CSCI E-106:Practice Midterm

Instructions

1-) Open book and open notes exam (textbooks (print or pdf), lecture slides, notes, practice exam, homework solutions, and TA slides, including all Rmd's*).

2-) You are allowed to use RStudio Cloud (https://rstudio.cloud.) , Microsoft Word, Power Point and PDF reader, and canvas on your laptop.

3-) You need to have a camera on your laptop. Proctorio is required to start this exam. If you are prompted for an access code, you must Configure Proctorio on your machine.

4-)Please read the list of recording and restrictions provided by Proctorio carefully before taking the exam

5-)Please pay attention any timing and technical warnings that popped up your screen The exam will be available from Friday October 23rd at 12 pm EST through Monday October 26th at 7:20 pm EST.

6-)Once you start the exam, you have to complete the exam in 3 hours or by Monday October 26th at 7:20 pm EST, whichever comes first.

7-)In order to receive full credit, please provide full explanations and calculations for each questions

8-)Make sure that you are familiar with the procedures for troubleshooting exam issues preview the document. Follow the protocol if there are any issues!

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 PM Q1 Data set. (40 Points)

a-) Build a regression model to predict Y as a function of X. Write down the regression model, Is the regression model significant? (5 points)

The model is significant. Y= -81432.946 + 158.95*X. R^2 is 67%.

```r
PM.Q1 <- read.csv("/cloud/project/Fall 2020/PM Q1.csv")
f.q1.a<-lm(Y~X,data=PM.Q1)
```

b-) Check all the assumptions related to the regression model and perform Brown-Forsythe Test. (10 points)

QQ plot indicates the departure from the normality. Residual vs. Fitted plot indicates hetoradasticity. Shapiro-Wilk normality test:

Ho: Data is normally distributed Ha: Data is NOT normally distributed

P value is $<$ 0.05, Reject Ho. Errors are Not normally distributed.

Ho: Error Variance is constant Ha: Error Variance is Not constant

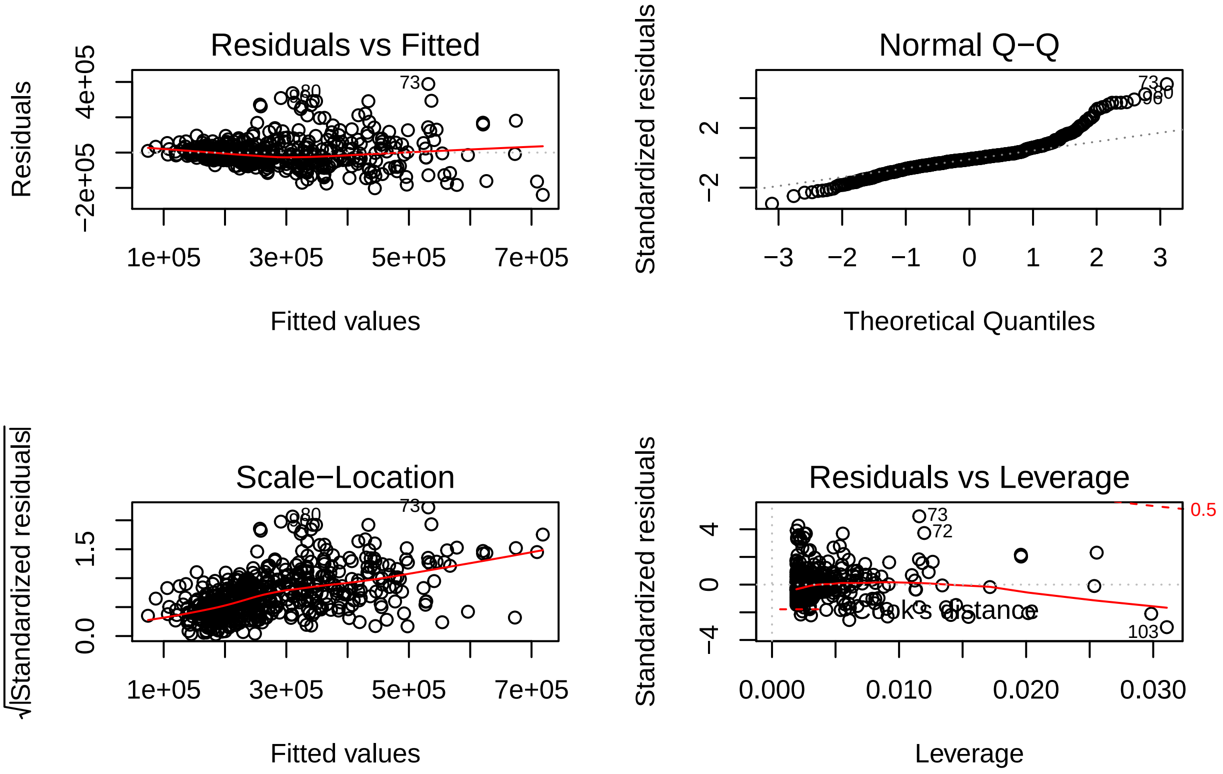
p-value of the test is 0.53, accept the null. Error variance is constant.

```
par(mfrow=c(2,2))
plot(f.q1.a)
ei<-f.q1.a$residuals
shapiro.test(ei)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ei
## W = 0.90048, p-value < 2.2e-16
```

```
#Brown-Forsythe Test using the median X to split the data

library(onewaytests)
```



```
bf.dat<-data.frame(ei,yhat=f.q1.a$fitted.values,ind=as.factor(I(PM.Q1$X<=median(PM.Q1$X))*1))
bf.test(ei~ind,data=bf.dat)
```

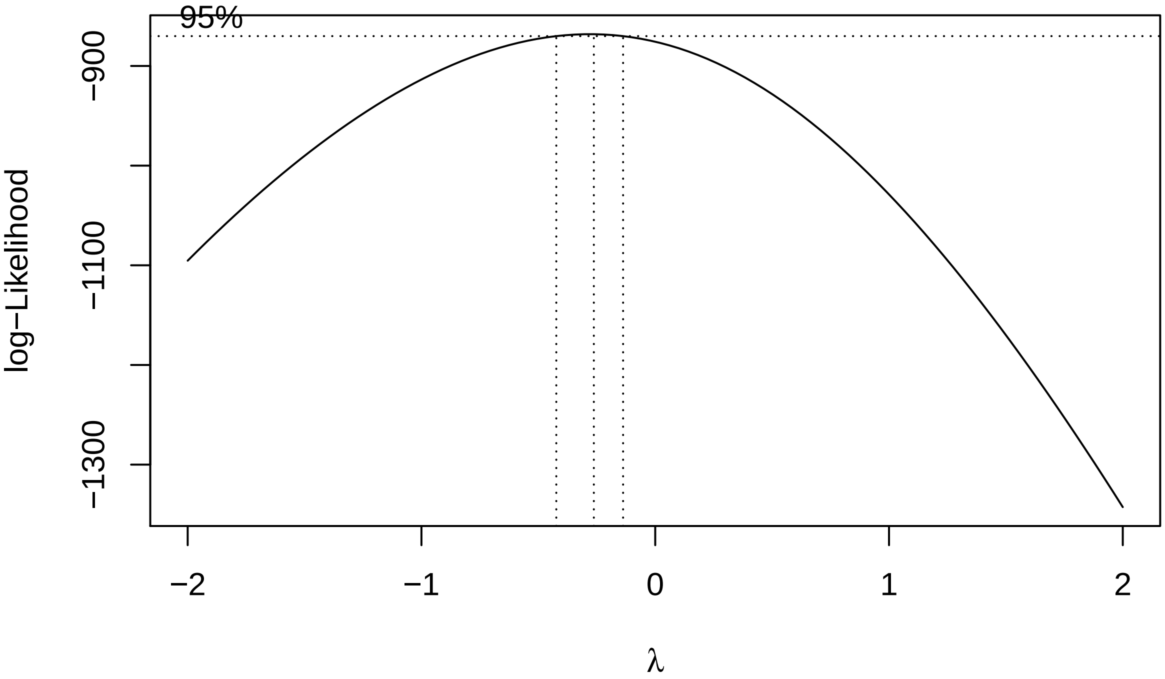
```
##
##   Brown-Forsythe Test (alpha = 0.05)
## -------------------------------------------------------------
##   data : ei and ind
##
##   statistic  : 0.389263
##   num df     : 1
##   denom df   : 312.1958
##   p.value    : 0.5331427
##
##   Result     : Difference is not statistically significant.
## -------------------------------------------------------------
```

c-)Do we need to transform Y? Use Box-Cox procedure to find out appropriate transformation of Y and justify your choice. (10 points)

The Box-Cox procedure is suggesting $\frac{1}{Y^{0.1}}$. However, this is complicated transformation and not easy to explain. As such, i will use the log transformation.

QQ plot looks much better. It is approximately normally distributed. Residual vs.Fitted Values do not indicate heterodasticty.
the new model is log(Y) = 11.28 + 0.0005 X, R^2 is increased to 70%.

```
library(MASS)
par(mfrow=c(1,1))
boxcox(f.q1.a,lambda = seq(-2,2,0.1))
```

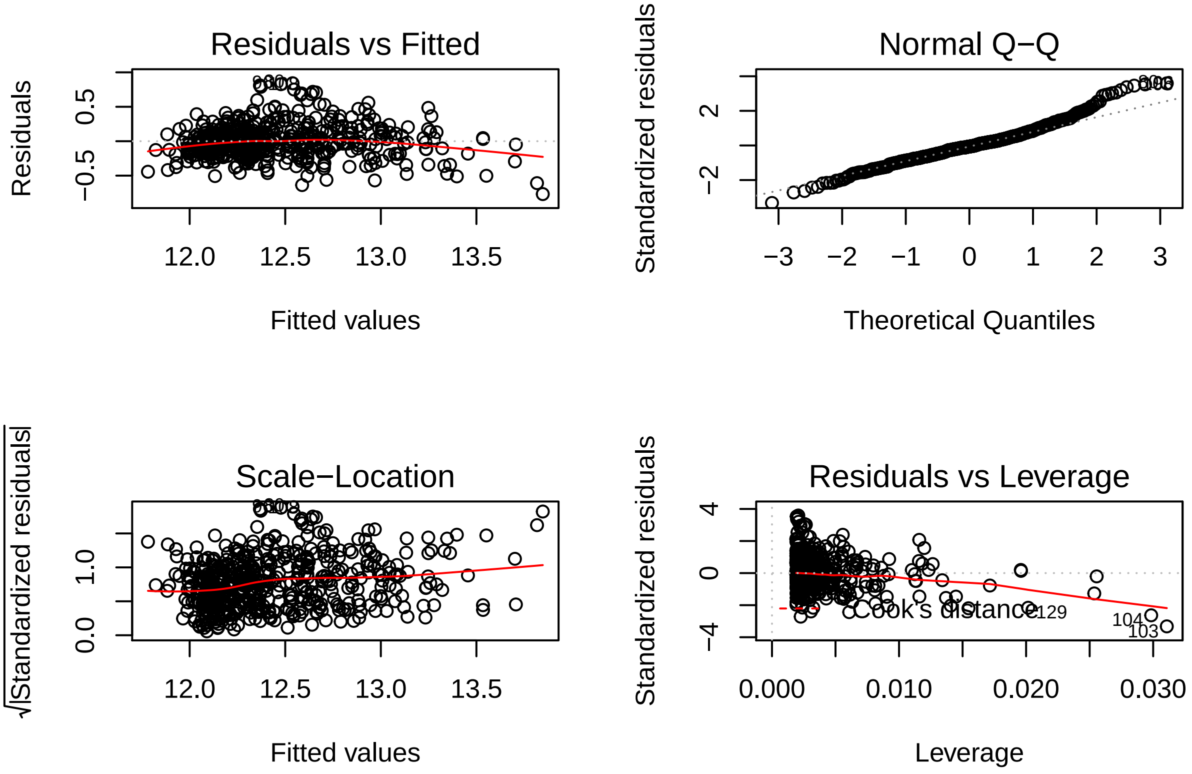


```
f.q1.a1<-lm(log(Y)~X,data=PM.Q1)
summary(f.q1.a1)
```

```
##
## Call:
## lm(formula = log(Y) ~ X, data = PM.Q1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.76772 -0.14834 -0.01448  0.11921  0.84182
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.128e+01  3.427e-02  329.24   <2e-16 ***
## X           5.097e-04  1.446e-05   35.24   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2347 on 520 degrees of freedom
## Multiple R-squared:  0.7049, Adjusted R-squared:  0.7043
## F-statistic:  1242 on 1 and 520 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(f.q1.a1)
```



d-) Use the final model to predict Y for three new X values, 1100, 3000 and 4900. which methods would use you to calculate the joint 90% confidence intervals? Justify your choice and calculate the confidence interval using the final model in part c. (15 points)

Log transformation is used for the final. We need to transform it back to original scale. The predicted values are 139,115.5 366,399.8 965,016.8

Bonferroni procedure will yield somewhat tighter prediction limits as (B   S). Please see below for the confidence intervals.

$84,160.65 \leq \hat{Y}_1 \leq 229,954.5 \; 221,828.97 \leq \hat{Y}_2 \leq 605,190.5 \; 580,722.72 \leq \hat{Y}_3 \leq 1,603,617.9$

```
X<-c(1100,3000,4900)
pred<-predict.lm(f.q1.a1,data.frame(X = c(X)),se.fit=TRUE)
fit<-exp(pred$fit)
fit
```

```
##        1        2        3
## 139115.5 366399.8 965016.8
```

```
B=qt(1-0.1/(2*3),520)
S=sqrt(3*qf(0.90,3,520))
cbind(B,S)
```

```
##             B        S
## [1,] 2.133716 2.506601
```

```
s.pred<-sqrt(pred$se.fit^2+pred$residual.scale^2)
exp(cbind(pred$fit-B*s.pred,pred$fit+B*s.pred))
```

```
##        [,1]      [,2]
## 1  84160.65  229954.5
## 2 221828.97  605190.5
## 3 580722.72 1603617.9
```

Problem 2

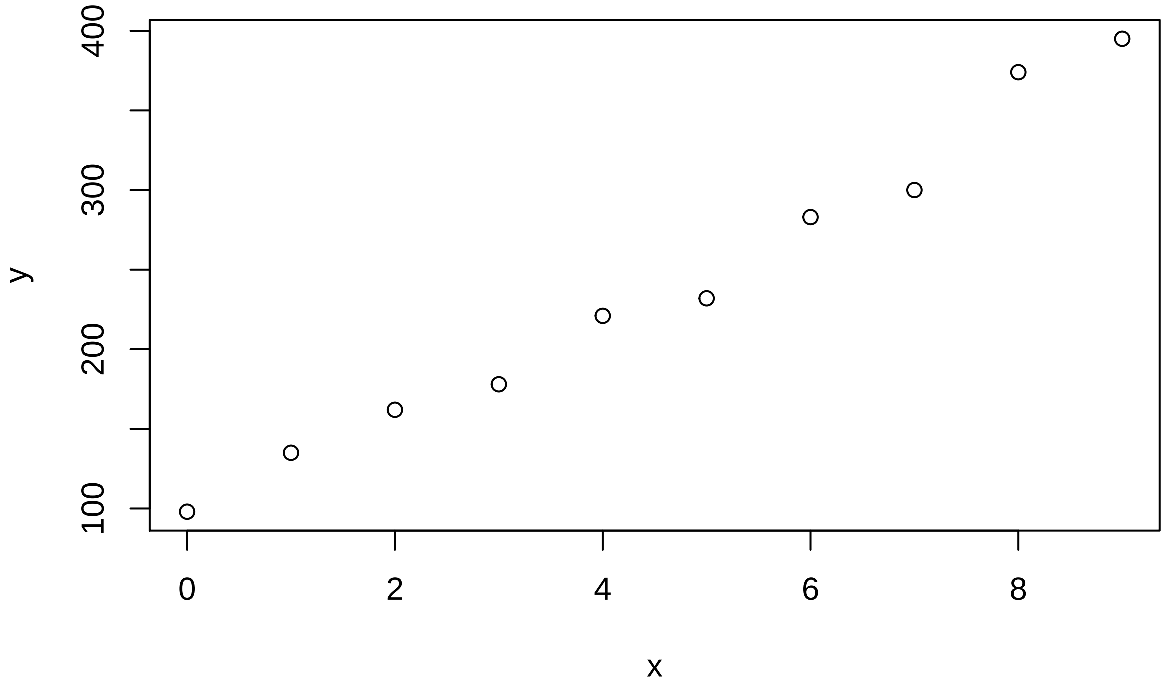
Copy and paste the data below in R. (30 points)

y=c(98,135,162,178,221,232,283,300,374,395)\ x=c(0,1,2,3,4,5,6,7,8,9)

a-) is the linear fit appropriate? If not, transform the data and find an appropriate fit. Comment on the model and regression model assumptions. (20 points)

The scatter plot indicates linear relationship between x and Y. However, QQ plot indicates a slight s function, which indicates slight departure from the normal distribution. Boxcox transfomration indicates, square root transformation. The model is Y=91.564 + 32.497X and R^2 is 98%. There is a slight increase of the power of the model, R^2 is 99% and the model is $\sqrt{Y} = 10.26 + 1.08X$.QQ plot shows the S shape, this is could be due to low sample size (10). However the final model is $\sqrt{Y} = 10.26 + 1.08X$.

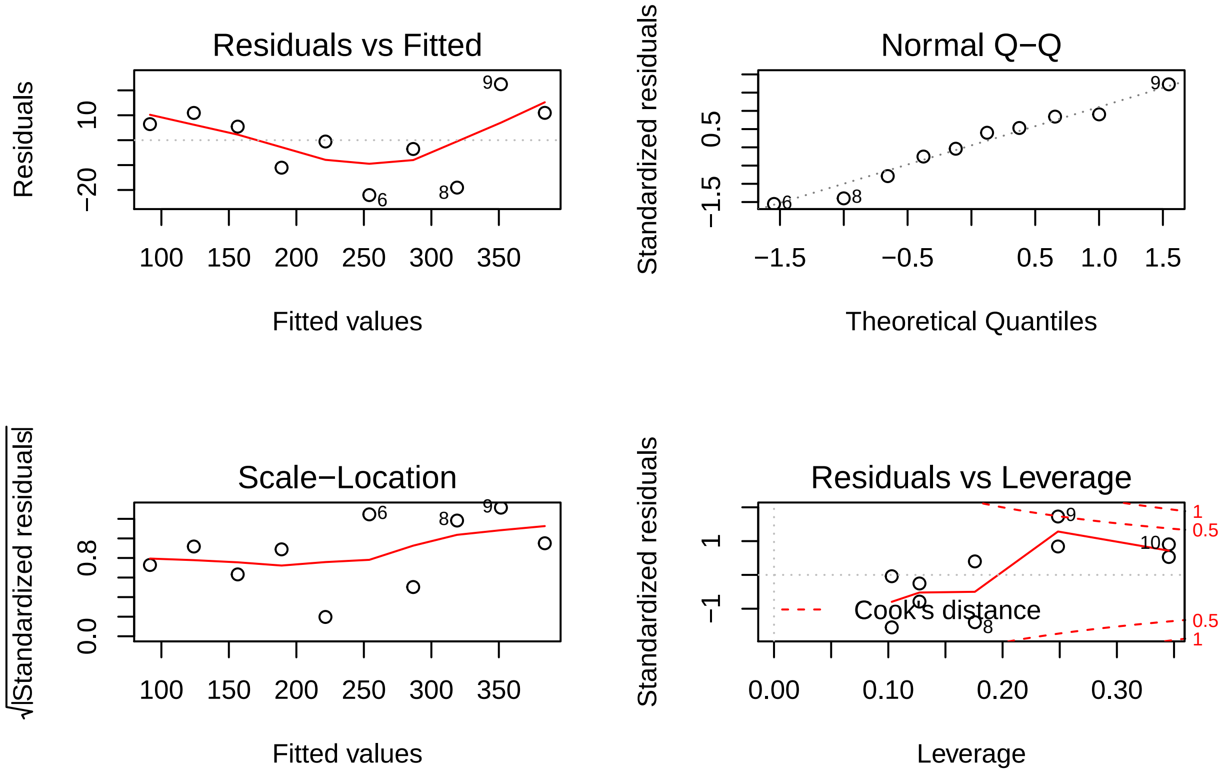
```
y=c(98,135,162,178,221,232,283,300,374,395)
x=c(0,1,2,3,4,5,6,7,8,9)
par(mfrow=c(1,1))
plot(x,y)
```



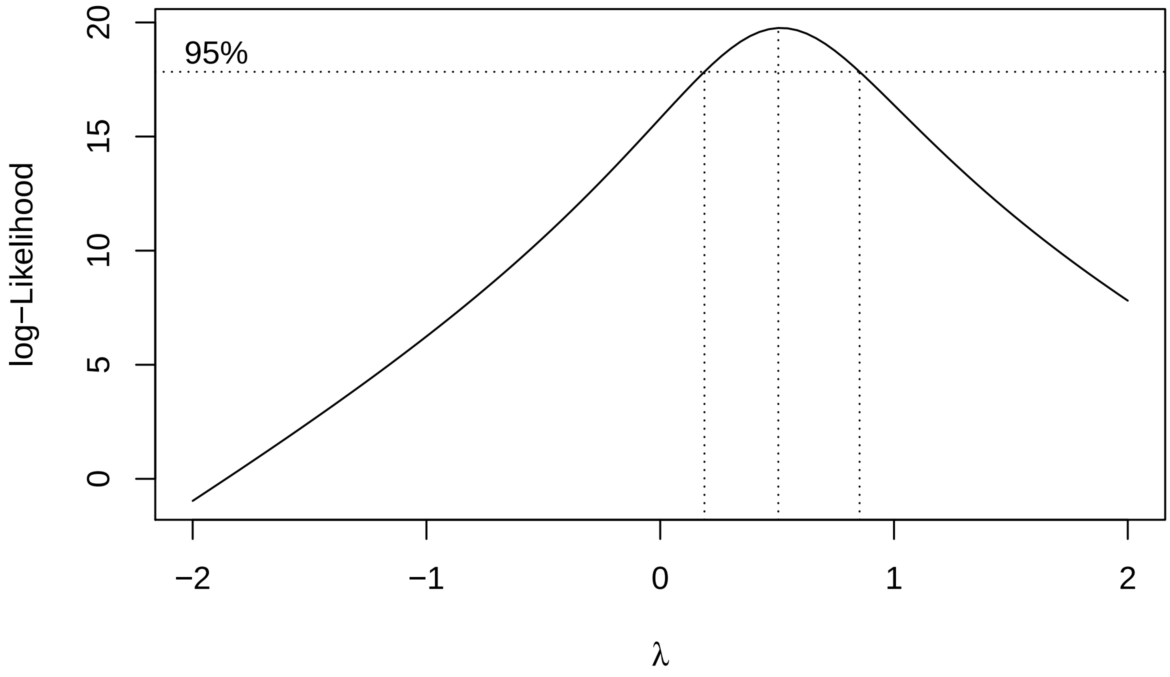
```
f2<-lm(y~x)
summary(f2)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## 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
```

```
par(mfrow=c(2,2))
plot(f2)
```



```r
par(mfrow=c(1,1))
boxcox(f2,lambda=seq(-2,2,0.1))
```



```r
f21<-lm(sqrt(y)~x)
summary(f21)
```

```
##
```

```
## Call:
## lm(formula = sqrt(y) ~ x)
##
## 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
```

b-) Predict Y when x=10, and calculate the prediction confidence interval for 90% confidence level (10 points)

See below, we need to transform it back to original scale.

```r
pred<-predict.lm(f21,data.frame(x = 10),interval =  "prediction",level=0.90)
pred^2
```

```
##        fit      lwr      upr
## 1 442.0022 408.3675 476.9679
```

Problem 3

Refer to the PM Q3 Data set. (30 Points)

Perform one factor analysis by finding the best variable to explain Y. Fit one variable regression model with Y as a dependent variable against remaining variables, as an independent variable one at time. Choice the best variable explain Y and comment on the QQ plot and error vs. fitted values graph for the model assumptions.

All variables are significant. The R^2s are below.

model with x1: R^2 is 0.2309 model with x2: R^2 is 0.692 model with x3: R^2 is 0.2836 model with x4: R^2 is 0.1281 model with x5: R^2 is 0.547 model with x6: R^2 is 0.3867

x2 is the most signficiant variable as it has the highest R^2. QQ plot roughly looks normal, however, observation 34 looks like an outlier. Residual vs. Fitted plot indicates heterodasticity and more test needed to for this.

```r
PM.Q3 <- read.csv("/cloud/project/Fall 2020/PM Q3.csv")
f1<-lm(y~x1,data=PM.Q3)
f2<-lm(y~x2,data=PM.Q3)
f3<-lm(y~x3,data=PM.Q3)
f4<-lm(y~x4,data=PM.Q3)
f5<-lm(y~x5,data=PM.Q3)
f6<-lm(y~x6,data=PM.Q3)
summary(f1)
```

```
##
## Call:
## lm(formula = y ~ x1, data = PM.Q3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -81.198 -32.819   0.367  20.865  68.795
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   95.405     22.346   4.269 0.000126 ***
## x1             7.069      2.093   3.377 0.001702 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.09 on 38 degrees of freedom
## Multiple R-squared:  0.2309, Adjusted R-squared:  0.2106
## F-statistic: 11.41 on 1 and 38 DF,  p-value: 0.001702
```

```
summary(f2)
```

```
##
## Call:
## lm(formula = y ~ x2, data = PM.Q3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -76.504 -10.248   1.365  11.301  43.470
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    44.86      13.86   3.237  0.00251 **
## x2            202.75      21.94   9.240 2.93e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.47 on 38 degrees of freedom
## Multiple R-squared:  0.692,  Adjusted R-squared:  0.6839
## F-statistic: 85.38 on 1 and 38 DF,  p-value: 2.926e-11
```

```
summary(f3)
```

```
##
## Call:
## lm(formula = y ~ x3, data = PM.Q3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -71.662 -25.127   0.554  25.618  78.102
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    93.68      20.04   4.676 3.63e-05 ***
## x3             55.84      14.40   3.879 0.000404 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.79 on 38 degrees of freedom
## Multiple R-squared:  0.2836, Adjusted R-squared:  0.2648
## F-statistic: 15.04 on 1 and 38 DF,  p-value: 0.000404
```

```
summary(f4)
```

```
##
## Call:
## lm(formula = y ~ x4, data = PM.Q3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -79.229 -26.205  -3.608  24.805  80.576
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  146.320     11.177  13.091 1.17e-15 ***
## x4             6.136      2.596   2.363   0.0233 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.49 on 38 degrees of freedom
## Multiple R-squared:  0.1281, Adjusted R-squared:  0.1052
## F-statistic: 5.585 on 1 and 38 DF,  p-value: 0.02333
```

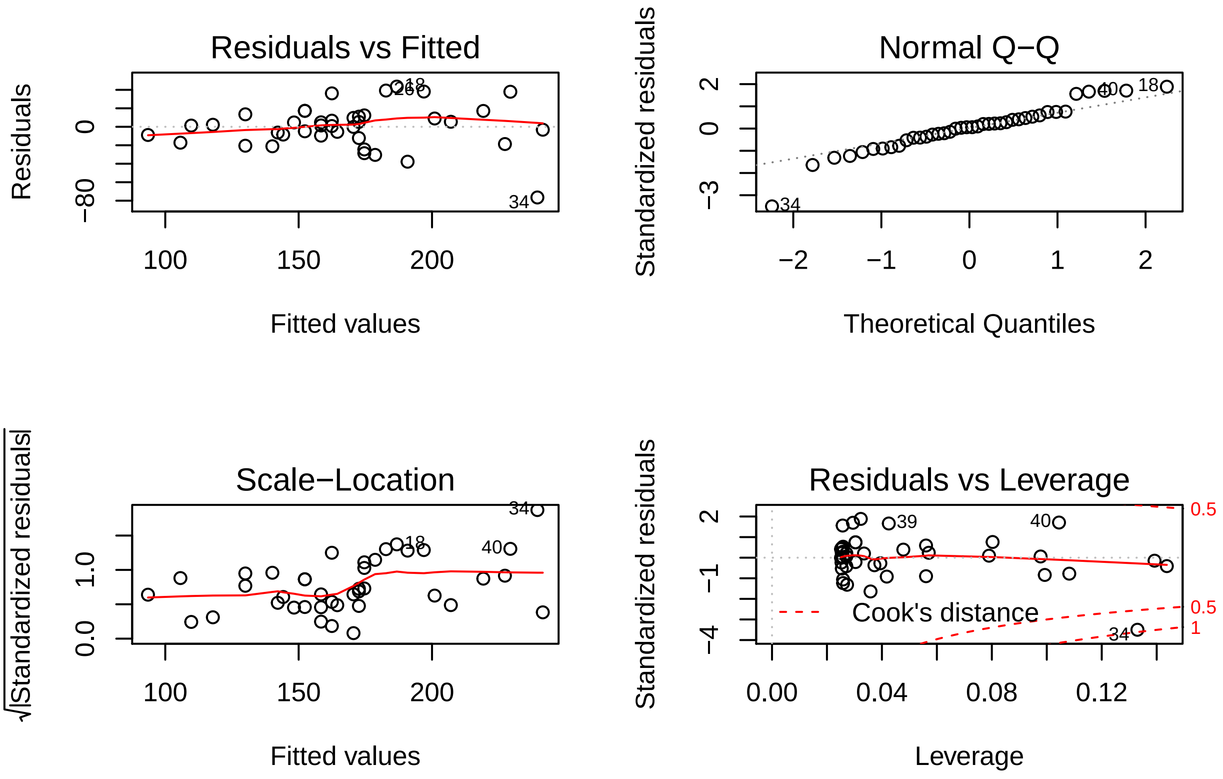
```
summary(f5)
```

```
##
## Call:
## lm(formula = y ~ x5, data = PM.Q3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.581 -19.592  -0.337  20.251  72.320
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    52.61      17.65   2.981    0.005 **
## x5            167.44      24.72   6.774 4.97e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.46 on 38 degrees of freedom
## Multiple R-squared:  0.547,  Adjusted R-squared:  0.5351
## F-statistic: 45.89 on 1 and 38 DF,  p-value: 4.973e-08
```

```
summary(f6)
```

```
##
## Call:
## lm(formula = y ~ x6, data = PM.Q3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -62.203 -16.180   1.339  18.162 109.671
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 133.4907     8.8189  15.137  < 2e-16 ***
## x6            1.1692     0.2388   4.895 1.84e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.12 on 38 degrees of freedom
## Multiple R-squared:  0.3867, Adjusted R-squared:  0.3706
## F-statistic: 23.96 on 1 and 38 DF,  p-value: 1.84e-05
```

```r
par(mfrow=c(2,2))
plot(f2)
```



Residuals vs Fitted
Residuals
0
−80
18
26
34
100
150
200
Fitted values
Normal Q−Q
Standardized residuals
2
0
−3
40
18
34
−2
−1
0
1
2
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
1.0
0.0
18
34
40
100
150
200
Fitted values
Residuals vs Leverage
Standardized residuals
2
−1
−4
39
40
34
Cook's distance
0.5
0.5
1
0.00
0.04
0.08
0.12
Leverage