> # 5126 HW3

> # 1.

> # Read table1.19

> t1.19 <- read.table("http://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets/KutnerData/Chapter%20%201%20Data%20Sets/CH01PR19.txt", header = FALSE)

> # Column names can be changed with

> names(t1.19)[1]<-paste("y")

> names(t1.19)[2]<-paste("x")

> t1.19

> # To use the column names without reference to t1.19 you need to attach the dataset:

> attach(t1.19)

>

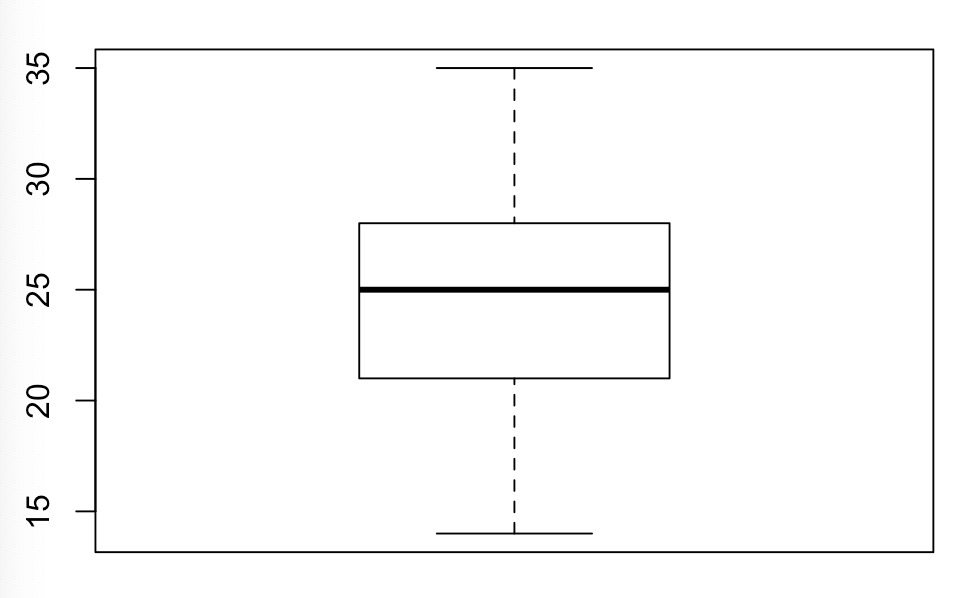
> # 1.a

> # Fitting the SLR model

> reg = lm(t1.19$y ~ t1.19$x)

> # Prepare a boxplot of the ACT scores (X variable):

> boxplot(t1.19$x)



> # All of the ACT scores are between the minimum and the maximum, so there are no outliers and not noteworthy features on the plot.

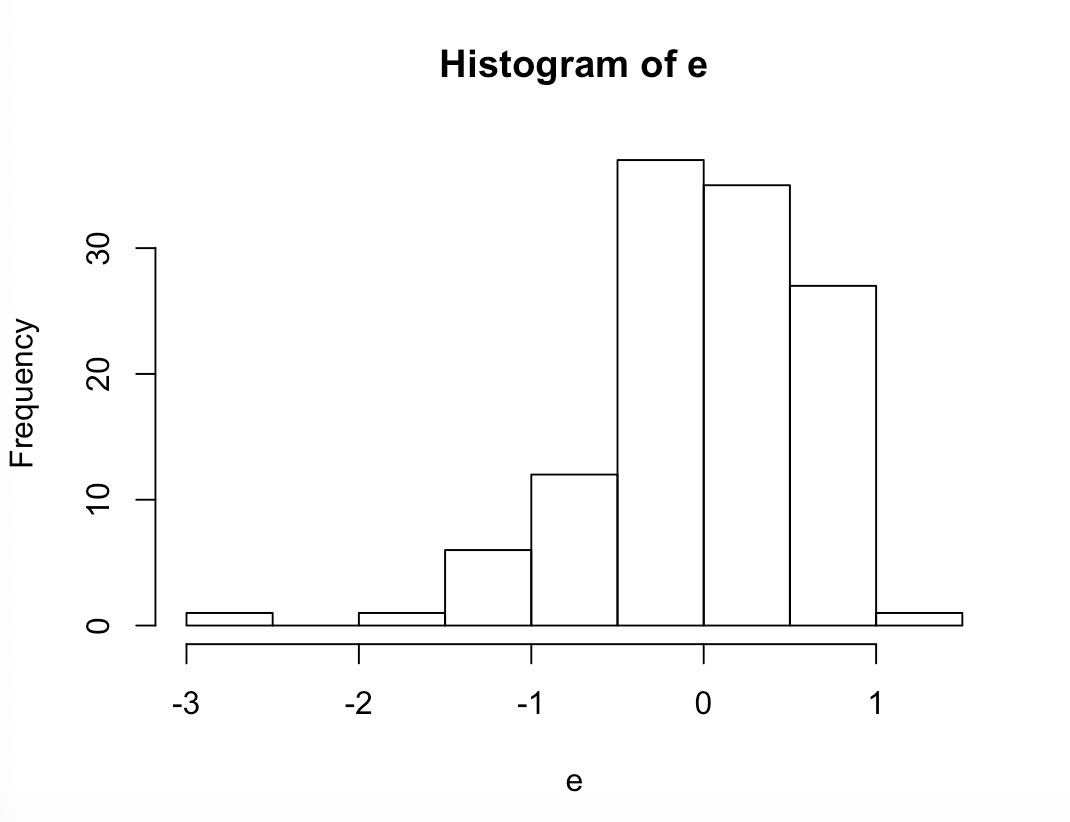
> # 1.b

> # Obtain the residuals:

> (e = resid(reg))

> # Prepare a histogram of the residuals:

> hist(e)



> # The plot provides the information that it is not symmetric, and most of the residuals are between -1 and 1, so there are some outliers.

> # 1.c

> # Plot the residuals against the fitted values y.hat

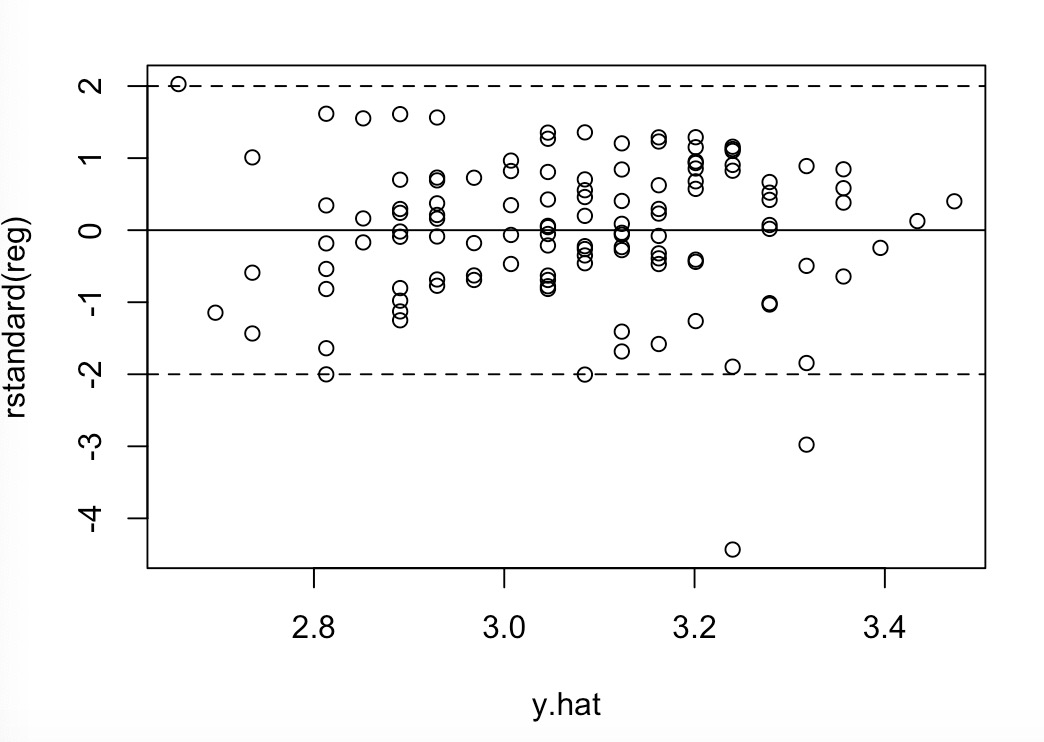
> y.hat = predict(reg)

> plot(y.hat,rstandard(reg))

> abline(h=0)

> abline(h = -2, lty =2)

> abline(h = 2, lty =2)



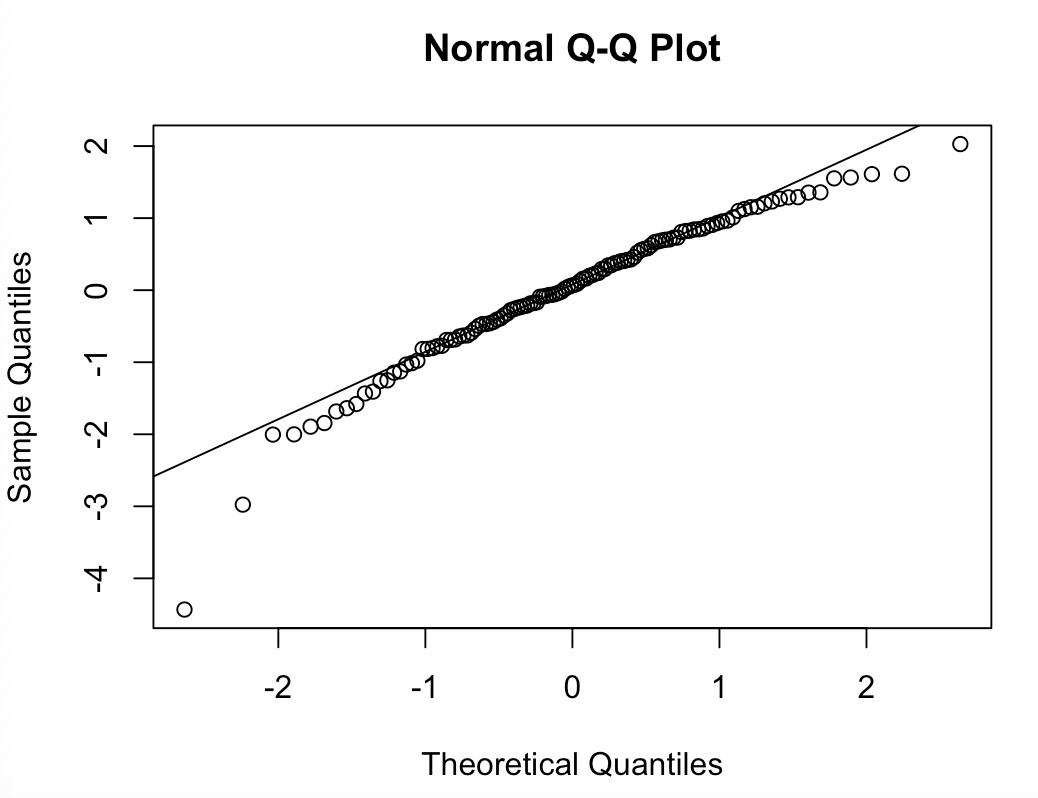
> # There are some outliers beyond [-2,2] and constant variance looks ok because it is scattered around zero randomly.

> # 1.d

> # Prepare a normal probability plot of the residuals

> qqnorm(rstandard(reg))

> qqline(rstandard(reg))

>

> # Kolmogorov-Smirnov test for normality

> ks.test(rstandard(reg), "pnorm")

One-sample Kolmogorov-Smirnov test

data: rstandard(reg)

D = 0.057725, p-value = 0.8188

alternative hypothesis: two-sided

> # 0.8188 is not less than 0.05

> # Do not reject Ho

> # That is, normality assumption holds

> # 1.e

> # Breusch-Pagan test for constant variance

> bptest(reg)

studentized Breusch-Pagan test

data: reg

BP = 0.29397, df = 1, p-value = 0.5877

> # 0.5877 is not less than 0.01

> # Do not reject Ho

> # The conclusion: constant variance assumption holds

> # My conclusion supports my preliminary findings in part c)

> # 2

> # Read table3.10

> t3.10 <- read.table("http://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets/KutnerData/Chapter%20%203%20Data%20Sets/CH03PR10.txt", header = FALSE)

> # Column names can be changed with

> names(t3.10)[1]<-paste("yi.hat")

> names(t3.10)[2]<-paste("ei.star")

> t3.10

> # To use the column names without reference to t1.19 you need to attach the dataset:

> attach(t3.10)

>

> # 2.a

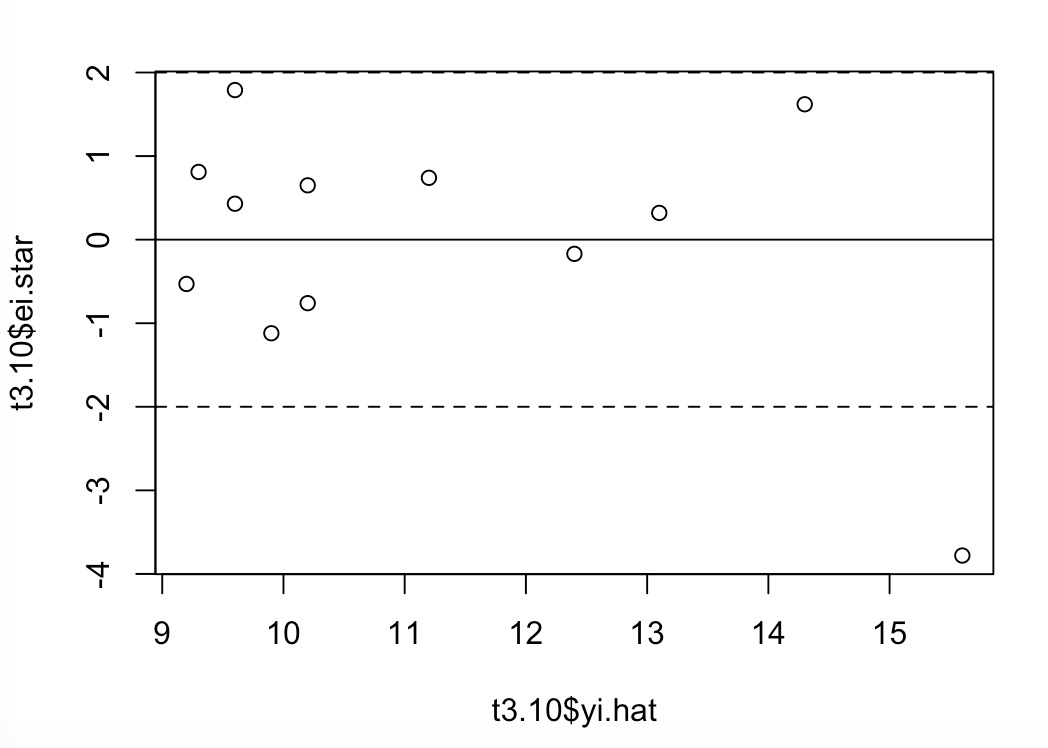
> # Plot the semistudentized residuals against the fitted values

> plot(t3.10$yi.hat,t3.10$ei.star)

> abline(h=0)

> abline(h = -2, lty =2)

> abline(h = 2, lty =2)



> # The plot suggests that there is one outlier beyond [-2,2], and the data set only contains 12 samples, which is a quite small data set.

>

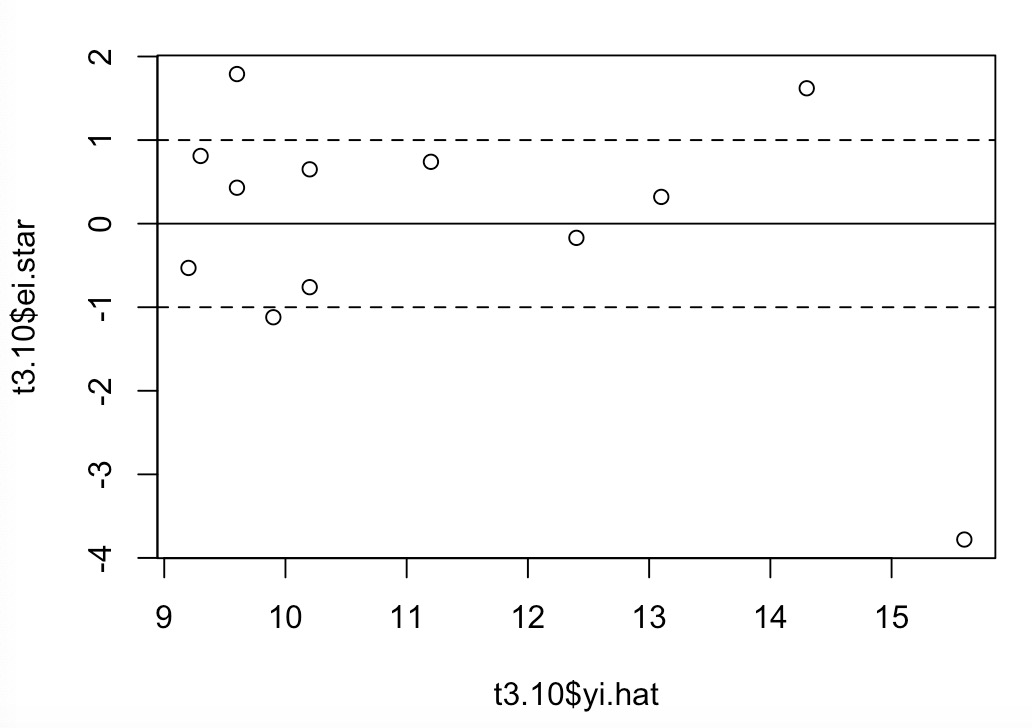
> # 2.b

> plot(t3.10$yi.hat,t3.10$ei.star)

> abline(h=0)

> abline(h = -1, lty =2)

> abline(h = 1, lty =2)



> # 4 out of 12 (1/3) semistudentized residuals are outside ±1 standard deviation

> # (1-68%)\*12= 3.84

> # 68%\*12= 8.16

> # since 68% are within ±1 standard deviation if the normal error model is appropriate

> # Approximately 4 semistudentized residuals are outside ±1 standard deviation and 8 semistudentized residuals are within ±1 standard deviation I would expect to see if the normal error model is appropriate

> # 3.

load("/Users/huchengxuan/Desktop/Master/HUDM5126/HW/HW3.RData")

> # Read data

> attach(data)

>

> # 3.a

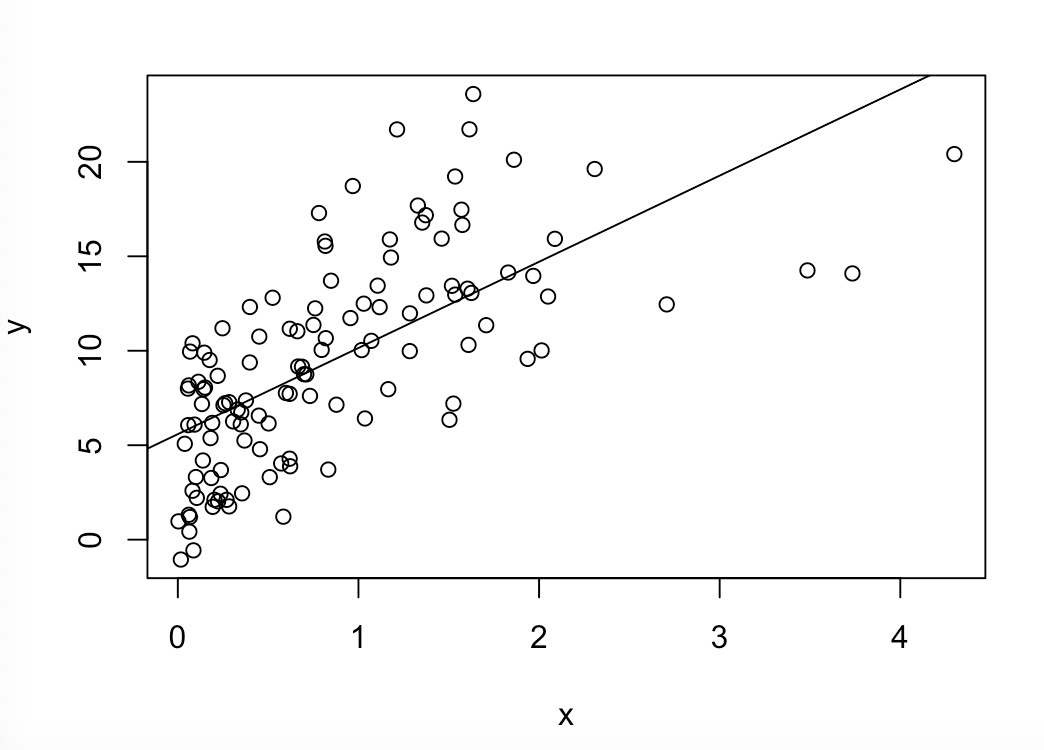
> # Prepare a scatterplot of X vs. Y overlaid with the estimated regression line

> plot(x,y)

>

> reg1 = lm(y~x)

> abline(reg1)



> # 3.b

> # Calculate the correlation coefficient between X and Y

> cor(x,y)

[1] 0.6642197

> # The correlation coefficient between X and Y is 0.6642197

> # There is moderate linear relationship between x and y

> # becasue the correlation coefficient is between -0.7 and 0.7

> # 3.c

> x.prime<-sqrt(x)

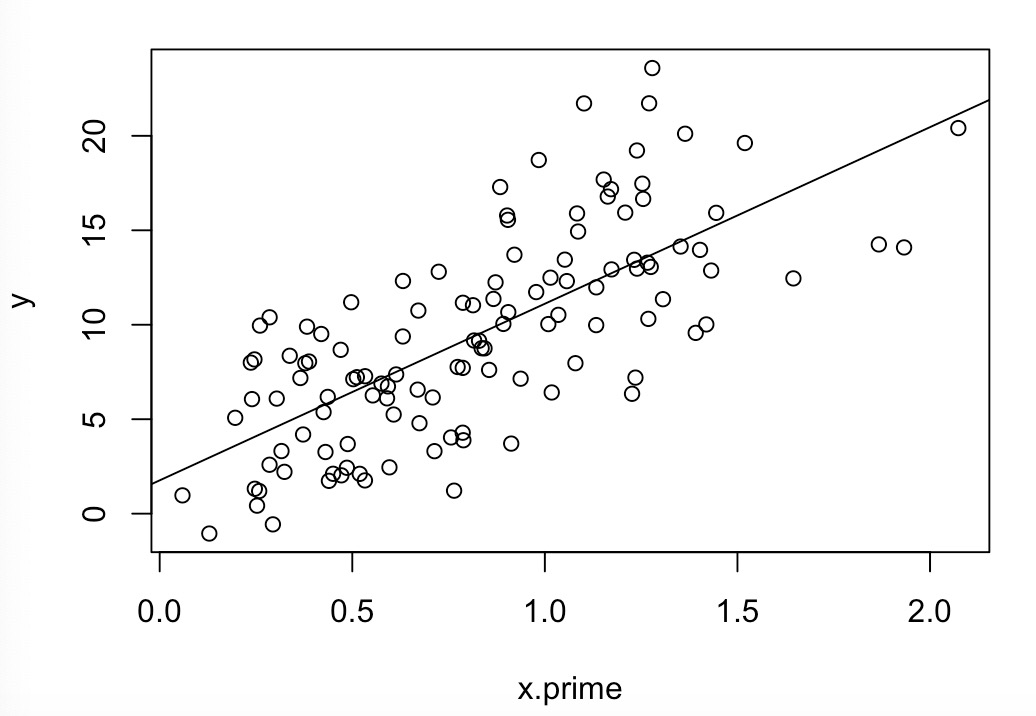
> # 3.d

> # Prepare a scatterplot of 𝑋′vs. Y overlaid with the estimated regression line

> plot(x.prime,y)

>

> abline(reg2)



> # 3.e

> # Calculate the correlation coefficient between 𝑋′ and Y

> cor(x.prime,y)

[1] 0.7171237

> # The correlation coefficient between X and Y is 0.7171237

> # There is strong linear relationship between x and y

> # because the correlation coefficient is not between -0.7 and 0.7

> # 3.f

> # Obtain the estimated linear regression function for the transformed data

> reg2 = lm(y~x.prime)

> summary(reg2)$coefficients

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.763567 0.7691637 2.292836 2.362860e-02

x.prime 9.349671 0.8364891 11.177278 3.188734e-20

>

> # y.hat = 1.7636 + 9.3497\*x.prime

> # 3.g

> # Obtain the residuals:

> (e = resid(reg2))

> # Plot the residuals vs. fitted values

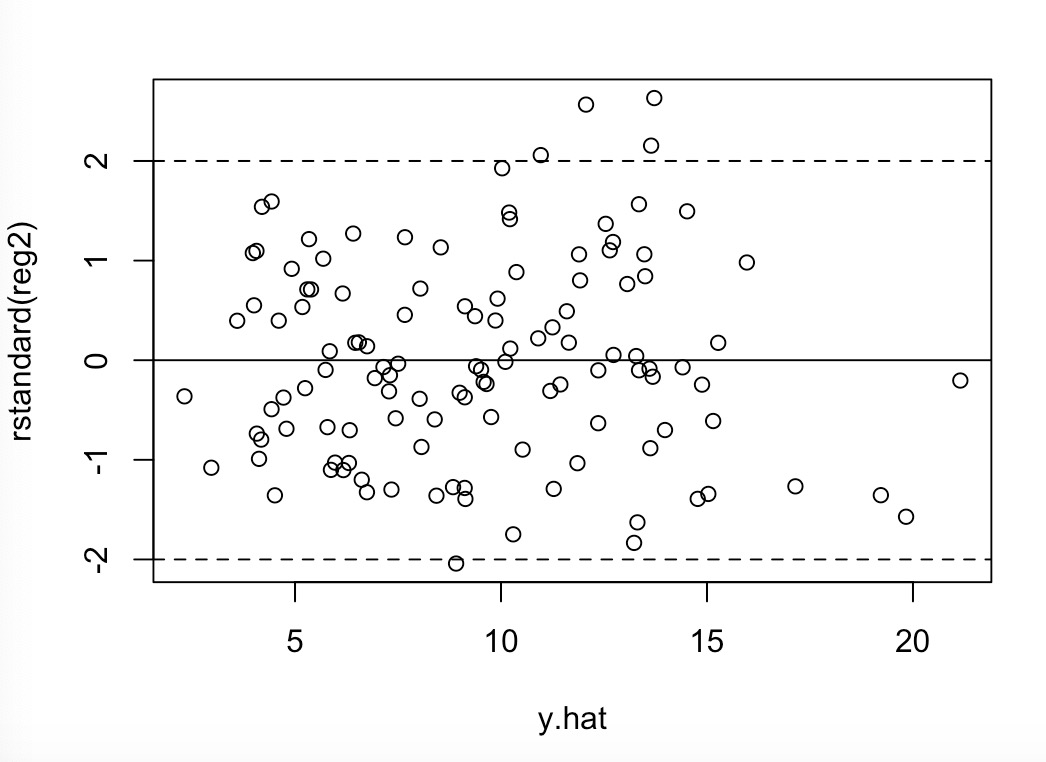
> y.hat = predict(reg2)

> plot(y.hat,rstandard(reg2))

> abline(h=0)

> abline(h = -2, lty =2)

> abline(h = 2, lty =2)



> # 3.h

> # The plot from part g) shows there are some outliers beyond [-2,2] and constant variance looks ok because it is randomly scattered around zero. There might be linearity.