Biostatistics: Theory and Applications in R (Virtual)

Week7_Session2_R_training7

Contents

Prof Dr Mohammed Abu Sayed Arfin Khan
Department of Forestry and Environmental Science
Shahjalal University of Science and Technology, Sylhet
+8801917174537, khan-for@sust.edu, nobelarfin@yahoo.com
Homepage | Google scholar | Researchgate | ORCID | Publons | BayCEER

```
#Set the working directory- getwd()/ setwd("Y:/")
getwd()
setwd("C:/Users/Fahmida Sultana/Desktop/R training/R_training_Class_07")


#install openxlsx package or xlsx package
library(openxlsx)
library(readxl)


#####import data set from xlsx file
study_1 <- read.xlsx("Tree_height.xlsx")
str(study_1)
```

#Summary statistics

```
library(Rmisc)
```

#Calculate the sum within a data set

```
sum1<-aggregate(biomass~treatment+year+plant,data=study_1,FUN=sum)
sum1
str(sum1)
```

#Calculate the mean, sd, se, ci within a data set

```
mean1 <- summarySE(study_1, measurevar="biomass", groupvars=c("treatment","year", "plant"),
na.rm=FALSE)
mean1
str(mean1)
```

#Shapiro Test

```
#or Shapiro-Wilk normality test
#Why is it used?
#To test if a sample follows a normal distribution.
#Normaly distributed= if p Value is > 0.05
#Not normaly distributed= if p Value is < 0.05

str(study_1)
shapiro.test(study_1$biomass)
```

#Kolmogorov and Smirnov Test

#is used to check whether 2 samples follow the same distribution.
#From different distributions if p < 0.05
#Both from normal distribution if p > 0.05

```
str(study_1)
ks.test(study_1$year, study_1$biomass)
```

#Chi Squared Test

#used to test if two categorical variables are dependent
#How to tell if x, y are independent?
#if p < 0.05, x and y are not independent.
#if p > 0.05, x and y are independent.

```
str(study_1)
chisq.test(table(study_1$year, study_1$biomass), correct = FALSE)

summary(table(study_1$year, study_1$biomass)) # performs a chi-squared test.
```

#Simple Correlation

```
study_2 <- read_excel("Tree_height.xlsx", sheet = "study_2")
str(study_2)
```

#Calculate correlation between y_axis= diameter and x_axis= elevation
#Correlation can take values between -1 to +1.

```
cor(study_2$Elevation_m, study_2$Diameter_cm)

cor.test(study_2$Elevation_m, study_2$Diameter_cm)
```

#Build Linear Model

```
m1 = lm((Diameter_cm)~Elevation_m, data=study_2)
m1
```

#what dose it mean in lm model

#Diameter_cm = Intercept + (β ∗ Elevation_m)
#Diameter_cm = 51.1743 + (-0.3888 ∗ Elevation_m)

t-statistic, R-square, F-statistic, p-value

#A large t-score, or t-value, indicates that the groups are different
#while a small t-score indicates that the groups are similar
#R-square= range 0-1, higher the better.
#F-statistic= higher the better
#p-value < 0.05, significant variation

summary(m1)

#AIC and BIC

#The Akaike's information criterion - AIC (Akaike, 1974)
#The Bayesian information criterion - BIC (Schwarz, 1978)
#measures of the goodness of fit of an estimated statistical model
#can also be used for model selection.

AIC(m1) # AIC => 123.7626
BIC(m1) # BIC => 125.8867

#How to know if the model is best fit for your data?

#R-Squared- Higher the better (> 0.70)
#Adj R-Squared- Higher the better
#F-Statistic- Higher the better
#Std. Error- Closer to zero the better
#t-statistic- Should be greater 1.96 for p-value to be less than 0.05
#AIC- Lower the better
#BIC- Lower the better

#ANOVA: linear model

```
#testing biomass vs year

par(mfrow = c(1,2))
C = lm((biomass)~year, data=study_1)
plot(fitted(C), resid(C), xlab = "fitted", ylab = "residuals")
qqnorm(resid(C), main = "")
qqline(resid(C), main = "", col = 2)
anova(C)
```

#posthoc test

```
TukeyHSD(aov((biomass) ~ as.factor(year), data = study_1))


#ANOVA
#linear model
#testing biomass vs treatment
str(study_1)
par(mfrow = c(1,2))
C = lm((biomass)~treatment, data=study_1)
plot(fitted(C), resid(C), xlab = "fitted", ylab = "residuals")
qqnorm(resid(C), main = "")
qqline(resid(C), main = "", col = 2)
anova(C)


#ANOVA
#linear model
#testing biomass vs plant

str(study_1)
par(mfrow = c(1,2))
C = lm((biomass)~plant, data=study_1)
plot(fitted(C), resid(C), xlab = "fitted", ylab = "residuals")
qqnorm(resid(C), main = "")
qqline(resid(C), main = "", col = 2)
anova(C)
```

#ANOVA: mixed effect model

```
library(nlme)
```

```
str(study_1)
par(mfrow = c(1,2))
C1 = lme((biomass) ~ treatment*year*plant, random = ~1|repetition, data = study_1)
plot(fitted(C1), resid(C1), xlab = "fitted", ylab = "residuals")
qqnorm(resid(C1), main = "")
qqline(resid(C1), main = "", col = 2)
anova(C1)
```

#posthoc test

```
TukeyHSD(aov((biomass) ~ as.factor(year), random = ~1|repetition, data = study_1))

TukeyHSD(aov((biomass) ~ treatment*as.factor(year), random = ~1|repetition, data = study_1))

TukeyHSD(aov((biomass) ~ treatment*as.factor(year)*plant, random = ~1|repetition, data = study_1))
```