GR5205 HW5 - Model Selection and Outlier Analysis

TONGHONG CHEN || tc2894

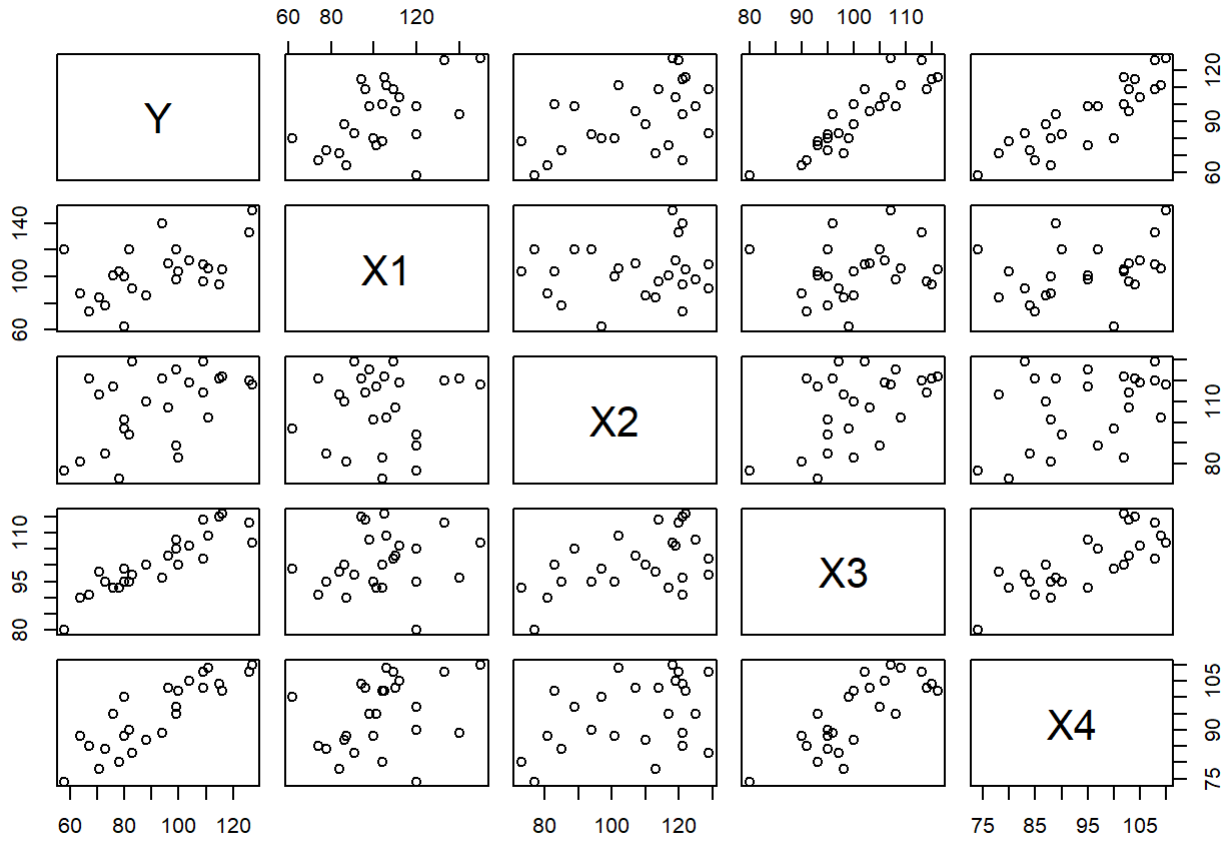
December 7, 2017

Problem 1

(a)(5p) Obtain the scatter plot matrix, and the correlation matrix of the X variables. What do the scatter plots suggest about the nature of the functional relationship between the response variable Y and each of the predictor variables? Are any serious multi-collinearity problems evident? Explain.

```
full=read.table("Homework_data_5_Problem1and2.txt",head=FALSE, col.names = c('Y','X1','X2',
'X3','X4'))

plot(full)
```



```
cor(full)
```

```
##            Y        X1        X2        X3        X4
## Y  1.0000000 0.5144107 0.4970057 0.8970645 0.8693865
## X1 0.5144107 1.0000000 0.1022689 0.1807692 0.3266632
## X2 0.4970057 0.1022689 1.0000000 0.5190448 0.3967101
## X3 0.8970645 0.1807692 0.5190448 1.0000000 0.7820385
## X4 0.8693865 0.3266632 0.3967101 0.7820385 1.0000000
```

From the scatter plot, X3 and X4 tends to have significant linear relationship with Y, while X2 has slight linear relationship with Y and X1 has insignificant linear relationship with Y.

However, X3 tends to have significant linear relationship with X4. Thus there is a serious multicolinearlity problem.

The correlation matrix has further strengthened my conclusion, since the correlation between Y and X3, Y and X4 are larger than 0.85. And meanwhile, the correlation between X3 and X4 is larger than 0.78

(b)(5p) Fit the multiple regression function containing all four predictor variables as first-order terms. Does it appear that all predictor variables should be retained?

```
Y=full$Y
X1=full$X1
X2=full$X2
X3=full$X3
X4=full$X4

regr = lm(Y~X1+X2+X3+X4)
regr
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4)
##
## Coefficients:
## (Intercept)           X1           X2           X3           X4
##  -124.38182     0.29573     0.04829     1.30601     0.51982
```

Y hat= -124.38182 + 0.29573X1 + 0.04829X2+1.30601X3+0.51982X4

While the four predictor variables are in the similar scale, the coefficient of X2 is only 0.04829 (very close to 0). Therefore, it might not be reasonable to retain all the variables.

(c)(10p) Using only first-order terms for the predictor variables in the pool of potential X variables, find the four best subset regression models according to the adjusted R2 criterion.

```
summary(lm(Y~X1))$adj.r.squared
```

```
## [1] 0.2326452
```

```
summary(lm(Y~X1+X2))$adj.r.squared
```

```
## [1] 0.4154853
```

```
summary(lm(Y~X1+X3))$adj.r.squared
```

```
## [1] 0.9269043
```

```
summary(lm(Y~X1+X4))$adj.r.squared
```

```
## [1] 0.7984716
```

```
summary(lm(Y~X1+X2+X3))$adj.r.squared
```

```
## [1] 0.9246779
```

```
summary(lm(Y~X1+X2+X4))$adj.r.squared
```

```
## [1] 0.8232664
```

```
summary(lm(Y~X1+X3+X4))$adj.r.squared
```

```
## [1] 0.9560482
```

```
summary(lm(Y~X1+X2+X3+X4))$adj.r.squared
```

```
## [1] 0.9554702
```

```
summary(lm(Y~X2+X3))$adj.r.squared
```

```
## [1] 0.7884436
```

```
summary(lm(Y~X2+X4))$adj.r.squared
```

```
## [1] 0.7635916
```

```
summary(lm(Y~X2+X3+X4))$adj.r.squared
```

```
## [1] 0.8616797
```

```
summary(lm(Y~X3+X4))$adj.r.squared
```

```
## [1] 0.8660988
```

Based on the adjusted R^2, the four best subset regression models are:

```
lm(Y~X1+X3+X4)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X3 + X4)
##
## Coefficients:
## (Intercept)           X1           X3           X4
##   -124.2000       0.2963       1.3570       0.5174
```

```
lm(Y~X1+X2+X3+X4)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4)
##
## Coefficients:
## (Intercept)           X1           X2           X3           X4
##  -124.38182      0.29573      0.04829      1.30601      0.51982
```

```
lm(Y~X1+X3)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X3)
##
## Coefficients:
## (Intercept)           X1           X3
##   -127.5957       0.3485       1.8232
```

```
lm(Y~X1+X2+X3)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3)
##
## Coefficients:
## (Intercept)           X1           X2           X3
##  -127.77378      0.34813      0.04353      1.77921
```

d)(10p) Using forward stepwise regression find the best subset of predictor variables to predict job proficiency. Use α limits of 0.05 and 0.10 for adding or deleting a variable, respectively (see Lecture 11, slides 11-12).

```
library(MASS)
Null = lm(Y ~ 1)
addterm(Null, scope = regr, test="F")
```

```
## Single term additions
##
## Model:
## Y ~ 1
##        Df Sum of Sq    RSS    AIC F Value     Pr(F)
## <none>              9054.0 149.30
## X1      1    2395.9 6658.1 143.62   8.276  0.008517 **
## X2      1    2236.5 6817.5 144.21   7.545  0.011487 *
## X3      1    7286.0 1768.0 110.47  94.782 1.264e-09 ***
## X4      1    6843.3 2210.7 116.06  71.198 1.699e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
NewMod = update( Null, .~. + X3)
addterm( NewMod, scope = regr, test="F" )
```

```
## Single term additions
##
## Model:
## Y ~ X3
##        Df Sum of Sq     RSS     AIC F Value     Pr(F)
## <none>              1768.02 110.469
## X1      1   1161.37  606.66  85.727  42.116 1.578e-06 ***
## X2      1     12.21 1755.81 112.295   0.153   0.69946
## X4      1    656.71 1111.31 100.861  13.001   0.00157 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
NewMod = update( NewMod, .~. + X1)
addterm( NewMod, scope = regr, test="F" )
```

```
## Single term additions
##
## Model:
## Y ~ X3 + X1
##        Df Sum of Sq    RSS    AIC F Value     Pr(F)
## <none>              606.66 85.727
## X2      1     9.937 596.72 87.314  0.3497 0.5605965
## X4      1   258.460 348.20 73.847 15.5879 0.0007354 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
NewMod = update( NewMod, .~. + X4)
addterm( NewMod, scope = regr, test="F" )
```

```
## Single term additions
##
## Model:
## Y ~ X3 + X1 + X4
##        Df Sum of Sq    RSS    AIC F Value  Pr(F)
## <none>              348.20 73.847
## X2      1     12.22 335.98 74.954  0.7274 0.4038
```

After forward stepwise regression, we choose the following model as the best subset model:

```
lm(Y~X1+X3+X4)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X3 + X4)
##
## Coefficients:
## (Intercept)           X1           X3           X4
##   -124.2000       0.2963       1.3570       0.5174
```

(e)(5p) How does the best subset according to forward stepwise regression compare with the best subset according to the adjusted R2 criterion from (c) above?

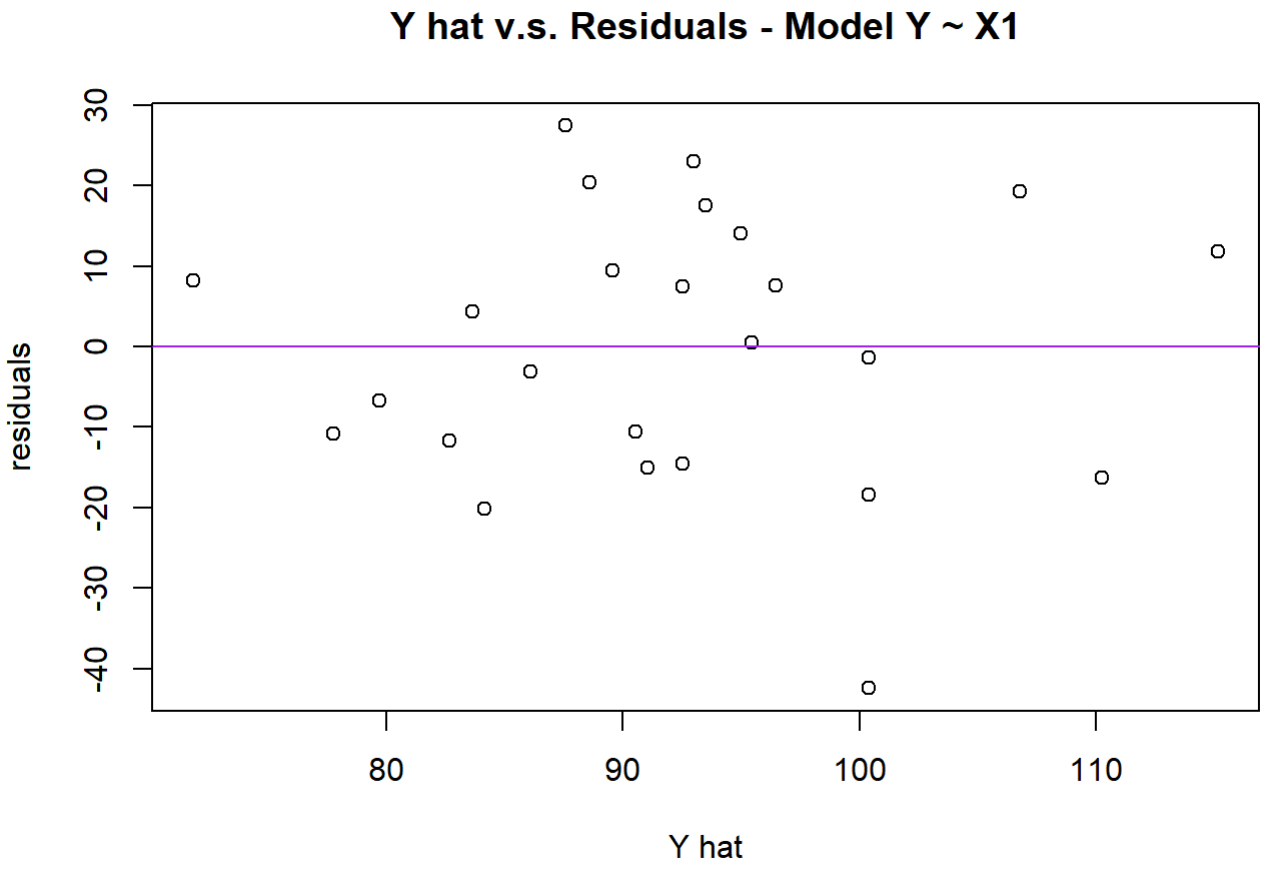
The model chosen by the forward stepwise regression is the same of the one chosen by the criteria of adjusted R^2.

Problem 2

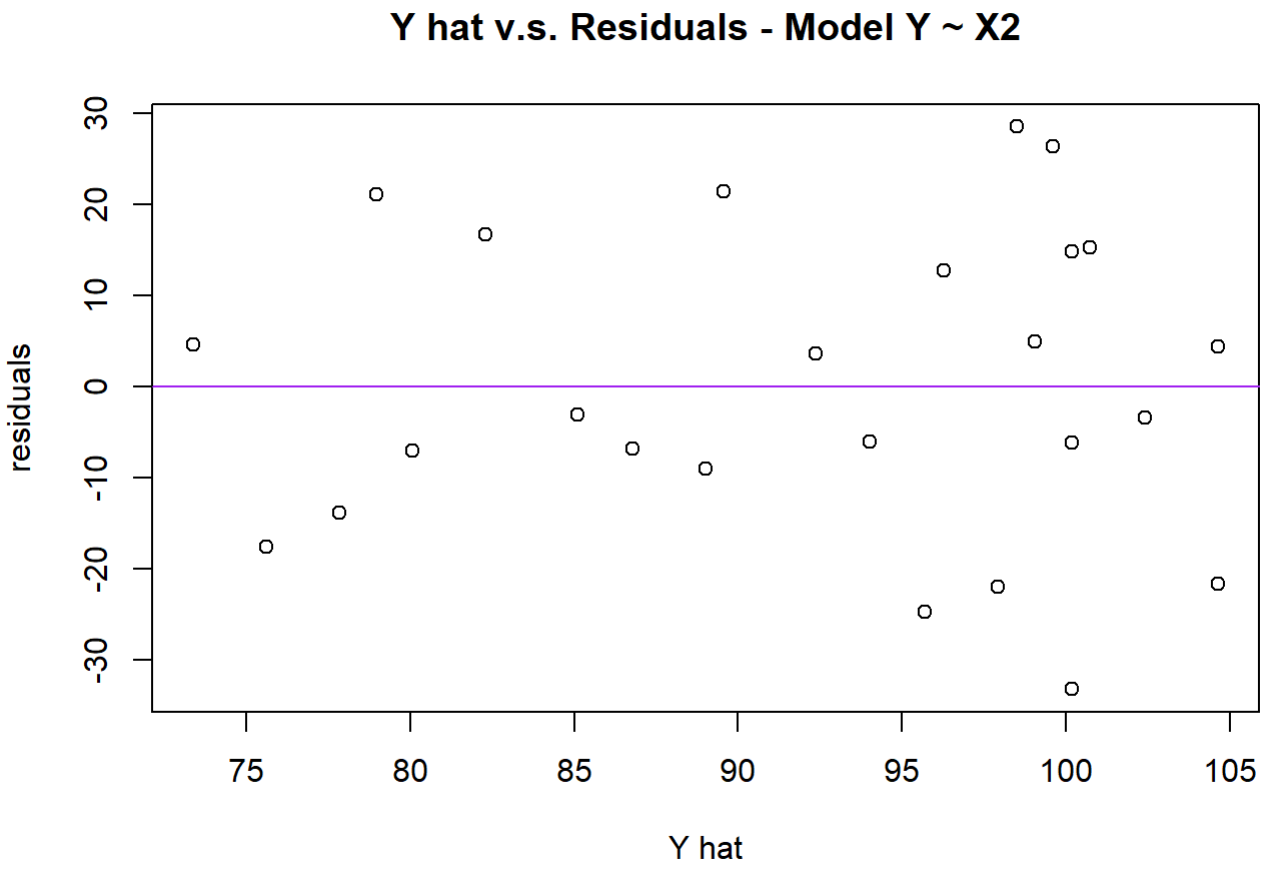
(a)(5p) Obtain the residuals and plot them separately against Y, each of the four predictor variables, and the cross-product term X1X3. On the basis of these plots, should any modifications in the

regression model be investigated?

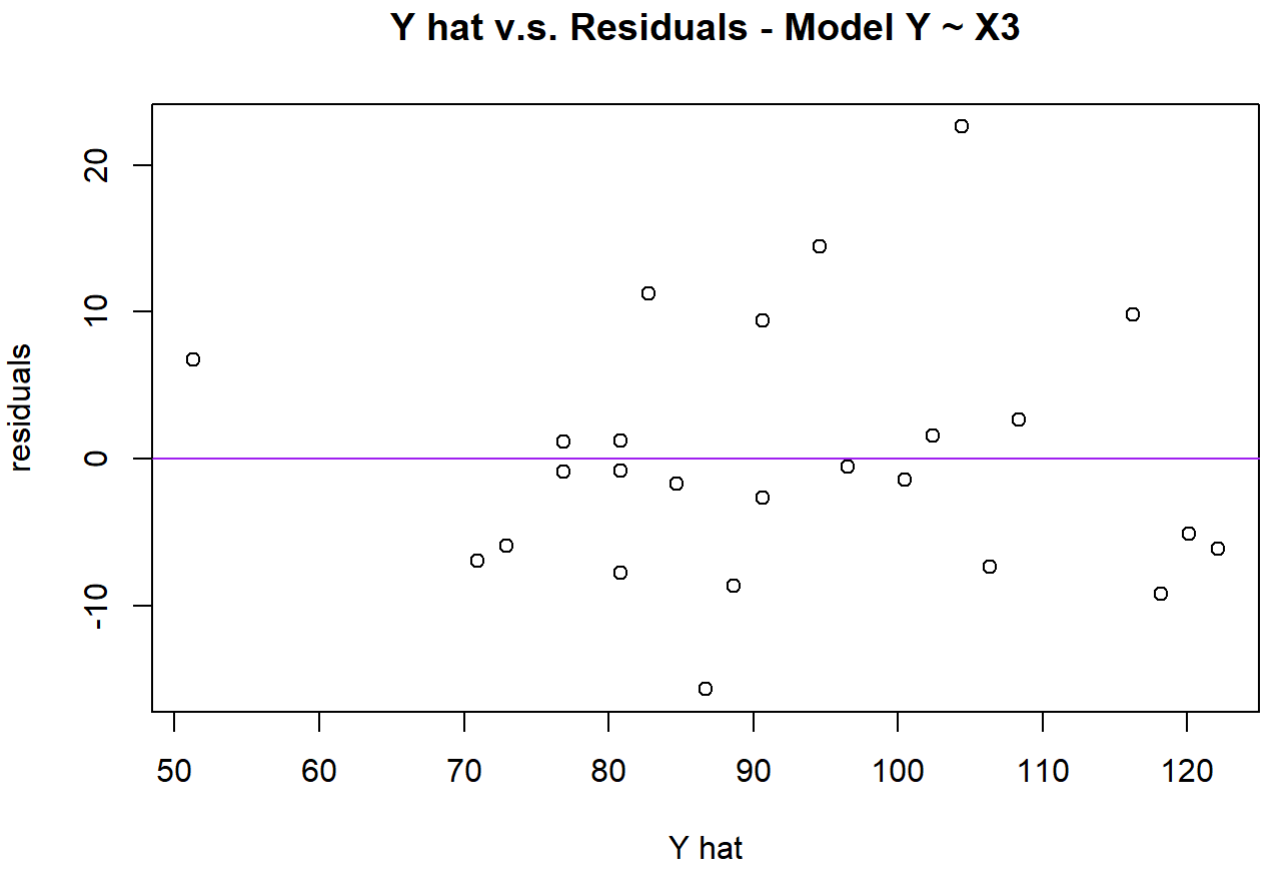
```
regr1 = lm(Y~X1)
plot(regr1$fitted.values, regr1$residuals, main="Y hat v.s. Residuals - Model Y ~ X1", xlab="Y h
at", ylab="residuals")
abline(0,0, col = "purple")
```



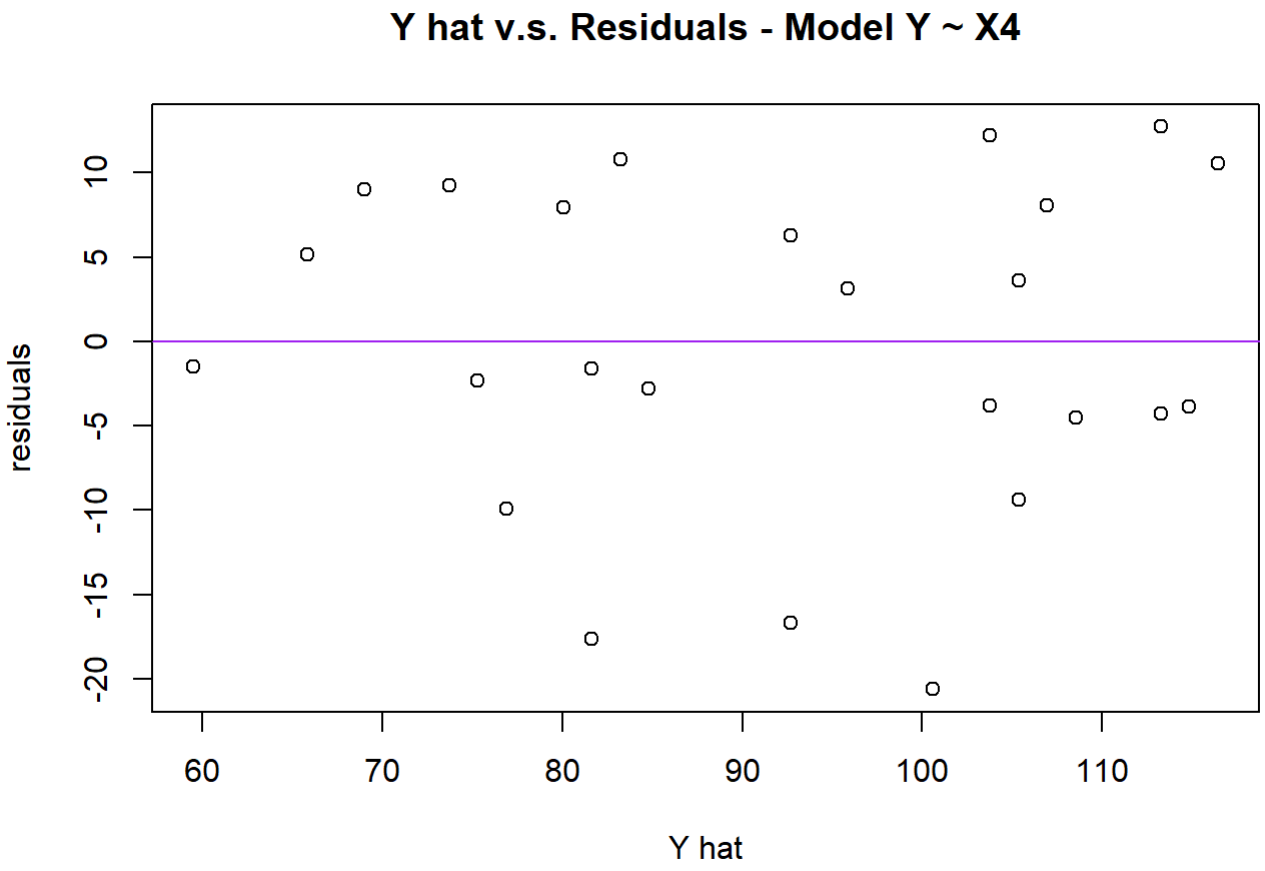

```
regr2 = lm(Y~X2)
plot(regr2$fitted.values, regr2$residuals, main="Y hat v.s. Residuals - Model Y ~ X2", xlab="Y h
at", ylab="residuals")
abline(0,0, col = "purple")
```



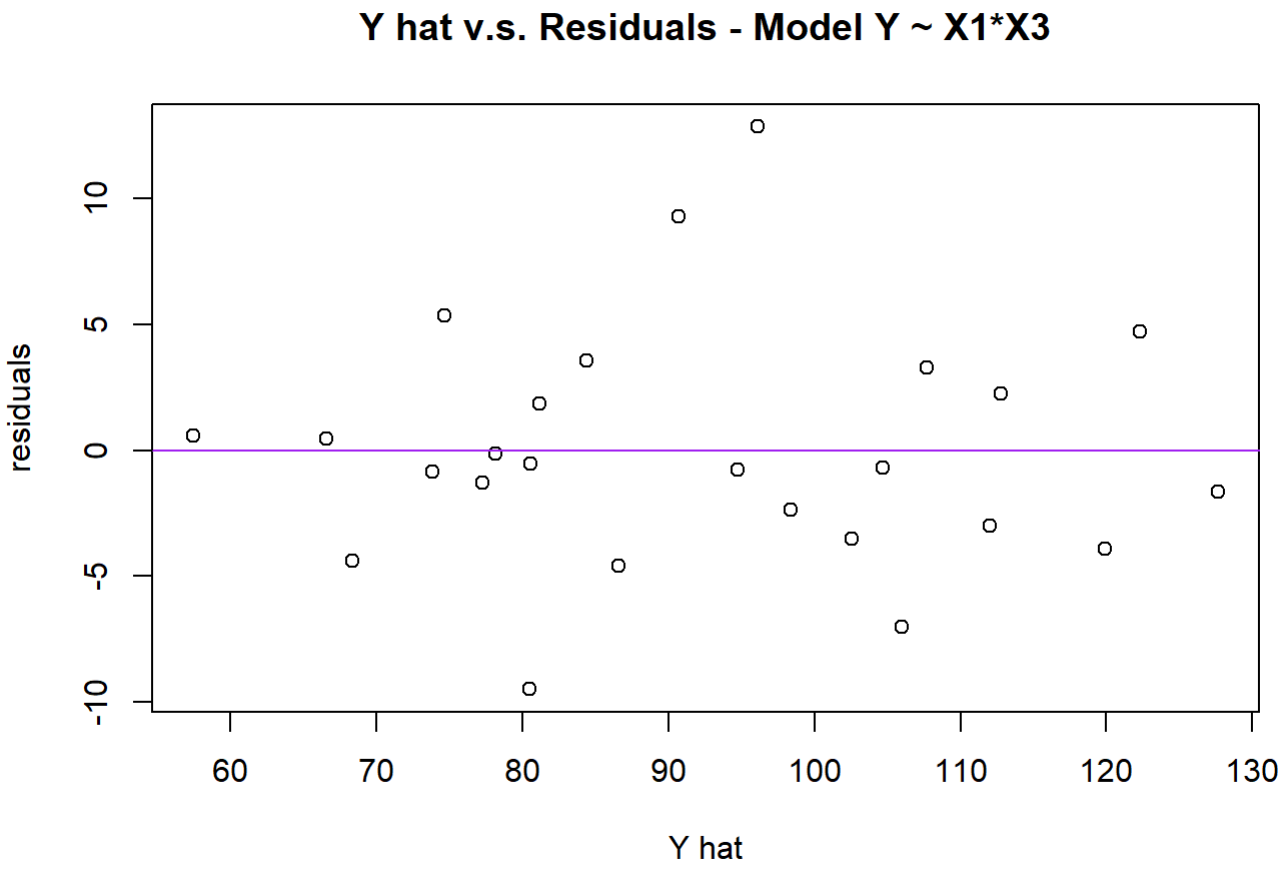

```
regr3 = lm(Y~X3)
plot(regr3$fitted.values, regr3$residuals, main="Y hat v.s. Residuals - Model Y ~ X3", xlab="Y h
at", ylab="residuals")
abline(0,0, col = "purple")
```




```
regr4 = lm(Y~X4)
plot(regr4$fitted.values, regr4$residuals, main="Y hat v.s. Residuals - Model Y ~ X4", xlab="Y h
at", ylab="residuals")
abline(0,0, col = "purple")
```




```
regr13 = lm(Y~X1*X3)
plot(regr13$fitted.values, regr13$residuals, main="Y hat v.s. Residuals - Model Y ~ X1*X3",
xlab="Y hat", ylab="residuals")
abline(0,0, col = "purple")
```



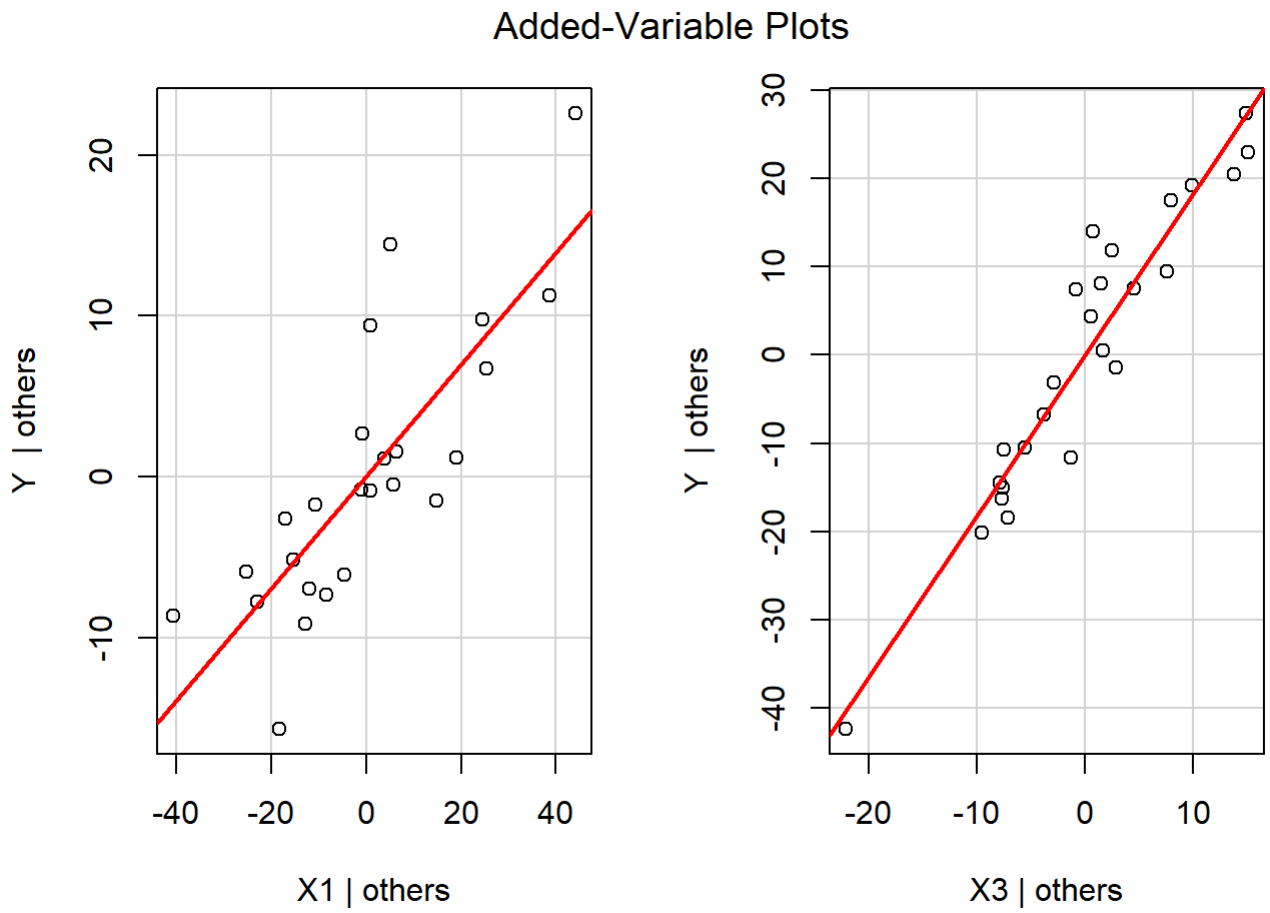
Since none of the residual plots above show significant autocorrelation between the residuals, no modification is needed.

(b)(10p) Prepare separate added-variable plots against e (X1 | X3) and e (X3 | X1). Do these plots suggest that any modifications in the model form are warranted?

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.4.3
```

```
avPlots(lm(Y~X1+X3))
```



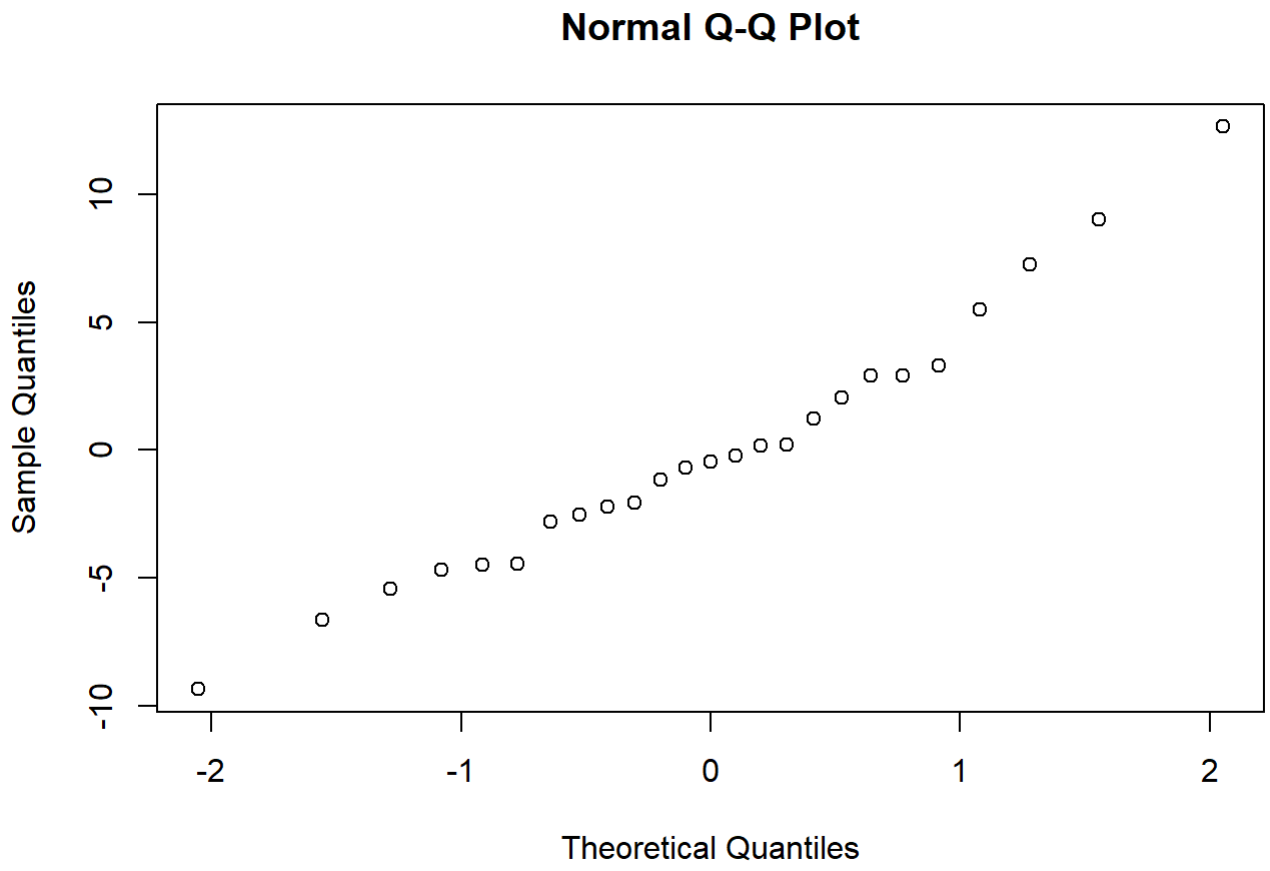
In the above graphs, the “others” in the first graph means “X3”, while the “others” in the first graph means “X1”.

Since the slope of the red line is equal to the estimate of the coefficient of the added variable in the full model, the plots suggest that we should include both X1 and X3 in the regression model, which probably helps improve the model accuracy.

(c)(10p) Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality. Test the reasonableness of the normality assumptions, using α = 0.1. What do you conclude?

The question does not specify the model we are going to plot the normal porbability plot, so I suppose it’s Y~X1+X3. The approach for other models should be the same as below:

```
regr1_3 = lm(Y~X1+X3)
qqp = qqnorm(regr1_3$residuals)
```



The coefficient of correlation between the ordered residuals in the plot and their expected values under normality is:

```
cor(qqp$x, qqp$y)
```

```
## [1] 0.9840739
```

From appendix table B.6 of the textbook, given level of significance of 10%, the critical value of t test with d.f.=26 is 0.967. Since 0.98 > 0.967, we conclude that the assumption of normality appears reasonable.

(d)(10p) Obtain the studentized deleted residuals and identify any outlying Y observations. Use the Bonferroni outlier test (a t-test with t (1 − α/(2N), N − P − 1) critical value) with α = 0.05. State the decision rule and conclusion.

```
studentized_resid = rstudent(regr1_3)
studentized_resid
```

```
##           1           2           3           4           5           6
##  0.64467817  1.19142379 -0.48361310 -0.22370364 -0.08715746  0.04177790
##           7           8           9          10          11          12
## -0.47669249 -0.93171502 -0.13263664 -1.33970402 -0.56015419  0.25316970
##          13          14          15          16          17          18
## -0.46576314 -1.97310552  0.57139728  2.82689084  1.85084184  1.67171059
##          19          20          21          22          23          24
## -0.87827393 -1.09086483  0.59249733 -0.96970269 -0.03905669  0.03516628
##          25
##  0.39192914
```

```
n=25
p=3
ifelse(abs(studentized_resid) > qt(p=1-0.05/(2*n),df=n-p-1), "outlier", "Non-outlier")
```

```
##             1             2             3             4             5
## "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier"
##             6             7             8             9            10
## "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier"
##            11            12            13            14            15
## "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier"
##            16            17            18            19            20
## "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier"
##            21            22            23            24            25
## "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier"
```

From the above result, observation 16 of Y has the highest studnetized residual. Given 5% level of significance, for model Y~X1+X3, we do not reject the hypothesis that any point is an outlier by using Bonferroni outlier test.

(e)(10p) Obtain the diagonal elements of the hat matrix. Using the rule of thumb (Xi is outlying case with regard to the X values, if hii > 2P/N), identify any outlying X observations.

```
hat_value = hatvalues(regr1_3)
hat_value
```

```
##          1          2          3          4          5          6
## 0.07065696 0.21418067 0.04600259 0.07240679 0.05797031 0.11297480
##          7          8          9         10         11         12
## 0.33658569 0.16503154 0.05876518 0.07258372 0.11674502 0.18227927
##         13         14         15         16         17         18
## 0.20825309 0.07879682 0.07578796 0.04353035 0.04043874 0.26349128
##         19         20         21         22         23         24
## 0.07482306 0.09578960 0.15795012 0.14979081 0.07389471 0.17143610
##         25
## 0.05983483
```

```
ifelse(hat_value > 2*p/n, "outlier", "Non-outlier")
```

```
##             1             2             3             4             5
## "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier"
##             6             7             8             9            10
## "Non-outlier"     "outlier" "Non-outlier" "Non-outlier" "Non-outlier"
##            11            12            13            14            15
## "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier"
##            16            17            18            19            20
## "Non-outlier" "Non-outlier"     "outlier" "Non-outlier" "Non-outlier"
##            21            22            23            24            25
## "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier"
```

From rule of thumb, observation 7 and 18 of X seem to be outliers.

(f)(10p) Case 7 and 18 appear to be moderately outlying with respect to their X values, and case 16 is reasonably far outlying with respect to its Y value. Obtain DF FITS, DFBETAS, and Cook's distance values for these cases to asses their influence. What do you conclude?

```
dfbetas_table = dfbetas(regr1_3)
influence_table = cbind(
  "DFFITS"  = dffits(regr1_3),
  "DFBETA_Intercept" = dfbetas_table[,1],
  "DFBETA_X1" = dfbetas_table[,2],
  "DFBETA_X3" = dfbetas_table[,3],
  "Cook's Dist." = cooks.distance(regr1_3))

influence_table[c(7,18,16),]
```

```
##        DFFITS DFBETA_Intercept  DFBETA_X1  DFBETA_X3 Cook's Dist.
## 7  -0.3395422      -0.24018835 -0.1511532 0.30330857   0.03982863
## 18  0.9998970      -0.46436969  0.8777907 0.11512640   0.30812948
## 16  0.6030725      -0.06867086  0.1519977 0.05116114   0.09199677
```

From the above table, none of the absolute value is larger than 1. Though observation 18 has DFFITS close to 1

Based on this result, we conclude that there is no outlier for model Y~X1+X3

(g)(10p) Obtain the variance inflation factors. What do they indicate?

```
library(car)
vif(regr1_3)
```

```
##       X1       X3
## 1.033781 1.033781
```

The VIF for X1 and X3 is relative small (<10), therefore we counclude that there is no multicollinearity.