HW5-Solutions

Problem 1

1- Refer to Plastic hardness data (20pts)

###a) Using matrix methods, obtain the following: (1) (X'X)-1, (2) b, (3) Yhat, (4) H, (5) SSE, (6) s2(b), (7) s2(pred) when Xh= 30. (10 pts) ###b) From part (a6), obtain the following: (1) s2{bo}; (2) s{bo, bl }; (3) s{bl }. (5pts) ###c) Obtain the matrix of the quadratic form for SSE. (5pts)

Solution:

a) Please see below._

```
library(knitr)
Plastic.Hardness <- read.csv("/cloud/project/Plastic Hardness.csv")
f1<-lm(Y~X,data=Plastic.Hardness)
X<-model.matrix(f1)
solve(t(X)%*%X)
```

```
##             (Intercept)           X
## (Intercept)    0.675000 -0.02187500
## X             -0.021875  0.00078125
```

```
Y<-as.matrix(Plastic.Hardness$Y,ncol=1)
Beta<-solve(t(X)%*%X)%*%t(X)%*%Y
Beta
```

```
##                   [,1]
## (Intercept) 168.600000
## X             2.034375
```

```
Yhat<-t(Beta)%*%t(X)
Yhat
```

```
##           1      2      3      4       5       6       7       8     9
## [1,] 201.15 201.15 201.15 201.15 217.425 217.425 217.425 217.425 233.7
##         10    11    12      13      14      15      16
## [1,] 233.7 233.7 233.7 249.975 249.975 249.975 249.975
```

```
H<- X%*%solve(t(X)%*%X)%*%t(X)
H
```

```
##         1      2      3      4     5     6     7     8     9    10    11
## 1   0.175  0.175  0.175  0.175 0.100 0.100 0.100 0.100 0.025 0.025 0.025
## 2   0.175  0.175  0.175  0.175 0.100 0.100 0.100 0.100 0.025 0.025 0.025
## 3   0.175  0.175  0.175  0.175 0.100 0.100 0.100 0.100 0.025 0.025 0.025
## 4   0.175  0.175  0.175  0.175 0.100 0.100 0.100 0.100 0.025 0.025 0.025
## 5   0.100  0.100  0.100  0.100 0.075 0.075 0.075 0.075 0.050 0.050 0.050
## 6   0.100  0.100  0.100  0.100 0.075 0.075 0.075 0.075 0.050 0.050 0.050
## 7   0.100  0.100  0.100  0.100 0.075 0.075 0.075 0.075 0.050 0.050 0.050
## 8   0.100  0.100  0.100  0.100 0.075 0.075 0.075 0.075 0.050 0.050 0.050
## 9   0.025  0.025  0.025  0.025 0.050 0.050 0.050 0.050 0.075 0.075 0.075
## 10  0.025  0.025  0.025  0.025 0.050 0.050 0.050 0.050 0.075 0.075 0.075
## 11  0.025  0.025  0.025  0.025 0.050 0.050 0.050 0.050 0.075 0.075 0.075
## 12  0.025  0.025  0.025  0.025 0.050 0.050 0.050 0.050 0.075 0.075 0.075
## 13 -0.050 -0.050 -0.050 -0.050 0.025 0.025 0.025 0.025 0.100 0.100 0.100
## 14 -0.050 -0.050 -0.050 -0.050 0.025 0.025 0.025 0.025 0.100 0.100 0.100
## 15 -0.050 -0.050 -0.050 -0.050 0.025 0.025 0.025 0.025 0.100 0.100 0.100
## 16 -0.050 -0.050 -0.050 -0.050 0.025 0.025 0.025 0.025 0.100 0.100 0.100
##       12     13     14     15     16
## 1  0.025 -0.050 -0.050 -0.050 -0.050
## 2  0.025 -0.050 -0.050 -0.050 -0.050
## 3  0.025 -0.050 -0.050 -0.050 -0.050
## 4  0.025 -0.050 -0.050 -0.050 -0.050
## 5  0.050  0.025  0.025  0.025  0.025
## 6  0.050  0.025  0.025  0.025  0.025
## 7  0.050  0.025  0.025  0.025  0.025
## 8  0.050  0.025  0.025  0.025  0.025
## 9  0.075  0.100  0.100  0.100  0.100
## 10 0.075  0.100  0.100  0.100  0.100
## 11 0.075  0.100  0.100  0.100  0.100
## 12 0.075  0.100  0.100  0.100  0.100
## 13 0.100  0.175  0.175  0.175  0.175
## 14 0.100  0.175  0.175  0.175  0.175
## 15 0.100  0.175  0.175  0.175  0.175
## 16 0.100  0.175  0.175  0.175  0.175
```

```
n<-dim(Plastic.Hardness)[1]
I<-diag(n)
SSE<-t(Y)%*%(I - H)%*%Y
SSE
```

```
##         [,1]
## [1,] 146.425
```

```
MSE<-SSE/(n-2)
s2b<-c(MSE)*solve(t(X) %*% X)
s2b
```

```
##             (Intercept)           X
## (Intercept)   7.0597768 -0.228789063
## X            -0.2287891  0.008171038
```

```
Xh<-matrix(c(1,30),ncol=1)
s2bpred<-c(MSE)*(t(Xh)%*%solve(t(X)%*%X)%*%Xh)
s2bpred
```

```
##           [,1]
## [1,] 0.6863672
```

b)

Solution:

See below

```
s2bo<-diag(s2b)[1]
s2bo
```

```
## (Intercept)
##    7.059777
```

```
cov.bo.b1<-s2b[1,2]
cov.bo.b1
```

```
## [1] -0.2287891
```

```
sb1<-sqrt(diag(s2b)[2])
sb1
```

```
##          X
## 0.09039379
```

```
SSE
```

```
##         [,1]
## [1,] 146.425
```

Problem 2

Refer to the Brand preference data. In a small-scale experimental study of the relation between degree of brand liking (Y) and moisture content (X1) and sweetness (X2) of the product. (25 pts)

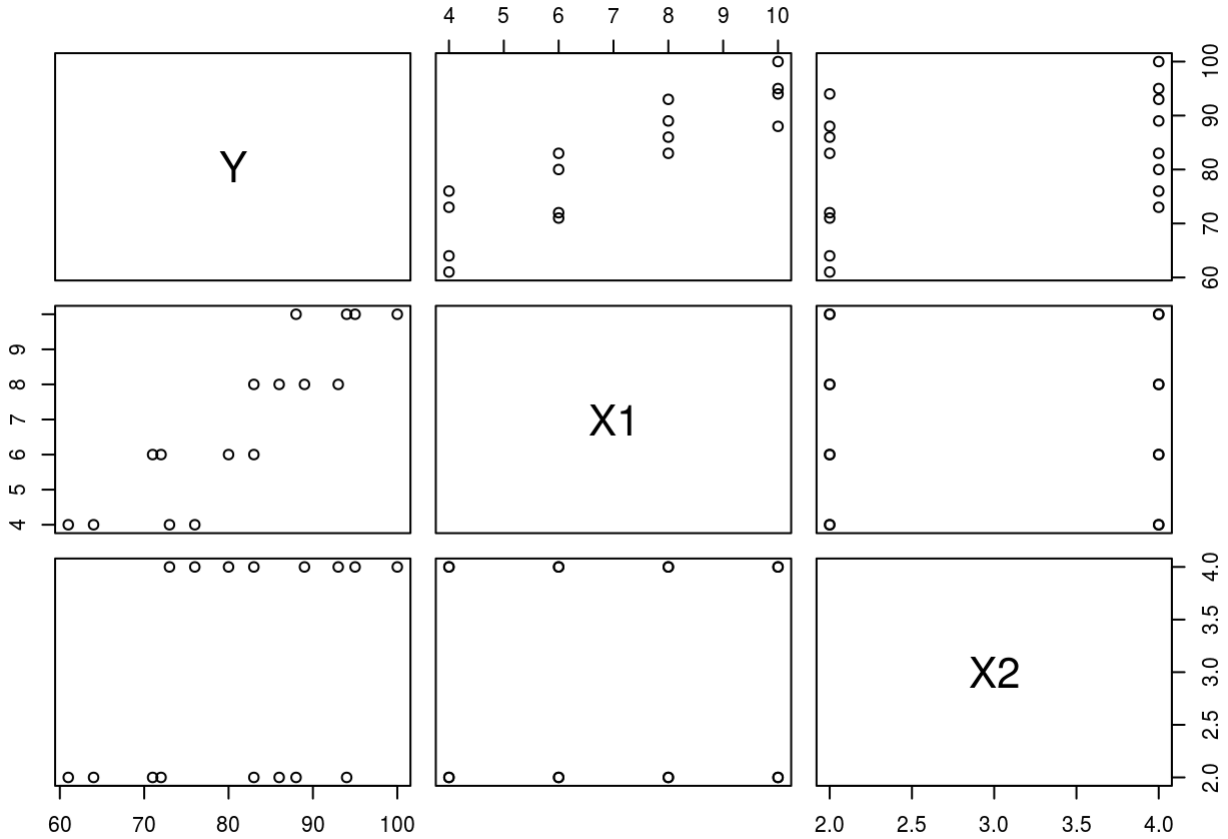
###a) Obtain the scatter plot matrix and the correlation matrix. What information do these diagnostic aids provide here? (5pts) ###b) Fit regression model to data. State the estimated regression function. Interpreted regression coefficients? (5pts) ###c) Obtain the residuals, and prepare box plot of the residuals. What information does this plot provide? (5pts) ###d) Plot the residuals against Yhat, X1, X2, and X1X2 on separate graphs. Also prepare a normal probability plot. Interpret the plots and summarize your findings. (5pts) ### e) Conduct the Breusch-Pagan test for constancy of the error variance. State the alternatives, decision rule, and conclusion. (5pts)

a)

Solution:

There is a strong positive correlation between Y and X1. X1 and X2 seem to be independent from each other. The correlation between Y and X2 is almost 0.4 which indicates that X1 is a stronger predictor.

```
Brand.Preference <- read.csv("/cloud/project/Brand Preference.csv")
plot(Brand.Preference)
```




```
cor(Brand.Preference)
```

```
##            Y        X1        X2
## Y  1.0000000 0.8923929 0.3945807
## X1 0.8923929 1.0000000 0.0000000
## X2 0.3945807 0.0000000 1.0000000
```

b)

Solution: The model is significant and both variables are significant. R square is 95% indicating strong fit.

```
f2<-lm(Y~X1+X2,data=Brand.Preference)
summary(f2)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2, data = Brand.Preference)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.400 -1.762  0.025  1.587  4.200
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.6500     2.9961  12.566 1.20e-08 ***
## X1            4.4250     0.3011  14.695 1.78e-09 ***
## X2            4.3750     0.6733   6.498 2.01e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.693 on 13 degrees of freedom
## Multiple R-squared:  0.9521, Adjusted R-squared:  0.9447
## F-statistic: 129.1 on 2 and 13 DF,  p-value: 2.658e-09
```

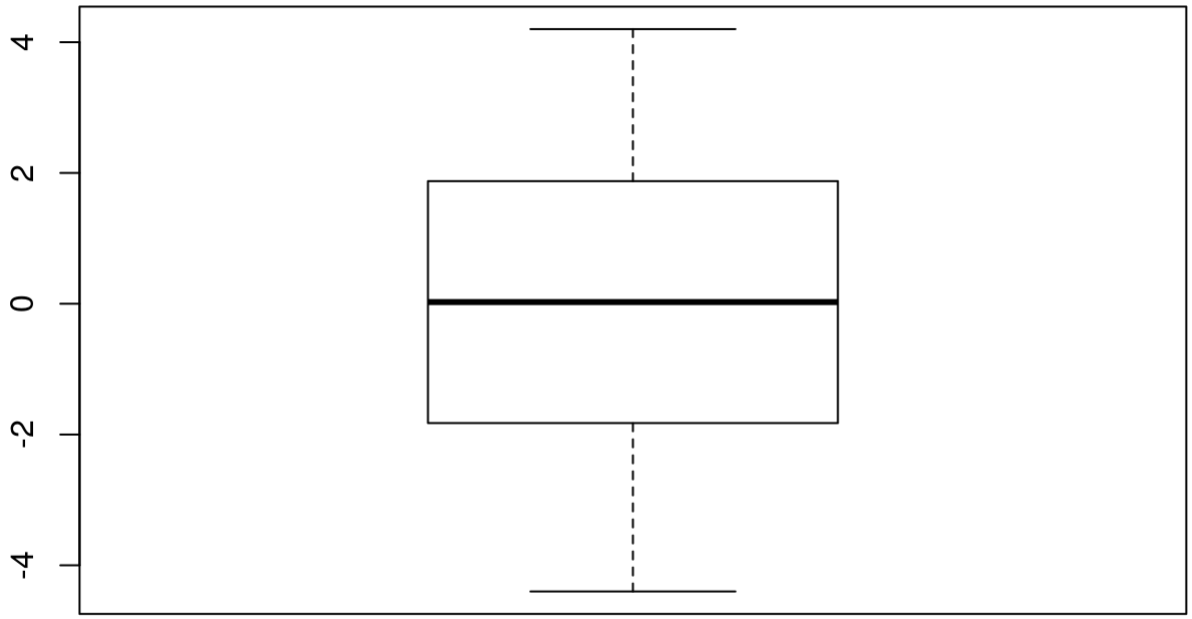
```
anova(f2)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X1         1 1566.45 1566.45 215.947 1.778e-09 ***
## X2         1  306.25  306.25  42.219 2.011e-05 ***
## Residuals 13   94.30    7.25
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c)

Solution: Residuals are normally distributed.

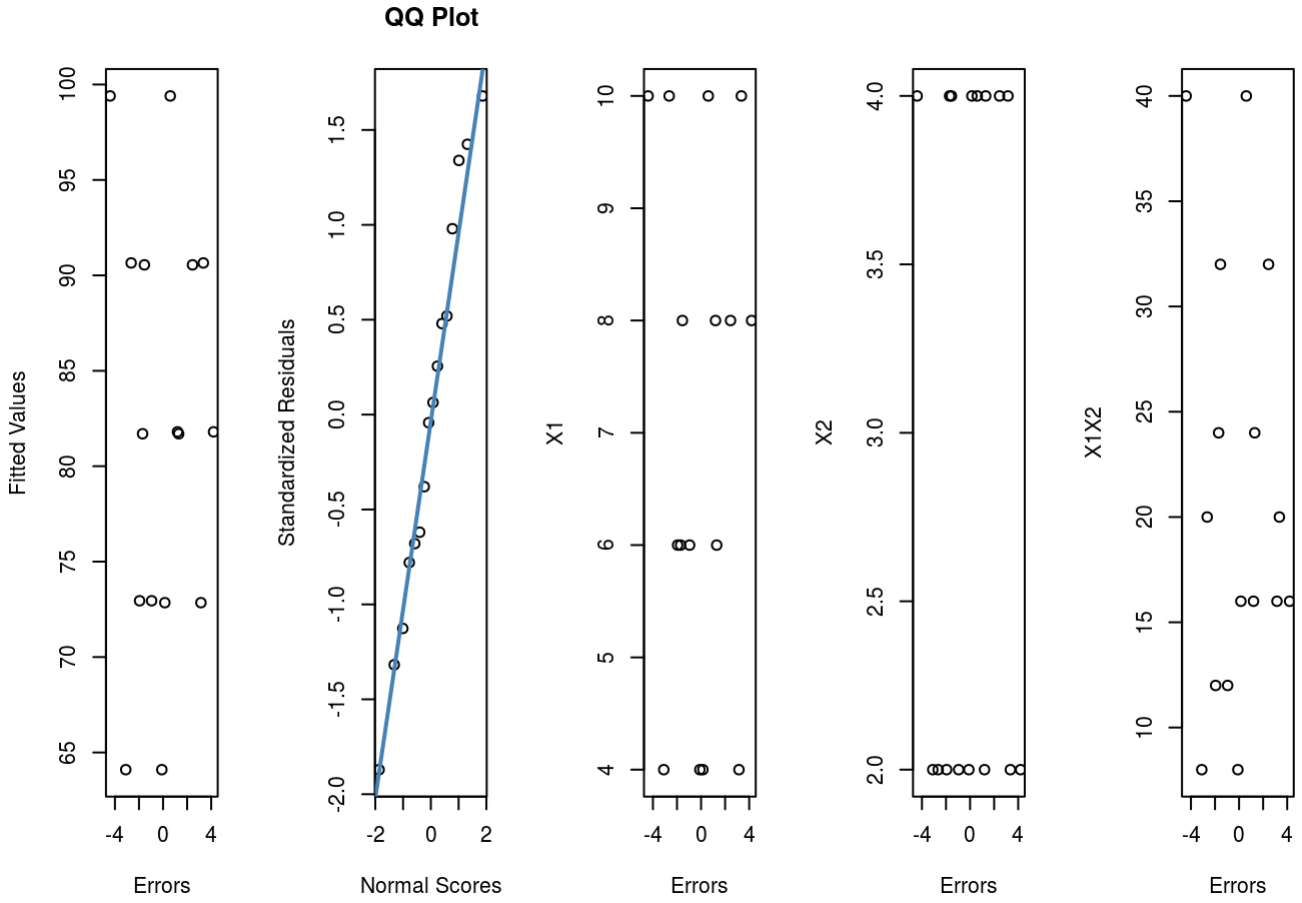
```
ei<-f2$residuals
boxplot(ei)
```



d) ### *Solution:* No intreaction

effect, errors are normally distributed.Error variances are constant.

```
X1<-Brand.Preference$X1
X2<-Brand.Preference$X2
X12<-I(X1*X2)
yhat<-f2$fitted.values
par(mfrow=c(1,5))
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)
plot(ei,X1,xlab="Errors",ylab="X1")
plot(ei,X2,xlab="Errors",ylab="X2")
plot(ei,X12,xlab="Errors",ylab="X1X2")
```



e) ### *Solution:* Conclude error variance constant.

```
ei2<-ei^2
g<-lm(ei2~X1+X2)
summary(g)
```

```
##
## Call:
## lm(formula = ei2 ~ X1 + X2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.724 -3.732 -1.961  2.987 11.276
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.1588     6.8599   0.169    0.868
## X1            0.9175     0.6894   1.331    0.206
## X2           -0.5625     1.5416  -0.365    0.721
##
## Residual standard error: 6.167 on 13 degrees of freedom
## Multiple R-squared:  0.1278, Adjusted R-squared:  -0.006434
## F-statistic: 0.9521 on 2 and 13 DF,  p-value: 0.4113
```

```
anova(g)
```

```
## Analysis of Variance Table
##
## Response: ei2
##           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<-67.34+5.06
SSE<- 94.30
chi.test<-(SSR/2)/((SSE/16)^2)
chi.test
```

```
## [1] 1.042138
```

```
1-pchisq(chi.test,2)
```

```
## [1] 0.5938854
```

Problem 3

###Refer to Problem 2 (Brand preference data) (20 pts) ###a) Test whether there is a regression relation, using α =0.01. State the alternatives, decision rule, and conclusion. What does your test imply about β1 and β2? (5pts) ###b) Estimate β1 and β2 jointly by the Bonferroni procedure, using a 99 percent family confidence coefficient. Interpret your results. (5pts) ###c) Obtain an interval estimate of E{Yh} when Xh1 = 5 and Xh2 = 4. Use a 99 percent confidence coefficient. Interpret your interval estimate. (5pts) ###d) Obtain a prediction interval for a new observation Yh(new) when Xh1 = 5 and Xh2 = 4. Use a 99 percent confidence coefficient. (5pts)

a)

Solution: H0: B1 = B2 = 0, Ha:The model is significant. see the anova table below, reject null, the model is significant. This test implies that either B1 or B2 is significant. The ANOVA table below shows that both B1 and B2 are significant.

```
anova(f2)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X1         1 1566.45 1566.45 215.947 1.778e-09 ***
## X2         1  306.25  306.25  42.219 2.011e-05 ***
## Residuals 13   94.30    7.25
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b)

Solution: please see below

```
confint(f2,level=1-0.01/2)[2:3,]
```

```
##      0.25 %  99.75 %
## X1 3.409483 5.440517
## X2 2.104236 6.645764
```

c)

Solution: Yhat=77.275. 99.9% confidence interval is 73.88<=Yhat<=80.67

```
predict.lm(f2,data.frame(X1=5,X2=4),interval = "confidence", level = 1-0.01)
```

```
##      fit      lwr      upr
## 1 77.275 73.88111 80.66889
```

c)

Solution: Yhat=77.275. 99.9% prediction confidence interval is 68.48<=Yhat<=86.07

```
predict.lm(f2,data.frame(X1=5,X2=4),interval = "prediction", level = 1-0.01)
```

```
##      fit      lwr      upr
## 1 77.275 68.48077 86.06923
```

Problem 4

Refer to Commercial properties data. The age (X1), operating expenses and taxes (X2), vacancy rates (X3), total square footage (X4), and rental rates (Y). (25pts)

a) Obtain the scatter plot matrix and the correlation matrix. Interpret these and state your principal findings. (5pts)

b) Fit regression model for four predictor variables to the data. State the estimated regression function. (5pts)

c) Obtain the residuals and prepare a QQ plot of the residuals. Does the distribution appear to be fairly symmetrical? (5pts)

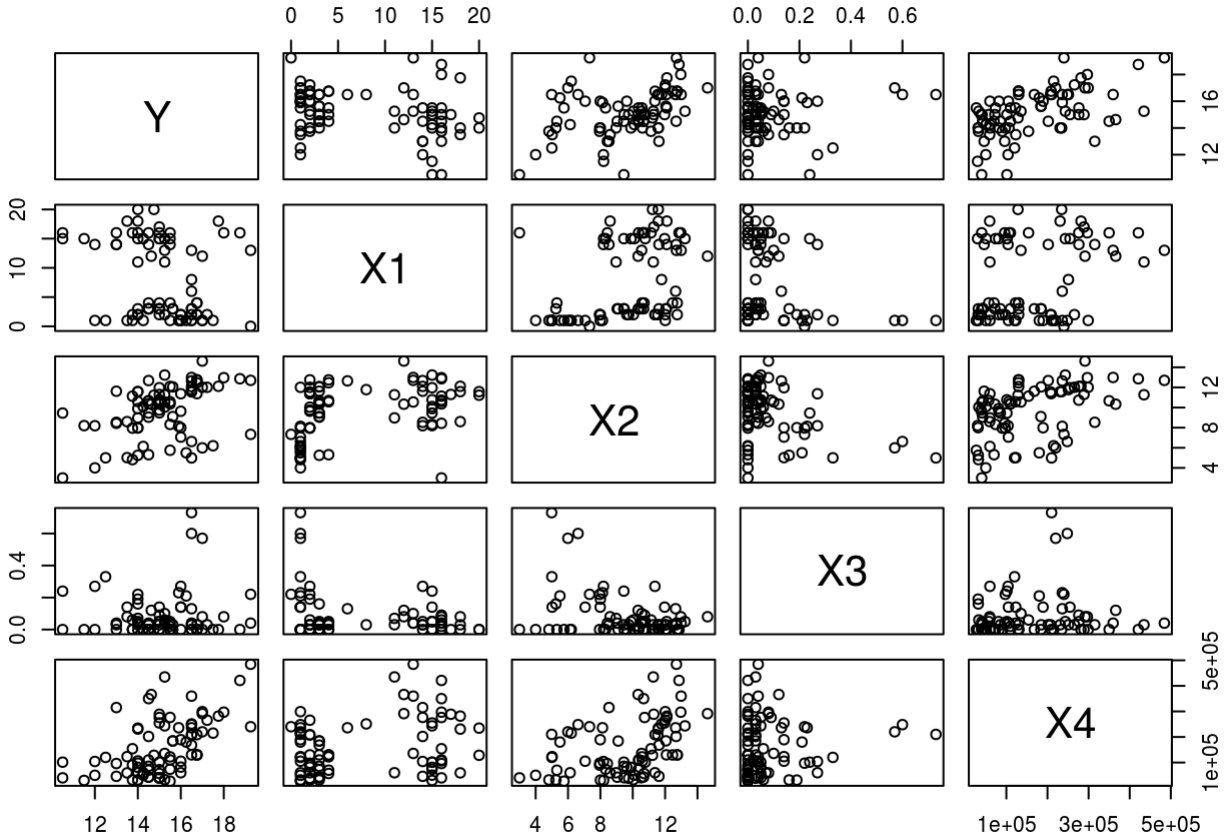
d) Plot the residuals against Y, each predictor variable, and each two-factor interaction terms on separate graphs. Also prepare a normal probability plot. Analyze yours plots and summarize your findings. (5pts)

e) Divide the 81 cases into two groups. placing the 40 cases with the smallest fitted values into group 1 and the remaining cases into group 2. Conduct the Brown-Forsythe test for constancy of the error variance, using α= .05. State the decision rule and conclusion. (5pts)

a)

Solution: X4 has the strongest linear relationship with Y. X2 has the second highest. X1 is the only variable that is negatively correlated with Y. there are possible outliers in X3.

```
Commercial.Properties <- read.csv("/cloud/project/Commercial Properties.csv")
plot(Commercial.Properties)
```




```
round(cor(Commercial.Properties),2)
```

```
##        Y    X1    X2    X3   X4
## Y   1.00 -0.25  0.41  0.07 0.54
## X1 -0.25  1.00  0.39 -0.25 0.29
## X2  0.41  0.39  1.00 -0.38 0.44
## X3  0.07 -0.25 -0.38  1.00 0.08
## X4  0.54  0.29  0.44  0.08 1.00
```

b)

Solution: please see below: All variables except X3 are significant. The RSquare is %58.

```
f4<-lm(Y~X1+X2+X3+X4,data=Commercial.Properties)
summary(f4)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4, data = Commercial.Properties)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1872 -0.5911 -0.0910  0.5579  2.9441
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.220e+01  5.780e-01  21.110  < 2e-16 ***
## X1          -1.420e-01  2.134e-02  -6.655 3.89e-09 ***
## X2           2.820e-01  6.317e-02   4.464 2.75e-05 ***
## X3           6.193e-01  1.087e+00   0.570     0.57
## X4           7.924e-06  1.385e-06   5.722 1.98e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.137 on 76 degrees of freedom
## Multiple R-squared:  0.5847, Adjusted R-squared:  0.5629
## F-statistic: 26.76 on 4 and 76 DF,  p-value: 7.272e-14
```

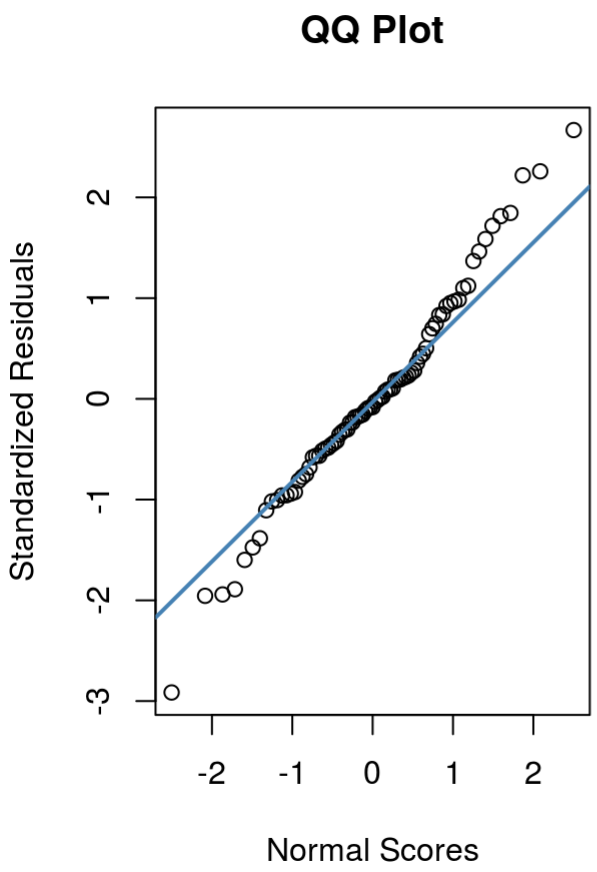
c)

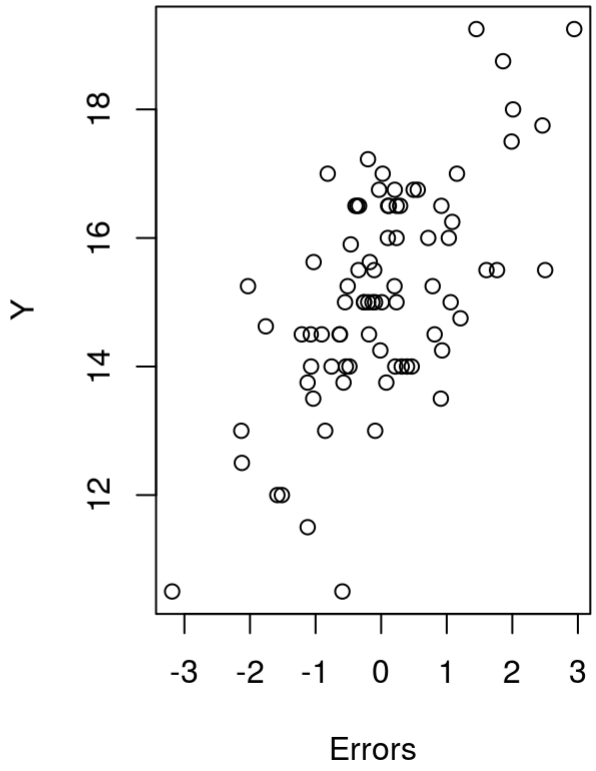
Solution: fairly symmetrical.

```
stdei<- rstandard(f4)
qqnorm(stdei,ylab="Standardized Residuals",xlab="Normal Scores", main="QQ Plot")
qqline(stdei,col = "steelblue", lwd = 2)
```

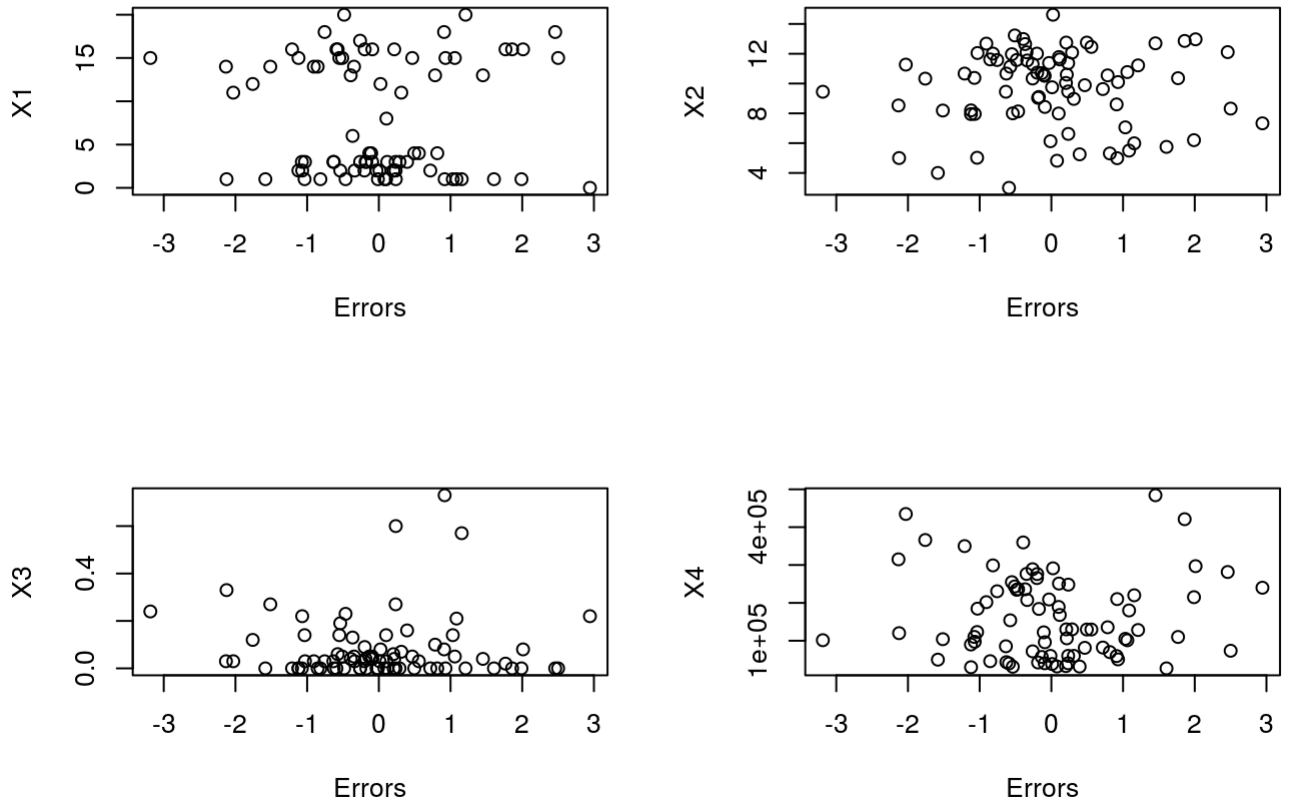
d) ### *Solution:* Graphs dont indicate any problems with the model assumptions. However, the plot with interaction terms indicate that this should be studied further.

```
ei<-f4$residuals
Y<-Commercial.Properties$Y
X1<-Commercial.Properties$X1
X2<-Commercial.Properties$X2
X3<-Commercial.Properties$X3
X4<-Commercial.Properties$X4
X12<-I(X1*X2)
X13<-I(X1*X3)
X14<-I(X1*X4)
X23<-I(X2*X3)
X24<-I(X2*X4)
X34<-I(X3*X4)
par(mfrow=c(1,2))
stdei<- rstandard(f4)
qqnorm(stdei,ylab="Standardized Residuals",xlab="Normal Scores", main="QQ Plot")
qqline(stdei,col = "steelblue", lwd = 2)
plot(ei,Y,xlab="Errors",ylab="Y")
```

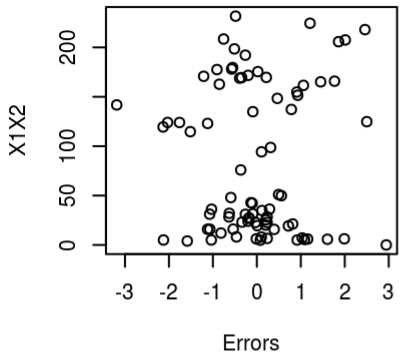


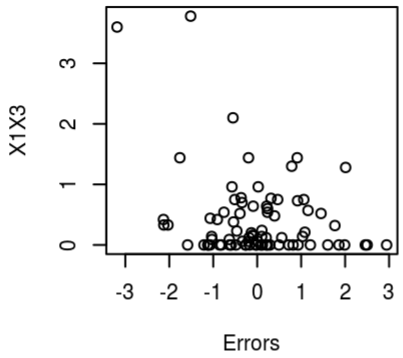


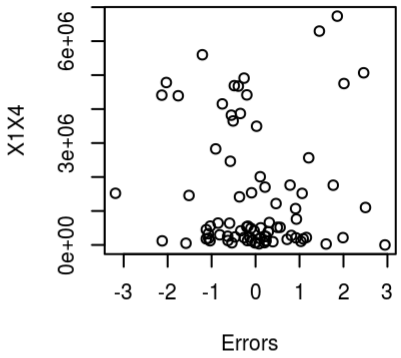

```
par(mfrow=c(2,2))
plot(ei,X1,xlab="Errors",ylab="X1")
plot(ei,X2,xlab="Errors",ylab="X2")
plot(ei,X3,xlab="Errors",ylab="X3")
plot(ei,X4,xlab="Errors",ylab="X4")
```



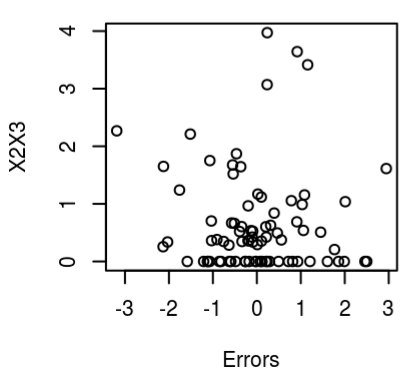

```
par(mfrow=c(2,3))
plot(ei,X12,xlab="Errors",ylab="X1X2")
plot(ei,X13,xlab="Errors",ylab="X1X3")
plot(ei,X14,xlab="Errors",ylab="X1X4")
plot(ei,X23,xlab="Errors",ylab="X2X3")
plot(ei,X24,xlab="Errors",ylab="X2X4")
plot(ei,X34,xlab="Errors",ylab="X3X4")
```

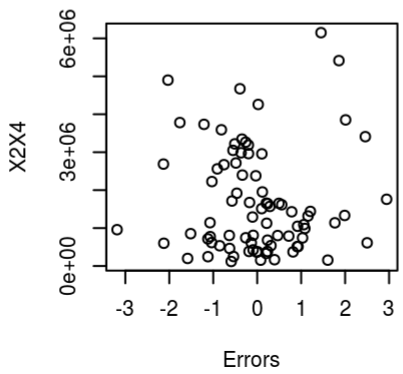


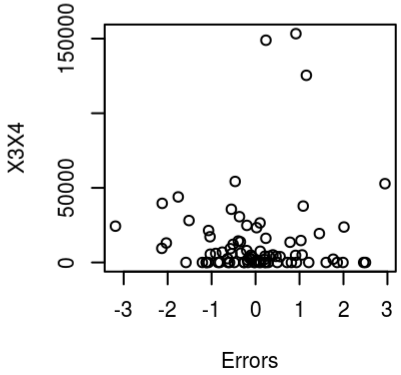




e) ### _Solution:Ho:Constant







Error Variance ### Ha: Non Constant Error Variance ### Accept Null, pvalue=0.37,error variance constant.

```
ei<-f4$residuals
DM<-data.frame(cbind(Commercial.Properties,ei))
DMS<-DM[order(DM$Y),]
DM1<-DMS[1:40,]
DM2<-DMS[41:81,]
M1<-median(DM1[,6])
M2<-median(DM2[,6])
N1<-length(DM1[,6])
N2<-length(DM2[,6])
d1<-abs(DM1[,6]-M1)
d2<-abs(DM2[,6]-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] -0.4729819
```

```
1-2*pt(T,df=N1+N2-2)
```

```
## [1] 0.3624696
```

Problem 5

Refer to Problem 4 (Commercial properties data) (10 pts).

a)Based on the data above in the table. Develop separate prediction intervals for the rental rates of these properties, using a 95 percent statement confidence coefficient in each case. Can the rental rates of these three properties be predicted fairly precisely? What is the family confidence level for the set of three predictions? (10pts)

a)

Solution: The prediction confidence intervals are below. it is 85% coverage

```
X5<-matrix(c(4,10,0.1,80000,6,11.5,0,120000,12,12.5,0.32,340000),byrow=T,ncol=4,nrow=3)
colnames(X5)<-c("X1","X2","X3","X4")
X5
```

```
##      X1   X2   X3     X4
## [1,]  4 10.0 0.10  80000
## [2,]  6 11.5 0.00 120000
## [3,] 12 12.5 0.32 340000
```

```
predict.lm(f4,data.frame(X5),interval = "prediction")
```

```
##        fit      lwr      upr
## 1 15.14850 12.85249 17.44450
## 2 15.54249 13.24504 17.83994
## 3 16.91384 14.53469 19.29299
```

```
0.95^3
```

```
## [1] 0.857375
```