Fall 2019 Midterm Solutions

Question 1

1-) The regression model we would like to study is:

$$Y_i = \beta_0 + \varepsilon_i$$

a-)write the likelihood function

$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(Y_i - \beta_0 - \lambda)^2}{2\sigma^2}\right)$$

b-)find the MLE estimations for b_0 and σ^2

Log-likelihood function

$$LogL = -\frac{n}{2}log2\pi - \frac{n}{2}log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \lambda)^2$$

$$\frac{\partial L}{\partial \beta_0} = \sum_{i=1}^{n}(Y_i - \beta_0 - \lambda) = 0$$

$$\beta_0 = \sum_{i=1}^{n} Y_i - n\beta_0 - n\lambda = 0$$

$$\beta_0 = \bar{Y} - \lambda$$

$$\frac{\partial L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(Y_i - \beta_0 - \lambda)^2 = 0$$

$$\sigma^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \lambda)^2}{n}$$

Problem 2

a-)Fit a regression model to predict Y. Obtain, the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show? (5pts)

b-) Calculate the simultaneous 90% confidence interval for β_0,and β_1 (5pts)

c-) Calculate the simultaneous 90% confidence intervals for the predicted new X values for 85 and 90. (5 pts)

d-) Conduct the Brown-Forsythe test to determine whether or not the error variance varies with the level of X. (10 pts)

e-) Use the Box-Cox procedure to find an appropriate power transformation and perform the transformation. Obtain, the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show? (15 pts)

a-)Fit a regression model to predict Y. Obtain, the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show? (5pts)

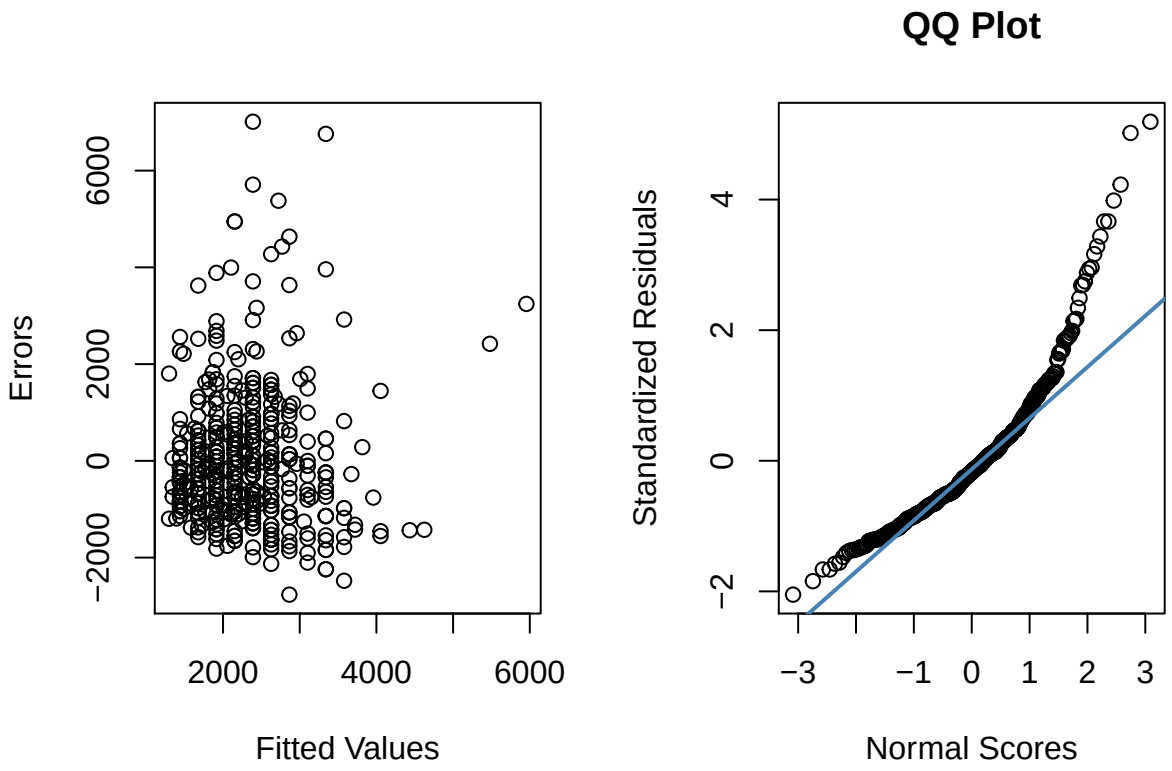
```
library(knitr)
question2 <- read.csv("/cloud/project/question2.csv")
f2<-lm(y~x,data=question2)
summary(f2)
```

```
##
## Call:
## lm(formula = y ~ x, data = question2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2765.3  -889.8  -239.8   536.8  7010.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1201.124    123.325    9.74   <2e-16 ***
## x             47.549      4.652   10.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1352 on 494 degrees of freedom
## Multiple R-squared:  0.1745, Adjusted R-squared:  0.1729
## F-statistic: 104.5 on 1 and 494 DF,  p-value: < 2.2e-16
```

```
anova(f2)
```

```
## Analysis of Variance Table
##
## Response: y
##            Df    Sum Sq   Mean Sq F value    Pr(>F)
## x           1 190820566 190820566  104.46 < 2.2e-16 ***
## Residuals 494 902418143   1826757
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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



Solution: R Square is 17% and the model is significant. Error vs. Fitted values graph indicates unequal variances. QQ plot indicates departures from normality.

b-) Calculate the simultaneous 90% confidence interval for b_0,and b_1 (5pts)

Solution: See below for the simultaneous 90% confidence interval for
b_0,and b_1

```
confint(f2,level=1-0.10/2)
```

```
##                 2.5 %    97.5 %
## (Intercept) 958.81911 1443.4296
## x            38.40798   56.6894
```

c-) Calculate the simultaneous 90% confidence intervals for the predicted new X values for 85 and 90. (5 pts)

Solution: see below for the simultaneous 90% confidence intervals for the predicted new X values for 85 and 90.

```
Xh<-c(85,90)
predict(f2,data.frame(x= c(Xh)),interval = "prediction", level = 1-0.10/2)
```

```
##        fit      lwr      upr
## 1 5242.763 2524.947 7960.580
## 2 5480.507 2752.805 8208.208
```

d-) Conduct the Brown-Forsythe test to determine whether or not the error variance varies with the level of X. (10 pts)

Solution:

H_0: the error variance is constant

H_A:the error variance is NOT constant

Based on the Brown Forsythe test, p-value is 0.00116009. Reject H_0, and conclude H_1, the error variance is NOT constant.

```
ei<-f2$residuals
M=median(question2$x)
DM<-data.frame(cbind(question2$y,question2$x,ei))
DM1<-DM[DM[,2]< M,]
DM2<-DM[DM[,2]>=M,]

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] -3.267609
```

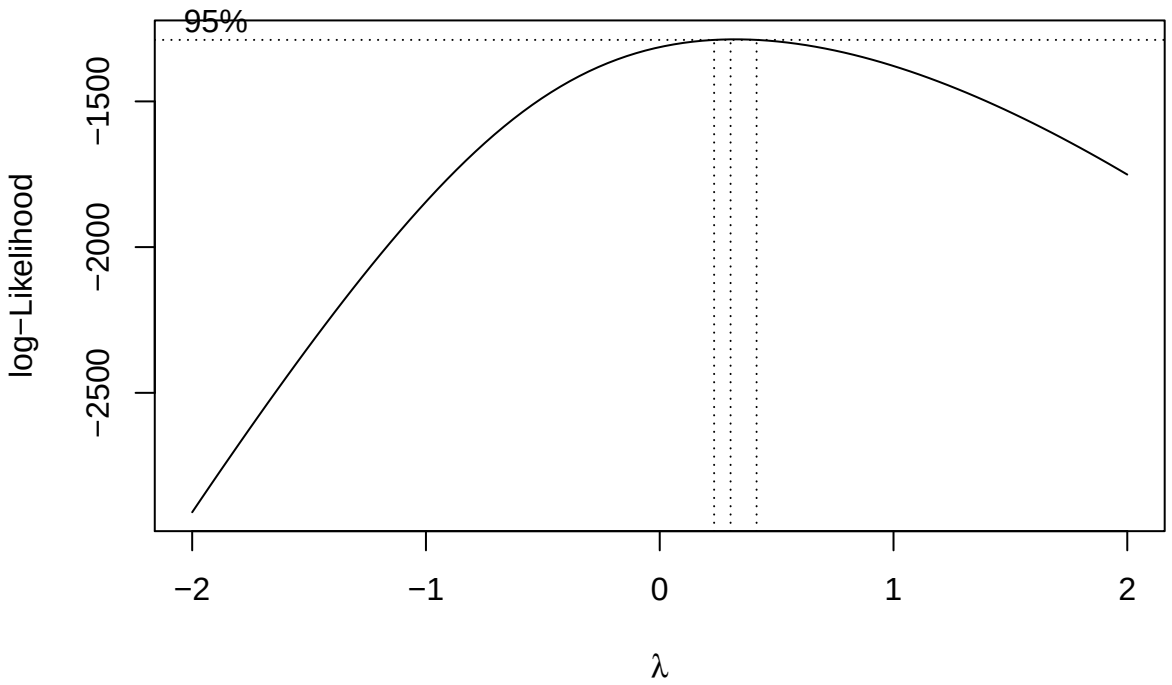
```
2*pt(T,df=N1+N2-2)
```

```
## [1] 0.00116009
```

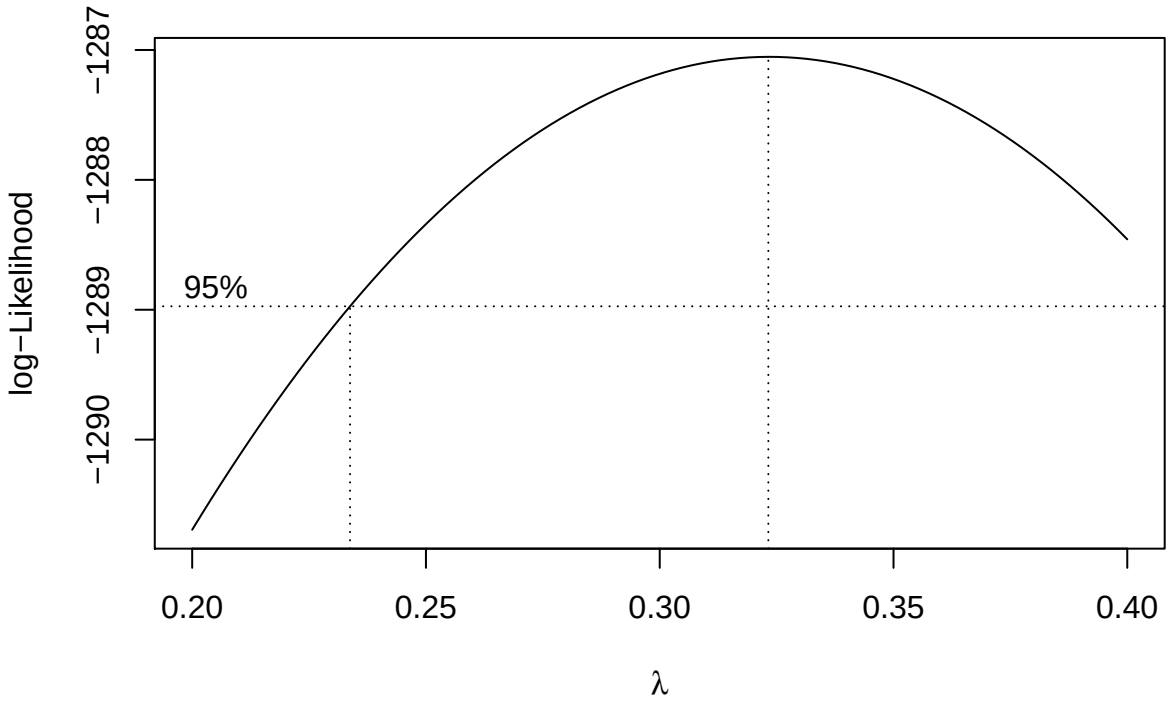
e-) Use the Box-Cox procedure to find an appropriate power transformation and perform the transformation. Obtain, the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show? (15 pts)

Solution: it looks like λ is 0.32. However, you could take λ 0.5 (square root transformation) or λ is 0 (log transformation) to make it easier. We used λ 0.32. After transformation, QQ plots look normal. However, error vs predicted graph still shows a V shape. Unequal variances still persists.

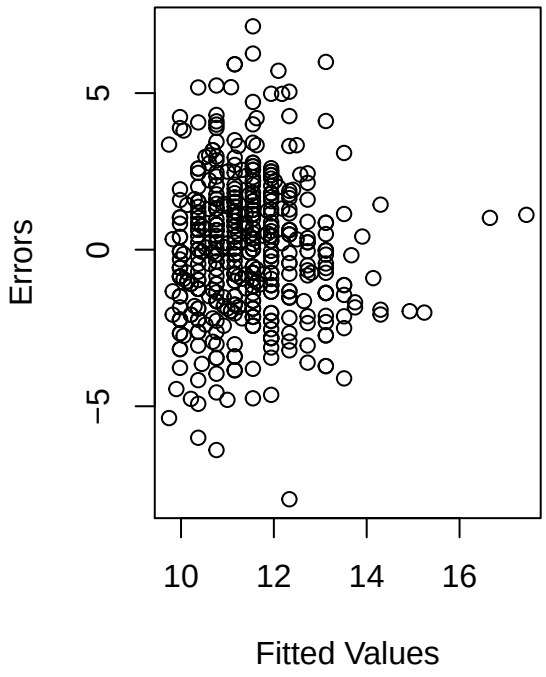
```
library(MASS)
boxcox(f2,lambda=seq(-2,2,by=0.1))
```

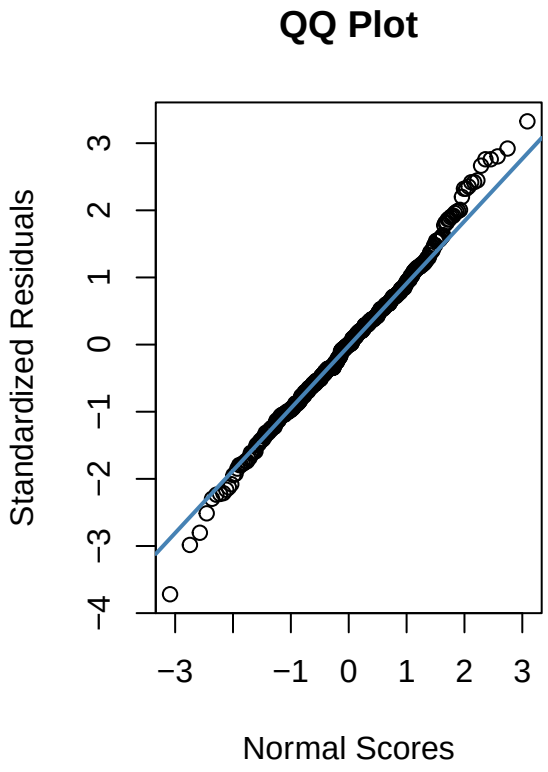



```
boxcox(f2,lambda=seq(0.2,0.4,by=0.1))
```




```
f2.1<-lm(y^0.32~x,data=question2)
ei<-f2.1$residuals
yhat<-f2.1$fitted.values
par(mfrow=c(1,2))
plot(yhat,ei,ylab="Errors",xlab="Fitted Values")
stdei<- rstandard(f2.1)
qqnorm(stdei,ylab="Standardized Residuals",xlab="Normal Scores", main="QQ Plot")
qqline(stdei,col = "steelblue", lwd = 2)
```





Problem 3

Refer to the question 2 data. (25 pts)

a) Create development sample and hold out sample. Development sample is a random sample of 70% of the data and hold out sample is the remainder 30% of the data. Use set.seed(1023) to select the samples. (10 pts)

b) Build the model on the development sample (a random sample of 70% of the data and use set.seed(1023) to select the sample). Obtain, the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show? (5 pts)

c) Calculate R Square on the hold out sample (hint: calculate SSE, SSR and SST on the hold out sample). (10 pts)

a) Create development sample and hold out sample. Development sample is a random sample of 70% of the data and hold out sample is the remainder 30% of the data. Use set.seed(1023) to select the samples. (10 pts)

Solution: please see below

```
set.seed(1023)
ind <- sample(1:nrow(question2), size =nrow(question2)*0.70)
dev <- question2[ind,]
holdout <- question2[-ind,]
```

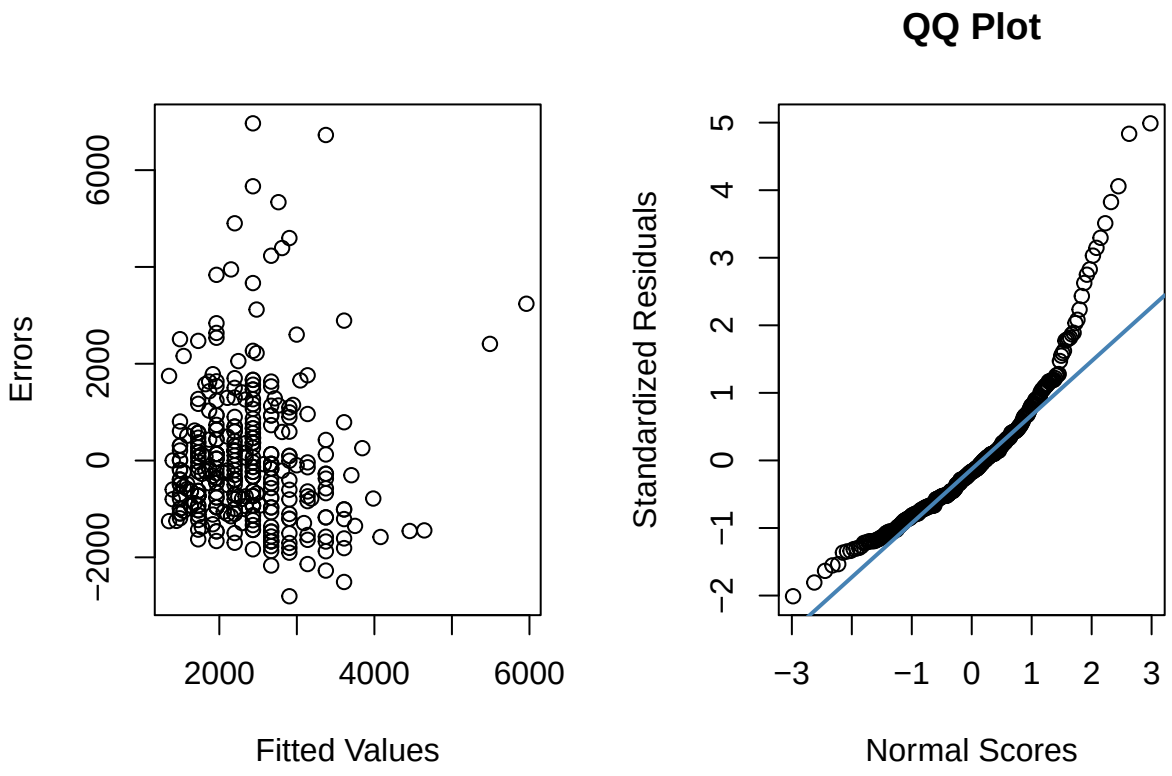
b) Build the model on the development sample (a random sample of 70% of the data and use set.seed(1023) to select the sample). Obtain, the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show? (5 pts)

Solution: R Square is 19% and the model is significant. Error vs. Fitted values graph indicates unequal variances. QQ plot indicates departures from normality. Same conclucions as question 2 par a.

```
f2.2<-lm(y~x,data=dev)
summary(f2.2)
```

```
##
## Call:
## lm(formula = y ~ x, data = dev)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2803.6  -933.3  -233.3   572.1  6966.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1257.562    146.831   8.565 3.65e-16 ***
## x             47.030      5.469   8.599 2.86e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1398 on 345 degrees of freedom
## Multiple R-squared:  0.1765, Adjusted R-squared:  0.1741
## F-statistic: 73.94 on 1 and 345 DF,  p-value: 2.858e-16
```

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



c) Calculate R Square on the hold out sample (hint: calculate SSE, SSR and SST on the hold out sample). (10 pts)

```
SST<-var(holdout$y)*(length(holdout$y)-1)
yhat<-predict(f2.2,holdout)
ei<-holdout$y-yhat
SSE<-sum(ei^2)
R.SQ=1-(SSE/SST)
R.SQ
```

```
## [1] 0.158099
```

Solution: R Square is 12.7%. it decreased from 19% on the holdout sample. Indicating that the model perfomance is not stable.

Problem 4.

Refer to question 4 data set. (15 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? (5 pts)

b)Conduct the Breusch-Pagan test to determine whether or not the error variance varies with the level of X. Use $alpha = .05$. State the alternatives. decision rule, and conclusion. Is your conclusion consistent with your preliminary findings in part (a)? (10 pts)

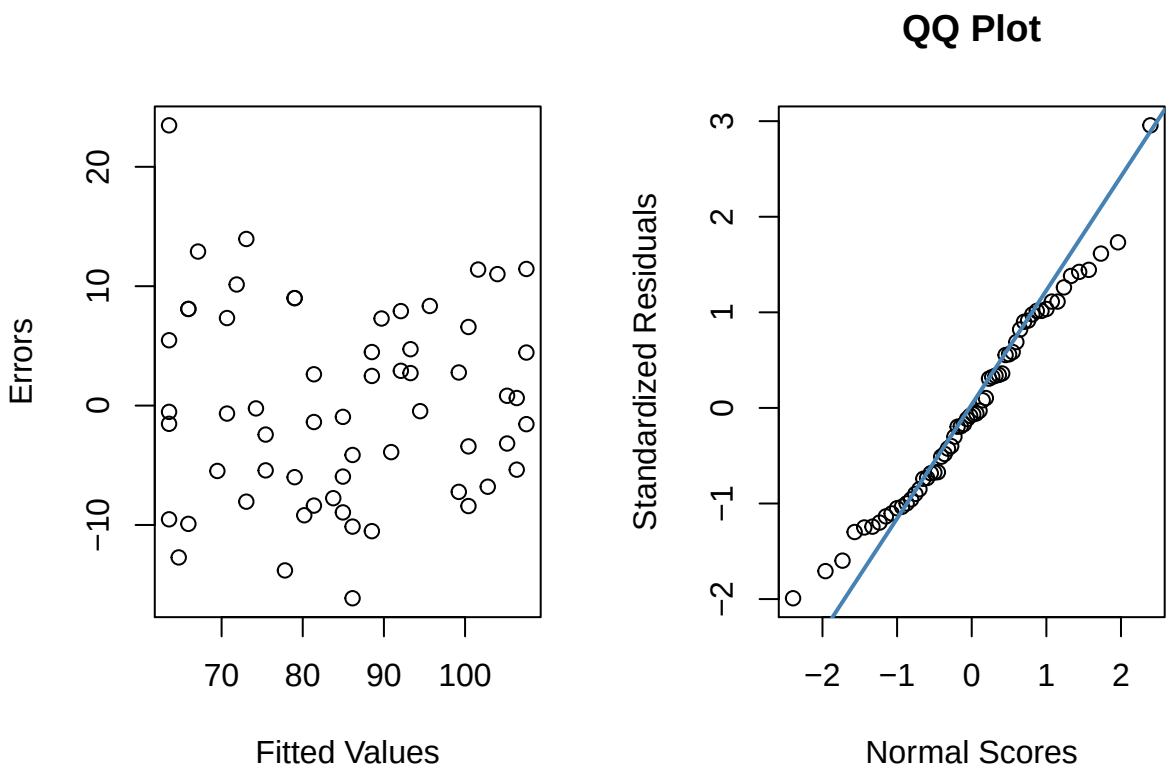
a)Fit a linear regression function. Obtain, the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show? (5 pts)

Solution: R Square is 75% and the model is significant. Error vs. Fitted values graph indicate unequal variances. QQ plot looks approximately normal.

```
question4 <- read.csv("/cloud/project/question4.csv")
f4<-lm(Y~X,data=question4)
summary(f4)
```

```
##
## Call:
## lm(formula = Y ~ X, data = question4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.1368  -6.1968  -0.5969   6.7607  23.4731
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 156.3466     5.5123   28.36   <2e-16 ***
## X            -1.1900     0.0902  -13.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.173 on 58 degrees of freedom
## Multiple R-squared:  0.7501, Adjusted R-squared:  0.7458
## F-statistic: 174.1 on 1 and 58 DF,  p-value: < 2.2e-16
```

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



b)Conduct the Breusch-Pagan test to determine whether or not the error variance varies with the level of X. Use $alpha = .05$. State the alternatives. decision rule, and conclusion. Is your conclusion consistent with your preliminary findings in part (a)? (10 pts)

Solution:

H_0: γ is 0 H_A: γ is NOT 0; R Square is 7% and the model is significant. Additionally, chi-square test (refer to equation 3.11 on the book) is rejected. Variances are not equal.

```
ei2<-(f4$residuals)^2
f4.1<-lm(ei2~question4$X)
summary(f4.1)
```

```
##
## Call:
## lm(formula = ei2 ~ question4$X)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -99.77 -43.63 -20.29  12.80 450.94
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -53.5326    56.0149  -0.956   0.3432
## question4$X   1.9690     0.9166   2.148   0.0359 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 83.05 on 58 degrees of freedom
```

```
## Multiple R-squared:  0.0737, Adjusted R-squared:  0.05773
## F-statistic: 4.615 on 1 and 58 DF,  p-value: 0.03589
```

```
anova(f4.1)
```

```
## Analysis of Variance Table
##
## Response: ei2
##             Df Sum Sq Mean Sq F value  Pr(>F)
## question4$X  1  31833   31833  4.6148 0.03589 *
## Residuals   58 400089    6898
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(f4)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X          1 11627.5 11627.5  174.06 < 2.2e-16 ***
## Residuals 58  3874.4    66.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
chi.test=(31833/2)/((3874.4/60)^2)
chi.test
```

```
## [1] 3.817167
```

```
pchisq(chi.test,1)
```

```
## [1] 0.9492701
```

Problem 5. The simple linear regression model was built on 45 observation

a) Complete the ANOVA table below (5pts)

b) Calculate R square, is the model statistically significant? (5pts)

a) Complete the ANOVA table below (5pts)

```
kable(data.frame(Source = c("Regression","Error","Total"), DF = c("","", ""), SS=c("","",""),MS = c("",
```

Source	DF	SS	MS	F
Regression				970
Error			80	
Total				

```
MSR=970*80
SSE= 43*80
SST=SSE+MSR
cbind(MSR,SSE,SST)
```

```
##        MSR  SSE   SST
```

```
## [1,] 77600 3440 81040
```

```
kable(data.frame(Source = c("Regression","Error","Total"), DF = c("1","43", "44"), SS=c("77600","3440",
```

Source	DF	SS	MS	F
Regression	1	77600	77600	970
Error	43	3440	80	
Total	44	81040		

b) Calculate R square, is the model statistically significant? (5pts)

```
R.SQ =77600/81040
R.SQ
```

```
## [1] 0.9575518
```

```
1-pf(970,1,43)
```

```
## [1] 0
```

Solution: R Square is 0.957%. Ho:B_1=0 Ha:B_1

F = 970, too large and pvalue=0.00, reject null, model is significant.