Regression Assignment 2

NWUDO CHIKAEZE FIDELIS JUNIOR / student number:202290064

2022-10-20

QUESTION 1

Experience with a certain type of plastic indicates that a relation exist between the hardness (in Brinell units) of items molded from the plastic (y) and the elapsed time (in hours) since termination of the molding process (x). Data on x and y from 16 items molded from 16 batches of plastic can be found on the D2L shell for Stat 3521. It is proposed to study the relation between x and y by means of regression analysis. Assume that the simple linear regression model is appropriate for this data.

(a) Construct a 98% confidence band for the estimated regression line. Construct a scatterplot of the data. Overlay a plot of the estimated regression line and a plot of the 98% confidence band on the scatterplot of the data. Determine the boundary values of the 98% confidence band for the regression line if the elapsed time is 29 hours 30minutes.

```
mydata <- read.table("C:/Users/HP PAVILION/Desktop/As2Prob1 Data.txt", header
= F)
x <- mydata$V2
y <- mydata$V1
xmn<-min(x)
xmx<-max(x)
ymn<-min(y)
ymx<-max(y)
plot(x,y, xlim = c(xmn,xmx), ylim = c(ymn,ymx), xlab = "elapsed time", ylab =
"hardness", col=1)
fit <- lm(y~x, data = mydata)
fit
```

```
##
## Call:
## lm(formula = y ~ x, data = mydata)
##
## Coefficients:
## (Intercept)            x
##     168.600        2.034
```

```
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x, data = mydata)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -5.1500 -2.2188  0.1625  2.6875  5.5750
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 168.60000    2.65702   63.45  < 2e-16 ***
## x             2.03438    0.09039   22.51 2.16e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.234 on 14 degrees of freedom
## Multiple R-squared:  0.9731, Adjusted R-squared:  0.9712
## F-statistic: 506.5 on 1 and 14 DF,  p-value: 2.159e-12
```

```
lines(x,fitted(fit),col=4)
CIs <- predict(fit, interval = "confidence", level = 0.98)
CIs <- data.frame(CIs)
class(CIs)
```

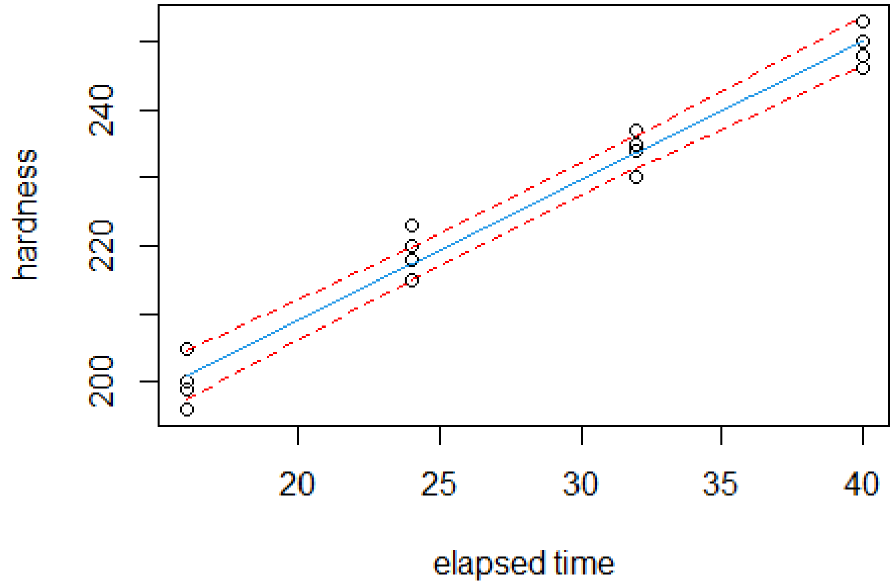
```
## [1] "data.frame"
```

```
CIs
```

```
##        fit      lwr      upr
## 1  201.150 197.5993 204.7007
## 2  201.150 197.5993 204.7007
## 3  201.150 197.5993 204.7007
## 4  201.150 197.5993 204.7007
## 5  217.425 215.1006 219.7494
## 6  217.425 215.1006 219.7494
## 7  217.425 215.1006 219.7494
## 8  217.425 215.1006 219.7494
## 9  233.700 231.3756 236.0244
## 10 233.700 231.3756 236.0244
## 11 233.700 231.3756 236.0244
## 12 233.700 231.3756 236.0244
## 13 249.975 246.4243 253.5257
## 14 249.975 246.4243 253.5257
## 15 249.975 246.4243 253.5257
## 16 249.975 246.4243 253.5257
```

```
CIs$V2 <- x
CIs
```

```
##        fit      lwr      upr V2
## 1  201.150 197.5993 204.7007 16
## 2  201.150 197.5993 204.7007 16
## 3  201.150 197.5993 204.7007 16
## 4  201.150 197.5993 204.7007 16
## 5  217.425 215.1006 219.7494 24
## 6  217.425 215.1006 219.7494 24
## 7  217.425 215.1006 219.7494 24
## 8  217.425 215.1006 219.7494 24
## 9  233.700 231.3756 236.0244 32
## 10 233.700 231.3756 236.0244 32
## 11 233.700 231.3756 236.0244 32
## 12 233.700 231.3756 236.0244 32
## 13 249.975 246.4243 253.5257 40
## 14 249.975 246.4243 253.5257 40
## 15 249.975 246.4243 253.5257 40
## 16 249.975 246.4243 253.5257 40

lines(CIs$V2, CIs$lwr, lty = 2, col="red")
lines(CIs$V2, CIs$upr, lty = 2, col= "red")
```



```
#the boundary values of the 98% confidence band for the regression line if
the elapsed time is 29 hours 30 minutes.
predict(fit, data.frame(x=29.5), interval = "confidence", conf.level = 0.98)

##        fit      lwr      upr
## 1 228.6141 226.8558 230.3724
```

QUESTION 1(b)

The plastic manufacturer has stated that the mean hardness should increase by 2 Brinell units per hour. Conduct a two-sided test to decide whether this standard is being satisfied. State the null and alternative hypotheses, compute the test statistic and p-value of the test.

null hypothesis Ho: B1 = 2, alternative hypothesis Ha: B1 =! 2

```
B1hat <- 2.03438
B1 <- 2
n <- 16
xbar <- sum(mydata$V2) / 16
SSxx <- sum((mydata$V2) ^ 2) - (n * (xbar^2))
SSE <- sum(resid(fit) ^ 2)
MSE <- SSE / (n-2)
VarOME <- MSE / SSxx
standarderror <- sqrt(VarOME)
to <- (B1hat - B1)/ standarderror
to

## [1] 0.3803358

t <- qt(0.99, 14)
t

## [1] 2.624494

P_value <- 2 * (1 -  pt(to, 14))
P_value

## [1] 0.7094043
```

Since the P_value(0.7094043) is greater than the significance level(0.02), we cannot reject the null hypothesis. (Alternatively, since Test statistic is less than my t-critical (0.3803358 < 2.624494), we cannot reject null hypothesis).Therefore, there is a significant relationship between time and hardness.

QUESTION 2

The number of galleys for a manuscript (x) and the dollar cost of correcting typo- graphical errors (y) in a random sample of recent orders handled by a firm specializing in technical manuscripts can be found on the D2L shell for Stat 6519. Assume that the regression model, yi = B1xi + E; is appropriate, with normally distributed independent error terms whose variance is variance = 16.

(a) Construct the ANOVA Table for the manuscript data based on the assumed model.

```
data2 <- read.table("C:/Users/HP PAVILION/Downloads/STAT6519 assignment 2
datasets/As2Prob2 Data.txt", header = T)
fit2 <- lm(data2$Y~data2$X-1)
summary(fit2)
```

```
##
## Call:
## lm(formula = data2$Y ~ data2$X - 1)
##
## Residuals:
##      1      2      3      4      5      6
##  2.501 -2.142  3.286 -0.999 -2.212  2.145
##
## Coefficients:
##         Estimate Std. Error t value Pr(>|t|)
## data2$X  17.9285     0.0577   310.7 6.56e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.535 on 5 degrees of freedom
## Multiple R-squared:  0.9999, Adjusted R-squared:  0.9999
## F-statistic: 9.653e+04 on 1 and 5 DF,  p-value: 6.555e-12
```

```
anova(fit2)
```

```
## Analysis of Variance Table
##
## Response: data2$Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## data2$X    1 620362  620362   96531 6.555e-12 ***
## Residuals  5     32       6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

QUESTION 2b

Test whether B1 = 0 against a one-sided alternative using a t-test at alpha = 0.05

null hypothesis Ho: B1 = 0,
alternative hypothesis Ha: B1 > 0

```
B1hat2 <- 17.9285
standarderror2 <- 0.0577

teststat <- B1hat2 / standarderror2
teststat
```

```
## [1] 310.7192
```

```
# since we have just one estimate, the degree of freedom is going to be n-1
t_critical <- qt(0.95, 5)
t_critical
```

```
## [1] 2.015048
```

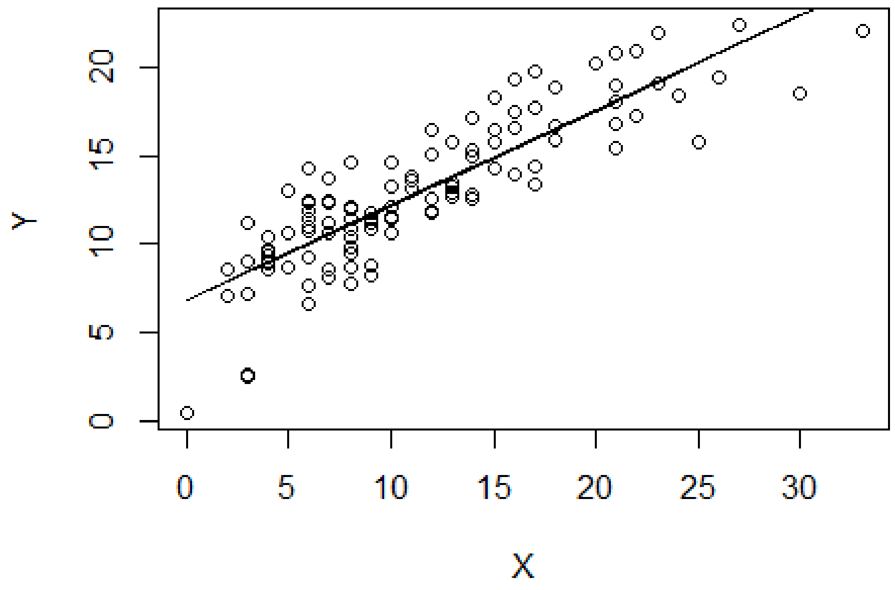
In conclusion, we reject Ho since test statistic is greater than t_critical (310.7192 > 2.015048).

QUESTION 3

In a manufacturing study, the production times for 111 recent production runs were obtained. The data is attached.

(a) Fit a linear model to the data.

```
data3<- read.table("C:/Users/HP PAVILION/Downloads/STAT6519 assignment 2
datasets/As2Prob3 Data.txt", header = T)
# n2 = number of observations
n2 <- 111
model3 <- lm(data3$Y~data3$X)
summary(model3)

##
## Call:
## lm(formula = data3$Y ~ data3$X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3535 -1.3154  0.0036  1.2405  4.2469
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.86349    0.39863   17.22   <2e-16 ***
## data3$X      0.53327    0.03028   17.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.118 on 109 degrees of freedom
## Multiple R-squared:   0.74,  Adjusted R-squared:  0.7376
## F-statistic: 310.2 on 1 and 109 DF,  p-value: < 2.2e-16

plot(data3$X, data3$Y, ylab = "Y", xlab = "X")
lines(data3$X, fitted(model3))
```

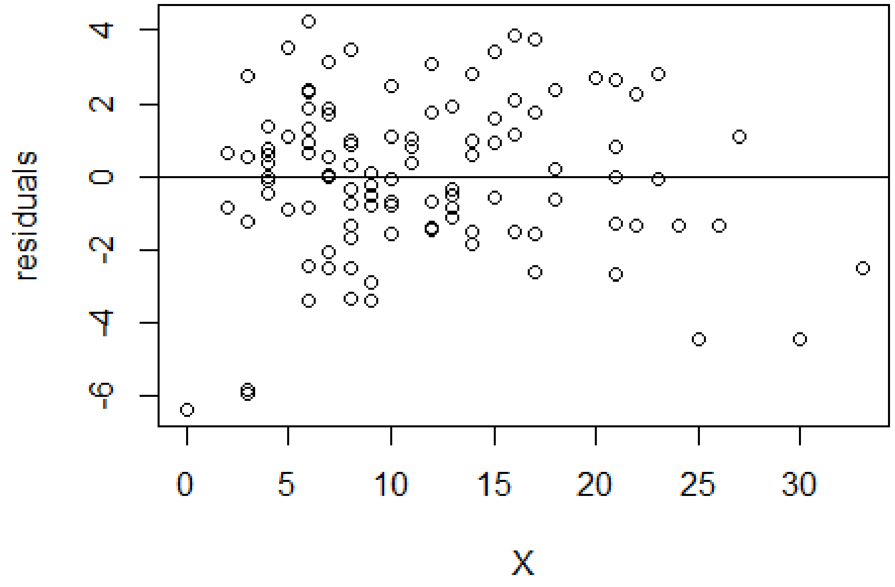


```
yHat <- 6.86349 + (0.53327 * data3$X)
```

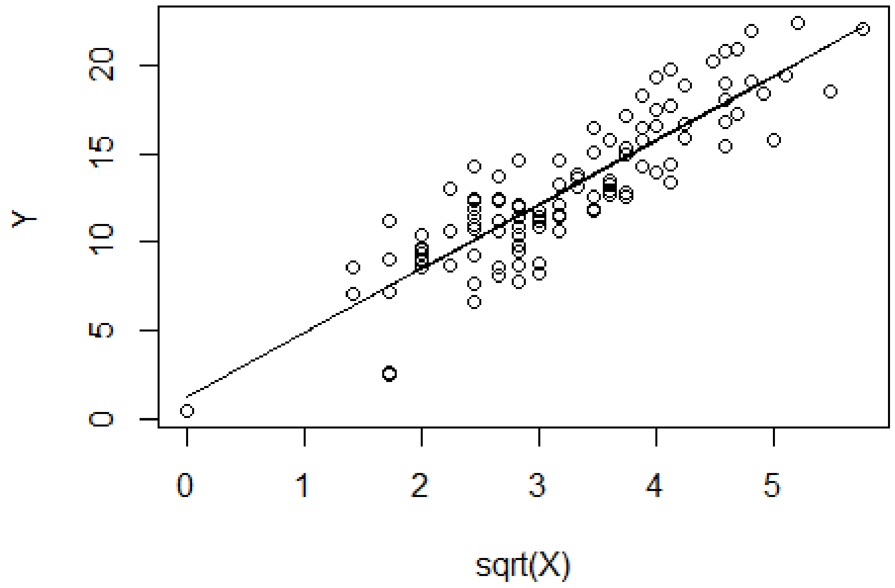
(b) Examine the appropriateness of the assumption of linearity. If the assumption of linearity is not valid, apply a suitable transformation to correct the problem.

(c) Fit a linear model to the transformed data and perform residual analysis to examine the validity of the model assumptions.

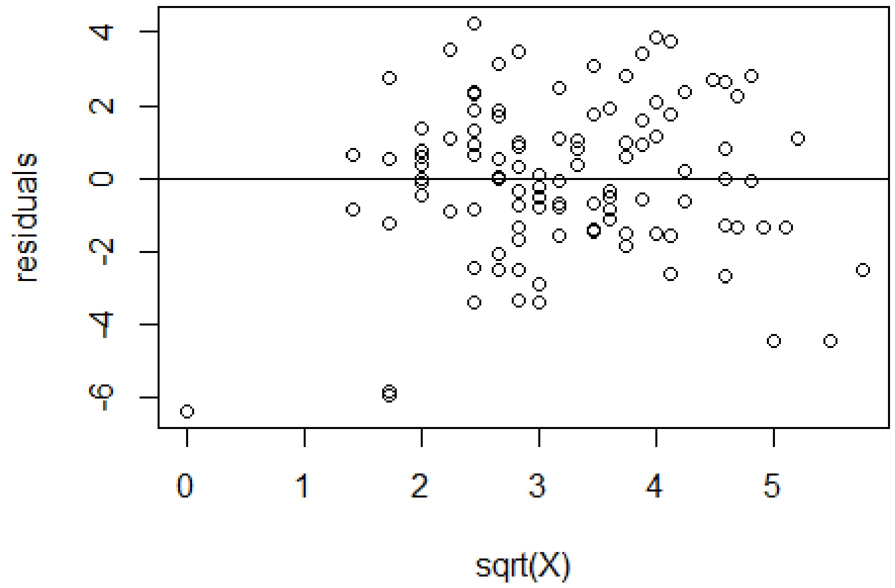
```
library(MASS)
resi <- resid(model3)
plot(data3$X, resi, xlab = "X", ylab = "residuals", abline(0,0))
```



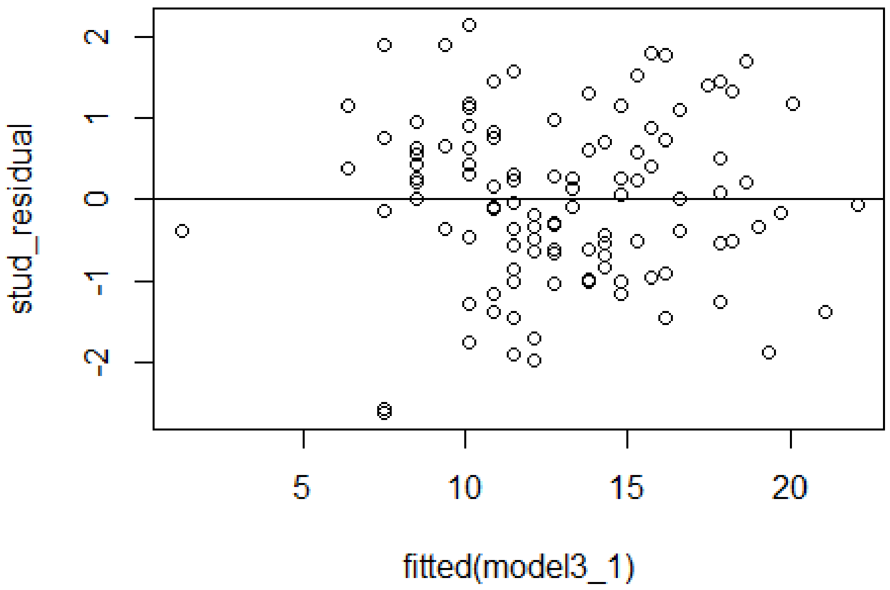
```
# square root transformation of the predictor variable

model3_1 <- lm(data3$Y~sqrt(data3$X))
summary(model3_1)

##
## Call:
## lm(formula = data3$Y ~ sqrt(data3$X))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0008 -1.2161  0.0383  1.3367  4.1795
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.2547     0.6389   1.964   0.0521 .
## sqrt(data3$X)   3.6235     0.1895  19.124   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.99 on 109 degrees of freedom
## Multiple R-squared:  0.7704, Adjusted R-squared:  0.7683
## F-statistic: 365.7 on 1 and 109 DF,  p-value: < 2.2e-16

plot(sqrt(data3$X), data3$Y, xlab = "sqrt(X)", ylab = "Y")
lines(sqrt(data3$X), fitted(model3_1))
```



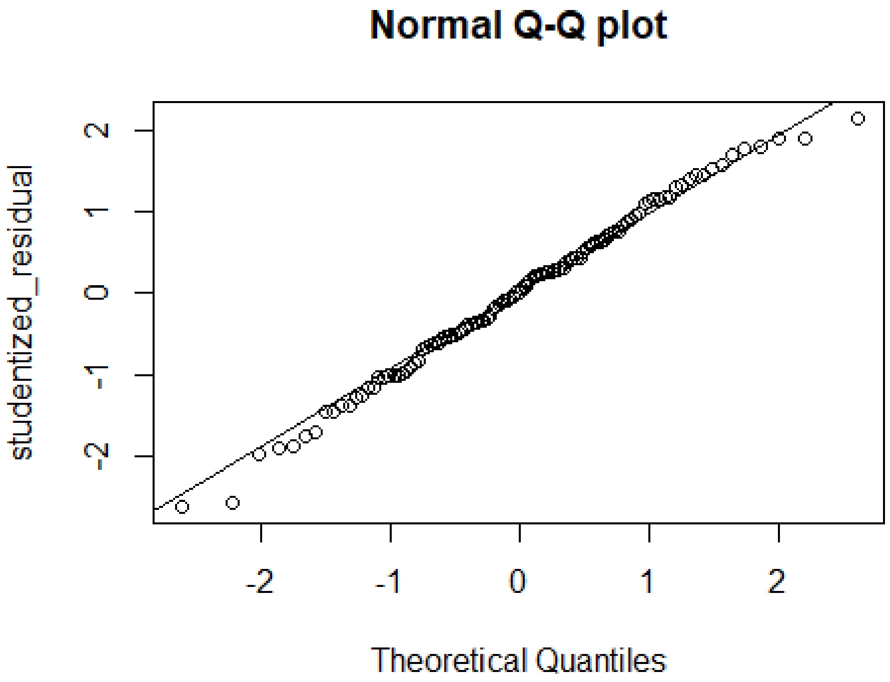
```
plot(sqrt(data3$X), resi, xlab = "sqrt(X)", ylab = "residuals", abline(0,0))
```



```
stud_residual <- studres(model3_1)
plot(fitted(model3_1), stud_residual, abline(0,0))
```



```r
qqnorm(y = studres(model3_1), main = "Normal Q-Q plot", xlab = "Theoretical
Quantiles", ylab = "studentized_residual", plot.it = TRUE)
qqline(y= studres(model3_1), distribution = qnorm)
```



After the square root transformation of the predictor axis, the semistudentized residuals fall within a horizontal band centered around the zero line, displaying no systematic pattern of positive and negative values. This shows that the assumption of linearity and constant variance is valid for these data sets.

Also, in the normal Q-Q plot, all the points fall approximately in the line. Then we can say it follows an approximate normal distribution

(d) Express the estimated regression function in the original units.

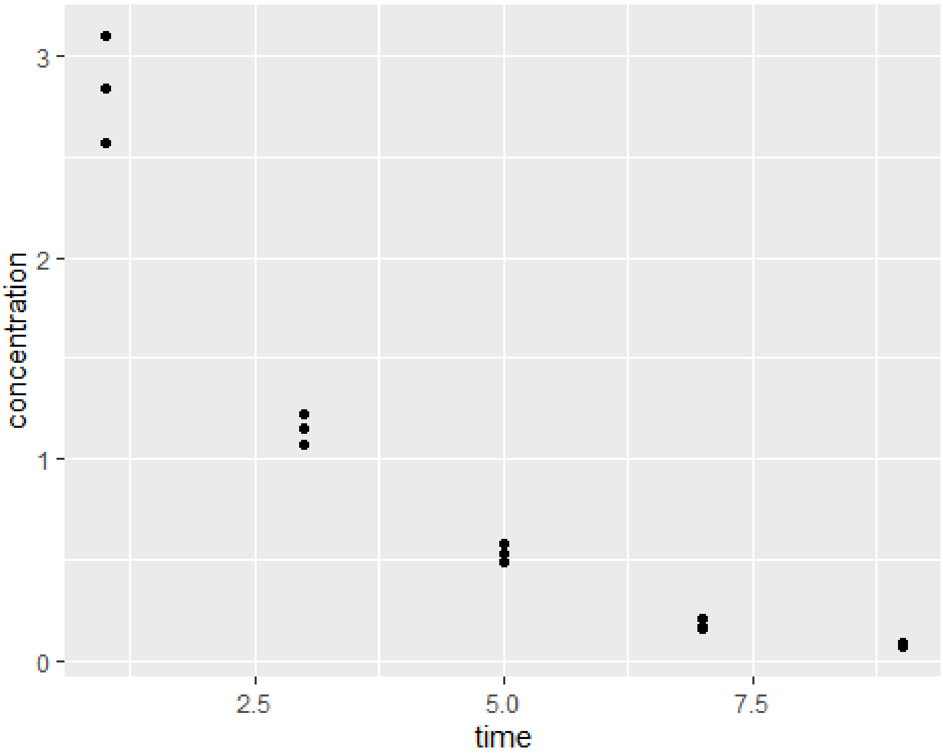
yhat = 1.2547 + (3.6235 * sqrt(X))

QUESTION 4

A chemist studied the concentration of a solution (y) over time (x). Fifteen identical solutions were prepared. The 15 solutions were randomly divided into five sets of three and the five sets were measured, respectively, after 1, 3, 5, 7 and 9 hours. The data is attached.

(a) Perform a lack of fit test to determine if a linear model is appropriate.

```
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse
1.3.2 ──
## ✔ ggplot2 3.4.0      ✔ purrr   0.3.5
## ✔ tibble  3.1.8      ✔ dplyr   1.0.10
## ✔ tidyr   1.2.1      ✔ stringr 1.4.1
## ✔ readr   2.1.3      ✔ forcats 0.5.2
## ── Conflicts ──────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ✖ dplyr::select() masks MASS::select()

library(MASS)
data4 <- read.table("C:/Users/HP PAVILION/Downloads/STAT6519 assignment 2
datasets/As2Prob4 Data.txt", header = T)
concentration <- data4$Y
time <- data4$X
ggplot(data = data4, mapping = aes(x = time, y = concentration)) +
geom_point()
```

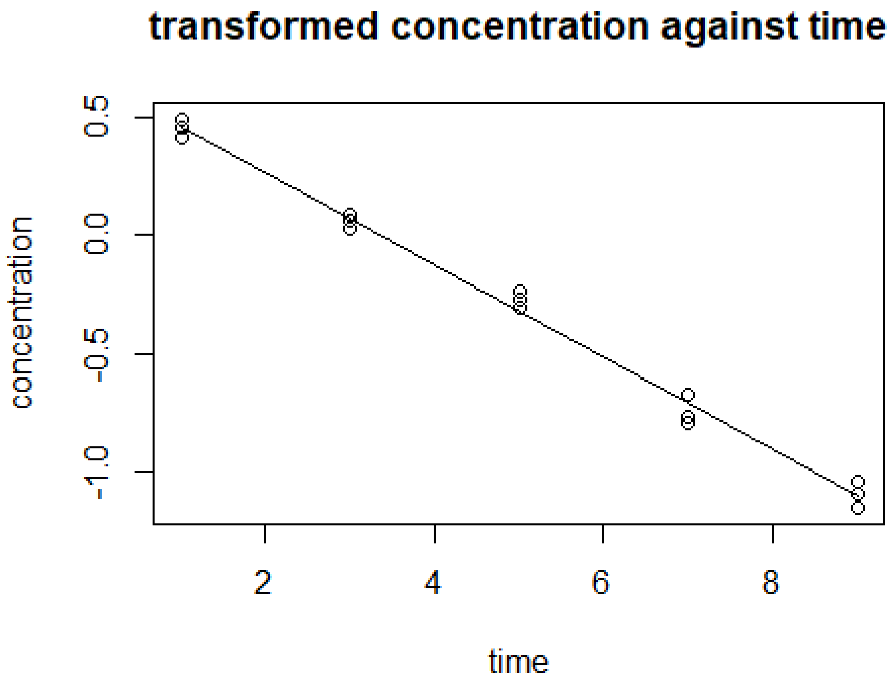


```
# fitting two regression models to the dataset

full <- lm(concentration ~ poly(time, 2), data = data4)

reduced <- lm(concentration ~ time, data = data4)

#lack of fit test
anova(full, reduced)

## Analysis of Variance Table
##
## Model 1: concentration ~ poly(time, 2)
## Model 2: concentration ~ time
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     12 0.38888
## 2     13 2.92465 -1   -2.5358 78.248 1.325e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

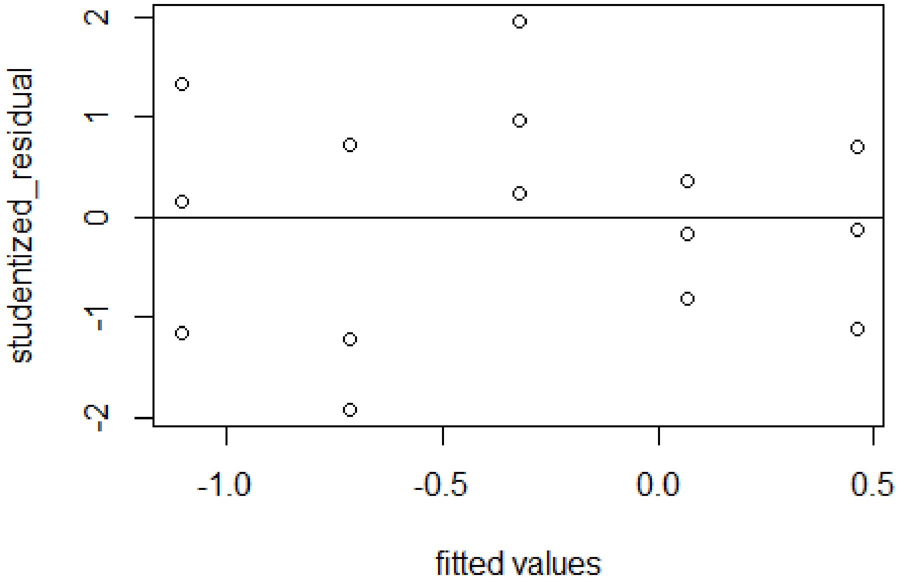
The F test-statistic turns out to be 78.248 and the corresponding p-value is 1.325e-06. Since the p-value is less than 0.05, we reject the null hypothesis of the test and conclude that the full model offers a statistically better fit than the reduced model.

(b) Apply the transformation y* = log10y. Obtain the estimated linear regression function for y* and examine the adequacy of the model.

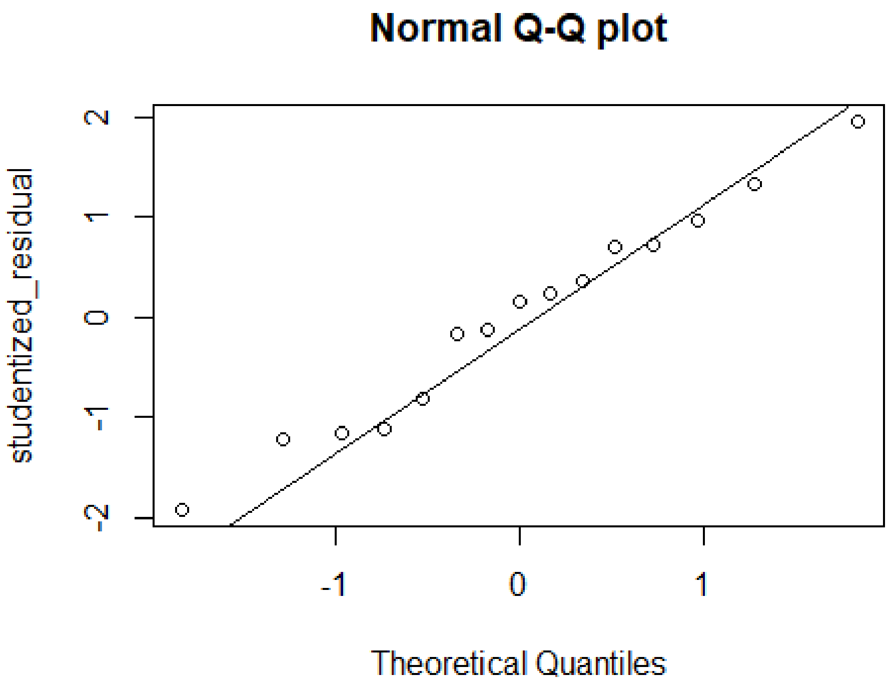
```
model4 <- lm(log10(data4$Y)~data4$X)
summary(model4)

##
## Call:
## lm(formula = log10(data4$Y) ~ data4$X)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.082958 -0.044421  0.006813  0.033512  0.085550
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.654880   0.026181   25.01 2.22e-12 ***
## data4$X     -0.195400   0.004557  -42.88 2.19e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04992 on 13 degrees of freedom
## Multiple R-squared:  0.993,  Adjusted R-squared:  0.9924
## F-statistic:  1838 on 1 and 13 DF,  p-value: 2.188e-15

plot(data4$X, log10(data4$Y), xlab = "time", ylab = "concentration", main =
"transformed concentration against time")
lines(data4$X, fitted(model4))
```



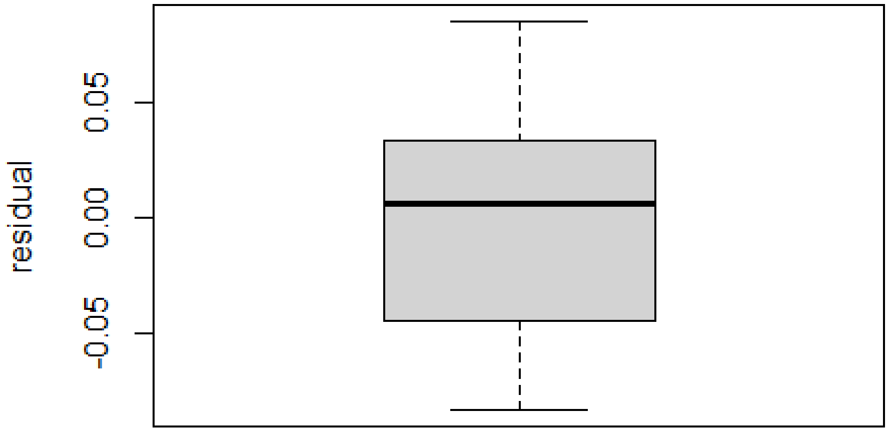
```r
# The estimated regression function is:
# yhat <- 0.654880 - (0.195400 * data4$X)
res <- resid(model4)
studentized_residual <- studres(model4)
plot(fitted(model4), studentized_residual, xlab = "fitted values", ylab =
"studentized_residual", abline(0,0))
```



```r
qqnorm(y = studres(model4), main = "Normal Q-Q plot", xlab = "Theoretical
Quantiles", ylab = "studentized_residual", plot.it = TRUE)
qqline(y= studres(model4), distribution = qnorm)
```



```
boxplot(resid(model4), ylab = "residual")
```



- Residuals fall within a horizontal band centered around the zero line, displaying no systematic pattern of positive and negative values. The residuals do not fan out as the predicted value increases. This shows that the assumption of linearity and constant variance is valid for these data sets.
- The box plot appears to indicate that the distribution of the observation is approximately symmetric with a median close to zero.
- In terms of normality, the points on the normal probability plot fall approximately on a straight line.

(c) Express the estimated regression function in the original units.

yhat = exp(0.65488 - 0.1954X)

QUESTION 5

Consider the attached data on brand preference with y = brand liking, x1 = moisture content and x2 = sweetness of the product. Assume that a multiple linear regression model is appropriate. Use the matrix approach to compute the following.

(a) XTX, XTY, and B

```
data5 <- read.table("C:/Users/HP PAVILION/Downloads/STAT6519 assignment 2
datasets/As2Prob5 Data.txt", header = T)
Y <- data5$Y
o.v <- rep(1, 16)
X1 <- data5$X1
X2 <- data5$X2
Xmat <- cbind(o.v, X1, X2)
XTX <- t(Xmat) %*% Xmat
XTX
```

```
##      o.v  X1  X2
## o.v   16 112  48
## X1   112 864 336
## X2    48 336 160
```

```
XTY <- t(Xmat) %*% Y
XTY
```

```
##     [,1]
## o.v 1308
## X1  9510
## X2  3994
```

```
XTX.inv <- solve(XTX)
be.v <- XTX.inv %*% XTY
be.v
```

```
##        [,1]
## o.v 37.650
## X1   4.425
## X2   4.375

#y.hat <- Xmat %*% be.v
#I.mt <- diag(rep(1,16))
#res <- Y - y.hat
#H.mat <- Xmat %*% XTX.inv %*% t(Xmat)
```

(5b)

Obtain a 99% prediction interval for a new observation ynew when xnew1 = 5 and xnew2 = 4.

```
#Yhatnew <- Bo + B1X1 + B2X2
Yhatnew <- 37.650 + (4.425 * 5) + ( 4.375 * 4)
ov1 <- rep(1,1)
Xnew1 <- c(5)
Xnew2 <- c(4)
XnewT<- cbind(ov1, Xnew1, Xnew2)
Xnew <- t(XnewT)
YTY <- t(Y) %*% Y
SSE <- YTY - (t(be.v) %*% XTY)
MSE <- SSE / (16-3)
Var_pred <- MSE + (MSE * (XnewT %*% XTX.inv %*% Xnew))
spred <- sqrt(Var_pred)
t.critical <- qt(0.9975, 14)
U.int <- Yhatnew + (t.critical * spred)
L.int <- Yhatnew - (t.critical * spred)
U.int

##          [,1]
## [1,] 86.98425

L.int

##          [,1]
## [1,] 67.56575

# The required interval for the new observation is (67.56575, 86.98452)
```

The required interval for the new observation is (67.56575, 86.98452)

(c) SSR, SST and SSE.

```
YTY <- t(Y) %*% Y
SSE <- YTY - (t(be.v) %*% XTY)
SSE
```

```
##      [,1]
## [1,] 94.3

MSE <- SSE / (16-3)
#var.est <- c(MSE) *  (XTX.inv)
J.mat <- o.v %*% t(o.v)
YTJY <- t(Y) %*% J.mat %*% Y
SST <- YTY - (c(1/16) * YTJY)
SST

##      [,1]
## [1,] 1967

SSR <- SST - SSE
SSR

##        [,1]
## [1,] 1872.7
```

(d) Conduct the Breush-Pagan test for constancy of error variance. Use alpha= 0.01

Null hypothesis Ho: Y1 = Y2 = 0 Alternative hypothesis Ha: Y1 = Y2 =! 0

```
xy.lm <- lm(Y ~ X1 + X2)
res <- resid(xy.lm)
sq.res <- res ^ 2
xy1.lm <- lm(sq.res ~ X1 + X2)
anova(xy1.lm)

## Analysis of Variance Table
##
## Response: sq.res
##           Df Sum Sq Mean Sq F value Pr(>F)
## X1         1  67.34  67.344  1.7710 0.2061
## X2         1   5.06   5.063  0.1331 0.7211
## Residuals 13 494.35  38.027

SSR.star <- 67.34 + 5.06
Xo <- ((16^2) / 2) * (SSR.star / (SSE^2))
Xo

##          [,1]
## [1,] 1.042138

chi.square <- qchisq(0.99, 2)
chi.square

## [1] 9.21034

pval <- 1 - pchisq(1.042138, 2)
pval
```

```
## [1] 0.5938853
```

- The test statistic is 1.042138 which is less than the chi square distribution 9.21034. So we cannot reject the null hypothesis suggesting that the constant variance assumption was not violated. Alternatively, P_value which is 0.5938853 is greater than significance level 0.01.