HW4-Solutions

Problem 1

Refer to the Production Time data set.(20 Pts)

a) Prepare a scatter plot of the data Does a linear relation appear adequate here? Would a transformation on X or Y be more appropriate here? Why? (4pts)

b) Use the transformation X' = $\sqrt{X}$ and obtain the estimated linear regression function for the transformed data. (4pts)

c) Plot the estimated regression line and the transformed data. Does the regression line appear to be a good fit to the transformed data? (4pts)

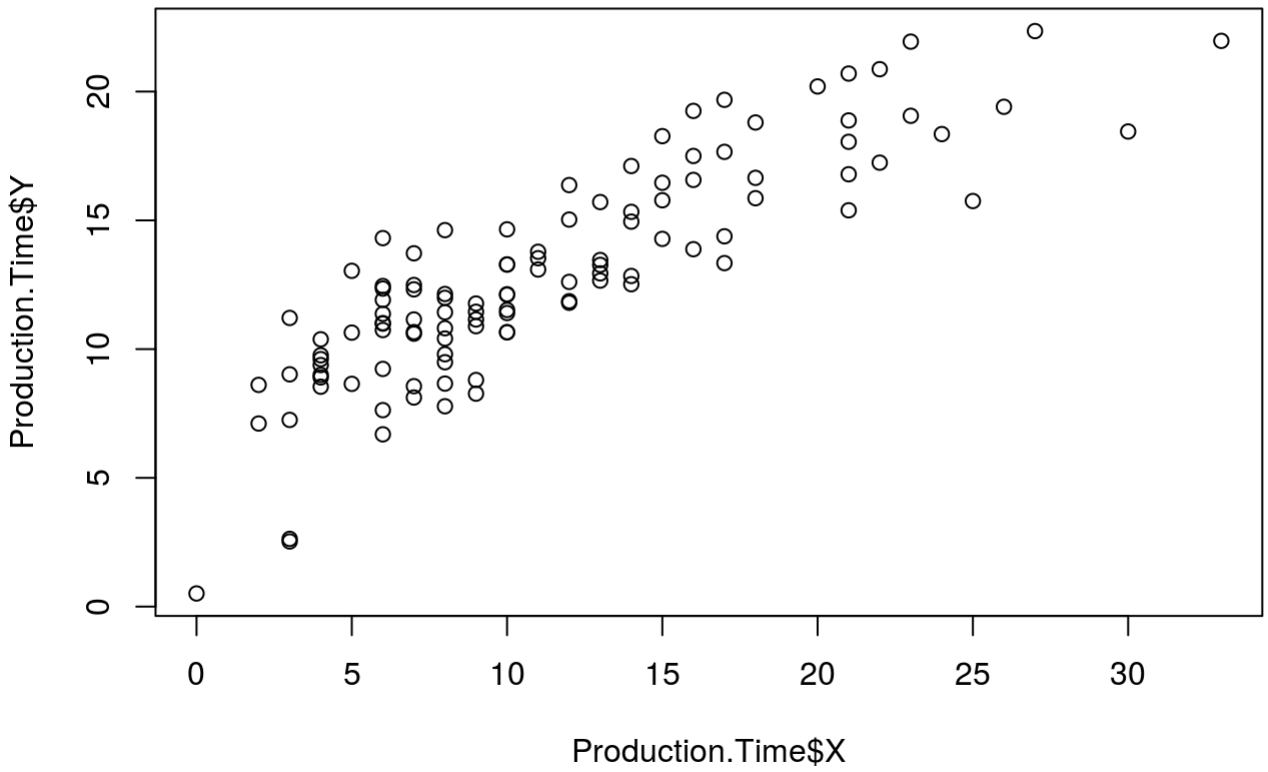
d) Obtain the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show? (4pts)

e) Express the estimated regression function in the original units. (4 pts)

Solution:

a) Plots indicate that errors are approximately normally distributed and have approximately constant variance. Therefore, X should be transformed, log or square root would be appropriate transformation based on the graph._

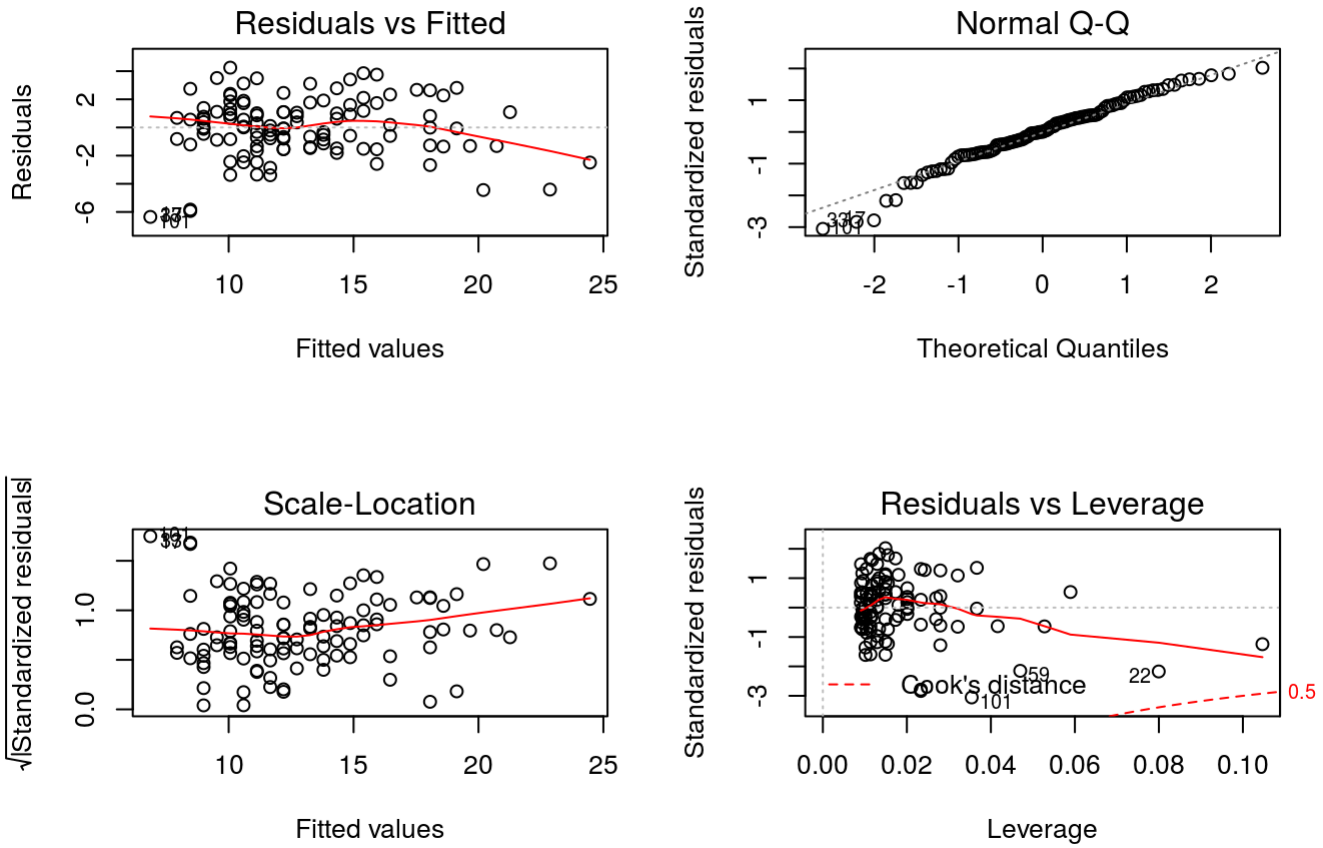
```
library(knitr)
Production.Time <- read.csv("/cloud/project/Production Time.csv")
par(mfrow=c(1,1))
plot(Production.Time$X,Production.Time$Y)
```




```
f1<-lm(Y~X,data=Production.Time)
summary(f1)
```

```
##
## Call:
## lm(formula = Y ~ X, data = Production.Time)
##
## 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 ***
## 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
```

```
par(mfrow=c(2,2))
plot(f1)
```



b)

Solution:

There is a slight increase in Rsquare, the new model is Y = 1.2547 + 3.6235 * SQRT(X)

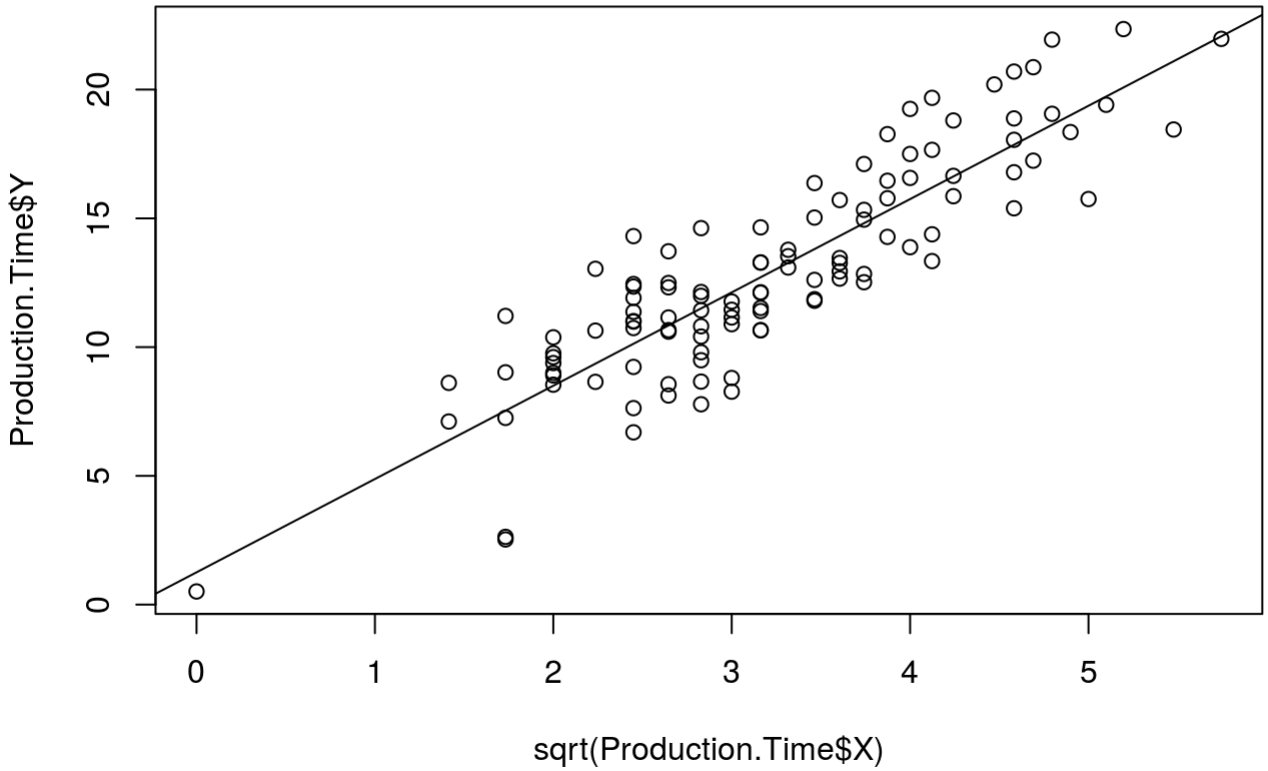
```
f1.2<-lm(Y~sqrt(X),data=Production.Time)
summary(f1.2)
```

```
##
## Call:
## lm(formula = Y ~ sqrt(X), data = Production.Time)
##
## 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(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
```

c)

Solution: it does fit well.

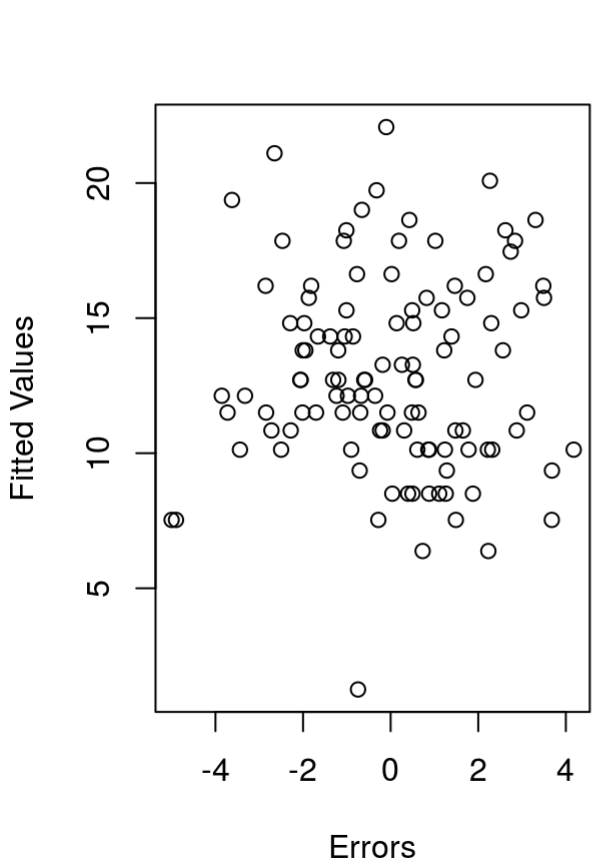
```
par(mfrow=c(1,1))
plot(sqrt(Production.Time$X),Production.Time$Y)
abline(f1.2)
```

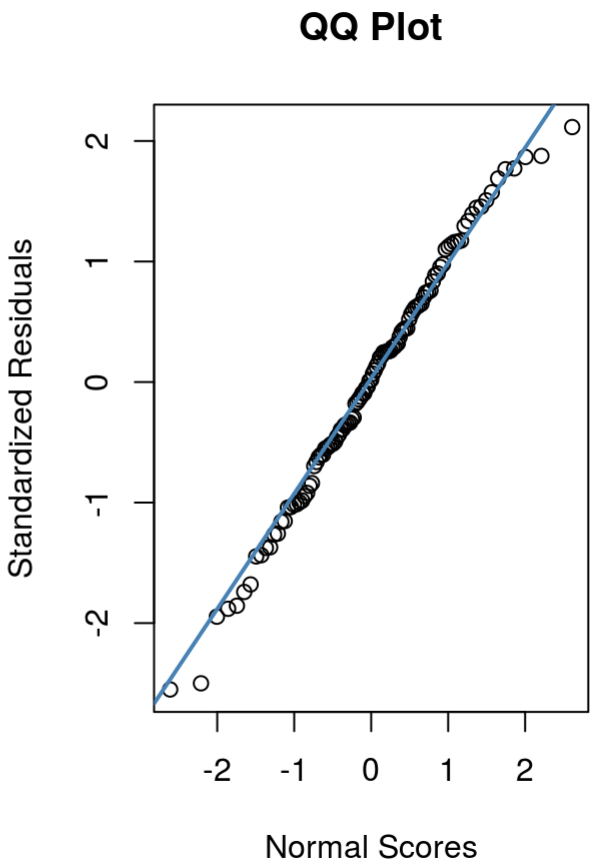


d)

Solution: it is a good fit.

```
ei<-f1.2$residuals
yhat<-f1.2$fitted.values
par(mfrow=c(1,2))
plot(ei,yhat,xlab="Errors",ylab="Fitted Values")
stdei<- rstandard(f1.2)
qqnorm(stdei,ylab="Standardized Residuals",xlab="Normal Scores", main="QQ Plot")
qqline(stdei,col = "steelblue", lwd = 2)
```





e)

Solution: Y = 1.2547 + 3.6235 * SQRT(X)

Problem 2

Refer to Solution Concentration data set. (20 pts)

a)Fit a linear regression function. Obtain, the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show?(3pts)

b)Prepare a scatter plot of the data. What transformation of Y might you try,to achieve constant variance and linearity? (3pts)

c)Use the Box-Cox procedure and standardization (3.36) to find an appropriate power transformation by using λ = -.2, -.1,0, .1, .2. What transformation of Y is suggested? (5pts)

d)Use the transformation Y' = log Y and obtain the estimated linear regression function for the transformed data. (5pts)

e)Plot the estimated regression line and the transformed data Does the regression line appear to be a good fit to the transformed data? (2 pts)

f)Obtain, the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show? (2pts)

g)Express the estimated regression function in the original units. (2pts)

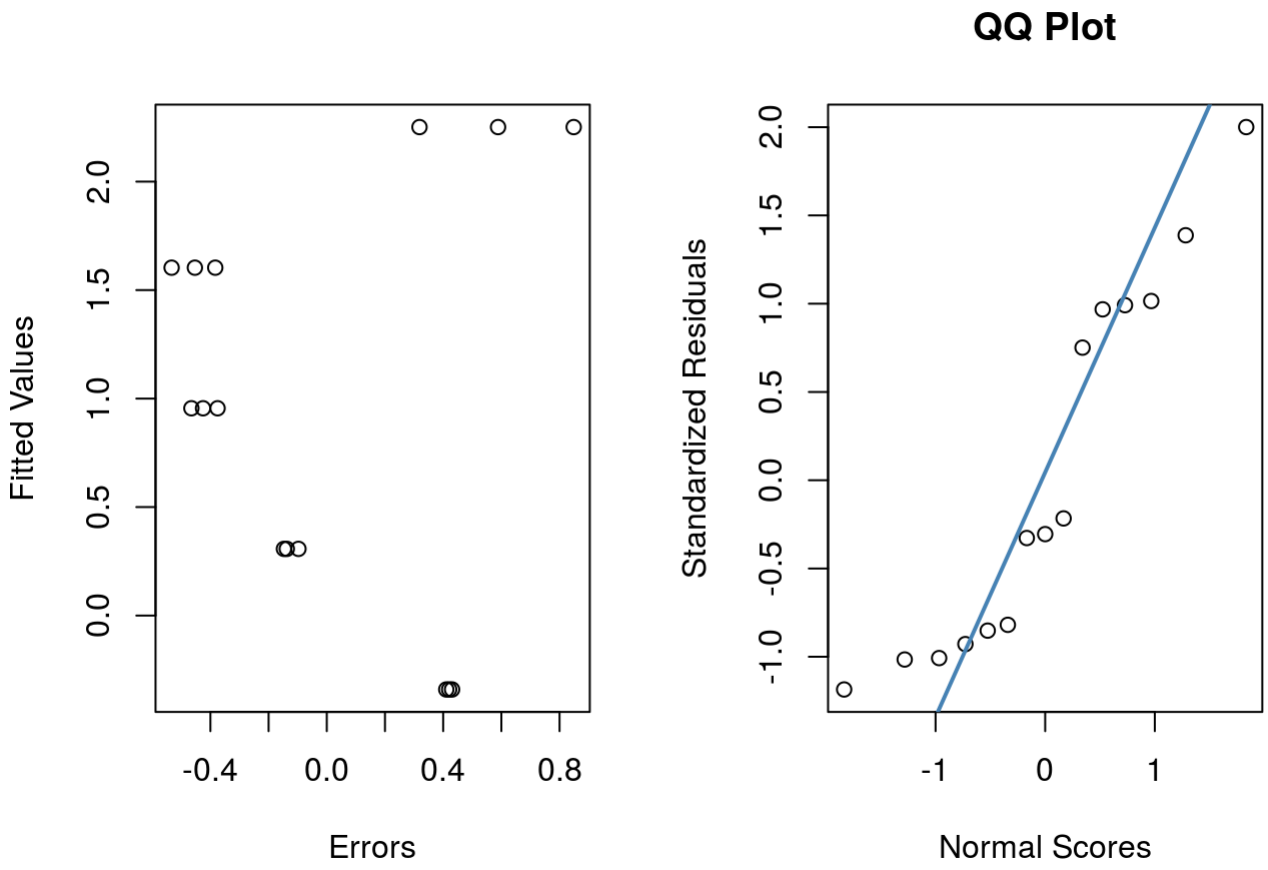
a)

Solution: QQ plot shows that the errors are not normally distributed. The errors vs fitted values plot indicate that variances are not equal. However, R square is 81% and the model is significant.

```
Solution.Concentration <- read.csv("/cloud/project/Solution Concentration.csv")
f2<-lm(Y~X,data=Solution.Concentration)
summary(f2)
```

```
##
## Call:
## lm(formula = Y ~ X, data = Solution.Concentration)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5333 -0.4043 -0.1373  0.4157  0.8487
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.5753     0.2487  10.354 1.20e-07 ***
## X            -0.3240     0.0433  -7.483 4.61e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4743 on 13 degrees of freedom
## Multiple R-squared:  0.8116, Adjusted R-squared:  0.7971
## F-statistic: 55.99 on 1 and 13 DF,  p-value: 4.611e-06
```

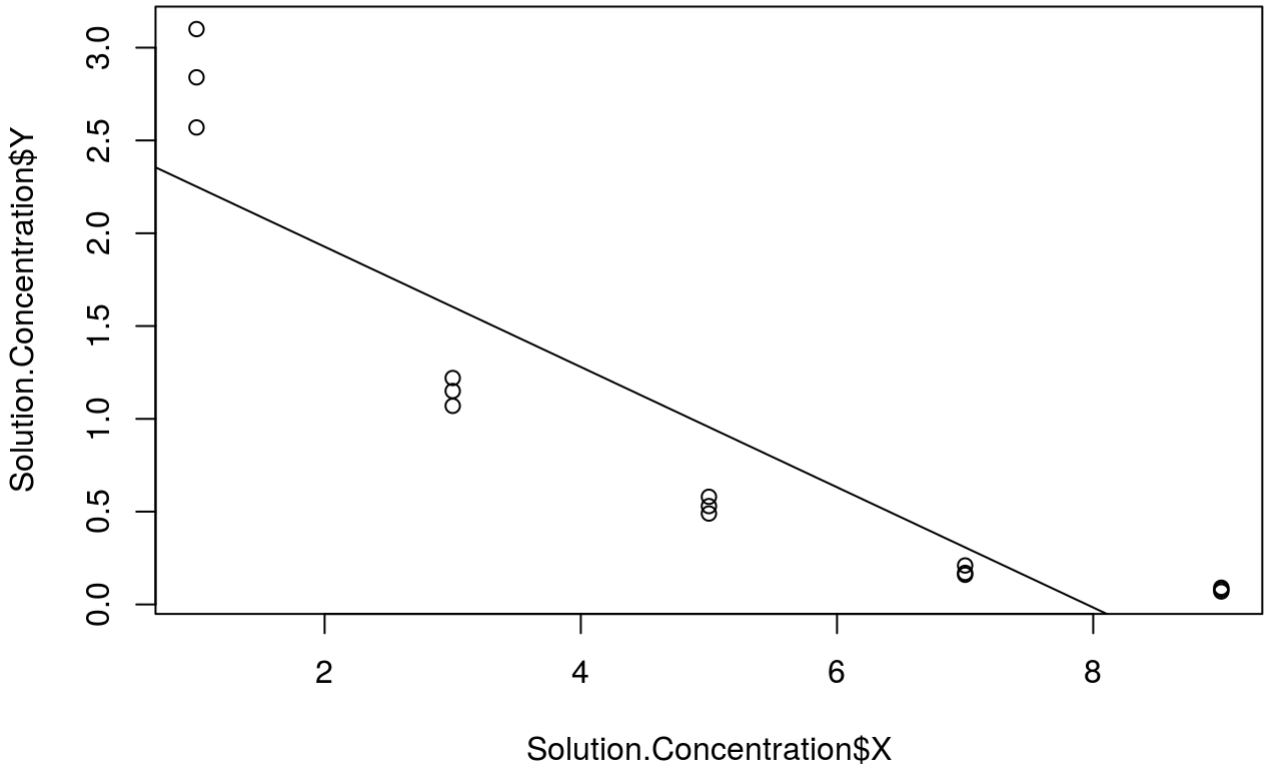
```
ei<-f2$residuals
yhat<-f2$fitted.values
par(mfrow=c(1,2))
plot(ei,yhat,xlab="Errors",ylab="Fitted Values")
stdei<- rstandard(f2)
qqnorm(stdei,ylab="Standardized Residuals",xlab="Normal Scores", main="QQ Plot")
qqline(stdei,col = "steelblue", lwd = 2)
```



b)

Solution: Log transform on Y.

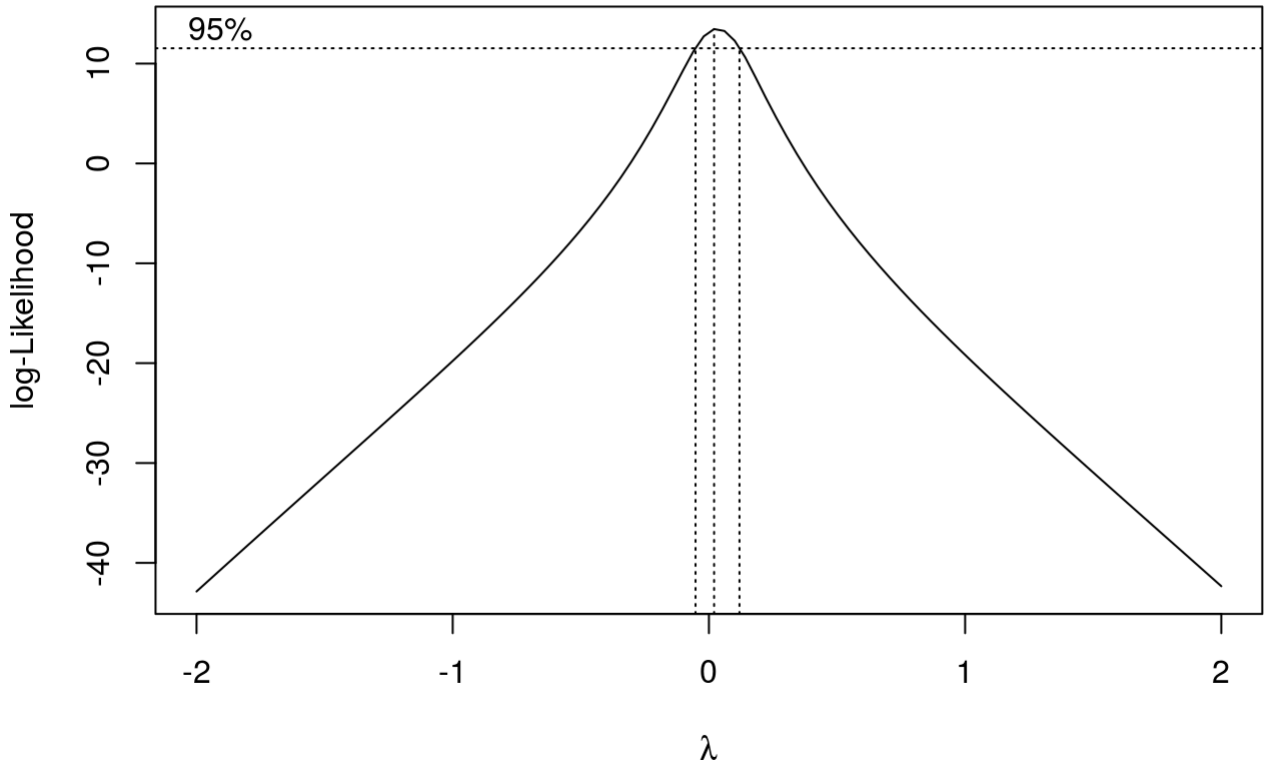
```
par(mfrow=c(1,1))
plot(Solution.Concentration$X,Solution.Concentration$Y)
abline(f2)
```



c) ### *Solution:* Boxcox

transformation indicate log transformation.

```
library(MASS)
boxcox(f2,lambda = seq(-2,2,by=0.1))
```



d) ### *Solution:* log(Y)= 1.50792

- 0.44993X, the r-square is almost 100%. It is a perfect fit.

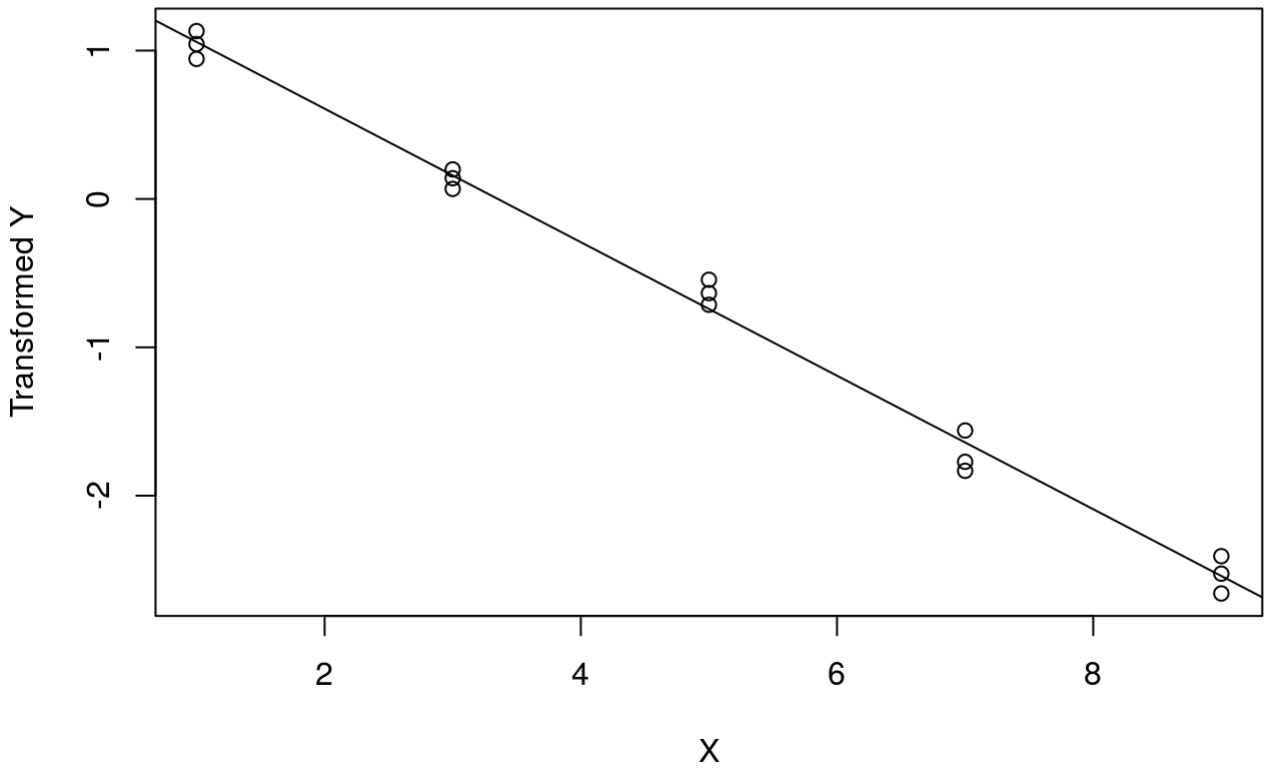
```
f2.1<-lm(log(Y)~X,data=Solution.Concentration)
summary(f2.1)
```

```
##
## Call:
## lm(formula = log(Y) ~ X, data = Solution.Concentration)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.19102 -0.10228  0.01569  0.07716  0.19699
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.50792    0.06028   25.01 2.22e-12 ***
## X           -0.44993    0.01049  -42.88 2.19e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.115 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
```

e)

Solution: It is a perfect fit.

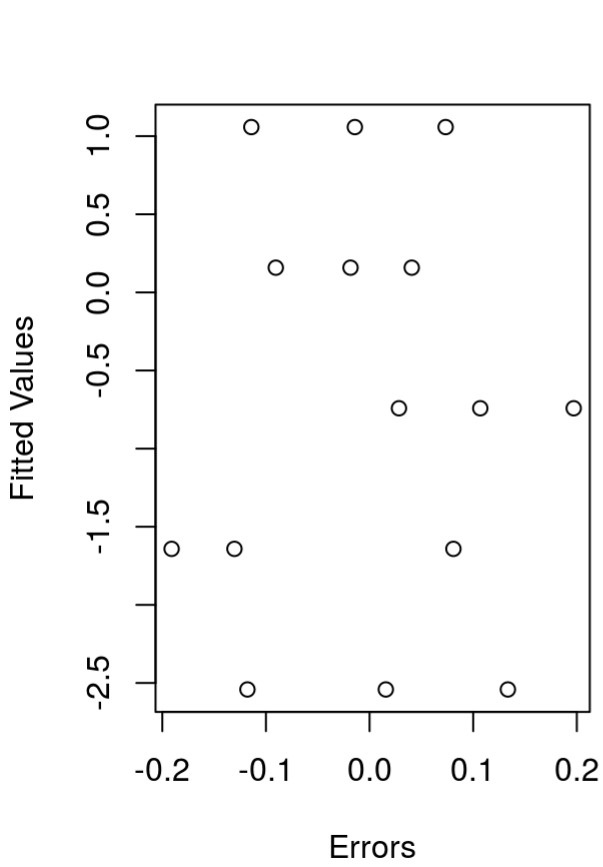
```
par(mfrow=c(1,1))
plot(Solution.Concentration$X,log(Solution.Concentration$Y),xlab="X",ylab="Transformed Y")
abline(f2.1)
```

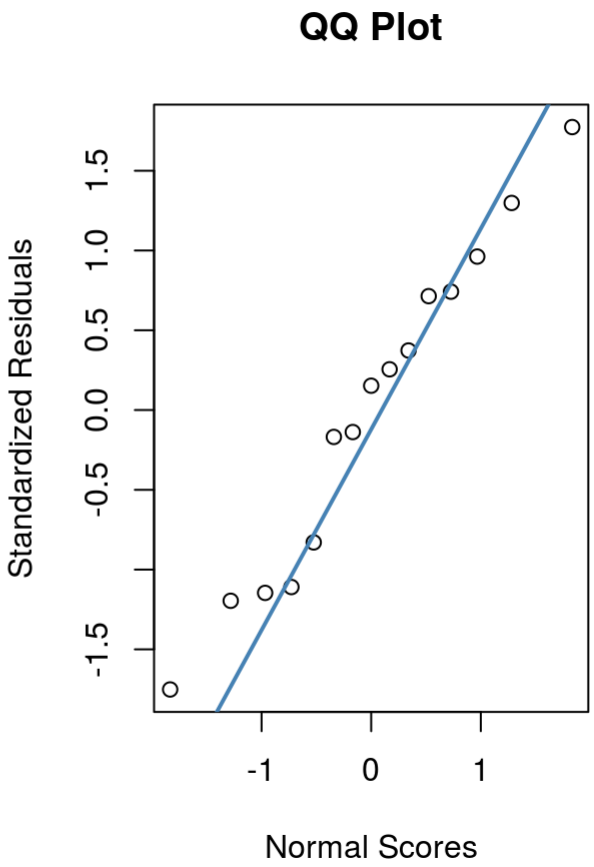


f) ### *Solution:* Error variances

look constant, errors are approximately normally distributed.

```
ei<-f2.1$residuals
yhat<-f2.1$fitted.values
par(mfrow=c(1,2))
plot(ei,yhat,xlab="Errors",ylab="Fitted Values")
stdei<- rstandard(f2.1)
qqnorm(stdei,ylab="Standardized Residuals",xlab="Normal Scores", main="QQ Plot")
qqline(stdei,col = "steelblue", lwd = 2)
```





g) ### *Solution:* log(Y)= 1.50792

- 0.44993X ==> Y=exp(1.50792 - 0.44993X)

Problem 3

###Refer to Crime rate data set. (25 pts) ### a) Fit a linear regression function. Obtain, the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show? (3pts) ### b) Conduct the Brown-Forsythe test to determine whether or not the error variance varies with the level of X. Divide the data into the two groups, X≤69, X > 69, and use α= .05. State the decision rule and conclusion. Does your conclusion support your preliminary findings in part (a)? (10 pts) ### c) Conduct the Breusch-Pagan test to determine whether or not the error variance varies with the level of X. Use α= .05. State the alternatives. decision rule, and conclusion. Is your conclusion consistent with your preliminary findings in part (a and b)? (12 pts)

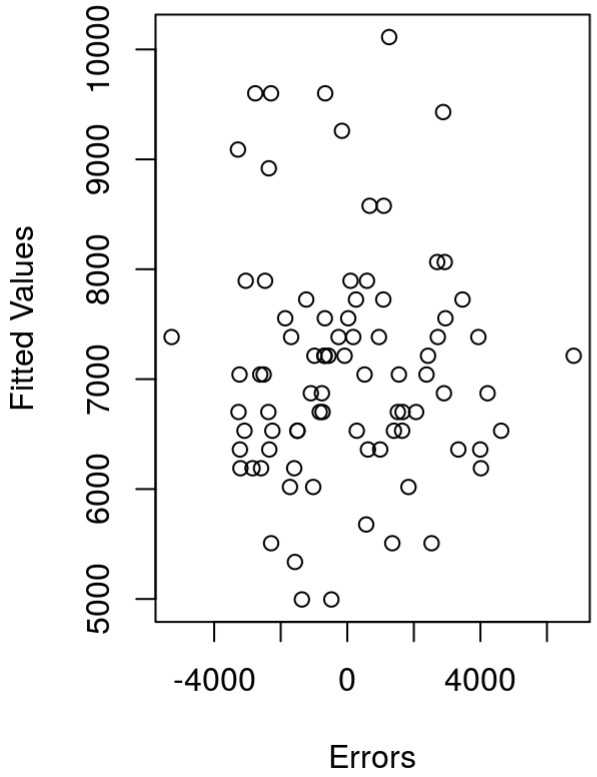
a)

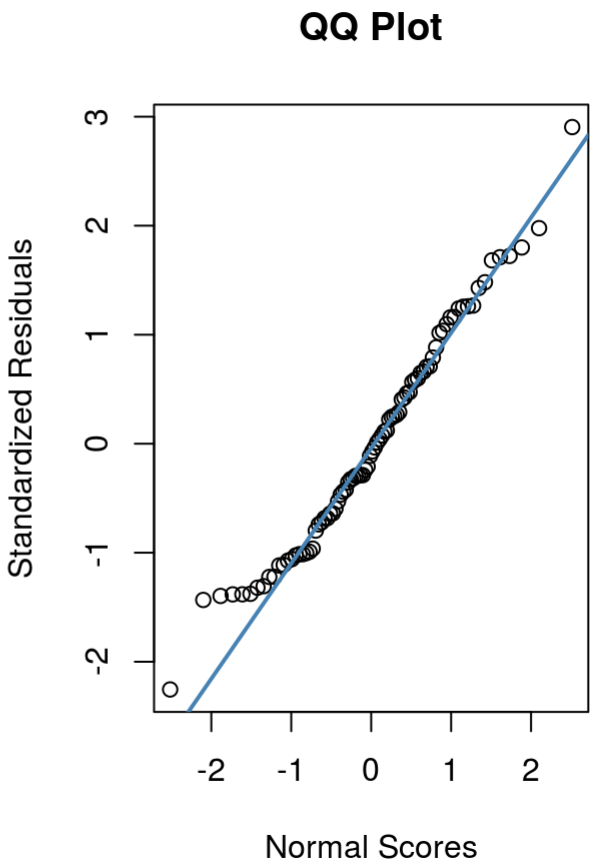
Solution: The model is significant, however Rsquare is low. The graphs do not indicate non constant variance or significant departures from the normal distribution.

```
Crime.Rate <- read.csv("/cloud/project/Crime Rate.csv")
f3<-lm(Y~X,data=Crime.Rate)
summary(f3)
```

```
##
## Call:
## lm(formula = Y ~ X, data = Crime.Rate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5278.3 -1757.5  -210.5  1575.3  6803.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20517.60    3277.64   6.260 1.67e-08 ***
## X            -170.58      41.57  -4.103 9.57e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2356 on 82 degrees of freedom
## Multiple R-squared:  0.1703, Adjusted R-squared:  0.1602
## F-statistic: 16.83 on 1 and 82 DF,  p-value: 9.571e-05
```

```
ei<-f3$residuals
yhat<-f3$fitted.values
par(mfrow=c(1,2))
plot(ei,yhat,xlab="Errors",ylab="Fitted Values")
stdei<- rstandard(f3)
qqnorm(stdei,ylab="Standardized Residuals",xlab="Normal Scores", main="QQ Plot")
qqline(stdei,col = "steelblue", lwd = 2)
```





b) ### *Solution:* Ho: Error variance is constant ### Ha: Error variance is NOT constant; T stat is less than 1, accept null.Error variances are constant.

```
ei<-f3$residuals
DM<-data.frame(cbind(Crime.Rate$X,Crime.Rate$Y,ei))
DM1<-DM[DM[,1]<=69,]
DM2<-DM[DM[,1]>69,]

M1<-median(DM1[,3])
M2<-median(DM2[,3])
N1<-length(DM1[,3])
N2<-length(DM2[,3])

d1<-abs(DM1[,3]-M1)
d2<-abs(DM2[,3]-M2)
s2<-sqrt((var(d1)*(N1-1)+var(d2)*(N2-1))/(N1+N2-2))
Den<- s2*sqrt(1/N1+1/N2)
Num<- mean(d1)-mean(d2)
T= Num/Den
T
```

```
## [1] -0.3550185
```

c)

Solution: Ho: Gamma is 0

Ha: Gamma is NOT 0; the slope is not significant. R square is almost zero, no need to perform the F test, accept null, the error variance is constant.

```
ei2<-(f3$residuals)^2
f3.1<-lm(ei2~Crime.Rate$X)
```

Problem 4

Refer to Plastic Hardness dataset.(15pts)

a) Fit a linear regression function. Obtain, the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show? (3 pts)

b) Obtain Bonferroni joint confidence intervals for β0 and β1, using a 90 percent family confidence coefficient. Interpret your confidence intervals. (3 pts)

c) Are bo and b1 positively or negatively correlated here? Is this reflected in your joint confidence intervals in part (b) (3 pts)

d) Management wishes to obtain interval estimates of the mean hardness when the elapsed time is 20, 30, and 40 hours, respectively. Calculate the desired confidence intervals using the Bonferroni procedure and a 90 percent family confidence coefficient. What is the meaning of the family confidence coefficient here? (3 pts)

e) The next two test items will be measured after 30 and 40 hours of elapsed time, respectively. Predict the hardness for each of these two items, using the most efficient procedure and a 90 percent family confidence coefficient. (3pts)

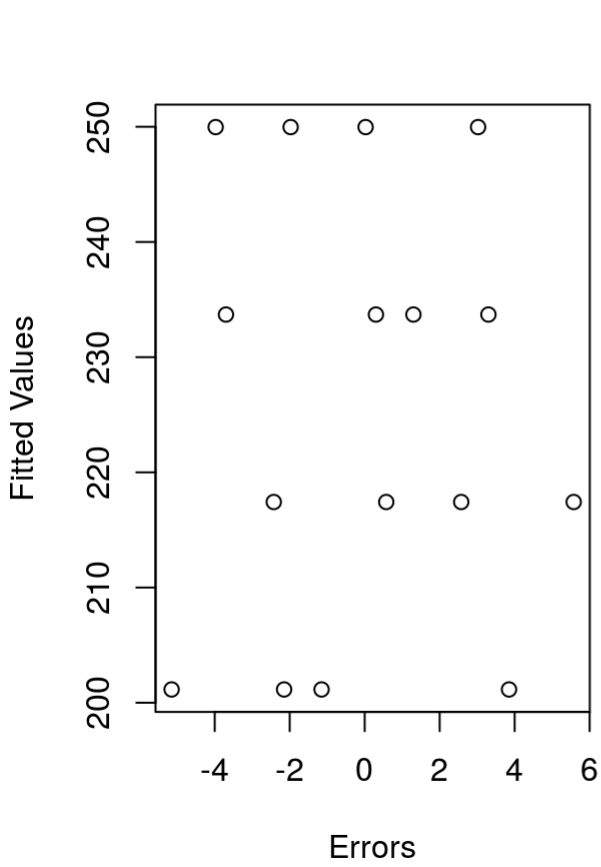
a)

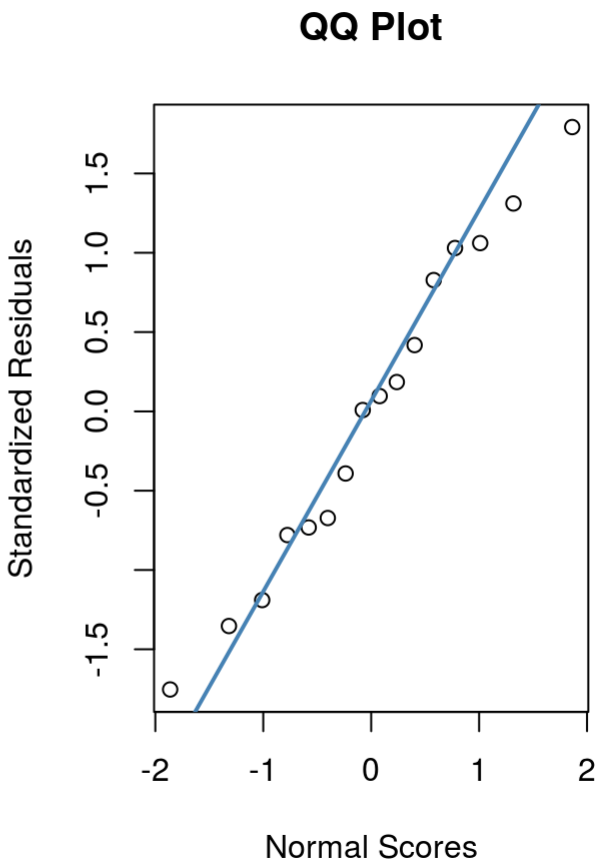
Solution: The model is significant. The Rsquare is 97%. The graphs do not indicate nonconstant variance or departure from the normal distribution.

```
Plastic.Hardness <- read.csv("/cloud/project/Plastic Hardness.csv")
f4<-lm(Y~X,data=Plastic.Hardness)
summary(f4)
```

```
##
## Call:
## lm(formula = Y ~ X, data = Plastic.Hardness)
##
## 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
```

```
ei<-f4$residuals
yhat<-f4$fitted.values
par(mfrow=c(1,2))
plot(ei,yhat,xlab="Errors",ylab="Fitted Values")
stdei<- rstandard(f4)
qqnorm(stdei,ylab="Standardized Residuals",xlab="Normal Scores", main="QQ Plot")
qqline(stdei,col = "steelblue", lwd = 2)
```





b) ### *Solution:* The confidence intervals are below:

```
confint(f4,level=1-0.1/2)
```

```
##                2.5 %    97.5 %
## (Intercept) 162.9013 174.29875
## X             1.8405   2.22825
```

c)

Solution: Yes, they are negatively correlated.

d)

Solution: The joint confidence intervals are below.

```
Xh<-c(20,30,40)
predict.lm(f4,data.frame(X= c(Xh)),interval = "confidence", level = 1-0.1/3)
```

```
##        fit      lwr      upr
## 1 209.2875 206.7277 211.8473
## 2 229.6312 227.6762 231.5863
## 3 249.9750 246.7824 253.1676
```

e)

_Solution: S = 2.340, B = t(975,14) = 2.145. Bonferroni is more efficent.

```
B<-qt(1-0.1/4,14)
S<-sqrt(2*qf(0.90,2,14))
cbind(B,S)
```

```
##             B        S
## [1,] 2.144787 2.335152
```

```
Xh<-c(30,40)
predict.lm(f4,data.frame(X= c(Xh)),interval = "prediction", level = 1-0.1/2)
```

```
##        fit      lwr      upr
## 1 229.6312 222.4710 236.7915
## 2 249.9750 242.4562 257.4938
```

Problem 5

Refer to the CDI data set. Consider the regression relation of number of active physicians to total population.(10 pts)

a) Obtain Bonferroni joint confidence intervals for β0 and β1 using a 95 percent family confidence coefficient. (2 pts)

b) An investigator has suggested that β0 should be -100 and β1 should be .0028. Do the joint confidence intervals in part (a) support this view? Discuss.(2 pts)

c) It is desired to estimate the expected number of active physicians for counties with total population of X = 500, 1000, 5000 thousand with family confidence coefficient .90. Which procedure, the Wolking-Hotelling or the Bonferroni, is more efficient here? (3pts)

d) Obtain the family of interval estimates required in part (c), using the more efficient procedure. Interpret your confidence intervals. (3pts)

a)

Solution: The regression model is significant. The Rsquare is %88. The joint confidence intervals are below.

```
CDI <- read.csv("/cloud/project/CDI.csv")
f5<-lm(Number.of.active.physicians~Total.population,data=CDI)
summary(f5)
```

```
##
## Call:
## lm(formula = Number.of.active.physicians ~ Total.population,
##     data = CDI)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1969.4  -209.2   -88.0    27.9  3928.7
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.106e+02  3.475e+01  -3.184  0.00156 **
## Total.population  2.795e-03  4.837e-05  57.793  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 610.1 on 438 degrees of freedom
## Multiple R-squared:  0.8841, Adjusted R-squared:  0.8838
## F-statistic:  3340 on 1 and 438 DF,  p-value: < 2.2e-16
```

```
confint(f5,level=1-0.05/2)
```

```
##                         1.25 %       98.75 %
## (Intercept)      -1.887833e+02 -32.486285498
## Total.population  2.686636e-03   0.002904214
```

b)

Solution: the joint confidence intervals: -188.8<bo<-32.5 and 0.0027<b1<0.0029

b0=100 and b1=0.0028 fall into the joint confidence intervals. Yes, it does support the investigator’s view.

c)

Solution: W = SQRT(2*2.314732)=2.15, B = t(0.97,438) = 1.838493 . Bonferroni is more efficent (B<W).

```
B<-qt(1-0.1/3,438)
W <- sqrt(2*qf(0.90,2,438))
cbind(B,W)
```

```
##             B        W
## [1,] 1.838493 2.151619
```

d)

Solution: See below for the confidence intervals

```
Xh<-c(500,1000,5000)
predict.lm(f5,data.frame(Total.population=c(Xh)),interval = "prediction", level = 1-0.1/3)
```

```
##          fit       lwr      upr
## 1 -109.23706 -1413.744 1195.269
## 2 -107.83935 -1412.344 1196.666
## 3  -96.65765 -1401.150 1207.834
```

#Problem 6 ##Refer to the SENIC data set. The average length of stay in a hospital (Y) is anticipated to be related to infection risk, available facilities and services, and routine chest X-ray ratio.(10 pts) ### a) Regress average length of stay on each of the three predictor variables. State the estimated regression functions. (3 pts) ### b) For each of the three fitted regression models, obtain the residuals and prepare a residual plot against X and a normal probability plot. Summarize your conclusions. (3 pts) ### c) Obtain the fitted regression function for the relation between length of stay and infection risk after deleting cases 47 (X47 = 6.5, Y47 = 19.56) and 112 (X112 = 5.9, Y112 = 17.94). From this fitted regression function obtain separate 95 percent prediction intervals for new Y observations at X = 6.5 and X = 5.9, respectively. Do observations Y47 and Y112 fall outside these prediction intervals? Discuss the significance of this. (4pts)

a)

Solution: see below for the regression coefficents.

```
SENIC <- read.csv("/cloud/project/SENIC.csv")
f6.1<-lm(Length.of.stay~Infection.risk,data=SENIC)
f6.2<-lm(Length.of.stay~Available.facilities.and.services,data=SENIC)
f6.3<-lm(Length.of.stay~Routine.chest.X.ray.ratio,data=SENIC)
coef<-rbind(f6.1$coefficients,f6.2$coefficients,f6.3$coefficients)
dimnames(coef)[[2]]<-c("bo","b1")
coef
```

```
##            bo         b1
## [1,] 6.336787 0.76042089
## [2,] 7.718767 0.04470767
## [3,] 6.566373 0.03775583
```

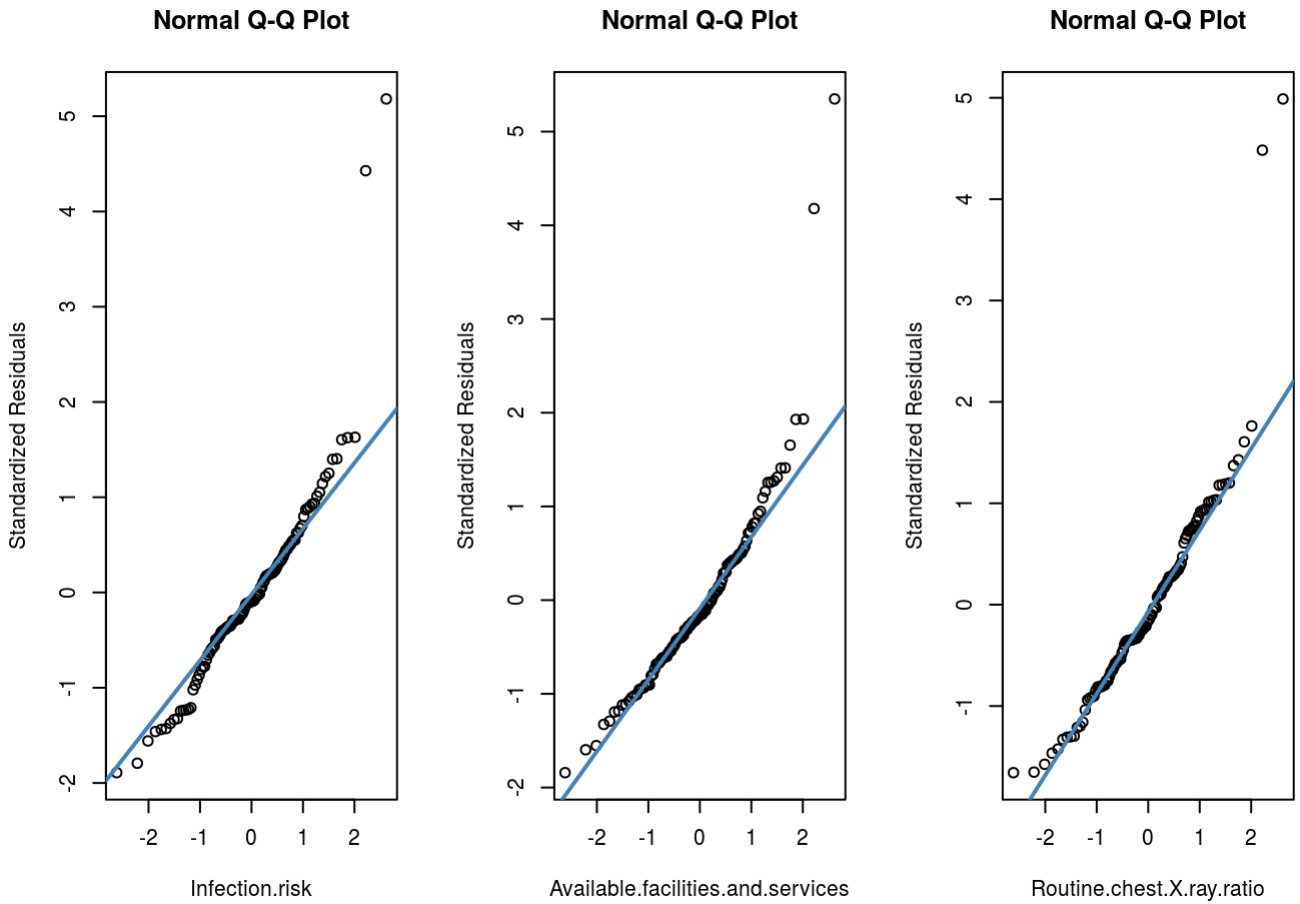
b)

Solution: There are outliers in the data, errors are approximately normally distributed.

```
par(mfrow=c(1,3))
stdei<- rstandard(f6.1)
qqnorm(stdei,ylab="Standardized Residuals",xlab="Infection.risk")
qqline(stdei,col = "steelblue", lwd = 2)

stdei<- rstandard(f6.2)
qqnorm(stdei,ylab="Standardized Residuals",xlab="Available.facilities.and.services")
qqline(stdei,col = "steelblue", lwd = 2)

stdei<- rstandard(f6.3)
qqnorm(stdei,ylab="Standardized Residuals",xlab="Routine.chest.X.ray.ratio")
qqline(stdei,col = "steelblue", lwd = 2)
```



c)

Solution: They fall outside of the prediction invervals. It indicates that these observations are outliers.

```
f6.11<-lm(Length.of.stay~Infection.risk,data=SENIC[-c(47,112),])
summary(f6.11)
```

```
##
## Call:
## lm(formula = Length.of.stay ~ Infection.risk, data = SENIC[-c(47,
##     112), ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89309 -0.67980 -0.08822  0.87180  3.07644
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.84922    0.40137  17.065  < 2e-16 ***
## Infection.risk  0.60975    0.08881   6.866 4.23e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.238 on 109 degrees of freedom
## Multiple R-squared:  0.3019, Adjusted R-squared:  0.2955
## F-statistic: 47.14 on 1 and 109 DF,  p-value: 4.233e-10
```

```
predict.lm(f6.11,data.frame(Infection.risk=c(6.5,5.9)),interval="prediction",level=0.95)
```

```
##        fit      lwr      upr
## 1 10.81259 8.318631 13.30654
## 2 10.44674 7.966822 12.92665
```