R for data analysis

Clinical Research Support Unit

Contents

Importing dataset (csv file)

Importing data set in CSV file named ‘data1’

Check your file extensions!! What type of file do you have?

```
getwd()
data1 <- read.csv(“location of the data/data1.csv", header = TRUE)

Install.packages(“readxl”)
library(readxl)
data1 <- read_excel (location of the data/data1.xlsx")
```

CRSU
Clinical Research Support Unit

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help

promotion_tenure - master - RStu...

Addins

Console

R 4.3.1

```
R version
Copyright
Platform:

R is free
You are we
Type 'lice

  Natural

R is a col
Type 'cont
'citation(

Type 'demo
'help.star
Type 'q()'

[Workspace
```

```
tical Computing

line help, or
help.

romotion_tenure/.RData]
```

- New File
- New Project...
- Open File... ⌘O
- Open File in New Column...
- Recent Files
- Open Project...
- Open Project in New Session...
- Recent Projects
- Import Dataset
  - From Text (base)...
  - From Text (readr)...
  - From Excel...
  - From SPSS...
  - From SAS...
  - From Stata...
- Save ⌘S
- Save As...
- Save All ⌥⌘S
- Print...
- Close ⌘W
- Close All ⇧⌘W
- Close All Except Current ⌥⇧⌘W
- Close Project
- Quit Session...

Environment History Connections Git Tutorial

Diff Commit Pull Push master

Staged	Status	Path
	? ?	.pdf
	? ?	.DS_Store
	? ?	.gitignore
	? ?	Appendix T_1.pdf
	? ?	Appendix T_2.pdf
	? ?	Appendix T_3.pdf
	? ?	cv.docx
	? ?	cv.html
	? ?	cv.log
	? ?	cv.tex
	? ?	cv_Daniel_Fuller.docx
	? ?	promotion_tenure.Rproj
	? ?	research.html

Files Plots Packages Help Viewer Presentation

R: Search Results Find in Topic

Search Results

The search string was **"tbl_regression"**

Vignettes:

gtsummary::tbl_regression Tutorial: tbl_regression HTML source R coc

Help pages:

gtsummary::add_global_p	Add the global p-values
gtsummary::add_n_regression	Add N to regression table
gtsummary::add_nevent_regression	Add event N to regression table
gtsummary::add_q	Add a column of q-values to account for multiple comparisons
gtsummary::bold_italicize_labels_levels	Bold or Italicize labels or levels in gtsummary tables

CRSU
Clinical Research Support Unit

Import Excel Data

File/URL:

~/Dropbox/Teaching/USask/Intro to R/2023 Intro To R/data1.xlsx

Data Preview:

id (double)	sex (double)	ethgrp (double)	weight (double)	age (double)	cvd (double)	stroke (double)	smoking (double)	Cancer (double)	ldl1 (double)	ldl2 (double)	gender (character)
1	1	3	34	39	0	0	0	1	107	106	f
2	1	3	39	42	0	1	0	1	110	109	f
3	0	2	63	63	1	1	0	0	111	109	m
4	1	2	44	39	0	1	1	0	107	108	f
5	0	2	47	45	1	1	0	0	107	106	m
6	1	2	47	40	0	1	1	1	108	106	f
7	0	2	57	47	1	1	0	1	108	109	m
8	1	2	39	44	0	0	0	0	109	109	f
9	0	3	48	44	1	1	0	0	110	107	m
10	1	1	47	53	0	1	0	0	108	110	f
11	0	2	34	39	0	1	0	1	108	110	m
12	0	3	37	39	0	1	1	1	106	108	m
13	1	3	47	47	1	0	0	1	107	109	f
14	0	2	47	42	1	1	0	0	111	109	m
15	0	3	39	26	1	1	0	0	106	107	m
16	0	3	47	36	1	1	1	0	110	108	m

Previewing first 50 entries.

Import Options:

Name: data1

Sheet: Default

Range: A1:D10

Max Rows:

Skip: 0

NA:

☑ First Row as Names

☑ Open Data Viewer

Code Preview:

```
library(readxl)
data1 <- read_excel("~/Dropbox/Teaching/USask/Intro to R/2023 Intro To R/data1.xlsx")
View(data1)
```

Reading Excel files using readxl

Import

R Markdown – Killer feature

https://yihui.org/

User created things are amazing

https://rmarkdown.rstudio.com/

Basic idea of R Markdown is to integrate statistical code, output, and interpretation into one generic, reproducible, and transferable file format

If you use it people will think you are magic

Default standard for R users doing exploratory data analysis

R Markdown – Killer feature

Based on generic Markdown
Very simple document formatting language

Markdown Quick Reference

R Markdown is an easy-to-write plain text format for creating dynamic documents and reports. See Using R Markdown to learn more.

Emphasis

```
*italic*   **bold**

_italic_   __bold__
```

Headers

```
# Header 1

## Header 2

### Header 3
```

Lists

Unordered List

```
* Item 1
* Item 2
    + Item 2a
    + Item 2b
```

CRSU
Clinical Research Support Unit

promotion_tenure - master - RStudio

promotion_tenure

New File
New Project...
Open File... ⌘O
Open File in New Column...
Recent Files
Open Project...
Open Project in New Session...
Recent Projects
Import Dataset
Save ⌘S
Save As...
Save All ⌥⌘S
Print...
Close ⌘W
Close All ⇧⌘W
Close All Except Current ⌥⇧⌘W
Close Project
Quit Session...

R Script ⇧⌘N
Quarto Document...
Quarto Presentation...
R Notebook
R Markdown...
Shiny Web App...
Plumber API...
C File
C++ File
Header File
Markdown File
HTML File
CSS File
JavaScript File
D3 Script
Python Script
Shell Script
SQL Script
Stan File
Text File
R Sweave
R HTML
R Documentation...

Console

R 4.3.1

```
R version
Copyright
Platform:

R is free
You are we
Type 'lice

  Natural

R is a col
Type 'cont
'citation(

Type 'demo
'help.star
Type 'q()'

[Workspace

> library(readxl)
> data1 <- read_excel("~/Dropbox/Teaching/USas                    R/data1.xlsx")
> View(data1)
>
```

Environment | History | Connections | Git | Tutorial

Import Dataset | 327 MiB | List

R | Global Environment

Data

cchs	5000 obs. of 3 variables
data1	200 obs. of 12 variables
freq_by_time	List of 9
freq_by_time_fit…	List of 9
freq_by_time_fit…	List of 9
freq_histo	List of 9
lm_type	List of 12
nb_type	List of 30
recruitment	199 obs. of 179 variables
recruitment_filt…	195 obs. of 177 variables

Files | Plots | Packages | Help | Viewer | Presentation

R: Search Results | Find in Topic

Search Results

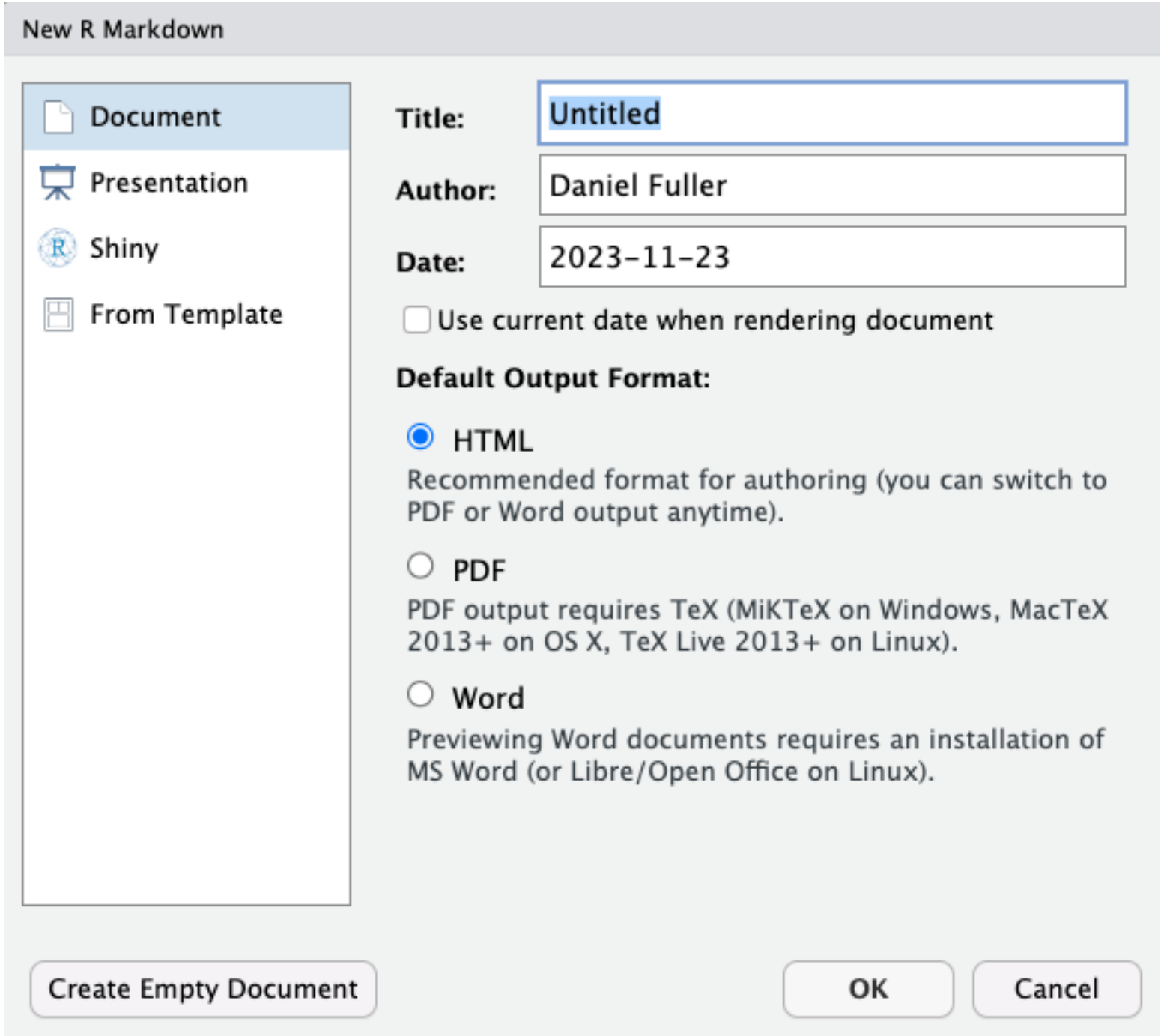
The search string was **"tbl_regression"**

Vignettes:

gtsummary::tbl_regression Tutorial: tbl_regression HTML source R code

Help pages:

gtsummary::add_global_p	Add the global p-values
gtsummary::add_n_regression	Add N to regression table
gtsummary::add_nevent_regression	Add event N to regression table
gtsummary::add_q	Add a column of q-values to account for multiple comparisons
gtsummary::bold_italicize_labels_levels	Bold or Italicize labels or levels in gtsummary tables



CRSU
Clinical Research Support Unit

promotion_tenure - master - RStudio

intro_to_R_day2.Rmd

```
---
title: "Intro to R - Day 2"
author: "Daniel Fuller"
date: "2023-11-23"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```{r cars}
summary(cars)
```

Including Plots
```

9:24 Chunk 1: setup R Markdown

Console Terminal Background Jobs

R 4.3.1 · ~/Dropbox/MUN/PnT/PnT/promotion_tenure/

```
R version 4.3.1 (2023-06-16) -- "Beagle Scouts"
Copyright (C) 2023 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin20 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
```

Environment History Connections Git Tutorial

Import Dataset 327 MiB List

R Global Environment

Data

cchs	5000 obs. of 3 variables
data1	200 obs. of 12 variables
freq_by_time	List of 9
freq_by_time_fit…	List of 9
freq_by_time_fit…	List of 9
freq_histo	List of 9
lm_type	List of 12
nb_type	List of 30
recruitment	199 obs. of 179 variables
recruitment_filt…	195 obs. of 177 variables

Files Plots Packages Help Viewer Presentation

R: Search Results Find in Topic

Search Results

The search string was **"tbl_regression"**

Vignettes:

gtsummary::tbl_regression Tutorial: tbl_regression HTML source R coc

Help pages:

gtsummary::add_global_p	Add the global p-values
gtsummary::add_n_regression	Add N to regression table
gtsummary::add_nevent_regression	Add event N to regression table
gtsummary::add_q	Add a column of q-values to account for multiple comparisons
gtsummary::bold_italicize_labels_levels	Bold or Italicize labels or levels in gtsummary tables

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help

promotion_tenure - master - RStudio

Go to file/function | Addins

intro_to_R_day2.Rmd

Knit on Save | Knit

Source | Visual

- Knit to HTML
- Knit to PDF
- Knit to Word
- Knit with Parameters...
- Knit Directory
- Clear Knitr Cache...

```
---
title: "Intro
author: "Dani
date: "2023-1
output: html_
---

```{r setup,
knitr::opts_chunk$set(echo = TRUE)
```

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details
Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code
the document. You can embed an R code chunk like this:

```{r cars}
summary(cars)
```

Including Plots
```

3:24 | Intro to R – Day 2

Setup a new Markdown File

```
# Importing data set in CSV file named 'data2'

data2 <- read.csv("location of the data/data2.csv", header = TRUE)

# merging data1 and data2

data1 <- dplyr::select(data1, id, sex, ethgrp, weight, age, cvd)
data2 <- dplyr::select(data2, id, stroke, smoking, Cancer, ldl1, ldl2, gender)

### joining/merging data
data_merge <- dplyr::full_join(data1, data2)

data_merge1 <- dplyr::full_join(data1, data2, by = join_by(id))

data_merge2 <- dplyr::full_join(data1, data2,by = join_by(id == id))
```

CRSU
Clinical Research Support Unit

Importing dataset, Data merging

```
# Importing data set in CSV file named 'data2'

data2 <- read.csv("location of the data/data2.csv", header = TRUE)

# merging data1 and data2

data1 <- dplyr::select(data1, id, sex, ethgrp, weight, age, cvd)
data2 <- dplyr::select(data2, id, stroke, smoking, Cancer, ldl1, ldl2, gender)

### joining/merging data
data_merge <- dplyr::full_join(data1, data2)

data_merge1 <- dplyr::full_join(data1, data2, by = join_by(id))

data_merge2 <- dplyr::full_join(data1, data2,by = join_by(id == id))
```

CRSU
Clinical Research Support Unit

Importing data named ‘test’

```
test <- read.csv("location of the data/test.csv", header = TRUE)

##test

head(test,10)

tail(test,10)
```

Summary statistics

```
# Need package vtable to describe summary statistics
# if vtable is already installed, need to run library only

library(vtable)

st(test)
```

Creating categorical variables

```
# to create categorial variable 'agecat' using age
summary(test$age)

test <- test %>%
        mutate(age_cat = case_when(
          age < 45 ~ "<45",
          age >= 45 & age < 50 ~ "45-49",
          age >= 50 & age < 59 ~ "50-59",
          age >= 60 & age < 65 ~ "60-64",
          TRUE ~ "65+"
        ))

count(test, age_cat)

table(test$age, test$age_cat)
```

CRSU
Clinical Research Support Unit

Chi-square test

```
# Check frequency distribution of a categorical variable
# Cross tabulation
# Chi-square test

# Check frequency distribution of gender
table(test$gender)

# Output:
#   gender
#   f      m
#  117  103
```

Chi-square test

```
# Cross tabulation of gender and stroke

table(test$gender, test$stroke)

# Output
#          stroke
# gender   0   1
#    f    50  67
#    m    24  79
```

Chi-square test

```
# Chi-square test between gender and stroke

chisq.test(test$gender, test$stroke)

# Output
# Pearson's Chi-squared test with Yates' continuity
correction

# data:  gender and stroke
# X-squared = 8.4179, df = 1, p-value = 0.003715
```

CRSU
Clinical Research Support Unit

Fisher’s exact test

fisher.test(test$gender, test$stroke)

```
# Output
# Fisher's Exact Test for Count Data

# data:  gender and stroke
# p-value = 0.002674
# alternative hypothesis: true odds ratio is not equal to 1
# 95 percent confidence interval:
# 1.318319 4.628560
# sample estimates:
# odds ratio
#  2.446347
```

T-test (for two independent samples)

```
# First, we check if data follow Normal distribution
# Normality test (if p>0.05: data are normally distributed)
# Want to compare mean ages between male and female

test_data_female <- filter(test, gender == "f")
shapiro.test(test_data_female$age)

test_data_male <- filter(test, gender == "m")
shapiro.test(test_data_male$age)

# Also check histogtam
hist_age_gender <- ggplot(test, aes(age)) +
  geom_histogram() +
  facet_wrap(~ gender)

plot(hist_age_gender)
```

T-test (for two independent samples)

help(t.test)
t.test(age ~ gender, data = test)

```
 Welch Two Sample t-test
data:  age by gender
t = -1.9624, df = 200.25, p-value = 0.05111
alternative hypothesis: true difference in means between group f and group m is
not equal to 0
95 percent confidence interval:
 -5.26625331  0.01274738
sample estimates:
mean in group f mean in group m
      51.11111        53.73786
```

#by default, unequal variance

CRSU
Clinical Research Support Unit

T-test (for two independent samples)

```
#check variance
var.test(age ~ gender, data = test)

F test to compare two variances
data:  age by gender
F = 0.70681, num df = 116, denom df = 102, p-value = 0.07034
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4830964 1.0293364
sample estimates:
ratio of variances
        0.7068141
```

CRSU
Clinical Research Support Unit

T-test (for two independent samples

variances are equal based on the test

t.test(age ~ gender, data = test, var.equal = TRUE)

```
Two Sample t-test

data:  age by gender
t = -1.984, df = 218, p-value = 0.04851
alternative hypothesis: true difference in means between group f and group m is
not equal to 0
95 percent confidence interval:
 -5.23613291 -0.01737302
sample estimates:
mean in group f mean in group m
      51.11111        53.73786
```

CRSU
Clinical Research Support Unit

Wilcoxon non-parametric test for independent samples

#non-parametric test if data are not normally distributed

wilcox.test(age ~ gender, data = test)

Wilcoxon rank sum test with continuity correction

data: age by gender
W = 5020, p-value = 0.03246
alternative hypothesis: true location shift is not equal to 0

CRSU
Clinical Research Support Unit

Paired t-test for two dependent samples test

#paired t-test for two dependent samples

t.test(test$ldl1, test$ldl2, paired = TRUE)

```
Paired t-test

data:  test$ldl1 and test$ldl2
t = -0.5298, df = 219, p-value = 0.5968
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -0.3003634  0.1730907
sample estimates:
mean difference
   -0.06363636
```

CRSU
Clinical Research Support Unit

Correlation analysis

```
# Pearson correlation coefficient

cor.test(test$ldl1, test$ldl2,  method = "pearson")

Pearson's product-moment correlation
data:  test$ldl1 and test$ldl2
t = 4.1684, df = 218, p-value = 4.425e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1446265 0.3899574
sample estimates:
      cor
0.2717003
```

Correlation Analysis

#Spearman correlation coefficient

```
cor.test(test$ldl1, test$ldl2,  method = "spearman")
```

```
Spearman's rank correlation rho
data:  test$ldl1 and test$ldl2
S = 1339899, p-value = 0.0002436
alternative hypothesis: true rho is not equal to 0
sample estimates:
     rho
0.2449701
```

CRSU
Clinical Research Support Unit

Linear Regression

Conduct linear regression model between dependent and independent variables
age is continuous, gender is categorical, ldl is continuous variable

```
linear_model <- lm(age ~ as.factor(gender) + ldl1, data = test)

summary(linear_model) ### Old School Way

install.packages("gtsummary")
library(gtsummary)

tbl_regression(linear_model)
```

Logistic regression model

```
logistic_model <- glm(Cancer ~ as.factor(gender) + ldl1 + smoking, data = test, family =
"binomial")

summary(logistic_model)

## odds ratios and 95% CI (Old School Way)
exp(cbind(OR = coef(logistic_model), confint(logistic_model)))

## odds ratios and 95% CI (New School way)
tbl_regression(logistic_model, exponentiate = TRUE)
```

CRSU
Clinical Research Support Unit

Analysis of variance (ANOVA)

```
# One-way ANOVA
# Pass arguments to aov() function for an ANOVA test

one.anova <- aov(age ~ ethgrp, data = test)
summary(one.anova)
```

Analysis of variance (ANOVA)

Non-parametric ANOVA

kruskal.test(age ~ ethgrp, data = test)

Kruskal-Wallis rank sum test

data: age by ethgrp
Kruskal-Wallis chi-squared = 15.426, df = 2, p-value = 0.000447

Acknowledgment
Jiwon Yoon
Prosanta Mondal