```
library(readxl)
data <- read_excel("2020 - assignment 3.xlsx", sheet = "habspecies", na =
"NA")
head(data)
str(data)
colnames(data)



par(mfrow=c(1,3))
plot(data$E, data$N,col=factor(data$minkp),pch=16, xlab= "East", ylab =
"North",main="mink")
plot(data$E, data$N,col=factor(data$martenp),pch=16, xlab= "East", ylab =
"North",main="marten")
plot(data$E, data$N,col=factor(data$otterp),pch=16, xlab= "East", ylab =
"North",main="otterp")

#install packages

install.packages("car")
install.packages("psych")
install.packages("carData")
install.packages("mvtnorm")


library (car)
library(carData)
library(psych)
library(lattice)
library(data.table)
library(plyr)
library(doBy)
library(afex)
library(multcomp)
library(lsmeans)
library(effects)
library(MASS)
library(plot3D)
library(rgl)
library(gmodels)

#Question 1

mytable1=xtabs(~minkp+martenp, data=data)
CrossTable(mytable1, expected=T, chisq=T, fisher = T, format = "SAS")
CrossTable(mytable1, expected=T, chisq=T,  format = "SAS")
chisq.test(mytable1, p=c(84.778,27.222,24.222,7.778), rescale.p=T, correct=F)

#Conclusion: In this case, the p < 0.05 would imply that we would reject the
#null hypothesis that both factors are independent, implying a significant
#association between Mink and Marten. Pearson's chi-square value
#(Chi^2 =  5.955739), which confirms that association between Mink and Marten
#is statistically significant.
#Biological Conclusion: Mink and marten co-occur more than expected by chance.
```

```
mytable3=xtabs(~minkp+otterp, data=data)
CrossTable(mytable3, expected=T, chisq=T, fisher = T, format = "SAS")
CrossTable(mytable3, expected=T, chisq=T,  format = "SAS")

#Conclusion: In this case, the p >0.05 would imply that we would not reject
#the null hypothesis that both factors are independent, implying there is no
#significant association between Mink and Otter. Pearson’s chi-square value
#(Chi^2 =  2.992208), which confirms that association between Mink and Otter
#is not statistically significant.
#Biological Conclusion: Mink and Otter co-occur less than expected by chance.


mytable5=xtabs(~martenp+otterp, data=data)
CrossTable(mytable5, expected=T, chisq=T, fisher = T, format = "SAS")
CrossTable(mytable5, expected=T, chisq=T,  format = "SAS")

#Conclusion: In this case, the p >0.05 would imply that we would not able to
#reject the null hypothesis that both factors are independent, implying no
#significant association between Marten and Otter. Pearson’s chi-square value
#(Chi^2 =  2.899), which confirms that association between Marten and Otter
#is not statistically significant.
#Biological Conclusion: Marten and Otter co-occur less than expected by
chance.



#Question 2

#created a data set for Figga river, to predict the current of Figga river

data.Figga <-subset(data, rivertype== "main")
data.FiggaRiver <-subset(data, river== "Figga")

# this data set is based on Current of Figga river, Figga river is the main
#type river and rest of the rivers are side rivers

# Model selection and statistical inference without log-transforming BD:
# plot before regression


###############################################################
#Calculations in R
###############################################################


data1 = lm(current ~ riverwidth * oceandist *riverdepth*bankheight,
data=data.Figga)
summary (data1)

# Answer: all variable individually and  with interation are not statistically
signifcant
```

```
# overall model p vaule (0.01725)  and F-value (2.242) indicates model is
statistically significant, this model
#has low R squre and adujusted R square is low [Multiple R-squared:0.407,
Adjusted R-squared:0.2254]

Anova(data1, type = 3)

# Answer:All variable individually and  with interation are not statistically
signifcant

summary(data1)$adj.r.squared
# Answer: [1] 0.225418 Adj R square is bit low that indicated model is not fit
well


###############################################################
#Stepwise Model reduction
###############################################################

full.model=lm(current ~ riverwidth * oceandist *riverdepth*bankheight,
data=data.Figga)
stepAIC(full.model)

#Start:  AIC=-185.76 #Step:  AIC=-189.66 #Step:  AIC=-191.44
#Step:  AIC=-192.54 #Step:  AIC=-192.67 #Step:  AIC=-194.56
#Step:  AIC=-195.3 #Step:  AIC=-195.47 #Step:  AIC=-195.78

#I generated 9 separaet model (reported above), based on AIC value, AIC=
-195.78 is the
#lowest AIC among all the model I generated. here is my final model reported
below


AIC(final)

#plot before regression

attach(data.Figga)
par(mfrow =c(2,2))
plot(current~riverwidth, main = "current vs riverwidth")
plot(current~oceandist, main = "current vs oceandist")
plot(current~riverdepth, main = "current vs riverdepth")
plot(current~bankheight, main = "current vs bankheight")
boxplot(current~riverdepth:bankheight, main = "current vs joint")


#linear preduction model

final=lm(current ~ riverwidth + oceandist + riverdepth +
           bankheight + riverwidth:bankheight + riverdepth:bankheight,
         data = data.Figga)
summary(final)
Anova(final, type = 3) # all effects are significant
summary(final)$adj.r.squared
```

```
# My final model is statistically significant based on P value (0.0005351) and
F statistic (4.749),
# both the value indicates model is overall statistically significant. In my
final model, I observed all the
#variable statistially insignificant at 5% level of significance except
"riverdepth:bankheight"
#(interation variable) which is significant at 5% level of significance. It
means river depth and bank height
# singificantly affect the current of Figga river. According to data Figga
river is the main river and rest of the
#rivers are side rivers. in the final model river weight, ocean distantce and
bank height got the negative
#coefficent which means Current of Figga River negatively associated with
river weight, ocean distantce
# and bank height. only riverwidth*bankheight interaction has positive
coefficent which means these two
#jointly positively related to current of Figga rive though they are
statistically insignificant at
#5% level of significance.

#Plot after regression
plot(allEffects(final))




# Model diagnosis, outliers and influential observations for "final"


residualPlots(final)
#linearity cannot be assumed

spreadLevelPlot(final)
# the plot suggests that variance increases

ncvTest(final)
# Based on ncvtest (Chisquare = 1.077289, Df = 1, p = 0.2993).
#Formal test rejects constant variance,based on p value model is statistically
insignificant

studres.final=rstudent(final) # studentized residuals
hist(studres.final,
     probability=T,
     col="lightgrey",
     xlim=c(-6,6),
     breaks=12,
     main="Distribution of Studentized Residuals",
     xlab="Studentized residuals")
xfit=seq(-6,6,length=100)
yfit=dnorm(xfit) # normal fit
lines(xfit, yfit, col="red",lwd=2)

# The distribution looks very good
```

```
shapiro.test(residuals(final))
# Formal test confirming normality since W (0.92) > 0.9

vif(final)
# variance inflation factors here river width, ocean distanct and river depth
are <5, which seems good
# but bankheight, riverwidth:bankheight riverdepth:bankheigh are >5 thats
doesnt seem good


outlierTest(final) # observation 11 is an outlier
#No Studentized residuals with Bonferroni p < 0.05
#Largest |rstudent|:
#rstudent unadjusted p-value Bonferroni p
#11 3.250354          0.0019367      0.12588

influenceIndexPlot(final,vars=c("Studentized","Bonf"))

influenceIndexPlot(final,vars="Cook") # but there are no influential
observations, so no problems there






# Conclusion

# The model "final" meets all assumptions, so we keep this model.
# It predicts Figga river current, whcih is jointly related to
riverwidth*bankheight
# this interaction effect jointly statistically significant

Anova(final, type = 3)
summary(final)

# What does this mean? We take a look to a graphical summary of the model
effects:

plot(allEffects(mod=model.fish.logBD0,partial.residuals=T),
smooth.residuals=F,residuals.color=adjustcolor("blue",alpha.f=0.5),residuals.pch=16)

# The significant riverwidth*bankheight effect indicates that the Figga RIver
current depends on joint
#interaction of river width and bank height of figga river. The rest of the
variables  individullay for exampl
# riverwidth, distance of ocean, river depth, and bank statistically
insignificant. Meaning, these variables
#individually doesn't effect the current of Figga river.

# My linear model confirms that current of Figga river jointly depends on
river width and bank height
#Yes, there is and interaction effect in the model, which is
riverwidth*bankheight (signicant at 5% level)
```

```
#log transformation:


data.Figga$log.current <- log10(data.Figga$current)
data.Figga$log.riverwidth <- log10(data.Figga$riverwidth)
data.Figga$log.oceandist <- log10(data.Figga$oceandist)
data.Figga$log.riverdepth <- log10(data.Figga$riverdepth)
data.Figga$log.bankheight <- log10(data.Figga$bankheight)






################################################################
#Multicollinearity
################################################################
fit1meancent=lm(current~scale(riverwidth, center=T, scale=F)*
                        scale(oceandist, center=T, scale=F)* scale(riverdepth,
center=T, scale=F)*
                        scale(bankheight, center=T, scale=F), data=data.Figga)
vif(fit1meancent)

summary(fit1meancent)
```