CSCI E-106:Assignment 4

Due Date: October 5, 2020 at 7:20 pm EST

Instructions

Students should submit their reports on Canvas. The report needs to clearly state what question is being solved, step-by-step walk-through solutions, and final answers clearly indicated. Please solve by hand where appropriate.

Please submit two files: (1) a R Markdown file (.Rmd extension) and (2) a PDF document, word, or html generated using knitr for the .Rmd file submitted in (1) where appropriate. Please, use RStudio Cloud for your solutions.

Problem 1

Refer to the Real estate sales data set. Obtain a random sample of 200 cases from the 522 cases in this data set (use set.seed(1023) before selecting the sample). Using the random sample, build a regression model to predict sales price (Y) as a function of finished square feet (X). The analysis should include an assessment of the degree to which the key regression assumptions are satisfied. If the regression assumptions are not met, include and justify appropriate remedial measures. Use the final model to predict sales price for two houses that are about to come on the market: the first has X = 1100 finished square feet and the second has X = 4900 finished square feet. Assess the strengths and weaknesses of the final model. (25 points)

Solutions:

QQ plot indicates departure from normality. Boxcox transformation showed that $\frac{1}{\sqrt{Y}}$ is the right transformation. After the transformation QQ plot indicates normality. The model coefficients are below. Both models are significant. R^2's are 67% and 68% respectively. Transformed model did not increase the model power.

	(Intercept)	Finished.square.feet
[original model]	-117240.711539118	177.2424966968304
[final model]	0.003231482	-0.0000005321774

Predicted house prices for X = 1100 and X = 4900 for both models are

	X=1100	X=4900
[original model]	$77,726	$751,247
[final model]	$142,820	$2,569,754

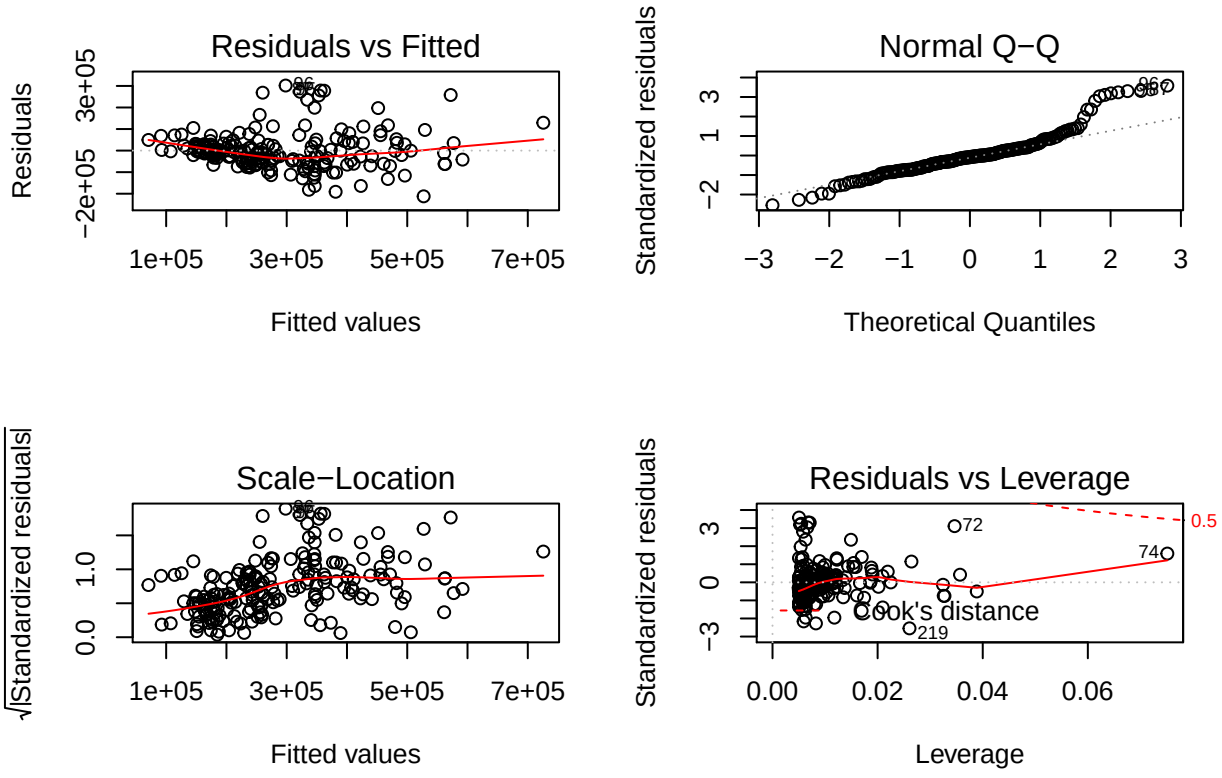
From the data set, the range for the sales price is between $84,000 and $920,000. The range for the finished square feet is between 980 and 5032. Estimates for the final model seem to over predict the house prices specifically for X=4900. The model on the full data set should be developed rather than 200 observations. This indicates the limitation of the simple linear regression model. More variables should be added to the regression model.

```
Real.Estate.Data <- read.csv("/cloud/project/Fall 2020/Real Estate Data.csv")
set.seed(1023)
ind<-sample(c(1:522),200)
```

```
samp<-Real.Estate.Data[ind,]
f<-lm(Sales.price~Finished.square.feet,data=samp)
summary(f)
```

```
##
## Call:
## lm(formula = Sales.price ~ Finished.square.feet, data = samp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -212213  -48875   -7133   29359  301784
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -1.172e+05  2.140e+04  -5.477  1.3e-07 ***
## Finished.square.feet  1.772e+02  9.025e+00  19.640  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 84460 on 198 degrees of freedom
## Multiple R-squared:  0.6608, Adjusted R-squared:  0.6591
## F-statistic: 385.7 on 1 and 198 DF,  p-value: < 2.2e-16
```

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

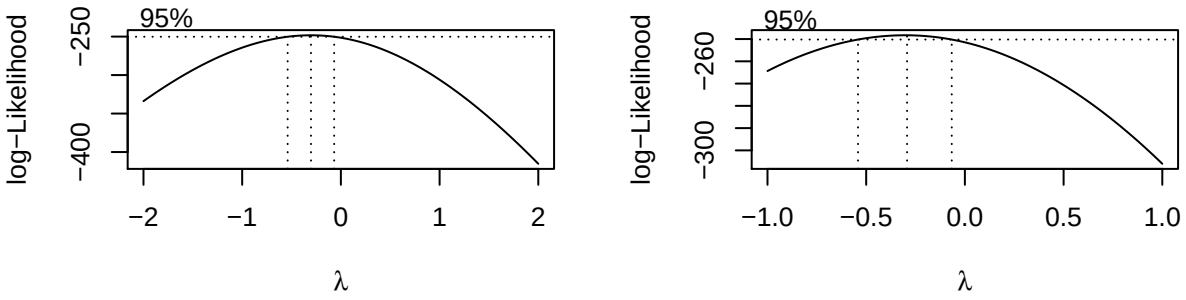



```
library(MASS)
boxcox(f,lambda=seq(-2,2,0.5))
boxcox(f,lambda=seq(-1,1,0.5))
```

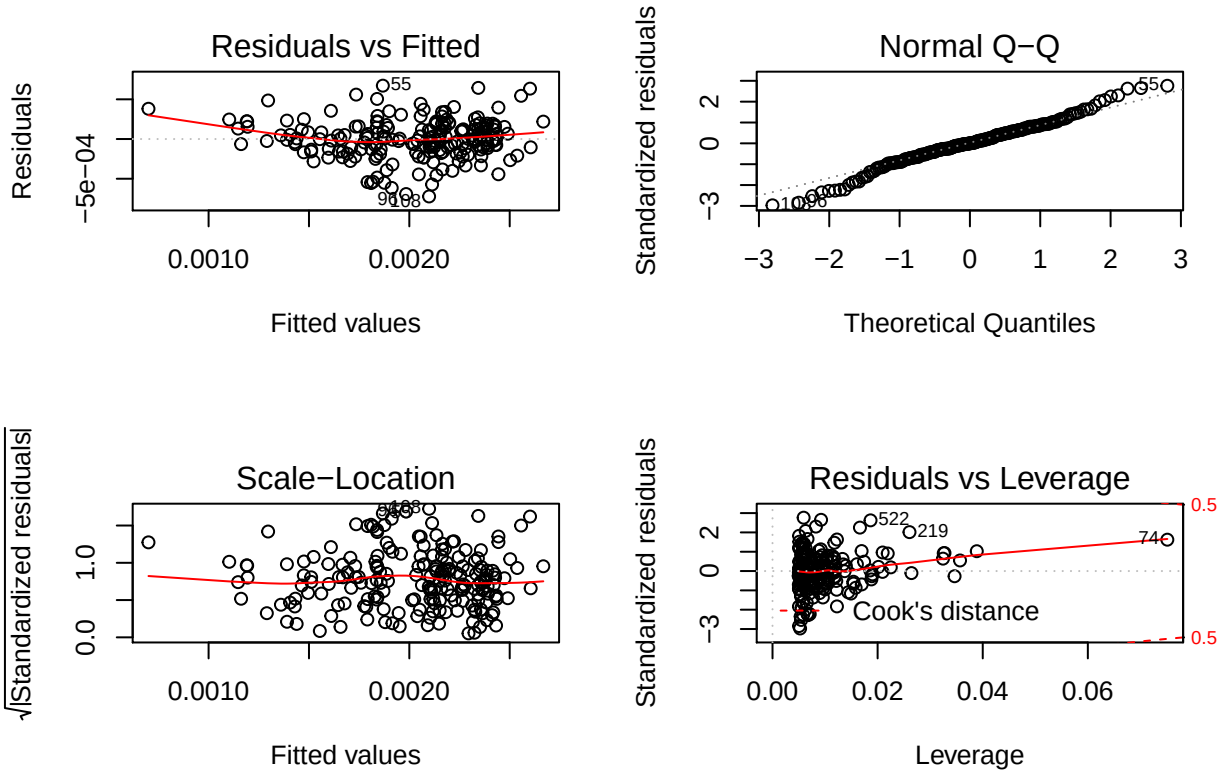
```
f1<-lm(1/sqrt(Sales.price)~Finished.square.feet,data=samp)
summary(f1)
```

```
##
## Call:
## lm(formula = 1/sqrt(Sales.price) ~ Finished.square.feet, data = samp)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -7.233e-04 -1.308e-04  1.200e-07  1.459e-04  6.720e-04
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.231e-03  6.184e-05   52.26   <2e-16 ***
## Finished.square.feet -5.322e-07  2.607e-08  -20.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.000244 on 198 degrees of freedom
## Multiple R-squared:  0.6779, Adjusted R-squared:  0.6762
## F-statistic: 416.6 on 1 and 198 DF,  p-value: < 2.2e-16
```

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




```
plot(f1)
```




```
x.pred<-data.frame(Finished.square.feet=c(1100,4900))
predict(f,x.pred)
```

```
##         1         2
##  77726.03 751247.52
```

```
(1/predict(f1,x.pred))^2
```

```
##         1         2
##  142820.9 2569754.3
```

```
summary(Real.Estate.Data$Sales.price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   84000  180000  229900  277894  335000  920000
```

```
summary(Real.Estate.Data$Finished.square.feet)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     980    1701    2061    2261    2636    5032
```

Problem 2

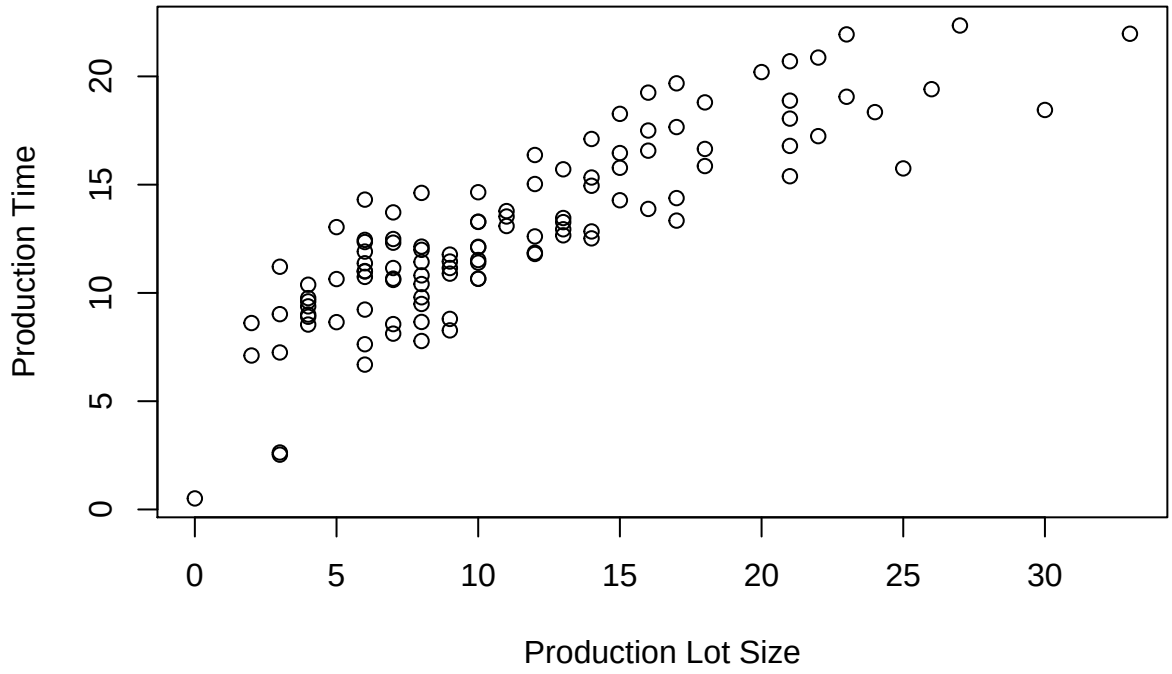
Refer to the Production time data. In a manufacturing study, the production times for 111 recent production runs were obtained. The production time in hours (Y) and the production lot size (X) are recorded for each run. (25 points, 5 points each)

Solutions

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?

From the graph, the relationship does not look linear. Square root transformation might be appropriate.

```
PT <- read.csv("/cloud/project/Fall 2020/Production Time Data.csv")
par(mfrow=c(1,1))
plot(PT$X,PT$Y,xlab="Production Lot Size",ylab="Production Time")
```



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

R^2 is %77 and the model is significant. Y= 1.2547 + 3.6235*X'

```
f<-lm(Y~sqrt(X),data=PT)
summary(f)
```

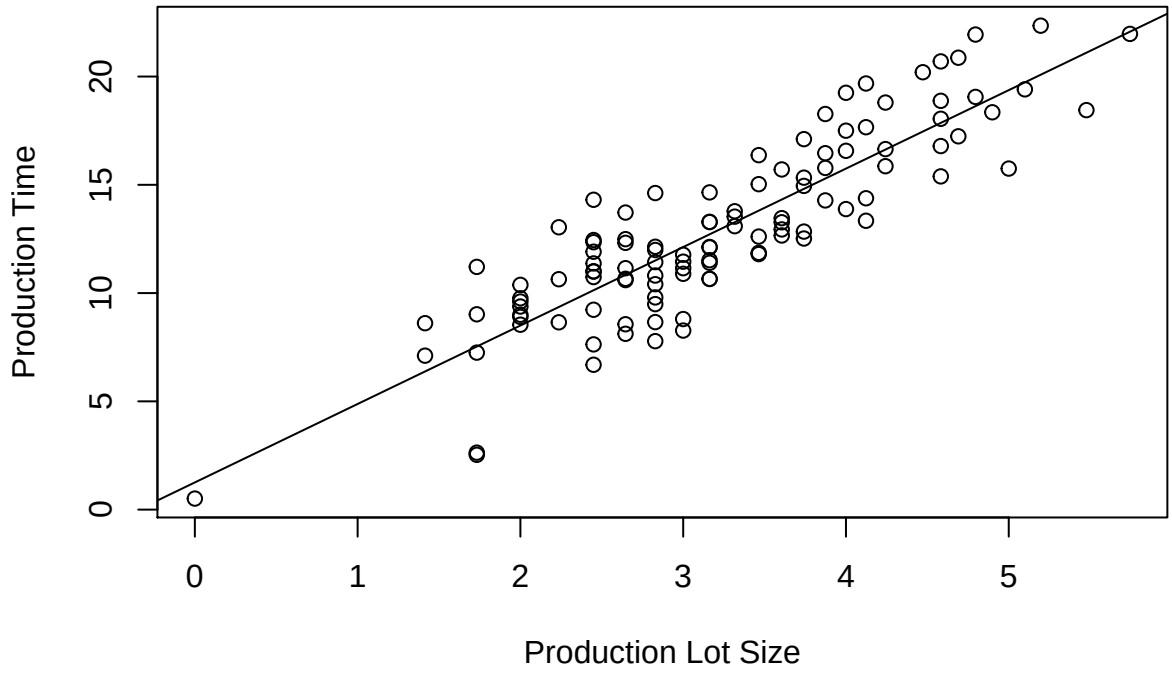
```
##
## Call:
## lm(formula = Y ~ sqrt(X), data = PT)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0008 -1.2161  0.0383  1.3367  4.1795
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.2547     0.6389   1.964   0.0521 .
## sqrt(X)       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-) Plot the estimated regression line and the transformed data. Does the regression line appear to be a good fit to the transformed data?

The graph showed the transformation is appropriate.

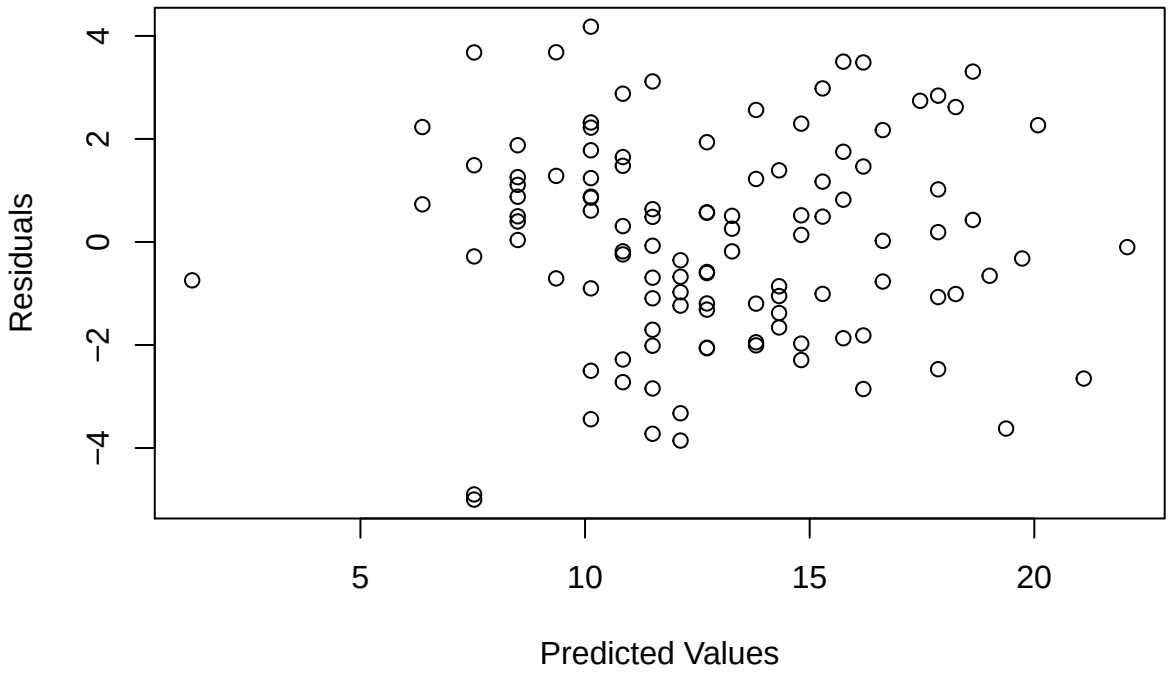
```
plot(sqrt(PT$X),PT$Y,xlab="Production Lot Size",ylab="Production Time")
abline(f)
```



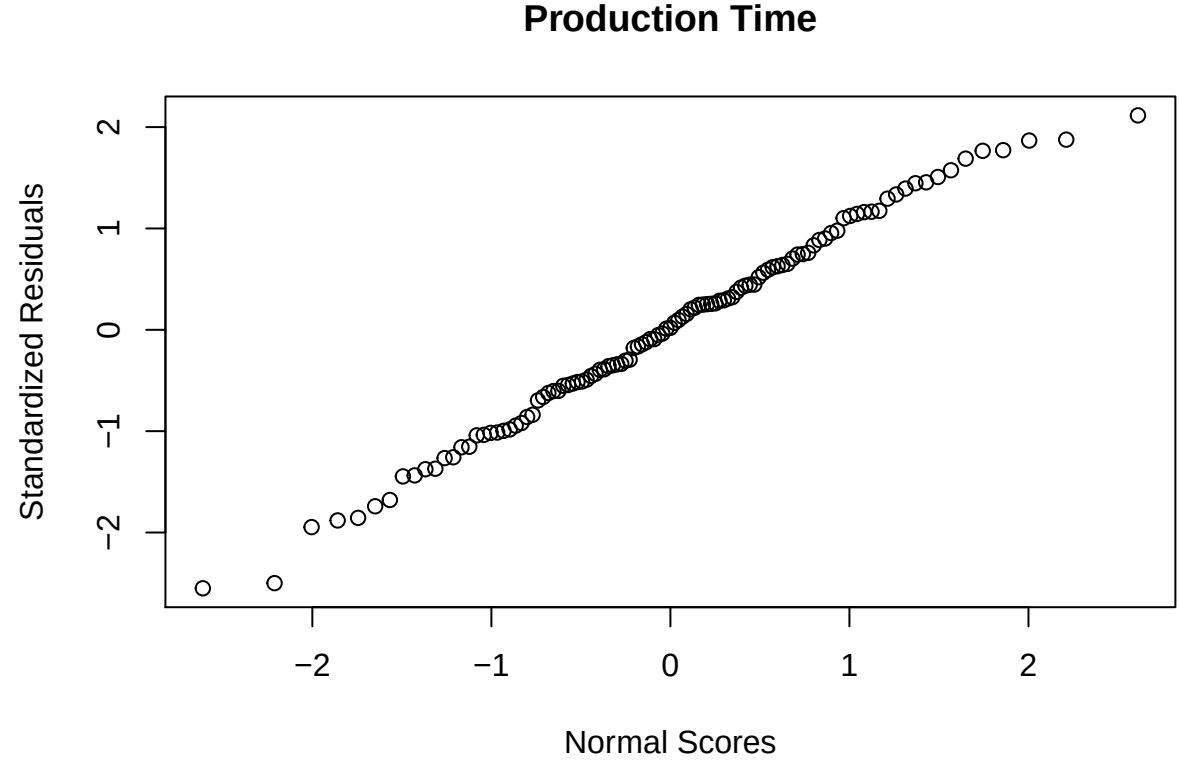
d-) Obtain the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show?

Plot show that error variances are equal and errors are normally distributed.

```
ei=f$residuals
yhat=f$fitted.values
plot(yhat,ei,xlab="Predicted Values",ylab="Residuals")
```




```
error.std = rstandard(f)
qqnorm(error.std,ylab="Standardized Residuals",xlab="Normal Scores",main="Production Time")
```



e-)Express the estimated regression function in the original units.

Y= 1.2547 + 3.6235*$\sqrt{X}$

Problem 3

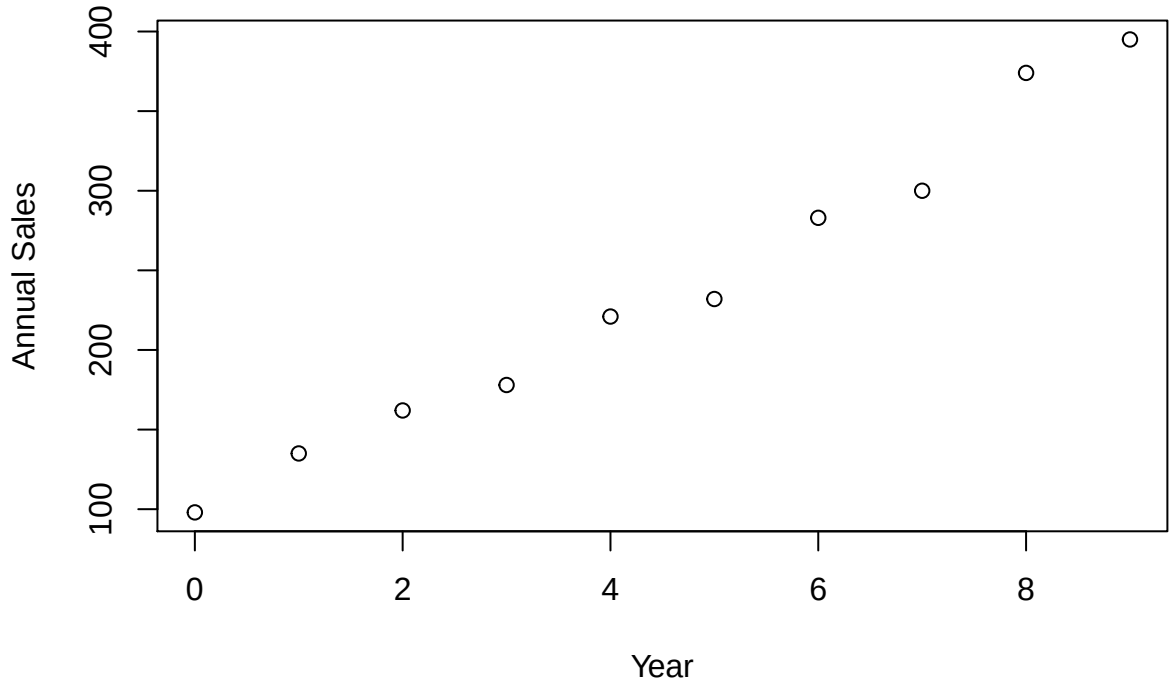
Refer to the Sales growth data. A marketing researcher studied annual sales of a product that had been introduced 10 years ago. The data are as follows, where X is the year (coded) and Y is sales in thousands.(25 points, 5 points each)

Solutions:

a-) Prepare a scatter plot of the data. Does a linear relation appear adequate here? Use the Box-Cox procedure and standardization to find an appropriate power transformation of Y. Evaluate SSE for $\lambda = .3, .4, .5, .6, .7$. What transformation of Y is suggested?

From the scatter plot, a linear relation appears to be adequate. R^2 is 98% and the model is significant. QQ plot indicates departure from the normality. From the boxcox algorithm, $Y^{'} = \sqrt{Y}$ is recommended.

```
SG <- read.csv("/cloud/project/Fall 2020/Sales Growth Data.csv")
plot(SG$X,SG$Y,xlab="Year",ylab="Annual Sales")
```

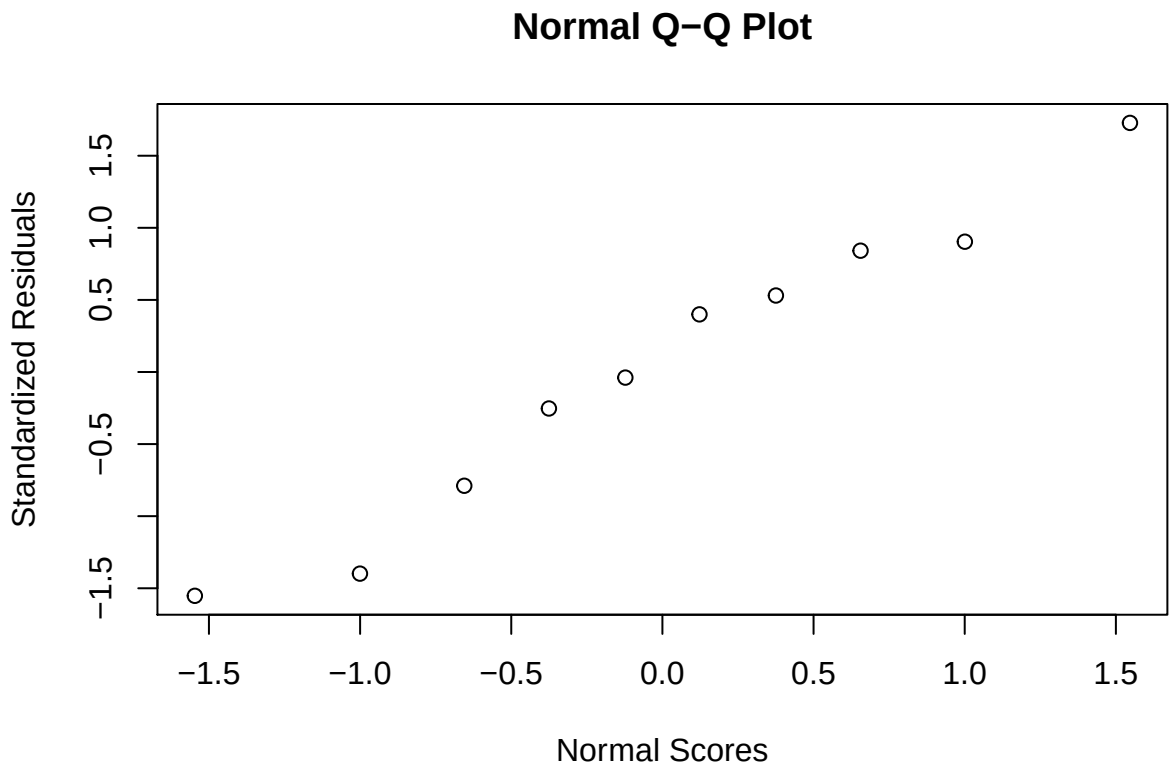



```
library(MASS)
f<-lm(Y~X,data=SG)
summary(f)
```

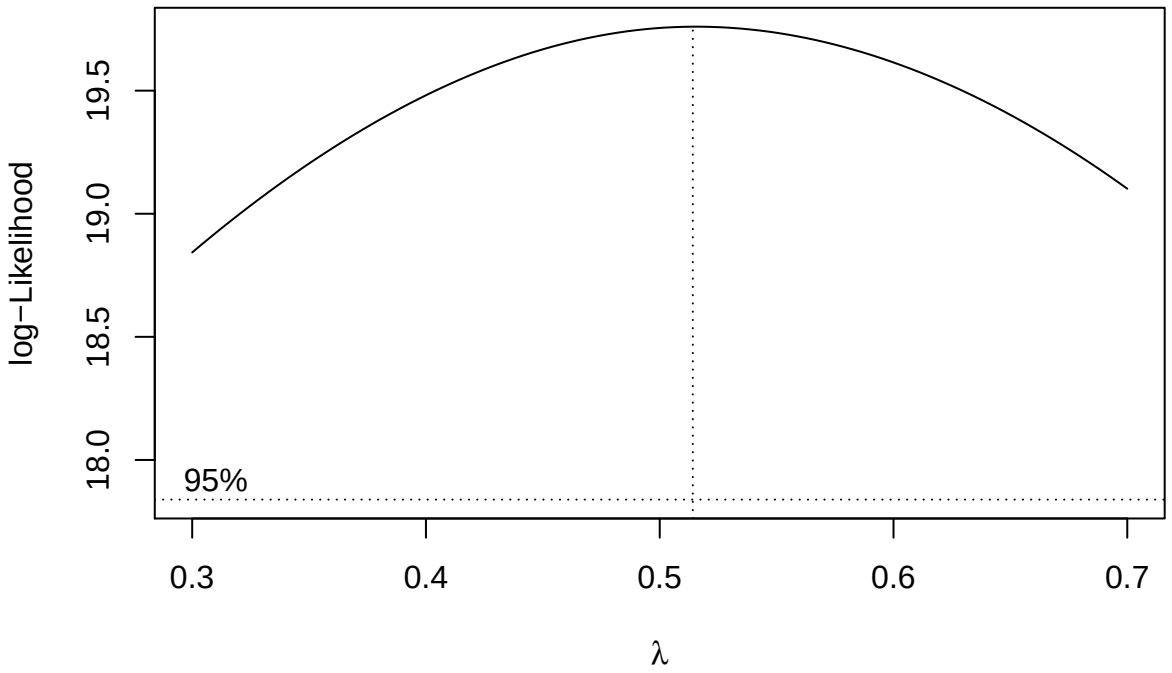
```
##
## Call:
## lm(formula = Y ~ X, data = SG)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.049  -9.177   2.446   9.814  22.461
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   91.564      8.814   10.39 6.38e-06 ***
```

```
## X             32.497      1.651   19.68 4.62e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15 on 8 degrees of freedom
## Multiple R-squared:  0.9798, Adjusted R-squared:  0.9772
## F-statistic: 387.4 on 1 and 8 DF,  p-value: 4.62e-08
```

```
error.std = rstandard(f)
qqnorm(error.std,ylab="Standardized Residuals",xlab="Normal Scores")
```




```
boxcox(f,lambda=seq(0.3,0.7,0.1))
```



b-) Use the transformation $Y^{'} = \sqrt{Y}$ and obtain the estimated linear regression function for the transformed data.

R^2 is 99.9%$ and the model is significant.

Yˆ{'}=10.26093 + 1.07629 X

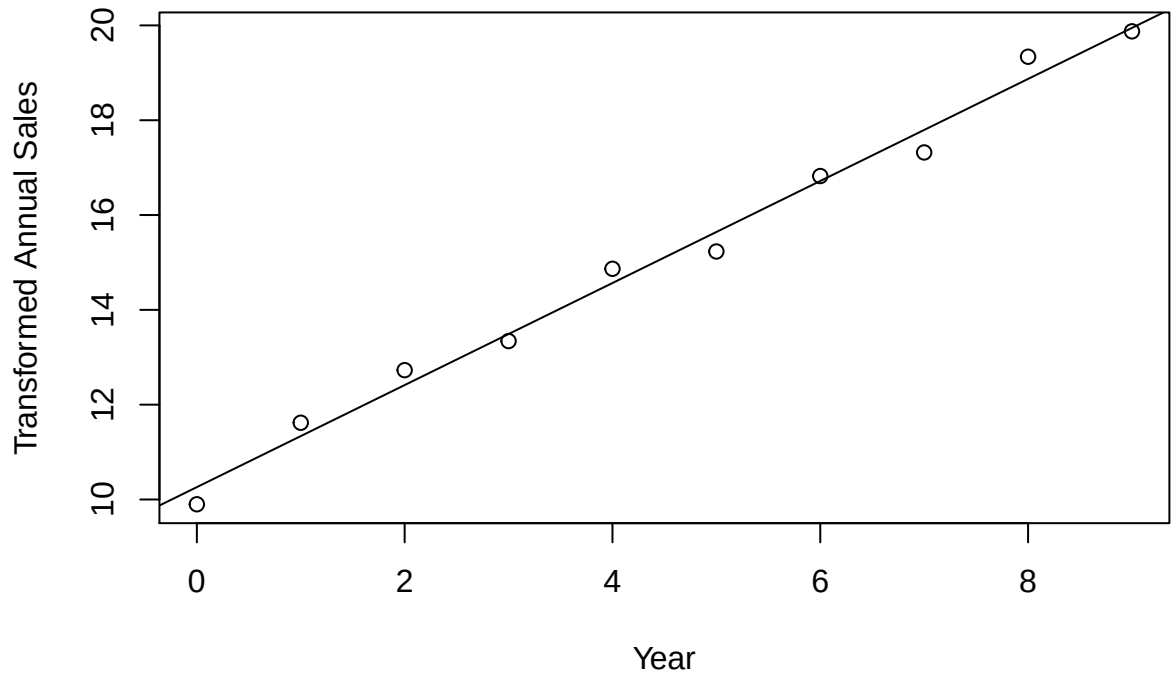
```
f1<-lm(sqrt(Y)~X,data=SG)
summary(f1)
```

```
##
## Call:
## lm(formula = sqrt(Y) ~ X, data = SG)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47447 -0.30811  0.01549  0.29541  0.46781
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.26093    0.21290   48.20 3.80e-11 ***
## X            1.07629    0.03988   26.99 3.83e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3622 on 8 degrees of freedom
## Multiple R-squared:  0.9891, Adjusted R-squared:  0.9878
## F-statistic: 728.4 on 1 and 8 DF,  p-value: 3.826e-09
```

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

Yes, the regression line appear to be a good fit to the transformed data.

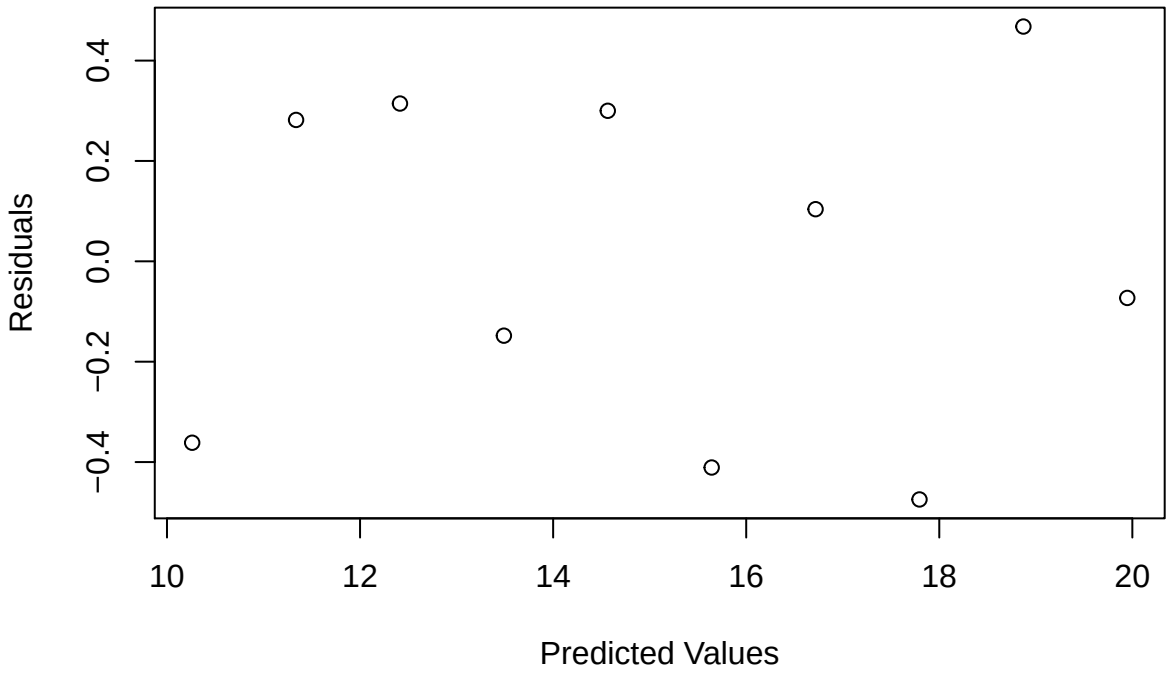
```
plot(SG$X,sqrt(SG$Y),xlab="Year",ylab="Transformed Annual Sales")
abline(f1)
```



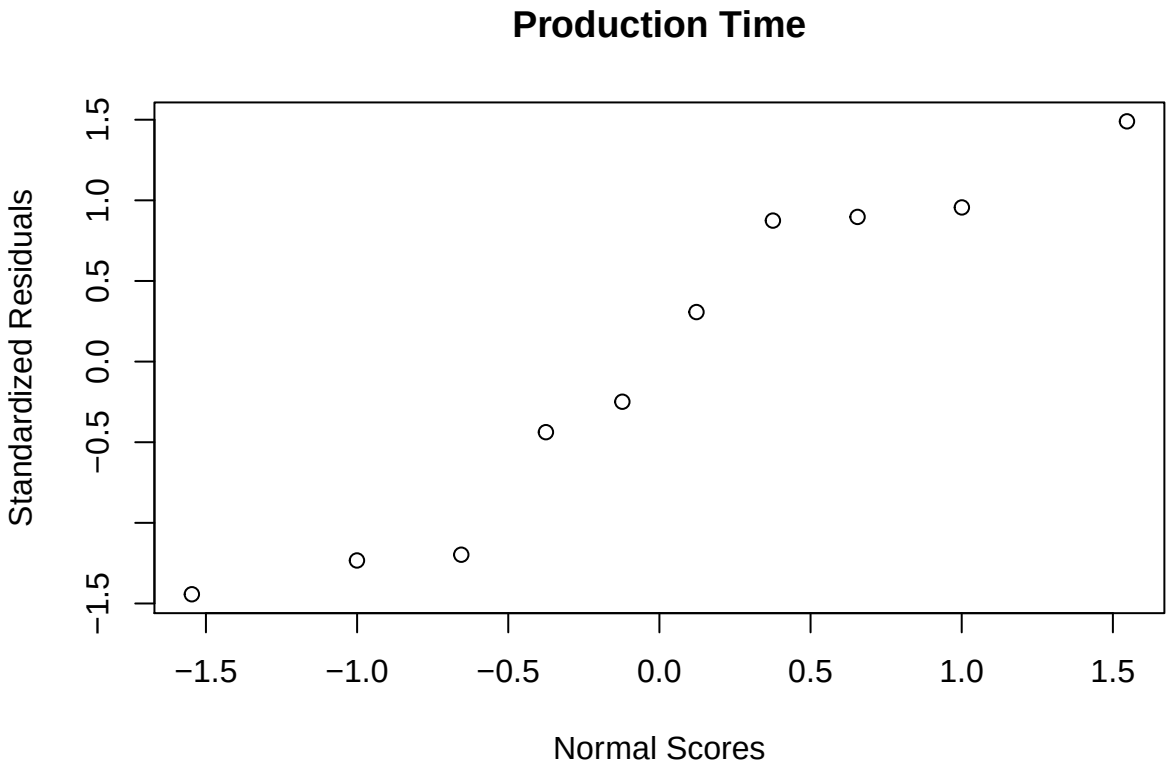
d-) Obtain the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show?

Errors have equal variances but QQ plot indicates heavy tails, further transformation is needed.

```
ei=f1$residuals
yhat=f1$fitted.values
plot(yhat,ei,xlab="Predicted Values",ylab="Residuals")
```




```
error.std = rstandard(f1)
qqnorm(error.std,ylab="Standardized Residuals",xlab="Normal Scores",main="Production Time")
```



e-) Express the estimated regression function in the original units.

$Y = (10.26093 + 1.07629X)^2$

Problem 4

The following data were obtained in a study of the relation between diastolic blood pressure (Y) and age (X) for boys 5 to 13 years old. (25 points)

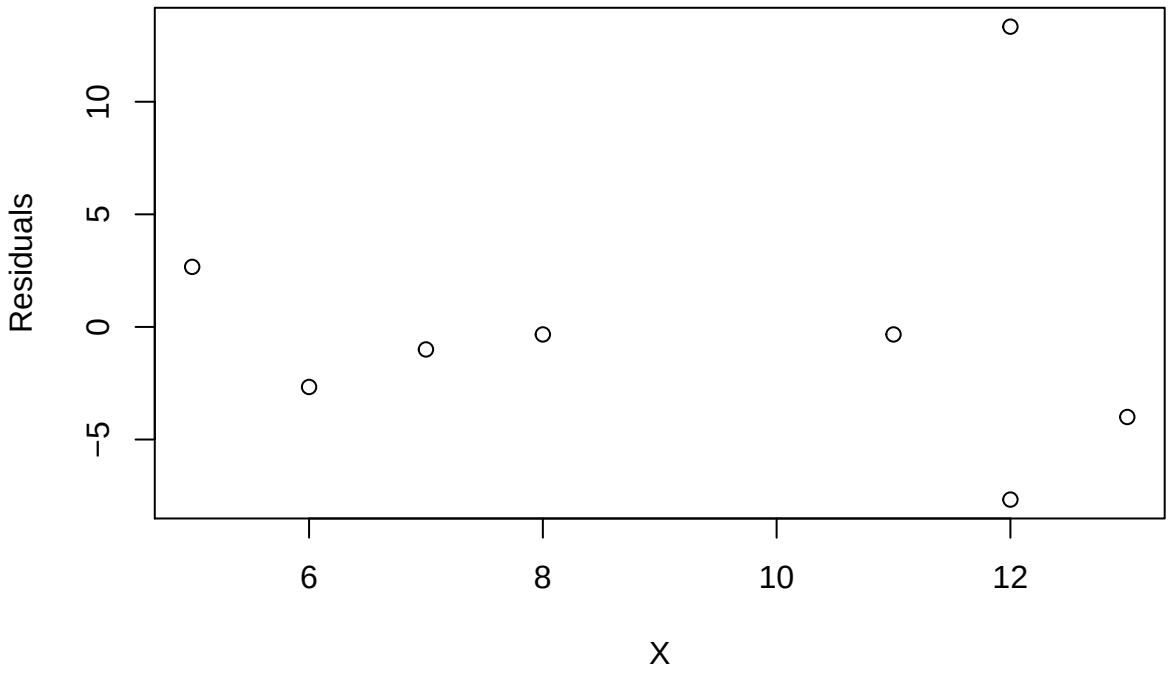
X<-c(5,8,11,7,13,12,12,6) Y<-c(63,67,74,64,75,69,90,60)

Solutions:

R^2 is 58% and the model is significant. Graph shows that there is an outlier in the data.

a-) Assuming normal error regression model is appropriate, obtain the estimated regression function and plot the residuals e_i against X_i. What does your plot residual plot show? (5 points)

```
X<-c(5,8,11,7,13,12,12,6)
Y<-c(63,67,74,64,75,69,90,60)
f<-lm(Y~X)
ei=f$residuals
plot(X,ei,xlab="X",ylab="Residuals")
```




```
summary(f)
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6667 -3.0000 -0.6667  0.4167 13.3333
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  48.6667     7.8869   6.171 0.000832 ***
## X             2.3333     0.8135   2.868 0.028487 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.683 on 6 degrees of freedom
## Multiple R-squared:  0.5783, Adjusted R-squared:  0.508
## F-statistic: 8.228 on 1 and 6 DF,  p-value: 0.02849
```

b-) Omit case 7 from the data and obtain the estimated regression function based on the remaining seven cases. Compare this estimated regression function to that obtained in part (a). What can you conclude about the effect of case 7? (10 points)

After deleting the case 7, R^2 is increased to 82%. the model is significant. This shows that Case 7 is an outlier. The coefficients are below. (Intercept) X [f,] 48.66667 2.333333 [f1,] 53.06796 1.621359

```
f1<-lm(Y[-7]~X[-7])
summary(f1)
```

```
##
## Call:
## lm(formula = Y[-7] ~ X[-7])
```

```
##
## Residuals:
##       1       2       3       4       5       6       7
##  1.8252  0.9612  3.0971 -0.4175  0.8544 -3.5243 -2.7961
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  53.0680     3.2136  16.514 1.49e-05 ***
## X[-7]         1.6214     0.3448   4.702  0.00533 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.645 on 5 degrees of freedom
## Multiple R-squared:  0.8156, Adjusted R-squared:  0.7787
## F-statistic: 22.11 on 1 and 5 DF,  p-value: 0.005327
```

```
rbind(f$coefficients,f1$coefficients)
```

```
##      (Intercept)        X
## [1,]    48.66667 2.333333
## [2,]    53.06796 1.621359
```

c-) Using your fitted regression function in part (b), obtain a 99 percent prediction interval for a new Y observation at X = 12. Does observation Y_7 fall outside this prediction interval? What is the significance of this? (10 points)

The confidence interval is below.

	fit	lwr	upr
1	72.52427	60.31266	84.73588

Y_7 is 90 and falls outside of this interval. This indicate that Y_7 is an outlier.

```
predict(f1,data.frame(X=12),interval = "prediction",level=0.99)
```

```
##        fit      lwr      upr
## 1 72.52427 60.31266 84.73588
```