

学籍番号:

公衆衛生学

疫学演習 2019-6-5 & 2019-6-12

氏名:

## Contents

1 問題1:両群間計量データの平均値を比較する	-1-
2 問題2:線形回帰モデル	-7-
3 問題3: $\chi^2$ 検定,オッズ比,ロジスティック回帰モデル	-20-
4 問題4:生存分析	-33-
5 参考図書:	-44-

### 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

- 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.

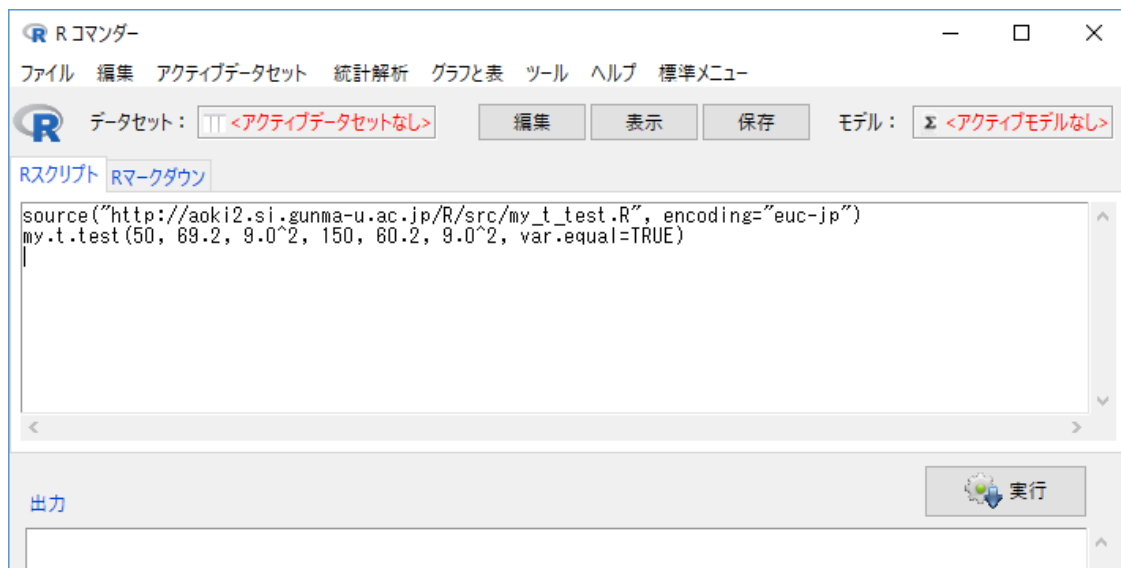
また,EZR で t 値,自由度 (degree of freedom)を使って P 値を計算する時,以下のコマンドを利用してください:

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

## 1.1 答え

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)
```



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.

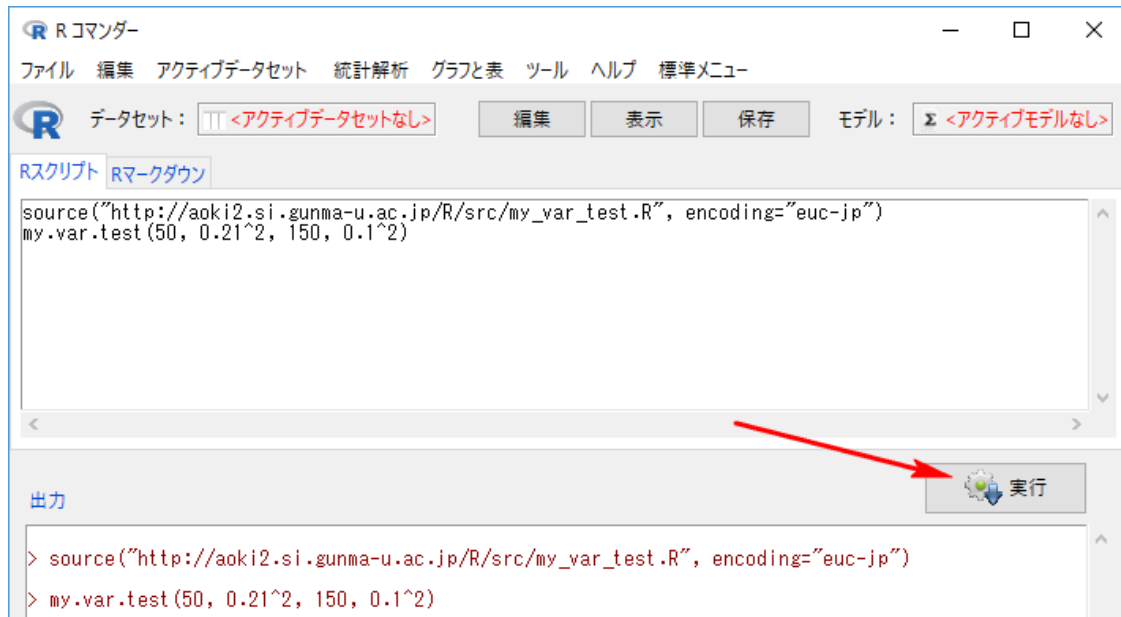
なお,EZR で F 値,両群の分散,両群それぞれの自由度 (df) を使って P 値を計算する時に,以下のコマンドを利用してください:

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

## 1.2 答え

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)
```



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:

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

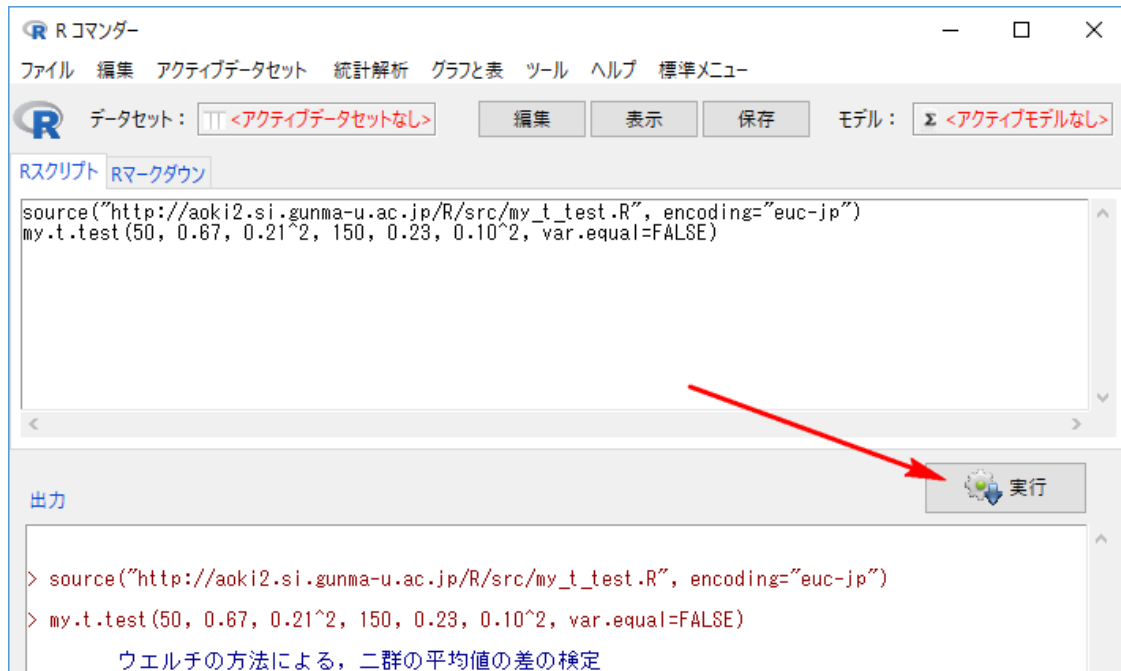
また,EZR で t 値,自由度 (df)を使って P 値を計算する時,以下のコマンドを利用してください:

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

### 1.3 答え

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)
```



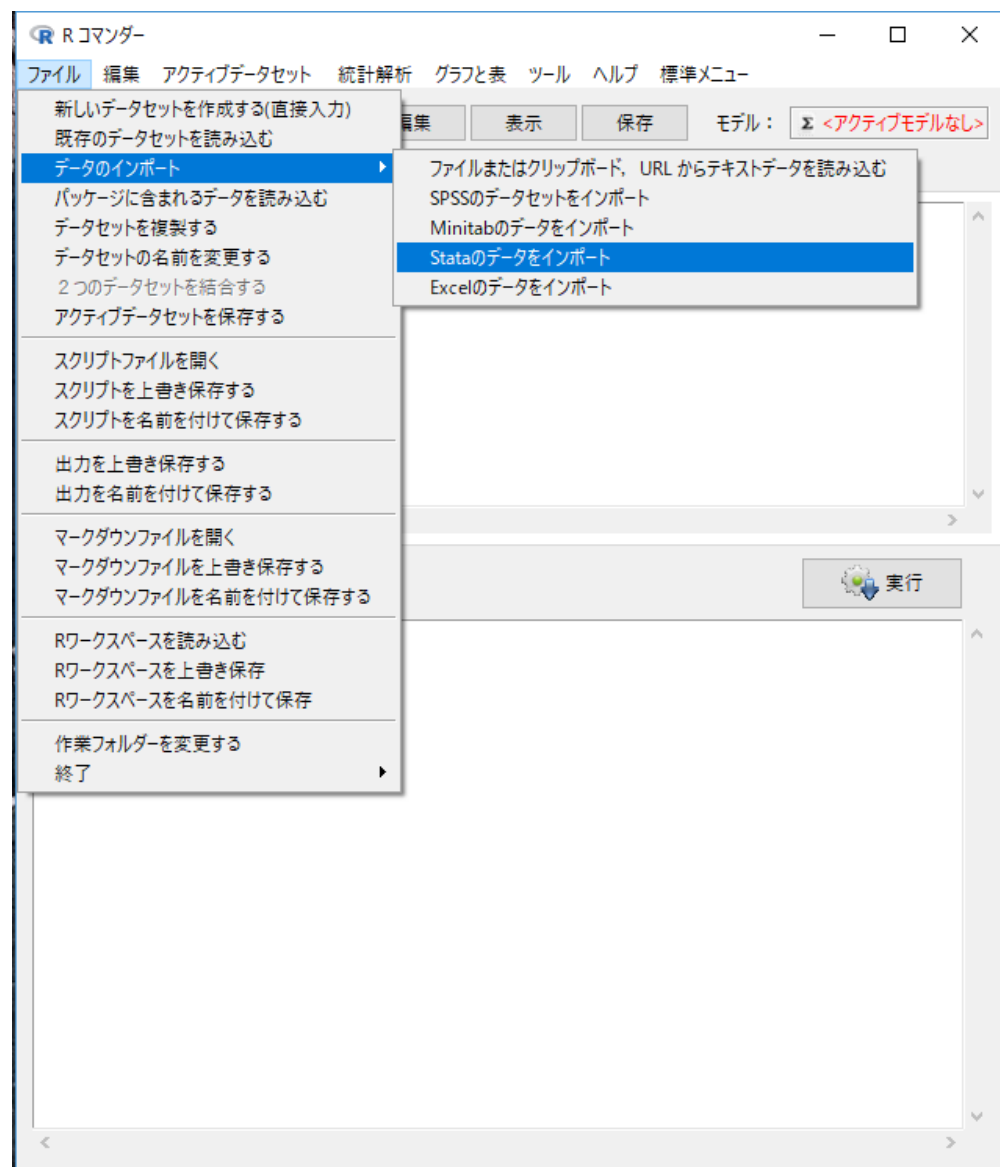
## 2 問題2:線形回帰モデル

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

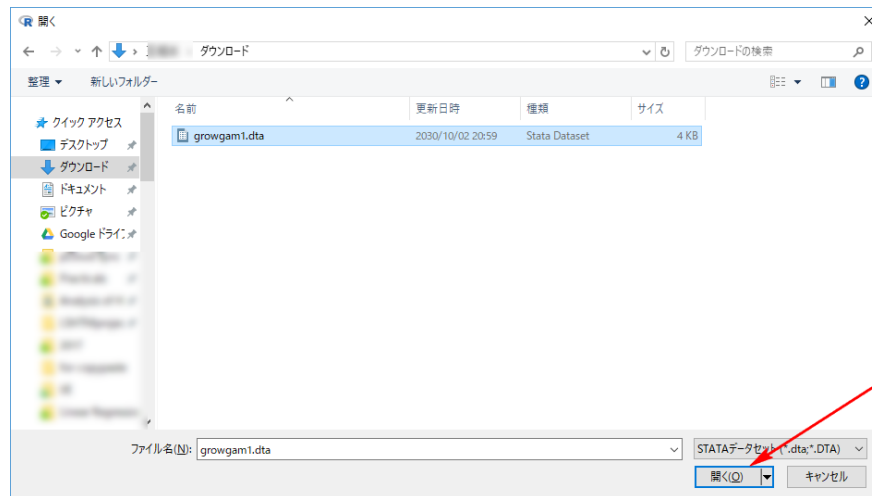




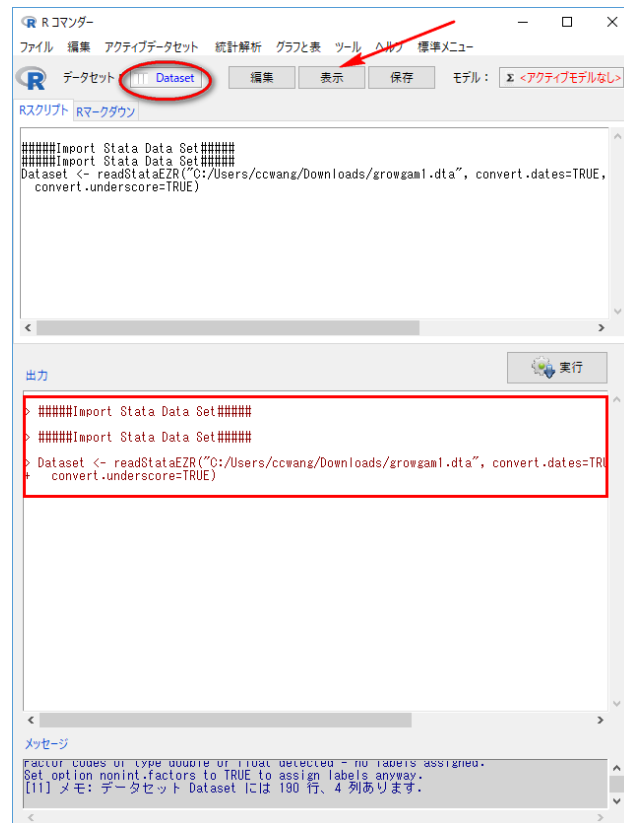
## 2.1.2 ステップ2



## 2.1.3 ステップ3



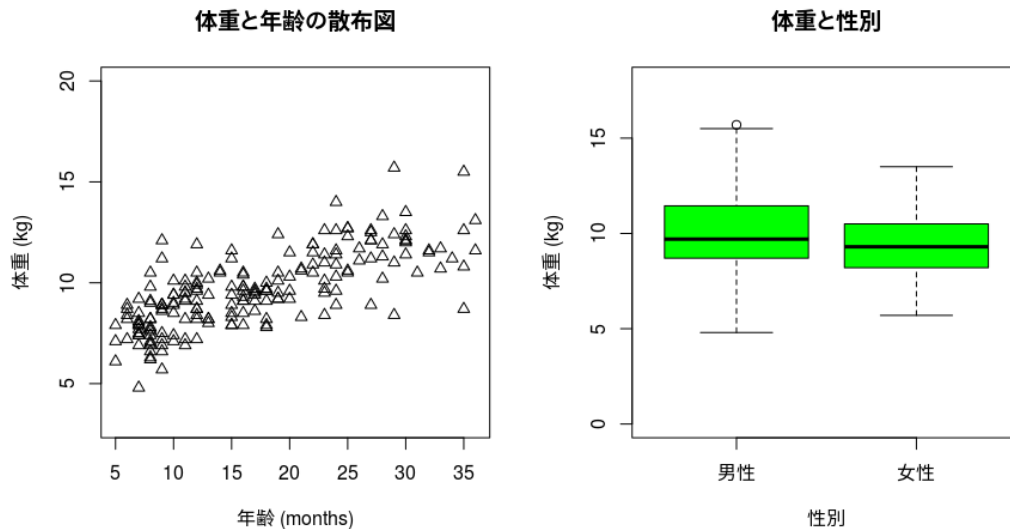
## 2.1.4 ステップ4



## 2.1.5 ステップ5

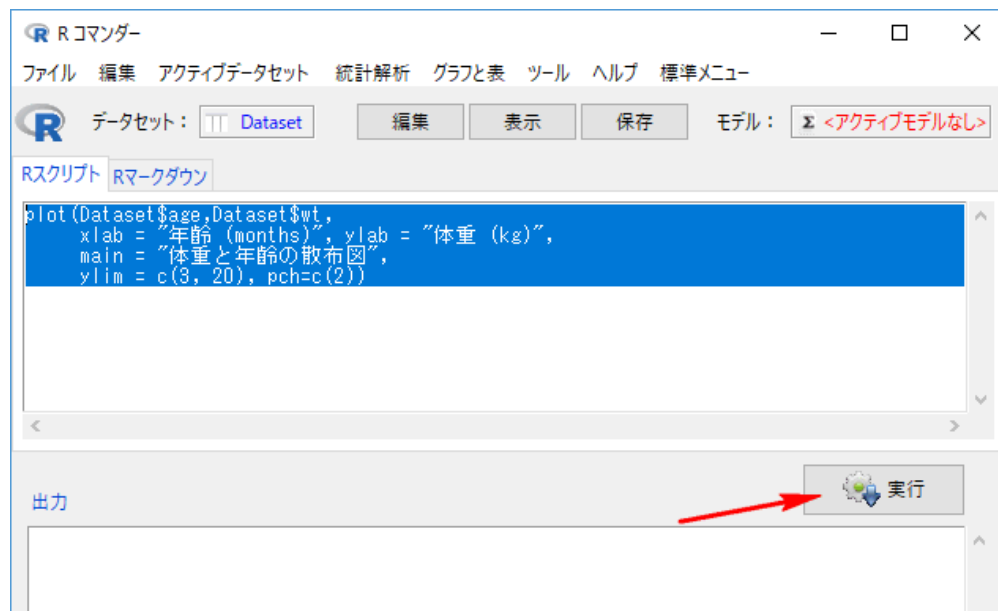
	sex	age	wt	len
1	2	23	8.4	73.2
2	2	22	10.9	84.4
3	2	6	7.2	68.7
4	1	24	10.3	83.7
5	1	14	10.5	79.2
6	2	18	9.6	75.8
7	2	30	11.4	84.4
8	1	24	11.4	84.8
9	1	17	9.4	74.3
10	2	27	12.5	82.6
11	2	16	9.1	74.1
12	1	30	12.0	86.4
13	1	18	7.8	71.9
14	2	15	8.5	77.6
15	2	13	8.0	72.2
16	1	11	9.1	72.4
17	2	8	10.5	71.2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



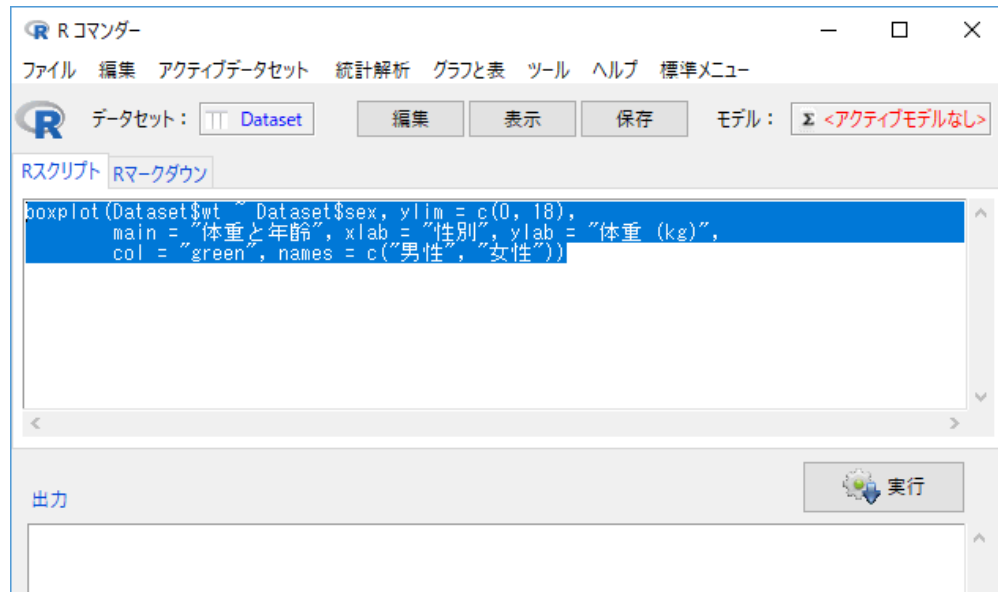
上記左のグラフを描くため,以下のコードをRスクリプトに入力して,実行をクリックしてください。

```
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 `cov(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$age)  
# 年齢の分散  
var(Dataset$age)  
# 体重の平均値  
mean(Dataset$wt)  
# 体重の分散  
var(Dataset$wt)  
# 体重と年齢の共分散 covariance  
cov(Dataset$wt, Dataset$age)
```

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:

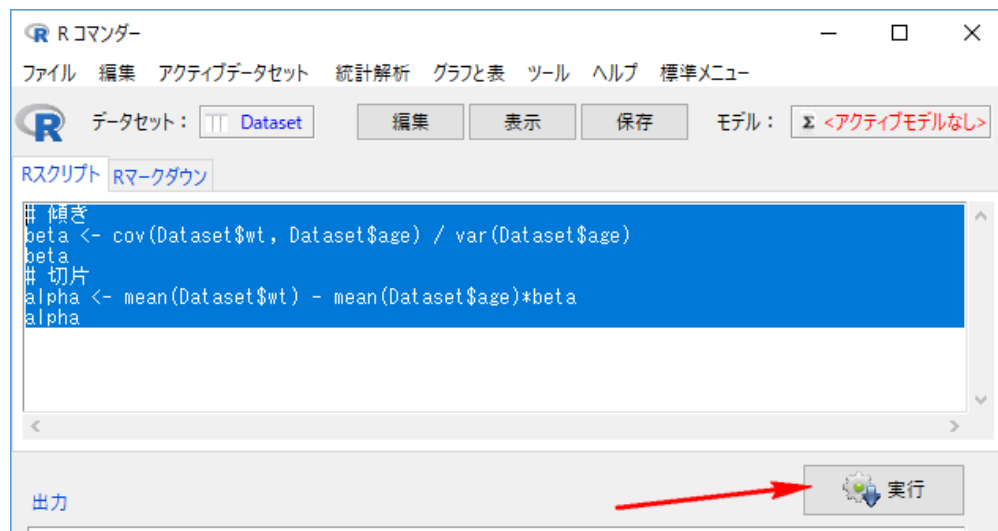
$$\begin{aligned}\text{Var}(X) &= \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n - 1} \\ &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}\end{aligned}$$

- 共分散 covariance:

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

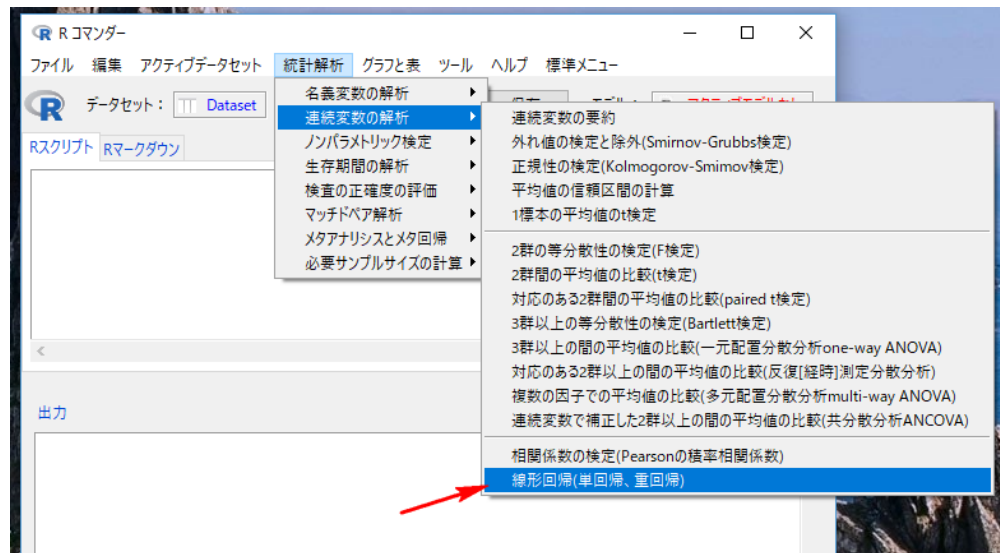
以下のコードをRスクリプトに入力して, 実行をクリックしてください. (結果を下の余白に記入すること)

```
# 傾き (slope)
beta <- cov(Dataset$wt, Dataset$age) / var(Dataset$age)
beta
# 切片 (intercept)
alpha <- mean(Dataset$wt) - mean(Dataset$age)*beta
alpha
```

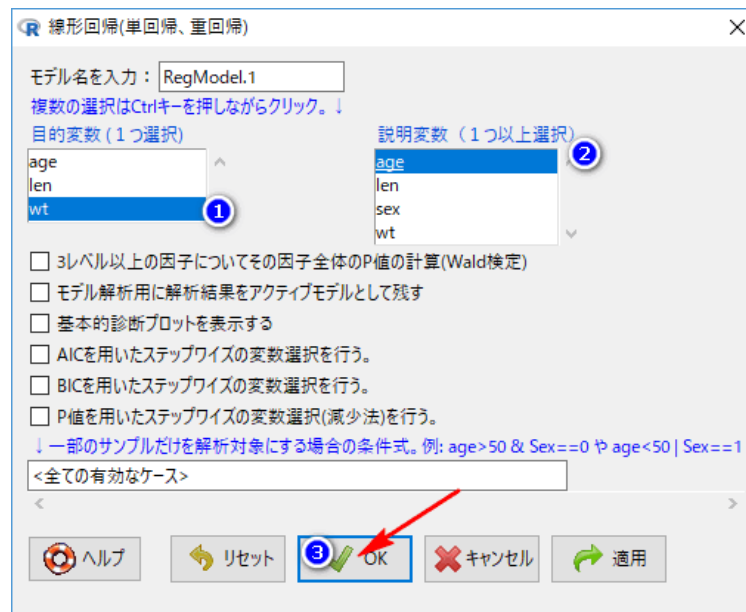


## 2.5 実際にEZRで線形モデルを作ってみよう:

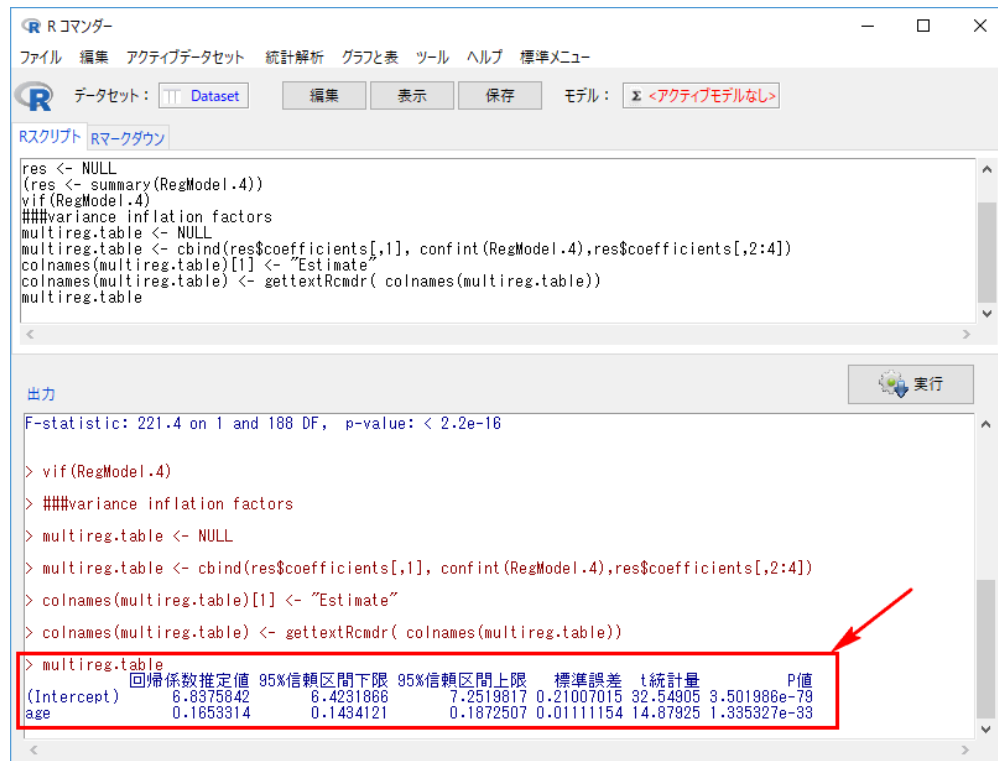
### 2.5.1 ステップ1



### 2.5.2 ステップ2



## 2.5.3 ステップ3



```

res <- NULL
(res <- summary(RegModel.4))
vif(RegModel.4)
###variance inflation factors
multireg.table <- NULL
multireg.table <- cbind(res$coefficients[,1], confint(RegModel.4), res$coefficients[,2:4])
colnames(multireg.table)[1] <- "Estimate"
colnames(multireg.table) <- gettextRcmdr( colnames(multireg.table))
multireg.table

```

出力

F-statistic: 221.4 on 1 and 188 DF, p-value: < 2.2e-16

```

> vif(RegModel.4)
> ###variance inflation factors
> multireg.table <- NULL
> multireg.table <- cbind(res$coefficients[,1], confint(RegModel.4), res$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)	6.8375842	6.4231866	7.2519817	0.21007015	32.54905	3.501986e-79
age	0.1653314	0.1434121	0.1872507	0.01111154	14.87925	1.335327e-33

推定された回帰係数自分の計算結果とは一致するかを確認してください。

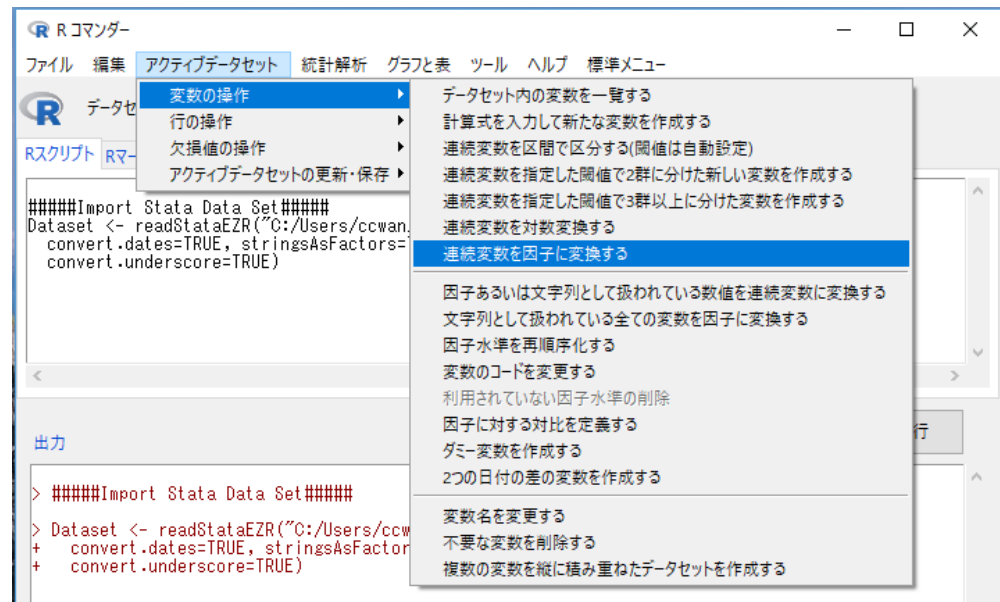
2.6 今まで計算した傾きと切片の数字を用いて、年齢と体重の関係を線形と考える場合の計算式を記入せよ。傾きと切片の計算結果の意味をそれぞれ記述せよ。

2.6.1 答え

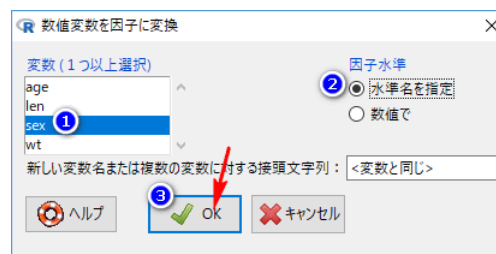
## 2.7 性別を説明変数に入れたモデルを作る

### 2.7.1 性別変数を因子 (factor) に変換する

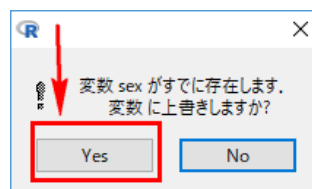
#### 2.7.1.1 ステップ1



#### 2.7.1.2 ステップ2

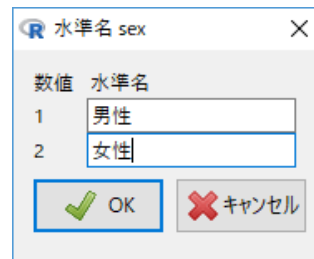


#### 2.7.1.3 ステップ3



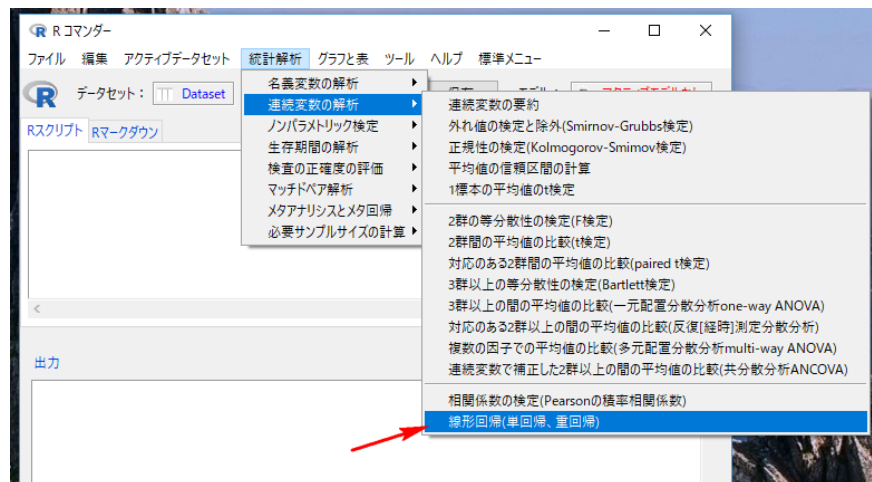


## 2.7.1.4 ステップ4-水準名に男性,女性を入力する

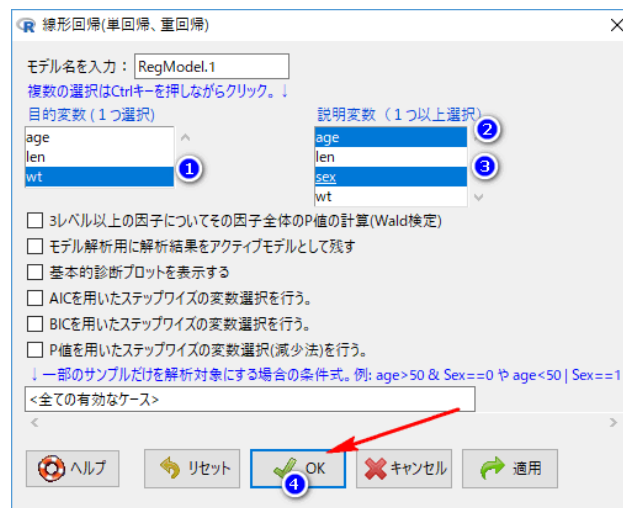


## 2.7.2 重回帰線形モデルを作る

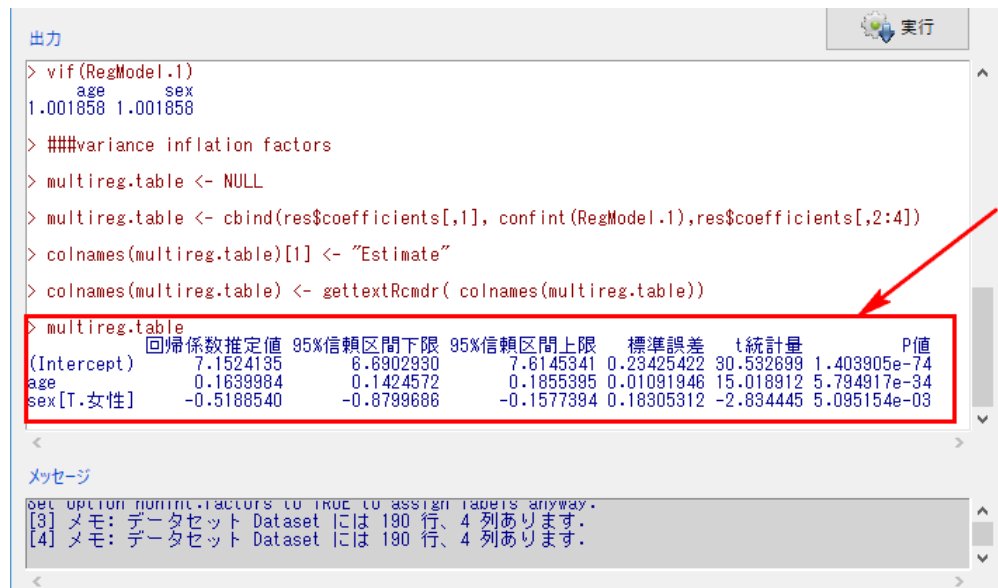
## 2.7.2.1 ステップ1



## 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



```

出力
> 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 nomint.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。

```

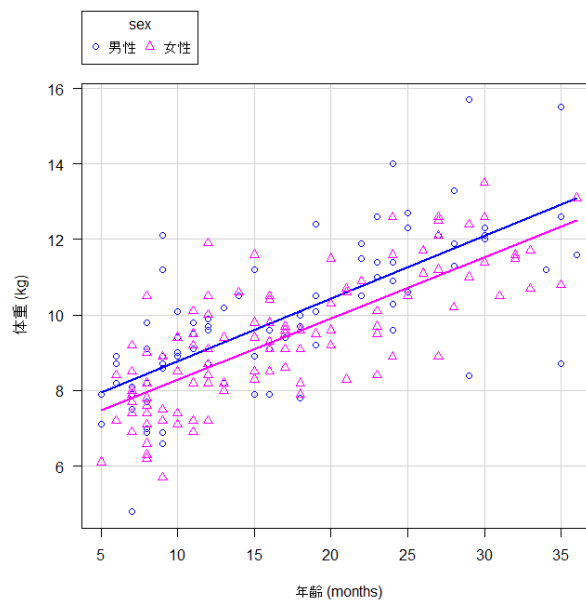
2.8 重回帰線形モデルの計算結果を用いて、体重の平均値を年齢と性別の線形モデルで表示せよ。各回帰係数の意味を説明せよ。

2.9 答え

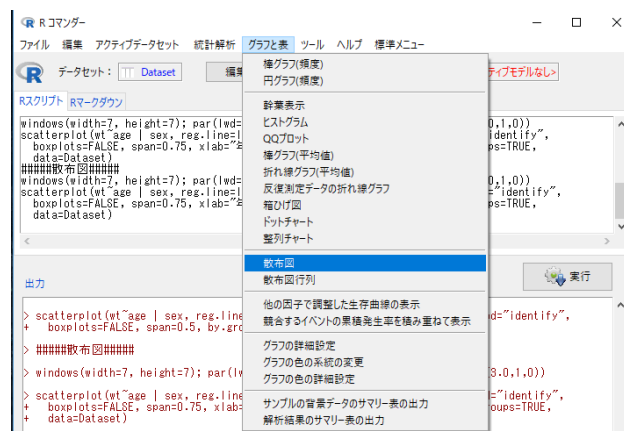
2.10 上記の重回帰線形モデルを用いて、年齢が34ヶ月の女の子の体重の予測値を計算せよ。

2.11 答え

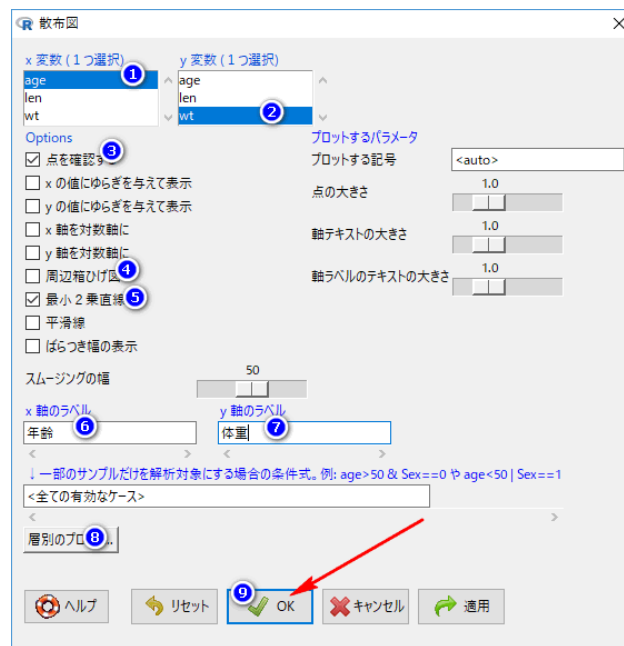
2.12 男女別の年齢と体重の散布図を描く



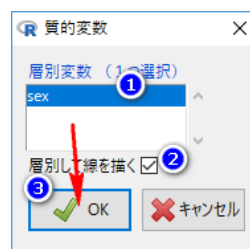
2.12.1 ステップ1



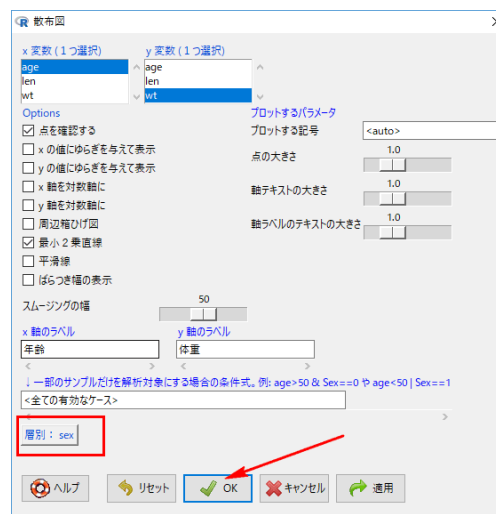
## 2.12.1.1 ステップ2



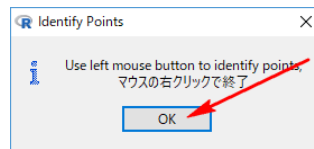
## 2.12.1.2 ステップ3



## 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, ロジスティック回帰モデル

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. 対象者は三年間観察され, 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			4161 (96.81%)
1			137 (3.19%)
合計	3971 (100%)	327 (100%)	4298 (100%)

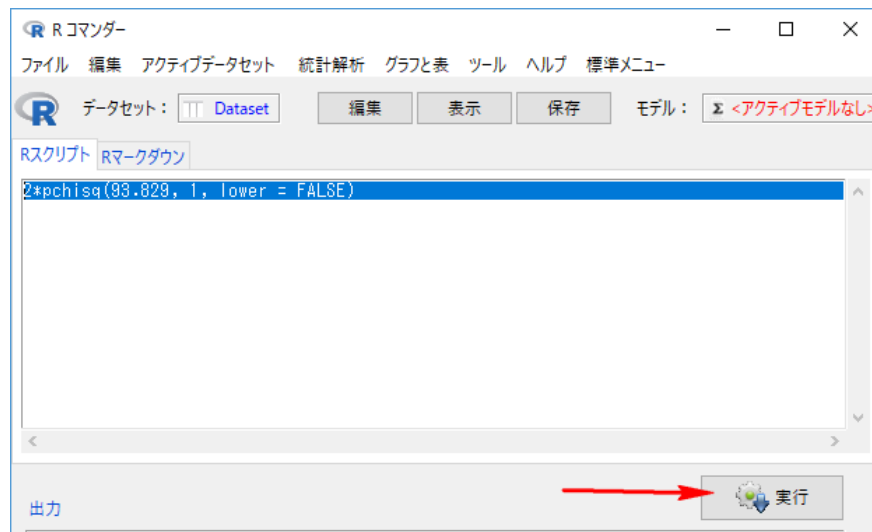
3.1.2 上記の2つの表の数字を使って  $\chi^2$  統計量を計算せよ

3.1.3 答え

3.1.4  $2 \times 2$  の分割表では,自由度は \_\_\_\_\_

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)**



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
a
```



3.1.5 視覚障害と死亡の関係を示すテーブルのデータを元に、下表を完成せよ:

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

では、視覚障害と死亡の関連を示すオッズ比を算出せよ:

OR =

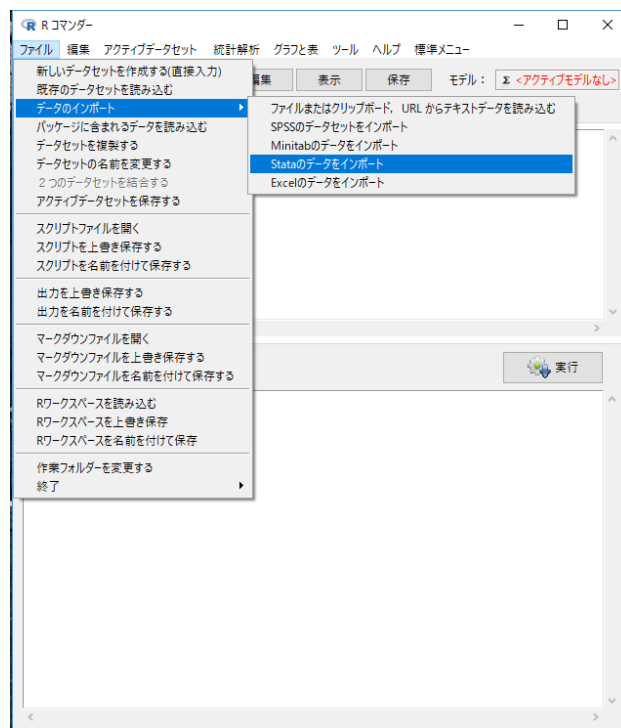
このオッズ比の対数を取った値  $\log(\text{OR})$  は:

$\log(\text{OR}) =$

3.1.6 EZRでロジスティック回帰モデルを作る

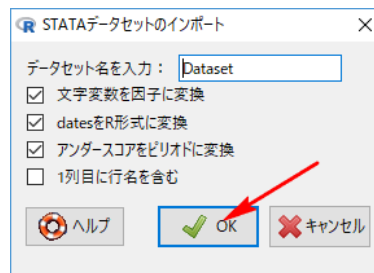
3.1.6.1 ステップ1 — データのインポート

1.

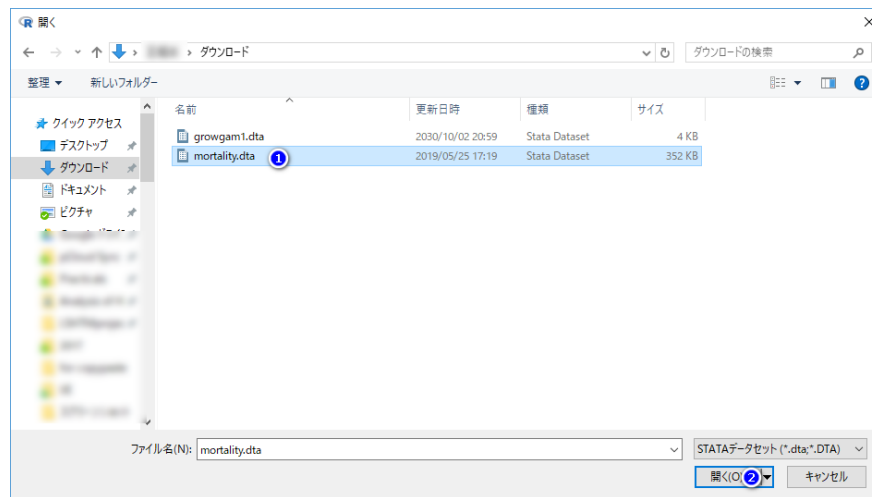




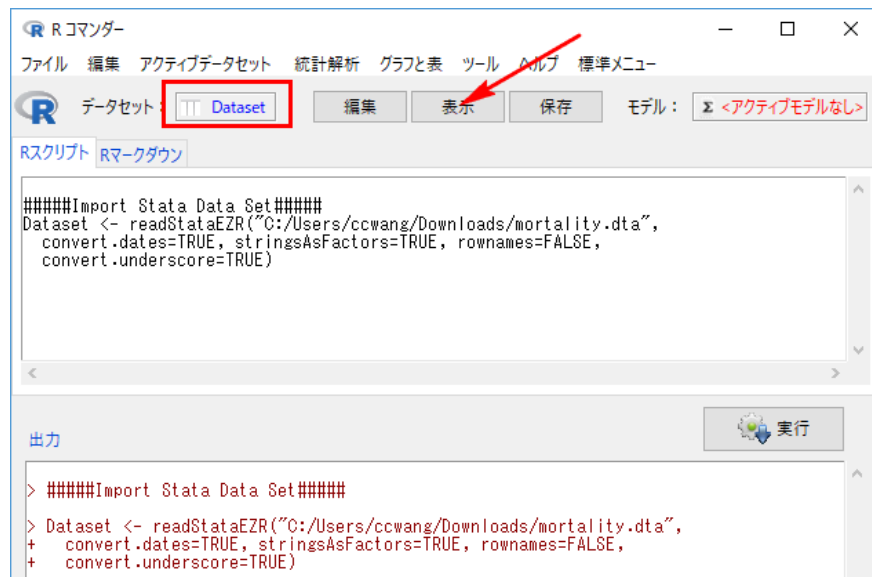
2.



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ーロジスティックモデルを作る

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。

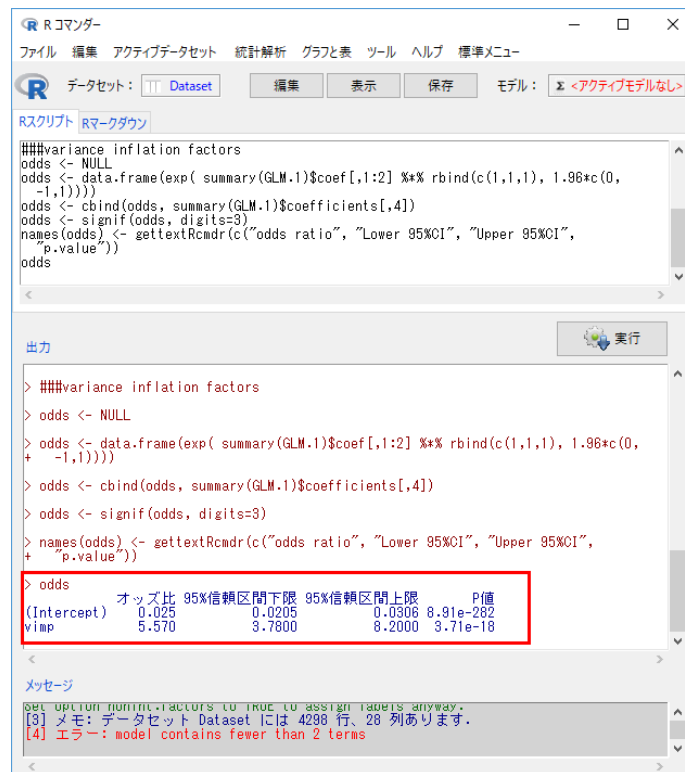
☐ P値を用いたステップワイズの変数選択(減少法)を行う。

↓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.



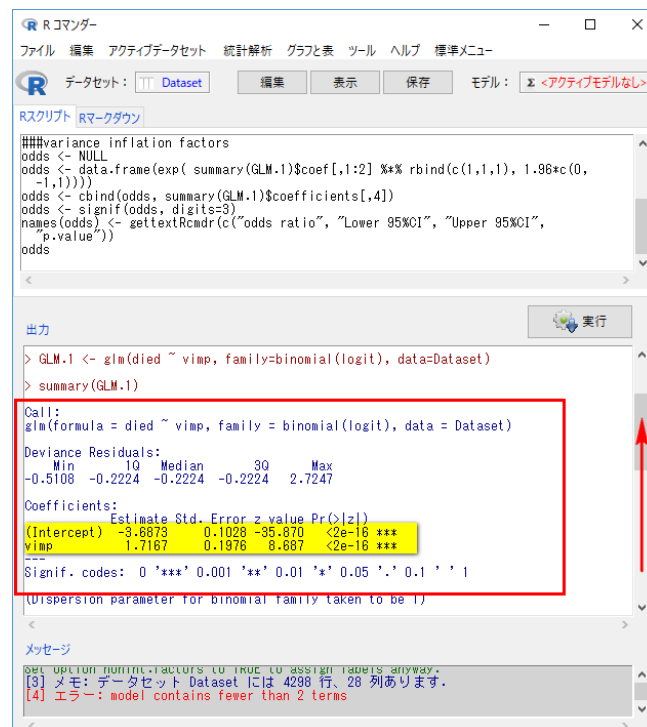
```
##variance inflation factors
odds <- NULL
odds <- data.frame(exp( summary(GLM.1)$coef[,1:2] %*% rbind(c(1,1,1), 1.96*c(0,
-1,1))))
odds <- cbind(odds, summary(GLM.1)$coefficients[,4])
odds <- signif(odds, digits=3)
names(odds) <- gettextRcmdr(c("odds ratio", "Lower 95%CI", "Upper 95%CI",
"p.value"))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

	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  
[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



```
> 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)	-3.8873	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  
[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										
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:

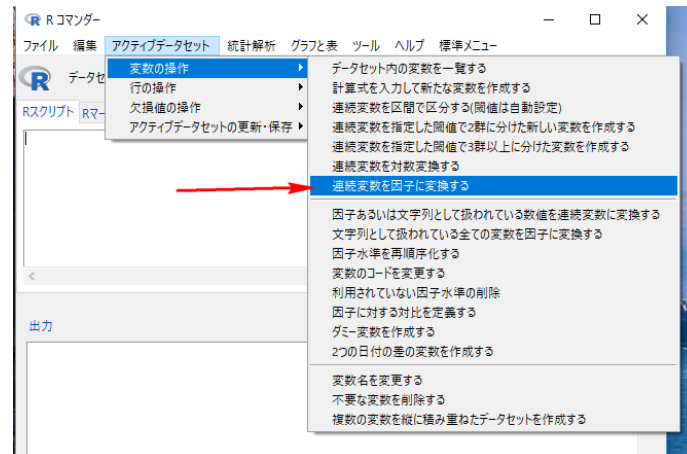
	オッズ		
年齢	視力正常	視覚障害	オッズ比
15-34	$29/2301 = 0.01260$		
35-54	$0.02990$		
55-64	$0.07075$		
65+	$0.16667$		

各年齢層では視覚障害と死亡との関連はどう変化しているか?

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

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

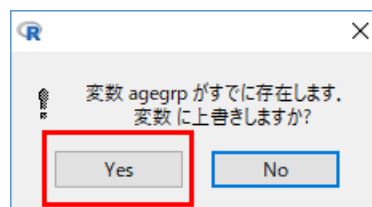
1.



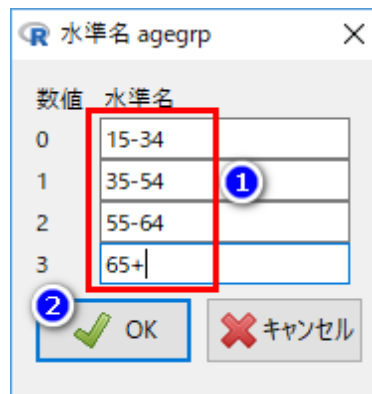
2.



3.

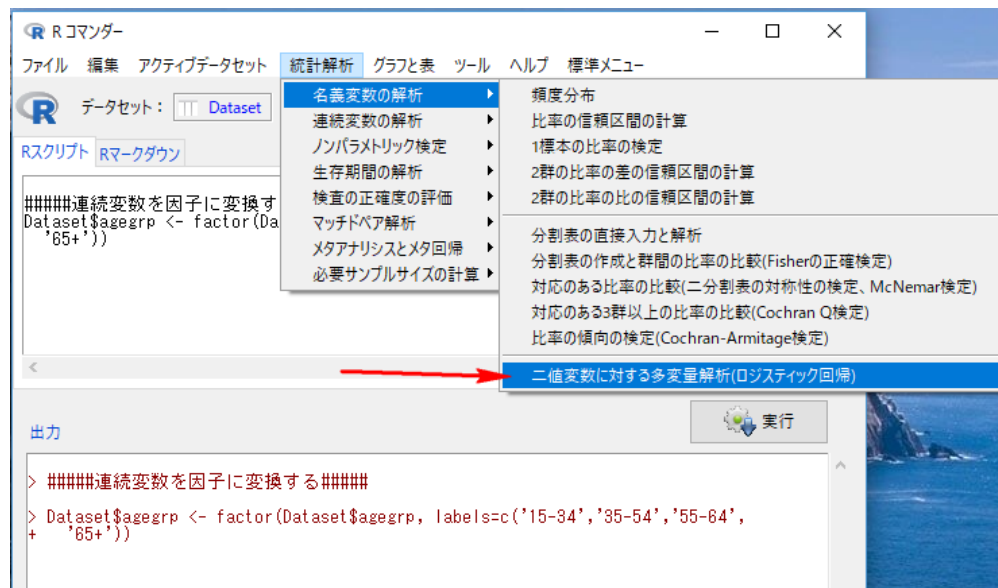


## 4. 水準名に各年齢グループの名前を入力する

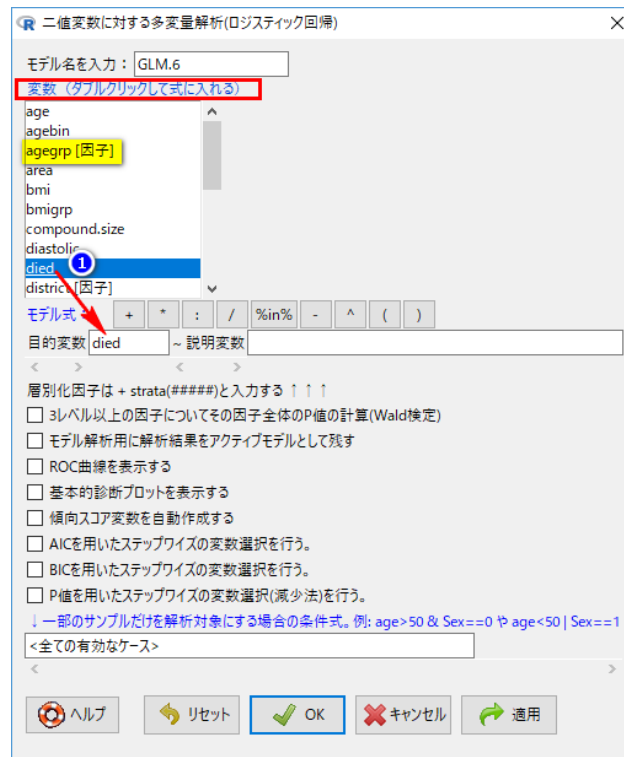


## 3.2.2.2 多変量ロジスティック回帰モデルを作る

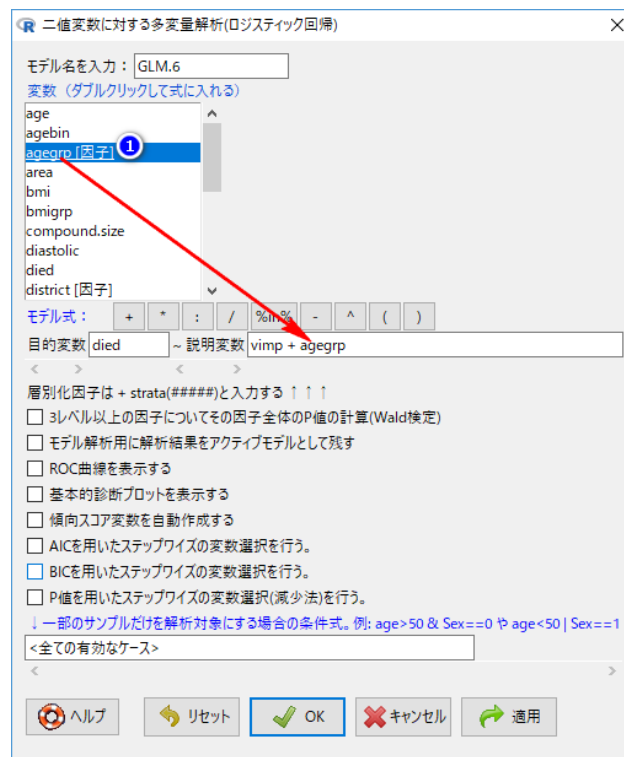
## 1.



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 も説明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

```

出力

```

> 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+]	9.9000	5.54000	17.7000	9.84e-15

メッセージ

```

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

```

3.2.2.3 単変量ロジスティック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

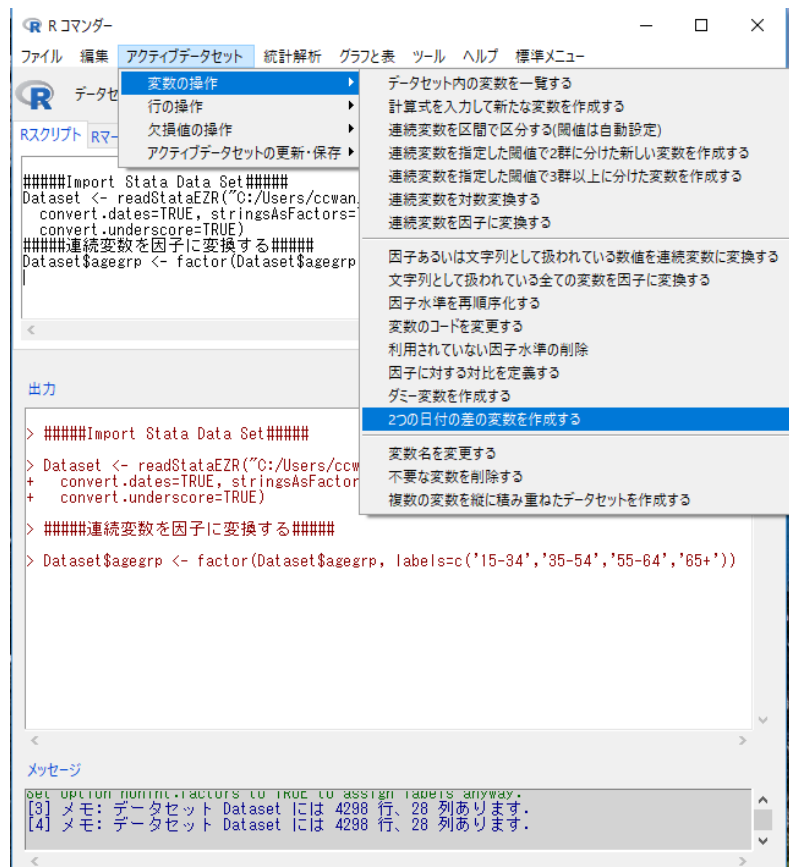
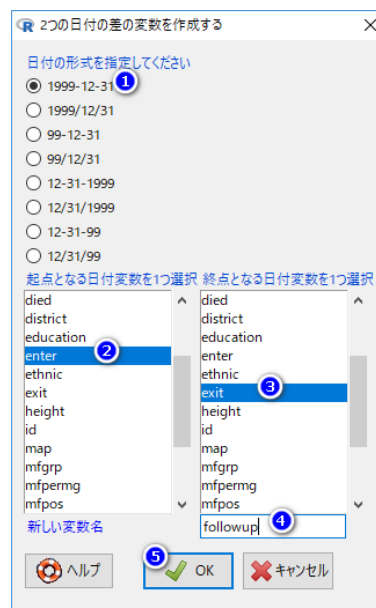
## 4 問題4:生存分析

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:

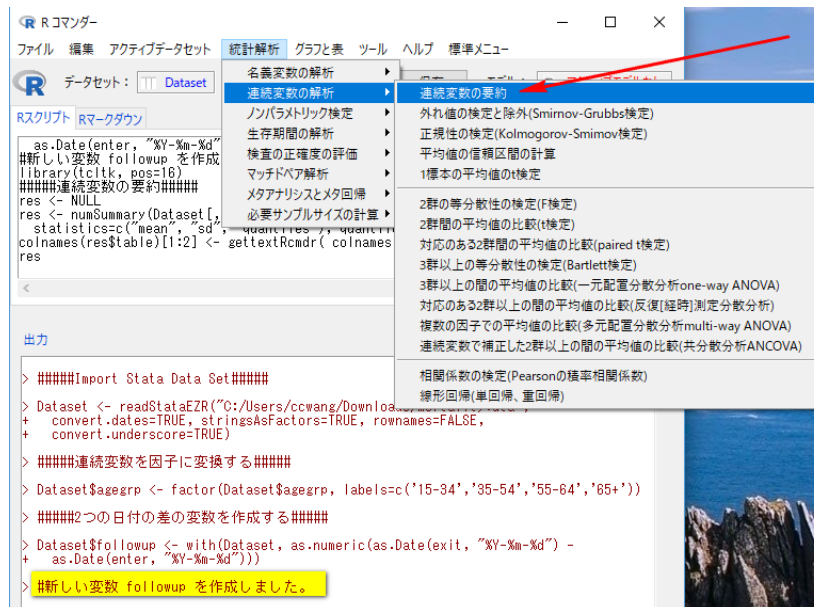
```
## # A tibble: 4,298 x 6
##       id agegrp enter      exit      vimp died
##   <dbl> <fct> <date>    <date>    <dbl+lbl> <dbl>
## 1     1  1 55-64 1989-05-09 1992-02-05 0 [Normal]      0
## 2     2  2 35-54 1989-05-09 1992-02-05 0 [Normal]      0
## 3     3  3 35-54 1989-05-09 1992-02-05 1 [Visually 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

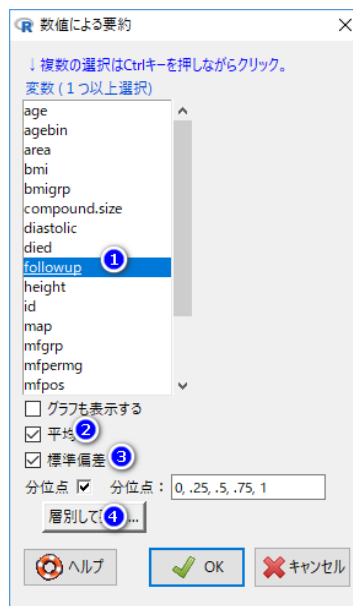
1.

2. 追跡期間の変数名を `followup` とする

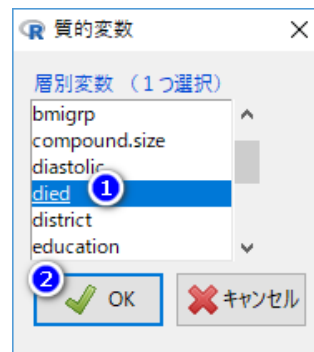
## 3. followup を作成したと確認メッセージが出る. 追跡期間の要約を調べる:



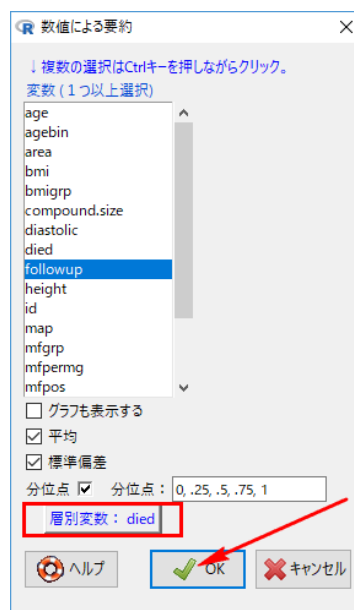
## 4. イベント別(死亡,打ち切り)により追跡期間を計算する



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?

```

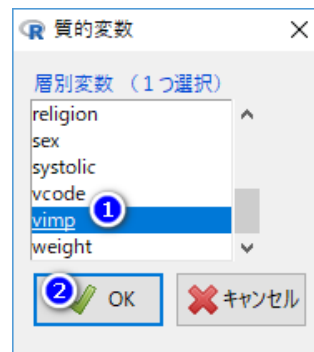
出力
> colnames(res$table)[1:2] <- gettextRcmdr( colnames(res$table)[1:2])
> res
  平均 標準偏差   0%  25%   50%   75%  100% data:n
0  977.2749 179.7676 24.0 909 1001 1088.5 1187 3971
1  929.7890 213.5089 22.5 896  987 1030.5 1158  327

#####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  991.5  137

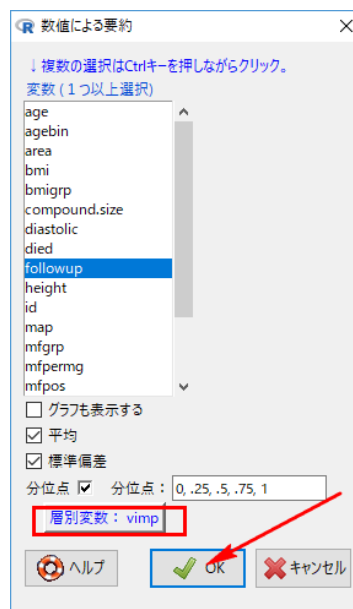
メッセージ
[5] メモ: データセット Dataset には 4298 行、28 列の Summary
[4] メモ: データセット Dataset には 4298 行、28 列あります。
[5] メモ: データセット Dataset には 4298 行、28 列あります。

```

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？

```

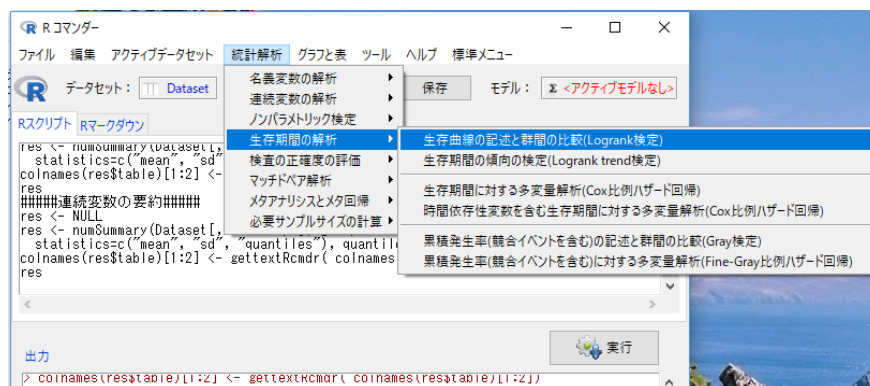
出力
+ as.Date(enter, "%Y-%m-%d"))
> ##新しい変数 followup を作成しました。
> library(tcltk, pos=16)
> #####連続変数の要約#####
> res <- NULL
> res <- numSummary(Dataset[, "followup"], groups=Dataset$vimp,
+   statistics=c("mean", "sd", "quantiles"), quantiles=c(0,.25,.5,.75,1))
> colnames(res$table)[1:2] <- gettextRcmdr( colnames(res$table)[1:2])
> res
  平均 標準偏差   0% 25%   50%   75% 100% data:n
0 977.2749 179.7876 24.0 909 1001 1089.5 1187 39711
1 829.7890 213.5093 22.5 896  997 1030.5 1158  327

メッセージ
[3] 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。

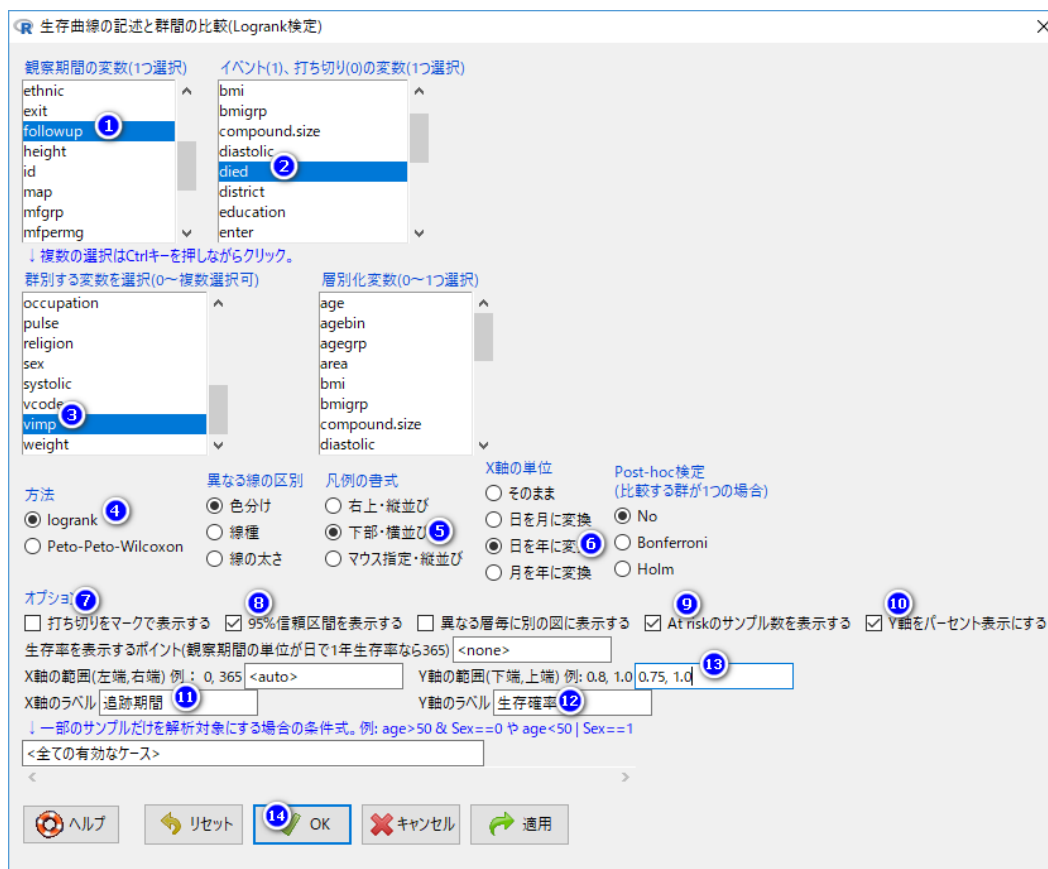
```

## 4.0.2 生存表と Kaplan-Meier グラフを作成する

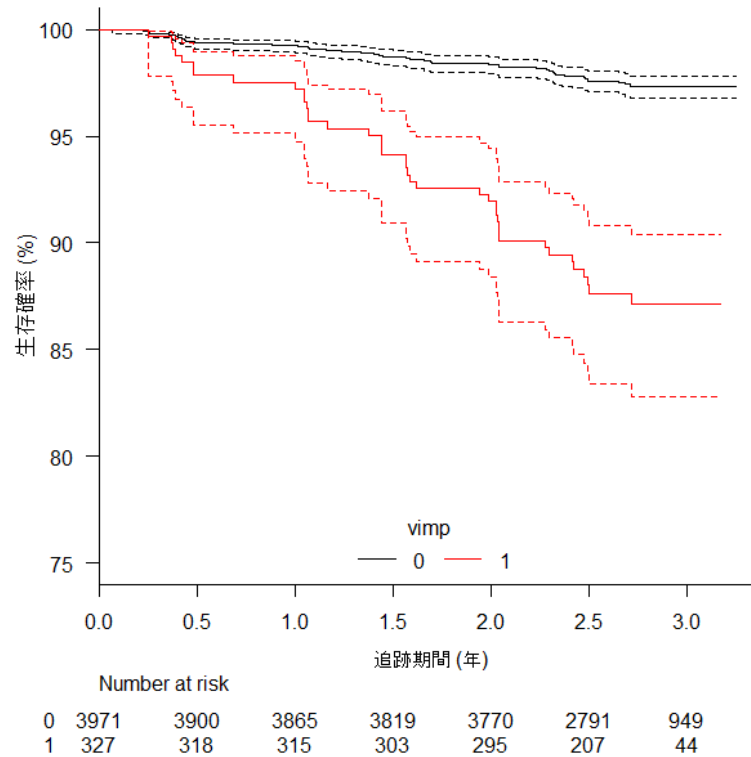
1.



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:

出力

実行

```
> summary(km)
Call: survfit(formula = Surv(followup/365.25, died == 1) ~ vimp,
data = Dataset, na.action = na.omit, conf.type = "log-log")
```

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
0.0657	3971	1	1.000	0.000252	0.998	1.000
0.2272	3959	1	0.999	0.000357	0.998	1.000
0.2300	3958	1	0.999	0.000437	0.998	1.000
0.2533	3955	1	0.999	0.000505	0.997	1.000
0.2546	3954	2	0.998	0.000618	0.997	0.999
0.2574	3951	2	0.998	0.000714	0.996	0.999
0.3559	3947	1	0.998	0.000757	0.996	0.999
0.3737	3941	1	0.997	0.000796	0.995	0.999
0.3888	3937	1	0.997	0.000837	0.995	0.998
0.3929	3933	1	0.997	0.000874	0.995	0.998
0.3997	3932	1	0.997	0.000910	0.994	0.998
0.4025	3931	2	0.996	0.000978	0.994	0.998
0.4175	3923	1	0.996	0.001010	0.993	0.998
0.4230	3921	1	0.996	0.001041	0.993	0.997
0.4298	3919	1	0.995	0.001071	0.993	0.997
0.4367	3913	1	0.995	0.001101	0.992	0.997
0.4408	3911	2	0.995	0.001158	0.992	0.997
0.4832	3907	1	0.994	0.001185	0.992	0.996
0.4846	3905	1	0.994	0.001212	0.991	0.996
0.4887	3901	1	0.994	0.001236	0.991	0.996
0.6790	3689	1	0.994	0.001264	0.991	0.996

メッセージ

[4] メモ: データセット Dataset には 4298 行、28 列があります。  
[5] メモ: データセット Dataset には 4298 行、29 列があります。

出力

実行

```
> summary(km)
Call: survfit(formula = Surv(followup/365.25, died == 1) ~ vimp,
data = Dataset, na.action = na.omit, conf.type = "log-log")
```

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
0.253	325	1	0.997	0.00307	0.978	1.000
0.372	324	1	0.994	0.00434	0.976	0.998
0.379	323	1	0.991	0.00530	0.972	0.997
0.392	322	1	0.988	0.00612	0.968	0.995
0.424	321	1	0.985	0.00683	0.963	0.994
0.478	320	1	0.982	0.00747	0.959	0.992
0.485	319	1	0.978	0.00805	0.955	0.990
0.683	316	1	0.975	0.00860	0.951	0.988
1.003	315	1	0.972	0.00911	0.947	0.985
1.047	314	1	0.969	0.00960	0.943	0.983
1.050	313	1	0.966	0.01005	0.940	0.981
1.061	312	1	0.963	0.01049	0.936	0.979
1.068	311	2	0.957	0.01130	0.928	0.974
1.165	309	1	0.954	0.01168	0.924	0.972
1.377	307	1	0.951	0.01205	0.921	0.969
1.439	306	1	0.947	0.01240	0.917	0.967
1.441	305	2	0.941	0.01307	0.909	0.962
1.563	303	1	0.938	0.01339	0.906	0.960
1.569	302	1	0.935	0.01371	0.902	0.957
1.574	301	1	0.932	0.01401	0.898	0.955
1.586	299	1	0.929	0.01430	0.895	0.952
1.619	298	1	0.926	0.01459	0.891	0.950
1.841	297	1	0.923	0.01487	0.888	0.947
1.988	296	1	0.919	0.01514	0.884	0.944

メッセージ

[4] メモ: データセット Dataset には 4298 行、28 列があります。  
[5] メモ: データセット Dataset には 4298 行、29 列があります。

## 4.0.3 Cox比例ハザードモデルを作る。

R コマンド

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

```
library(survival, pos=1)
km <- NULL
km.summary.table <- NULL
km <- survfit(Surv(followup/365.25, died == 1) ~ vimp,
na.omit, conf.type="log-log")
summary(km)
len <- nchar("vimp")
legend <- substring(names(km$strata), len+2)
windows(width=7, height=7, par(lwd=1, las=1, fam=
))
mar <- par("mar")
mar[1] <- mar[1] + length(km$strata) + 0.5
mar[2] <- mar[2] + 2
par(mar=mar)
opar <- par(mar = mar)
on.exit(par(opar))
plot(km, bty="n", col=rep(1:32, each=3), lty=1, lwd=1, conf.int=TRUE,
main="Kaplan-Meier survival plot")
```

統計解析

- 名義変数の解析
- 連続変数の解析
- ノパラメトリック検定
- 生存期間の解析
- 検査の正確度の評価
- マッチペア解析
- メタアナリシスとメタ回帰
- 必要サンプルサイズの計算
- 生存曲線の記述と群間の比較(Logrank検定)
- 生存期間の傾向の検定(Logrank trend検定)
- 生存期間に対する多変量解析(Cox比例ハザード回帰)
- 時間依存性変数を含む生存期間に対する多変量解析(Cox比例ハザード回帰)
- 累積発生率(競合イベントを含む)の記述と群間の比較(Gray検定)
- 累積発生率(競合イベントを含む)に対する多変量解析(Fine-Gray比例ハザード回帰)

保存 モデル: < アクティブモデルなし >

## 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。

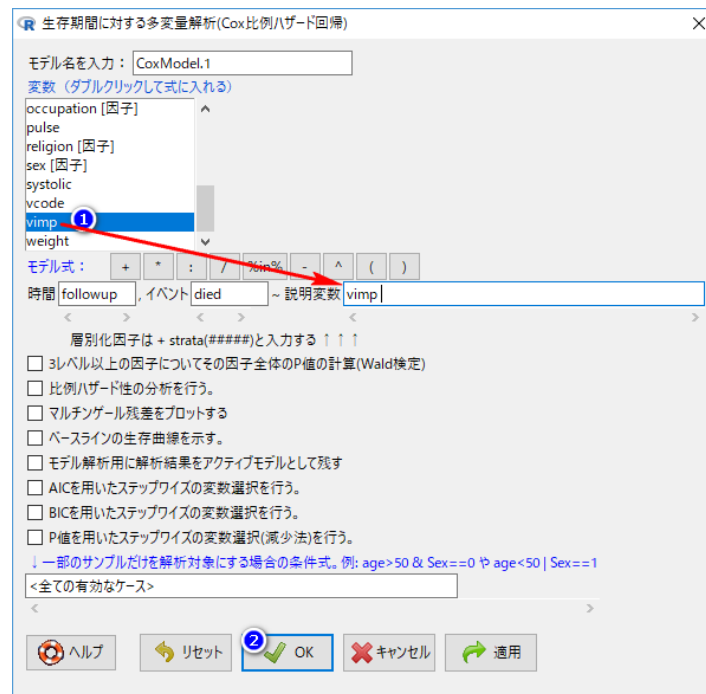
☐ P値を用いたステップワイズの変数選択(減少法)を行う。

↓ 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

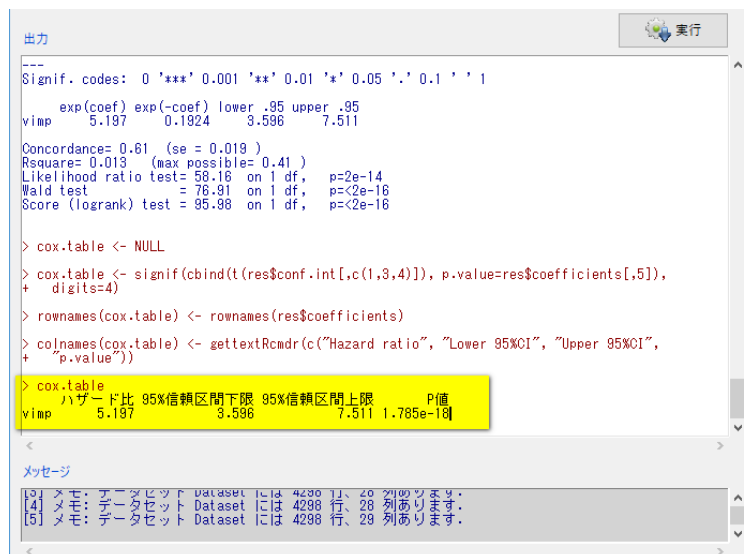
<全ての有効なケース>

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.



## 4.0.3.1 答え

## 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: CoxModel.2

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: + \* : / %in% - ^ ( )

時間: followup, イベント: died ~ 説明変数: vimp + agegrp

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。

```

## 4.0.4 答え

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!」.