Regression 2

Mehrab Atighi

5/3/2021

here we can see all of the regression 2 codes with exmaple and exercise

First

solve 5 bottom questions:

1)

a) According to the Bottem data make main model woth 18 points and get summary function?

```
#Exercise one:
rm(list=ls())
#intruducing Data:
x<-c(1.5,1.7,2,2.2,2.5,2.5,2.7,2.9,3,3.5,3.8,4.2,4.3,4.6,4,5.1,5.2,5.5)
y<-c(1,2.5,3.5,3,3.1,3.6,2.2,3.9,4,4,4.2,4.1,4.8,1.2,5.1,5.1,4.8,5.3)

A=3.4;B=9.5;C=9.5
a=8;b=8;c=2.5
#Make Models and get needed information:
#Make First Regression Model(with main Data)
#a)
fit1<-lm(y~x)
#Get needed information from this Model
summary(fit1)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1998 -0.0290  0.3070  0.5749  1.0834
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.4616     0.7204   2.029  0.05945 .
## x             0.6387     0.1996   3.200  0.00558 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.025 on 16 degrees of freedom
## Multiple R-squared:  0.3902, Adjusted R-squared:  0.3521
## F-statistic: 10.24 on 1 and 16 DF,  p-value: 0.005578
```

now we can see that or p_value for x is lower than 0.05 and the betha0 = 1.4616,

betha1 = 0.6387, the Residual standard error = 1.025,

R-squared = 0.3902,t(betha1) = 3.200

b) According to the Bottem data make main model woth 18 points+A&a and get summary function?

```
#b)
#make Second regssion Model(With A point & main Data)
x[19]= A ; y[19]= a
fit2<-lm(y~x)
#Get needed information from this Model
summary(fit2)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4296 -0.2435  0.0813  0.3591  4.1368
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.6914     1.0036   1.685   0.1102
## x             0.6387     0.2789   2.290   0.0351 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.432 on 17 degrees of freedom
## Multiple R-squared:  0.2358, Adjusted R-squared:  0.1908
## F-statistic: 5.245 on 1 and 17 DF,  p-value: 0.03506
```

now we can see that or p_value for x is lower than 0.05 and the betha0 = 1.6914,

betha1 = 0.6387, the Residual standard error = 1.432,

R-squared = 0.2358,t(betha1) = 2.290

c) According to the Bottem data make main model woth 18 points+B&b and get summary function?

```
#c)
#make third regssion Model(With B point & main Data)
x[19]= B ; y[19]= b
fit3<-lm(y~x)
```

```
#Get needed information from this Model
summary(fit3)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2633 -0.0281  0.2221  0.5561  1.0464
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.3223     0.5251   2.518   0.0221 *
## x             0.6828     0.1270   5.375 5.03e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9973 on 17 degrees of freedom
## Multiple R-squared:  0.6296, Adjusted R-squared:  0.6078
## F-statistic:  28.9 on 1 and 17 DF,  p-value: 5.034e-05
```

now we can see that or p_value for x is lower than 0.05 and the betha0 = 1.3223,

betha1 = 0.6828, the Residual standard error = 0.9973,

R-squared = 0.6296,t(betha1)= 5.375

d) According to the Bottem data make main model woth 18 points+C&c and get summary function?

```
#d)
#make forth regssion Model(With C point & main Data)
x[19]= C ; y[19]= c
fit4<-lm(y~x)
#Get needed information from this Model
summary(fit4)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5206 -0.5277  0.4463  0.7961  1.4797
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.9518     0.6646   4.442 0.000358 ***
## x             0.1671     0.1608   1.040 0.313055
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.262 on 17 degrees of freedom
## Multiple R-squared:  0.05978,    Adjusted R-squared:  0.004475
## F-statistic: 1.081 on 1 and 17 DF,  p-value: 0.3131
```

now we can see that or p_value for x is not lower than 0.05 and the betha0 = 2.9518,

betha1 = 0.1671, the Residual standard error = 1.262,

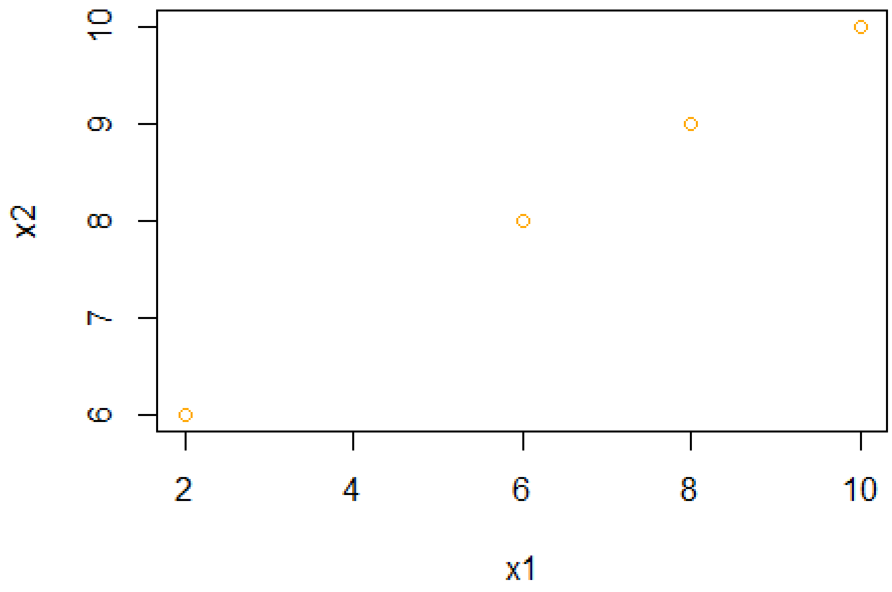
R-squared = 0.05978,t(betha1)=1.040

End.

---

2) show that the bottem datas variables (x1,x2) have linear relation.

```
#Exercise Two:
x1<-c(2,8,6,10)
x2<-c(6,9,8,10)
#produce variables plot:
par(mfrow=c(1,1))
plot(x1,x2,col="orange")
```



We can see a possetive relation between X1 & X2, so we can say when they have a linear relation. it will be possible to make a mistake make a regression model. End.

---

3) This question involves the use of simple linear regression on the Auto data set.

a) Use the lm() function to perform a simple linear regression with mpg as the response and horsepower as the predictor. Use the summary() function to print the results. Comment on the output. For example:

```
#Exercise Third(eigth of Book):
#library Data:
library("ISLR")
```

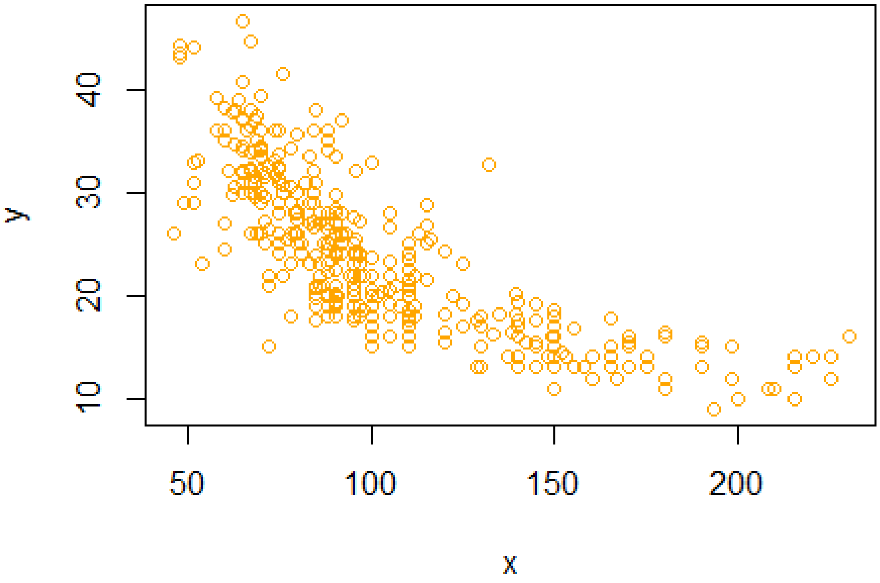
```
## Warning: package 'ISLR' was built under R version 4.0.4
```

```
y<-Auto$mpg
x<-Auto$horsepower
#a)
fit1<-lm(y~x)
summary(fit1)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5710  -3.2592  -0.3435   2.7630  16.9240
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.935861   0.717499   55.66   <2e-16 ***
## x           -0.157845   0.006446  -24.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.906 on 390 degrees of freedom
## Multiple R-squared:  0.6059, Adjusted R-squared:  0.6049
## F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

i. Is there a relationship between the predictor and the response?

```
#i)
plot(x, y, col = "orange", type = "p")
```



We can see a nagative relation between X & Y its fixed for x > 150 .

ii. How strong is the relationship between the predictor and the response?

```
#ii
summary(fit1)[8]

## $r.squared
## [1] 0.6059483

anova(fit1)[5]

##               Pr(>F)    
## x         < 2.2e-16 ***
## Residuals              
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

the r.squared is equal to 0.6059483.

the p value of the Anova of our model is very lower than 0.05.

So we can say that this linear regression model is significant and the r. squared show that its not very strong relation!.

iii. Is the relationship between the predictor and the response positive or negative?

```
#iii)
cor(x,y)
```

```
## [1] -0.7784268
```

the correlation value is equal to -0.7784268 , so we have a nagative relation between X & Y

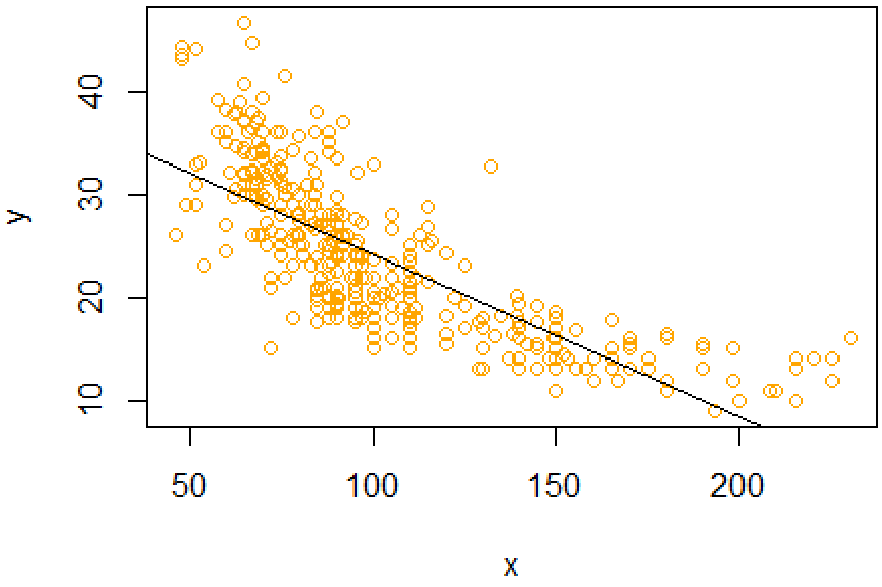
iv. What is the predicted mpg associated with a horsepower of 98? What are the associated 95 % confidence and prediction intervals?

```
#iv)
#make a f(x) function and calculate f(98):
f<-function(X){
  f=as.numeric(fit1$coefficients[1])+ as.numeric((X*fit1$coefficients[2]))
  return(f)
}

f(98)

## [1] 24.46708

#predictioin confidence interval 95% for f(98) predict:
predict(fit1,newdata = as.data.frame(x<-c(98)),interval = "confidence")

##        fit      lwr      upr
## 1 24.46708 23.97308 24.96108
```

clearly we can see the lower and upper limit of our confidence interval and the prediction value.

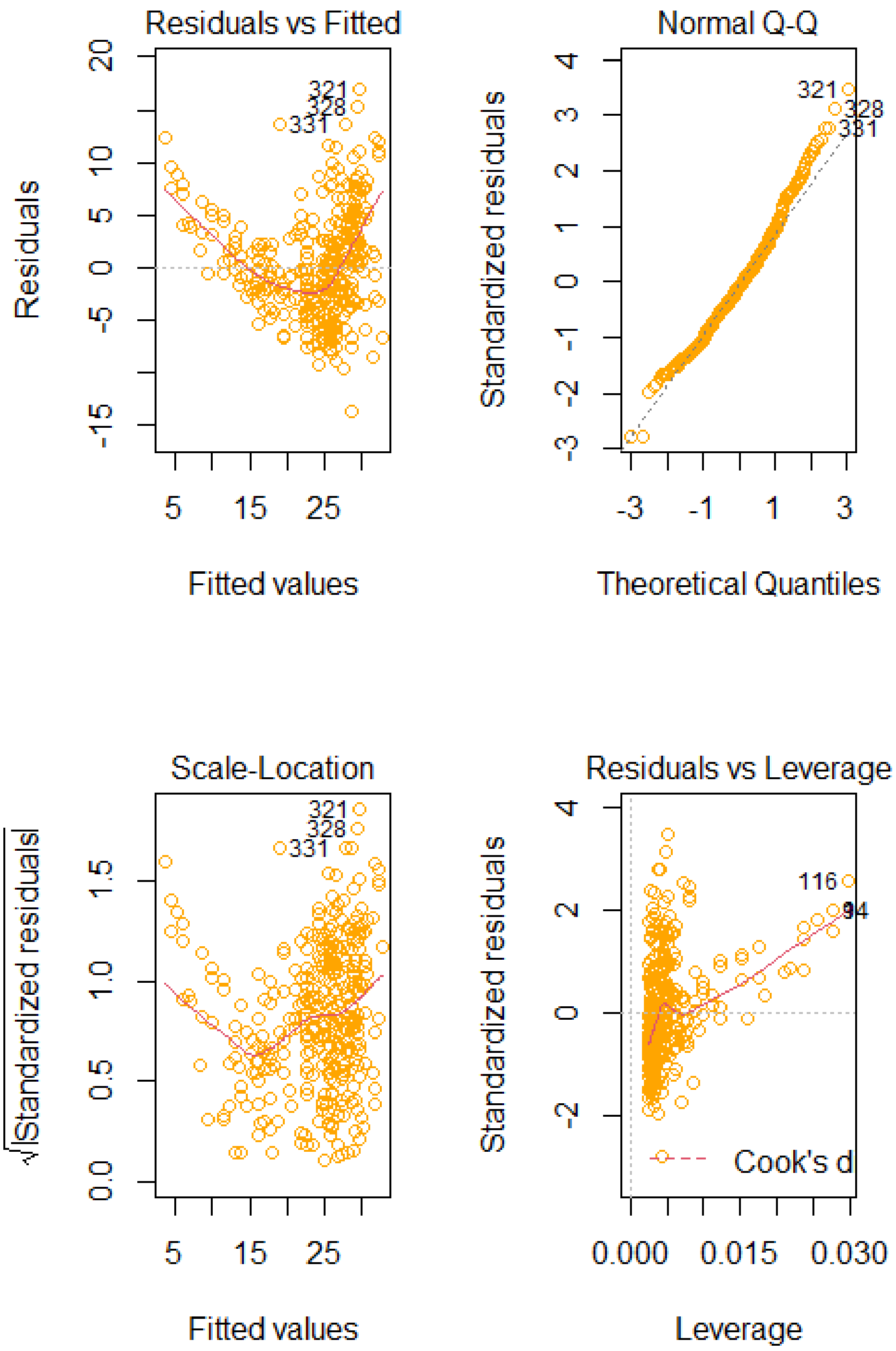
b) Plot the response and the predictor. Use the abline() function to display the least squares regression line.

```
#b)
x<-Auto$horsepower
y<-Auto$mpg
plot(x, y, col = "orange", type = "p")
abline(fit1,col="black")
```



c) Use the plot() function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

```
#c)
par(mfrow= c(1,2))
plot(fit1,col="orange")
```



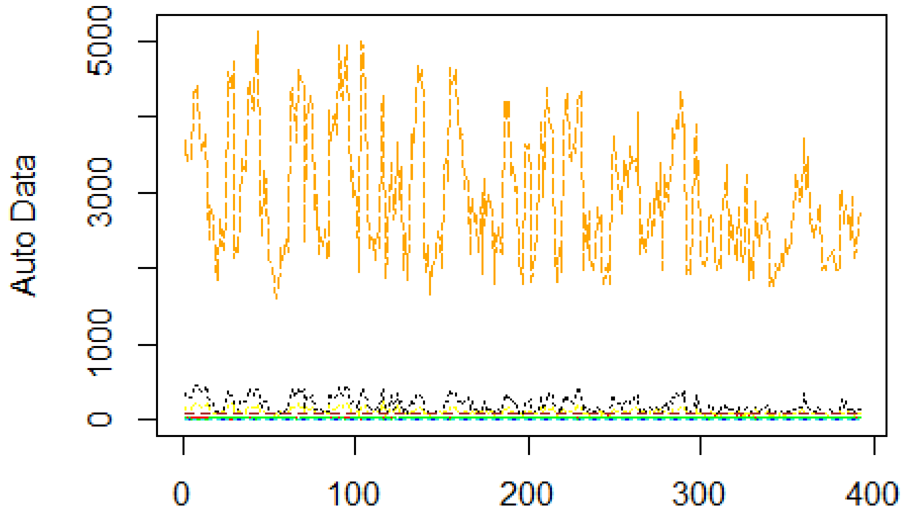
i think that we have three out lievs points and there is no high leverge points. and the QQplots show the normality distrubtion.

End.

4) This question involves the use of multiple linear regression on the Auto data set.

a) Produce a scatterplot matrix which includes all of the variables in the data set.

```
#Exercise forth(ninth of Book):
#a):
matplot(Auto[1:8],col =
c("Red","Blue","Black","yellow","orange","Green","Brown",85),ylab = "Auto
Data",type = "l")
```



b) Compute the matrix of correlations between the variables using the function cor(). You will need to exclude the name variable, cor() which is qualitative.

```
#b)
(as.matrix(cor(Auto[1:8])))

##                     mpg  cylinders displacement horsepower     weight
## mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
## cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
## displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
## horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
## weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
## acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
## year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199
## origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054
##              acceleration       year     origin
## mpg             0.4233285  0.5805410  0.5652088
## cylinders      -0.5046834 -0.3456474 -0.5689316
## displacement   -0.5438005 -0.3698552 -0.6145351
## horsepower     -0.6891955 -0.4163615 -0.4551715
## weight         -0.4168392 -0.3091199 -0.5850054
## acceleration    1.0000000  0.2903161  0.2127458
## year            0.2903161  1.0000000  0.1815277
## origin          0.2127458  0.1815277  1.0000000
```

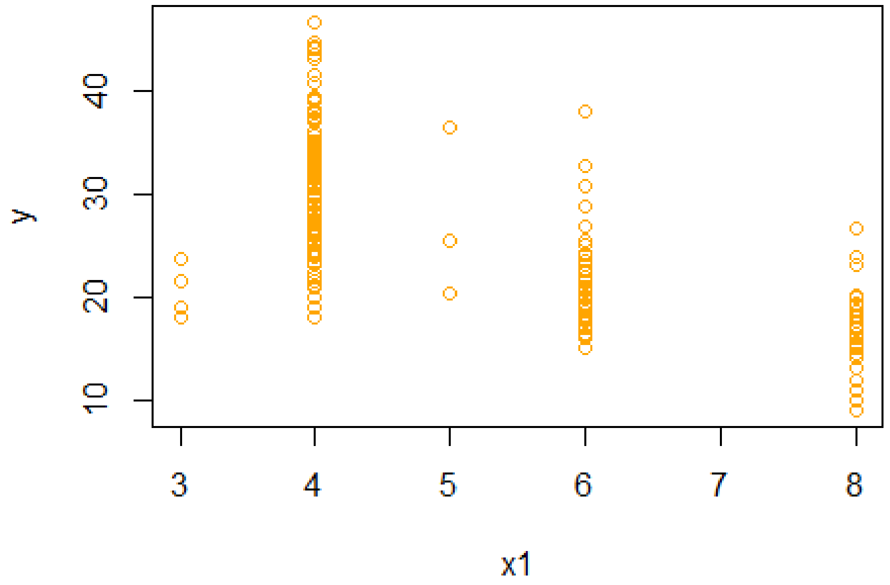
c) Use the lm() function to perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Use the summary() function to print the results. Comment on the output. For instance:

```
#c)
y<-Auto$mpg
x1<-Auto$cylinders
x2<-Auto$displacement
x3<-Auto$horsepower
x4<-Auto$weight
x5<-Auto$acceleration
x6<-Auto$year
x7<-Auto$origin
fit1<-lm(y~x1+x2+x3+x4+x5+x6+x7)
summary(fit1)
```

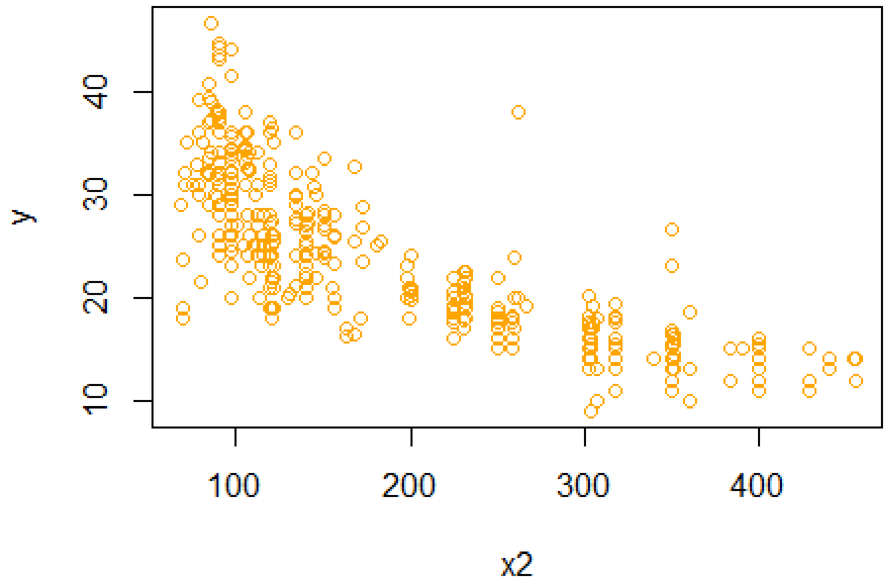
```
## 
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -9.5903 -2.1565 -0.1169  1.8690 13.0604 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -17.218435   4.644294  -3.707  0.00024 ***
## x1           -0.493376   0.323282  -1.526  0.12780    
## x2            0.019896   0.007515   2.647  0.00844 ** 
## x3           -0.016951   0.013787  -1.230  0.21963    
## x4           -0.006474   0.000652  -9.929  < 2e-16 ***
## x5            0.080576   0.098845   0.815  0.41548    
## x6            0.750773   0.050973  14.729  < 2e-16 ***
## x7            1.426141   0.278136   5.127 4.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.328 on 384 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182 
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

i. Is there a relationship between the predictors and the response?

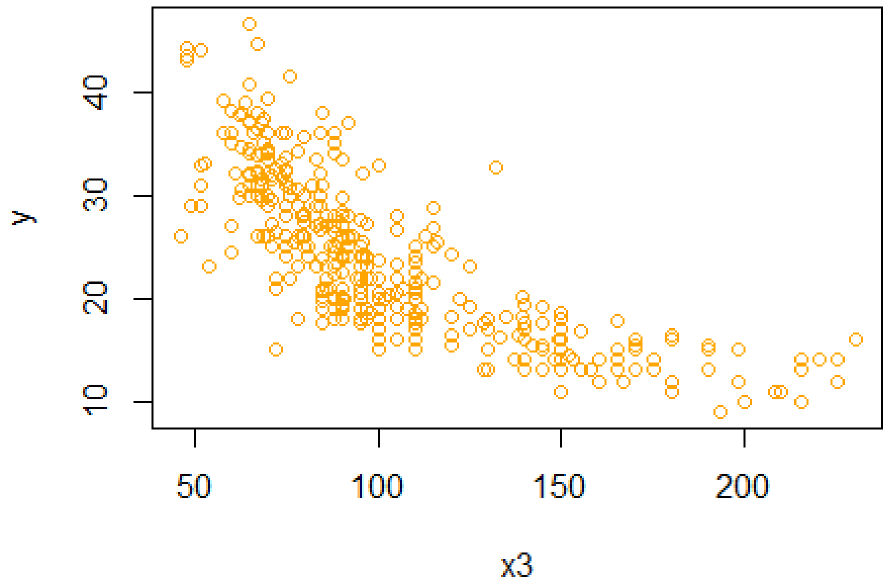
```
#i)
plot(x1,y,col="orange")
```



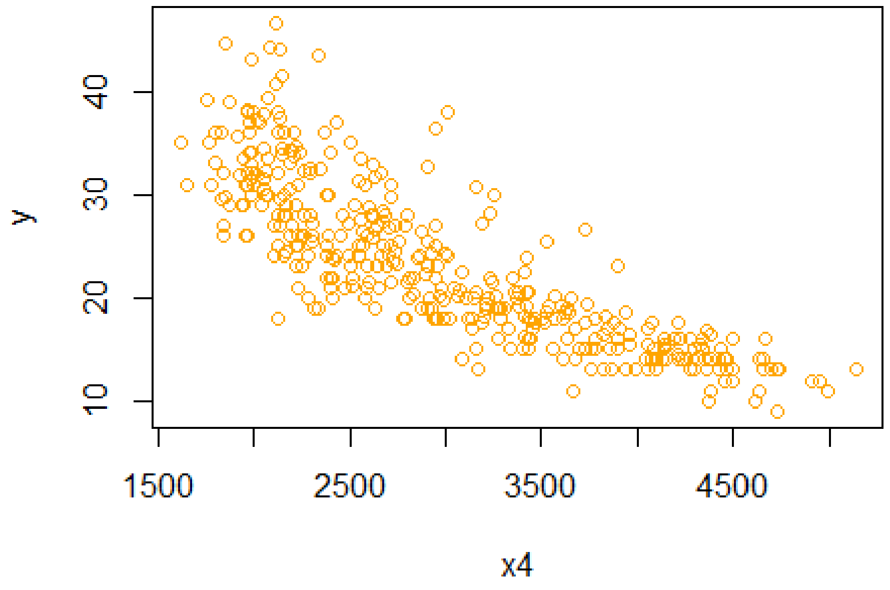
```
plot(x2,y,col="orange")
```



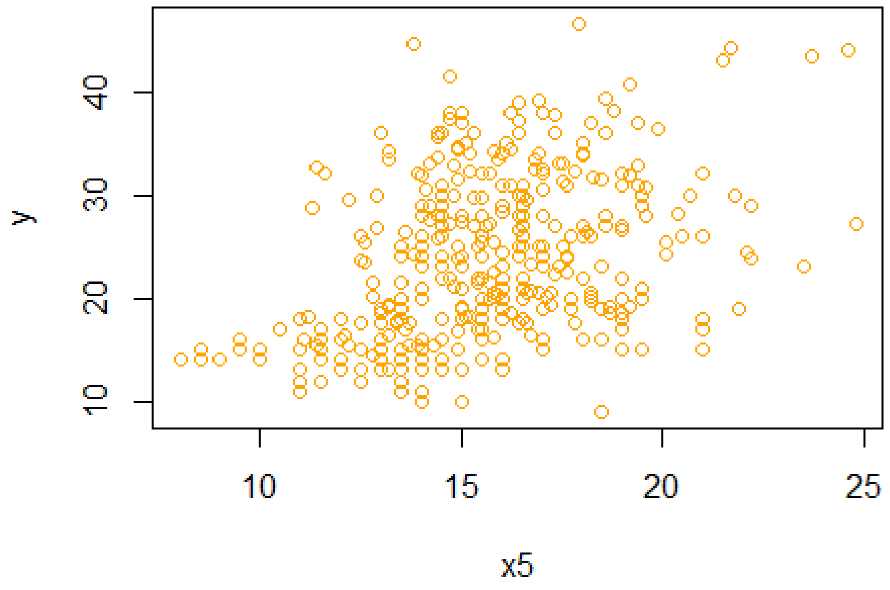
```
plot(x3,y,col="orange")
```



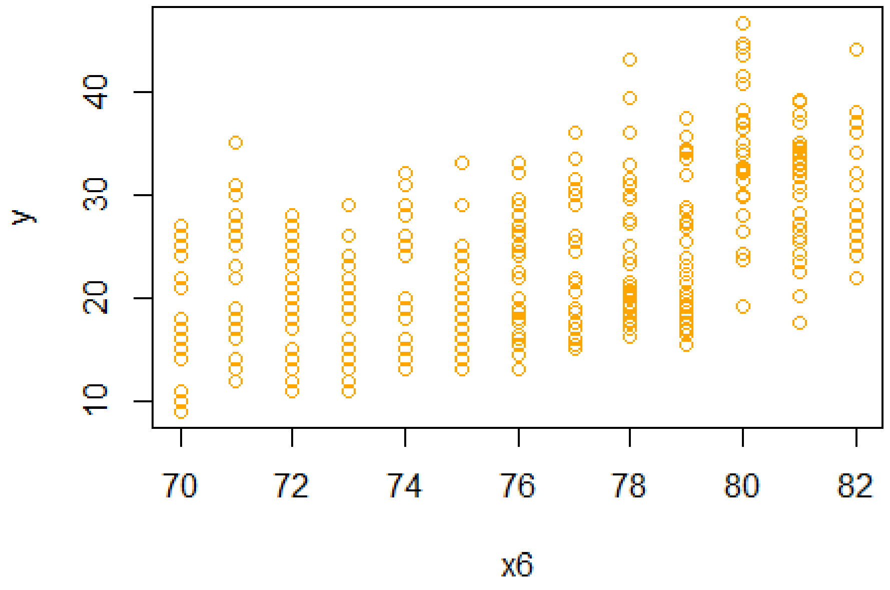
```
plot(x4,y,col="orange")
```



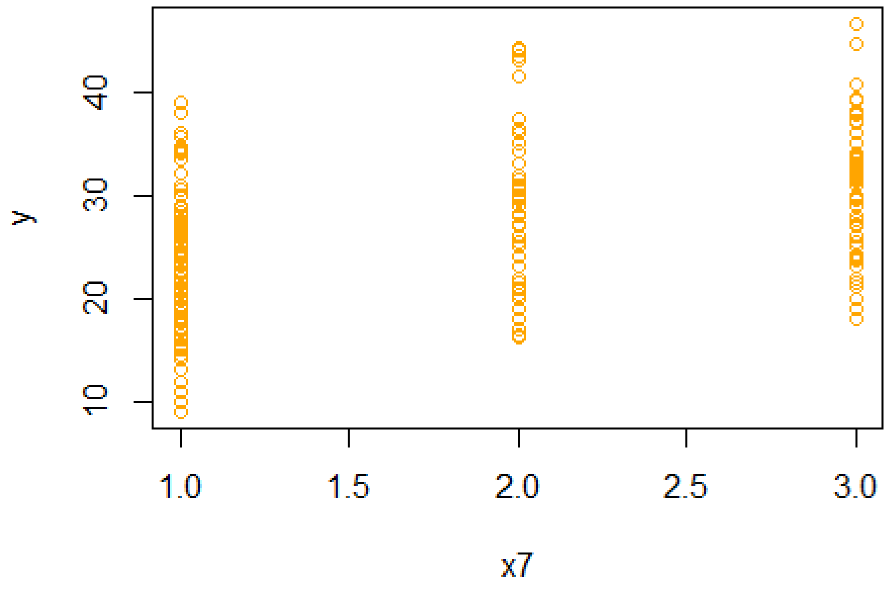
```
plot(x5,y,col="orange")
```



```
plot(x6,y,col="orange")
```



```
plot(x7,y,col="orange")
```



```
summary(fit1)[8]
```

```
## $r.squared
## [1] 0.8214781
```

X2,X3,X4 have negative linear relationship with resonse.

X5 have posetvie relation with response.

ii. Which predictors appear to have a statistically significant relationship to the response?

```
#ii)
summary(fit1)[4]

## $coefficients
##                  Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept) -17.218434622 4.6442941494 -3.707438 2.401841e-04
## x1           -0.493376319 0.3232823146 -1.526147 1.277965e-01
## x2            0.019895644 0.0075150792  2.647430 8.444649e-03
## x3           -0.016951144 0.0137868914 -1.229512 2.196328e-01
## x4           -0.006474043 0.0006520478 -9.928787 7.874953e-21
## x5            0.080575838 0.0988449567  0.815174 4.154780e-01
## x6            0.750772678 0.0509731223 14.728795 3.055983e-39
## x7            1.426140495 0.2781360924  5.127492 4.665681e-07

anova(fit1)

## Analysis of Variance Table
##
## Response: y
##            Df  Sum Sq Mean Sq   F value    Pr(>F)
## x1          1 14403.1 14403.1 1300.6838 < 2.2e-16 ***
## x2          1  1073.3  1073.3   96.9293 < 2.2e-16 ***
## x3          1   403.4   403.4   36.4301 3.731e-09 ***
## x4          1   975.7   975.7   88.1137 < 2.2e-16 ***
## x5          1     1.0     1.0    0.0872    0.7679
## x6          1  2419.1  2419.1  218.4609 < 2.2e-16 ***
## x7          1   291.1   291.1   26.2912 4.666e-07 ***
## Residuals 384  4252.2    11.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

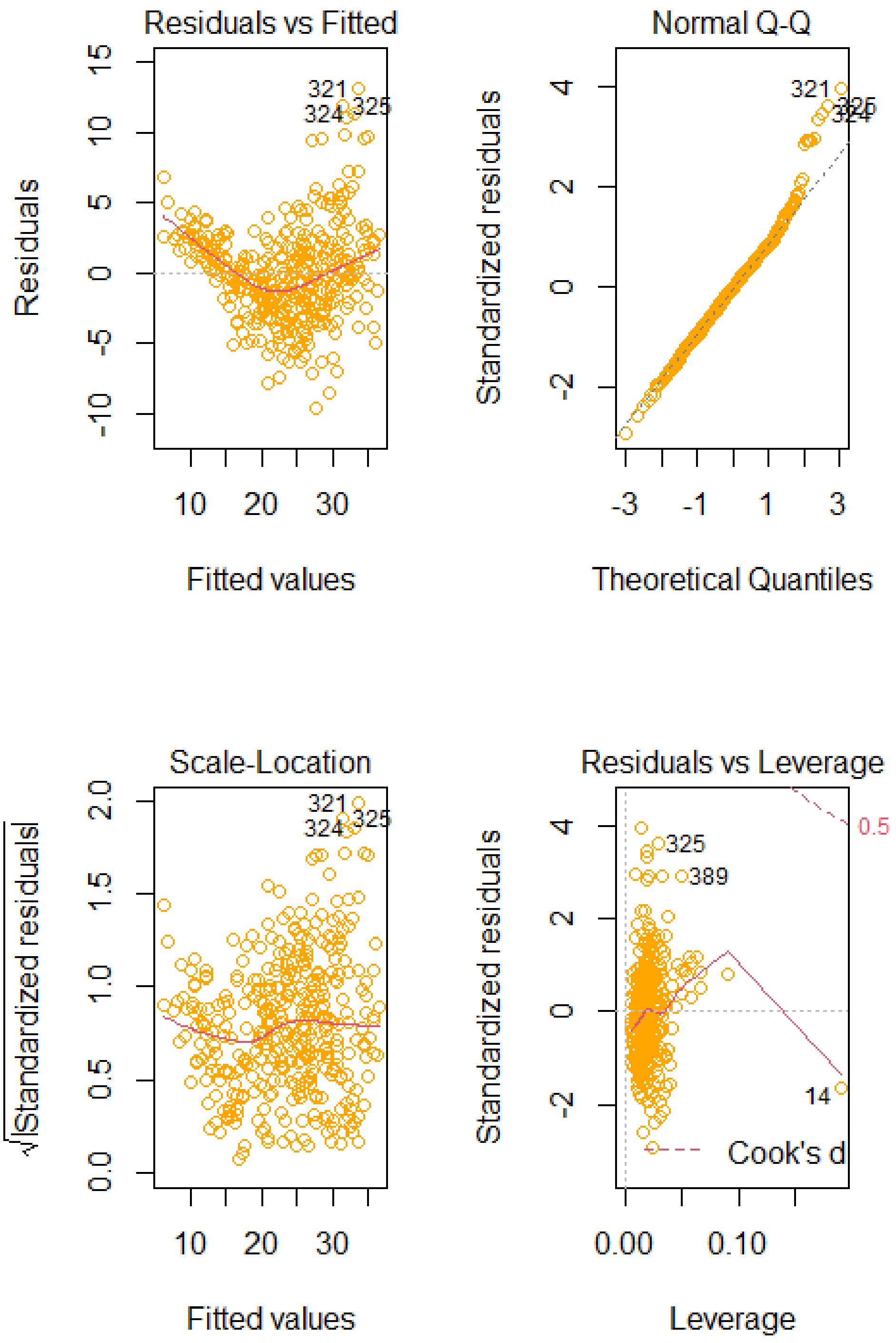
all of the predictors have statistically significant relationship to the response Except X5.

iii. What does the coefficient for the year variable suggest?

```
#iii)
coefficients(fit1)[7]

##        x6
## 0.7507727
```

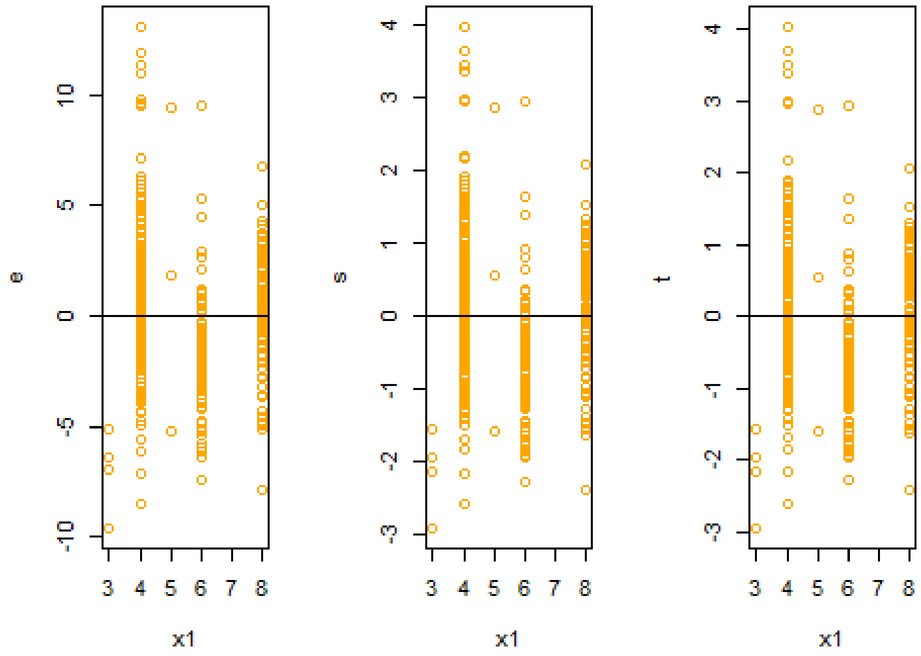
d) Use the plot() function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit.Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

```r
#d)
par(mfrow= c(1,2))
plot(fit1,col="Orange")
```

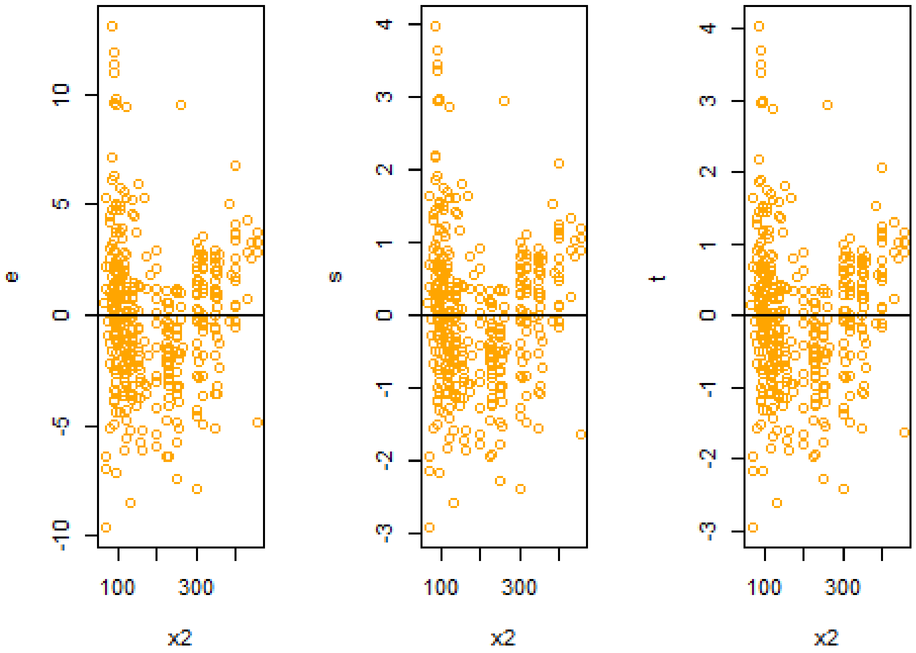


```
e<-residuals(fit1)
s<-rstandard(fit1)
t<-rstudent(fit1)
par(mfrow=c(1,3))
```

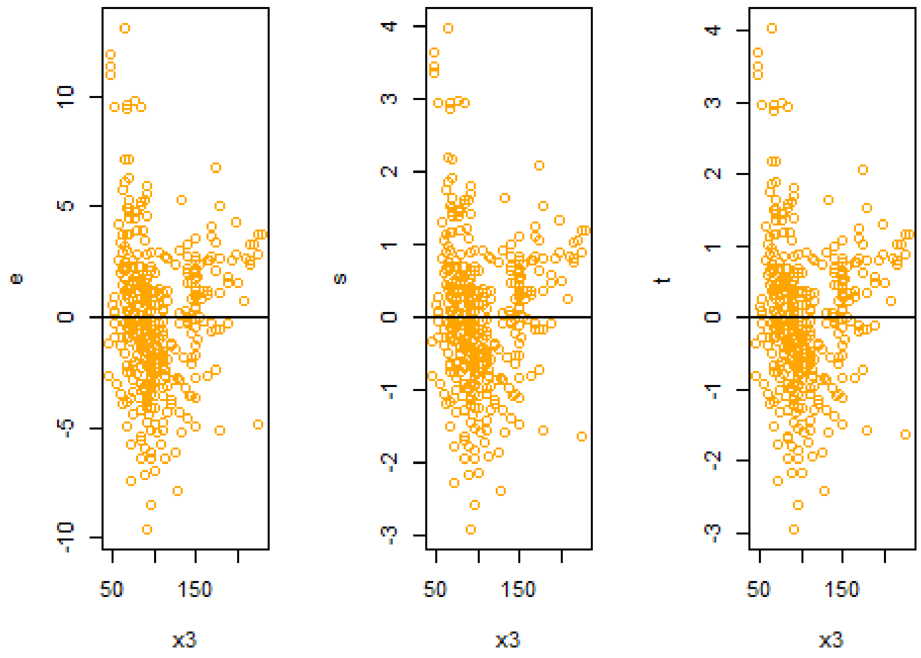
```
plot(x1,e,col="Orange")
abline(h=0,col="black")
plot(x1,s,col="Orange")
abline(h=0,col="black")
plot(x1,t,col="Orange")
abline(h=0,col="black")
```



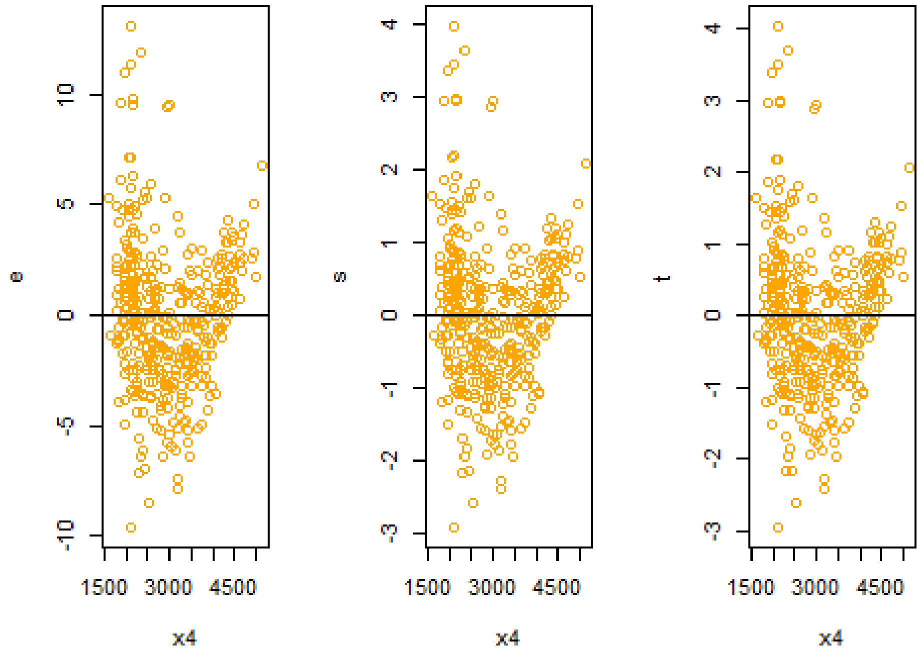
```
plot(x2,e,col="Orange")
abline(h=0,col="black")
plot(x2,s,col="Orange")
abline(h=0,col="black")
plot(x2,t,col="Orange")
abline(h=0,col="black")
```



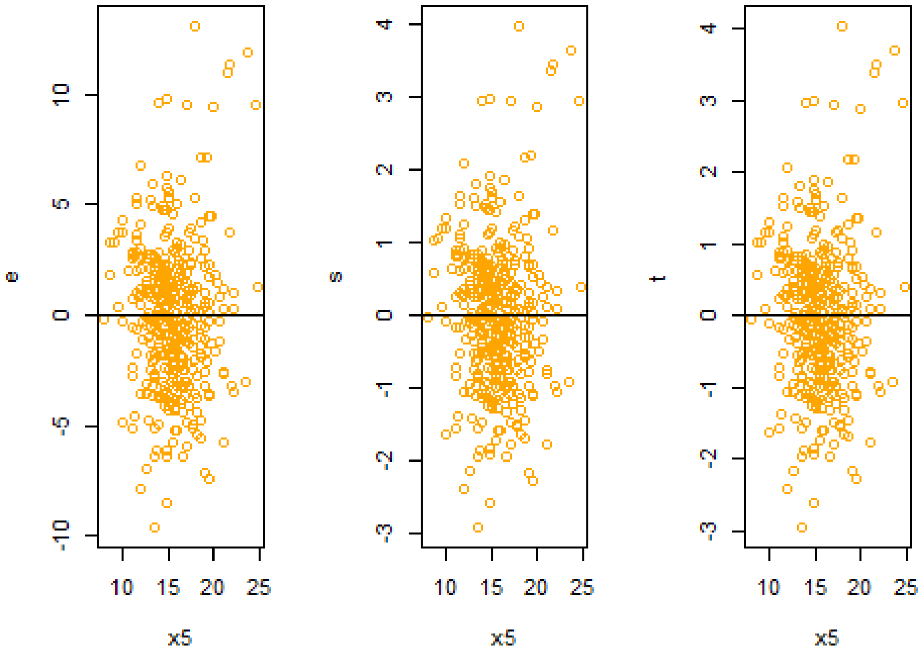
```
plot(x3,e,col="Orange")
abline(h=0,col="black")
plot(x3,s,col="Orange")
abline(h=0,col="black")
plot(x3,t,col="Orange")
abline(h=0,col="black")
```



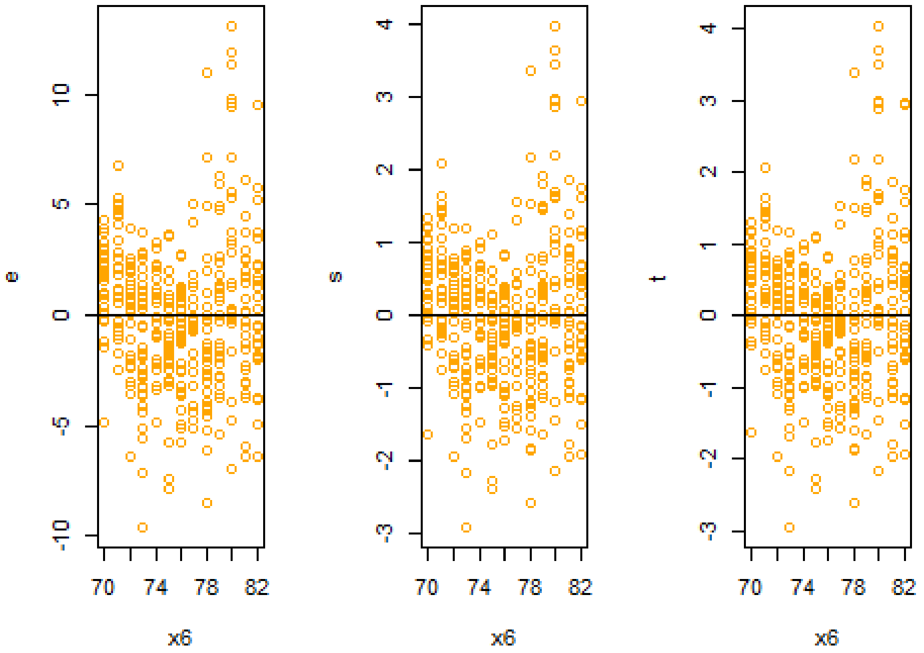
```
plot(x4,e,col="Orange")
abline(h=0,col="black")
plot(x4,s,col="Orange")
abline(h=0,col="black")
plot(x4,t,col="Orange")
abline(h=0,col="black")
```



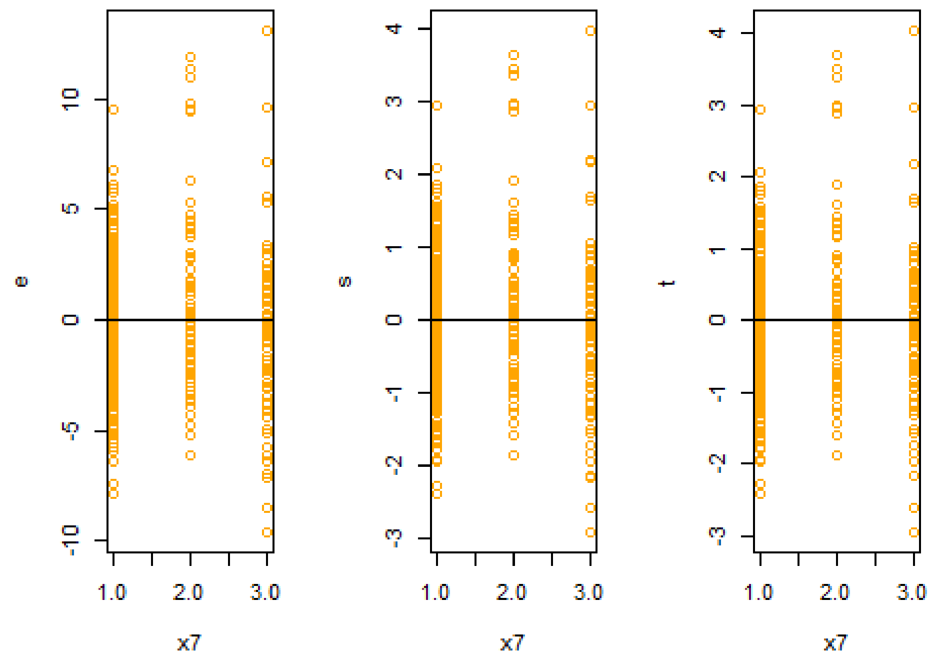
```
plot(x5,e,col="Orange")
abline(h=0,col="black")
plot(x5,s,col="Orange")
abline(h=0,col="black")
plot(x5,t,col="Orange")
abline(h=0,col="black")
```



```
plot(x6,e,col="Orange")
abline(h=0,col="black")
plot(x6,s,col="Orange")
abline(h=0,col="black")
plot(x6,t,col="Orange")
abline(h=0,col="black")
```



```
plot(x7,e,col="Orange")
abline(h=0,col="black")
plot(x7,s,col="Orange")
abline(h=0,col="black")
plot(x7,t,col="Orange")
abline(h=0,col="black")
```



at the top of the QQplot we can see some unusually points that are not in y=x linear.

we can see that we dont have fixed standard devation and i think that its posetive function of our predictors.

and we can see some outlievs and High leverage points in our model.

e) Use the * and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

```
#e)
full.fit<-
lm(y~x1+x2+x3+x4+x5+x6+x7+x1*x2+x1*x3+x1*x4+x1*x5+x1*x6+x1*x7+x2*x3+x2*x4+x3*
x5+x3*x6+x3*x7+x4*x5+x4*x6+x4*x7+x5*x6+x5*x7+x6*x7)
summary(full.fit)

##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x1 * x2 +
##     x1 * x3 + x1 * x4 + x1 * x5 + x1 * x6 + x1 * x7 + x2 * x3 +
##     x2 * x4 + x3 * x5 + x3 * x6 + x3 * x7 + x4 * x5 + x4 * x6 +
##     x4 * x7 + x5 * x6 + x5 * x7 + x6 * x7)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5267 -1.4631  0.0026  1.3259 11.3919
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.748e+01  5.028e+01   1.541  0.12414
## x1          -5.346e+00  5.546e+00  -0.964  0.33567
## x2          -1.642e-02  3.542e-02  -0.464  0.64326
## x3           4.370e-01  3.443e-01   1.269  0.20519
## x4          -2.289e-02  1.439e-02  -1.591  0.11251
## x5          -4.954e+00  2.142e+00  -2.313  0.02127 *
## x6           1.802e-01  5.804e-01   0.310  0.75639
## x7          -1.713e+01  6.869e+00  -2.494  0.01308 *
## x1:x2       -4.154e-03  4.452e-03  -0.933  0.35143
## x1:x3        1.165e-02  1.588e-02   0.734  0.46363
## x1:x4        7.368e-04  7.933e-04   0.929  0.35364
## x1:x5        1.346e-01  9.843e-02   1.367  0.17241
## x1:x6       -1.544e-03  6.482e-02  -0.024  0.98101
## x1:x7        6.039e-01  4.263e-01   1.417  0.15747
## x2:x3       -1.977e-04  2.104e-04  -0.940  0.34791
## x2:x4        1.939e-05  1.034e-05   1.876  0.06150 .
## x3:x5       -7.266e-03  3.641e-03  -1.996  0.04669 *
## x3:x6       -5.333e-03  3.949e-03  -1.350  0.17772
## x3:x7       -2.921e-03  2.914e-02  -0.100  0.92020
## x4:x5        1.719e-04  1.801e-04   0.954  0.34051
## x4:x6        7.777e-05  1.776e-04   0.438  0.66165
## x4:x7        8.030e-04  1.146e-03   0.701  0.48376
## x5:x6        4.802e-02  2.521e-02   1.905  0.05754 .
## x5:x7        4.607e-01  1.449e-01   3.179  0.00161 **
## x6:x7        7.784e-02  7.101e-02   1.096  0.27372
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.714 on 367 degrees of freedom
## Multiple R-squared:  0.8865, Adjusted R-squared:  0.8791
## F-statistic: 119.4 on 24 and 367 DF,  p-value: < 2.2e-16
```

```
anova(full.fit)
```

```
## Analysis of Variance Table
##
## Response: y
##            Df  Sum Sq Mean Sq   F value    Pr(>F)
## x1          1 14403.1 14403.1 1955.3643 < 2.2e-16 ***
## x2          1  1073.3  1073.3  145.7173 < 2.2e-16 ***
## x3          1   403.4   403.4   54.7667 9.343e-13 ***
## x4          1   975.7   975.7  132.4645 < 2.2e-16 ***
## x5          1     1.0     1.0    0.1312 0.7174471
## x6          1  2419.1  2419.1  328.4201 < 2.2e-16 ***
## x7          1   291.1   291.1   39.5245 9.203e-10 ***
## x1:x2       1   596.8   596.8   81.0189 < 2.2e-16 ***
## x1:x3       1   370.8   370.8   50.3382 6.715e-12 ***
## x1:x4       1    36.0    36.0    4.8862 0.0276888 *
## x1:x5       1     4.9     4.9    0.6609 0.4167787
## x1:x6        1    97.9    97.9   13.2872 0.0003058 ***
## x1:x7        1    42.5    42.5    5.7724 0.0167756 *  
## x2:x3        1    76.8    76.8   10.4282 0.0013530 ** 
## x2:x4        1     5.1     5.1    0.6931 0.4056486    
## x3:x5        1    44.7    44.7    6.0664 0.0142361 *  
## x3:x6        1    90.8    90.8   12.3300 0.0005013 ***
## x3:x7        1    33.3    33.3    4.5227 0.0341142 *  
## x4:x5        1     4.2     4.2    0.5665 0.4521240    
## x4:x6        1     7.7     7.7    1.0495 0.3062885    
## x4:x7        1    45.2    45.2    6.1410 0.0136574 *  
## x5:x6        1    17.7    17.7    2.4024 0.1220083    
## x5:x7        1    65.7    65.7    8.9155 0.0030173 ** 
## x6:x7        1     8.9     8.9    1.2016 0.2737230    
## Residuals 367  2703.3     7.4                       
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

now we can see that or p_value for predictors are signifact when is lower than 0.05.

the bethaj for j=0,1,2,3,4,5,6,7 is equal to the Pr(>|t|) column.

the Residual standard error = 2.714, R-squared = 0.8791.

f) Try a few different transformations of the variables, such as log(X), √X, X2. Comment on your findings.

```
#f)
#the log(X) transformation:
y<-Auto$mpg
x1<-log(Auto$cylinders)
x2<-log(Auto$displacement)
x3<-log(Auto$horsepower)
x4<-log(Auto$weight)
x5<-log(Auto$acceleration)
x6<-log(Auto$year)
x7<-log(Auto$origin)
fit4<-lm(y~x1+x2+x3+x4+x5+x6+x7)
summary(fit4)

## 
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -9.5987 -1.8172 -0.0181  1.5906 12.8132 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -66.5643    17.5053  -3.803 0.000167 ***
## x1            1.4818     1.6589   0.893 0.372273    
## x2           -1.0551     1.5385  -0.686 0.493230    
## x3           -6.9657     1.5569  -4.474 1.01e-05 ***
## x4          -12.5728     2.2251  -5.650 3.12e-08 ***
## x5           -4.9831     1.6078  -3.099 0.002082 ** 
## x6           54.9857     3.5555  15.465  < 2e-16 ***
## x7            1.5822     0.5083   3.113 0.001991 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.069 on 384 degrees of freedom
## Multiple R-squared:  0.8482, Adjusted R-squared:  0.8454 
## F-statistic: 306.5 on 7 and 384 DF,  p-value: < 2.2e-16
```

now we can see that or p_value for predictors are signifact when is lower than 0.05.

the bethaj for j=0,1,2,3,4,5,6,7 is equal to the Pr(>|t|) column.

the Residual standard error =3.069 , R-squared =0.8482 .

```
#the second root of X transformation:
y<-Auto$mpg
x1<-sqrt(Auto$cylinders)
x2<-sqrt(Auto$displacement)
x3<-sqrt(Auto$horsepower)
x4<-sqrt(Auto$weight)
x5<-sqrt(Auto$acceleration)
x6<-sqrt(Auto$year)
x7<-sqrt(Auto$origin)
fit5<-lm(y~x1+x2+x3+x4+x5+x6+x7)
summary(fit5)
```

```
## 
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -9.5250 -1.9822 -0.1111  1.7347 13.0681 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -49.79814    9.17832  -5.426 1.02e-07 ***
## x1           -0.23699    1.53753  -0.154   0.8776    
## x2            0.22580    0.22940   0.984   0.3256    
## x3           -0.77976    0.30788  -2.533   0.0117 *  
## x4           -0.62172    0.07898  -7.872 3.59e-14 ***
## x5           -0.82529    0.83443  -0.989   0.3233    
## x6           12.79030    0.85891  14.891  < 2e-16 ***
## x7            3.26036    0.76767   4.247 2.72e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
```

```
## Residual standard error: 3.21 on 384 degrees of freedom
## Multiple R-squared:  0.8338, Adjusted R-squared:  0.8308
## F-statistic: 275.3 on 7 and 384 DF,  p-value: < 2.2e-16
```

now we can see that or p_value for predictors are signifact when is lower than 0.05.

the bethaj for j=0,1,2,3,4,5,6,7 is equal to the Pr(>|t|) column.

the Residual standard error = 3.21, R-squared = 0.8338.

```
#the X power to two transformation:
y<-Auto$mpg
x1<-(Auto$cylinders)^2
x2<-(Auto$displacement)^2
x3<-(Auto$horsepower)^2
x4<-(Auto$weight)^2
x5<-(Auto$acceleration)^2
x6<-(Auto$year)^2
x7<-(Auto$origin)^2
fit6<-lm(y~x1+x2+x3+x4+x5+x6+x7)
summary(fit6)

##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6786 -2.3227 -0.0582  1.9073 12.9807
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.208e+00  2.356e+00   0.513 0.608382
## x1          -8.829e-02  2.521e-02  -3.502 0.000515 ***
## x2           5.680e-05  1.382e-05   4.109 4.87e-05 ***
## x3          -3.621e-05  4.975e-05  -0.728 0.467201
## x4          -9.351e-07  8.978e-08 -10.416  < 2e-16 ***
## x5           6.278e-03  2.690e-03   2.334 0.020130 *
## x6           4.999e-03  3.530e-04  14.160  < 2e-16 ***
## x7           4.129e-01  6.914e-02   5.971 5.37e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.539 on 384 degrees of freedom
## Multiple R-squared:  0.7981, Adjusted R-squared:  0.7944
## F-statistic: 216.8 on 7 and 384 DF,  p-value: < 2.2e-16
```

now we can see that or p_value for predictors are signifact when is lower than 0.05.

the bethaj for j=0,1,2,3,4,5,6,7 is equal to the Pr(>|t|) column.

the Residual standard error = 3.539, R-squared = 0.7981.

End.

---

5) This question should be answered using the Carseats data set.

a) Fit a multiple regression model to predict Sales using Price, Urban, and US.

```
#Exercise fivth(Ten of Book):
#introducing the variables:for classifactions qualitative variables(Yes=2
, No=1)

#a)
y<-Carseats$Sales
x1<-Carseats$Price
x2<-as.integer(Carseats$Urban)
x3<-as.integer(Carseats$US)
fit1<-lm(y~x1+x2+x3)
```

yes =2 and no =1

b) Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative!

```
#b)
summary(fit1)

##
## Call:
## lm(formula = y ~ x1 + x2 + x3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.864812   0.841681  14.097  < 2e-16 ***
## x1          -0.054459   0.005242 -10.389  < 2e-16 ***
## x2          -0.021916   0.271650  -0.081    0.936
## x3           1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

just X3 have posetive relationship with response.

c) Write out the model in equation form, being careful to handle the qualitative variables properly.

```
#c)
f<-function(X1,X2,X3){
  f<-(as.numeric(coef(fit1)[1])+ (as.numeric(coef(fit1)[2])*X1)+
(as.numeric(coef(fit1)[3])*X2)+ (as.numeric(coef(fit1)[4])*X3))
  return(f)
}
```

d) For which of the predictors can you reject the null hypothesis H0 : βj = 0?

```
#d)
anova(fit1)

## Analysis of Variance Table
##
## Response: y
##            Df  Sum Sq Mean Sq  F value    Pr(>F)    
## x1          1  630.03  630.03 103.0603 < 2.2e-16 ***
## x2          1    0.10    0.10   0.0158    0.9001    
## x3          1  131.31  131.31  21.4802  4.86e-06 ***
## Residuals 396 2420.83    6.11                       
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova(fit1)[5]

##           Pr(>F)    
## x1        <2e-16 ***
## x2        0.9001    
## x3        <2e-16 ***
## Residuals           
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

now we can see that all of our predictors are signifact Except X2.

e) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

```
#e)
Reduce.fit<-lm(Sales~US+Price,data = Carseats)
```

f) How well do the models in (a) and (e) fit the data?

```
#f)
summary(Reduce.fit)

##
## Call:
## lm(formula = Sales ~ US + Price, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9269 -1.6286 -0.0574  1.5766  7.0515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
## USYes        1.19964    0.25846   4.641 4.71e-06 ***
## Price       -0.05448    0.00523 -10.416  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.469 on 397 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2354
## F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

the R squared is fixed(0.2393) when we remove the Urban perdictors form our model.

the residuals standard error just is lower(0.003) than the full model.

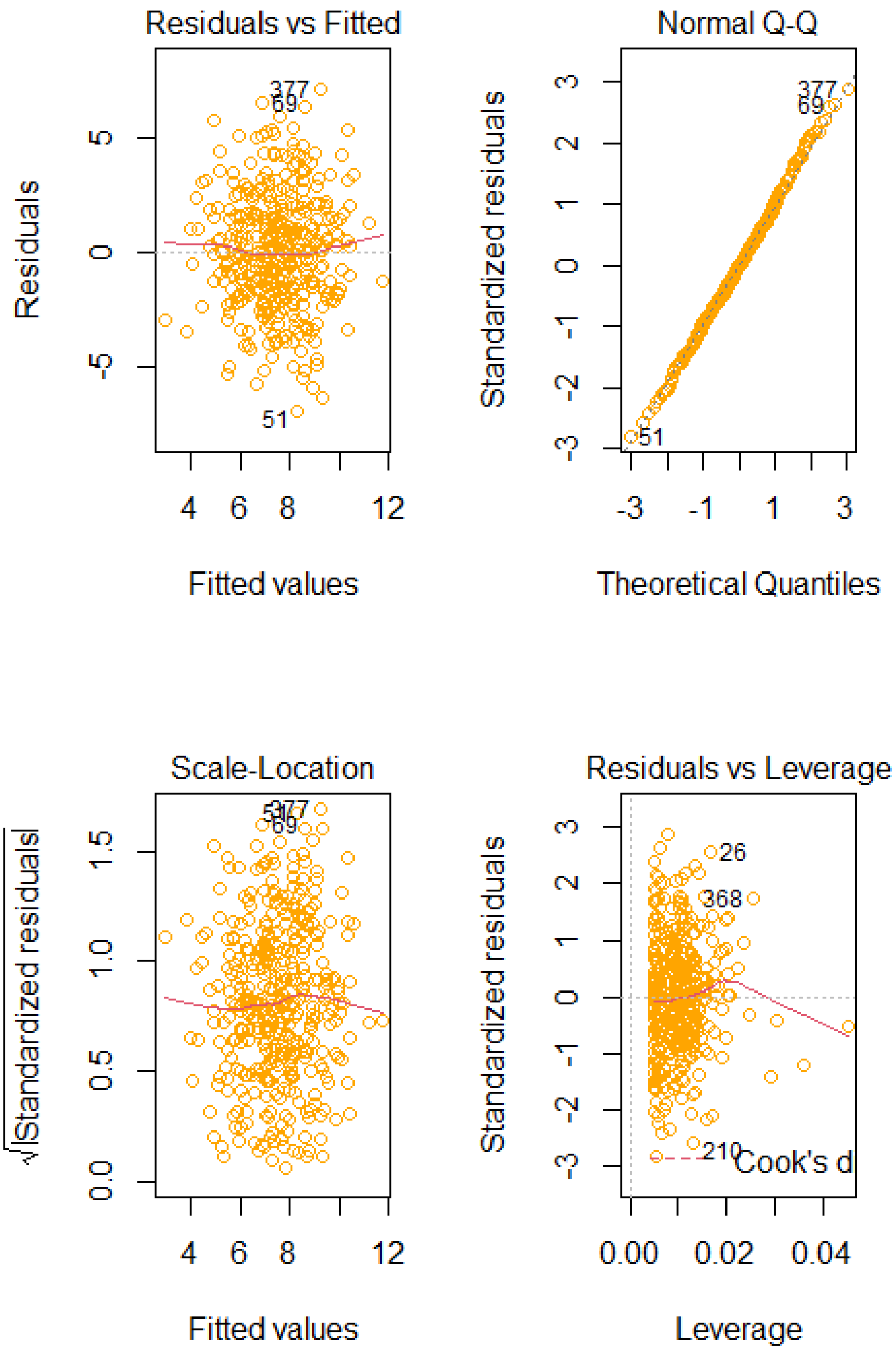
g) Using the model from (e), obtain 95 % confidence intervals for the coefficient(s).

```
#g)
confint(fit1,level = 0.95)
```

```
##                   2.5 %      97.5 %
## (Intercept) 10.21009150 13.51953328
## x1          -0.06476419 -0.04415351
## x2          -0.55597316  0.51214085
## x3           0.69130419  1.70984121
```

h) Is there evidence of outliers or high leverage observations in the model from (e)?

```
#h)
par(mfrow=c(1,2))
plot(fit1,col="orange")
```



i think that our residuals have standard normal distrubtion.

at the first devation is low then more after that again lower.

we have some lievs andHigh leverage points.

End.

Second

Introducing data to R

```
#after installing the msme packages we should library that`
library("msme")

## Loading required package: MASS

## Loading required package: lattice

data("medpar")
#we want to see small sample of our data:
head(medpar[,1:6])

##   los hmo white died age80 type
## 1   4   0     1    0     0    1
## 2   9   1     1    0     0    1
## 3   3   1     1    1     1    1
## 4   9   0     1    0     0    1
## 5   1   0     1    1     1    1
## 6   4   0     1    1     0    1
```

We know that the los column indicates the length of nights the person has been hospitalized.

hmo column indicates whether the person was covered by insurance or not (yes=1 ,No=0)

white column indicates whether the person is white or not (yes=1 , No=0)

died column indicates whether the person died within 48 hours of hospitalization or not (yes=1 ,No=0)

age80 column ididactes wheter the person age is more equal 80 or not (yes=1 ,No=0)

type coulmn idicates the person's kind of hospitalization (Optional=1 ,Instant=2 ,Emergency=3)

Solve:

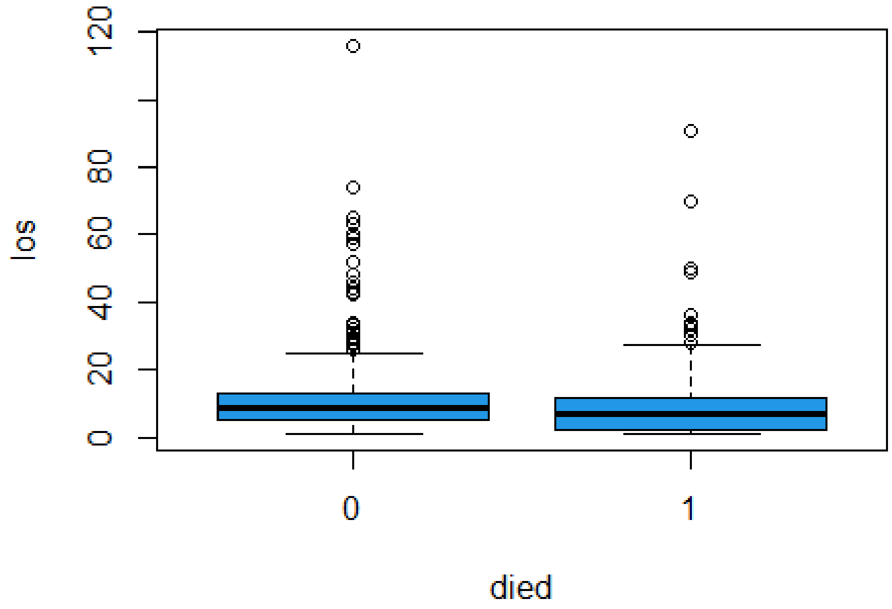
chek correlation between the variables:

```
cor(medpar[,1:6])

##                los           hmo       white          died       age80
## los     1.00000000 -5.832123e-02 -0.06779545 -1.037458e-01 -0.03303782
## hmo    -0.05832123  1.000000e+00  0.05435482 -4.371603e-05 -0.03853239
## white -0.06779545  5.435482e-02  1.00000000  3.830089e-02  0.04647059
## died  -0.10374584 -4.371603e-05  0.03830089  1.000000e+00  0.13167978
## age80 -0.03303782 -3.853239e-02  0.04647059  1.316798e-01  1.00000000
## type   0.25511584 -1.127590e-01 -0.07471925  8.975658e-02 -0.03005332
##             type
## los    0.25511584
## hmo   -0.11275902
## white -0.07471925
## died   0.08975658
## age80 -0.03005332
## type   1.00000000
```

According to this matrix we can say that we dont have any significant dependence and correlation between variables.

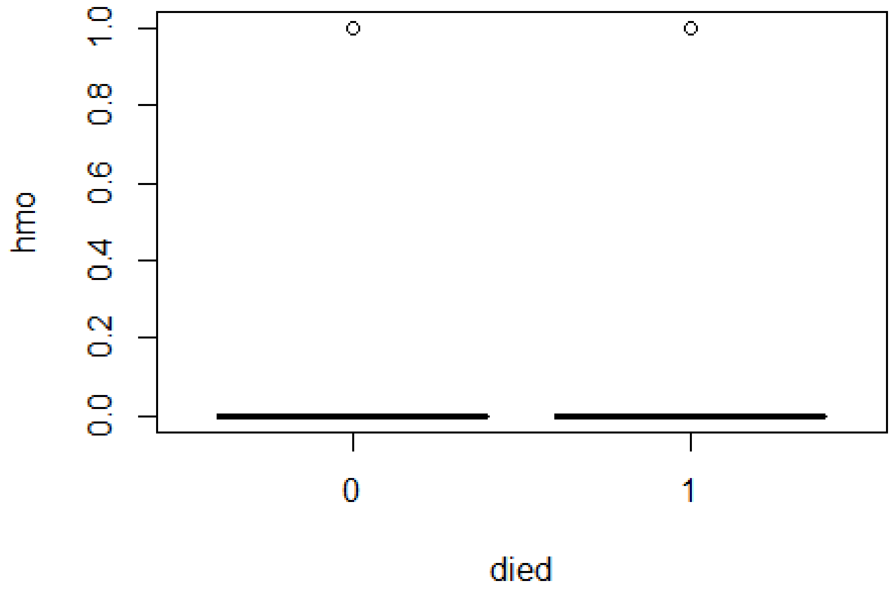
chek the relation beween response and variables with Boxplots:

```
#at the first we should attach the data
attach(medpar)
#now we want to see the Box plot of each variable with our response:
boxplot(los~died ,col=4)
```



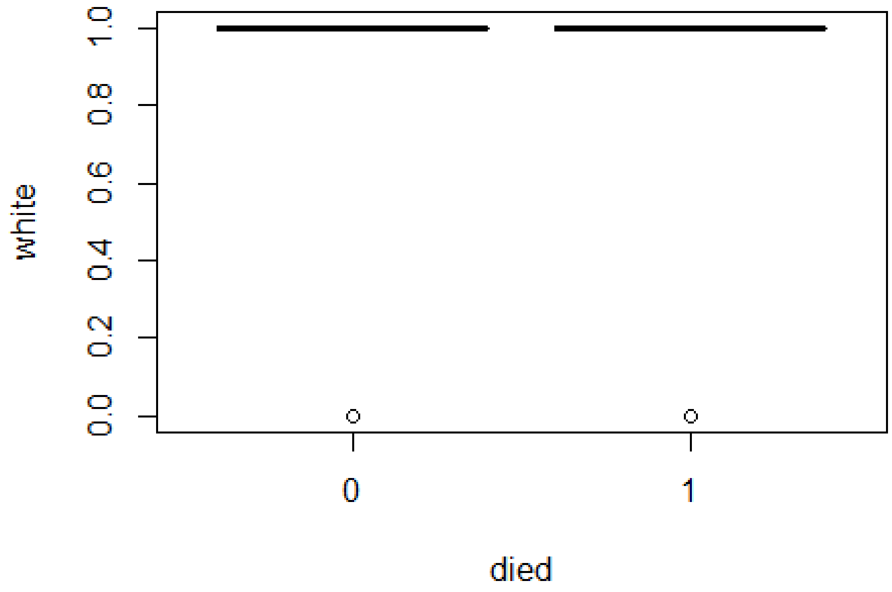
According to this Boxplot we can say that median of los for death and live persons are equal but for more los value we have more live persons.

```
boxplot(hmo~died ,col=3)
```



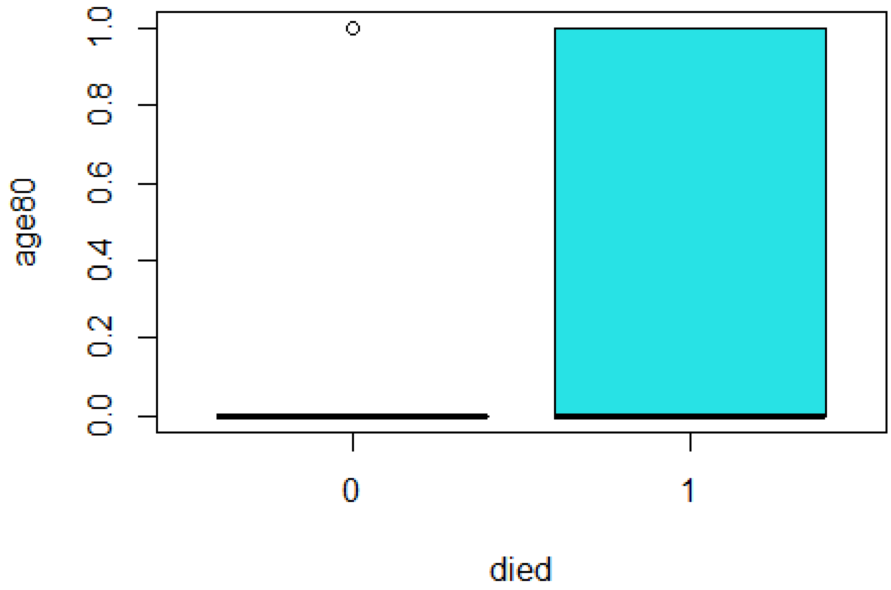
According to top Boxplot we can say that the majority of those admitted did not have insurance coverage.

```
boxplot(white~died ,col=2)
```



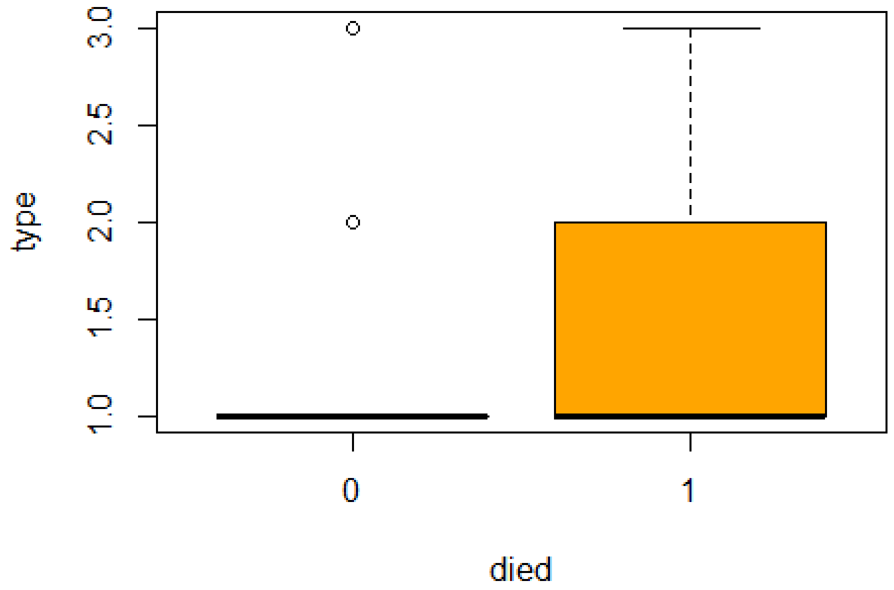
According to top Boxplot we can say that the majority of those admitted were whith white skin.

```
boxplot(age80~died ,col=85)
```



According to top Boxplot we can say that the majority of those admitted that death, are more and equal 80 years old.

```
boxplot(type~died, col="orange")
```



According to top Boxplot we can say that the majority of those admitted that their hospitalization were instant and emergency includ the majority of death.

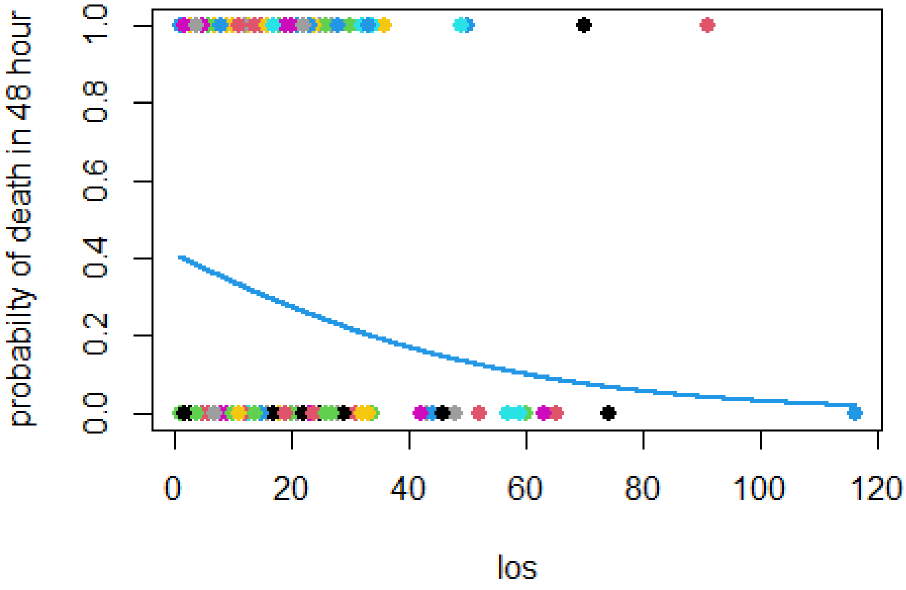
Logstics Regression with severan variables and univariables

Logstics regression for each variable and response:

```
fit1<-glm(died~los,family = binomial)
coef(fit1)

## (Intercept)         los 
## -0.36170695 -0.03048316

summary(fit1)

## 
## Call:
## glm(formula = died ~ los, family = binomial)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.0160  -0.9449  -0.8767   1.3614   2.5212  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -0.361707   0.088436  -4.090 4.31e-05 ***
## los         -0.030483   0.007691  -3.964 7.38e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1922.9  on 1494  degrees of freedom
## Residual deviance: 1904.6  on 1493  degrees of freedom
## AIC: 1908.6
##
## Number of Fisher Scoring iterations: 4
```

According to the summary and coef function outputs we can say that we have nagative relationship(betha los = -0.03048316) and our p-value for signifacting H0 (Betha0 = Betha1 = 0)is equal to 7.38e-05 and its less than alpha(0.05) so we say the H0 reject and the regression is signifact.

```
seq1<-data.frame(los=seq(min(los),max(los),len=10^4))
seq1$died=predict(fit1,newdata= seq1,type="response")
plot(los,died,col= c(1:length(los)),pch=19,cex=1.1,ylab="probabilty of death
in 48 hour")
lines(died~los , seq1 ,col=4,lwd=2)
```



According to top plot we can see the probabilty of death in 48 hour for los value between (0, 35) is value between (0.2 , 0.4) and for los more equal than 60 day is less than 0.1 .

```
fit2<-glm(died~hmo,family = binomial)
coef(fit2)
```

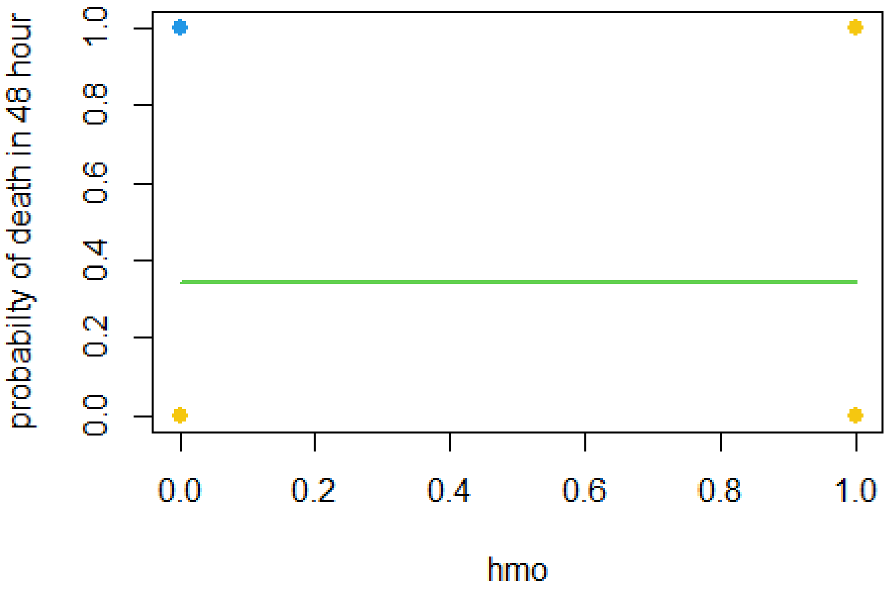
```
##   (Intercept)           hmo
## -0.6492752962 -0.0002512619
```

```
summary(fit2)
```

```
##
## Call:
## glm(formula = died ~ hmo, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9169  -0.9169  -0.9169   1.4626   1.4627
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.6492753  0.0594332 -10.924   <2e-16 ***
## hmo         -0.0002513  0.1486501  -0.002    0.999
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1922.9  on 1494  degrees of freedom
## Residual deviance: 1922.9  on 1493  degrees of freedom
## AIC: 1926.9
##
## Number of Fisher Scoring iterations: 4
```

According to the summary and coef function outputs we can say that we have very Weak nagative relationship(betha hmo= -0.0002512619) and our p-value for signifacting H0 (Betha0 = Betha1 = 0)is equal to 0.999 and its not less than alpha(0.05) so we say the H0 accept and the regression is not signifact. i think its not good variable for our regression.

```
seq2<-data.frame(hmo=seq(min(hmo),max(hmo),len=10^4))
seq2$died=predict(fit2,newdata= seq2,type="response")
plot(hmo,died,col= c(1:length(hmo)),pch=19,cex=1.1,ylab="probabilty of death
in 48 hour")
lines(died~hmo , seq2 ,col=3,lwd=2)
```



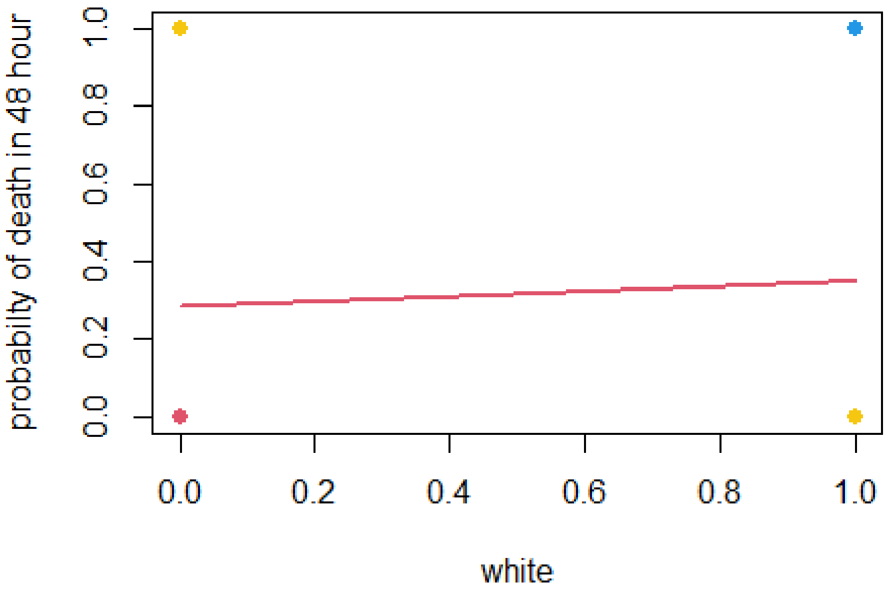
According to top plot we can see the fix line about probabilty of death is equal to 0.34 and its fix when the hmo variable change, so we can say its not good variable for our regression.

```
fit3<-glm(died~white,family = binomial)
coef(fit3)

## (Intercept)       white 
##  -0.9273406   0.3025126

summary(fit3)

## 
## Call:
## glm(formula = died ~ white, family = binomial)
## 
## Deviance Residuals: 
##    Min      1Q  Median      3Q     Max  
## -0.926  -0.926  -0.926   1.452   1.588  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  -0.9273     0.1969  -4.710 2.48e-06 ***
## white         0.3025     0.2049   1.476     0.14    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 1922.9  on 1494  degrees of freedom
## Residual deviance: 1920.6  on 1493  degrees of freedom
## AIC: 1924.6
##
## Number of Fisher Scoring iterations: 4
```

According to the summary and coef function outputs we can say that we have positive relationship(betha white= 0.3025126) and our p-value for signifacting H0 (Betha0 = Betha1 = 0)is equal to 0.14 and its less than alpha(0.05) so we say the H0 reject and the regression is signifact.

```
seq3<-data.frame(white=seq(min(white),max(white),len=10^4))
seq3$died=predict(fit3,newdata= seq3,type="response")
plot(white,died,col= c(1:length(white)),pch=19,cex=1.1,ylab="probabilty of
death in 48 hour")
lines(died~white , seq3 ,col=2,lwd=2)
```



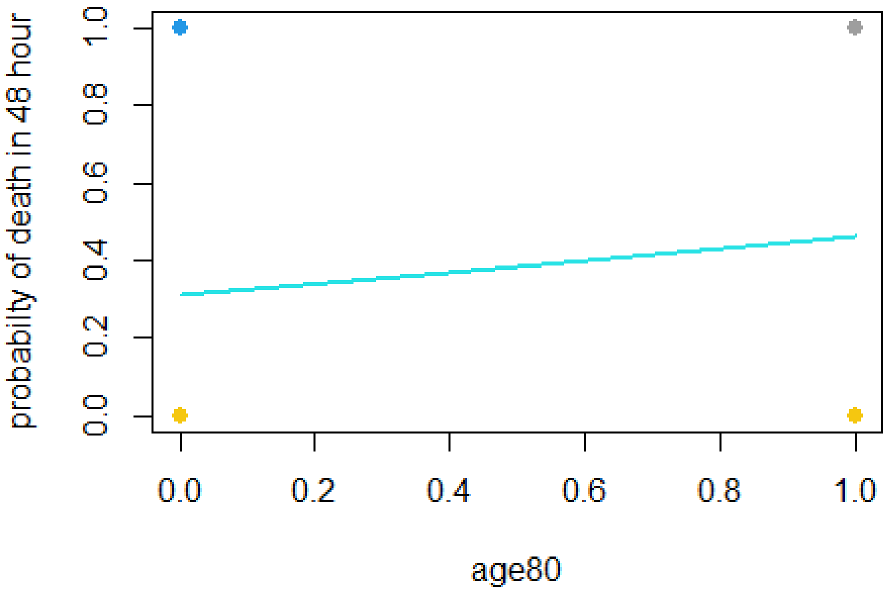
According to top plot we can see the line probabilty of death is between(0.283,0.36) and its apprixmimately fix when the white variable change, so we can say its not good variable for our regression.

```
fit4<-glm(died~age80,family = binomial)
coef(fit4)

## (Intercept)       age80
##  -0.8007213   0.6428183
```

```
summary(fit4)

##
## Call:
## glm(formula = died ~ age80, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1111  -0.8612  -0.8612   1.2452   1.5308
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.80072    0.06336 -12.639  < 2e-16 ***
## age80        0.64282    0.12732   5.049 4.45e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1922.9  on 1494  degrees of freedom
## Residual deviance: 1897.7  on 1493  degrees of freedom
## AIC: 1901.7
##
## Number of Fisher Scoring iterations: 4
```

According to the summary and coef function outputs we can say that we have positive relationship(betha age80= 0.64282) and our p-value for signifacting H0 (Betha0 = Betha1 = 0)is equal to 4.45e-07 and its less than alpha(0.05) so we say the H0 reject and the regression is signifact.

```
seq4<-data.frame(age80=seq(min(age80),max(age80),len=10^4))
seq4$died=predict(fit4,newdata= seq4,type="response")
plot(age80,died,col= c(1:length(age80)),pch=19,cex=1.1,ylab="probabilty of
death in 48 hour")
lines(died~age80 , seq4 ,col=85,lwd=2)
```



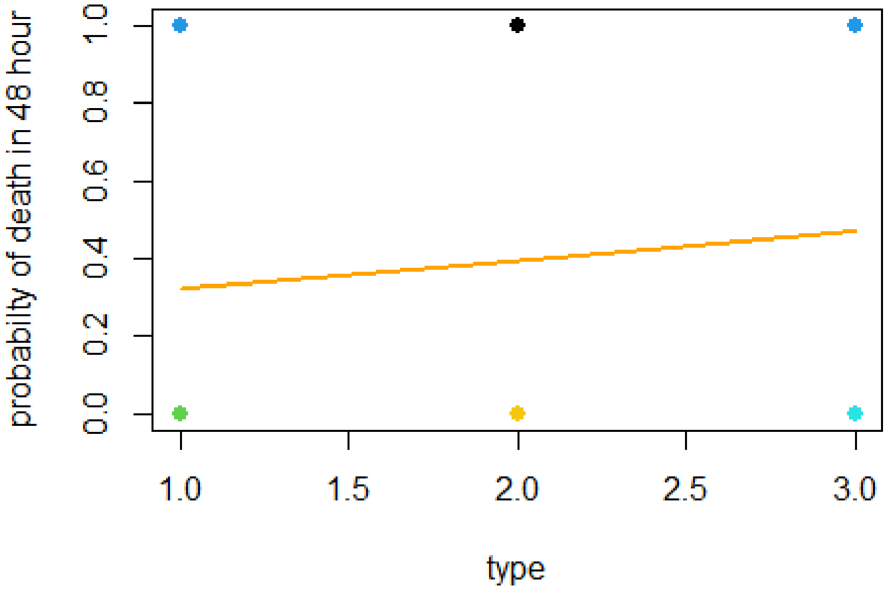
According to top plot we can see the line with positive slope about probabilty of death is between (0.309 ,0.46) and its increasing when the age80 variable increase.

```
fit5<-glm(died~type,family = binomial)
coef(fit5)

## (Intercept)        type
##  -1.0613241   0.3121013

summary(fit5)

##
## Call:
## glm(formula = died ~ type, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1248  -0.8799  -0.8799   1.3678   1.5075
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.06132    0.13245  -8.013 1.12e-15 ***
## type         0.31210    0.09055   3.447 0.000568 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 1922.9  on 1494  degrees of freedom
## Residual deviance: 1911.1  on 1493  degrees of freedom
## AIC: 1915.1
##
## Number of Fisher Scoring iterations: 4
```

According to the summary and coef function outputs we can say that we have positive relationship(betha type= 0.3121013) and our p-value for signifacting H0 (Betha0 = Betha1 = 0)is equal to 0.000568 and its less than alpha(0.05) so we say the H0 reject and the regression is signifact.

```
seq5<-data.frame(type=seq(min(type),max(type),len=10^4))
seq5$died=predict(fit5,newdata= seq5,type="response")
plot(type,died,col= c(1:length(type)),pch=19,cex=1.1,ylab="probabilty of
death in 48 hour")
lines(died~type , seq5 ,col="Orange",lwd=2)
```



According to top plot we can see the line with positive slope about probabilty of death is between (0.320 ,0.468) and its increasing when the age80 variable increase.

Logstics regression for all variable and response(Multiple logstic regression)

```
full.fit<-glm(died~los+hmo+white+age80+type,family = binomial)
coef(full.fit)

## (Intercept)         los         hmo       white       age80        type
## -1.32099883 -0.03679831  0.06156626  0.25907749  0.65167023  0.46900157
```

```
summary(full.fit)

##
## Call:
## glm(formula = died ~ los + hmo + white + age80 + type, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5196  -0.8928  -0.7923   1.2790   2.3160
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.320999   0.253700  -5.207 1.92e-07 ***
## los         -0.036798   0.007891  -4.663 3.11e-06 ***
## hmo          0.061566   0.152732   0.403    0.687
## white        0.259077   0.210062   1.233    0.217
## age80        0.651670   0.129545   5.030 4.89e-07 ***
## type         0.469002   0.097186   4.826 1.39e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1922.9  on 1494  degrees of freedom
## Residual deviance: 1856.2  on 1489  degrees of freedom
## AIC: 1868.2
##
## Number of Fisher Scoring iterations: 4
```

According to this models we can say that just the (hmo,white) variable p-values is more than 0.05 and its not good for our model and we should remove it, the others variable have positive relationships with response expect los.

Reduce logstics model

```
reduce.fit<-glm(died~los+age80+type,family = binomial)
coef(reduce.fit)

## (Intercept)         los       age80        type
##  -1.0540990  -0.0373605   0.6566579   0.4576727

summary(reduce.fit)

##
## Call:
## glm(formula = died ~ los + age80 + type, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5085  -0.8845  -0.8018   1.2856   2.2566
##
## Coefficients:
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.054099   0.147929  -7.126 1.04e-12 ***
## los         -0.037360   0.007876  -4.743 2.10e-06 ***
## age80        0.656658   0.129178   5.083 3.71e-07 ***
## type         0.457673   0.096384   4.748 2.05e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1922.9  on 1494  degrees of freedom
## Residual deviance: 1858.0  on 1491  degrees of freedom
## AIC: 1866
##
## Number of Fisher Scoring iterations: 4
```

We can see that the reduce models outputs show that the (hmo,white) variable wasnt important and dont have effects on response.

the Reduce model is best model with out any bad variable and all of them are signifact and we have good predict for our response,

we can see the logstics predicton function here for reduce model:

y = died , x1 = los , x2 = age80 , x3 = type.

End.

---

thrid

Introducing data to R

```
#now we want to attach our data to R from notepad or txt.
Heart <- read.table("F://lessons//regression2//exercise//2-
2//data.txt",sep=",",head=T,row.names=1)
#now we want to see some of our data
head(Heart)
```

```
##   sbp tobacco  ldl adiposity famhist typea obesity alcohol age chd
## 1 160   12.00 5.73     23.11 Present    49   25.30   97.20  52   1
## 2 144    0.01 4.41     28.61  Absent    55   28.87    2.06  63   1
## 3 118    0.08 3.48     32.28 Present    52   29.14    3.81  46   0
## 4 170    7.50 6.41     38.03 Present    51   31.99   24.26  58   1
## 5 134   13.60 3.50     27.78 Present    60   25.99   57.34  49   1
## 6 132    6.20 6.47     36.21 Present    62   30.77   14.14  45   0
```

```
#the fivth column of our dataset or famhist is not numeric and we get labe (
present = 1 ,Absent = 0)
Heart[,5]=ifelse(Heart[,5]=="Absent",0,1)
```

sbp column indicates systolic blood pressure.

tobacco column indicates cumulative tobacco (Kg).

ldl column indicates low densiity lipoprotein cholesterol.

adiposity coulmn it is a concept of body masses that are often in the form of fat.

famhist column indicates family history of heart disease (Present, Absent).

typea column indicates type-A behavior.

obesity column it is a concept of obesity that has different types and according to its coefficient, different information can be obtained.

alcohol column indicates current alcohol consumption.

age column indicates age at onset.

chd column indicates response, coronary heart disease.

Solve

chek correlation between the variables:

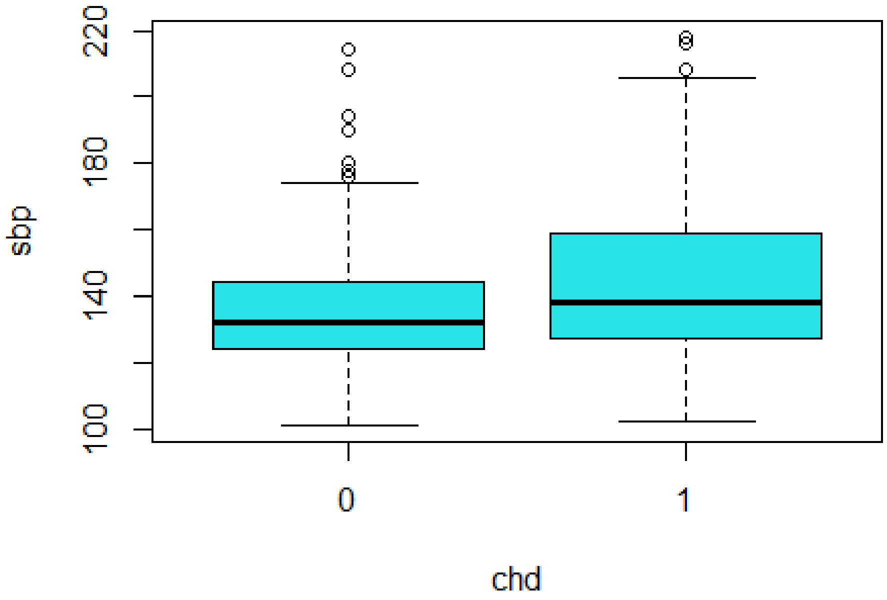
```
cor(Heart)
```

```
##                   sbp     tobacco         ldl   adiposity    famhist
## sbp        1.00000000  0.21224652  0.15829633  0.35650008 0.08564531
## tobacco    0.21224652  1.00000000  0.15890546  0.28664037 0.08860143
## ldl        0.15829633  0.15890546  1.00000000  0.44043175 0.16135306
## adiposity  0.35650008  0.28664037  0.44043175  1.00000000 0.18172101
## famhist    0.08564531  0.08860143  0.16135306  0.18172101 1.00000000
## typea     -0.05745431 -0.01460788  0.04404758 -0.04314364 0.04480858
## obesity    0.23806661  0.12452941  0.33050586  0.71655625 0.11559508
## alcohol    0.14009559  0.20081339 -0.03340340  0.10033013 0.08051969
## age        0.38877060  0.45033016  0.31179923  0.62595442 0.23966742
## chd        0.19235411  0.29971754  0.26305268  0.25412139 0.27237273
##                 typea    obesity     alcohol        age        chd
## sbp       -0.05745431 0.23806661  0.14009559  0.3887706 0.19235411
## tobacco   -0.01460788 0.12452941  0.20081339  0.4503302 0.29971754
## ldl        0.04404758 0.33050586 -0.03340340  0.3117992 0.26305268
## adiposity -0.04314364 0.71655625  0.10033013  0.6259544 0.25412139
## famhist    0.04480858 0.11559508  0.08051969  0.2396674 0.27237273
## typea      1.00000000 0.07400610  0.03949794 -0.1026063 0.10315583
## obesity    0.07400610 1.00000000  0.05161957  0.2917771 0.10009508
## alcohol    0.03949794 0.05161957  1.00000000  0.1011246 0.06253068
## age       -0.10260632 0.29177713  0.10112465  1.0000000 0.37297334
## chd        0.10315583 0.10009508  0.06253068  0.3729733 1.00000000
```

chek the relation beween response and variables with Boxplots:

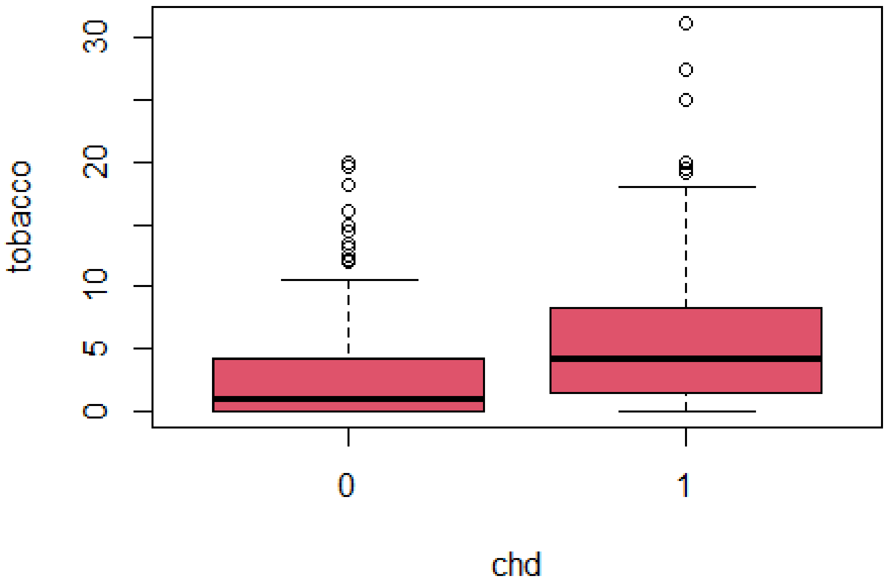
```
#at the first we should attach the data
attach(Heart)
```

```
#now we want to see the Box plot of each variable with our response:
boxplot(sbp~chd,col=85)
```



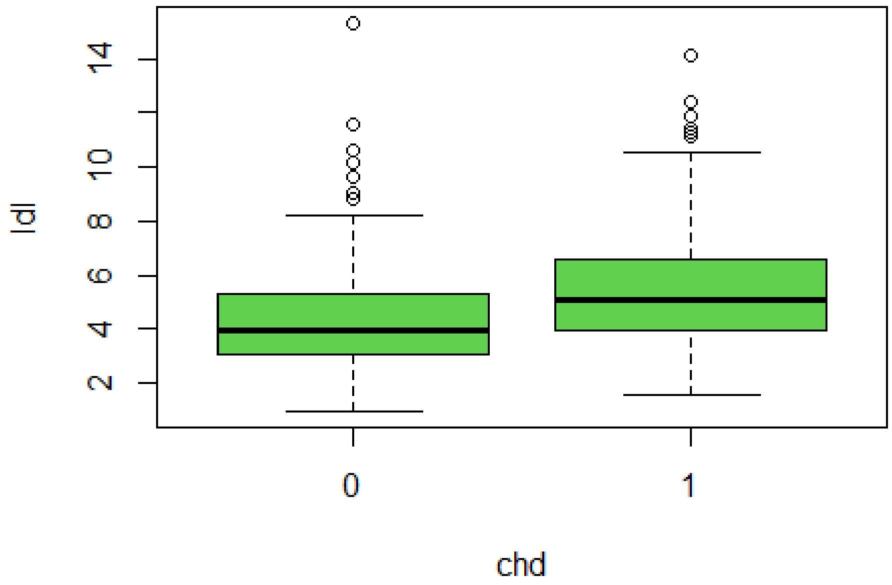
According to this Boxplot we can say that median of sbp for Having and not having coronary heart disease are equal but for more sbp values we have more cornary heart disease.

```
boxplot(tobacco~chd,col=2)
```



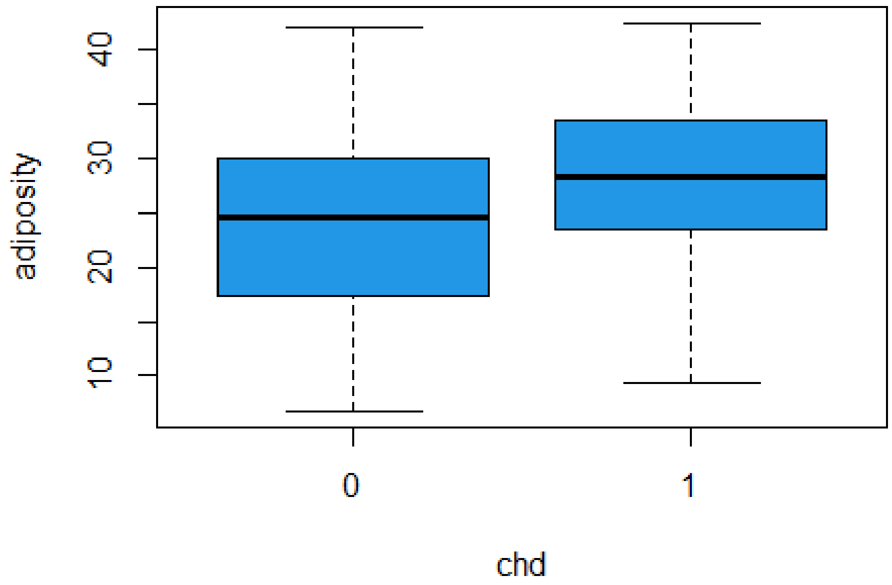
According to this Boxplot we can say that median of tovacco for Having and not having coronary heart disease are not equal and for more tobacco values we have more cornary heart disease.

```
boxplot(ldl~chd,col=3)
```



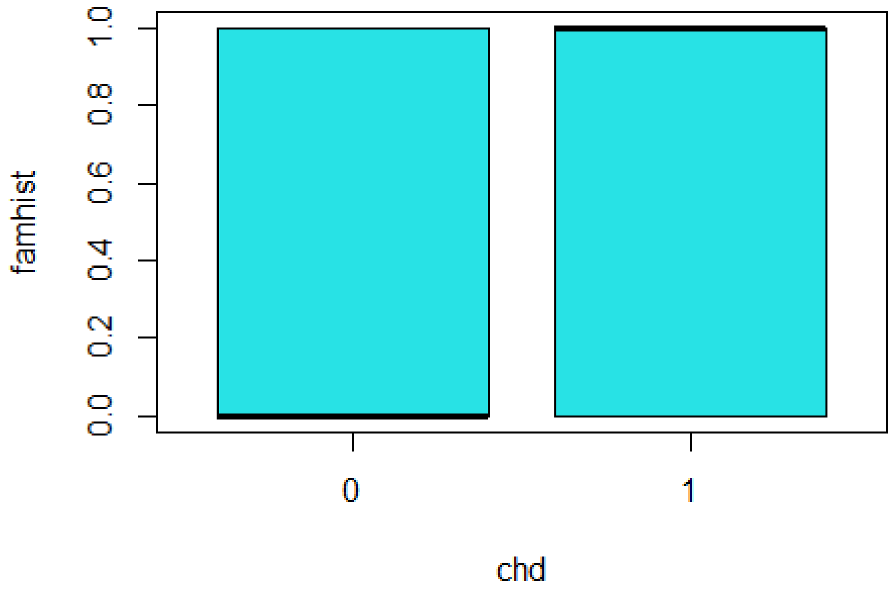
According to this Boxplot we can say that median of ldl for Having and not having coronary heart disease are not equal and for more ldl values we have more cornary heart disease.

```
boxplot(adiposity~chd,col=4)
```



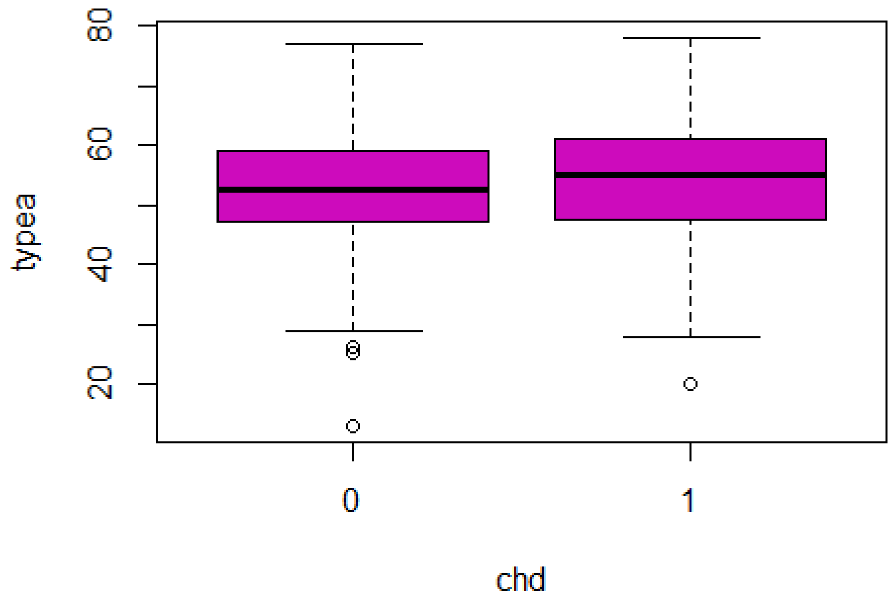
According to this Boxplot we can say that median of adiposity for Having and not having coronary heart disease are not equal and for adiposity values between (0,25) we can say that we dont have cornary heart disease and for more adiposity vlaues we have more and more cornary heart disease.

```
boxplot(famhist~chd,col=5)
```



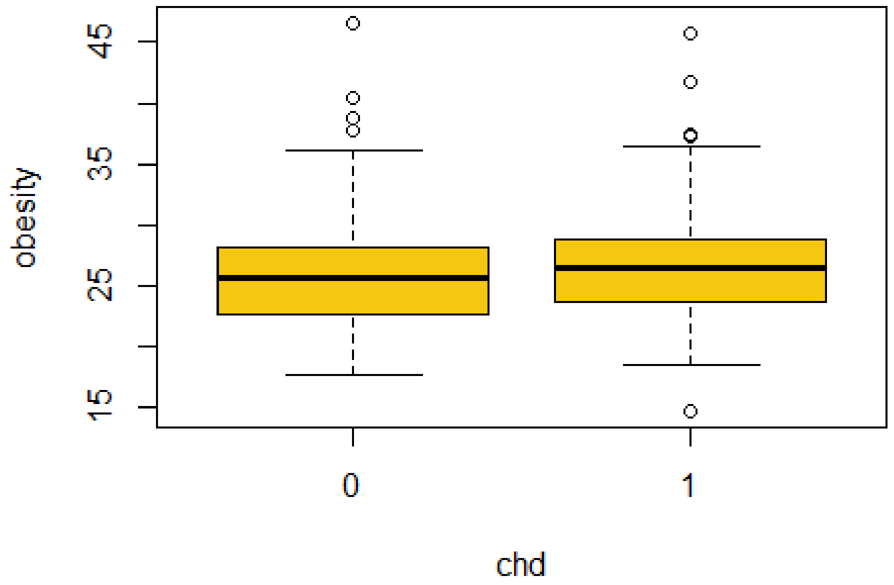
According to this Boxplot we can say that median of famhist for Having and not having coronary heart disease are not equal but for more famhist values we have more cornary heart disease.

```
boxplot(typea~chd,col=6)
```



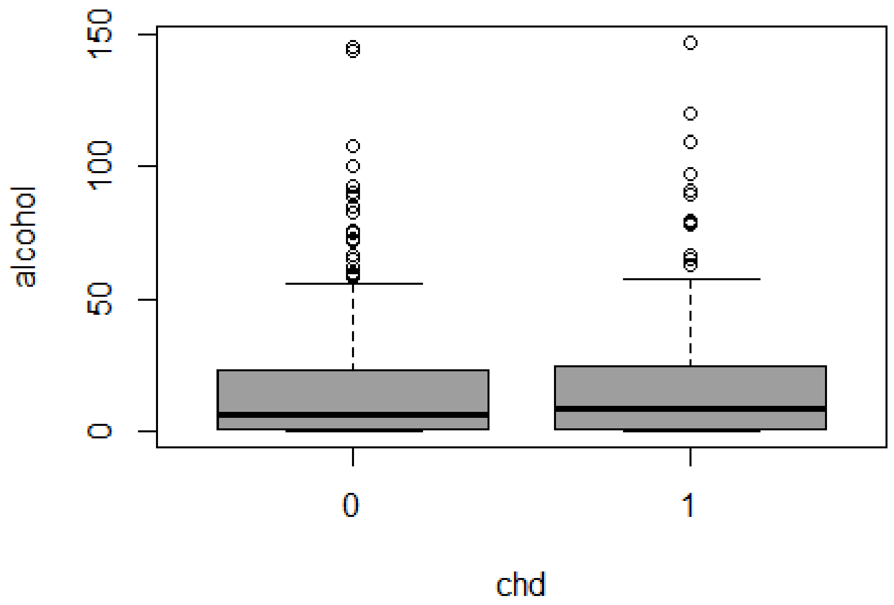
According to this Boxplot we can say that median of typea for Having and not having coronary heart disease are equal.

```
boxplot(obesity~chd,col=7)
```



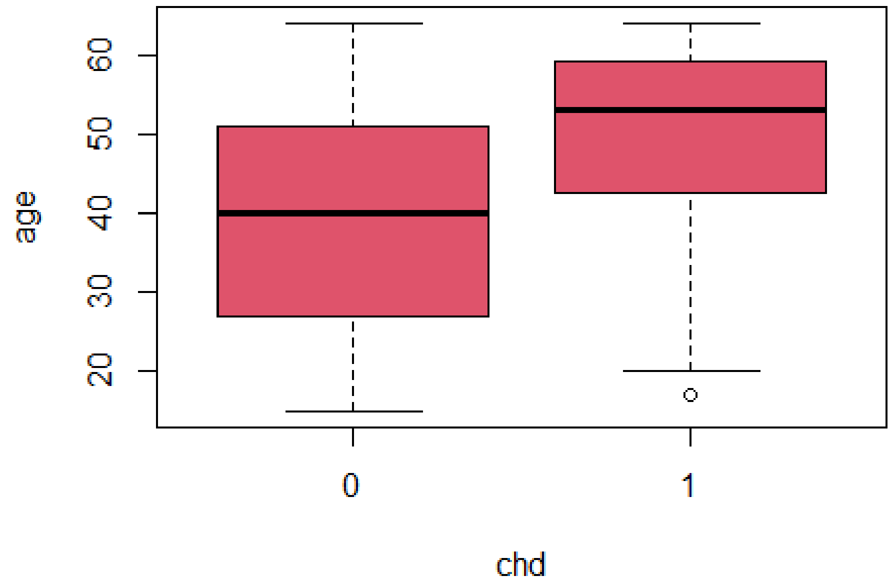
According to this Boxplot we can say that median of obesity for Having and not having coronary heart disease are equal.

```
boxplot(alcohol~chd,col=8)
```



According to this Boxplot we can say that median of alcohol for Having and not having coronary heart disease are equal.

```
boxplot(age~chd,col=10)
```



According to this Boxplot we can say that median of age for Having and not having coronary heart disease are not equal and for age values between (1,40) we can say that we dont have cornary heart disease and for more adiposity vlaues we have more and more cornary heart disease.

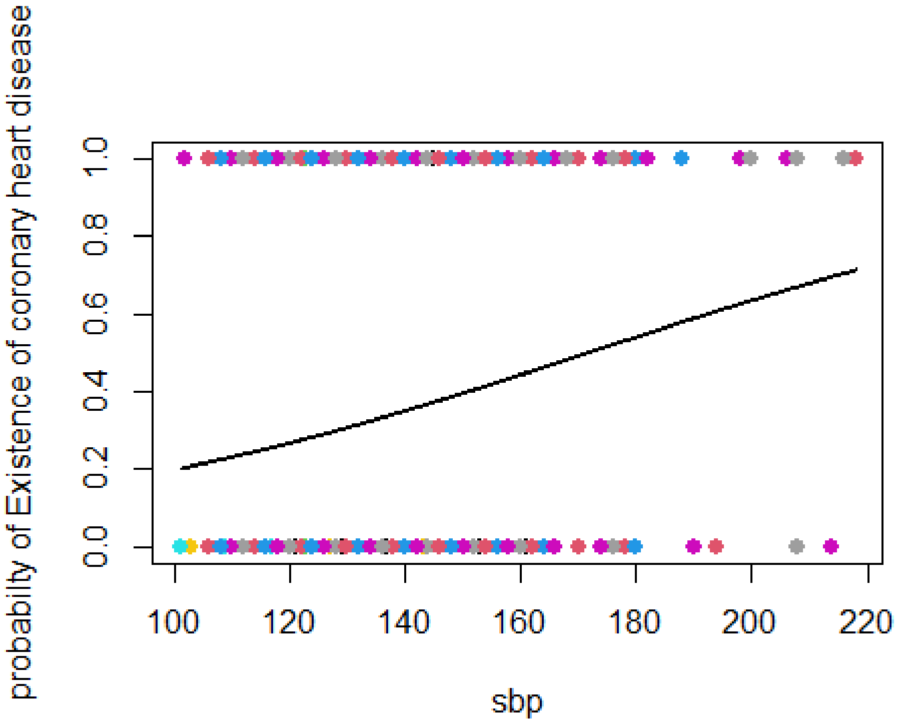
Logstics Regression with severan variables and univariables

Logstics regression for each variable and response:

```
fit1<-glm(chd~sbp,family = binomial)
coef(fit1)

## (Intercept)         sbp 
## -3.35271610  0.01950936

summary(fit1)

## 
## Call:
## glm(formula = chd ~ sbp, family = binomial)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.5405  -0.8982  -0.8009   1.3113   1.7836  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -3.352716   0.687698  -4.875 1.09e-06 ***
## sbp          0.019509   0.004863   4.012 6.02e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 596.11  on 461  degrees of freedom
## Residual deviance: 579.32  on 460  degrees of freedom
## AIC: 583.32
##
## Number of Fisher Scoring iterations: 4
```

According to the summary and coef function outputs we can say that we have posetive relationship (betha sbp = 0.019509) and our p-value for signifacting H0 (Betha0 = Betha1 = 0)is equal to 6.02e-05 and its less than alpha(0.05) so we say the H0 reject and the regression is signifact.

```
seq1<-data.frame(sbp=seq(min(sbp),max(sbp),len=10^4))
seq1$chd=predict(fit1,newdata= seq1,type="response")
plot(sbp,chd,col=sbp,pch=19,cex=1.1,ylab="probabilty of Existence of coronary
heart disease")
lines(chd~sbp , seq1 ,col=1,lwd=2)
```



According to top plot we can see the probabilty of Existence of coronary heart disease for sbp large or big values its more and more and we have posetive relationship for example sbp is equal 220 our probabilty is equal 0.8.

```
fit2<-glm(chd~tobacco,family = binomial)
coef(fit2)
```

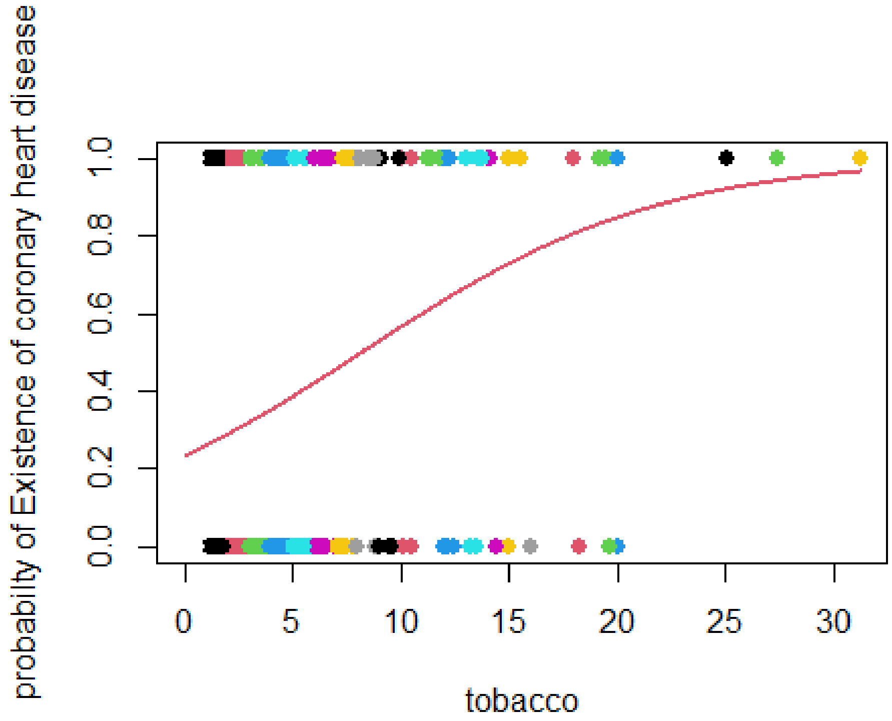
```
## (Intercept)     tobacco
##  -1.1894300   0.1452696
```

```
summary(fit2)
```

```
##
## Call:
## glm(formula = chd ~ tobacco, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9397  -0.8467  -0.7290   1.1997   1.7060
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.18943    0.13900  -8.557  < 2e-16 ***
## tobacco      0.14527    0.02476   5.866 4.46e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 596.11  on 461  degrees of freedom
## Residual deviance: 554.65  on 460  degrees of freedom
## AIC: 558.65
##
## Number of Fisher Scoring iterations: 4
```

According to the summary and coef function outputs we can say that we have posetive relationship (betha tobacco = 0.14527) and our p-value for signifacting H0 (Betha0 = Betha1 = 0)is equal to 4.46e-09 and its less than alpha(0.05) so we say the H0 reject and the regression is signifact.

```
seq2<-data.frame(tobacco=seq(min(tobacco),max(tobacco),len=10^4))
seq2$chd=predict(fit2,newdata= seq2,type="response")
plot(tobacco,chd,col=tobacco,pch=19,cex=1.1,ylab="probabilty of Existence of
coronary heart disease")
lines(chd~tobacco , seq2 ,col=2,lwd=2)
```



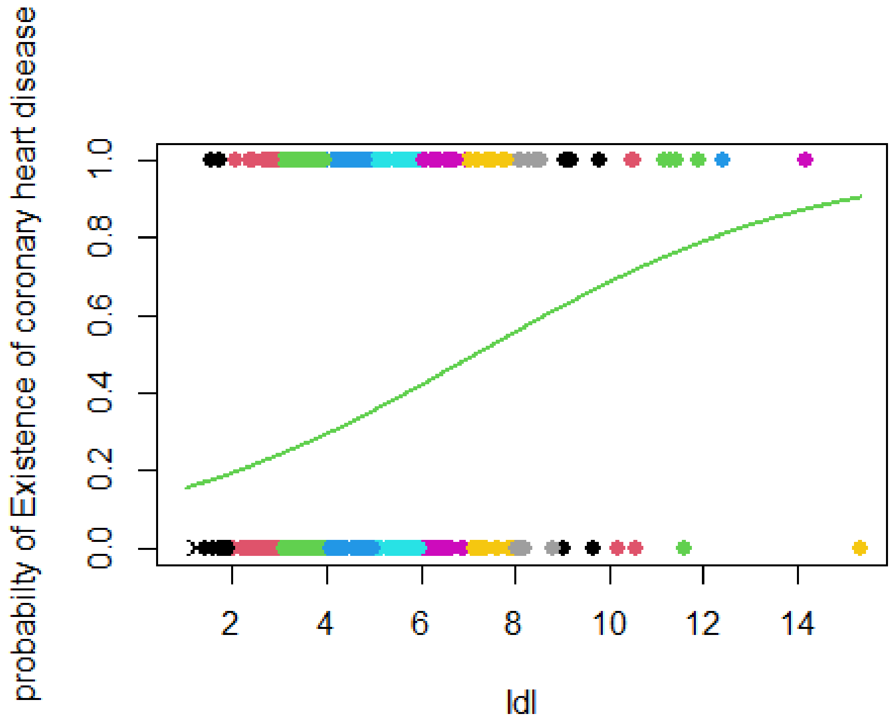
According to top plot we can see the probabilty of Existence of coronary heart disease for tobacco large or big its more and more values and we have posetive relationship , for example tobacco is more and equal 25 our probabilty is more than 0.95.

```
fit3<-glm(chd~ldl,family = binomial)
coef(fit3)

## (Intercept)         ldl
##  -1.9686681   0.2746613

summary(fit3)

##
## Call:
## glm(formula = chd ~ ldl, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1647  -0.8948  -0.7426   1.2688   1.8637
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.96867    0.27308  -7.209 5.63e-13 ***
## ldl          0.27466    0.05164   5.319 1.04e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 596.11  on 461  degrees of freedom
## Residual deviance: 564.28  on 460  degrees of freedom
## AIC: 568.28
##
## Number of Fisher Scoring iterations: 4
```

According to the summary and coef function outputs we can say that we have posetive relationship (betha tobacco = 0.27466) and our p-value for signifacting H0 (Betha0 = Betha1 = 0)is equal to 1.04e-07 and its less than alpha(0.05) so we say the H0 reject and the regression is signifact.

```
seq3<-data.frame(ldl=seq(min(ldl),max(ldl),len=10^4))
seq3$chd=predict(fit3,newdata= seq3,type="response")
plot(ldl,chd,col=ldl,pch=19,cex=1.1,ylab="probabilty of Existence of coronary
heart disease")
lines(chd~ldl, seq3 ,col=3,lwd=2)
```



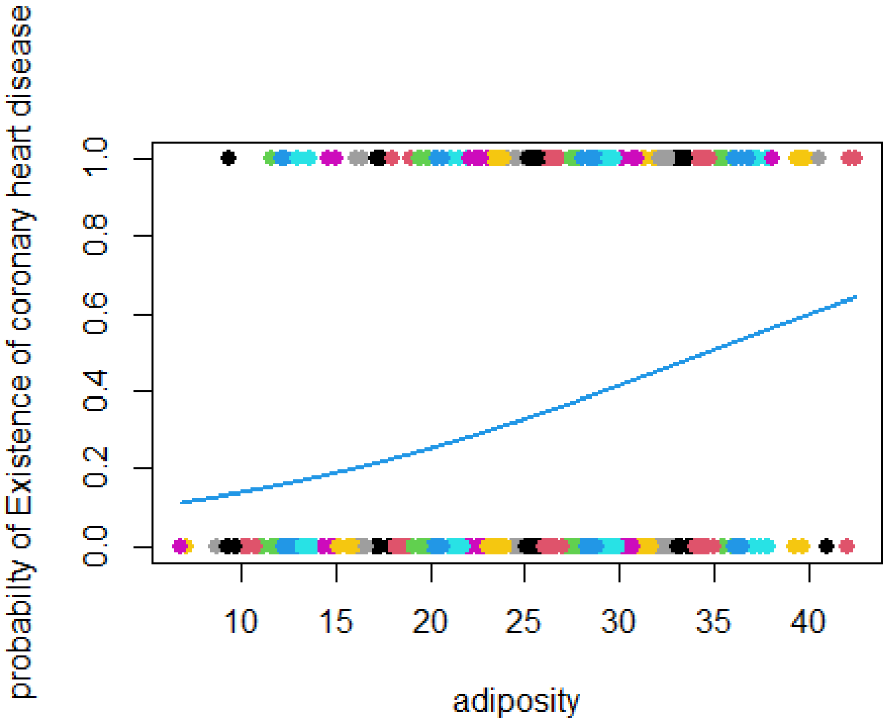
According to top plot we can see the probabilty of Existence of coronary heart disease for ldl large or big values its more and more and we have posetive relationship, for example tobacco is more and equal 12 our probabilty is more than 0.8.

```
fit4<-glm(chd~adiposity,family = binomial)
coef(fit4)
```

```
## (Intercept)   adiposity 
## -2.56922635  0.07410225

summary(fit4)

## 
## Call:
## glm(formula = chd ~ adiposity, family = binomial)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.4170  -0.9554  -0.6960   1.2228   2.0081  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -2.56923    0.38712  -6.637 3.21e-11 ***
## adiposity    0.07410    0.01396   5.308 1.11e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 596.11  on 461  degrees of freedom
## Residual deviance: 565.05  on 460  degrees of freedom
## AIC: 569.05
## 
## Number of Fisher Scoring iterations: 4
```

According to the summary and coef function outputs we can say that we have posetive relationship (betha adiposity = 0.07410) and our p-value for signifacting H0 (Betha0 = Betha1 = 0)is equal to 1.11e-07 and its less than alpha(0.05) so we say the H0 reject and the regression is signifact.

```
seq4<-data.frame(adiposity=seq(min(adiposity),max(adiposity),len=10^4))
seq4$chd=predict(fit4,newdata= seq4,type="response")
plot(adiposity,chd,col=adiposity,pch=19,cex=1.1,ylab="probabilty of Existence 
of coronary heart disease")
lines(chd~adiposity, seq4 ,col=4,lwd=2)
```



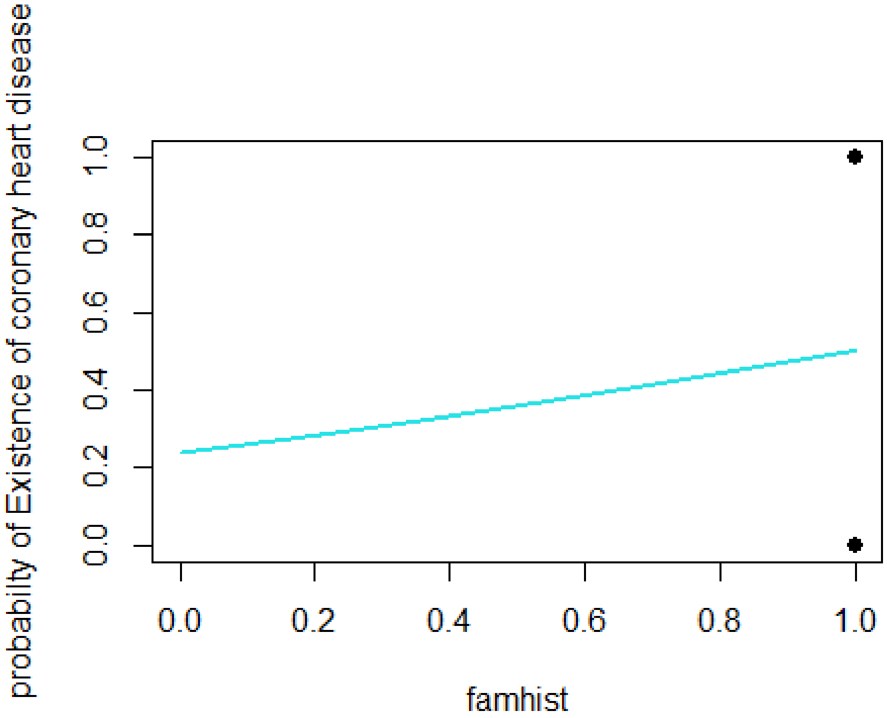
According to top plot we can see the probabilty of Existence of coronary heart disease for adiposity large or big values its more and more and we have posetive relationship, for example adiposity is more and equal 35 our probabilty is more than 0.5.

```
fit5<-glm(chd~famhist,family = binomial)
coef(fit5)

## (Intercept)     famhist
##   -1.168993    1.168993

summary(fit5)

##
## Call:
## glm(formula = chd ~ famhist, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1774  -0.7356  -0.7356   1.1774   1.6968
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.1690     0.1431  -8.169 3.12e-16 ***
## famhist       1.1690     0.2033   5.751 8.85e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 596.11  on 461  degrees of freedom
## Residual deviance: 561.89  on 460  degrees of freedom
## AIC: 565.89
##
## Number of Fisher Scoring iterations: 4
```

According to the summary and coef function outputs we can say that we have posetive relationship (betha famhist = 1.1690) and our p-value for signifacting H0 (Betha0 = Betha1 = 0)is equal to 8.85e-09 and its less than alpha(0.05) so we say the H0 reject and the regression is signifact.

```
seq5<-data.frame(famhist=seq(min(famhist),max(famhist),len=10^4))
seq5$chd=predict(fit5,newdata= seq5,type="response")
plot(famhist,chd,col=famhist,pch=19,cex=1.1,ylab="probabilty of Existence of
coronary heart disease")
lines(chd~famhist, seq5 ,col=5,lwd=2)
```



According to top plot we can see the probabilty of Existence of coronary heart disease for famhist large or big values its more and more and we have posetive relationship. but i think that its not good variable for our multiple logsticks regression.

```
fit6<-glm(chd~typea,family = binomial)
coef(fit6)
```

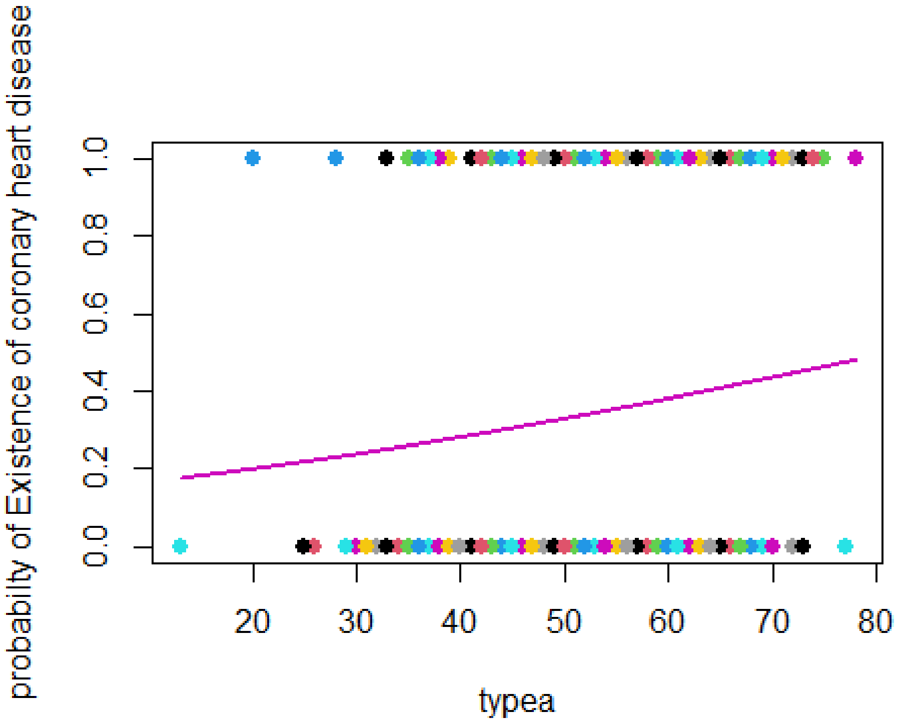
```
## (Intercept)       typea
## -1.84469292  0.02263029
```

```
summary(fit6)
```

```
##
## Call:
## glm(formula = chd ~ typea, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1343  -0.9440  -0.8602   1.3874   1.7967
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.84469    0.56068  -3.290   0.0010 **
## typea        0.02263    0.01026   2.205   0.0275 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 596.11  on 461  degrees of freedom
## Residual deviance: 591.12  on 460  degrees of freedom
## AIC: 595.12
##
## Number of Fisher Scoring iterations: 4
```

According to the summary and coef function outputs we can say that we have weak posetive relationship (betha typea = 0.02263) and our p-value for signifacting H0 (Betha0 = Betha1 = 0)is equal to 0.0275 and its less than alpha(0.05) so we say the H0 reject and the regression is signifact.

```
seq6<-data.frame(typea=seq(min(typea),max(typea),len=10^4))
seq6$chd=predict(fit6,newdata= seq6,type="response")
plot(typea,chd,col=typea,pch=19,cex=1.1,ylab="probabilty of Existence of
coronary heart disease")
lines(chd~typea, seq6 ,col=6,lwd=2)
```



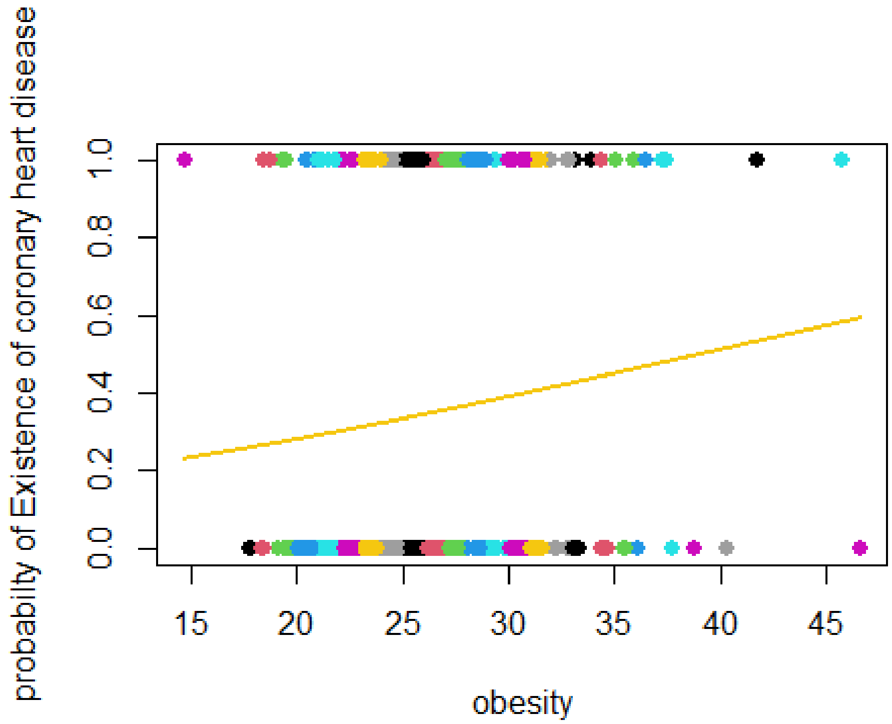
According to top plot we can see the probabilty of Existence of coronary heart disease for typea large or big values its more and more and we have posetive relationship . but i think that its not good variable for our multiple logsticks regression.

```
fit7<-glm(chd~obesity,family = binomial)
coef(fit7)

## (Intercept)     obesity
## -1.92831227  0.04941705

summary(fit7)

##
## Call:
## glm(formula = chd ~ obesity, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3396  -0.9257  -0.8558   1.4021   1.7116
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.92831    0.61692  -3.126  0.00177 **
## obesity      0.04942    0.02318   2.132  0.03302 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 596.11  on 461  degrees of freedom
## Residual deviance: 591.53  on 460  degrees of freedom
## AIC: 595.53
##
## Number of Fisher Scoring iterations: 4
```

According to the summary and coef function outputs we can say that we have weak posetive relationship (betha obesity = 0.04942) and our p-value for signifacting H0 (Betha0 = Betha1 = 0)is equal to 0.03302 and its less than alpha(0.05) so we say the H0 reject and the regression is signifact.

```
seq7<-data.frame(obesity=seq(min(obesity),max(obesity),len=10^4))
seq7$chd=predict(fit7,newdata= seq7,type="response")
plot(obesity,chd,col=obesity,pch=19,cex=1.1,ylab="probabilty of Existence of
coronary heart disease")
lines(chd~obesity, seq7 ,col=7,lwd=2)
```



According to top plot we can see the probabilty of Existence of coronary heart disease for obesity large or big values its more and more and we have posetive relationship. but i think that its not good variable for our multiple logsticks regression.

```
fit8<-glm(chd~alcohol,family = binomial)
coef(fit8)
```

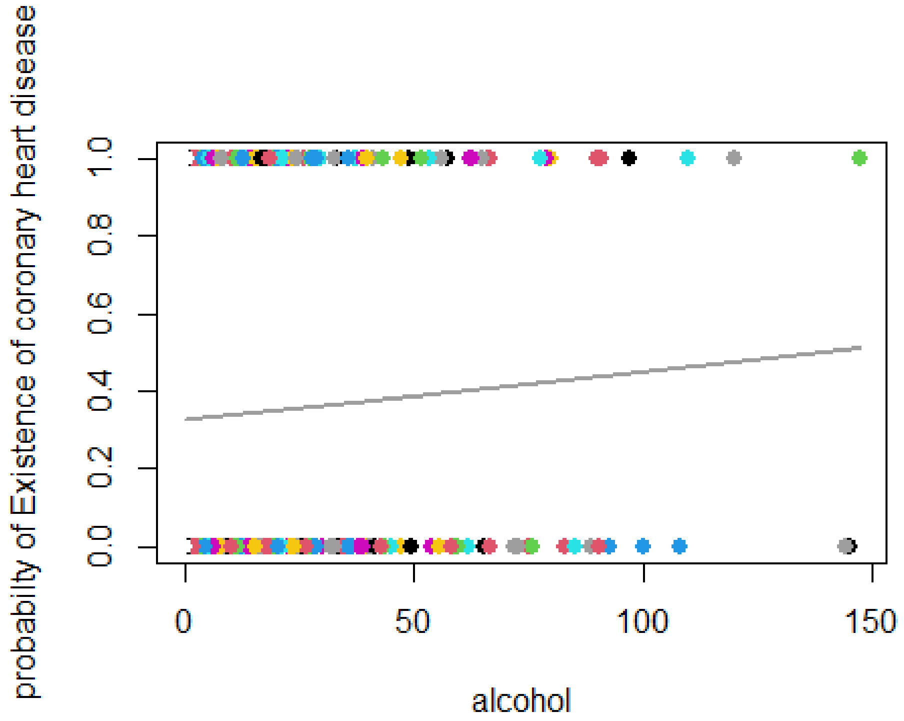
```
##  (Intercept)      alcohol 
## -0.726007546  0.005197845
```

```
summary(fit8)
```

```
## 
## Call:
## glm(formula = chd ~ alcohol, family = binomial)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.1898  -0.9104  -0.8884   1.4415   1.4971  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -0.726008   0.120004  -6.050 1.45e-09 ***
## alcohol      0.005198   0.003892   1.336    0.182    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 596.11  on 461  degrees of freedom
## Residual deviance: 594.35  on 460  degrees of freedom
## AIC: 598.35
## 
## Number of Fisher Scoring iterations: 4
```

According to the summary and coef function outputs we can say that we have weak posetive relationship (betha alcohol = 0.005198) and our p-value for signifacting H0 (Betha0 = Betha1 = 0)is equal to 0.182 and its more than alpha(0.05) so we say the H0 accept and the regression isnt signifact.

```
seq8<-data.frame(alcohol=seq(min(alcohol),max(alcohol),len=10^4))
seq8$chd=predict(fit8,newdata= seq8,type="response")
plot(alcohol,chd,col=alcohol,pch=19,cex=1.1,ylab="probabilty of Existence of
coronary heart disease")
lines(chd~alcohol, seq8 ,col=8,lwd=2)
```



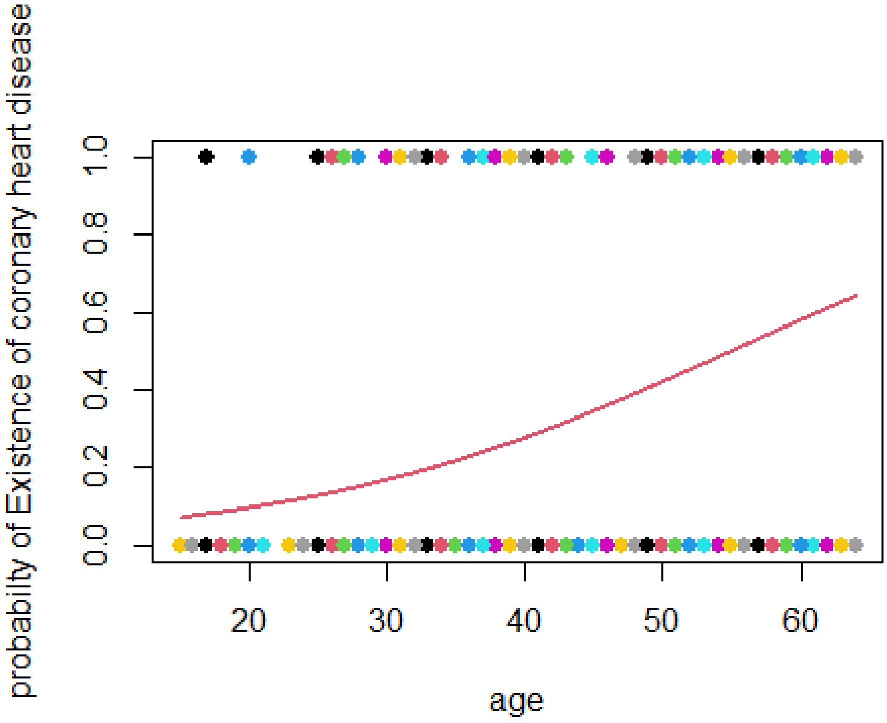
According to top plot we can see the probabilty of Existence of coronary heart disease for alcohol large or big values its more and more and we have posetive relationship. but i think that its not good variable for our multiple logsticks regression.

```
fit9<-glm(chd~age,family = binomial)
coef(fit9)

## (Intercept)         age
## -3.52171034  0.06410803

summary(fit9)

##
## Call:
## glm(formula = chd ~ age, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4321  -0.9215  -0.5392   1.0952   2.2433
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.521710   0.416031  -8.465  < 2e-16 ***
## age          0.064108   0.008532   7.513 5.76e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 596.11  on 461  degrees of freedom
## Residual deviance: 525.56  on 460  degrees of freedom
## AIC: 529.56
##
## Number of Fisher Scoring iterations: 4
```

According to the summary and coef function outputs we can say that we have weak posetive relationship (betha age = 0.064108) and our p-value for signifacting H0 (Betha0 = Betha1 = 0)is equal to 5.76e-14 and its less than alpha(0.05) so we say the H0 reject and the regression is signifact.

```
seq9<-data.frame(age=seq(min(age),max(age),len=10^4))
seq9$chd=predict(fit9,newdata= seq9,type="response")
plot(age,chd,col=age,pch=19,cex=1.1,ylab="probabilty of Existence of coronary
heart disease")
lines(chd~age, seq9 ,col=10,lwd=2)
```



According to top plot we can see the probabilty of Existence of coronary heart disease for age large or big values its more and more and we have posetive relationship , for example adiposity is more and equal 55 our probabilty is more than 0.5.

Logstics regression for all variable and response(Multiple logstic regression)

```
full.fit<-glm(chd~.,data=Heart,family = binomial)
coef(full.fit)
```

```
##   (Intercept)           sbp       tobacco           ldl     adiposity
## -6.1507208650  0.0065040171  0.0793764457  0.1739238981  0.0185865682
##       famhist         typea       obesity       alcohol           age
##  0.9253704194  0.0395950250 -0.0629098693  0.0001216624  0.0452253496
```

```
summary(full.fit)
```

```
##
## Call:
## glm(formula = chd ~ ., family = binomial, data = Heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7781  -0.8213  -0.4387   0.8889   2.5435
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.1507209  1.3082600  -4.701 2.58e-06 ***
## sbp          0.0065040  0.0057304   1.135 0.256374
## tobacco      0.0793764  0.0266028   2.984 0.002847 **
## ldl          0.1739239  0.0596617   2.915 0.003555 **
## adiposity    0.0185866  0.0292894   0.635 0.525700
## famhist      0.9253704  0.2278940   4.061 4.90e-05 ***
## typea        0.0395950  0.0123202   3.214 0.001310 **
## obesity     -0.0629099  0.0442477  -1.422 0.155095
## alcohol      0.0001217  0.0044832   0.027 0.978350
## age          0.0452253  0.0121298   3.728 0.000193 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 596.11  on 461  degrees of freedom
## Residual deviance: 472.14  on 452  degrees of freedom
## AIC: 492.14
##
## Number of Fisher Scoring iterations: 5
```

According to this models we can say that the (sbp,adiposity,obesity,alcohol) variables p-values is more than 0.05 and they are not good for our model and we should remove them, the others variable have positive relationships with response expect obesity.

Reduce logstics model

```
Reduce.fit<-glm(chd~tobacco+ldl+famhist+typea+age,data=Heart,family =
binomial)
coef(Reduce.fit)
```

```
## (Intercept)     tobacco         ldl     famhist       typea         age
## -6.44644451  0.08037533  0.16199164  0.90817526  0.03711521  0.05046038
```

```
summary(Reduce.fit)
```

```
##
## Call:
## glm(formula = chd ~ tobacco + ldl + famhist + typea + age, family =
binomial,
##     data = Heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9165  -0.8054  -0.4430   0.9329   2.6139
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.44644    0.92087  -7.000 2.55e-12 ***
## tobacco      0.08038    0.02588   3.106  0.00190 **
## ldl          0.16199    0.05497   2.947  0.00321 **
## famhist      0.90818    0.22576   4.023 5.75e-05 ***
## typea        0.03712    0.01217   3.051  0.00228 **
## age          0.05046    0.01021   4.944 7.65e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 596.11  on 461  degrees of freedom
## Residual deviance: 475.69  on 456  degrees of freedom
## AIC: 487.69
##
## Number of Fisher Scoring iterations: 5
```

We can see that the reduce models outputs show that the (sbp,adiposity,obesity,alcohol) variable were not important and dont have effects on response.

the Reduce model is best model without any bad variable and all of them are signifact and we have good predict for our response,

we can see the logstics predicton function here for reduce model:

y = chd , x1 = tobacco , x2 = ldl , x3 = famhist , x4 = typea , x5= age.

End.