**STAT 512**

**HW 2**

**Name: Haoran Zhang**

**Section: 010**

**Time: 1:30-2:45 pm Tuesday/Thursday**

1. (2) A member of a student team playing an interactive marketing game received the following computer output when studying the relation between advertising expenditures (X) and sales (Y) for one of the team’s products:

Estimated regression equation: Ŷ = 350.7 - .18X

Two-sided P-value for estimated slope: .91

The student stated: “the message I get here is that the more we spend on advertising this product, the fewer units we sell!” Comment.

Assume that the significance level is 0.05, so the given p-value, which is 0.91, is larger than , therefore we could not reject the null hypothesis. As a result, the statement of this student is incorrect, because there is no evidence to show that .

2. (4) Refer to the problem 5 in the homework1, use R to generate confidence band for

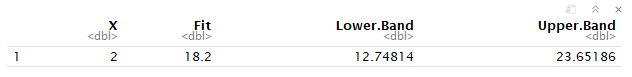
a) mean value prediction, b) single value prediction, c) mean of 3 new values prediction, and d) working-hoteling confidence band. Comment on their difference.

Assume that the significance level is 0.01

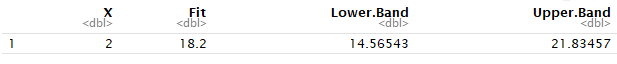
1. The result is shown as below:



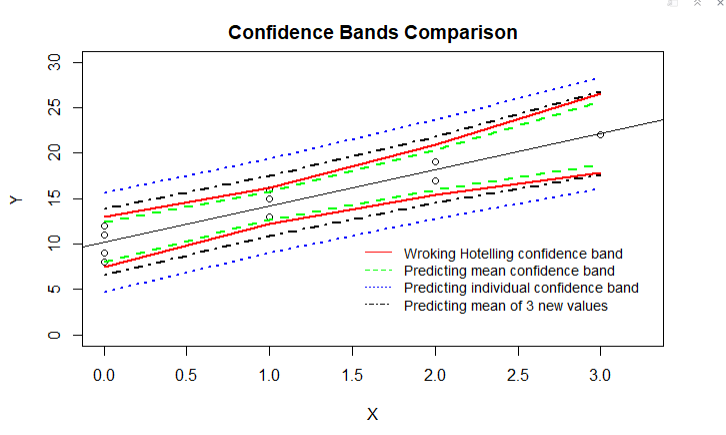
1. The result is shown as below:



1. The result is shown as follows:



1. The plot is shown as below:



Comments:

The confidence interval (CI) of next single shipment is always wider than the confidence interval of mean value, because when the CI of new single shipment is predicted, both variances from the difference of sample and difference of should be considered. In contrast, only the variance of should be considered when calculating the CI of mean value.

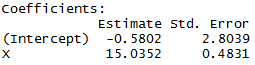
The confidence interval of mean of next 3 new shipments is wider than CI of mean value, but narrower than CI of next single shipment, because the variance of mean of 3 new shipments includes the variance from different and the variance from predicting next 3 new samples, which is smaller than variance of single sample.

For Working-Hotelling band, the calculation formula is:

, where

Since the Working-Hotelling confidence band should encompass the entire regression line rather than a single , the should be larger than . Therefore, as shown in the plot, the Working-Hotelling bands are a little wider apart than the mean value prediction band.

3. (19) Tri-City Office Equipment Corporation sells an imported copier on franchise basis and performs preventive maintenance and repair service on this copier. Data have been collected from 45 recent calls on users to perform routine preventive maintenance service; for each call, X is the number of copiers serviced and Y is the total number of minutes spent by the service person. Assume that simple linear regression model is appropriate. The following shows partial result.





1. (3) Estimate the change in the mean service time when the number of copiers serviced increases by one. Use a 90 percent confidence interval. Interpret your confidence interval.

It is 90 percent confident that if the number of copiers serviced increment by one, the change in the mean service time will drop into the range of .

1. (4) Conduct a t test to determine whether or not there is a linear association between X and Y here (i.e., ; At a significant level of 0.05, state the hypothesis, reject region, estimate the p value, and state the conclusion of your test.

Therefore,

Since , The rejection region is:

Besides, the corresponding p-value is <0.001

As a result, since , so it is evident that the null hypothesis should be rejected, and there is a linear association between X and Y.

1. (4) The manufacturer has suggested that the mean required time should not increase by more 14 minutes for each additional copier that is serviced on a service call. Conduct a test to test whether this standard is being satisfied by Tri-City. At a significant level of 05. State the hypothesis, reject region, estimate the p value, and state the conclusion of your test.

Therefore, using t test:

Since , The rejection region is:

Besides, the corresponding p-value is 0.01892

As a result, since , so it is evident that the null hypothesis should be rejected, and this standard is satisfied by Tri-City.

1. (2) Does give any relevant information here about the “start-up” time on calls—i.e. about time required before service work is begun on the copiers at a customer location?

The intercept does not give any relevant information about the “start-up” time on calls, because it is not related to “start-up” time.

Specifically,

Therefore, using t test:

Since , The rejection region is:

Besides, the corresponding p-value is 0.8371

As a result, since , so it fails to reject the null hypothesis, and it cannot prove that .

1. (4) In order to perform the following hypothesis test (,

complete the following ANOVA table for the data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Source** | **degrees of freedom** | **Sum of Squares** | **Mean Square** | **F-value** |
| **Model** | **\_\_1\_\_** | **76960** | **76960** | **968.5449** |
| **Error** | **\_\_43\_\_** | **3416.7540** | **79.4594** |  |
| **Corrected Total** | **\_\_44\_** |  |  |  |

According to the F value and degree of freedom, use F tale to estimate the P-value of the test.

According to the F table, when , the p-value < 0.001.

f) (2) Compare the F test statistic obtained here and demonstrate numerically its equivalence to the T test statistic in b).

For t test,

For F test,

Therefore, .

4. (17) Refer to the problem 2 in homework 1.

a) (6) Complete the ANVOA table for the hypothesis test.

ACT and GPA score are not associated ACT and GPA score are associated

Or equivalently,

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Source** | **degrees of freedom** | **Sum of Squares** | **Mean Square** | **F-value** |
| **Model** | **\_\_1\_\_\_** | **0.009976** | **0.009976** | **0.1887** |
| **Error** | **\_\_8\_\_\_** | **0.4228** | **0.05285** |  |
| **Corrected Total** | **\_\_9\_\_\_** |  |  |  |

Estimate the p value and state the conclusion

Given that ,

According to the F table, when .

Therefore, it is evident that the null hypothesis should not be rejected, and there should be no linear association between ACT and GPA scores.

b) (4) what is ? Perform a hypothesis test on the correlation, compute the test statistic and estimate p value, then state the conclusion. Use significant level of 0.05.

is the coefficient of determination, it represents the proportionate reduction of total variation associated with the predictor variable X. The formula is . Since , . When larger is, the more the total variation of Y is reduced.

Since , The rejection region is:

The corresponding p-value is 0.6755.

Since , this null hypothesis should not be rejected. Therefore, there is no linear association between ACT and GPA scores.

c) (2) Compare 4a) to 3e), which model seems to be a better fit? Discuss the models based on MSE and

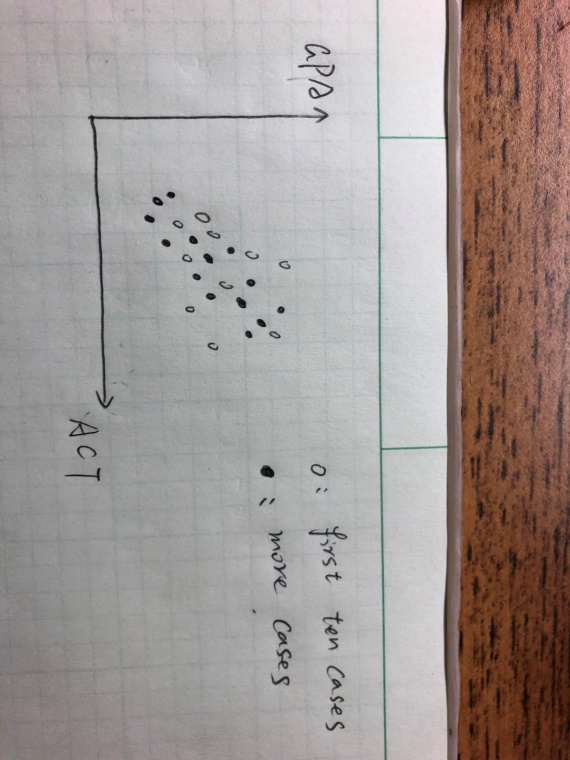
for 3e), and

for 4a), and

The best model is the model with highest and lowest MSE, but 3e) model has high and high MSE; while 4a) model has low and low MSE. In my opinion, 3e) model is better, since and low MSE means that the degree of linear association between X and Y is really high, but with a significant variance.

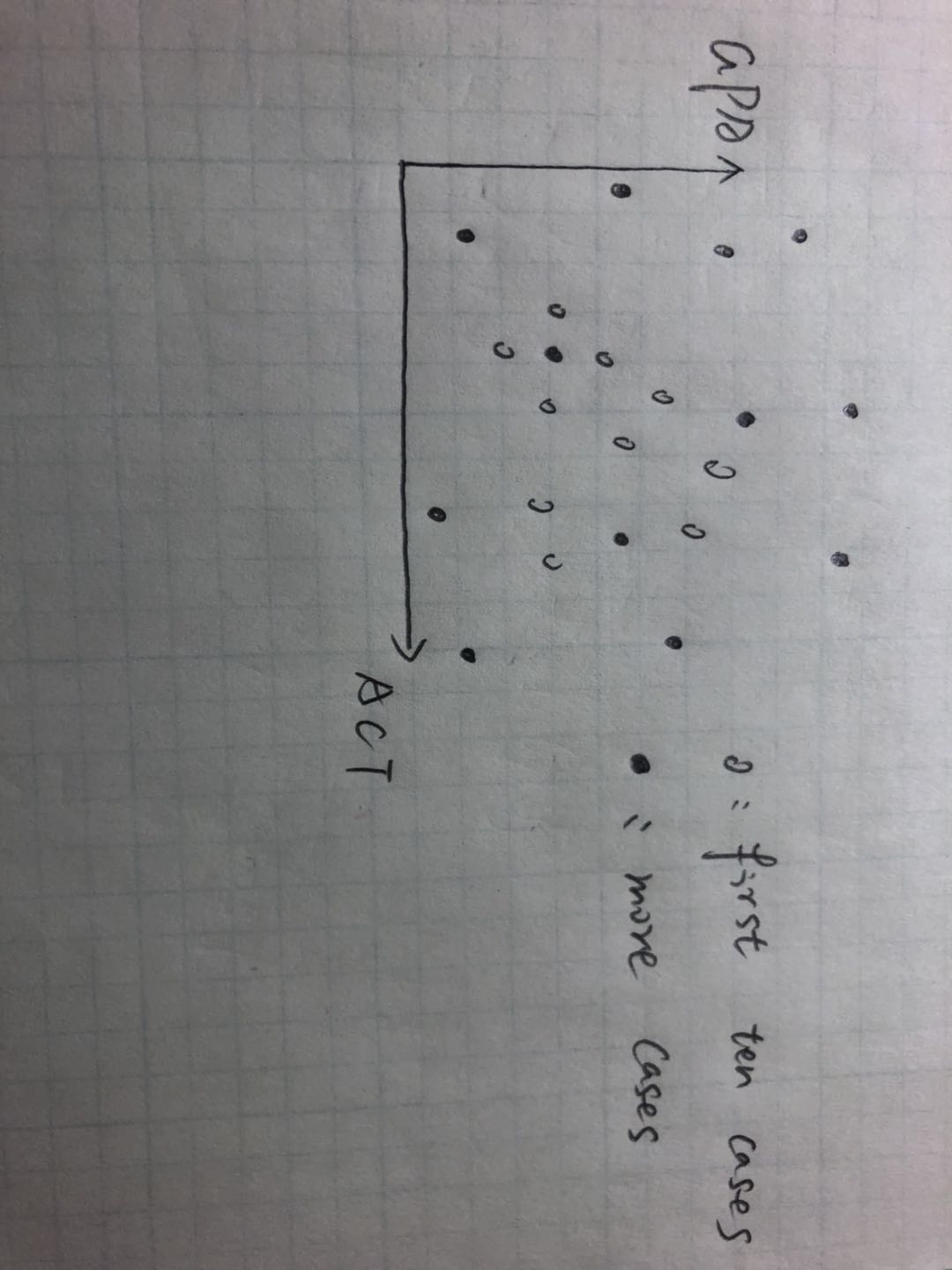
d) (3) The GPA data used in this problem is actually the first 10 cases of a larger data set, and has a very small , is it possible that for the complete set n>10, will not be zero? Could not be zero for the first 10 cases, yet equal to zero for all 30 cases? If applicable, sketch two scatter plots to demonstrate the two situations.

Yes, the could not be zero when the complete set n>10.



As shown in the above plot, when those black data points added, the degree of linear association between ACT and GPA scores will be much higher, therefore the resultant will not be zero.

Yes, could be zero when the complete set n>10.



As shown in the above plot, when those black data points included in the data set, could be zero, since the whole plot is a random scatter plot, which means there is no evidence that there is a linear association between ACT and GPA scores.

e) (2) Use R to compute a 95% CI for the population coefficient.

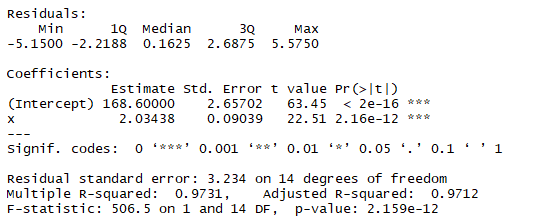
.

5. (Use R to complete this problem)(8) Experience with a certain type of plastic indicates that a relation exists between the hardness (measured in Brinell units) of items molded from the plastic, and the elapsed time since termination of the molding process.

Sixteen batches of the plastic were made, and one test item was molded from each batch. Each test item was randomly assigned to one of four predetermined time levels (X=16, 24, 32, or 40 hours), and the hardness (Y) was measured after the assigned elapsed. Data is in plastic.csv

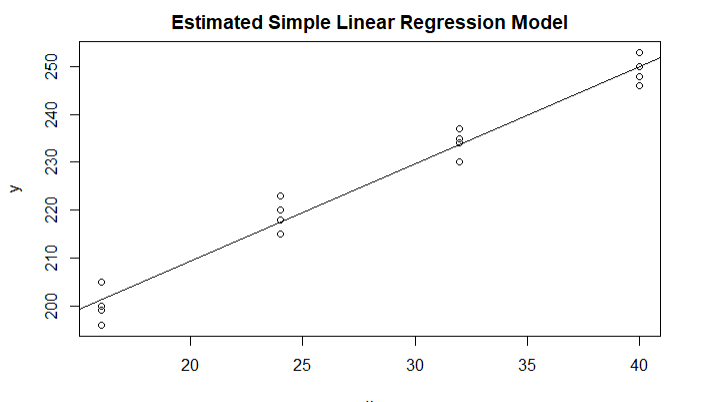
1. (2) Obtain the estimated simple linear regression model (SLR or SLM). Plot the estimated regression function and the data. Does a linear regression function appear to be a good fit?

The data of estimated simple linear regression model is:



The fitted equation is:

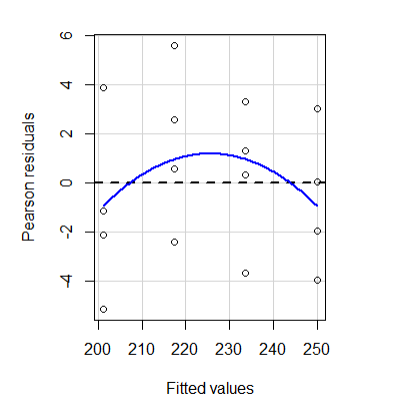
The plot of the regression model is shown as below:



The estimated linear model seems to be a good fit, since is close to 1, which means the linear association between Y and X is apparent, although the MSE is higher than expected, which means the variance is significant.

1. (2) Plot the residuals against the fitted values to ascertain whether any departures from regression model are evident. State your findings

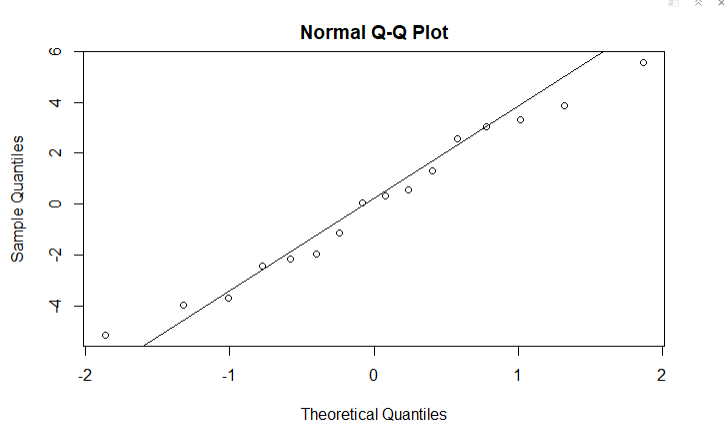
The residuals against the fitted values is shown as follows:



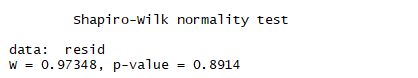
The departure from the fitted model is not evident, since no apparent pattern can be concluded from this plot. More data could improve diagnostics with the residual plot.

1. (2) Plot a normal probability plot of the residuals. Perform a Shapiro-wilk normality test on the residuals. Does the normality assumption appear to be reasonable here?

The normal probability plot of the residuals is shown as below:



The result of Shapiro-wilk normality test is:



Since p-value of Shapiro-wilk normality test equals to 0.894, which is larger than 0.05, the residuals are normally plotted.

1. (2) Use the Brown-Forsythe test to determine whether or not the error variance varies with the level of X. Divide the data into two groups, , use . Does your conclusion support your preliminary findings in part b)?

The results are shown as below:



Since p-value >0.05, it is evident that departure from the model is not significant, which is the same as the result of part b).

Hw2\_Coding

Haoran Zhang

09/14/2018

## hw2 problem 2a)  
## Loading library and read the file  
library(ALSM)  
##library(alr4)  
flights <- read.csv(file="C:/Users/Haoran Zhang/Downloads/airfreight.csv",  
 header=TRUE,sep=",")  
## plot(Y~X, data = flights)  
  
## Linear regression model  
## Problem 2 (a)  
flights.mod <- lm(Y~X, data = flights)  
summary(flights.mod)

##   
## Call:  
## lm(formula = Y ~ X, data = flights)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.2 -1.2 0.3 0.8 1.8   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.2000 0.6633 15.377 3.18e-07 \*\*\*  
## X 4.0000 0.4690 8.528 2.75e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.483 on 8 degrees of freedom  
## Multiple R-squared: 0.9009, Adjusted R-squared: 0.8885   
## F-statistic: 72.73 on 1 and 8 DF, p-value: 2.749e-05

anova(flights.mod)

## Analysis of Variance Table  
##   
## Response: Y  
## Df Sum Sq Mean Sq F value Pr(>F)   
## X 1 160.0 160.0 72.727 2.749e-05 \*\*\*  
## Residuals 8 17.6 2.2   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

new <- data.frame(X = 2) ## make it new observation at X = 2  
ci.reg(flights.mod, new, type='m', alpha=0.01) ## mean value prediction

## X Fit Lower.Band Upper.Band  
## 1 2 18.2 15.97429 20.42571

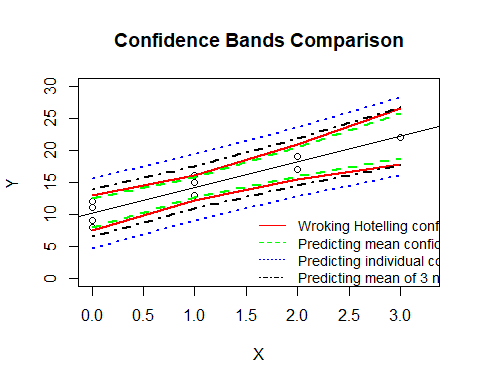
## problem 2b)  
ci.reg(flights.mod, new, type='n',alpha=0.01) ## single value prediction

## X Fit Lower.Band Upper.Band  
## 1 2 18.2 12.74814 23.65186

## problem 2c)  
ci.reg(flights.mod, new, type='nm', m=3, alpha=0.01) ## 3 new values prediction

## X Fit Lower.Band Upper.Band  
## 1 2 18.2 14.56543 21.83457

## problem 2d)  
x<-data.frame(X=sort(unique(flights$X)))  
cim<-ci.reg(flights.mod, x, type='m',alpha=0.01) ## mean value prediction  
cin<-ci.reg(flights.mod, x, type='n',alpha=0.01) ## single value prediction  
cinm<-ci.reg(flights.mod, x, type='nm', m=3, alpha=0.01) ## 3 new values prediction  
ciw<-ci.reg(flights.mod, x, type='w',alpha=0.01) ##working-hotel CI at each x  
  
##plot all 4 bands  
plot(Y~X, flights, ylim=c(0,30), xlim=c(0.0,3.25), main="Confidence Bands Comparison")  
abline(lm(Y~X, flights))  
lines(ciw$X, ciw$Lower.Band,col="red",lwd=2, lty=1)  
lines(ciw$X, ciw$Upper.Band, col="red", lwd=2, lty=1)  
lines(ciw$X, cim$Lower.Band,col="green", lwd=2, lty=2)  
lines(ciw$X, cim$Upper.Band, col="green", lwd=2, lty=2)  
lines(ciw$X, cin$Lower.Band,col="blue", lwd=2, lty=3)  
lines(ciw$X, cin$Upper.Band, col="blue", lwd=2, lty=3)  
lines(ciw$X, cinm$Lower.Band,col="black", lwd=2, lty=4)  
lines(ciw$X, cinm$Upper.Band, col="black", lwd=2, lty=4)  
  
legend(x=1.5, y=11, legend=c("Wroking Hotelling confidence band","Predicting mean confidence band","Predicting individual confidence band","Predicting mean of 3 new values"), lty=c(1,2,3,4), col=c("red","green","blue","black"),cex=0.9, bty="n")



## problem 3b)  
alpha=0.05  
n=45  
pvalue\_t<-qt(1-0.5\*alpha,n-2)  
pvalue\_f<-qf(1-alpha,1,n-2)  
2\*pt(-abs(31.1232),df=n-2)

## [1] 4.009341e-31

## problem 3c)  
alpha=0.05  
n=45  
pvalue\_t<-qt(1-0.5\*alpha,n-2)  
pvalue\_f<-qf(1-alpha,1,n-2)  
pt(-abs(2.1428),df=n-2)

## [1] 0.01891548

## problem 3d)  
alpha=0.05  
n=45  
pvalue\_t<-qt(1-0.5\*alpha,n-2)  
pvalue\_f<-qf(1-alpha,1,n-2)  
2\*pt(-abs(0.2069),df=n-2)

## [1] 0.8370646

## problem 4a)  
##read files and plot  
Grades <- read.csv(file="C:/Users/Haoran Zhang/Desktop/GPA.csv",  
 header=TRUE,sep=",")  
##plot(GPA~ACT, data = Grades)  
  
##linear regression model  
Grades.mod <- lm(GPA~ACT, data = Grades)  
summary(Grades.mod)

##   
## Call:  
## lm(formula = GPA ~ ACT, data = Grades)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.21921 -0.19204 -0.03343 0.12015 0.36008   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.90826 0.63244 4.598 0.00176 \*\*  
## ACT 0.01015 0.02335 0.434 0.67544   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2299 on 8 degrees of freedom  
## Multiple R-squared: 0.02305, Adjusted R-squared: -0.09907   
## F-statistic: 0.1887 on 1 and 8 DF, p-value: 0.6754

anova(Grades.mod)

## Analysis of Variance Table  
##   
## Response: GPA  
## Df Sum Sq Mean Sq F value Pr(>F)  
## ACT 1 0.00998 0.009976 0.1887 0.6754  
## Residuals 8 0.42283 0.052854

##problem 4b)  
alpha=0.05  
n=10  
qt(1-0.5\*alpha,n-2)

## [1] 2.306004

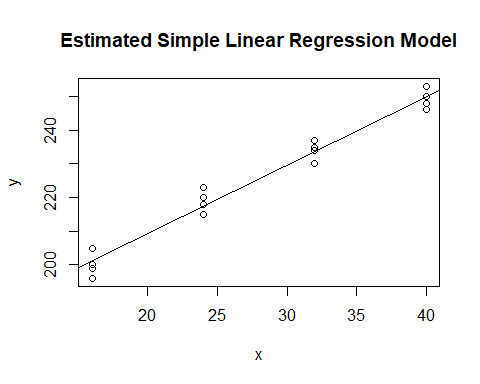
2\*pt(-abs(0.4344),df=n-2)

## [1] 0.6754766

##problem 4e)  
cor.test(Grades$GPA,Grades$ACT, conf.leve=0.95)

##   
## Pearson's product-moment correlation  
##   
## data: Grades$GPA and Grades$ACT  
## t = 0.43445, df = 8, p-value = 0.6754  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.5283062 0.7132660  
## sample estimates:  
## cor   
## 0.1518212

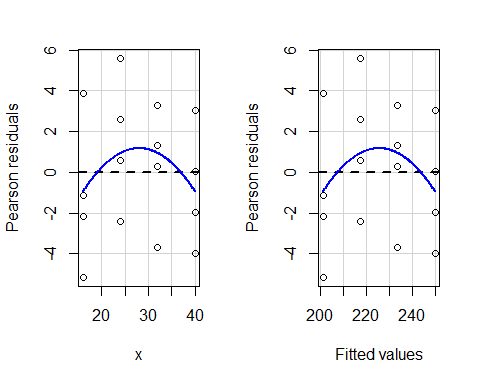
## hw2 problem 5 a)  
##read files and plot  
Plastic<- read.csv(file="C:/Users/Haoran Zhang/Downloads/plastic.csv",  
 header=TRUE,sep=",")  
plot(y~x, data = Plastic, main="Estimated Simple Linear Regression Model")  
Plastic.mod<-lm(y~x, data=Plastic)  
abline(Plastic.mod)



summary(Plastic.mod)

##   
## Call:  
## lm(formula = y ~ x, data = Plastic)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.1500 -2.2188 0.1625 2.6875 5.5750   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 168.60000 2.65702 63.45 < 2e-16 \*\*\*  
## x 2.03438 0.09039 22.51 2.16e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.234 on 14 degrees of freedom  
## Multiple R-squared: 0.9731, Adjusted R-squared: 0.9712   
## F-statistic: 506.5 on 1 and 14 DF, p-value: 2.159e-12

## problem 5 b)  
resid<-residuals(Plastic.mod)  
residualPlots(Plastic.mod)

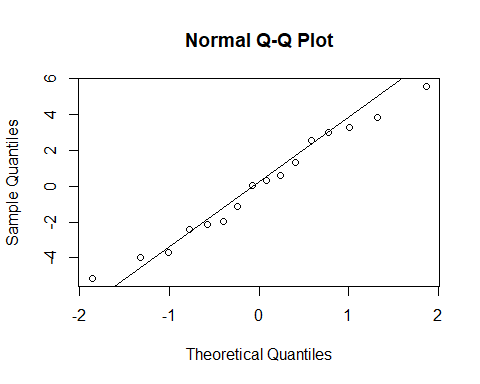


## Test stat Pr(>|Test stat|)  
## x -1.1752 0.2610  
## Tukey test -1.1752 0.2399

## problem 5 c)  
shapiro.test(resid)

##   
## Shapiro-Wilk normality test  
##   
## data: resid  
## W = 0.97348, p-value = 0.8914

qqnorm(resid)  
qqline(resid)



## problem 5 d)

library(ALSM)   
g<-rep(1,16)  
g[Plastic$x<=24]=0 #form two groups  
bftest(Plastic.mod,g)

## t.value P.Value alpha df  
## [1,] 0.8557853 0.4065253 0.05 14