STAT291-Assignment-3

2355170 - Abdullah Burkan Bereketoglu

1/18/2022

Contents

Introduction

Welcome to my homework!!!

In this homework I, Burkan, prepared a nice pdf to show all the parts of the questions in different parts and so.

Question 1

In this question we are expected to read a data set from a .txt file by keeping their respective column names, then turning these names into respective gender's name such as "age.male", "age.female".

Later on we will continue with plotting the separate gender's with their traits (age, fat) with proper plot properties, then calculate correlation between age and fat for both male and females and these values should be written on the plot.

Lastly, we will continue with bringing both plots in one page with par() function.

Question 1.A

Here we will do the separation of male and female and naming the column names to further make the plot in part B.

```
age_fat =
  read.table(
    "C:/users/Wilkins Inc/OneDrive/Desktop/hw3/agefat.txt",
                               header= TRUE,
                               stringsAsFactors = FALSE)

males = age_fat[which(age_fat$Gender == "m"),]
females = age_fat[which(age_fat$Gender == "f"),]

names(males) = c("age.male","fat.male","gender.male")
names(females) = c("age.female","fat.female","gender.female")
```

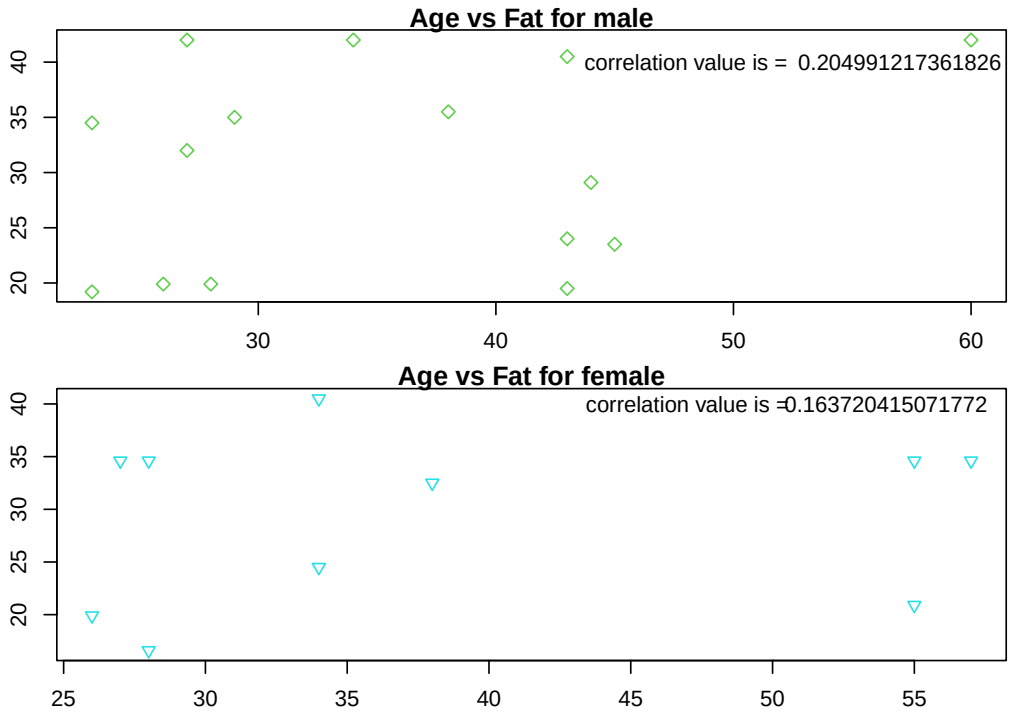
The deed is done in the abovementioned code by the which function of the R language.

Question 1.B

Here we will continue with the scatter-plot and correlation calculation between age and fat for both males and females which will be written on their plots.

Here we will also show both plots in one page.

```
cor.male = cor(males$age.male, males$fat.male)
cor.female = cor(females$age.female, females$fat.female)

par(mfrow = c(2,1),mar = c(2.4, 2.4, 1, 1), cex = 0.8)

plot(males$age.male, males$fat.male,
     main = "Age vs Fat for male",
     xlab = "fat of male",
     ylab = "age of male",
     col = 3,
     pch = 5)
text(48,40,"correlation value is =", cex = 1)
text(57,40, cor.male, cex = 1)

plot(females$age.female, females$fat.female,
     main = "Age vs Fat for female",
     xlab = "fat of female",
     ylab = "age of female",
     col = 5,
     pch = 6)
text(47,40,"correlation value is =", cex = 1)
text(54,40, cor.female, cex = 1)
```



Results of the correlation function is really low so we can say there is no or really low correlation between fat and age both in female and male.

Question 2

In this question we are firstly going to read the two .csv files named (USD_TRY and EUR_TRY) to R. Which shows daily currency exchange prices for last 6 months. First variable in the data set is Date and the second is the highest price that happened in that day.

What is going to be done is to make two line graphs one for each that shows compared to TRY value as USD/TRY, EUR/TRY or x/TRY and so. also there should be different colors, proper titling legend and so.

```
usd_try =
  read.csv(
    "C:/users/Wilkins Inc/OneDrive/Desktop/hw3/USD_TRY.csv",
                    header = FALSE,
                    dec = ",")

eur_try =
  read.csv(
    "C:/users/Wilkins Inc/OneDrive/Desktop/hw3/EUR_TRY.csv",
                    header = FALSE,
                    dec = ",")

names(usd_try) = c("Date", "ExchangeRate")
names(eur_try) = c("Date", "ExchangeRate")

usd_try$Date = as.Date(usd_try$Date, "%d.%m.%Y")
eur_try$Date = as.Date(eur_try$Date, "%d.%m.%Y")

head(usd_try, n = 5)
```

```
##         Date ExchangeRate
## 1 2022-01-12      138.543
## 2 2022-01-11      139.636
## 3 2022-01-10      140.410
## 4 2022-01-07      140.133
## 5 2022-01-06      139.521
```

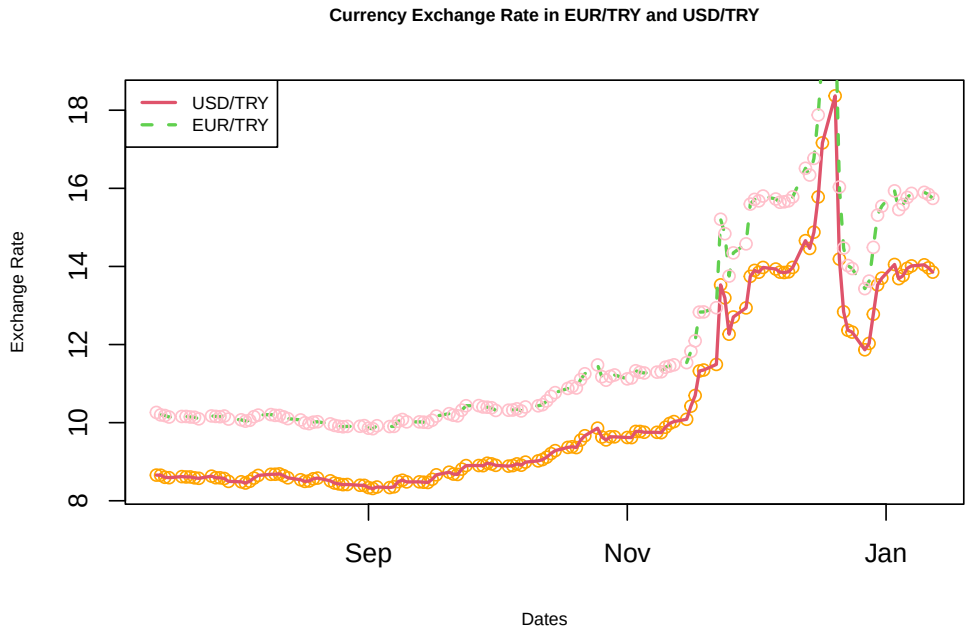
```
head(eur_try, n = 5)
```

```
##         Date ExchangeRate
## 1 2022-01-12      157.422
## 2 2022-01-11      158.428
## 3 2022-01-10      158.975
## 4 2022-01-07      158.704
## 5 2022-01-06      157.670
```

As it is seen from the data sets for both usd_try and eur_try exchange rates are given the exchange rate of that day multiplied by 10 so to find the correct exchange rate we need to divide it by 10.

```
usd_try$ExchangeRate = as.numeric(usd_try$ExchangeRate)

eur_try$ExchangeRate = as.numeric(eur_try$ExchangeRate)

usd_try$ExchangeRate = usd_try$ExchangeRate/10

eur_try$ExchangeRate = eur_try$ExchangeRate/10

plot(usd_try$Date, usd_try$ExchangeRate,
     main = "Currency Exchange Rate in EUR/TRY and USD/TRY",
     xlab =  "Dates",
     ylab = "Exchange Rate",
     cex.main = 0.7,
     cex.lab = 0.7,
     col = "orange")
lines(usd_try$Date,usd_try$ExchangeRate,
```

```
      lty = 1,
      col = 2,
      lwd = 2,
      cex = 0.7)
lines(eur_try$Date,eur_try$ExchangeRate,
      lty = 2,
      col = 3,
      lwd = 2,
      cex = 0.7)
points(eur_try$Date,
       eur_try$ExchangeRate,
       col = "pink")
legend("topleft",
       legend = c("USD/TRY","EUR/TRY"),
       lty = c(1,2),
       col = c(2,3),
       lwd = 2,
       bty = "o",
       cex = 0.7)
```


Currency Exchange Rate in EUR/TRY and USD/TRY
USD/TRY
EUR/TRY
Exchange Rate
8
10
12
14
16
18
Sep
Nov
Jan
Dates

Question 3

In this question, the question givs us details on human pregnancies length from start which is conception to birth approximates a normal distribution with mean of 266 days and its standard deviation of 16 days.

Question 3.A

In the first part of the given question with mean 266 and std 16 what is the probability for a pregnant person to have their pregnancy last between 240 and 270 days?

```
q3.a.result <-
  pnorm(240,
        mean = 266,
        sd = 16,
        lower.tail = FALSE)-pnorm(270,
                                  mean = 266,
                                  sd = 16,
                                  lower.tail = FALSE)

q3.a.result
```

```
## [1] 0.546625
```

Here we get the result of 0.546625 which indicates half of the time the pregnancy can be in the range of 240 to 270 days.

Question 3.B

In the first part of the given question with mean 266 and std 16 what is the probability for a pregnant person to have their pregnancy last more than 250 days?

```
q3.b.result <- pnorm(250,mean = 266, sd = 16,
                     lower.tail = FALSE)

q3.b.result
```

```
## [1] 0.8413447
```

Here for the 3.B we see that pregnancy to last more than 250 days is highly possible with the rate of 84% or 0.8413557 as the result.

Question 3.C

In the first part of the given question with mean 266 and std 16 what is the probability for a pregnant person to have their pregnancy last less than 235 days?

```
q3.c.result <- pnorm(235,mean = 266, sd = 16,
                     lower.tail = TRUE)

q3.c.result
```

```
## [1] 0.02634213
```

In the last part of the question we end up with the result of 0.02634213 for pregnancy in human to last less than 235. Which indicates that pregnancy most of the time take more than 235 and this may be an indicator of some issues with mother or the baby since these probabilities show that being less than 235 is somewhat exceptional in medicine since p for medical studies most of the time $p<0.05$, so that may indicate some problems (but not always and we cannot know it just by looking at the probability).

Question 4

In this question, we the students are expected to write hypothesis for each part, check if the hypothesis at certain confidence level is significant by providing the outputs, then interpreting them.

Question 4.A

Here we are dealing with a Salmonella-related illness which is attributed to a factory that produces ice cream. Scientists then measured the level of Salmonella in 12 randomly sampled batches of the ice cream from the factory.

To find the result first we set a seed that will be our student ID with last 4 digits of it. My student ID is 2355170, so last 4 digits are 5170.

Let's do it step by step.

Question 4.a.1

Firstly, we will do comparison with critical value and after the results comment on the result. Since our null hypothesis is mean being 0.35 MPN/g and trying to find an evidence if it's higher than the 0.35 we will pick mu_0 as 0.35, our alpha level is 0.05 and df is 12-1 which is 11. If the mean of our sample in t score is bigger than t critical that will mean that our mean is bigger than the 0.35.

Basically what we are trying to do is to see that if we can reject the null hypothesis(H0) and make our hypothesis(H1) that is mean level of Salmonella is higher than 0.35 MPN/g will be accepted.

```r
seed <- 5170 # last 4 digits of your student ID
set.seed(seed)
x <- rnorm(12, 0.45, 0.25)

mu_0 = 0.35
alpha = 0.05
degrees.of.freedom = length(x) - 1
mean(x)
```

```
## [1] 0.5060549
```

```r
t.score =
  qt(p = alpha, df = degrees.of.freedom,
             lower.tail = F)

cat("the critical t value for our function",
      "at one tail with 11 degrees of freedom",
      "is ",
    round(t.score,4),
    ".",
    sep = "\n")
```

```
## the critical t value for our function
## at one tail with 11 degrees of freedom
## is
## 1.7959
## .
```

```r
t_score = (mean(x)-mu_0)/(sd(x)/sqrt(12))

cat("the t value that we want to test for",
      "our function at one tail with",
      "11 degrees of freedom is ",
    round(t_score,4),
    ".",
    sep = "\n")
```

```
## the t value that we want to test for
## our function at one tail with
## 11 degrees of freedom is
## 1.9694
## .
```

```
if(t_score - t.score > 0){
  cat("The null hypothesis is rejected",
        "and the mean level of Salmonella",
        "is more than 0.35 MPN/g.", sep = "\n")
} else if(t_score < 0 && abs(t_score) - t.score > 0){
  cat("The null hypothesis is rejected",
        "and the mean level of Salmonella",
        "is less than 0.35 MPN/g.", sep = "\n")
} else{
  cat("The null hypothesis is not rejected",
        "and there is no evidence to change",
        "the mean from 0.35 MPN/g .", sep = "\n")
}
```

```
## The null hypothesis is rejected
## and the mean level of Salmonella
## is more than 0.35 MPN/g.
```

Here we see that our mean result which is 0.5060549 for seed 5170 is evidently can reject the null hypothesis and accept our alternative hypothesis which is mean level of Salmonella is more than 0.35 MPN/g. But now let's look at from the perspective of P-value, since we now did Critical Value test.

Question 4.a.2

Here in this part of the question we will test our null hypothesis, if we can reject it at $p<0.05$ level since our alpha is given as 0.05. Now let's look at the P-Value of the given data sample.

```r
p_value = pt(t_score, degrees.of.freedom, lower.tail = FALSE)

cat("This is t score",t_score,
    "This is t critical",t.score,
    "This is p value",p_value,
    "This is alpha level", alpha, sep = "\n")
```

```
## This is t score
## 1.969369
## This is t critical
## 1.795885
## This is p value
## 0.03730602
## This is alpha level
## 0.05
```

```r
if(t_score > t.score && p_value < alpha){
  cat("The null hypothesis is rejected",
        "and the mean level of Salmonella",
        "is more than 0.35 MPN/g.", sep = "\n")
} else if(t_score < t.score && p_value < alpha){
  cat("The null hypothesis is rejected",
      "and the mean level of Salmonella",
      "is less than 0.35 MPN/g.", sep = "\n")
} else{
  cat("The null hypothesis is not rejected",
        "and there is no evidence to change",
        "the mean from 0.35 MPN/g .", sep = "\n")
}
```

```
## The null hypothesis is rejected
## and the mean level of Salmonella
## is more than 0.35 MPN/g.
```

Since we have the t_score from the formula applied, we can find p_value from the t_score that we achieved. Then we can make a comparison since p_value can be

also below alpha level in the other tail we should also analyze the t_score with t critical score so that we can see if the result indicates more than or less than from our tested hypothesis.

From the printed values we can see that t score is higher than t critical and p value is below alpha 0.05 level hence we can both reject the hypothesis and can say that the mean level is more than 0.35 MPN/g.

Question 4.a.3

Lastly, let's compare our results that we achieved from the P-value comparison and Critical Value comparison.

```
result.t = t.test(x,alternative = "greater" ,
                  mu = 0.35,
                  conf.level = 0.95)

result.t
```

```
##
##  One Sample t-test
##
## data:  x
## t = 1.9694, df = 11, p-value = 0.03731
## alternative hypothesis: true mean is greater than 0.35
## 95 percent confidence interval:
##  0.3637471       Inf
## sample estimates:
## mean of x
## 0.5060549
```

```
if(round(result.t$p.value,5)
   == round(p_value,5)
   &&
   round(result.t$statistic,4) ==
   round(t_score,4)){
  cat("Our findings fit to the t-test function result.")
```

```
} else{
  cat("Our findings differ from the
      t-test function hence may have some errors.")
}
```

```
## Our findings fit to the t-test function result.
```

From the resulting information we can deduce that our results fit to the t-test and it also indicates that our null hypothesis is rejected and the mean level is greater than the 0.35 MPN/g .

Question 4.B

In this part of the question we are dealing with a researchers claim of a certain type of training that is applied to track athletes their track time is improved. For that this researcher gathered 15 volunteer athletes and recorded their times before the training and after few months of that the researcher applied the training program.

So let's find if there is a significant improvement in mean track time. We will look at confidence level of 0.95 which is alpha 0.05. will look if its less since the improved time will be lower than the previous track time (due to more speed you will cover more grounds per seconds).

```
before <- c(101.32, 108.00, 102.08, 95.07, 101.85,
101.87, 99.41, 100.25, 97.94, 106.18,
103.45, 98.90, 96.49 , 92.64, 100.64)

after <- c(93.99, 96.05, 110.36, 105.65, 95.82,
89.15, 94.15, 100.74, 99.14, 103.96,
106.13, 95.42, 102.63, 93.95, 96.22)

mean.before = mean(before)
mean.after = mean(after)

mu_0_b = mean.before
alpha = 0.05
```

```r
degrees.of.freedom.b = length(before) - 1

t.score_b =
  qt(p = alpha, df = degrees.of.freedom.b,
                lower.tail = T)

t_score_b =
  (mean.after-mu_0_b)/(sd(after)/sqrt(length(after)))

p_value_b =
  pt(t_score_b, degrees.of.freedom.b, lower.tail = TRUE)


cat("This is t score",t_score_b,
    "This is t critical",t.score_b,
    "This is p value",p_value_b,
    "This is alpha level", alpha, sep = "\n")
```

```
## This is t score
## -1.007282
## This is t critical
## -1.76131
## This is p value
## 0.1654463
## This is alpha level
## 0.05
```

```r
if(t_score_b > t.score_b && p_value_b < alpha){
  cat("The null hypothesis is rejected",
        "and the mean level of Salmonella",
        "is more than ", mean.before,
      ".", sep = "\n")
} else if(t_score_b < t.score_b && p_value_b < alpha){
  cat("The null hypothesis is rejected",
      "and the mean level of Salmonella",
      "is less than ",mean.before ,
```

```
      ".",sep = "\n")
} else{
  cat("The null hypothesis is not rejected",
        "and there is no evidence to change",
        "the mean from ",mean.before ,
      ".", sep = "\n")
}
```

```
## The null hypothesis is not rejected
## and there is no evidence to change
## the mean from
## 100.406
## .
```

Since our selected alpha level is 0.05, from the results we can deduce that p is not below 0.05 and it is actually 0.1654463 which is not even acceptable at 0.1 level. So this indicates that we cannot reject H0 and say this training program has a significant improvement on the track times of athletes' .

Question 4.C

TO a herd of 25 dairy cows a feeding test is made to compare two diets, A and B. A sample of 13 cows from this 25 dairy cows are randomly selected from the herd are fed diet A and the remaining 12 cows are fed with diet B. From the observations made over a three-week period, average daily milk production (in L) and they are recorded for each cow. These two samples are assumed to have equal variances; now we will try to investigate any evidence of difference in true mean milk yields for the two diets.

```
diet_A = c(44,44,56,46,47,38,58,53,49,35,46,30,41)
diet_B = c(35,47,55,29,40,39,32,41,42,57,51,39)

sd.a = sd(diet_A)
sd.b = sd(diet_B)
```

```r
degrees.of.freedom.c = length(diet_A)-1 + length(diet_B)-1

s_pooled =
  (((length(diet_A)-1)*sd.a^2) + ((length(diet_B)-1)*sd.b^2))/
  (length(diet_A)+length(diet_B)-2)

t_score_c =
  (mean(diet_A)-mean(diet_B))/
  sqrt((s_pooled/length(diet_A))+
         (s_pooled/length(diet_B)))


t.score_c =
  qt(p = alpha, df = degrees.of.freedom.c,
     lower.tail = F)


p_value_c =
  pt(t_score_c,
     degrees.of.freedom.c,
     lower.tail = FALSE)*2
# since two tail


cat("This is t score",t_score_c,
    "This is t critical",t.score_c,
    "This is p value",p_value_c,
    "This is alpha level", alpha, sep = "\n")
```

```
## This is t score
## 0.8675501
## This is t critical
## 1.713872
## This is p value
## 0.3946023
## This is alpha level
## 0.05
```

```r
if(t_score_c > t.score_c && p_value_c < alpha){
  cat("The null hypothesis is rejected",
        "true difference in means",
        " is not equal to 0 ",
      ".", sep = "\n")
} else if(t_score_c < t.score_c && p_value_c < alpha){
  cat("The null hypothesis is rejected",
        "true difference in means",
        " is not equal to 0 ",
      ".", sep = "\n")
} else{
  cat("The null hypothesis is not rejected",
        "true difference in means",
        " is equal to 0 ",
      ".", sep = "\n")
}
```

```
## The null hypothesis is not rejected
## true difference in means
##  is equal to 0
## .
```

Since our selected alpha level is 0.05, from the results we can deduce that p is not below 0.05 and it is actually 0.3946023 which is not even acceptable at 0.1, 0.2 for two-tail. So this indicates that we cannot reject H0 and say there is no evidence of a difference in true mean milk yields for the two diets.

Question 5

Here in this question there will be some statistical answers to the data set of "Carseats" dat is under "ISLR" package in R. Also we should use response variable as Sales.

Question 5.A

Here in this part we will fit a multiple regression model to predict Sales by using Price. With using summary() function, we will look at the information about our model, later on continue with commenting our results.

```
library(ISLR)

data(Carseats)

model_1 <- lm(Sales ~ Price, data = Carseats)
summary(model_1)
```

```
##
## Call:
## lm(formula = Sales ~ Price, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5224 -1.8442 -0.1459  1.6503  7.5108
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.641915   0.632812  21.558   <2e-16 ***
## Price       -0.053073   0.005354  -9.912   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.532 on 398 degrees of freedom
## Multiple R-squared:  0.198,  Adjusted R-squared:  0.196
## F-statistic: 98.25 on 1 and 398 DF,  p-value: < 2.2e-16
```

```
#extra
head(predict(model_1,
             interval = "confidence",
             level = 0.95),
     n=5)
```

```
##        fit      lwr      upr
## 1 7.273153 7.020328 7.525978
## 2 9.236855 8.811255 9.662455
## 3 9.396074 8.944481 9.847667
## 4 8.493832 8.175864 8.811801
## 5 6.848569 6.568449 7.128689
```

As it can be seen from the Sales = 13.641915 - 0.053073 * Price is our function that gives sales value from corresponding price value. Also from the summary statistics from Pr(>|t|) of Price we can say that the estimate value which is Beta1 and is = -0.053073 is statistically significant.

The RMSE or Residual standard error (Typical Error), shows average error of our model in a sense so our model is off in average 2.5 predictions for our data set. Which is actually pretty good. This RMSE is mathematically calculated directly from sqrt(SSE/n-2)

Now we have multiple R-square, it shows us that the percentage of variation in the response variable (sales) is explained by the variation in the explanatory variable (price) could be though as R^2 = SSM/SST or 1-(SSE/SST). In a good linear model, we would want most of the time more than 0.5, but since it is 0.198 can be said that prices are not that good at explaining the variation of sales, which can also be seen in the plot in part B.

Similarly adjusted R^2 is low too, which is the R^2 method that is robust against statistically insignificant data, since R^2 will increase if we add infinitely many data the adjusted one may not. That also indicates that the power of explanation of price is not that high.

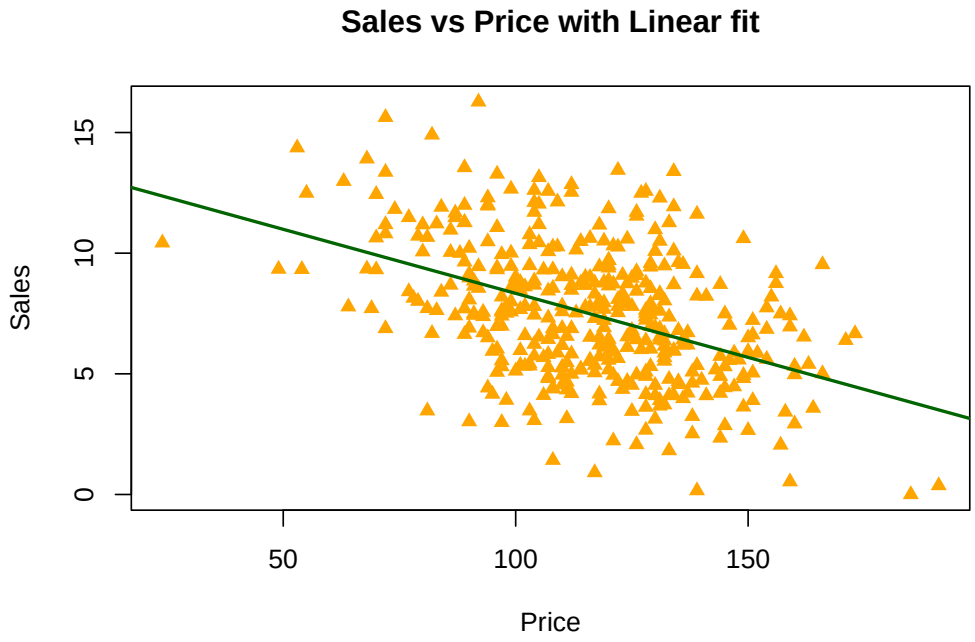
Lastly let's comment on F-statistics and p-value, F-statistics mathematically showing us that how well the overall model is doing compared to the error, which is actually (MeanSquareModel/MeanSquareError), so high values indicate that model is explaining more than compared the error if F is high then the model's explanatory power compared to error is higher than lower F values.

A p-value as all p-values if less than 0.1,0.05,0.01 in most cases 0.05 is picked shows us that overall model is significant which is good that since our model overall is significant.

Question 5.B

Here in this part we will draw a scatter plot Price vs Sales, and we will add our regression line on the plot. We will add custom title, axis labels, colors, line width for the regression line etc.

```
plot(Carseats$Sales ~ Carseats$Price,
     main = "Sales vs Price with Linear fit",
     xlab = "Price",
     ylab = "Sales",
     col = "orange",
     pch = 17)

abline(model_1, lwd = 2, col = "darkgreen")
```



As it is visually can be seen from the plot, the explanation done in part A that gave example about how the model is not explaining all the variability can be seen easily. Since data varies a lot but not in 1 dimensions but in many dimensions, it is hard for a linear model to explain them all perfectly.

Question 5.C

Here in this part we will use "US" variable as independent variable and construct a model for the again response variable, Sales.

```
model_2 <- lm(Sales ~ US, data = Carseats)
summary(model_2)
```

```
##
## Call:
## lm(formula = Sales ~ US, data = Carseats)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.497 -1.929 -0.105  1.836  8.403
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.8230     0.2335   29.21  < 2e-16 ***
## USYes         1.0439     0.2908    3.59 0.000372 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.783 on 398 degrees of freedom
## Multiple R-squared:  0.03136,    Adjusted R-squared:  0.02893
## F-statistic: 12.89 on 1 and 398 DF,  p-value: 0.0003723
```

Here again as we discussed in part A, pr(>|t|) here too shows us that our Beta1 is statistically significant.

The RMSE or Residual standard error (Typical Error), shows average error of our model in a sense so our model is off in average 2.783 predictions for our data set. Which is actually pretty good. This RMSE is mathematically calculated directly from sqrt(SSE/n-2)

Now we have multiple R-square, it shows us that the percentage of variation in the response variable (sales) is explained by the variation in the explanatory variable (price) could be though as R^2 = SSM/SST or 1-(SSE/SST). In a good linear model, we would want most of the time more than 0.5, but since it is 0.03136 can be said that US is not that good at explaining the variation of sales.

Similarly adjusted R^2 is low too, which is the R^2 method that is robust against statistically insignificant data, since R^2 will increase if we add infinitely many data the adjusted one may not. That also indicates that the power of explanation of price is not that high.

Lastly let's comment on F-statistics and p-value, F-statistics mathematically showing us that how well the overall model is doing compared to the error, which is actually (MeanSquareModel/MeanSquareError), so high values indicate that model is explaining more than compared the error if F is high then the model's explanatory power compared to error is higher than lower F values. This F statistic is way

lower than Sales with Price model but still it is below 0.05 so overall model is significant, but explanatory power is low.