 29    47 35-54 1989-05-18 1992-02-05 0 [Normal]      0
## 30    48 15-34 1989-05-16 1992-02-08 0 [Normal]      0
## 31    55 55-64 1989-05-14 1992-02-08 1 [Visually impaired] 0
## 32    56 15-34 1989-05-14 1992-02-06 0 [Normal]      0
## 33    59 15-34 1989-05-14 1992-02-08 0 [Normal]      0
## 34    62 15-34 1989-05-14 1992-02-08 0 [Normal]      0
## 35    63 15-34 1989-05-14 1992-02-05 0 [Normal]      0
## # ... with 4,263 more rows
```

## 4.0.1 EZRでは、追跡期間を計算するために、日付の差を取る必要がある

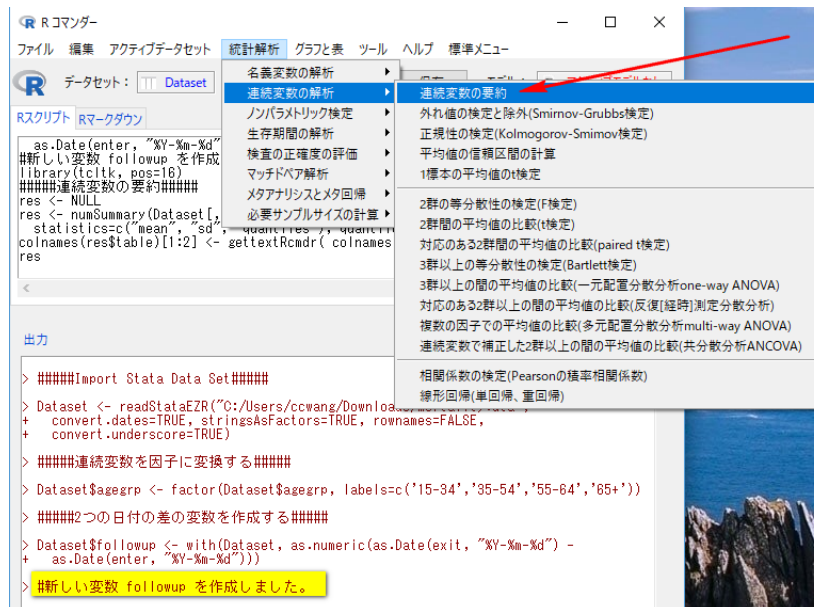
1.



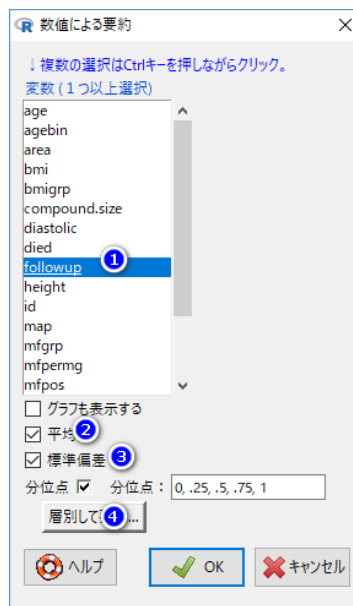
2. 追跡期間の変数名を followup とする



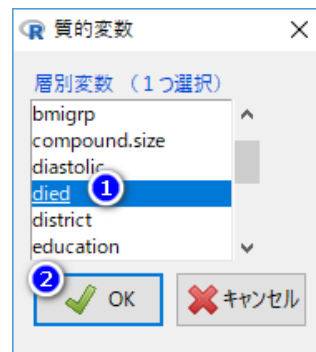
## 3. followup を作成したと確認メッセージが出る. 追跡期間の要約を調べる:



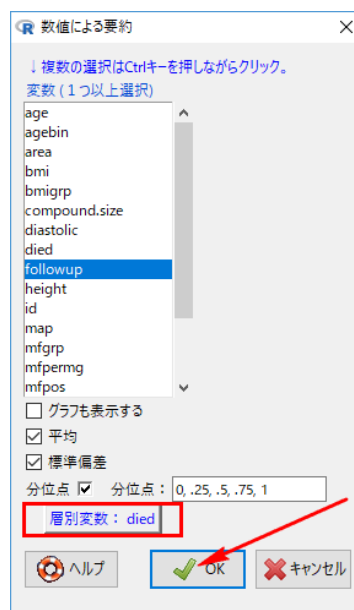
## 4. イベント別(死亡, 打ち切り)により追跡期間を計算する



5.



6.



7. 単純要約を見ると,死亡者と生存者の追跡期間(日数)の平均値(と中央値)はどっちが長い?

```

出力
> colnames(res$table)[1:2] <- gettextRcmdr( colnames(res$table)[1:2])
> res
  平均 標準偏差   0% 25%   50%   75% 100% data:n
0 977.2749 179.7676 24.0 909 1001 1088.5 1187 3971
1 929.7890 213.5089 22.5 896  987 1030.5 1158  327

#####連続変数の要約#####
> res <- NULL
> res <- numSummary(Dataset[, "followup"], groups=Dataset$died,
+   statistics=c("mean", "sd", "quantiles"), quantiles=c(0,.25,.5,.75,1))
> colnames(res$table)[1:2] <- gettextRcmdr( colnames(res$table)[1:2])
> res
  平均 標準偏差   0% 25%   50%   75% 100% data:n
0 988.0888 159.3268 22.5 910.0 1001.0 1090 1187.0 4161
1 535.4891 285.6341 24.0 250.5  547.5  817  891.5  137

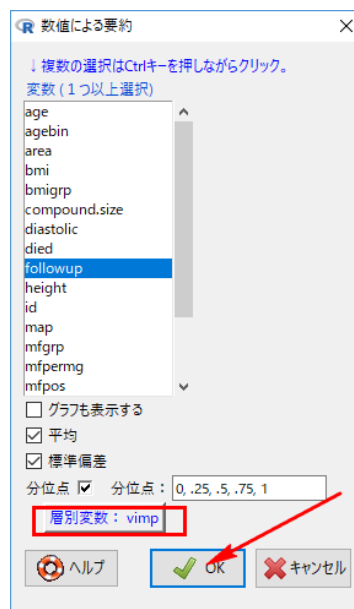
メッセージ
[5] メモ: データセット Dataset には 4298 行、28 列の Summary
[4] メモ: データセット Dataset には 4298 行、28 列あります。
[5] メモ: データセット Dataset には 4298 行、28 列あります。

```

8. 層別変数を `vimp` (視覚障害) に変更し, 追跡期間の要約を比較してみる



9.



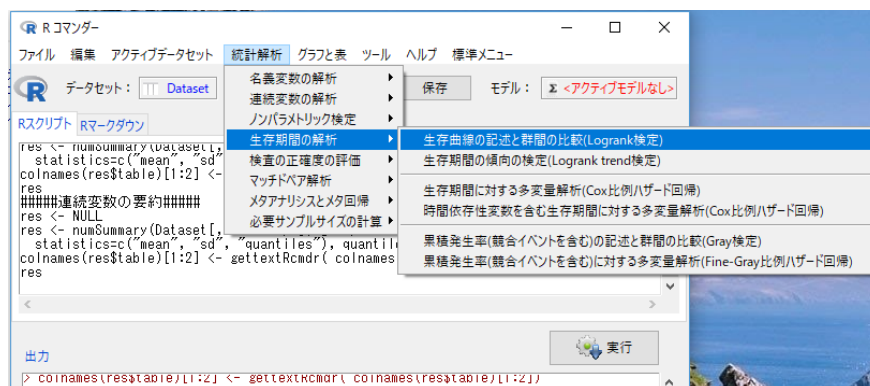
10. 視覚障害者と視力正常者と比べ, 追跡期間の違いはあるか?



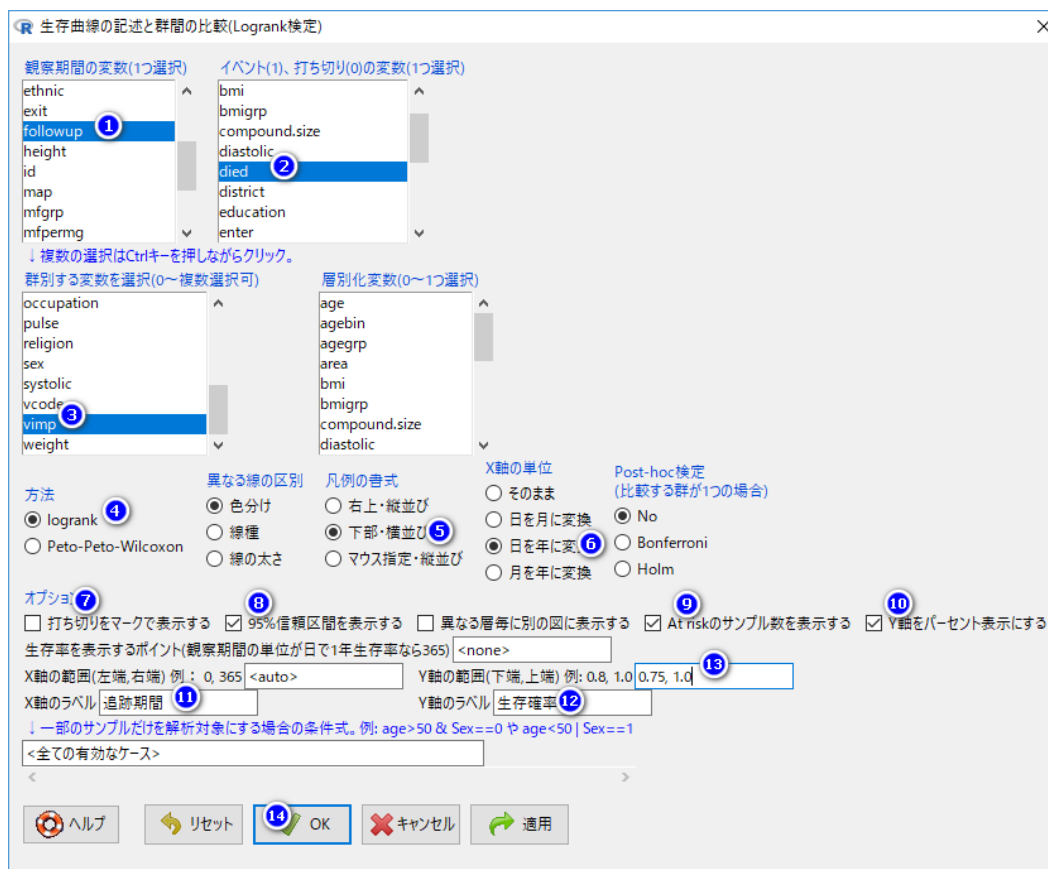


## 4.0.2 生存表と Kaplan-Meier グラフを作成する

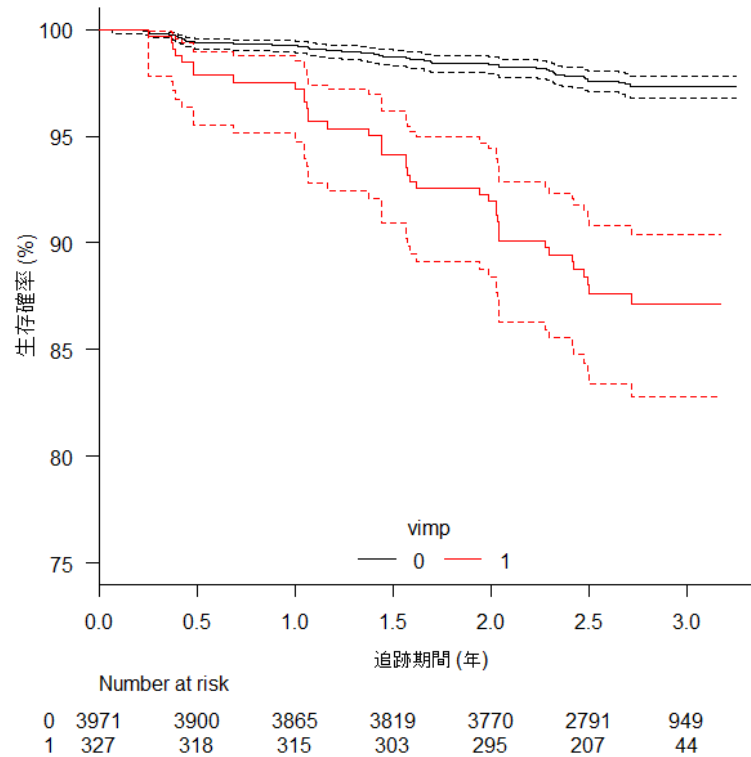
1.



2.



3. 視覚障害者(赤線)と視力正常者(黒線)のカプランマイヤー生存曲線:(点線は95%信頼区間を示す)



視覚障害者の生存率は視力正常者より低いことが分かる。

4. log-rank 検定の結果も同時に示される。 $p = < 2e - 16 < 0.00001$ の結果から、「視覚障害者と視力正常者の生存曲線が等しい」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。

```

出力
> res <- NULL
> (res <- survdiff(Surv(followup,died==1)~vimp, data=Dataset, rho=0, na.action =
+ na.omit))
Call:
survdiff(formula = Surv(followup, died == 1) ~ vimp, data = Dataset,
na.action = na.omit, rho = 0)

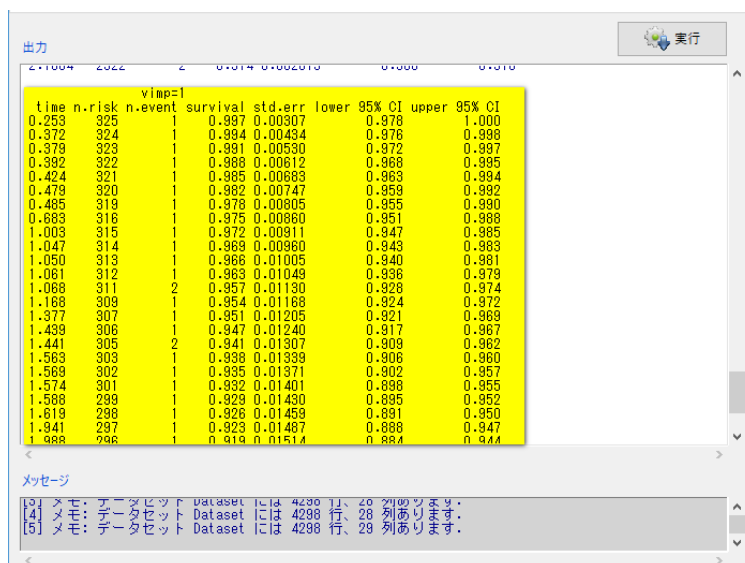
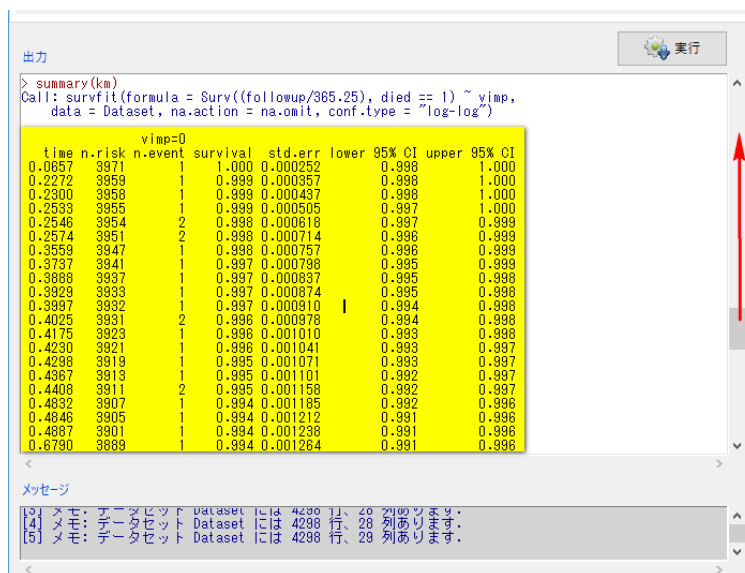
      N Observed Expected (O-E)^2/E (O-E)^2/V
vimp=0 3971      97    128.9      7.06      96
vimp=1  327      40     10.1     88.92      96

Chisq= 96 on 1 degrees of freedom, p= <2e-16 |

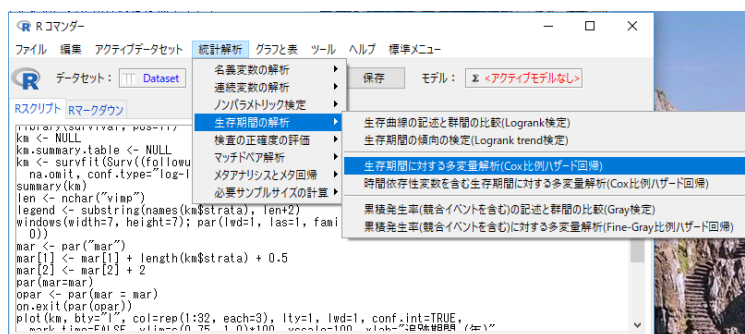
> km.summary.table <- summary.km(survfit=km, survdiff=res)
> km.summary.table
      サンプル数 生存期間中央値 95%信頼区間      P値
vimp=0      3971             NA      NA-NA 1.16e-22
vimp=1       327             NA      NA-NA

```

## 5. 生存曲線を作成するための両群の生存率表も確認できる:



## 4.0.3 Cox比例ハザードモデルを作る。



## 2. followupを時間へ

生存期間に対する多変量解析(Cox比例ハザード回帰)

モデル名を入力: CoxModel.1

変数 (ダブルクリックして式に入れる)

ethnic (因子)  
exit  
followup  
height  
id  
map  
mfgrp  
mfperm

モデル式: + \* : / %in% - ^ ( )

時間 followup, イベント died ~ 説明変数

層別化因子は + strata(####)と入力する ↑ ↑ ↑

☐ 3レベル以上の因子についてその因子全体のP値の計算(Wald検定)

☐ 比例ハザード性の分析を行う。

☐ マルチンゲール残差をプロットする

☐ ベースラインの生存曲線を示す。

☐ モデル解析用に解析結果をアクティブモデルとして残す

☐ AICを用いたステップワイズの変数選択を行う。

☐ BICを用いたステップワイズの変数選択を行う。

☐ P値を用いたステップワイズの変数選択(減少法)を行う。

↓ 一部のサンプルだけを解析対象にする場合の条件式。例: age>50 & Sex==0 や age<50 | Sex==1

<全ての有効なケース>

ヘルプ リセット OK キャンセル 適用

## 3. diedをイベントへ

生存期間に対する多変量解析(Cox比例ハザード回帰)

モデル名を入力: CoxModel.1

変数 (ダブルクリックして式に入れる)

bmi  
bmigrp  
compound.size  
diastolic  
died  
distnet (因子)  
education (因子)  
enter

モデル式: + \* : / %in% - ^ ( )

時間 followup, イベント died ~ 説明変数

層別化因子は + strata(####)と入力する ↑ ↑ ↑

☐ 3レベル以上の因子についてその因子全体のP値の計算(Wald検定)

☐ 比例ハザード性の分析を行う。

☐ マルチンゲール残差をプロットする

☐ ベースラインの生存曲線を示す。

☐ モデル解析用に解析結果をアクティブモデルとして残す

☐ AICを用いたステップワイズの変数選択を行う。

☐ BICを用いたステップワイズの変数選択を行う。

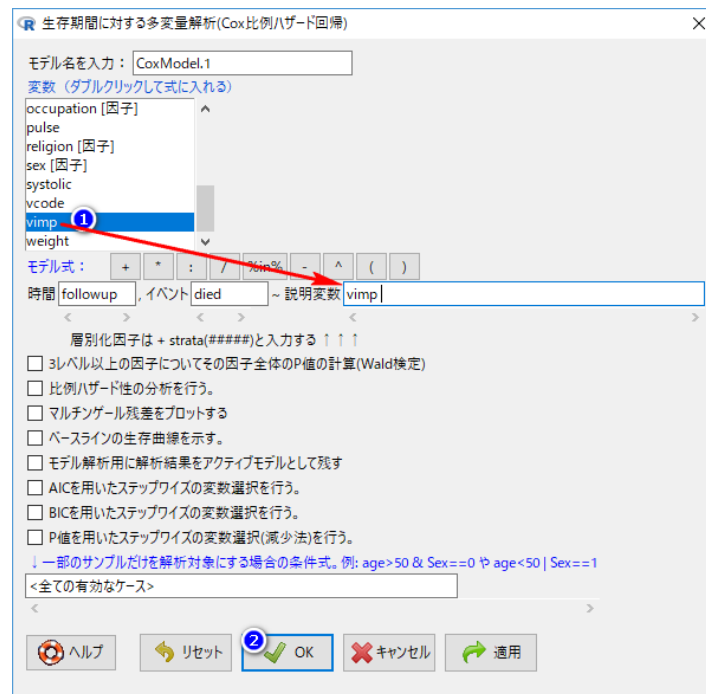
☐ P値を用いたステップワイズの変数選択(減少法)を行う。

↓ 一部のサンプルだけを解析対象にする場合の条件式。例: age>50 & Sex==0 や age<50 | Sex==1

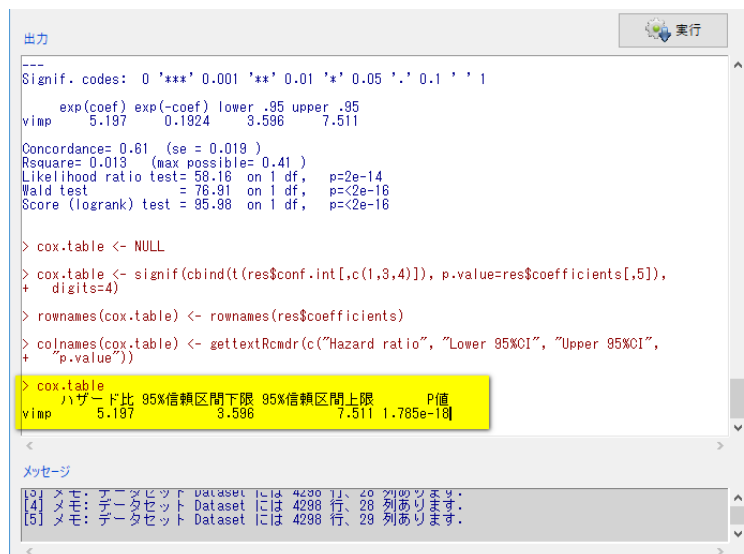
<全ての有効なケース>

ヘルプ リセット OK キャンセル 適用

## 4. vimp (視覚障害)を説明変数へ,OKをクリックする



## 5. 単変量ハザード比,及び信頼区間の意味を説明せよ.



## 6. 年齢調整ハザード比を求めよ.

生存期間に対する多変量解析(Cox比例ハザード回帰)

モデル名を入力:

変数 (ダブルクリックして式に入れる)

age  
agebin  
agegrp [因子] ①  
area  
bmi  
bmigrp  
compound.size  
diastolic

モデル式:

時間:  イベント:  ~ 説明変数

層別化因子は + strata(####)と入力する ↑↑↑

☐ 3レベル以上の因子についてその因子全体のP値の計算(Wald検定)

☐ 比例ハザード性の分析を行う。

☐ マルチンゲール残差をプロットする

☐ ベースラインの生存曲線を示す。

☐ モデル解析用に解析結果をアクティブモデルとして残す

☐ AICを用いたステップワイズの変数選択を行う。

☐ BICを用いたステップワイズの変数選択を行う。

☐ P値を用いたステップワイズの変数選択(減少法)を行う。

↓一部のサンプルだけを解析対象にする場合の条件式。例: age>50 & Sex==0 や age<50 | Sex==1

<全ての有効なケース>

ヘルプ リセット OK キャンセル 適用

## 7. 年齢調整ハザード比, 及び信頼区間の意味を説明せよ.

出力

	2.298	0.4351	1.456	3.627
agegrp [1.35-54]	2.298	0.4351	1.456	3.627
agegrp [1.35-64]	5.134	0.1948	2.971	8.872
agegrp [1.65+]	8.810	0.1135	5.035	15.413

Concordance= 0.727 (se = 0.023 )  
 Rsquare= 0.027 (max possible= 0.41 )  
 Likelihood ratio test= 118.8 on 4 df, p=<2e-16  
 Wald test = 137 on 4 df, p=<2e-16  
 Score (logrank) test = 194 on 4 df, p=<2e-16

```

> cox.table <- NULL
> cox.table <- signif(cbind(res$conf.int[,c(1,3,4)], res$coefficients[,5]), digits=4)
> cox.table <- data.frame(cox.table)
> colnames(cox.table) <- gettextRcmdr(c("Hazard ratio", "Lower 95%CI", "Upper 95%CI",
+ "p.value"))
> cox.table

```

	ハザード比	95%信頼区間下限	95%信頼区間上限	P値
vimp	2.098	1.370	3.212	6.531e-04
agegrp [1.35-54]	2.298	1.456	3.627	3.497e-04
agegrp [1.35-64]	5.134	2.971	8.872	4.612e-09
agegrp [1.65+]	8.810	5.035	15.410	2.462e-14

メッセージ

[0] メモ: データセット Dataset には 4298 行、28 列あります。  
 [4] メモ: データセット Dataset には 4298 行、28 列あります。  
 [5] メモ: データセット Dataset には 4298 行、28 列あります。

5 参考図書:

1. 中澤 港,「Rによる保健医療データ解析演習」, (<http://minato.sip21c.org/msb/medstatbookx.pdf>)
2. 新谷 歩,「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!」.