Homework 4 and 5 combined

Mitesh Ranmal Jain

11/16/2019

library(data.table)  
library(PMCMR)

## PMCMR is superseded by PMCMRplus and will be no longer maintained. You may wish to install PMCMRplus instead.

library(userfriendlyscience)

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

## Registered S3 methods overwritten by 'lme4':  
## method from  
## cooks.distance.influence.merMod car   
## influence.merMod car   
## dfbeta.influence.merMod car   
## dfbetas.influence.merMod car

library(PASWR2)

## Loading required package: lattice

##   
## Attaching package: 'lattice'

## The following object is masked from 'package:userfriendlyscience':  
##   
## oneway

## Loading required package: ggplot2

library(psych)

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

library(BSDA)

##   
## Attaching package: 'BSDA'

## The following objects are masked from 'package:PASWR2':  
##   
## normarea, nsize, ntester, SIGN.test, tsum.test, z.test, zsum.test

## The following object is masked from 'package:datasets':  
##   
## Orange

## Question 1

Prob\_Type\_1 <- (1 - (0.2+0.3))  
  
Prob\_Type\_2 <- (1 - (0.2 + 0.1 + 0.1 + 0.2 + 0.1))

#### Answer:

Probability of Type I error is 0.5. Probability of type II error is

## Question 2

Question2 <- read.csv("Question2.csv")  
smoke\_break\_avg <- mean(Question2$Minutes)  
smoke\_break\_std <- sd(Question2$Minutes)  
prev\_break\_avg = 32  
prev\_break\_sd = 8  
n <- nrow(Question2)  
  
t.test(Question2$Minutes, alternative = "less", mu = 32)

##   
## One Sample t-test  
##   
## data: Question2$Minutes  
## t = -3.8247, df = 109, p-value = 0.0001093  
## alternative hypothesis: true mean is less than 32  
## 95 percent confidence interval:  
## -Inf 30.82116  
## sample estimates:  
## mean of x   
## 29.91818

We reject and accept mean time will be less than 32 minutes.

z <-(smoke\_break\_avg - prev\_break\_avg)/(prev\_break\_sd/sqrt(n))  
  
p\_sample <- pnorm(z)  
  
z\_critical <- qnorm(0.5)  
  
x\_critical <- 32 + z\_critical\*prev\_break\_sd/sqrt(n)  
  
Type\_2\_Error\_Prob <- 1 - pnorm(x\_critical, mean = 30, prev\_break\_sd/sqrt(n))

#### Answer:

## Question 3

mean\_sales\_9 = 153.60  
sd\_sales\_9 = 25.57  
n\_sales\_9 = 20  
  
mean\_sales\_10 = 151.50  
sd\_sales\_10 = 30.39  
n\_sales\_10 = 20  
  
mean\_sales\_11 = 133.25  
sd\_sales\_11 = 25.03  
n\_sales\_11 = 20  
  
grand\_mean = (mean\_sales\_9 + mean\_sales\_10 + mean\_sales\_11)/3  
total\_n = 60  
  
sst = (20\*((mean\_sales\_9 - grand\_mean)^2)) + (20\*((mean\_sales\_10 - grand\_mean)^2)) + (20\*((mean\_sales\_11 - grand\_mean)^2))  
  
sse = (19\*((sd\_sales\_9)^2)) + (19\*((sd\_sales\_10)^2)) + (19\*((sd\_sales\_11)^2))  
  
mst = sst/2  
  
mse = sse/(60-3)  
  
f\_statistic = mst/mse  
  
f\_critical = qf(p = 0.05, df1 = 2, df2 = 57, lower.tail = FALSE)  
  
if(f\_statistic > f\_critical){  
 print("The manager should conclude that the sales volume WILL differ significantly if the new product is priced $9 or $11")  
} else {  
 print("The manager should conclude that the sales volume WILL NOT differ significantly if the new product is priced $9 or $11")  
}

## [1] "The manager should conclude that the sales volume WILL differ significantly if the new product is priced $9 or $11"

grand\_mean = (mean\_sales\_9 + mean\_sales\_10)/2  
total\_n = 40  
  
sst = (20\*((mean\_sales\_9 - grand\_mean)^2)) + (20\*((mean\_sales\_10 - grand\_mean)^2))  
  
sse = (19\*((sd\_sales\_9)^2)) + (19\*((sd\_sales\_10)^2))  
  
mst = sst/2  
  
mse = sse/(40-2)  
  
f\_statistic = mst/mse  
  
f\_critical = qf(p = 0.05, df1 = 1, df2 = 38, lower.tail = FALSE)  
  
if(f\_statistic > f\_critical){  
 print("The manager should conclude that the sales volume WILL differ significantly if the new product is priced $9.")  
} else {  
 print("The manager should conclude that the sales volume WILL NOT differ significantly if the new product is priced $9.")  
}

## [1] "The manager should conclude that the sales volume WILL NOT differ significantly if the new product is priced $9."

grand\_mean = (mean\_sales\_11 + mean\_sales\_10)/2  
total\_n = 40  
  
sst = (20\*((mean\_sales\_11 - grand\_mean)^2)) + (20\*((mean\_sales\_10 - grand\_mean)^2))  
  
sse = (19\*((sd\_sales\_11)^2)) + (19\*((sd\_sales\_10)^2))  
  
mst = sst/2  
  
mse = sse/(40-2)  
  
f\_statistic = mst/mse  
  
P\_10\_11 <- pf(f\_statistic, 1, 38, lower.tail = FALSE)  
  
if(P\_10\_11 > 0.05){  
 print("The manager should conclude that the sales volume WILL differ significantly if the new product is priced $11.")  
} else {  
 print("The manager should conclude that the sales volume WILL NOT differ significantly if the new product is priced $11.")  
}

## [1] "The manager should conclude that the sales volume WILL differ significantly if the new product is priced $11."

#### Answer:

The sales volume will differ significantly if the product is priced at $11.

## Question 4

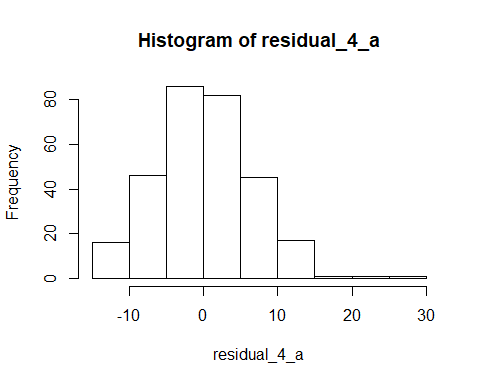
Question4 <- read.csv("Question4.csv")  
  
Cereal <- factor(Question4$Cereal)  
Age <- Question4$Age  
Income <- Question4$Income  
Education <- Question4$Education

### Part a.

model\_4\_a <- lm(Age ~ Cereal)  
residual\_4\_a <- resid(model\_4\_a)  
predict\_4\_a <- predict(model\_4\_a)  
  
## Test for normality  
  
nortest::ad.test(residual\_4\_a)

##   
## Anderson-Darling normality test  
##   
## data: residual\_4\_a  
## A = 0.74779, p-value = 0.051

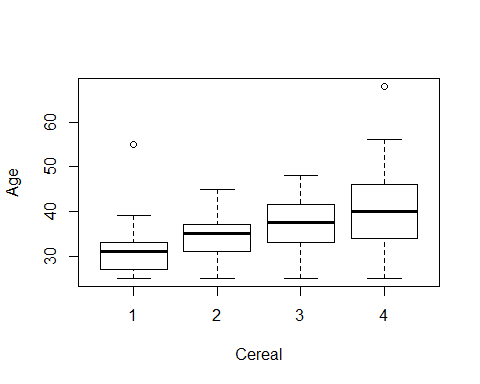
hist(residual\_4\_a)



## Test for equal variances  
  
bartlett.test(Age ~ Cereal)

##   
## Bartlett test of homogeneity of variances  
##   
## data: Age by Cereal  
## Bartlett's K-squared = 36.285, df = 3, p-value = 6.518e-08

boxplot(Age ~ Cereal)



oneway.test(Age ~ Cereal, var.equal = FALSE)

##   
## One-way analysis of means (not assuming equal variances)  
##   
## data: Age and Cereal  
## F = 25.154, num df = 3.00, denom df = 132.24, p-value = 6.11e-13

p-value is less than 0.05, hence we can conclude that there are differences between the ages of the consumers of the 4 different cereals.

posthocTGH(Age, Cereal, method = "games-howell")

## n means variances  
## 1 63 31 28  
## 2 81 34 23  
## 3 40 37 31  
## 4 111 40 72

## Registered S3 methods overwritten by 'ufs':  
## method from   
## grid.draw.ggProportionPlot userfriendlyscience  
## pander.associationMatrix userfriendlyscience  
## pander.dataShape userfriendlyscience  
## pander.descr userfriendlyscience  
## pander.normalityAssessment userfriendlyscience  
## print.CramersV userfriendlyscience  
## print.associationMatrix userfriendlyscience  
## print.confIntOmegaSq userfriendlyscience  
## print.confIntV userfriendlyscience  
## print.dataShape userfriendlyscience  
## print.descr userfriendlyscience  
## print.ggProportionPlot userfriendlyscience  
## print.meanConfInt userfriendlyscience  
## print.multiVarFreq userfriendlyscience  
## print.normalityAssessment userfriendlyscience  
## print.scaleDiagnosis userfriendlyscience  
## print.scaleStructure userfriendlyscience  
## print.scatterMatrix userfriendlyscience

## diff ci.lo ci.hi t df p  
## 2-1 3.1 0.88 5.4 3.6 126 <.01  
## 3-1 6.1 3.16 9.0 5.5 80 <.01  
## 4-1 8.6 5.91 11.3 8.2 170 <.01  
## 3-2 3.0 0.24 5.7 2.9 68 .03  
## 4-2 5.5 3.00 8.0 5.7 180 <.01  
## 4-3 2.6 -0.57 5.7 2.1 105 .15

#### Answer:

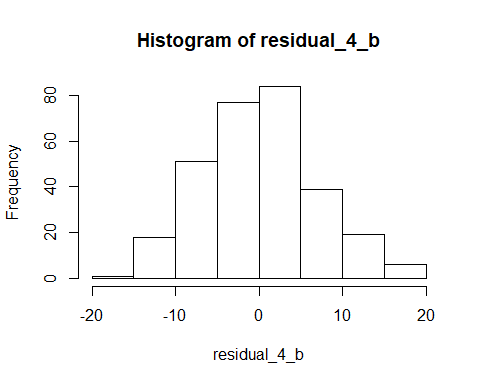
There are differences between the ages of the consumers of the four cereals

### Part b.

model\_4\_b <- lm(Income ~ Cereal)  
residual\_4\_b <- resid(model\_4\_b)  
predict\_4\_b <- predict(model\_4\_b)  
  
## Test for normality  
  
nortest::ad.test(residual\_4\_b)

##   
## Anderson-Darling normality test  
##   
## data: residual\_4\_b  
## A = 0.38897, p-value = 0.3826

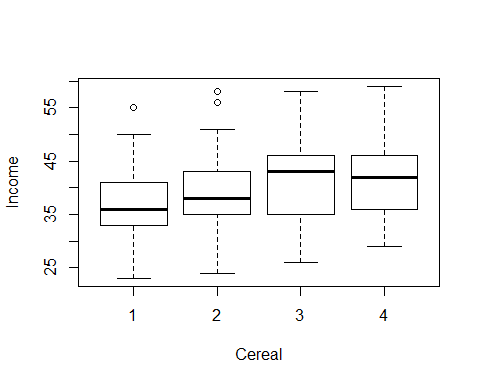
hist(residual\_4\_b)



## Test for equal variances  
  
bartlett.test(Income ~ Cereal)

##   
## Bartlett test of homogeneity of variances  
##   
## data: Income by Cereal  
## Bartlett's K-squared = 2.9232, df = 3, p-value = 0.4036

boxplot(Income ~ Cereal)

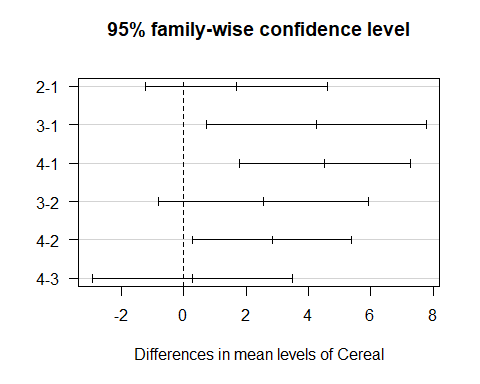


ANOVA\_4\_b <- aov(Income ~ Cereal)  
summary(ANOVA\_4\_b)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Cereal 3 1007 335.8 7.372 8.9e-05 \*\*\*  
## Residuals 291 13256 45.6   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

p-value is less than 0.05, hence we can conclude that there are differences between the incomes of the consumers of the 4 different cereals.

plot(TukeyHSD(ANOVA\_4\_b), las = 1)



#### Answer:

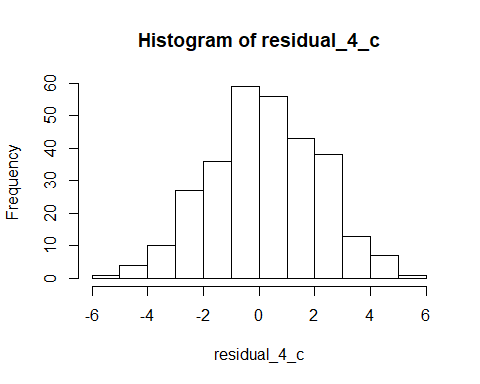
There are differences between the incomes of the consumers of the four cereals.

### Part c.

model\_4\_c <- lm(Education ~ Cereal)  
residual\_4\_c <- resid(model\_4\_c)  
predict\_4\_c <- predict(model\_4\_c)  
  
## Test for normality  
  
nortest::ad.test(residual\_4\_c)

##   
## Anderson-Darling normality test  
##   
## data: residual\_4\_c  
## A = 0.87762, p-value = 0.02435

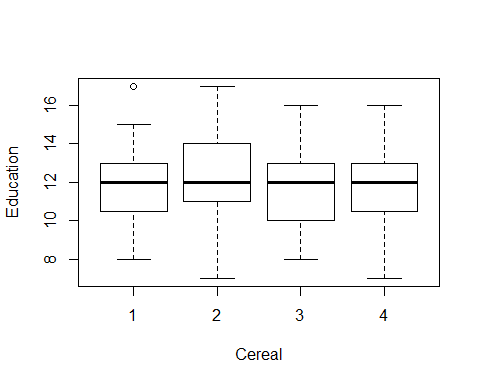
hist(residual\_4\_c)



## Test for equal variances  
  
bartlett.test(Education ~ Cereal)

##   
## Bartlett test of homogeneity of variances  
##   
## data: Education by Cereal  
## Bartlett's K-squared = 1.3761, df = 3, p-value = 0.7111

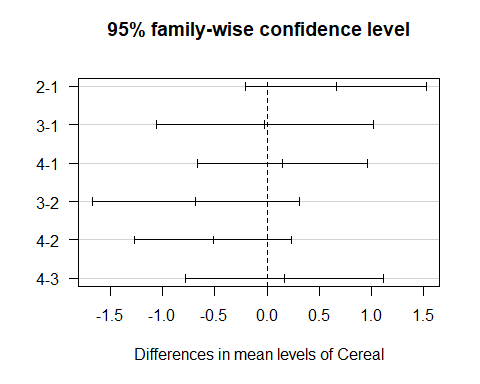
boxplot(Education ~ Cereal)



ANOVA\_4\_c <- aov(Education ~ Cereal)  
summary(ANOVA\_4\_c)

## Df Sum Sq Mean Sq F value Pr(>F)  
## Cereal 3 21.7 7.236 1.824 0.143  
## Residuals 291 1154.2 3.966

plot(TukeyHSD(ANOVA\_4\_c), las = 1)



#### Answer:

There are NO differences between the education of the consumers of the four cereals.

### Part d.

For Years, the following cereals differ:   
Cereal 2 and 1  
Cereal 3 and 1  
Cereal 4 and 1  
Cereal 4 and 2

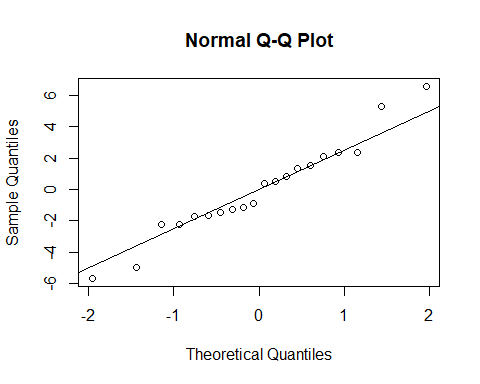
For Income, the following cereals differ:  
Cereal 3 and 1  
Cereal 4 and 1  
Cereal 4 and 2

## Question 5

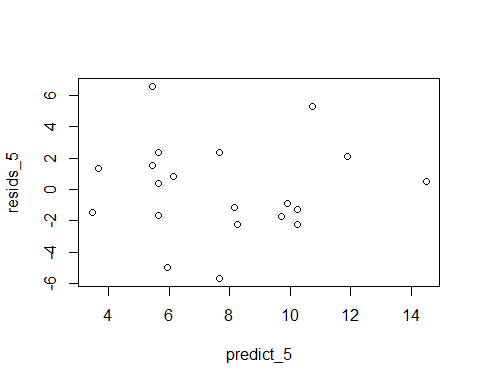
Question5 <- read.csv("Question5.csv")  
  
Group <- factor(Question5$Group)  
Diet <- factor(Question5$Treatment)  
Weight\_Loss <- Question5$WeightLoss  
  
model\_5 <- lm(Weight\_Loss ~ Group + Diet)  
resids\_5 <- residuals(model\_5)  
predict\_5 <- predict(model\_5)  
  
## Normality Tests  
  
nortest::ad.test(resids\_5)

##   
## Anderson-Darling normality test  
##   
## data: resids\_5  
## A = 0.33216, p-value = 0.4816

qqnorm(resids\_5)  
qqline(resids\_5)



## Variance Tests  
  
plot(resids\_5 ~ predict\_5)



car::leveneTest(Weight\_Loss ~ Diet)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 3 0.7001 0.5656  
## 16

car::leveneTest(Weight\_Loss ~ Group)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 4 0.5625 0.6935  
## 15

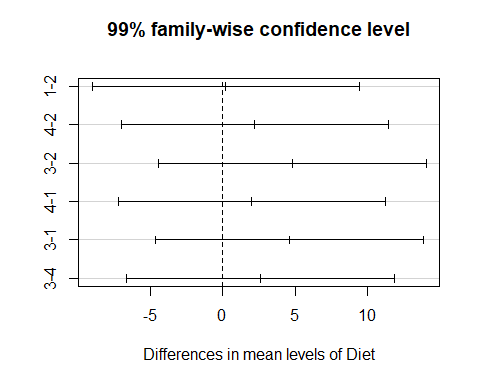
ANOVA\_5 <- aov(Weight\_Loss ~ Group + Diet)  
summary(ANOVA\_5)

## Df Sum Sq Mean Sq F value Pr(>F)  
## Group 4 83.2 20.80 1.475 0.270  
## Diet 3 74.8 24.93 1.768 0.207  
## Residuals 12 169.2 14.10

TukeyHSD(ANOVA\_5, which = 'Diet', ordered = TRUE, conf.level = 0.99)

## Tukey multiple comparisons of means  
## 99% family-wise confidence level  
## factor levels have been ordered  
##   
## Fit: aov(formula = Weight\_Loss ~ Group + Diet)  
##   
## $Diet  
## diff lwr upr p adj  
## 1-2 0.2 -9.038802 9.438802 0.9997731  
## 4-2 2.2 -7.038802 11.438802 0.7915816  
## 3-2 4.8 -4.438802 14.038802 0.2337741  
## 4-1 2.0 -7.238802 11.238802 0.8335433  
## 3-1 4.6 -4.638802 13.838802 0.2641935  
## 3-4 2.6 -6.638802 11.838802 0.6992112

plot(TukeyHSD(ANOVA\_5, which = 'Diet', ordered = TRUE, conf.level = 0.99))

 ### Answer:

Randomized Block Design Model was used for this experiment. The p-values for the treatment(Diet) and blocks(Group) is NOT statitistically significant at 1% confidence level. So we can conclude that there is NO difference among the 4 diets and blocking was NOT meaningful. From this we can conclude that the experimental design was NOT sound.

## Question 6

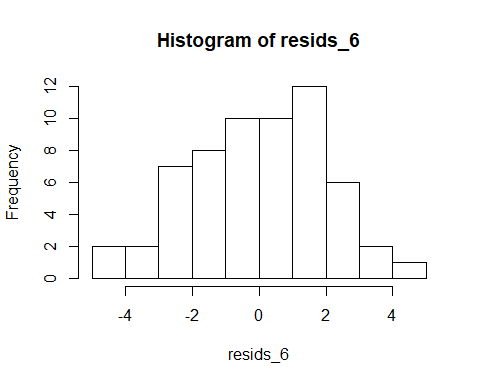
Question6 <- read.csv("Question6.csv")  
  
stack\_6\_business <- Question6[, c(1,2)]  
stack\_6\_arts <- Question6[, c(1,3)]  
stack\_6\_science <- Question6[, c(1,4)]  
  
stack\_6\_business["Degree"] = "Business"  
stack\_6\_arts["Degree"] = "Arts"  
stack\_6\_science["Degree"] = "Science"  
  
names(stack\_6\_arts)[2] <- "Commission"  
names(stack\_6\_business)[2] <- "Commission"  
names(stack\_6\_science)[2] <- "Commission"  
  
stack\_6 <- rbind(stack\_6\_arts, stack\_6\_business, stack\_6\_science)  
  
Group\_6 <- factor(stack\_6$Group)  
Commission <- stack\_6$Commission  
Degree <- factor(stack\_6$Degree)

#### Part a.

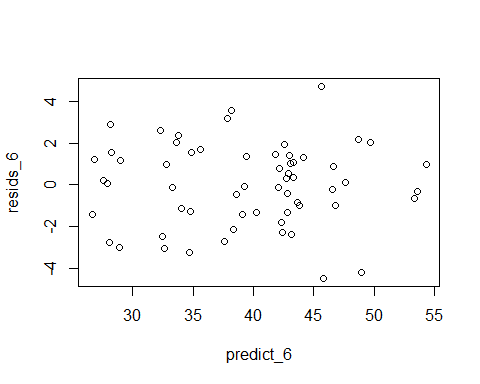
model\_6 <- lm(Commission ~ Group\_6 + Degree)  
resids\_6 <- residuals(model\_6)  
predict\_6 <- predict(model\_6)  
  
## Normality Tests  
  
nortest::ad.test(resids\_6)

##   
## Anderson-Darling normality test  
##   
## data: resids\_6  
## A = 0.23463, p-value = 0.7838

hist(resids\_6)



## Variance Tests  
  
plot(resids\_6 ~ predict\_6)



car::leveneTest(Commission ~ Degree)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 2 0.4146 0.6626  
## 57

car::leveneTest(Commission ~ Group\_6)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 19 0.456 0.9658  
## 40

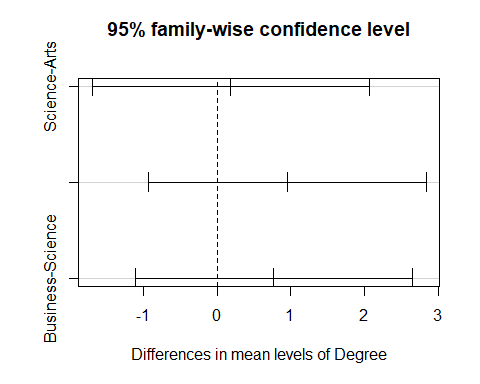
Anova\_6 <- aov(Commission ~ Group\_6 + Degree)  
summary(Anova\_6)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Group\_6 19 3020.3 158.96 26.64 2.4e-16 \*\*\*  
## Degree 2 10.3 5.13 0.86 0.431   
## Residuals 38 226.7 5.97   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

TukeyHSD(Anova\_6, which = 'Degree', ordered = TRUE)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
## factor levels have been ordered  
##   
## Fit: aov(formula = Commission ~ Group\_6 + Degree)  
##   
## $Degree  
## diff lwr upr p adj  
## Science-Arts 0.185 -1.6987645 2.068764 0.9688969  
## Business-Arts 0.955 -0.9287645 2.838764 0.4395209  
## Business-Science 0.770 -1.1137645 2.653764 0.5833203

plot(TukeyHSD(Anova\_6, which = 'Degree', ordered = TRUE))

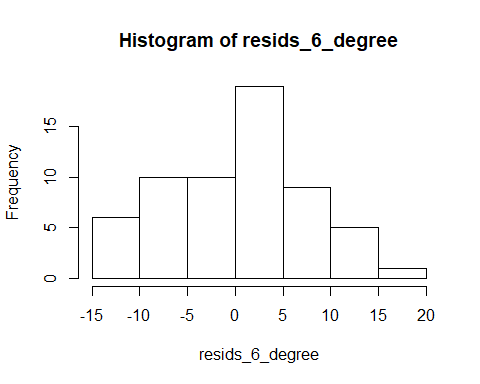
 #### Answer:

#### Part b.

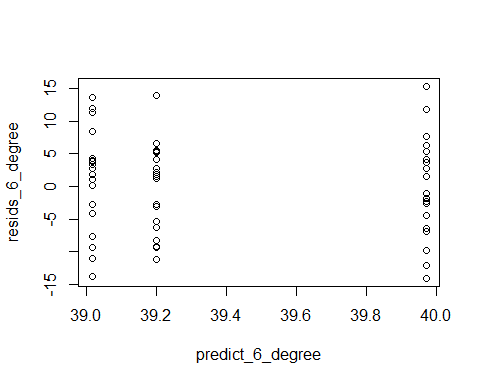
model\_6\_degree <- lm(Commission ~ Degree)  
resids\_6\_degree <- residuals(model\_6\_degree)  
predict\_6\_degree <- predict(model\_6\_degree)  
  
## Normality Tests  
  
nortest::ad.test(resids\_6\_degree)

##   
## Anderson-Darling normality test  
##   
## data: resids\_6\_degree  
## A = 0.505, p-value = 0.1954

hist(resids\_6\_degree)



## Variance Tests  
  
plot(resids\_6\_degree ~ predict\_6\_degree)



car::leveneTest(Commission ~ Degree)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 2 0.4146 0.6626  
## 57

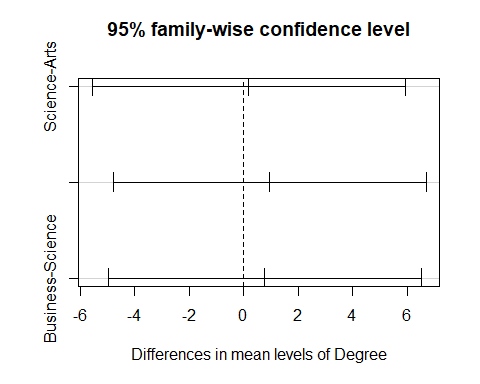
Anova\_6\_Degree <- aov(Commission ~ Degree)  
summary(Anova\_6\_Degree)

## Df Sum Sq Mean Sq F value Pr(>F)  
## Degree 2 10 5.13 0.09 0.914  
## Residuals 57 3247 56.97

TukeyHSD(Anova\_6\_Degree, which = 'Degree', ordered = TRUE)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
## factor levels have been ordered  
##   
## Fit: aov(formula = Commission ~ Degree)  
##   
## $Degree  
## diff lwr upr p adj  
## Science-Arts 0.185 -5.558483 5.928483 0.9966933  
## Business-Arts 0.955 -4.788483 6.698483 0.9156674  
## Business-Science 0.770 -4.973483 6.513483 0.9442990

plot(TukeyHSD(Anova\_6\_Degree, which = 'Degree', ordered = TRUE))

 #### Answer: If only degree is taken into account there is NO sufficient evidence to allow the recruiter to conclude that there are differences in sales ability between the holders of the three types of degrees.

## Question 7

### Test for differences between the levels of Age

### Test for differences between the levels of Gender

### Test for interaction between Age and Gender

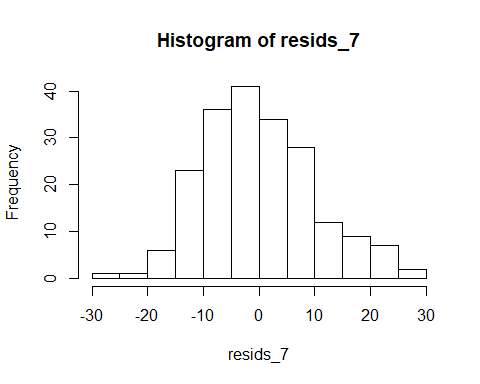
Question7 <- read.csv("Question7.csv")  
  
test <- Question7  
Question7[1:20, 1] = "Males"  
Question7[21:40, 1] = "Females"  
  
Age1 <- Question7[, c(1,2)]  
Age1["Age"] = "Age1"  
names(Age1)[2] = "MilesDriven"  
names(Age1)[1] = "Gender"  
  
Age2 <- Question7[, c(1,3)]  
Age2["Age"] = "Age2"  
names(Age2)[2] = "MilesDriven"  
names(Age2)[1] = "Gender"  
  
Age3 <- Question7[, c(1,4)]  
Age3["Age"] = "Age3"  
names(Age3)[2] = "MilesDriven"  
names(Age3)[1] = "Gender"  
  
Age4 <- Question7[, c(1,5)]  
Age4["Age"] = "Age4"  
names(Age4)[2] = "MilesDriven"  
names(Age4)[1] = "Gender"  
  
Age5 <- Question7[, c(1,6)]  
Age5["Age"] = "Age5"  
names(Age5)[2] = "MilesDriven"  
names(Age5)[1] = "Gender"  
  
stack\_7 <- rbind(Age1, Age2, Age3, Age4, Age5)  
  
Gender <- factor(stack\_7$Gender)  
MilesDriven <- as.numeric(stack\_7$MilesDriven)  
Age\_7 <- factor(stack\_7$Age)  
  
model\_7 <- lm(MilesDriven ~ Gender + Age\_7 + Gender\*Age\_7)  
resids\_7 <- residuals(model\_7)  
predicts\_7 <- predict(model\_7)  
  
## Test for normality  
  
nortest::ad.test(resids\_7)

##   
## Anderson-Darling normality test  
##   
## data: resids\_7  
## A = 1.0324, p-value = 0.01003

shapiro.test(resids\_7)

##   
## Shapiro-Wilk normality test  
##   
## data: resids\_7  
## W = 0.98121, p-value = 0.008857

hist(resids\_7)



## The data is normal

## Variance test  
  
car::leveneTest(MilesDriven ~ Gender)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)   
## group 1 3.9121 0.04933 \*  
## 198   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

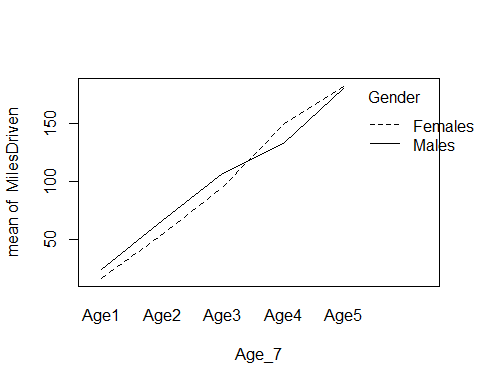
car::leveneTest(MilesDriven ~ Age\_7)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 4 0 1  
## 195

## ANOVA  
  
ANOVA\_7 <- aov(MilesDriven ~ Gender + Age\_7 + Gender\*Age\_7)  
summary(ANOVA\_7)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Gender 1 354 354 3.297 0.071 .   
## Age\_7 4 640000 160000 1491.066 < 2e-16 \*\*\*  
## Gender:Age\_7 4 5908 1477 13.765 7.08e-10 \*\*\*  
## Residuals 190 20388 107   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

interaction.plot(Age\_7, Gender, MilesDriven)

 #### Answer: We do NOT have enough evidence to conclude that males and female drivers differ in the number of miles they drive. We CAN infer that there are differences between the age categories in the number of miles they drive

## Question 8

Observed\_8 = c(0,20,83,52)  
Prob\_8 = c(0.105,0.219,0.533,0.143)  
  
chisq.test(x = Observed\_8,p = Prob\_8)

##   
## Chi-squared test for given probabilities  
##   
## data: Observed\_8  
## X-squared = 62.165, df = 3, p-value = 2.026e-13

#### Answer:

At least 1 frequency after the contest differs from the proportions before the contest.

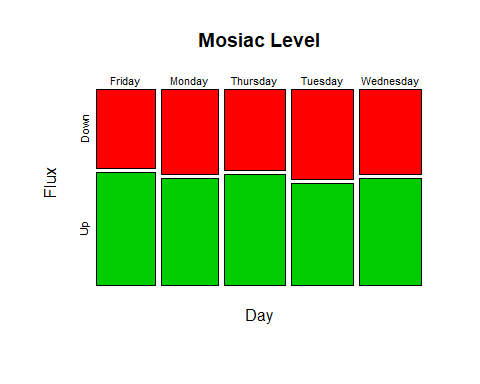
## Question 9

: Market Direction and Day of the week are independent. : Market Direction and Day of the week are not independent.

stack\_9\_data <- data.frame(c(rep(c('Down'),5),  
 rep(c('Up'),5)),   
 rep(c('Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday'),2),  
 c(42,49,46,43,41,  
 53,55,58,59,58))  
  
colnames(stack\_9\_data) <- c("Flux", "Day", "Return")  
  
stack\_9\_data <- stack\_9\_data[rep(seq\_len(nrow(stack\_9\_data)), stack\_9\_data$Return),]  
  
chisq.test(stack\_9\_data$Flux, stack\_9\_data$Day)

##   
## Pearson's Chi-squared test  
##   
## data: stack\_9\_data$Flux and stack\_9\_data$Day  
## X-squared = 0.81886, df = 4, p-value = 0.9359

mosaicplot(~ Day + Flux, data = stack\_9\_data,  
 main = "Mosiac Level", xlab = "Day", ylab = "Flux",  
 col = c (18,3))

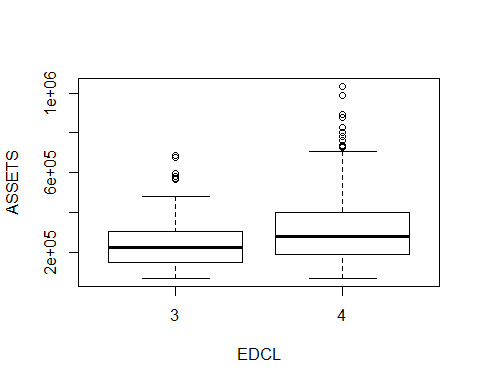
 #### Answer: p-value is greater than 0.05 and we fail to reject i.e. Market Direction and Day of the week are independent.

## Question 10

: The two population locations are the same

: The location of the assets of households with college degrees is to the right of the location of assets of households with some college

Question10 <- read.csv("Question10.csv")  
  
Question10$EDCL <- factor(Question10$EDCL)  
  
Question10 <- Question10[Question10$EDCL == "4"|Question10$EDCL == "3",]  
  
EDCL <- factor(Question10$EDCL)  
ASSETS <- Question10$ASSET  
  
ASSETS <- as.numeric(gsub(",", "", ASSETS))  
  
boxplot(ASSETS ~ EDCL)



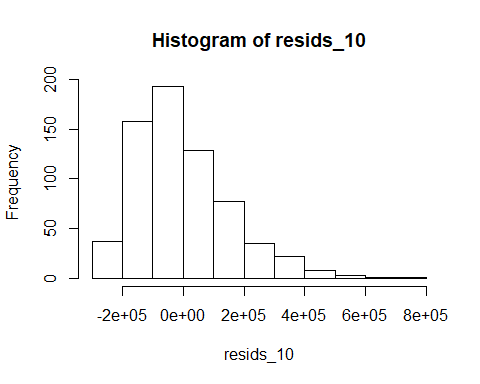
tapply(ASSETS, EDCL, sd)

## 3 4   
## 121541.6 168671.7

model\_10 <- lm(ASSETS ~ EDCL)  
resids\_10 <- residuals(model\_10)  
preds\_10 <- predict(model\_10)  
  
nortest::ad.test(resids\_10)

##   
## Anderson-Darling normality test  
##   
## data: resids\_10  
## A = 9.4906, p-value < 2.2e-16

hist(resids\_10)



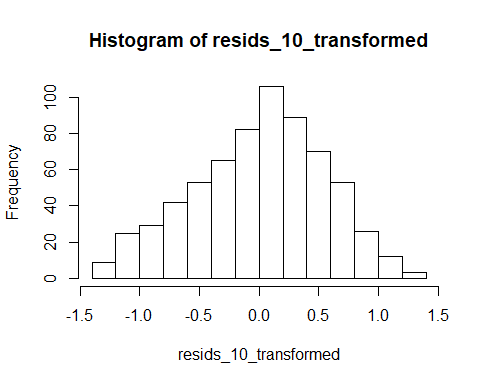
car::leveneTest(ASSETS ~ EDCL)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)   
## group 1 16.613 5.141e-05 \*\*\*  
## 662   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### Transform using log  
  
Transformed\_ASSETS <- log(ASSETS)  
  
model\_10\_transformed <- lm(Transformed\_ASSETS ~ EDCL)  
resids\_10\_transformed <- residuals(model\_10\_transformed)  
preds\_10\_transformed <- predict(model\_10\_transformed)  
  
nortest::ad.test(resids\_10\_transformed)

##   
## Anderson-Darling normality test  
##   
## data: resids\_10\_transformed  
## A = 1.7156, p-value = 0.0002125

## The data is not normally distributed  
  
hist(resids\_10\_transformed)



## Test for variance  
  
car::leveneTest(Transformed\_ASSETS ~ EDCL)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 1 0.9384 0.333  
## 662

## The variances are almost equal  
  
wilcox.test(Transformed\_ASSETS ~ EDCL, alt = "less", paired = FALSE)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: Transformed\_ASSETS by EDCL  
## W = 36122, p-value = 9.613e-08  
## alternative hypothesis: true location shift is less than 0

### Answer:

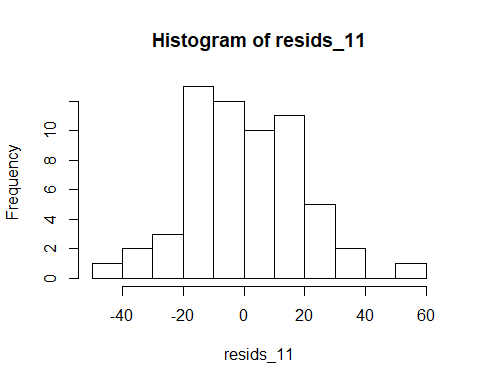
There is enough evidence to conclude that heads of households with college degrees have more assets than those who have some college i.e. we reject .

## Question 11

Question11 <- read.csv("Question11.csv")  
stacked\_11 <- stack(Question11)  
  
wait\_time <- stacked\_11$values  
campus <- stacked\_11$ind  
  
model\_11 <- lm(wait\_time ~ campus)  
resids\_11 <- residuals(model\_11)  
predicts\_11 <- predict(model\_11)  
  
## Normality Test  
  
nortest::ad.test(resids\_11)

##   
## Anderson-Darling normality test  
##   
## data: resids\_11  
## A = 0.29114, p-value = 0.5973

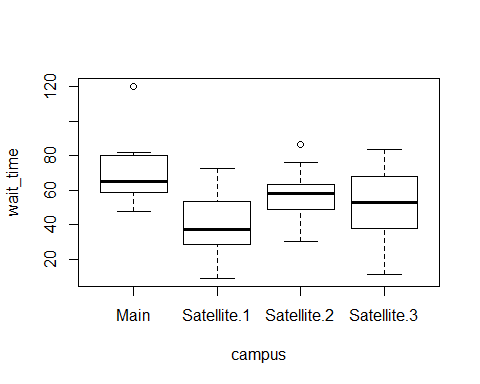
hist(resids\_11)



## Test for constant variance  
  
bartlett.test(wait\_time ~ campus)

##   
## Bartlett test of homogeneity of variances  
##   
## data: wait\_time by campus  
## Bartlett's K-squared = 1.766, df = 3, p-value = 0.6224

boxplot(wait\_time ~ campus)



ANOVA\_11 <- aov(wait\_time ~ campus)  
summary(ANOVA\_11)

## Df Sum Sq Mean Sq F value Pr(>F)   
## campus 3 6312 2104.1 6.372 0.000859 \*\*\*  
## Residuals 56 18493 330.2   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

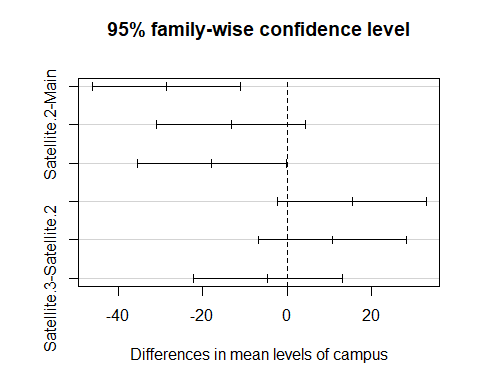
## There is significant difference in the mean wait time at the different campuses  
  
posthocTGH(wait\_time, campus, method = "games-howell")

## n means variances  
## Main 15 70 331  
## Satellite.1 15 41 376  
## Satellite.2 15 57 205  
## Satellite.3 15 52 408  
##   
## diff ci.lo ci.hi t df p  
## Satellite.1-Main -28.6 -47.3 -9.8 4.16 28 <.01  
## Satellite.2-Main -13.3 -29.7 3.1 2.22 27 .14  
## Satellite.3-Main -17.9 -37.1 1.3 2.54 28 .07  
## Satellite.2-Satellite.1 15.3 -1.8 32.4 2.46 26 .09  
## Satellite.3-Satellite.1 10.7 -9.0 30.5 1.48 28 .46  
## Satellite.3-Satellite.2 -4.6 -22.2 13.0 0.72 25 .89

TukeyHSD(ANOVA\_11, las = 1)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = wait\_time ~ campus)  
##   
## $campus  
## diff lwr upr p adj  
## Satellite.1-Main -28.588667 -46.158986 -11.0183478 0.0003814  
## Satellite.2-Main -13.281333 -30.851652 4.2889855 0.1997139  
## Satellite.3-Main -17.870667 -35.440986 -0.3003478 0.0447878  
## Satellite.2-Satellite.1 15.307333 -2.262986 32.8776522 0.1086697  
## Satellite.3-Satellite.1 10.718000 -6.852319 28.2883188 0.3785196  
## Satellite.3-Satellite.2 -4.589333 -22.159652 12.9809855 0.8998737

plot(TukeyHSD(ANOVA\_11, las = 1))



#### Answer:

There is evidence of a difference in waiting time.

## Question 12

Question12 <- read.csv("Question12.csv")  
  
stack\_12 <- melt(data = Question12, id.vars = c("Person"), measure.vars = c("Original", "New.Recipe.1", "New.Recipe..2"))

## Warning in melt(data = Question12, id.vars = c("Person"), measure.vars =  
## c("Original", : The melt generic in data.table has been passed a data.frame  
## and will attempt to redirect to the relevant reshape2 method; please note that  
## reshape2 is deprecated, and this redirection is now deprecated as well. To  
## continue using melt methods from reshape2 while both libraries are attached,  
## e.g. melt.list, you can prepend the namespace like reshape2::melt(Question12).  
## In the next version, this warning will become an error.

Rating <- stack\_12$value  
Person <- factor(stack\_12$Person)  
Recipe <- factor(stack\_12$variable)  
  
model\_12 <- lm(Rating ~ Recipe)  
resids\_12 <- residuals(model\_12)  
predicts\_12 <- predict(model\_12)  
  
nortest::ad.test(resids\_12)

##   
## Anderson-Darling normality test  
##   
## data: resids\_12  
## A = 0.59315, p-value = 0.1188

## The data is normal

## Check for variance  
bartlett.test(Rating ~ Recipe)

##   
## Bartlett test of homogeneity of variances  
##   
## data: Rating by Recipe  
## Bartlett's K-squared = 0.2875, df = 2, p-value = 0.8661

## The data satisies constant variance requirement

ANOVA\_12 <- aov(Rating ~ Recipe + Person)  
summary(ANOVA\_12)

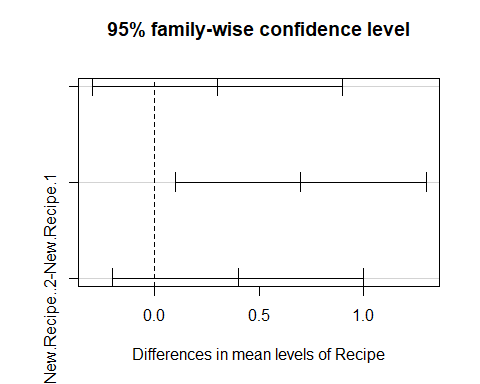
## Df Sum Sq Mean Sq F value Pr(>F)   
## Recipe 2 4.93 2.467 4.064 0.0252 \*   
## Person 19 56.98 2.999 4.941 1.43e-05 \*\*\*  
## Residuals 38 23.07 0.607   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## There is significant difference in ratings at 5% significance level

TukeyHSD(ANOVA\_12, which = 'Recipe', ordered = TRUE)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
## factor levels have been ordered  
##   
## Fit: aov(formula = Rating ~ Recipe + Person)  
##   
## $Recipe  
## diff lwr upr p adj  
## New.Recipe.1-Original 0.3 -0.30087128 0.9008713 0.4502703  
## New.Recipe..2-Original 0.7 0.09912872 1.3008713 0.0192397  
## New.Recipe..2-New.Recipe.1 0.4 -0.20087128 1.0008713 0.2482763

plot(TukeyHSD(ANOVA\_12, which = 'Recipe', ordered = TRUE))



#### Answer:

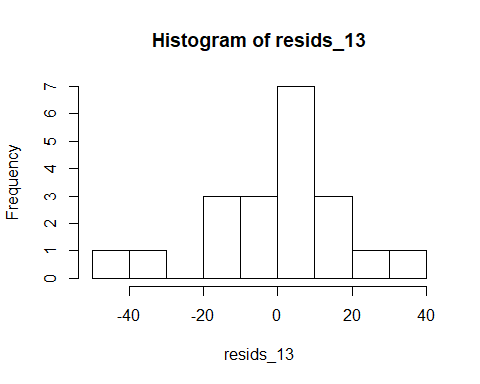
The original recipe and the New Recipe 2 differ in ratings.

## Question 13

Question13 <- read.csv("Question13.csv")  
  
## Test for normality  
  
stack\_13 <- melt(Question13, id.vars = c("Year"), measure.vars = c("Growth", "Value"))

## Warning in melt(Question13, id.vars = c("Year"), measure.vars = c("Growth", :  
## The melt generic in data.table has been passed a data.frame and will attempt  
## to redirect to the relevant reshape2 method; please note that reshape2 is  
## deprecated, and this redirection is now deprecated as well. To continue using  
## melt methods from reshape2 while both libraries are attached, e.g. melt.list,  
## you can prepend the namespace like reshape2::melt(Question13). In the next  
## version, this warning will become an error.

Year <- stack\_13$Year  
Investment <- stack\_13$variable  
Annual\_Return <- stack\_13$value  
  
model\_13 <- lm(Annual\_Return ~ Year + Investment)  
resids\_13 <- residuals(model\_13)  
predicts\_13 <- predict(model\_13)  
  
hist(resids\_13)



nortest::ad.test(resids\_13)

##   
## Anderson-Darling normality test  
##   
## data: resids\_13  
## A = 0.38241, p-value = 0.3647

## The data is normally distributed

## Test for variance  
  
bartlett.test(Annual\_Return ~ Investment)

##   
## Bartlett test of homogeneity of variances  
##   
## data: Annual\_Return by Investment  
## Bartlett's K-squared = 0.089085, df = 1, p-value = 0.7653

## The test is satisfied

cor.test( as.numeric(Investment), as.numeric(Annual\_Return))

##   
## Pearson's product-moment correlation  
##   
## data: as.numeric(Investment) and as.numeric(Annual\_Return)  
## t = -0.2902, df = 18, p-value = 0.775  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.4957906 0.3859335  
## sample estimates:  
## cor   
## -0.0682419

#### Answer:

The performance of the returns of these two funds were NOT similar over the 2007-2016 timeframe.

## Question 14

Question14 <- read.csv("Question14.csv")  
  
cor.test(x = Question14$JOBLOSE, y = Question14$HRS1, method = "spearman")

## Warning in cor.test.default(x = Question14$JOBLOSE, y = Question14$HRS1, :  
## Cannot compute exact p-value with ties

##   
## Spearman's rank correlation rho  
##   
## data: Question14$JOBLOSE and Question14$HRS1  
## S = 151103791, p-value = 0.002339  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## 0.09609079

#### Answer:

There is a weak positive correlation between long hours worked and chances of losing one’s job.

## Question 15

Question15 <- read.csv("Question15.csv")  
  
SIGN.test(x = Question15$European,  
 y = Question15$Domestic,  
 alternative = "greater",  
 conf.level = 0.90)

##   
## Dependent-samples Sign-Test  
##   
## data: Question15$European and Question15$Domestic  
## S = 30, p-value = 0.000236  
## alternative hypothesis: true median difference is greater than 0  
## 90 percent confidence interval:  
## 0 Inf  
## sample estimates:  
## median of x-y   
## 0.5   
##   
## Achieved and Interpolated Confidence Intervals:   
##   
## Conf.Level L.E.pt U.E.pt  
## Lower Achieved CI 0.8775 0 Inf  
## Interpolated CI 0.9000 0 Inf  
## Upper Achieved CI 0.9225 0 Inf

#### Answer:

The data provide sufficient evidence to infer that the European brand is perceived to be more effective than Domestic brand as the p-value is less than 0.90.

## Question 16

Question16 <- read.csv("Question16.csv")  
   
stack\_\_16 <- melt(data = Question16, id.vars = c("Title"), measure.vars = c("Amazon", "BN"))

## Warning in melt(data = Question16, id.vars = c("Title"), measure.vars =  
## c("Amazon", : The melt generic in data.table has been passed a data.frame and  
## will attempt to redirect to the relevant reshape2 method; please note that  
## reshape2 is deprecated, and this redirection is now deprecated as well. To  
## continue using melt methods from reshape2 while both libraries are attached,  
## e.g. melt.list, you can prepend the namespace like reshape2::melt(Question16).  
## In the next version, this warning will become an error.

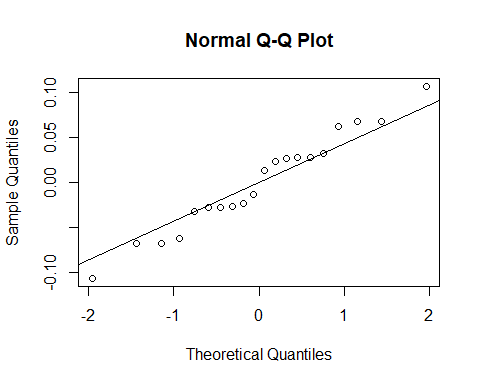
Title <- stack\_\_16$Title  
Retailer <- stack\_\_16$variable  
Cost <- stack\_\_16$value  
  
model\_16 <- lm(Cost ~ Retailer + Title)  
resids\_16 <- residuals(model\_16)  
predict\_16 <- predict(model\_16)  
  
## Normality Tests  
  
nortest::ad.test(resids\_16)

##   
## Anderson-Darling normality test  
##   
## data: resids\_16  
## A = 0.29375, p-value = 0.5649

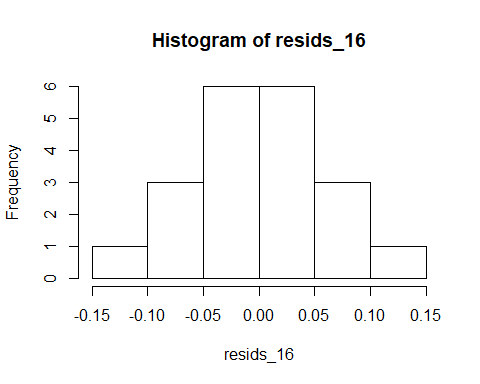
shapiro.test(resids\_16)

##   
## Shapiro-Wilk normality test  
##   
## data: resids\_16  
## W = 0.97321, p-value = 0.8207

qqnorm(resids\_16)  
qqline(resids\_16)



hist(resids\_16)



## Variance Tests  
  
bartlett.test(Cost ~ Retailer)

##   
## Bartlett test of homogeneity of variances  
##   
## data: Cost by Retailer  
## Bartlett's K-squared = 2.2269e-07, df = 1, p-value = 0.9996

car::leveneTest(Cost ~ Retailer)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 1 1e-04 0.9912  
## 18

### Constant variance requirement satisfied for Diet  
  
ANOVA\_16 <- aov(Cost ~ Retailer + Title)  
summary(ANOVA\_16)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Retailer 1 0.1 0.09 14.57 0.00411 \*\*   
## Title 9 394.2 43.80 6901.73 3.51e-16 \*\*\*  
## Residuals 9 0.1 0.01   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

TukeyHSD(ANOVA\_16, which = 'Retailer', ordered = TRUE)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
## factor levels have been ordered  
##   
## Fit: aov(formula = Cost ~ Retailer + Title)  
##   
## $Retailer  
## diff lwr upr p adj  
## BN-Amazon 0.136 0.05540453 0.2165955 0.0041072

#### Answer:

The 𝑝-values for both Retailer and Title are statistically significant. So, we conclude that there is a difference among the 2 retailers AND blocking was meaningful.

## Question 17

: The data comes from a normal distribution. : The data does not follow a normal distribution.

observed\_17 <- c(10, 18, 48, 16, 8)  
prob\_1 <- pnorm(-1.5)  
prob\_2 <- pnorm(-0.5) - pnorm(-1.5)  
prob\_3 <- pnorm(0.5) - pnorm(-0.5)  
prob\_4 <- pnorm(1.5) - pnorm(0.5)  
prob\_5 <- pnorm(1.5, lower.tail = FALSE)  
  
probability\_17 <- c(prob\_1, prob\_2, prob\_3, prob\_4, prob\_5)  
  
chisq.test(x = observed\_17, p = probability\_17)

##   
## Chi-squared test for given probabilities  
##   
## data: observed\_17  
## X-squared = 8.7104, df = 4, p-value = 0.06876

#### Answer:

As p-value is less than 0.05, we fail to reject i.e. we can conclude that the data comes from a normal distribution.