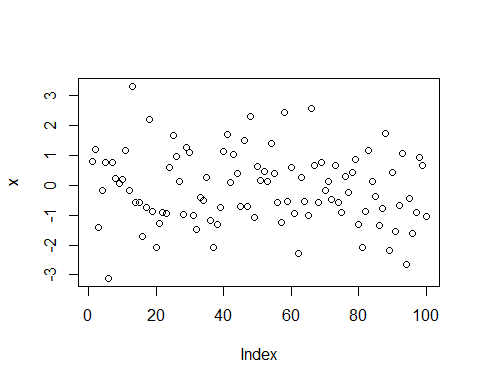
HW8

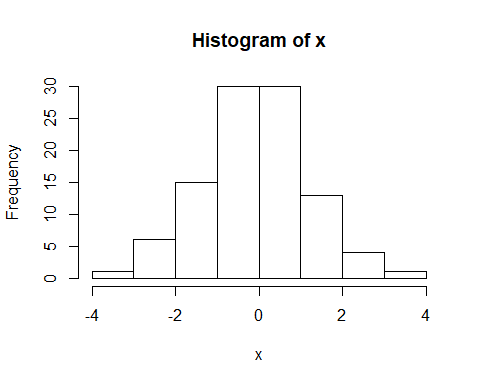
Naimon Hafiz

March 31, 2019

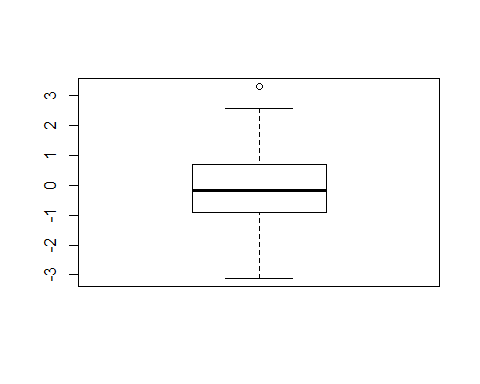
#1.  
x<-rnorm(100)##100 random numbers from a standard normal distribution  
plot(x)



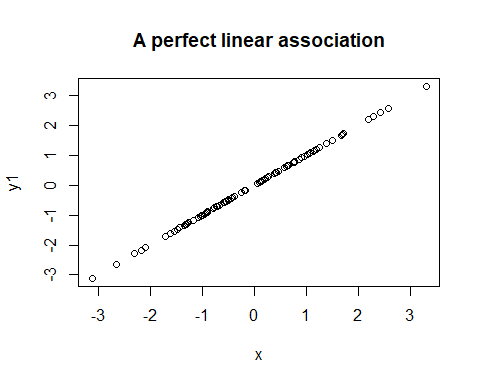
hist(x)



boxplot(x)



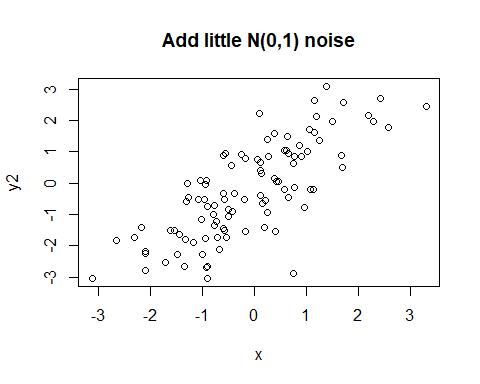
#2  
y1<-0+1\*x  
y2<- 0+1\*x+rnorm(length(x),0,sqrt(1)) #sigma^2=1  
y3<- 0+1\*x +rnorm(length(x),0,sqrt(0.1)) #sigma^2=0.1  
y4<- 0+1\*x +rnorm(length(x),0,sqrt(2)) #sigma^2=2  
y5<- 0+1\*x +rnorm(length(x),0,sqrt(3)) #sigma^2=3  
y6<- 0+1\*x +rnorm(length(x),0,sqrt(4)) #sigma^2=4  
  
  
##3 & 4  
plot(x,y1, main="A perfect linear association")



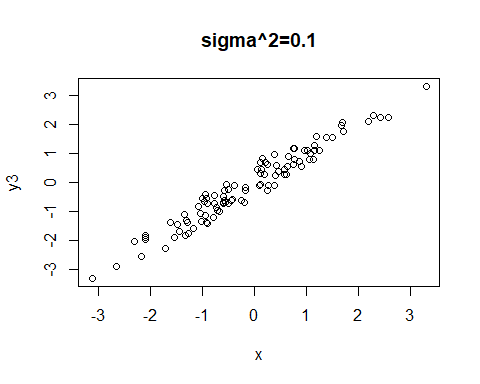
r.1=cor(x,y1) ## correalation coefficient r=1  
R2.1<-r.1^2 # Coefficient of determination R=1  
R2.1a<-summary(lm(y1~x))$r.squared ## Coefficient of determination 1

## Warning in summary.lm(lm(y1 ~ x)): essentially perfect fit: summary may be  
## unreliable

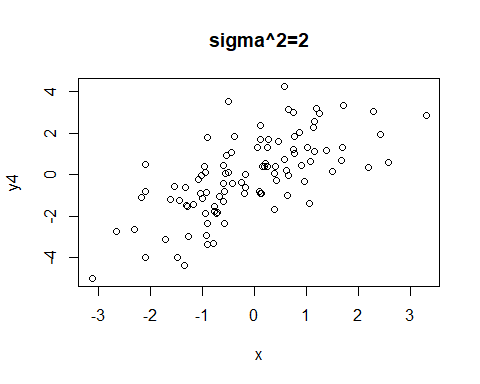
plot(x,y2, main="Add little N(0,1) noise")



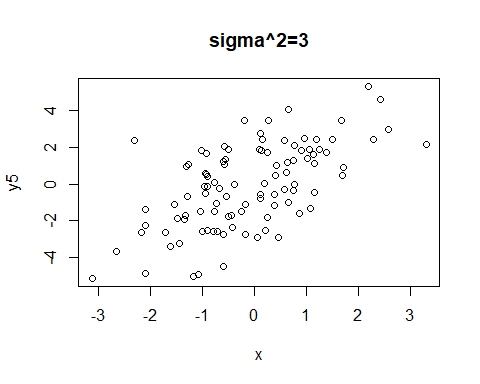
r.2=cor(x,y2) ## correalation coefficient r=7127166  
R2.2<-r.2^2 # Coefficient of determination R=0.507965  
R2.2a<-summary(lm(y2~x))$r.squared ## Coefficient of determination 0.507965  
  
  
plot(x,y3, main="sigma^2=0.1")



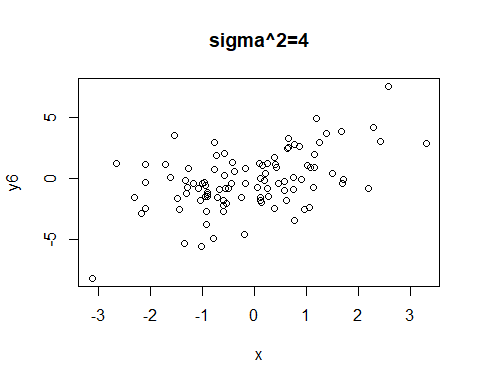
r.3=cor(x,y3) ## correalation coefficient r=0.9473147  
R.3<-r.3^2 # Coefficient of determination R=0.8974052  
R.3a<-summary(lm(y3~x))$r.squared ## Coefficient of determination 0.8974052  
  
  
plot(x,y4, main="sigma^2=2")



r.4=cor(x,y4) ## correalation coefficient r=0.5802353  
R.4<-r.4^2 # Coefficient of determination R=0.336673  
R.4a<-summary(lm(y4~x))$r.squared ## Coefficient of determination 0.336673  
  
plot(x,y5, main="sigma^2=3")



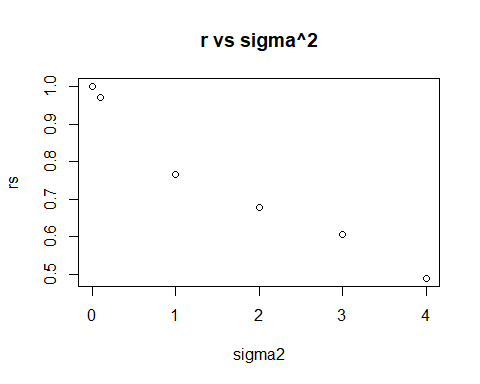
r.5=cor(x,y5) ## correalation coefficient r=0.4626949  
R.5<-r.5^2 # Coefficient of determination R=0.2140866  
R.5a<-summary(lm(y5~x))$r.squared ## Coefficient of determination 0.2140866  
  
  
  
plot(x,y6, main="sigma^2=4")



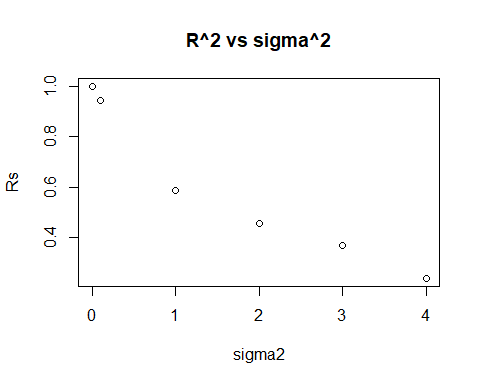
r.6=cor(x,y6) ## correalation coefficient r=0.4188453  
R.6<-r.6^2 # Coefficient of determination R=0.1754314  
R.6a<-summary(lm(y6~x))$r.squared ## Coefficient of determination 0.1754314  
r.6

## [1] 0.48943

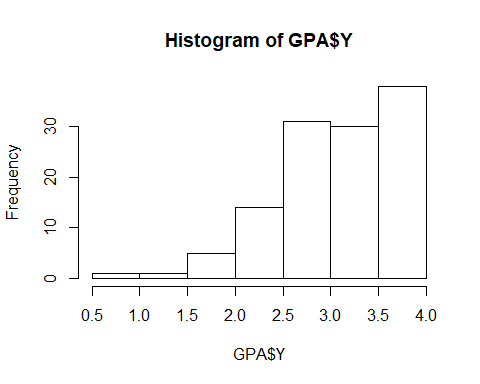
##5  
  
sigma2 <- c(0,1,0.1,2,3,4) # a vector of the sigma^2 values I chose  
rs <- c(r.1, r.2, r.3, r.4, r.5, r.6) # a vector of all the r's  
Rs <- c(R2.1, R2.2, R.3, R.4, R.5, R.6) # a vector of all the R^2 values  
  
plot(sigma2, rs, main="r vs sigma^2")



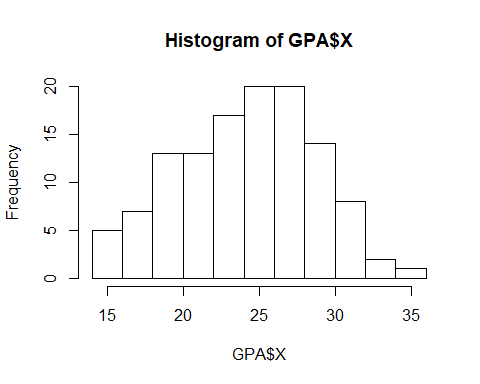
#correlation co-efficient,r and sigma square have invers relationship. When  
#sigma\_square is low r is high. And when sigma\_square is high, r is low.  
  
plot(sigma2, Rs, main="R^2 vs sigma^2")



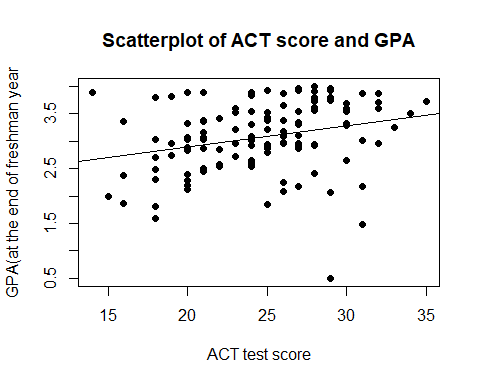
#Coefficient of determination,R and sigma square have opposite relationship.  
#When sigma\_square is low, R is high and when sigma\_square is high,R is low  
  
  
##Simple Linear Regression  
#1. Here Y= GPA at the end of the freshman year  
#X= AcT score  
#The regression model will be  
  
#GPA= Beta\_0+Beta\_1\*AcT score+Epsilon  
  
#The model Assumptions are--i.The response is linearly related to the Beta's.  
#ii.Once we have accounted for the relationship between the DV and IVs,   
#all that should be left is random noise.  
#iii.The data are an independent sample  
#iv.The variance of the data sigma\_square is not a function of any of the   
#independent variables, which is called homoscedasticity.  
  
#2.  
GPA <- read.table("C:/Users/naimo/Desktop/823 spring/data/data/CH01PR19.txt")  
   
# View and clean the data  
View(GPA)  
names(GPA) <- c("Y","X")  
attach(GPA)  
  
#Visualize the data-  
hist(GPA$Y)



hist(GPA$X)



#variable Y does not look bellshaped but variable X looks kind of bellshaped.  
  
# Plot to make sure a linear relationship makes sense  
plot(Y~X, xlab = "ACT test score", ylab ="GPA(at the end of freshman year", pch = 16,  
 main = "Scatterplot of ACT score and GPA")  
m1 <- lm(Y~X)  
abline(m1)



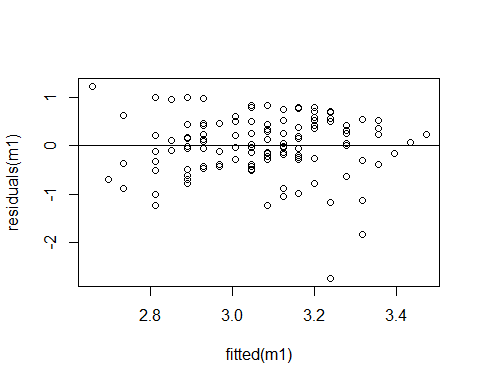
m1s<-summary(m1)  
names(m1)

## [1] "coefficients" "residuals" "effects" "rank"   
## [5] "fitted.values" "assign" "qr" "df.residual"   
## [9] "xlevels" "call" "terms" "model"

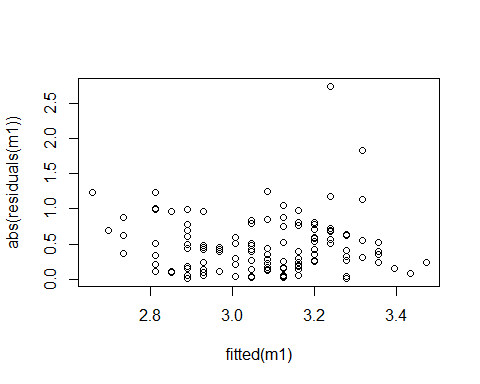
names(m1s)

## [1] "call" "terms" "residuals" "coefficients"   
## [5] "aliased" "sigma" "df" "r.squared"   
## [9] "adj.r.squared" "fstatistic" "cov.unscaled"

#Check the assumption of variance  
plot(fitted(m1),residuals(m1))  
abline(h=0)



#From the residuals plot, we see that the random noise are cluster above and  
#below and they are symetry.So they look constant.  
  
plot(fitted(m1),abs(residuals(m1)))



#Formal test of nonconstant variance  
  
#null-constant variance exists  
#alternavtive-Constant variance does not exist  
  
require(car)

## Loading required package: car

## Loading required package: carData

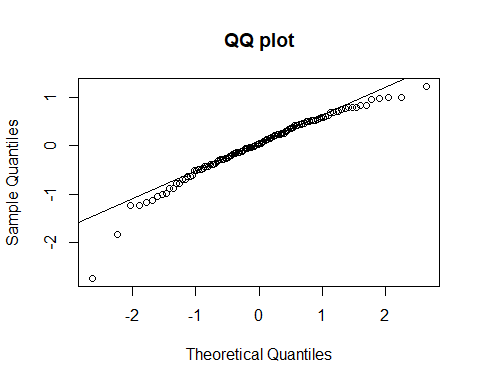
ncvTest(m1)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 0.6392758, Df = 1, p = 0.42397

## Here the p value of ncvtest is bigger than .05. So we fail to reject the null  
#and conclude that the variance of Y is same(constant) regardless the values of   
#predictors.  
  
#another test of nonconstant variance is-  
  
summary(lm(abs(residuals(m1))~fitted(m1)))

##   
## Call:  
## lm(formula = abs(residuals(m1)) ~ fitted(m1))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.46383 -0.29169 -0.05703 0.21015 2.26704   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 0.49554 0.64949 0.763 0.447  
## fitted(m1) -0.00696 0.21095 -0.033 0.974  
##   
## Residual standard error: 0.3996 on 118 degrees of freedom  
## Multiple R-squared: 9.226e-06, Adjusted R-squared: -0.008465   
## F-statistic: 0.001089 on 1 and 118 DF, p-value: 0.9737

##Here the slope is close to zero and the p value of slope is bigger than .05   
#so we fail ro reject the null that its actually 0 and conclude that the slope  
#is 0. so the assumption of constant variance is met.  
  
  
##Check the assumption of normality  
  
qqnorm(residuals(m1),main="QQ plot")  
qqline(residuals(m1))



#The qq plot looks normal except its head and tail part. The head and tail  
#has deviated little from the QQ line. We will perform a shapiro-wilk test  
#to check further.  
  
#Formal test of normality  
  
#null-normal  
#alternative-not normal  
  
shapiro.test(residuals(m1))

##   
## Shapiro-Wilk normality test  
##   
## data: residuals(m1)  
## W = 0.95249, p-value = 0.0003304

#From the output, the p-value < 0.05 so wereject the null and  
#conclude that the distribution of the data are not normally distributed.   
  
##Check the assumption of error correlation  
plot(residuals(m1))  
abline(h=0)  
#the residuals are cluster around 0.  
summary(lm(residuals(m1)[-1]~-1+residuals(m1)[-120]))#check

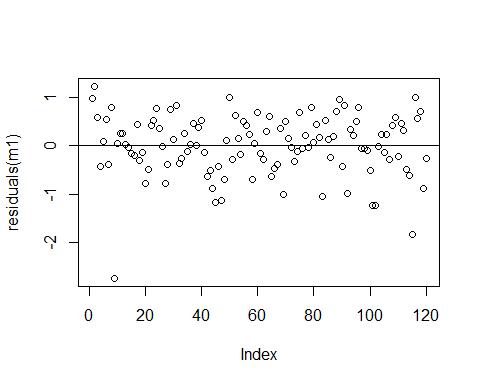
##   
## Call:  
## lm(formula = residuals(m1)[-1] ~ -1 + residuals(m1)[-120])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.79902 -0.33970 0.03432 0.45083 1.15591   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## residuals(m1)[-120] 0.07385 0.09092 0.812 0.418  
##   
## Residual standard error: 0.615 on 118 degrees of freedom  
## Multiple R-squared: 0.00556, Adjusted R-squared: -0.002867   
## F-statistic: 0.6598 on 1 and 118 DF, p-value: 0.4183

#Durbin-Watson test of autocorrelation  
## null-there is no auto-Correlation among residuals,they are independent  
#alter-there is auto-correlation among residuals  
library(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric



dwtest(m1)

##   
## Durbin-Watson test  
##   
## data: m1  
## DW = 1.8307, p-value = 0.1758  
## alternative hypothesis: true autocorrelation is greater than 0

#Here p value is greater than .05 and DW is close to 2 so we fail to reject  
#the null and conclude that there is no autocorrelation  
  
  
# So This model has met the constant variance assumption,the assumption of   
#uncorrelated errors but the normality has not been met.  
  
#If any of these assumptions is violated (i.e., if there are nonlinear   
#relationships between dependent and independent variables or the errors   
#exhibit correlation, heteroscedasticity, or non-normality), then the forecasts,   
#confidence intervals, and scientific insights yielded by a regression model   
#may be (at best) inefficient or (at worst) seriously biased or   
#misleading.   
  
  
#3. We are assuming that the regression model is appropriate..  
 m1=lm(Y~X)# fit model  
summary(m1)

##   
## Call:  
## lm(formula = Y ~ X)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.74004 -0.33827 0.04062 0.44064 1.22737   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.11405 0.32089 6.588 1.3e-09 \*\*\*  
## X 0.03883 0.01277 3.040 0.00292 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6231 on 118 degrees of freedom  
## Multiple R-squared: 0.07262, Adjusted R-squared: 0.06476   
## F-statistic: 9.24 on 1 and 118 DF, p-value: 0.002917

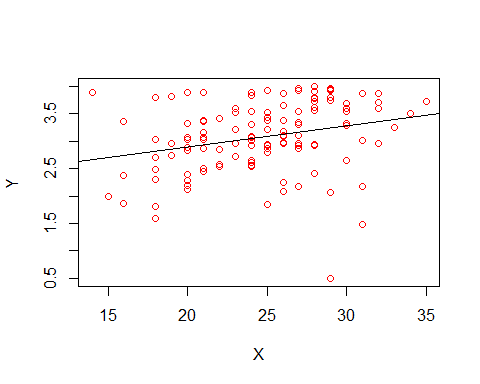
m1

##   
## Call:  
## lm(formula = Y ~ X)  
##   
## Coefficients:  
## (Intercept) X   
## 2.11405 0.03883

#Here the Beta\_0 is 2.11405 and Beta\_1 is 0.03883.  
#Estimated regression Equation is  
#Y=2.11405+0.03883X   
  
#Interpretation-  
#If act score is 0 then the GPA of freshmen will be 2.11(beta\_0)  
  
# If ACT score increases by 1 unit, The GPA of freshmen increases by  
#.0388(beta\_1) unit.  
  
  
#4.  
m1=lm(Y~X)# fit model  
summary(m1)

##   
## Call:  
## lm(formula = Y ~ X)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.74004 -0.33827 0.04062 0.44064 1.22737   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.11405 0.32089 6.588 1.3e-09 \*\*\*  
## X 0.03883 0.01277 3.040 0.00292 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6231 on 118 degrees of freedom  
## Multiple R-squared: 0.07262, Adjusted R-squared: 0.06476   
## F-statistic: 9.24 on 1 and 118 DF, p-value: 0.002917

with(GPA,plot(X,Y,col="red"))  
abline(m1)



#Yes the estimated regression function appears close to fit the data but   
#not perfect  
  
  
  
#5.Here R\_square is .0726. So 7.26 percentage of the variation in freshman GPA   
#is explained by ACT score.  
  
  
#6  
confint(m1)

## 2.5 % 97.5 %  
## (Intercept) 1.47859015 2.74950842  
## X 0.01353307 0.06412118

## 0.01353307 < Beta\_1 < 0.06412118  
  
#If repeated samples were taken and the 95% confidence interval   
#computed for each sample,95% of the intervals would contain the Beta\_1 and it  
#will be between 0.01353307 and 0.06412118  
  
#  
#If the confidence interval for beta\_1 contains 0, then  
#it can be concluded there is no significant evidence of a linear   
#relationship between predictor x and response y in the population.  
  
#7.null- Beta=0  
#alternative- Beta not equal 0  
  
t.test(X,Y, mu = 0, paired = FALSE,  
 conf.level = 0.99)

##   
## Welch Two Sample t-test  
##   
## data: X and Y  
## t = 52.493, df = 123.94, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 99 percent confidence interval:  
## 20.57193 22.72997  
## sample estimates:  
## mean of x mean of y   
## 24.72500 3.07405

#Here our test supports the alternative, which is true diffrence in means  
#is not equal to 0  
  
# The p value of the t test is 2.2e-16. Which is very low and lower than .01.  
# We reject the null and conclude that there is a significant evidence of a linear   
#relationship between predictor x and response y in the population.  
  
  
  
##Multiple candidate predictor  
  
#1.  
##install.packages("faraway")  
library("faraway")

## Warning: package 'faraway' was built under R version 3.5.3

##   
## Attaching package: 'faraway'

## The following objects are masked from 'package:car':  
##   
## logit, vif

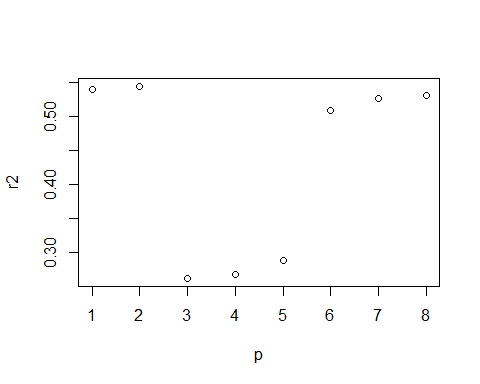
attach(prostate)  
str(prostate)

## 'data.frame': 97 obs. of 9 variables:  
## $ lcavol : num -0.58 -0.994 -0.511 -1.204 0.751 ...  
## $ lweight: num 2.77 3.32 2.69 3.28 3.43 ...  
## $ age : int 50 58 74 58 62 50 64 58 47 63 ...  
## $ lbph : num -1.39 -1.39 -1.39 -1.39 -1.39 ...  
## $ svi : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ lcp : num -1.39 -1.39 -1.39 -1.39 -1.39 ...  
## $ gleason: int 6 6 7 6 6 6 6 6 6 6 ...  
## $ pgg45 : int 0 0 20 0 0 0 0 0 0 0 ...  
## $ lpsa : num -0.431 -0.163 -0.163 -0.163 0.372 ...

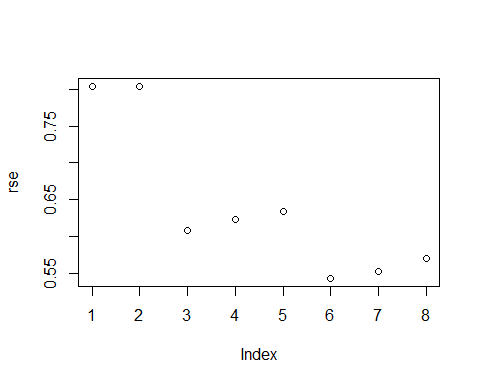
prostate

## lcavol lweight age lbph svi lcp gleason pgg45 lpsa  
## 1 -0.5798185 2.7695 50 -1.386294 0 -1.38629 6 0 -0.43078  
## 2 -0.9942523 3.3196 58 -1.386294 0 -1.38629 6 0 -0.16252  
## 3 -0.5108256 2.6912 74 -1.386294 0 -1.38629 7 20 -0.16252  
## 4 -1.2039728 3.2828 58 -1.386294 0 -1.38629 6 0 -0.16252  
## 5 0.7514161 3.4324 62 -1.386294 0 -1.38629 6 0 0.37156  
## 6 -1.0498221 3.2288 50 -1.386294 0 -1.38629 6 0 0.76547  
## 7 0.7371641 3.4735 64 0.615186 0 -1.38629 6 0 0.76547  
## 8 0.6931472 3.5395 58 1.536867 0 -1.38629 6 0 0.85442  
## 9 -0.7765288 3.5395 47 -1.386294 0 -1.38629 6 0 1.04732  
## 10 0.2231436 3.2445 63 -1.386294 0 -1.38629 6 0 1.04732  
## 11 0.2546422 3.6041 65 -1.386294 0 -1.38629 6 0 1.26695  
## 12 -1.3470736 3.5987 63 1.266948 0 -1.38629 6 0 1.26695  
## 13 1.6134299 3.0229 63 -1.386294 0 -0.59784 7 30 1.26695  
## 14 1.4770487 2.9982 67 -1.386294 0 -1.38629 7 5 1.34807  
## 15 1.2059708 3.4420 57 -1.386294 0 -0.43078 7 5 1.39872  
## 16 1.5411591 3.0611 66 -1.386294 0 -1.38629 6 0 1.44692  
## 17 -0.4155154 3.5160 70 1.244155 0 -0.59784 7 30 1.47018  
## 18 2.2884862 3.6494 66 -1.386294 0 0.37156 6 0 1.49290  
## 19 -0.5621189 3.2677 41 -1.386294 0 -1.38629 6 0 1.55814  
## 20 0.1823216 3.8254 70 1.658228 0 -1.38629 6 0 1.59939  
## 21 1.1474025 3.4194 59 -1.386294 0 -1.38629 6 0 1.63900  
## 22 2.0592388 3.5010 60 1.474763 0 1.34807 7 20 1.65823  
## 23 -0.5447272 3.3759 59 -0.798508 0 -1.38629 6 0 1.69562  
## 24 1.7817091 3.4516 63 0.438255 0 1.17865 7 60 1.71380  
## 25 0.3852624 3.6674 69 1.599388 0 -1.38629 6 0 1.73166  
## 26 1.4469190 3.1246 68 0.300105 0 -1.38629 6 0 1.76644  
## 27 0.5128236 3.7197 65 -1.386294 0 -0.79851 7 70 1.80006  
## 28 -0.4004776 3.8660 67 1.816452 0 -1.38629 7 20 1.81645  
## 29 1.0402767 3.1290 67 0.223144 0 0.04879 7 80 1.84845  
## 30 2.4096442 3.3759 65 -1.386294 0 1.61939 6 0 1.89462  
## 31 0.2851789 4.0902 65 1.962908 0 -0.79851 6 0 1.92425  
## 32 0.1823216 6.1076 65 1.704748 0 -1.38629 6 0 2.00821  
## 33 1.2753628 3.0374 71 1.266948 0 -1.38629 6 0 2.00821  
## 34 0.0099503 3.2677 54 -1.386294 0 -1.38629 6 0 2.02155  
## 35 -0.0100503 3.2169 63 -1.386294 0 -0.79851 6 0 2.04769  
## 36 1.3083328 4.1198 64 2.171337 0 -1.38629 7 5 2.08567  
## 37 1.4231083 3.6571 73 -0.579818 0 1.65823 8 15 2.15756  
## 38 0.4574248 2.3749 64 -1.386294 0 -1.38629 7 15 2.19165  
## 39 2.6609586 4.0851 68 1.373716 1 1.83258 7 35 2.21375  
## 40 0.7975072 3.0131 56 0.936093 0 -0.16252 7 5 2.27727  
## 41 0.6205765 3.1420 60 -1.386294 0 -1.38629 9 80 2.29757  
## 42 1.4422020 3.6826 68 -1.386294 0 -1.38629 7 10 2.30757  
## 43 0.5822156 3.8660 62 1.713798 0 -0.43078 6 0 2.32728  
## 44 1.7715568 3.8969 61 -1.386294 0 0.81093 7 6 2.37491  
## 45 1.4861397 3.4095 66 1.749200 0 -0.43078 7 20 2.52172  
## 46 1.6639261 3.3928 61 0.615186 0 -1.38629 7 15 2.55334  
## 47 2.7278528 3.9954 79 1.879465 1 2.65676 9 100 2.56879  
## 48 1.1631508 4.0351 68 1.713798 0 -0.43078 7 40 2.56879  
## 49 1.7457155 3.4980 43 -1.386294 0 -1.38629 6 0 2.59152  
## 50 1.2208299 3.5681 70 1.373716 0 -0.79851 6 0 2.59152  
## 51 1.0919233 3.9936 68 -1.386294 0 -1.38629 7 50 2.65676  
## 52 1.6601310 4.2348 64 2.073172 0 -1.38629 6 0 2.67759  
## 53 0.5128236 3.6336 64 1.492904 0 0.04879 7 70 2.68444  
## 54 2.1270405 4.1215 68 1.766442 0 1.44692 7 40 2.69124  
## 55 3.1535904 3.5160 59 -1.386294 0 -1.38629 7 5 2.70471  
## 56 1.2669476 4.2801 66 2.122262 0 -1.38629 7 15 2.71800  
## 57 0.9745596 2.8651 47 -1.386294 0 0.50078 7 4 2.78809  
## 58 0.4637340 3.7647 49 1.423108 0 -1.38629 6 0 2.79423  
## 59 0.5423243 4.1782 70 0.438255 0 -1.38629 7 20 2.80639  
## 60 1.0612565 3.8512 61 1.294727 0 -1.38629 7 40 2.81241  
## 61 0.4574248 4.5245 73 2.326302 0 -1.38629 6 0 2.84200  
## 62 1.9974177 3.7197 63 1.619388 1 1.90954 7 40 2.85359  
## 63 2.7757088 3.5249 72 -1.386294 0 1.55814 9 95 2.85359  
## 64 2.0347056 3.9170 66 2.008214 1 2.11021 7 60 2.88200  
## 65 2.0731719 3.6230 64 -1.386294 0 -1.38629 6 0 2.88200  
## 66 1.4586150 3.8362 61 1.321756 0 -0.43078 7 20 2.88759  
## 67 2.0228712 3.8785 68 1.783391 0 1.32176 7 70 2.92047  
## 68 2.1983351 4.0509 72 2.307573 0 -0.43078 7 10 2.96269  
## 69 -0.4462871 4.4085 69 -1.386294 0 -1.38629 6 0 2.96269  
## 70 1.1939225 4.7804 72 2.326302 0 -0.79851 7 5 2.97298  
## 71 1.8640801 3.5932 60 -1.386294 1 1.32176 7 60 3.01308  
## 72 1.1600209 3.3411 77 1.749200 0 -1.38629 7 25 3.03735  
## 73 1.2149127 3.8254 69 -1.386294 1 0.22314 7 20 3.05636  
## 74 1.8389611 3.2367 60 0.438255 1 1.17865 9 90 3.07501  
## 75 2.9992262 3.8491 69 -1.386294 1 1.90954 7 20 3.27526  
## 76 3.1411305 3.2638 68 -0.051293 1 2.42037 7 50 3.33755  
## 77 2.0108950 4.4338 72 2.122262 0 0.50078 7 60 3.39283  
## 78 2.5376572 4.3548 78 2.326302 0 -1.38629 7 10 3.43560  
## 79 2.6483002 3.5821 69 -1.386294 1 2.58400 7 70 3.45789  
## 80 2.7794402 3.8232 63 -1.386294 0 0.37156 7 50 3.51304  
## 81 1.4678743 3.0704 66 0.559616 0 0.22314 7 40 3.51601  
## 82 2.5136561 3.4735 57 0.438255 0 2.32728 7 60 3.53076  
## 83 2.6130067 3.8888 77 -0.527633 1 0.55962 7 30 3.56530  
## 84 2.6775910 3.8384 65 1.115142 0 1.74920 9 70 3.57094  
## 85 1.5623463 3.7099 60 1.695616 0 0.81093 7 30 3.58768  
## 86 3.3028493 3.5190 64 -1.386294 1 2.32728 7 60 3.63099  
## 87 2.0241931 3.7317 58 1.638997 0 -1.38629 6 0 3.68009  
## 88 1.7316555 3.3690 62 -1.386294 1 0.30010 7 30 3.71235  
## 89 2.8075938 4.7181 65 -1.386294 1 2.46385 7 60 3.98434  
## 90 1.5623463 3.6951 76 0.936093 1 0.81093 7 75 3.99360  
## 91 3.2464910 4.1018 68 -1.386294 0 -1.38629 6 0 4.02981  
## 92 2.5329028 3.6776 61 1.348073 1 -1.38629 7 15 4.12955  
## 93 2.8302678 3.8764 68 -1.386294 1 1.32176 7 60 4.38515  
## 94 3.8210036 3.8969 44 -1.386294 1 2.16905 7 40 4.68444  
## 95 2.9074474 3.3962 52 -1.386294 1 2.46385 7 10 5.14312  
## 96 2.8825636 3.7739 68 1.558145 1 1.55814 7 80 5.47751  
## 97 3.4719665 3.9750 68 0.438255 1 2.90417 7 20 5.58293

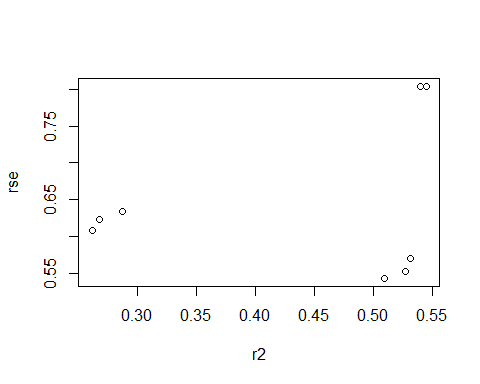
prostate$svi[prostate$svi==0]<-NA # [transforming 0 to NA]  
prostate$ppg45[prostate$pgg45==0]<-NA #[transforming 0 to NA]  
  
#2  
n1 <- lm(lcavol~lpsa, data = prostate)  
sigma1 <- summary(n1)$sigma   
R2.1 <- summary(n1)$r.squared  
  
n2 <- lm(lcavol~lpsa + lweight, data = prostate)#addind lweight   
sigma2 <- summary(n2)$sigma   
R2.2 <- summary(n2)$r.squared  
  
n3 <- lm(lcavol~lpsa + lweight+svi, data = prostate)#adding svi   
sigma3 <- summary(n3)$sigma   
R2.3 <- summary(n3)$r.squared  
  
  
n4 <- lm(lcavol~lpsa + lweight+svi+lbph, data = prostate) #adding lbph  
sigma4 <- summary(n4)$sigma   
R2.4 <- summary(n4)$r.squared  
  
  
n5 <- lm(lcavol~lpsa + lweight+svi+lbph+age, data = prostate)#adding age  
sigma5 <- summary(n5)$sigma   
R2.5 <- summary(n5)$r.squared  
  
  
n6 <- lm(lcavol~lpsa + lweight+svi+lbph+age+lcp, data = prostate)#adding lcp   
sigma6 <- summary(n6)$sigma   
R2.6 <- summary(n6)$r.squared  
  
  
n7 <- lm(lcavol~lpsa + lweight+svi+lbph+age+lcp+pgg45, data = prostate)  
#adding pgg45  
sigma7 <- summary(n7)$sigma   
R2.7 <- summary(n7)$r.squared  
  
n8 <- lm(lcavol~lpsa + lweight+svi+lbph+age+lcp+pgg45+gleason, data = prostate)  
#adding gleason  
sigma8 <- summary(n8)$sigma   
R2.8 <- summary(n8)$r.squared  
  
#4  
r2=c(R2.1,R2.2,R2.3,R2.4,R2.5,R2.6,R2.7,R2.8)  
p=c(1,2,3,4,5,6,7,8)  
plot(p,r2)



#For model 1,2,3,6,7 & 8 the R sqaure is higher than model 3,4 & 5  
  
#5  
rse<-c(sigma1,sigma2,sigma3,sigma4,sigma5,sigma6,sigma7,sigma8)  
plot(rse)



#For model 1 and model 2 sigma square is very high.The more predictor variable  
# we add the more it gets lower. Sigma square is the lowest for model 6.  
plot(r2,rse)



#ANOVA  
require(faraway)   
data(coagulation)  
?coagulation

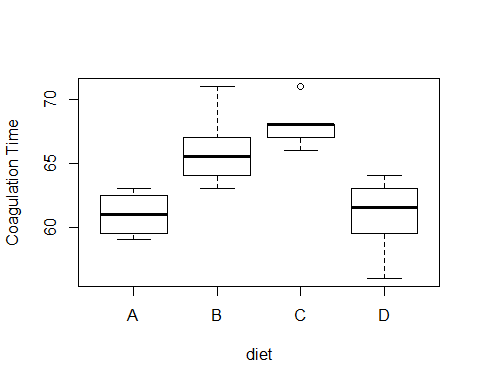
## starting httpd help server ...

## done

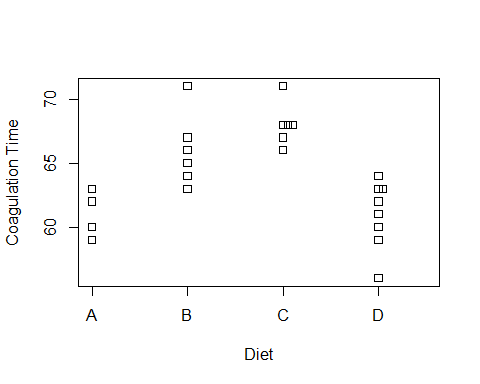
str(coagulation)

## 'data.frame': 24 obs. of 2 variables:  
## $ coag: num 62 60 63 59 63 67 71 64 65 66 ...  
## $ diet: Factor w/ 4 levels "A","B","C","D": 1 1 1 1 2 2 2 2 2 2 ...

plot(coag~diet, coagulation, ylab="Coagulation Time")



with(coagulation, stripchart(coag~diet,vertical=TRUE, method = "stack",  
 xlab="Diet",ylab="Coagulation Time"))



###We can see some skewness in diet C from the box plot and it is right skewed.  
#an outlier is defined as a data point that is   
#located outside the fences of the boxplot.We can see a outlier in diet C.  
#Nonconstant variance not present because of skewness and outliers  
  
#2  
g <- lm(coag~diet, coagulation)  
summary(g)

##   
## Call:  
## lm(formula = coag ~ diet, data = coagulation)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.00 -1.25 0.00 1.25 5.00   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.100e+01 1.183e+00 51.554 < 2e-16 \*\*\*  
## dietB 5.000e+00 1.528e+00 3.273 0.003803 \*\*   
## dietC 7.000e+00 1.528e+00 4.583 0.000181 \*\*\*  
## dietD 2.991e-15 1.449e+00 0.000 1.000000   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.366 on 20 degrees of freedom  
## Multiple R-squared: 0.6706, Adjusted R-squared: 0.6212   
## F-statistic: 13.57 on 3 and 20 DF, p-value: 4.658e-05

model.matrix(g)

## (Intercept) dietB dietC dietD  
## 1 1 0 0 0  
## 2 1 0 0 0  
## 3 1 0 0 0  
## 4 1 0 0 0  
## 5 1 1 0 0  
## 6 1 1 0 0  
## 7 1 1 0 0  
## 8 1 1 0 0  
## 9 1 1 0 0  
## 10 1 1 0 0  
## 11 1 0 1 0  
## 12 1 0 1 0  
## 13 1 0 1 0  
## 14 1 0 1 0  
## 15 1 0 1 0  
## 16 1 0 1 0  
## 17 1 0 0 1  
## 18 1 0 0 1  
## 19 1 0 0 1  
## 20 1 0 0 1  
## 21 1 0 0 1  
## 22 1 0 0 1  
## 23 1 0 0 1  
## 24 1 0 0 1  
## attr(,"assign")  
## [1] 0 1 1 1  
## attr(,"contrasts")  
## attr(,"contrasts")$diet  
## [1] "contr.treatment"

##intercept,where Beta\_zero=mean is the expected response for the reference   
#diet.5.000e+00 is diffrence in the expected response for diet B compared to   
#the refernce diet.  
#7.000e+00 is diffrence in the expected response for diet C  
#compared to the reference diet.  
#2.991e-15 is diffrence in the expected response  
#for diet D compared to the reference diet  
  
#3  
  
gi <- lm(coag~diet -1, coagulation)  
model.matrix(gi)

## dietA dietB dietC dietD  
## 1 1 0 0 0  
## 2 1 0 0 0  
## 3 1 0 0 0  
## 4 1 0 0 0  
## 5 0 1 0 0  
## 6 0 1 0 0  
## 7 0 1 0 0  
## 8 0 1 0 0  
## 9 0 1 0 0  
## 10 0 1 0 0  
## 11 0 0 1 0  
## 12 0 0 1 0  
## 13 0 0 1 0  
## 14 0 0 1 0  
## 15 0 0 1 0  
## 16 0 0 1 0  
## 17 0 0 0 1  
## 18 0 0 0 1  
## 19 0 0 0 1  
## 20 0 0 0 1  
## 21 0 0 0 1  
## 22 0 0 0 1  
## 23 0 0 0 1  
## 24 0 0 0 1  
## attr(,"assign")  
## [1] 1 1 1 1  
## attr(,"contrasts")  
## attr(,"contrasts")$diet  
## [1] "contr.treatment"

summary(gi)

##   
## Call:  
## lm(formula = coag ~ diet - 1, data = coagulation)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.00 -1.25 0.00 1.25 5.00   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## dietA 61.0000 1.1832 51.55 <2e-16 \*\*\*  
## dietB 66.0000 0.9661 68.32 <2e-16 \*\*\*  
## dietC 68.0000 0.9661 70.39 <2e-16 \*\*\*  
## dietD 61.0000 0.8367 72.91 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.366 on 20 degrees of freedom  
## Multiple R-squared: 0.9989, Adjusted R-squared: 0.9986   
## F-statistic: 4399 on 4 and 20 DF, p-value: < 2.2e-16

#interpretation  
#when x1 increases by 1 unit coagulation time y increases by 61 and x2,x3 and x4   
#remain uncahnged.  
#when x2 increases by 1 unit coagulation time y increases by 66 and x1,  
#x3 and x4 remain uncahnged.  
#when x3 increases by 1 unit coagulation time y increases by 68 and x1,x2   
#and x4 remain uncahnged.  
#when x3 increases by 1 unit coagulation time y increases by 61 and x1,x2   
#and x4 remain uncahnged.  
  
#for this problem i prefer to fit a model without intercept. Because in this   
#we can compare each diet with restall of the diets.  
  
#4.  
#H0 : µ1 = µ2 = µ3 = µ4, H1 : µi 6= µj for some i 6= j.   
  
gnull <- lm(coag~1, coagulation)  
model.matrix(gnull)

## (Intercept)  
## 1 1  
## 2 1  
## 3 1  
## 4 1  
## 5 1  
## 6 1  
## 7 1  
## 8 1  
## 9 1  
## 10 1  
## 11 1  
## 12 1  
## 13 1  
## 14 1  
## 15 1  
## 16 1  
## 17 1  
## 18 1  
## 19 1  
## 20 1  
## 21 1  
## 22 1  
## 23 1  
## 24 1  
## attr(,"assign")  
## [1] 0

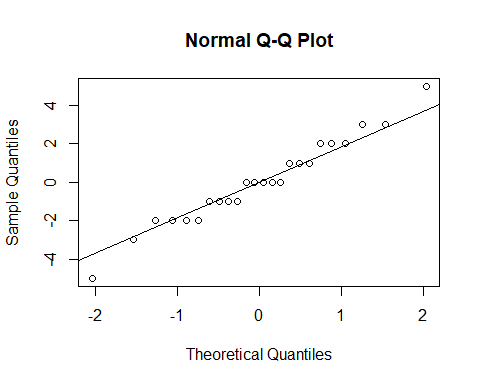
anova(gnull,gi)

## Analysis of Variance Table  
##   
## Model 1: coag ~ 1  
## Model 2: coag ~ diet - 1  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 23 340   
## 2 20 112 3 228 13.571 4.658e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

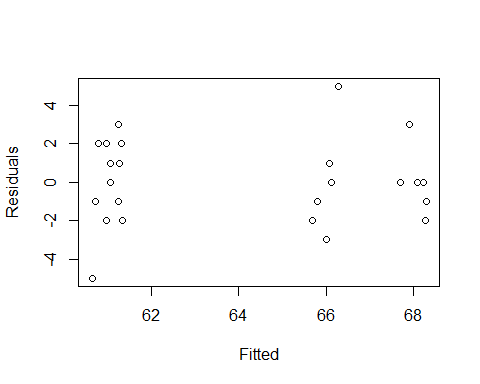
#The F statistics and p value for this test are 13.571 and 4.658e-05.  
#Since P value is lower than .05, so we reject the null and conclude that  
#there is a diffrence in at least one of the group means.  
  
#5.  
options(contrasts=c("contr.sum","contr.poly"))  
gs <- lm(coag~diet,coagulation)  
summary(gs)

##   
## Call:  
## lm(formula = coag ~ diet, data = coagulation)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.00 -1.25 0.00 1.25 5.00   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 64.0000 0.4979 128.537 < 2e-16 \*\*\*  
## diet1 -3.0000 0.9736 -3.081 0.005889 \*\*   
## diet2 2.0000 0.8453 2.366 0.028195 \*   
## diet3 4.0000 0.8453 4.732 0.000128 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.366 on 20 degrees of freedom  
## Multiple R-squared: 0.6706, Adjusted R-squared: 0.6212   
## F-statistic: 13.57 on 3 and 20 DF, p-value: 4.658e-05

#Interpretation:The overall average response accross all four diets is  
#64.The deviation from the overall mean for diet 1 is -3, so the response for  
# Diet 1 is 61.  
#The deviation from the overall mean for diet 2 is 2,so the mean response   
#for diet 2 is 66  
#The deviation from the overall mean for diet 3 is 4,so the mean response for   
#diet 3 is 68  
#The deviation from the overall mean for diet 4 is -3,so the mean response for   
#diet 3 is 68  
  
#6.  
qqnorm(residuals(g))  
qqline(residuals(g))



plot(jitter(fitted(g)),residuals(g), xlab = "Fitted", ylab = "Residuals")



##the qqnorm plot and qqline of residuals look normal but we will confirm it  
#by shapiro wilk test.  
#null-normal  
#alternative-not normal  
shapiro.test(residuals(g))

##   
## Shapiro-Wilk normality test  
##   
## data: residuals(g)  
## W = 0.97831, p-value = 0.8629

#conclusion-Since the p value of shapiro wilk is higher than .05 so we fail to  
#reject the null and conclude that normality of residuals exist.  
  
#7  
require(car)  
#null-all variances are equal  
#alternative-variances are not equal  
leveneTest(coag~diet, coagulation)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 3 0.6492 0.5926  
## 20

# The p value of levenetest is higher that .05. So we fail to reject the null  
#and conclude that the assumption of equal variance is met.