67— title: “STAC67 Project” output: html_document date: “2023-04-01” —

STAC67 Project

Salman (1004991690)- Model Diagonostics and Presentation.

Billy (1007501080)- Background research, Presentation and Limitations.

Priyal (1007311703)- Model Building, Model Validation and Exploratory Data Analysis.

R Markdown

Abtract

Hypertension, or high blood pressure, is a common and serious health problem that affects millions of people worldwide. High blood pressure can lead to a number of health complications, including heart disease, stroke, and kidney failure. Although there are several factors that contribute to the development of hypertension, such as age, genetics, and lifestyle habits, the exact causes of hypertension are still not fully understood. Therefore, it is important to study the potential risk factors for hypertension, and to identify effective interventions to prevent or manage this condition.The Blood Pressure Data is a valuable resource for studying the risk factors for hypertension, and for evaluating the effectiveness of different interventions for hypertension prevention and management. The dataset includes a variety of variables, such as race, alcohol use, treatment status, body mass index, stress level, salt intake level, childbearing potential, income level, and education level, that have been shown to be associated with hypertension in previous research. By analyzing the relationships between these variables and blood pressure levels, we can gain insights into the complex mechanisms underlying hypertension, and identify potential targets for interventions.Moreover, the Blood Pressure Data is particularly valuable because it includes information on both treated and untreated hypertensive patients. This allows us to examine the effects of different treatment strategies on blood pressure levels, and to compare the effectiveness of different medications, lifestyle changes, and other interventions for hypertension management.Overall, the Blood Pressure Data has the potential to inform the development of effective strategies for hypertension prevention and management, and to improve the health outcomes of millions of people worldwide.

Variable Description

sbp : Systolic Blood Pressure.

gender : M = Male, F = Female.

married : Y = Married, N = Not Married.

smoke : Smoking Status, Y = Smoker, N = Non-Smoker.

exercise : Exercise level, 1 = Low, 2 = Medium, 3 = High.

age : Continuous variable (years).

weight : Continuous variable (lbs).

height : Continuous variable (inches).

overwt : Overweight, 1 = Normal, 2 = Overweight, 3 = Obese.

race : Categorical variable taking values 1, 2, 3, 4.

alcohol : Alcohol Use, 1 = Low, 2 = Medium, 3 = High.

trt : Treatment (for hypertension), Y = Treated, N = Untreated.

bmi : Body Mass Index, (Weight/Height^2) x 703.

stress : Stress Level, 1 = Low, 2 = Medium, 3 = High.

salt： Salt (NaCl) Intake Level， 1 = Low, 2 = Medium, 3 = High.

chldbear: Childbearing Potential, 1 = Male, 2 = Able Female, 3 = Unable Female.

income : Income Level, 1 = Low, 2 = Medium, 3 = High.

educatn : Education Level, 1 = Low, 2 = Medium, 3 = High.

```
library("readxl")
data = read_excel("BloodPressure.xlsx")
colnames(data)
```

```
##  [1] "sbp"      "gender"   "married"  "smoke"    "exercise" "age"
##  [7] "weight"   "height"   "overwt"   "race"     "alcohol"  "trt"
## [13] "bmi"      "stress"   "salt"     "chldbear" "income"   "educatn"
```

```
summary(data)
```

```
##       sbp           gender            married             smoke
##  Min.   : 67.0   Length:500         Length:500         Length:500
##  1st Qu.:130.0   Class :character   Class :character   Class :character
##  Median :140.5   Mode  :character   Mode  :character   Mode  :character
##  Mean   :145.0
##  3rd Qu.:162.2
##  Max.   :224.0
##     exercise          age           weight          height          overwt
##  Min.   :1.000   Min.   :18.0   Min.   : 90.0   Min.   :54.00   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:28.0   1st Qu.:133.0   1st Qu.:60.00   1st Qu.:1.000
##  Median :2.000   Median :40.0   Median :168.0   Median :65.00   Median :2.000
##  Mean   :1.948   Mean   :40.2   Mean   :166.6   Mean   :65.33   Mean   :2.034
##  3rd Qu.:3.000   3rd Qu.:52.0   3rd Qu.:198.0   3rd Qu.:70.00   3rd Qu.:3.000
##  Max.   :3.000   Max.   :64.0   Max.   :249.0   Max.   :77.00   Max.   :3.000
##       race          alcohol           trt             bmi
##  Min.   :1.000   Min.   :1.000   Min.   :0.000   Min.   :11.00
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.000   1st Qu.:21.00
##  Median :1.000   Median :2.000   Median :0.000   Median :27.00
##  Mean   :1.424   Mean   :2.026   Mean   :0.202   Mean   :27.66
##  3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:0.000   3rd Qu.:33.00
##  Max.   :4.000   Max.   :3.000   Max.   :1.000   Max.   :53.00
##      stress           salt          chldbear         income         educatn
##  Min.   :1.000   Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.00   1st Qu.:1.000   1st Qu.:1.000
##  Median :2.000   Median :2.000   Median :2.00   Median :2.000   Median :2.000
##  Mean   :2.046   Mean   :2.022   Mean   :1.77   Mean   :1.962   Mean   :1.998
##  3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:2.00   3rd Qu.:3.000   3rd Qu.:3.000
##  Max.   :3.000   Max.   :3.000   Max.   :3.00   Max.   :3.000   Max.   :3.000
```

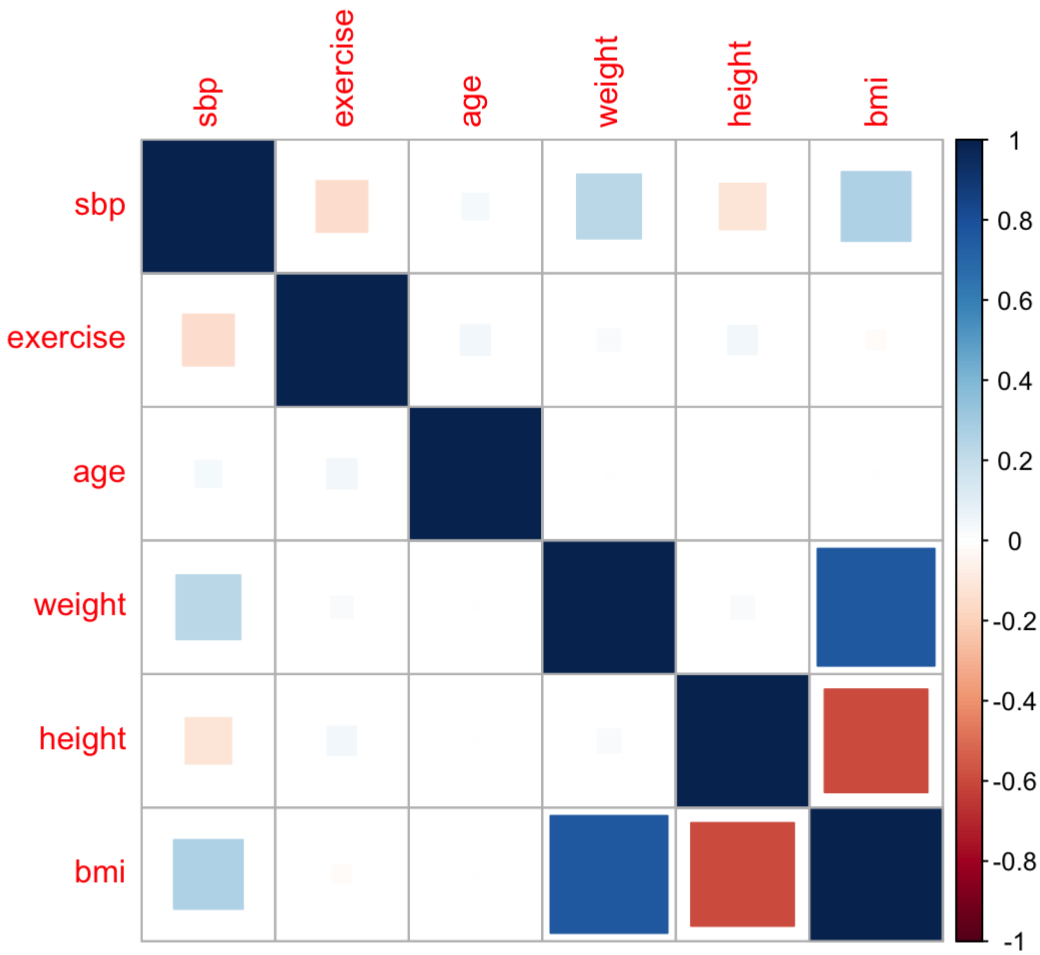
Analysis of Quantative Variables

Correlation plot for quantitative Variables

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
quant_var = c("sbp","exercise","age", "weight", "height", "bmi")
df1 <- data[ ,quant_var]
corrplot(cor(df1),method="square")
```




```
cor(df1)
```

```
##                 sbp    exercise          age       weight       height
## sbp       1.00000000 -0.14537399  0.037463336  0.230277555 -0.116917759
## exercise -0.14537399  1.00000000  0.047921023  0.025433338  0.044683669
## age       0.03746334  0.04792102  1.000000000 -0.002432779 -0.000918395
## weight    0.23027755  0.02543334 -0.002432779  1.000000000  0.028305097
## height   -0.11691776  0.04468367 -0.000918395  0.028305097  1.000000000
## bmi       0.26666927 -0.01782191  0.001822463  0.768325838 -0.594317652
##                  bmi
## sbp       0.266669272
## exercise -0.017821909
## age       0.001822463
## weight    0.768325838
## height   -0.594317652
## bmi       1.000000000
```

There is no significant problem of multi-collinearity between our quantitative variables.

#Coding Binary Variables as 0 and 1.

```
#qual_var = c("married","gender","smoke")
#df2 <- data[ ,qual_var]
data$married<-ifelse(data$married=="Y",1,0)
data$gender<-ifelse(data$gender=="F",1,0)
data$smoke<-ifelse(data$smoke=="Y",1,0)
data
```

```
## # A tibble: 500 × 18
##      sbp gender married smoke exercise   age weight height overwt  race alcohol
##    <dbl>  <dbl>   <dbl> <dbl>    <dbl> <dbl>  <dbl>  <dbl>  <dbl> <dbl>   <dbl>
##  1   133      1       0     0        3    60    159     56      3     1       2
##  2   115      0       0     1        1    55    107     65      1     1       2
##  3   140      0       0     1        1    18    130     59      2     1       1
##  4   132      0       1     0        2    19    230     57      3     2       3
##  5   133      0       0     0        2    58    201     74      2     1       3
##  6   138      1       0     0        3    55    166     67      2     1       1
##  7   133      1       1     0        1    22    188     66      3     1       3
##  8    67      1       1     0        3    52    123     67      1     1       2
##  9   138      0       1     0        1    46    106     73      1     1       3
## 10   130      0       1     1        3    38    166     72      1     1       1
## # … with 490 more rows, and 7 more variables: trt <dbl>, bmi <dbl>,
## #   stress <dbl>, salt <dbl>, chldbear <dbl>, income <dbl>, educatn <dbl>
```

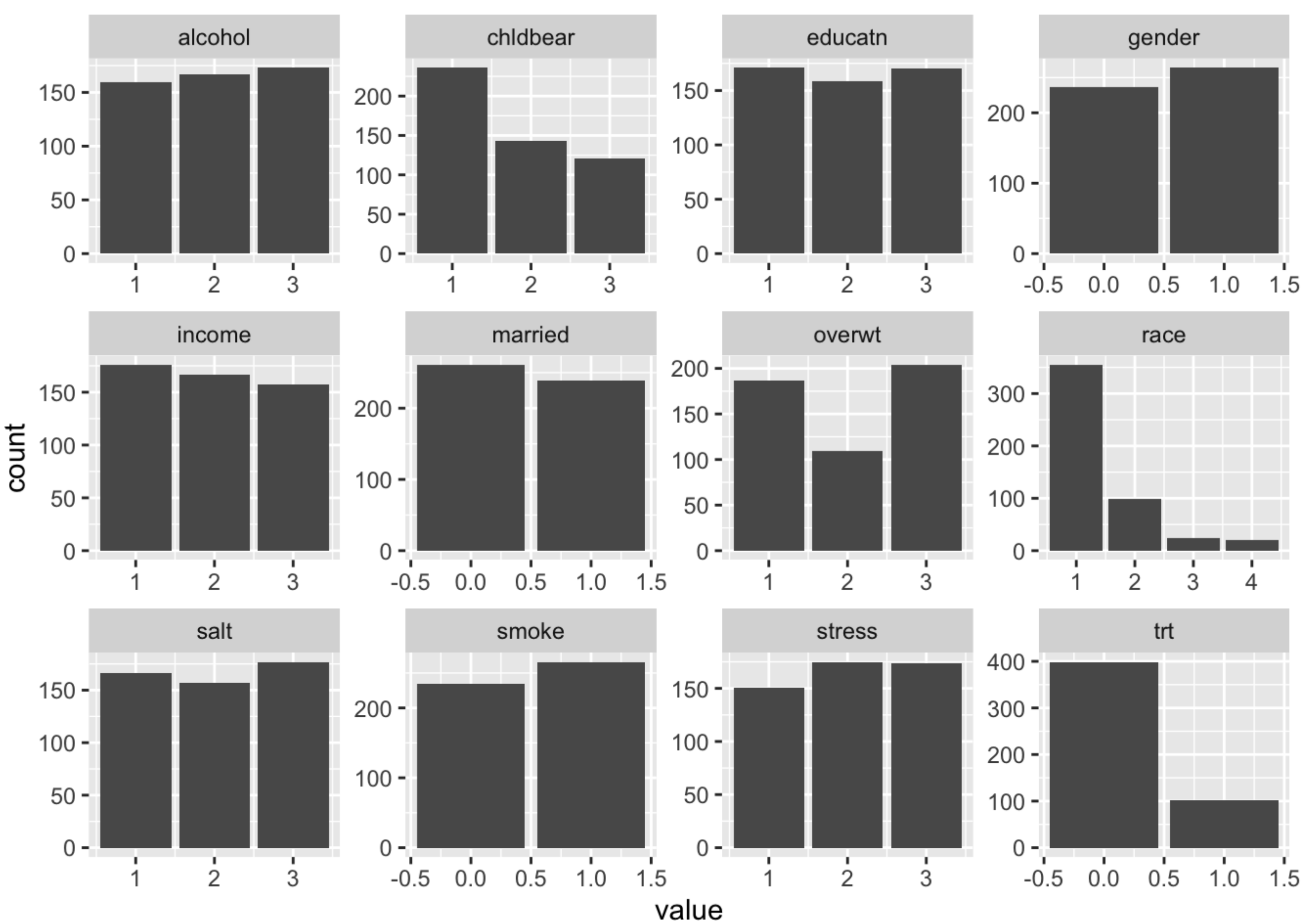
Analysis of Qualitative/Categorical Variables \

Does type of Qualitative determine if there is a systolic blood pressure (sbp) or not?

```
library(tidyverse)
```

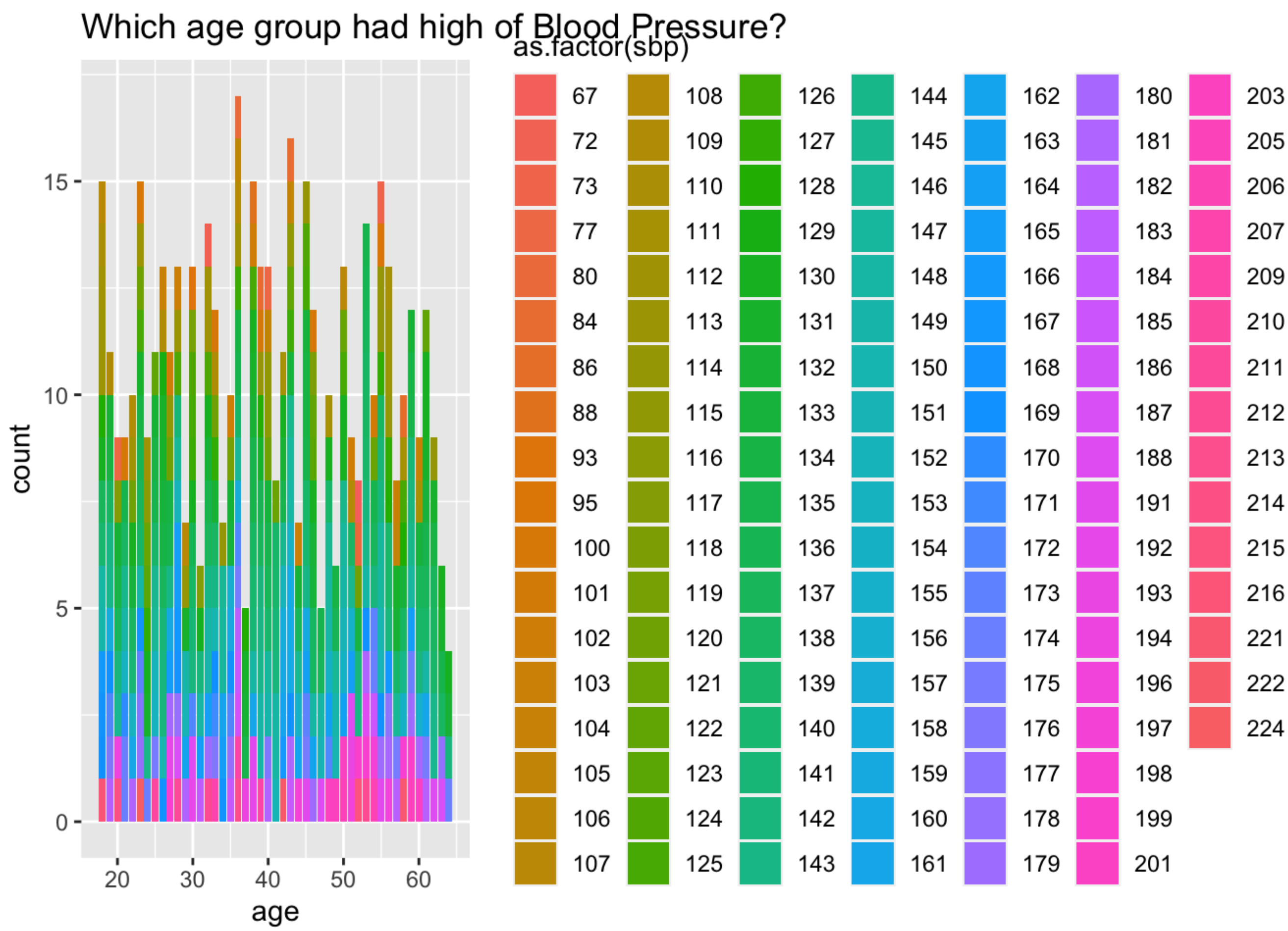
```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.1     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.1     ✔ tibble    3.2.1
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
qual_var = c("married","gender","smoke","overwt","race","alcohol","trt", "stress", "salt", "chldbear", "income", "educatn")
data[,qual_var] %>% gather() %>% ggplot(aes(value)) + facet_wrap(~key,scales = "free") + geom_bar()
```



Which age group had most number of relapse

```
df3 <- data
w = c(5,8,10,12,15,18,20,22,25,28,30,32,35,38,40,42,45,48,50,52,55,58,60 ,62,65,68,70,7
2,75,78,80,82,85,Inf)
df3$Age.Group = cut(data$age,breaks = w)
ggplot(df3) +
geom_bar(aes(x = age, fill = as.factor(sbp))) +
ggtitle("Which age group had high of Blood Pressure?")
```



It looks like that age between 35 to 40 had high blood pressure among others.

#Model Building.

1) Step AIC for Main effect model

```
quant_var = c("sbp","exercise","age", "weight", "height", "bmi")
qual_var = c("married","gender","smoke","overwt","race","alcohol","trt", "stress", "sal
t", "chldbear", "income", "educatn")

fit.simple <- lm(data$sbp ~ 1, data = data)

fit.complex <- lm(data$sbp ~ data$exercise + data$age+ data$weight+ data$height + data$b
mi + factor(data$married) + factor(data$gender) + factor(data$smoke) + factor(data$overw
t) + factor(data$race) + factor(data$alcohol) + factor(data$trt) + factor(data$stress) +
factor(data$salt) + factor(data$chldbear) + factor(data$income) + factor(data$educatn))
library(MASS)
stepAIC(fit.simple, scope = list(upper = fit.complex, lower = fit.simple), direction =
"both")
```

Let us validate our main effect model using AIC, BIC Rsq and AdjRsq

```
library("SciViews")
library(leaps)
allreg<- regsubsets((data$sbp) ~ data$bmi + factor(data$smoke) + factor(data$trt) +
    factor(data$alcohol) + data$exercise + data$height + data$age, nbest = 7, data = dat
a)
aprout = summary(allreg)
n = dim(data)[1]
pprime = apply(aprout$which, 1, sum)
aprout$aic <- aprout$bic - log((n))* pprime + 2 * pprime
df3<- with(aprout, round(cbind(which,rsq, adjr2, cp, bic, aic), 3))
```

Therefore, we end up choosing the best model with 7 terms -92.609 aic, -58.892 bic, cp = 7.132 is also near p'=p+1 that is 9. However, by looking at our r square and adjusted r square values. That is with 7 terms age = 1, height = 1, intercept =1, bmi = 1, smoke = 1, trt =1, factor(data$alcohol)2 = 0, factor(data$alcohol)3 = 1, exercise = 1.

2) Step AIC for Interaction model

```
quant_var = c("sbp","exercise","age", "weight", "height", "bmi")
qual_var = c("married","gender","smoke","overwt","race","alcohol","trt", "stress", "sal
t", "chldbear", "income", "educatn")

fit.simple.1 <- lm(data$sbp ~ 1, data = data)
fit.complex.1 <- lm(data$sbp ~ data$exercise * data$age* data$weight* data$height * data
$bmi * factor(data$married) * factor(data$gender) * factor(data$smoke) * factor(data$ove
rwt) * factor(data$race) * factor(data$alcohol) * factor(data$trt))
library(MASS)
stepAIC(fit.simple.1, scope = list(upper = fit.complex.1, lower = fit.simple.1), directi
on = "both")
```

Let us validate our interaction effect model using AIC, BIC Rsq and AdjRsq

```
#install.packages("SciViews") # ask whether ln or not
library("SciViews")
library(leaps)
allreg<- regsubsets(data$sbp ~ data$bmi + factor(data$smoke) + factor(data$trt) +
    factor(data$alcohol) + data$exercise + data$height + factor(data$married) +
    data$bmi:factor(data$trt) + data$bmi:data$exercise + factor(data$trt):factor(data$al
cohol) +
    factor(data$smoke):factor(data$trt) + factor(data$trt):data$exercise, nbest = 12, da
ta = data)
aprout = summary(allreg)
p.prime = apply(aprout$which, 1, sum)
aprout$aic <- aprout$bic - log(n)* p.prime + 2 * p.prime
df5<- with(aprout, round(cbind(which,rsq, adjr2, cp, bic, aic), 3))
```

Therefore, we end up choosing the best model with 8 terms and -110.179 aic, -72.248 bic, cp = 12.916 is also near p'=p+1. That is with 6 terms factor(data$trt)1 : data$exercise = 0, factor(data$smoke)1 : factor(data$trt) = 1, factor(data$trt)1 : factor(data$alcohol)3 = 1, factor(data$trt)1 : factor(data$alcohol)2 = 0, data$bmi : data$exercise = 1, data$bmi : factor(data$trt)1 =1, factor(data$married)1 = 0, data$height = 0, data$bmi = 1, factor(data$smoke)1 = 1, factor(data$trt)1 = 0, factor(data$alcohol)2 = 0, factor(data$alcohol)3 = 1, data$exercise = 1. So we decide upon a few interactions such as exercise and trt, bmi and exercise, bmi and trt. We remove marriage also.

3) Step AIC for Power model

```
quant_var = c("sbp","exercise","age", "weight", "height", "bmi")
qual_var = c("married","gender","smoke","overwt","race","alcohol","trt", "stress", "sal
t", "chldbear", "income", "educatn")
trt <- factor(data$trt)
fit.simple.2 <- lm(data$sbp ~ 1, data = data)
fit.complex.2 <- lm(data$sbp ~ data$exercise + data$bmi + factor(data$trt) +I(data$exerc
ise^2) +I(data$bmi^2) +
                      I(trt^2) + I(data$exercise * data$bmi) + I(data$exercise * trt) +
I(data$bmi * trt) + I(data$exercise * data$bmi*trt), data = data )
library(MASS)
stepAIC(fit.simple.2, scope = list(upper = fit.complex.2, lower = fit.simple.2), directi
on = "both")
```

Let us validate our interaction effect model using AIC, BIC Rsq and AdjRsq

```
library("SciViews")
library(leaps)
allreg<- regsubsets(data$sbp ~ data$exercise + data$bmi + factor(data$trt) +I(data$exerc
ise^2) +I(data$bmi^2) +
                      I(trt^2) + I(data$exercise * data$bmi) + I(data$exercise * trt) +
I(data$bmi * trt) + I(data$exercise * data$bmi*trt), nbest = 10, data = data)
aprout = summary(allreg)
p.prime = apply(aprout$which, 1, sum)
aprout$aic <- aprout$bic - log(n)* p.prime + 2 * p.prime
df5<- with(aprout, round(cbind(which,rsq, adjr2, cp, bic, aic), 3))
```

I(data$exercise * trt)I(data$bmi * trt) I(data$exercise * data$bmi) might be significant.

4) Let us try a combination model from the insight that we drew from the above models

```
quant_var = c("sbp","exercise","age", "weight", "height", "bmi")
qual_var = c("married","gender","smoke","overwt","race","alcohol","trt", "stress", "sal
t", "chldbear", "income", "educatn")
trt <- factor(data$trt)
fit.simple.2 <- lm(data$sbp ~ 1, data = data)

#main effect
fit.complex <- lm(data$sbp ~ data$exercise + data$age+ data$weight+ data$height + data$b
mi + factor(data$married) + factor(data$gender) + factor(data$smoke) + factor(data$overw
t) + factor(data$race) + factor(data$alcohol) + factor(data$trt) + factor(data$stress) +
factor(data$salt) + factor(data$chldbear) + factor(data$income) + factor(data$educatn))

#interaction effect
fit.complex.1 <- lm(data$sbp ~ data$bmi + factor(data$smoke) + factor(data$trt) + factor
(data$alcohol) + data$exercise + data$height + factor(data$married) + data$bmi:factor(da
ta$trt) + data$bmi:data$exercise + factor(data$trt):factor(data$alcohol) + factor(data$s
moke):factor(data$trt) + factor(data$trt):data$exercise,data = data)

#power effect
fit.complex.2 <- lm(data$sbp ~ data$exercise + data$bmi + factor(data$trt) +I(data$exerc
ise^2) +I(data$bmi^2) +
    I(trt^2) + I(data$exercise * data$bmi) + I(data$exercise * trt) + I(data$bmi * trt)
+ I(data$exercise * data$bmi*trt), data = data )

data$exercise = data$exercise - mean(data$exercise)
data$bmi = data$bmi - mean(data$bmi)
trt = trt - mean(trt)
```

```
## Warning in mean.default(trt): argument is not numeric or logical: returning NA
```

```
## Warning in Ops.factor(trt, mean(trt)): '-' not meaningful for factors
```

```
fit.final.complex <- lm(data$sbp ~ data$exercise + factor(data$alcohol) + factor(data$sm
oke) + data$bmi + factor(data$trt) + data$height +
    I(data$exercise^2) + data$exercise:data$bmi + data$exercise:factor(data$trt) + data
$bmi:factor(data$trt) + factor(data$alcohol):factor(data$trt) + factor(data$smoke):facto
r(data$trt) , data = data )

library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
stepAIC(fit.final.complex, scope = list(upper = fit.final.complex, lower = fit.simple.
2), direction = "both")
```

```
## Start:  AIC=3220.43
## data$sbp ~ data$exercise + factor(data$alcohol) + factor(data$smoke) +
##     data$bmi + factor(data$trt) + data$height + I(data$exercise^2) +
##     data$exercise:data$bmi + data$exercise:factor(data$trt) +
##     data$bmi:factor(data$trt) + factor(data$alcohol):factor(data$trt) +
##     factor(data$smoke):factor(data$trt)
##
##                                           Df Sum of Sq    RSS    AIC
## <none>                                                 295217 3220.4
## - data$exercise:factor(data$trt)          1    1877.8 297094 3221.6
## - factor(data$smoke):factor(data$trt)     1    2498.3 297715 3222.6
## - I(data$exercise^2)                      1    2593.5 297810 3222.8
## - factor(data$alcohol):factor(data$trt)  2    4003.0 299220 3223.2
## - data$height                             1    3093.3 298310 3223.6
## - data$exercise:data$bmi                  1    3273.8 298490 3223.9
## - data$bmi:factor(data$trt)               1    4799.2 300016 3226.5
```

```
##
## Call:
## lm(formula = data$sbp ~ data$exercise + factor(data$alcohol) +
##     factor(data$smoke) + data$bmi + factor(data$trt) + data$height +
##     I(data$exercise^2) + data$exercise:data$bmi + data$exercise:factor(data$trt) +
##     data$bmi:factor(data$trt) + factor(data$alcohol):factor(data$trt) +
##     factor(data$smoke):factor(data$trt), data = data)
##
## Coefficients:
##                            (Intercept)
##                                97.4142
##                          data$exercise
##                                -6.8144
##                  factor(data$alcohol)2
##                                 2.1086
##                  factor(data$alcohol)3
##                                15.4646
##                    factor(data$smoke)1
##                                13.6790
##                               data$bmi
##                                 1.3606
##                      factor(data$trt)1
##                                 2.5753
##                            data$height
##                                 0.5115
##                     I(data$exercise^2)
##                                 5.2443
##                 data$exercise:data$bmi
##                                 0.3648
##         data$exercise:factor(data$trt)1
##                                 5.5876
##              data$bmi:factor(data$trt)1
##                                -0.9723
## factor(data$alcohol)2:factor(data$trt)1
##                                -5.4389
## factor(data$alcohol)3:factor(data$trt)1
##                               -16.4322
##   factor(data$smoke)1:factor(data$trt)1
##                               -11.6243
```

Our final model consists of main effect terms, interactions terms and power term. We can conclude from the above that following interactions between the variables are significant and a 1 unit increase in the blood pressure could be due to the following interaction terms - \ excercise and bmi \ exercise and trt that is treatment for hypertension \ bmi and trt \ alcohol and trt\ smoke and trt\ as well as a power term of excercise.

Let us validate our final effect model using AIC, BIC Rsq and AdjRsq

```
library("SciViews")
library(leaps)
all.reg<- regsubsets(data$sbp ~ data$exercise + factor(data$alcohol) + factor(data$smok
e) + data$bmi + factor(data$trt) + data$height +
    I(data$exercise^2) + data$exercise:data$bmi + data$exercise:factor(data$trt) + data
$bmi:factor(data$trt) + factor(data$alcohol):factor(data$trt) + factor(data$smoke):facto
r(data$trt), nbest = 14, data = data)
approut = summary(all.reg)
p..prime = apply(approut$which, 1, sum)
n = dim(data)[1]
approut$aic <- approut$bic - log(n) * p..prime + 2 * p..prime
df9<- with(approut, round(cbind(which,rsq, adjr2, cp, bic, aic), 3))
```

Model Validation statistics

R_square and adjusted R_Square were calculated through model summaries

```
#fit full model
full_model <- lm(data$sbp ~ ., data = data)
library(olsrr)
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:MASS':
##
##     cement
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
ols_mallows_cp(fit.complex, full_model)
```

```
## [1] 11.32296
```

```
ols_mallows_cp(fit.complex.1, full_model)
```

```
## [1] -11.33609
```

```
ols_mallows_cp(fit.complex.2, full_model)
```

```
## [1] 37.94582
```

```
ols_mallows_cp(fit.final.complex, full_model)
```

```
## [1] -13.45559
```

```
AIC(fit.complex)
```

```
## [1] 4679.159
```

```
AIC(fit.complex.1)
```

```
## [1] 4643.663
```

```
AIC(fit.complex.2)
```

```
## [1] 4690.097
```

```
AIC(fit.final.complex)
```

```
## [1] 4641.367
```

```
BIC(fit.complex)
```

```
## [1] 4792.954
```

```
BIC(fit.complex.1)
```

```
## [1] 4711.096
```

```
BIC(fit.complex.2)
```

```
## [1] 4736.458
```

```
BIC(fit.final.complex)
```

```
## [1] 4708.8
```

```
#summary(fit.complex)
#summary(fit.complex.1)
#summary(fit.complex.2)
```

```
install.packages("MPV", repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/1k/y75_pwxj5jzgpc03v1l_d7nm0000gn/T//RtmprSfE6y/downloaded_packages
```

```
library(MPV)
```

```
## Loading required package: lattice
```

```
## Loading required package: KernSmooth
```

```
## KernSmooth 2.23 loaded
## Copyright M. P. Wand 1997-2009
```

```
## Loading required package: randomForest
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
##
## Attaching package: 'MPV'
```

```
## The following object is masked from 'package:olsrr':
##
##     cement
```

```
## The following object is masked from 'package:MASS':
##
##     cement
```

```
PRESS(fit.complex)
```

```
## [1] 337647.3
```

```
PRESS(fit.complex.1)
```

```
## [1] 310144.6
```

```
PRESS(fit.complex.2)
```

```
## [1] 342420.6
```

```
PRESS(fit.final.complex)
```

```
## [1] 308767.2
```

As per the trend, our r square and adjusted r squared are increasing, and AIC, BIC, and press values are decreasing hence satisfying our criteria.

Cross Validation -

```
#Spliting Data
bp.samp = sample(1:length(data$sbp),350,replace = FALSE)
#model building dataset
bp.cv.in = data[bp.samp,]
#validation dataset
bp.cv.out = data[-bp.samp,]
#fit model for training set (used final complex model)
fit.cv.in.complex = lm(bp.cv.in$sbp ~ exercise + factor(alcohol) + factor(smoke) + bmi +
factor(trt) + height + I(exercise^2) + exercise:bmi + exercise:factor(trt) +
    bmi:factor(trt) + factor(alcohol):factor(trt) +
    factor(smoke):factor(trt), data = bp.cv.in)

anova(fit.cv.in.complex)
```

```
## Analysis of Variance Table
##
## Response: bp.cv.in$sbp
##                              Df Sum Sq Mean Sq F value    Pr(>F)
## exercise                      1   4660  4660.3  7.8026  0.005517 **
## factor(alcohol)               2   4847  2423.5  4.0576  0.018148 *
## factor(smoke)                 1  17431 17431.4 29.1850 1.250e-07 ***
## bmi                           1  21475 21475.2 35.9554 5.227e-09 ***
## factor(trt)                   1   9980  9980.2 16.7095 5.458e-05 ***
## height                        1   2307  2306.7  3.8620  0.050217 .
## I(exercise^2)                 1   2080  2080.3  3.4829  0.062879 .
## exercise:bmi                  1   1644  1643.8  2.7521  0.098061 .
## exercise:factor(trt)          1    724   724.0  1.2122  0.271691
## bmi:factor(trt)               1   6404  6404.0 10.7220  0.001169 **
## factor(alcohol):factor(trt)   2   1730   864.9  1.4481  0.236485
## factor(smoke):factor(trt)     1   1927  1927.1  3.2265  0.073356 .
## Residuals                   335 200087   597.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

here MSE = 639.2 from the model-building dataset

```
##### Compute MSPE
fit.cv.out.complex = lm(bp.cv.out$sbp ~ exercise + factor(alcohol) + factor(smoke) + bmi
+ factor(trt) + height + I(exercise^2) + exercise:bmi + exercise:factor(trt) + bmi:facto
r(trt) + factor(alcohol):factor(trt) + factor(smoke):factor(trt), data = bp.cv.out)
pred.cv.out = predict(fit.cv.out.complex,bp.cv.out)
delta.cv.out = bp.cv.out$sbp[-bp.samp]-pred.cv.out
```

```
## Warning in bp.cv.out$sbp[-bp.samp] - pred.cv.out: longer object length is not a
## multiple of shorter object length
```

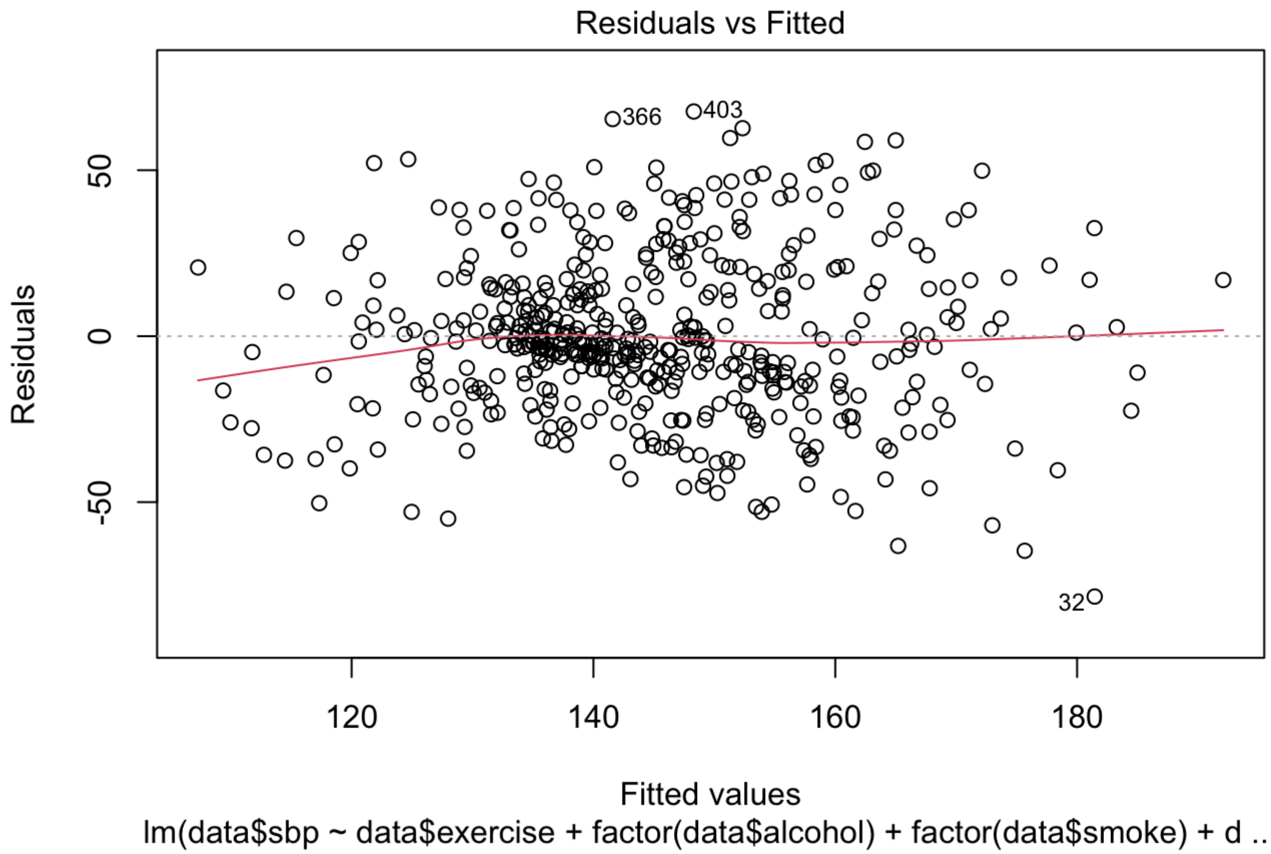
```
n.star = dim(bp.cv.out)[1]
MSPE <- sum((delta.cv.out)^2)/n.star
MSPE
```

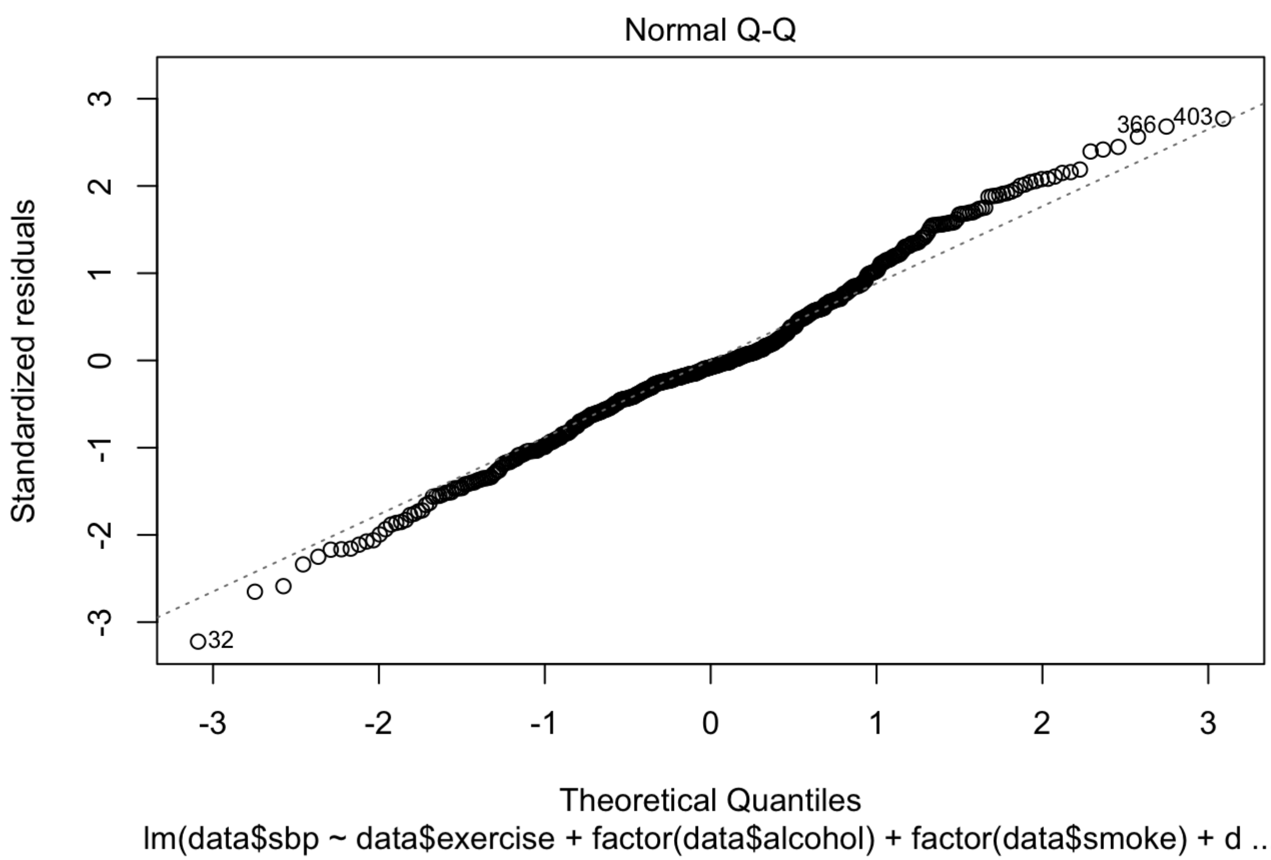
```
## [1] 758.9266
```

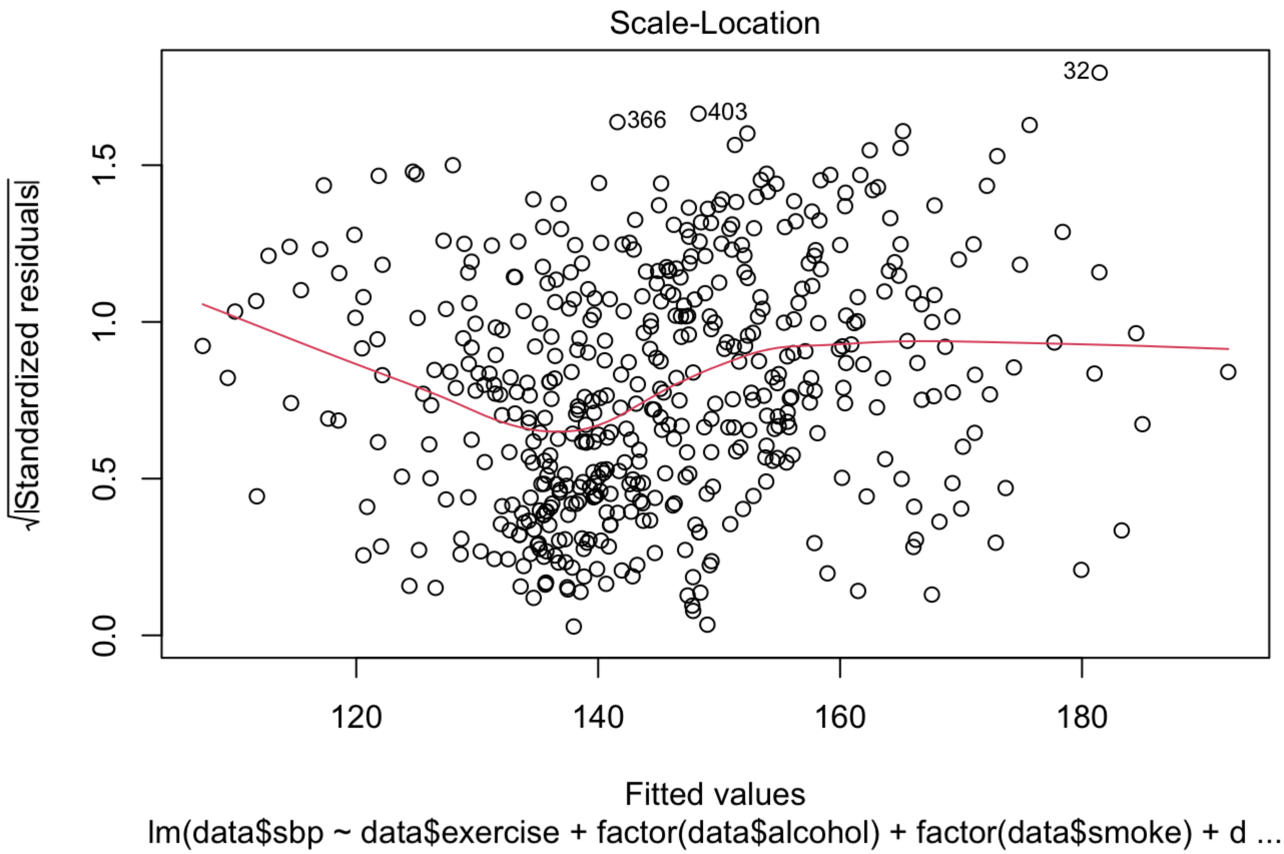
##here MSPE is = 659 , which is close to the MSE we got previously, hence we can validate the model.

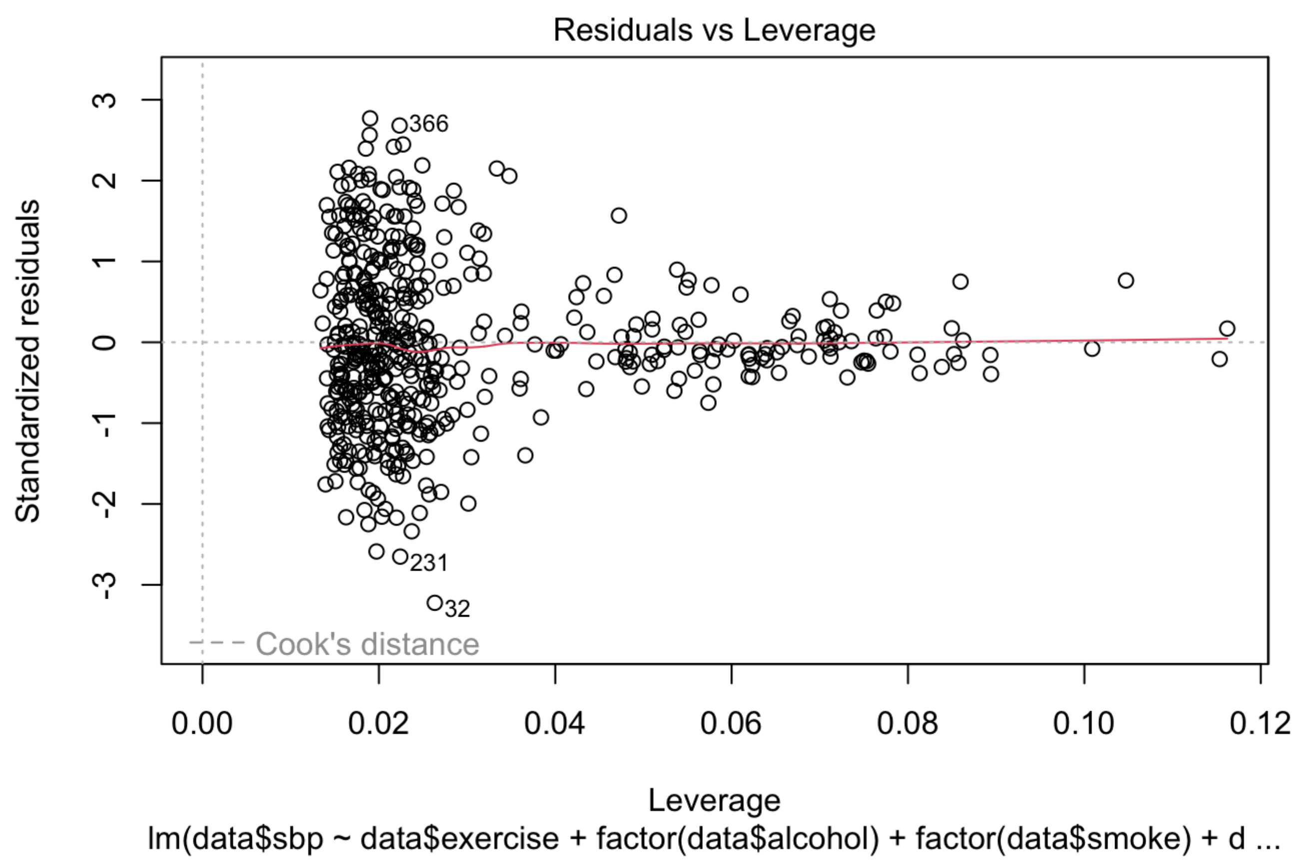
Model Diagnostics

```
plot(fit.final.complex)
```


Residuals vs Fitted
Residuals
50
0
-50
366
403
32
120
140
160
180
Fitted values
lm(data$sbp ~ data$exercise + factor(data$alcohol) + factor(data$smoke) + d ...


Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
-3
366
403
32
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(data$sbp ~ data$exercise + factor(data$alcohol) + factor(data$smoke) + d ...


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
120
140
160
180
Fitted values
lm(data$sbp ~ data$exercise + factor(data$alcohol) + factor(data$smoke) + d ...
32
366
403


Residuals vs Leverage
Standardized residuals
-3
-2
-1
0
1
2
3
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Leverage
lm(data$sbp ~ data$exercise + factor(data$alcohol) + factor(data$smoke) + d ...
366
231
32
Cook's distance

From the graphs above, we see that our model is fairly randomly scattered and therefore it satisfy the linearity assumption. Also, the QQ plot has a slight departure on the tail area but we know it can't be perfectly lined up and we could safely say our model follows a normal distribution. The Scale-location graph almost follow a horizontal line with the observation scattered randomly which give us a strong belief our model have equal error variances.

```
# Studentized deleted residuals for final model
t.final.complex = rstudent(fit.final.complex)
alpha = 0.05
n = dim(data)[1]
p.prime = length(coef(fit.final.complex))
t.final.complex.crit = qt(1-alpha/(2*n), n - p.prime -1)
t.final.complex.crit
```

```
## [1] 3.923262
```

```
which(abs(t.final.complex) > t.final.complex.crit)
```

```
## named integer(0)
```

From the code above, we see that there are no observations larger than the studentized residual. in other words, our model does not have outlying observations in terms of Y.

```
# Outlying X observations for final model
hii.final.complex = hatvalues(fit.final.complex)
which(hii.final.complex > 2*p.prime/n)
```

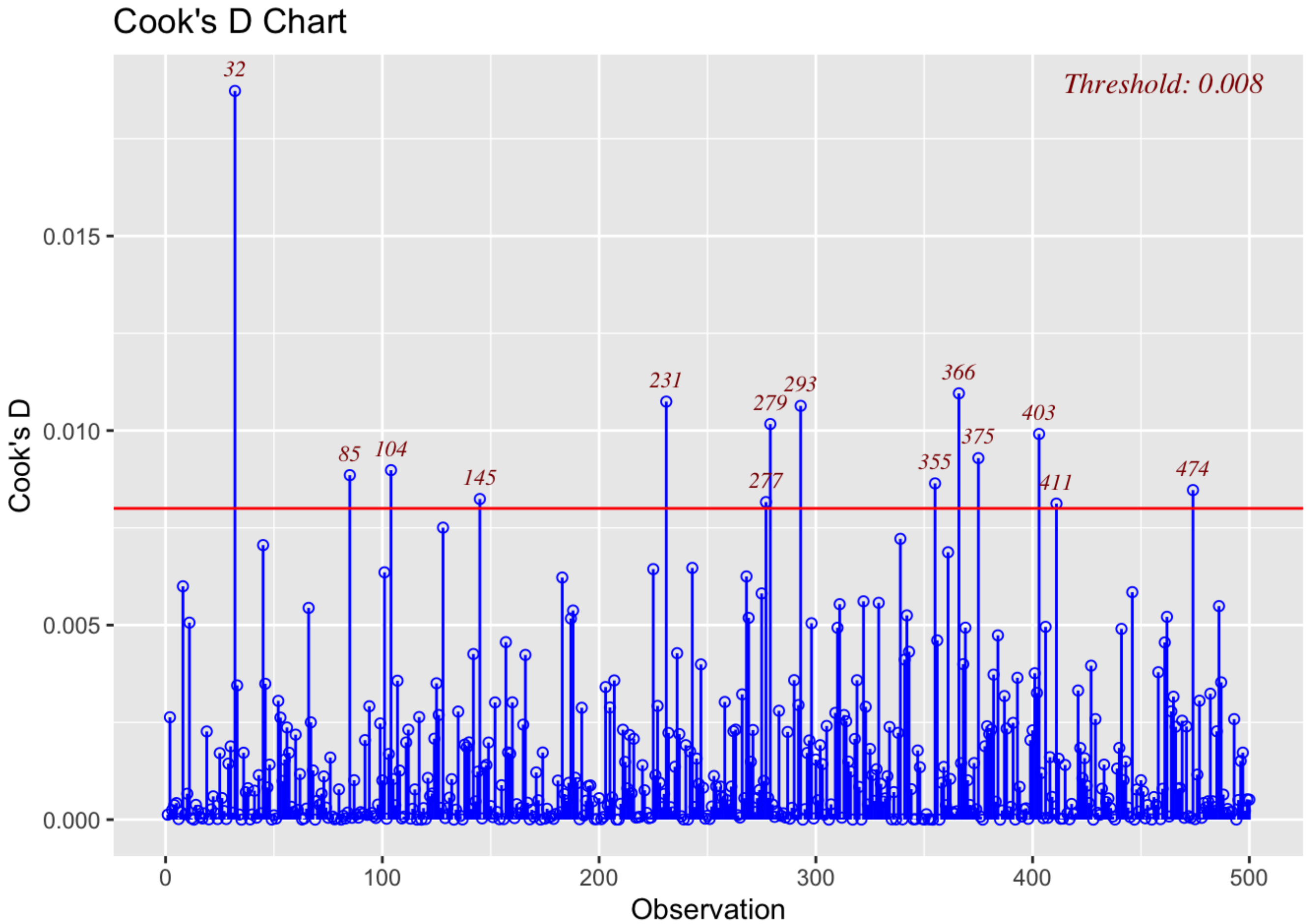
```
##   4   6   9  10  14  15  23  28  39  40  54  58  61  69  80  83  84  86  87  91
##   4   6   9  10  14  15  23  28  39  40  54  58  61  69  80  83  84  86  87  91
##  95 102 113 116 118 130 131 136 150 156 163 170 176 182 193 195 209 210 219 234
##  95 102 113 116 118 130 131 136 150 156 163 170 176 182 193 195 209 210 219 234
## 241 248 249 250 254 340 373 415 416 437 439 460 461 484 487 495 496
## 241 248 249 250 254 340 373 415 416 437 439 460 461 484 487 495 496
```

We see that there are 39 observations that are considered to be outliers in terms of the value of X.

```
# Influential observations for the final model
influence.measures(fit.final.complex)
```

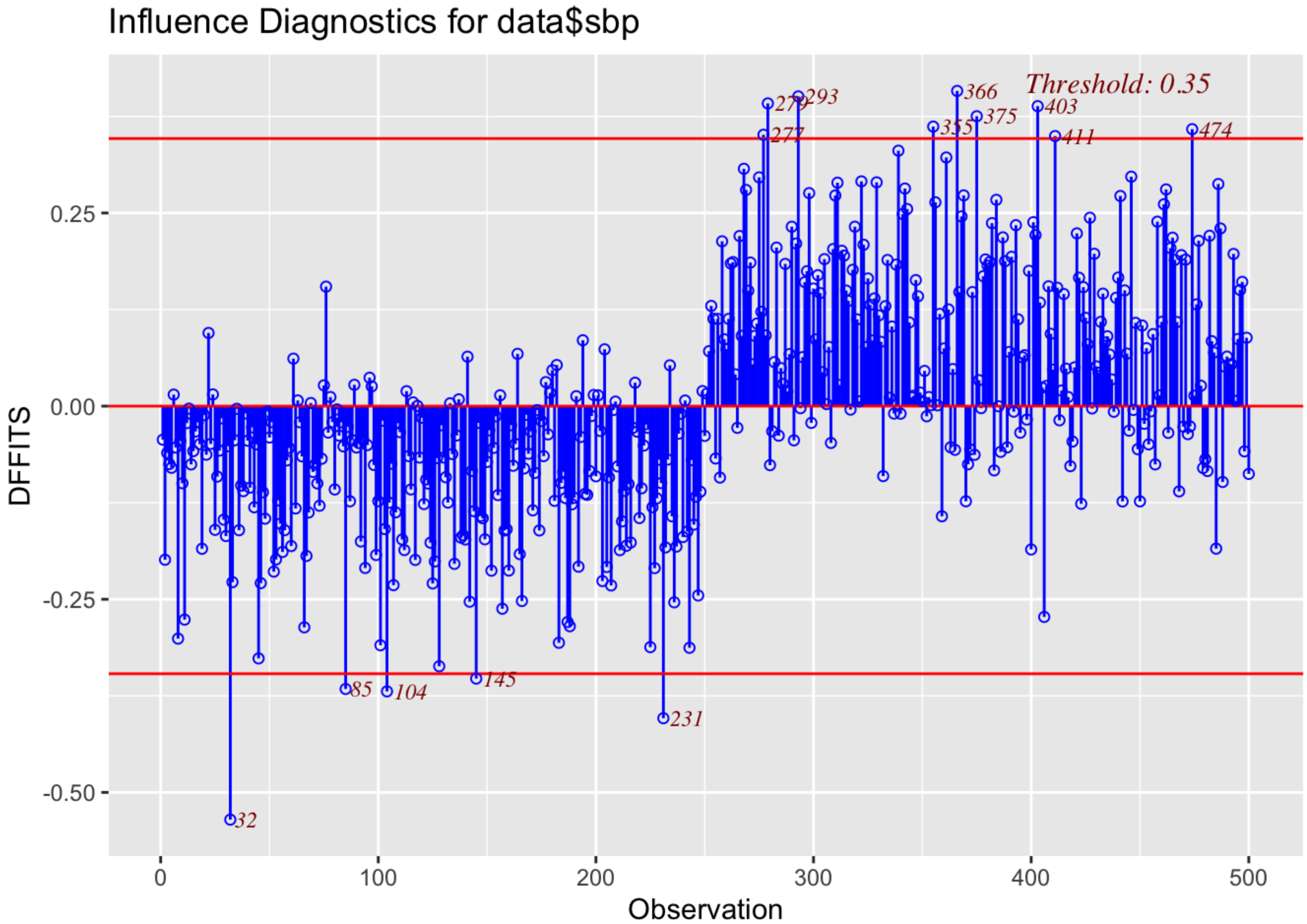
from above, we can see that the function "influence.measures" susbect but not necessary true that there are "*" 72 influencing observations. that is, there are 72 observations that influence the slope of the model.

```
# Graphical Diagonistics
library(olsrr)
library(ggpubr)
ols_plot_cooksd_chart(fit.final.complex)
```



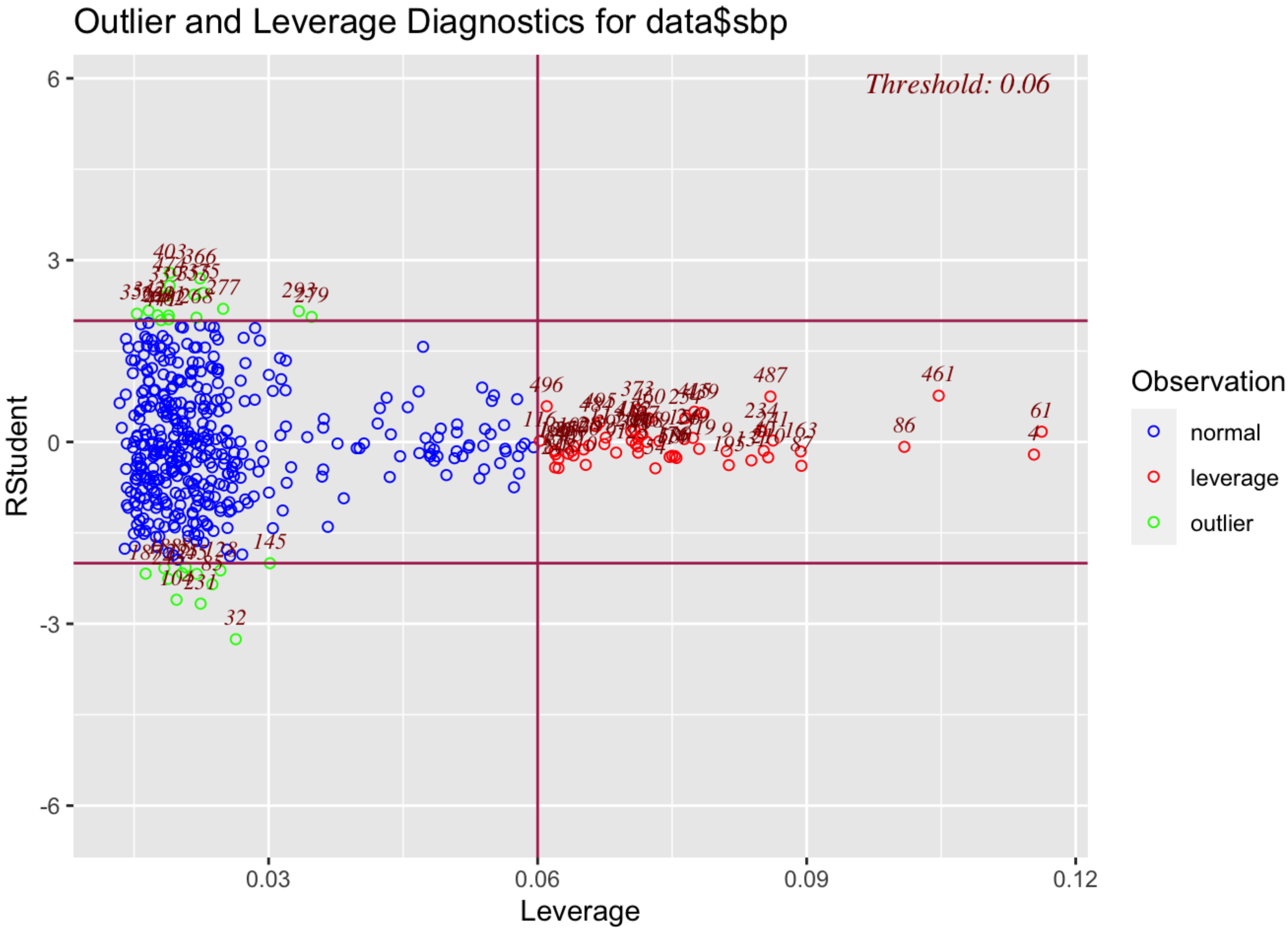
The Cook's distance graph above has a threshold of 4/n = 4/500 = 0.008. It shows that observation 32 is extremely influencing the model and it also shows other observation that are close to the threshold and might be influencing as well

```
ols_plot_dffits(fit.final.complex)
```



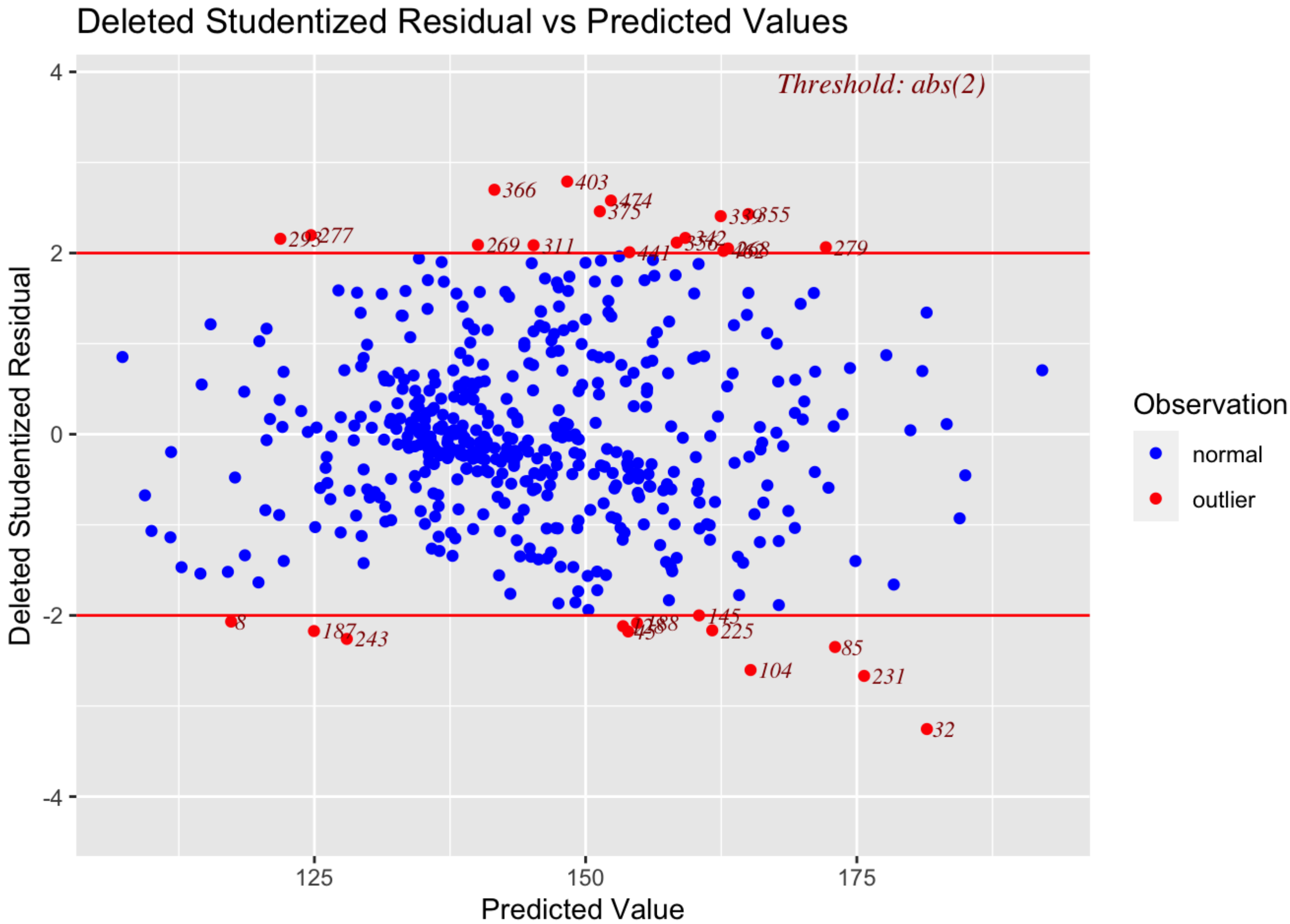
The graph above has a threshold of 2*sqrt(p'/n) = 0.37 and ot also agrees with cook's distance graph that observation 32 highly influence the model.

```
ols_plot_resid_lev(fit.final.complex)
```



This graph shows the suspected outliers observations in terms of X.

```
ols_plot_resid_stud_fit(fit.final.complex)
```



This graph shows the suspected outliers observations in terms of Y.