Homework 2

YuanJian Zhou

Oct 25 2019

Contents

Question 1

Import necessary packages

```
library(tidyverse) # for ggplot
```

```
## Warning in as.POSIXlt.POSIXct(Sys.time()): unable to identify current timezone 'U':
## please set environment variable 'TZ'
```

```
library(car) # for EDA plots
library(corrplot) # for correlation plots
library(leaps) # for subset regression
library(lmtest) # for ramsey test
library(DAAG) # for cross-validation
library(gridExtra) # for combination of ggplots
library(broom) # better output
library(MASS) # for stepAIC
library(multcomp) # for glht
library(emmeans) # for ANOVA
library(effects)
```

Import data

```
setwd("C:/Users/zyj37/Desktop/MAE/ECON 430")
h = read.csv("Homework/2/heart.csv")
names(h) = c("age","sex","chest_pain","bps","chol",
             "bloodsugar","ecg","max_heartrate","exercise_angina","oldpeak",
             "slope","n_major_vessel","thal","target")
h2 = h # keep the original dataset unchanged
```

Clean the data and change category variables to factor ones

```
# The following codes are based on the data description on the UCI website and the comment named "The u
h2 = subset(h2,(h2$n_major_vessel!=4 & h2$thal!=0))
h2 = h2 %>%
  mutate(sex_f = factor(sex,labels =c("female","male")),
         chest_pain_f = factor(chest_pain,levels=c(0,2,1,3),
         labels=c("Asymptomatic","Non_Angina","Atypical_Angina","Typical_Angina")),
         bloodsugar_f = factor(bloodsugar,labels=c("<120",">120")),
         ecg_f = factor(ecg,levels=c(0,1,2),labels=c("Left_Hypertrophy","Normal","Abnormality")),
         exercise_angina_f = factor(exercise_angina,labels=c("No","Yes")),
         slope_f = factor(slope,levels=c(2,1,0),
                          labels=c("Up","Flat","Down")),
         thal_f = factor(thal,levels=c(2,1,3),
                         labels=c("Normal","Fix","Reversable")),
         n_major_vessel_f = factor(n_major_vessel),
         target_f = factor(target,labels=c("Yes","No")))
```

(a)

Five numbers summary

```
summary(h2[,-c(2,3,6,7,9,11,12,13,14)])
```

```
##       age             bps             chol       max_heartrate
```

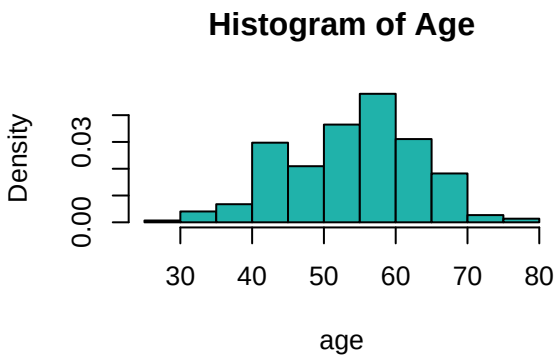
```
##  Min.   :29.00   Min.   : 94.0   Min.   :126.0   Min.   : 71.0
##  1st Qu.:48.00   1st Qu.:120.0   1st Qu.:211.0   1st Qu.:133.0
##  Median :56.00   Median :130.0   Median :242.5   Median :152.5
##  Mean   :54.52   Mean   :131.6   Mean   :247.2   Mean   :149.6
##  3rd Qu.:61.00   3rd Qu.:140.0   3rd Qu.:275.2   3rd Qu.:166.0
##  Max.   :77.00   Max.   :200.0   Max.   :564.0   Max.   :202.0
##     oldpeak         sex_f              chest_pain_f bloodsugar_f
##  Min.   :0.000   female: 95   Asymptomatic   :141   <120:253
##  1st Qu.:0.000   male  :201   Non_Angina     : 83   >120: 43
##  Median :0.800                Atypical_Angina: 49
##  Mean   :1.059                Typical_Angina : 23
##  3rd Qu.:1.650
##  Max.   :6.200
##               ecg_f     exercise_angina_f slope_f           thal_f
##  Left_Hypertrophy:145   No :199           Up  :138   Normal    :163
##  Normal           :147   Yes: 97           Flat:137   Fix       : 18
##  Abnormality      :  4                     Down: 21   Reversable:115
##
##
##
##  n_major_vessel_f target_f
##  0:173            Yes:136
##  1: 65            No :160
##  2: 38
##  3: 20
##
##
```

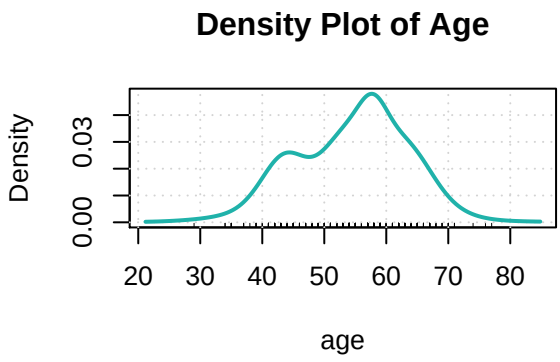
- We can see the summaries for each variable. I just refer to several of these.The mean of the age is about 54 years old with a maxmium of 77 and a minimum of 29; The composition of sex is a little unbalanced; Some of these variables are so technical that I can not understand the meaning even after googling.

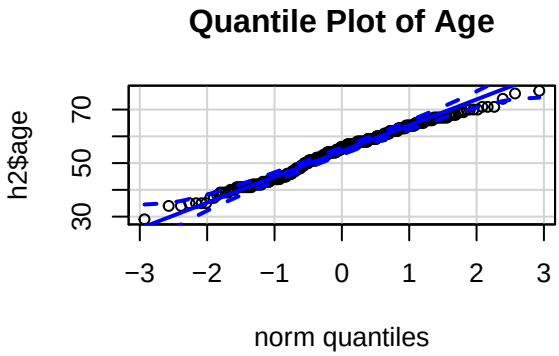
Explorary Descriptive Analysis for each variable

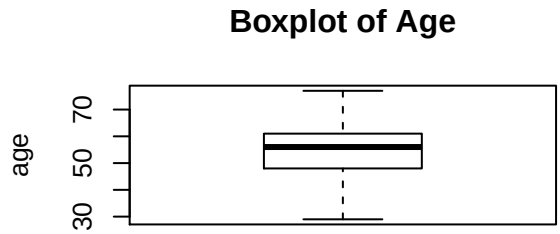
Age

```
par(mfrow=c(2,2))
age_hist = hist(h2$age,freq=F,breaks="FD",main="Histogram of Age",col="lightseagreen",xlab="age")
densityPlot(~age,data=h2,main="Density Plot of Age",col="lightseagreen")
qqPlot(h2$age,main="Quantile Plot of Age",id=F)
Boxplot(~age,data=h2,main="Boxplot of Age",id=F)
```





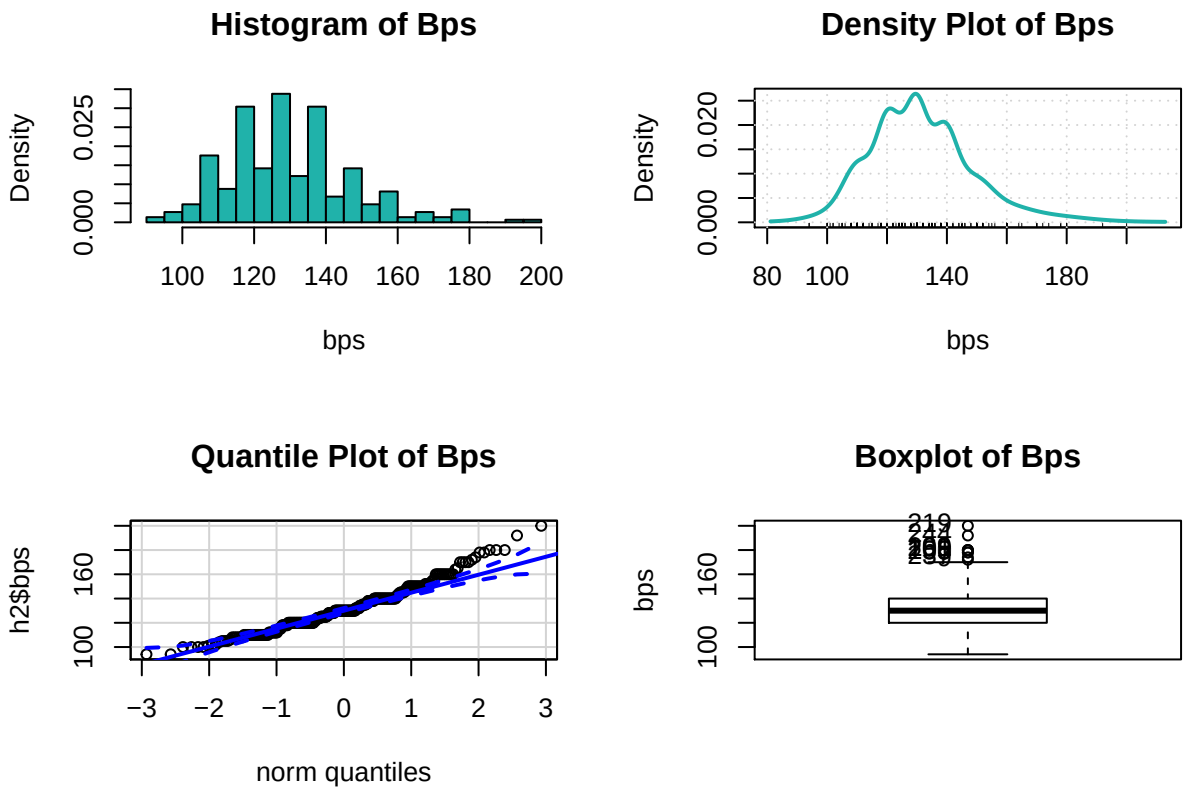




- We can see from these graphs that the variable is not so bad in normality. One problem is that it has two relative peaks. From the boxplot we can know that there are no outliers and it looks symmetric.

bps

```
par(mfrow=c(2,2))
bps_hist = hist(h2$bps,freq=F,breaks="FD",main="Histogram of Bps",col="lightseagreen",xlab="bps")
densityPlot(~bps,data=h2,main="Density Plot of Bps",col="lightseagreen")
qqPlot(h2$bps,main="Quantile Plot of Bps",id=F)
Boxplot(~bps,data=h2,main="Boxplot of Bps")
```

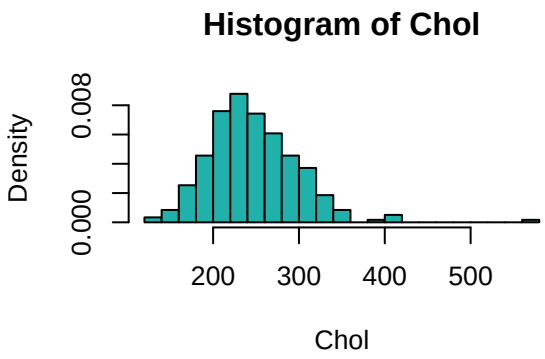


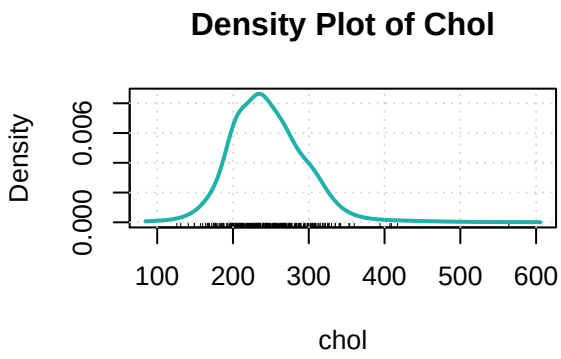

```
## [1] "9"   "100" "109" "199" "219" "237" "244" "255" "261"
```

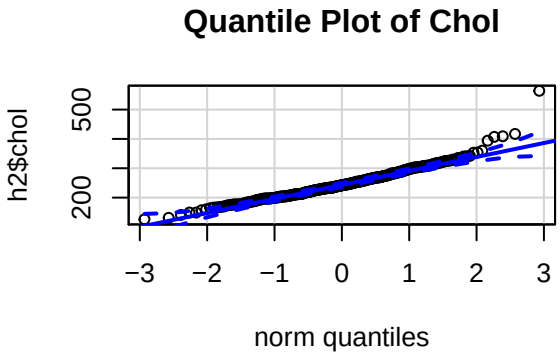
- We can see from the histogram that the varible is not so well normally distributed and the density plot shows that it is skewed. From the quantitle plot and boxplot we can know that there are a lot of outliers, while the problem we want to figure out is that whether they have heart disease, so may be these outliers are useful since people with higher bps will be more likely to get a heart disease.

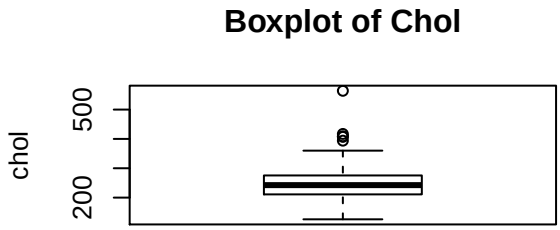
chol

```
par(mfrow=c(2,2))
chol_hist = hist(h2$chol,freq=F,breaks="FD",main="Histogram of Chol",col="lightseagreen",xlab="Chol")
densityPlot(~chol,data=h2,main="Density Plot of Chol",col="lightseagreen")
qqPlot(h2$chol,main="Quantile Plot of Chol",id=F)
Boxplot(~chol,data=h2,main="Boxplot of Chol",id=F)
```





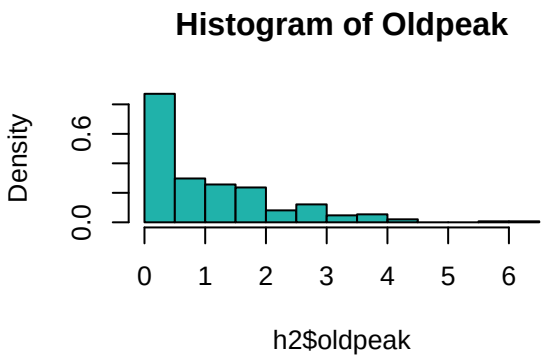


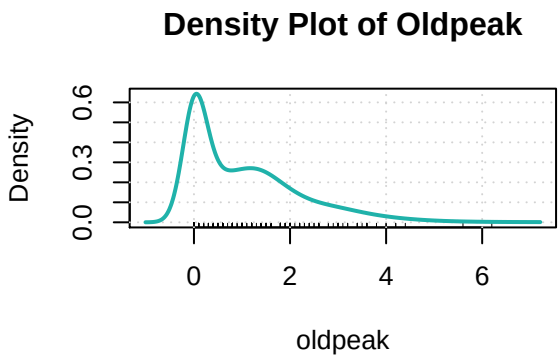


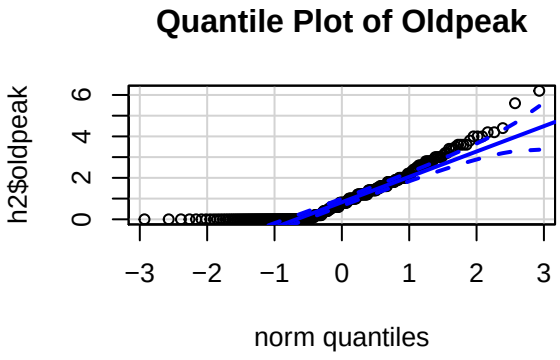
- Histogram of this variable looks so well, but the other three ones shows that there are a lot of outliers with relative big values, comments of these are as above.

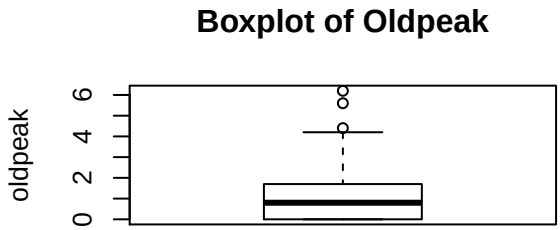
oldpeak

```
par(mfrow=c(2,2))
oldpeak_hist = hist(h2$oldpeak,freq=F,breaks="FD",main="Histogram of Oldpeak",col="lightseagreen")
densityPlot(~oldpeak,data=h2,main="Density Plot of Oldpeak",col="lightseagreen")
qqPlot(h2$oldpeak,main="Quantile Plot of Oldpeak",id=F)
Boxplot(~oldpeak,data=h2,main="Boxplot of Oldpeak",id=F)
```





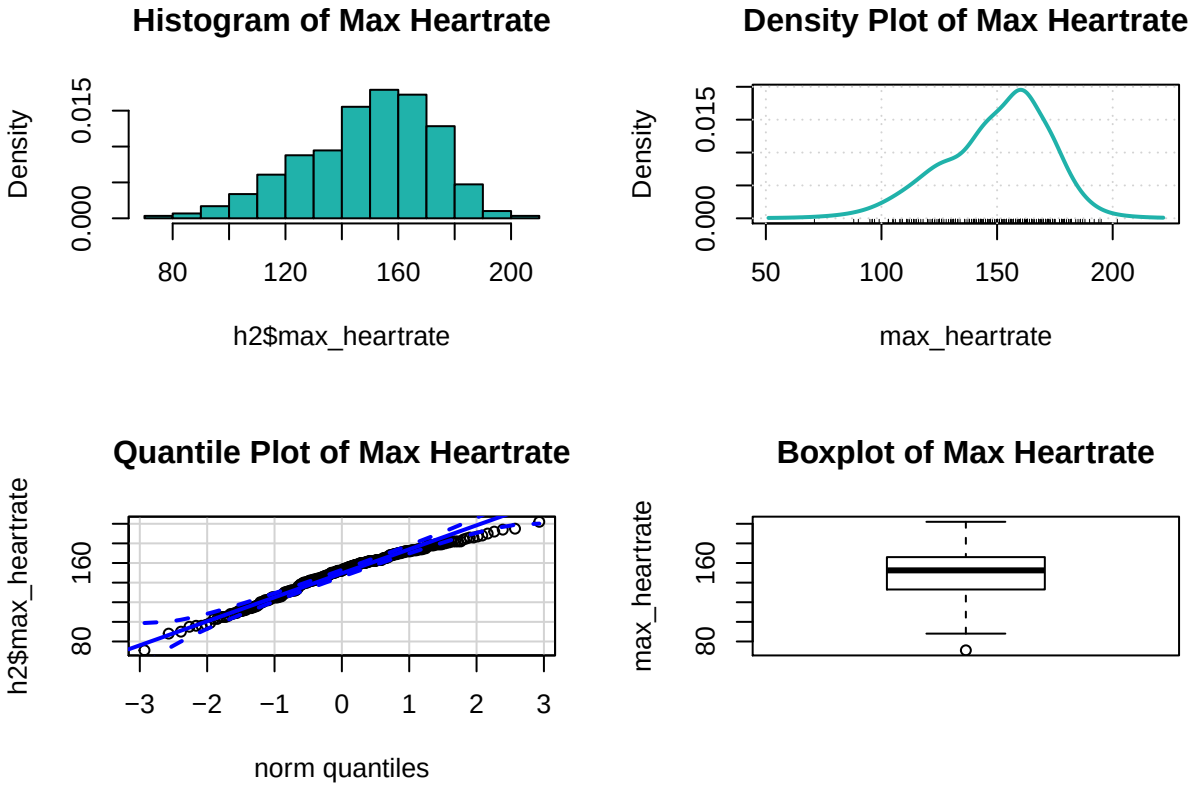




- Every graph of these graphs looks not so good, I think it is becasue the range of it is relatively small and it has too many 0 values. So it is not normally distributed and have lots of outliers.

max_heartrate

```
par(mfrow=c(2,2))
hist_max_heartrate = hist(h2$max_heartrate,freq=F,breaks="FD",main="Histogram of Max Heartrate",col="li
densityPlot(~max_heartrate,data=h2,main="Density Plot of Max Heartrate",col="lightseagreen")
qqPlot(h2$max_heartrate,main="Quantile Plot of Max Heartrate",id=F)
Boxplot(~max_heartrate,data=h2,main="Boxplot of Max Heartrate",id=F)
```

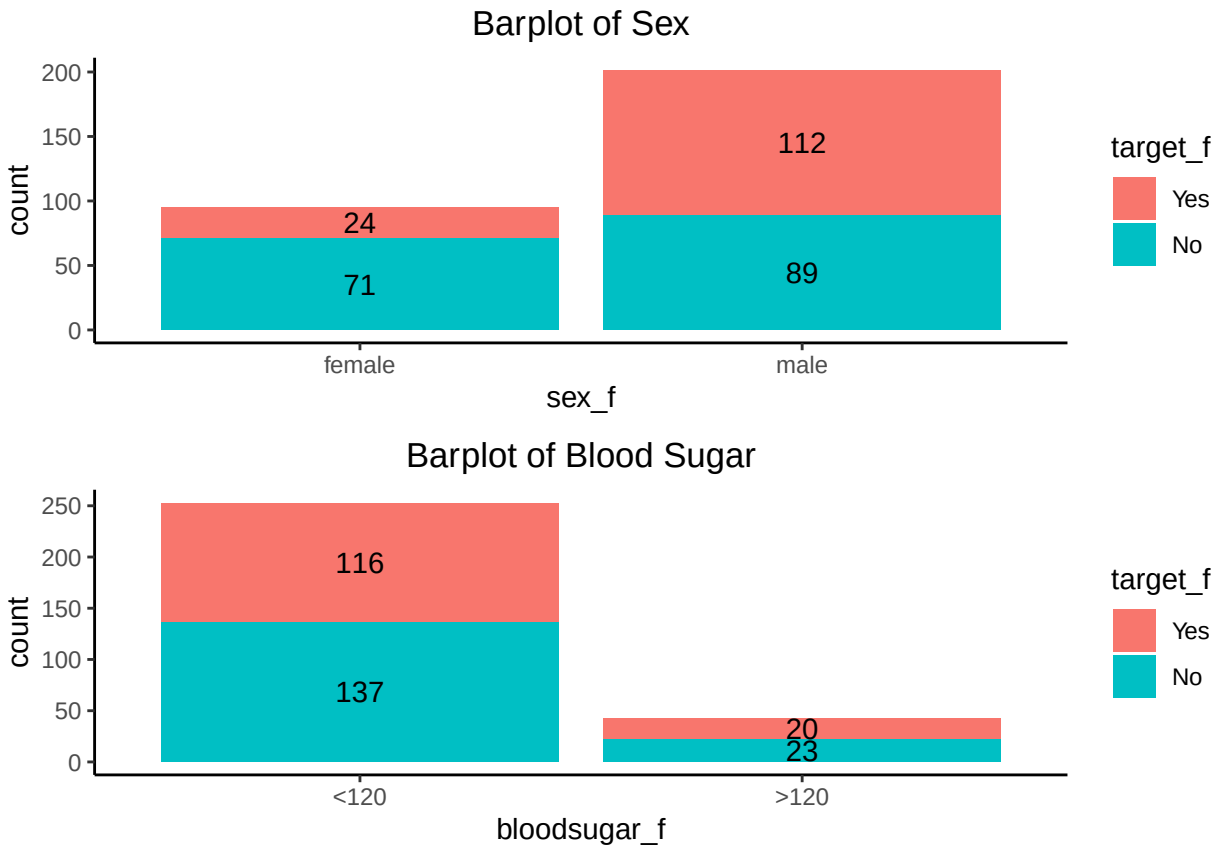


- This variable is distrbuted not so bad. The main problem is that it's skewed, but the quantile plot and boxplot look quite well.

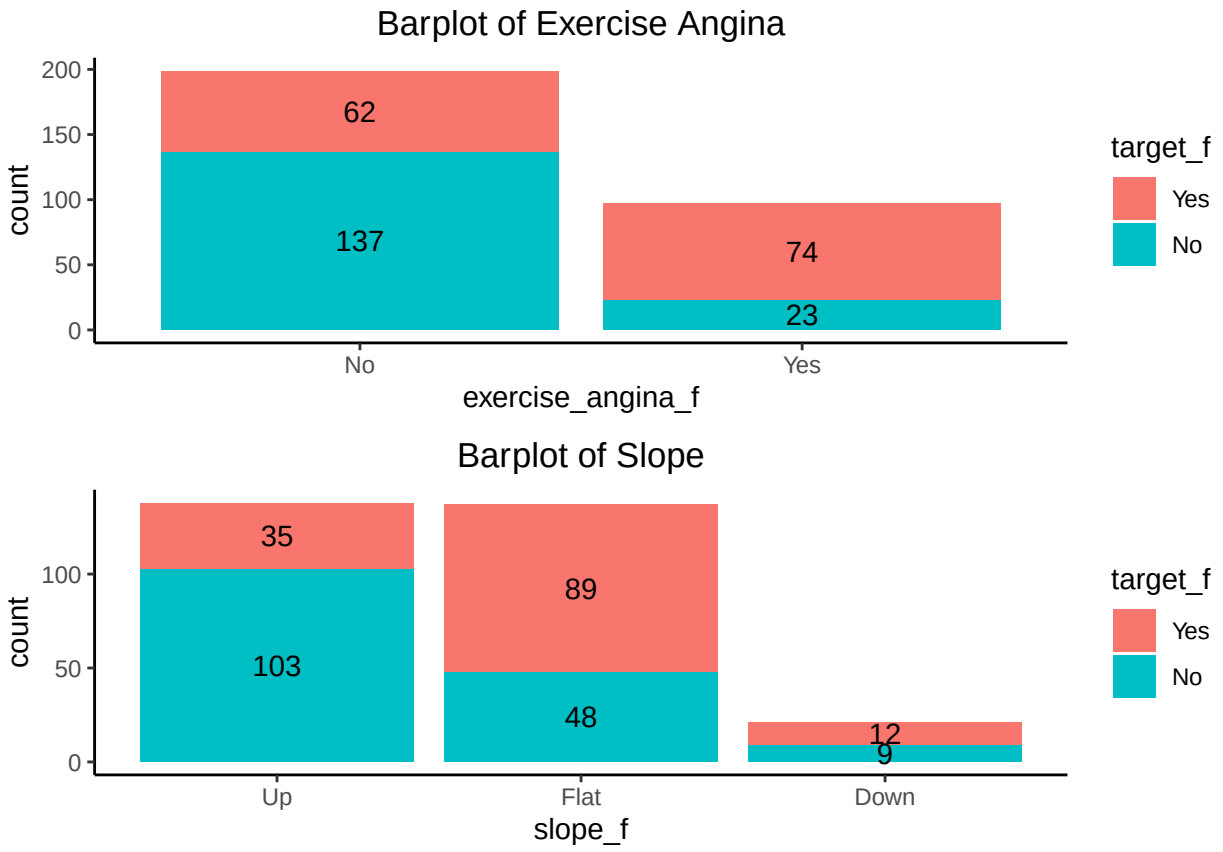
category variables

```
## sex
g1 = ggplot(h2,aes(sex_f,fill=target_f))+
  geom_bar()+
  geom_text(stat = "count",aes(label =..count..),position=position_stack(0.5))+
  theme_classic()+
  ggtitle("Barplot of Sex")+
  theme(plot.title = element_text(hjust = 0.5))
## chest_pain
g2 = ggplot(h2,aes(x =chest_pain_f,fill=target_f)) +
  geom_bar()+
  geom_text(stat = "count",aes(label =..count..),position=position_stack(0.5))+
  theme_classic()+
  ggtitle("Barplot of Chest Pain")+
  theme(plot.title = element_text(hjust = 0.5))
## blood sugar
g3 = ggplot(h2,aes(x =bloodsugar_f,fill=target_f)) +
  geom_bar()+
  geom_text(stat = "count",aes(label =..count..),position=position_stack(0.5))+
  theme_classic()+
  ggtitle("Barplot of Blood Sugar")+
```

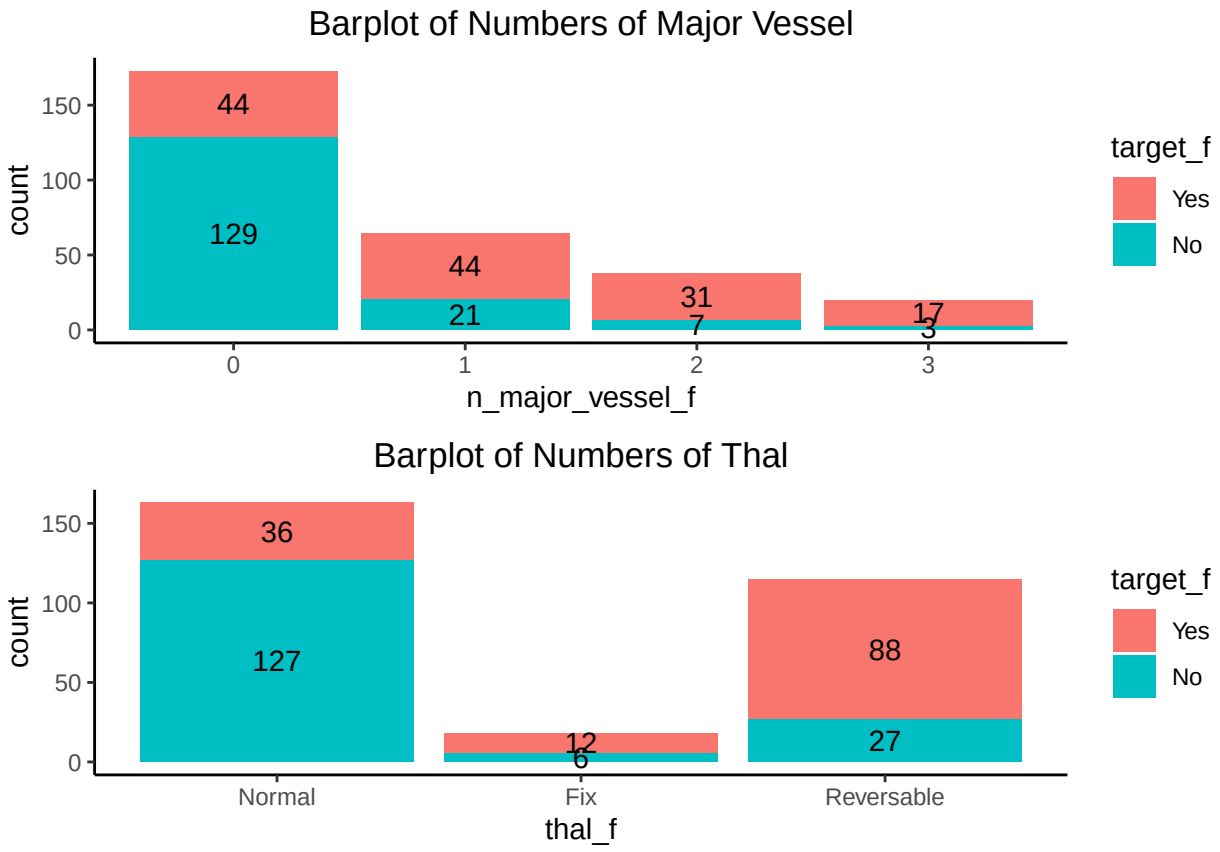
```
  theme(plot.title = element_text(hjust = 0.5))
## ecg
g4 = ggplot(h2,aes(x =ecg_f,fill=target_f)) +
  geom_bar()+
  geom_text(stat = "count",aes(label =..count..),position=position_stack(0.5))+
  theme_classic()+
  ggtitle("Barplot of ECG")+
  theme(plot.title = element_text(hjust = 0.5))
## exercise_angina
g5 = ggplot(h2,aes(x =exercise_angina_f,fill=target_f)) +
  geom_bar()+
  geom_text(stat = "count",aes(label =..count..),position=position_stack(0.5))+
  theme_classic()+
  ggtitle("Barplot of Exercise Angina")+
  theme(plot.title = element_text(hjust = 0.5))
## slope
g6 = ggplot(h2,aes(x =slope_f,fill=target_f)) +
  geom_bar()+
  geom_text(stat = "count",aes(label =..count..),position=position_stack(0.5))+
  theme_classic()+
  ggtitle("Barplot of Slope")+
  theme(plot.title = element_text(hjust = 0.5))
## n_major_vessel
g7 = ggplot(h2,aes(x =n_major_vessel_f,fill=target_f)) +
  geom_bar()+
  geom_text(stat = "count",aes(label =..count..),position=position_stack(0.5))+
  theme_classic()+
  ggtitle("Barplot of Numbers of Major Vessel")+
  theme(plot.title = element_text(hjust = 0.5))
## thal
g8 = ggplot(h2,aes(x =thal_f,fill=target_f)) +
  geom_bar()+
  geom_text(stat = "count",aes(label =..count..),position=position_stack(0.5))+
  theme_classic()+
  ggtitle("Barplot of Numbers of Thal")+
  theme(plot.title = element_text(hjust = 0.5))
## target
g9 = ggplot(h2,aes(x=target_f)) +
  geom_bar(aes(fill=target_f)) +
  geom_text(stat ='count',aes(label=..count..),vjust=-0.5)+
  theme_classic()+
  ggtitle("Barplot of Target")+
  theme(plot.title = element_text(hjust = 0.5))

grid.arrange(g1,g3)
```



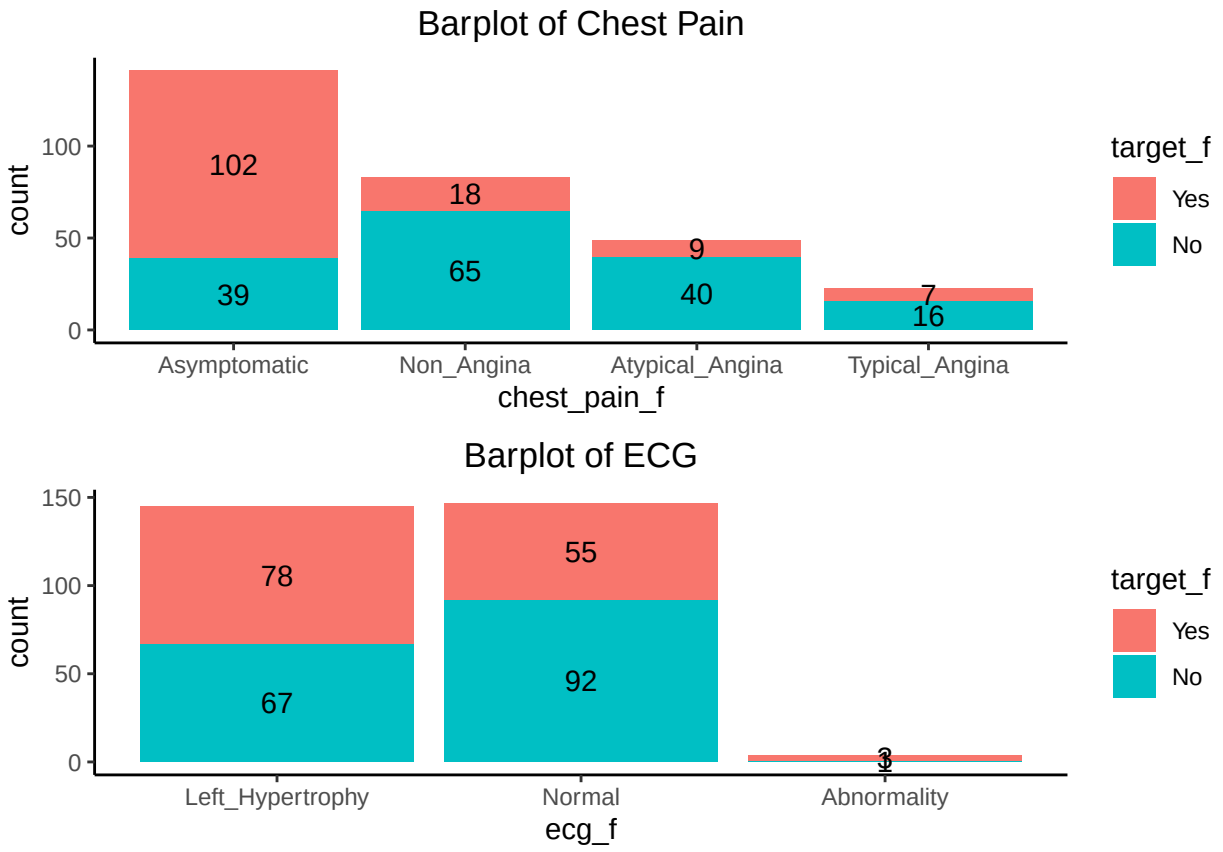

```
grid.arrange(g5,g6)
```



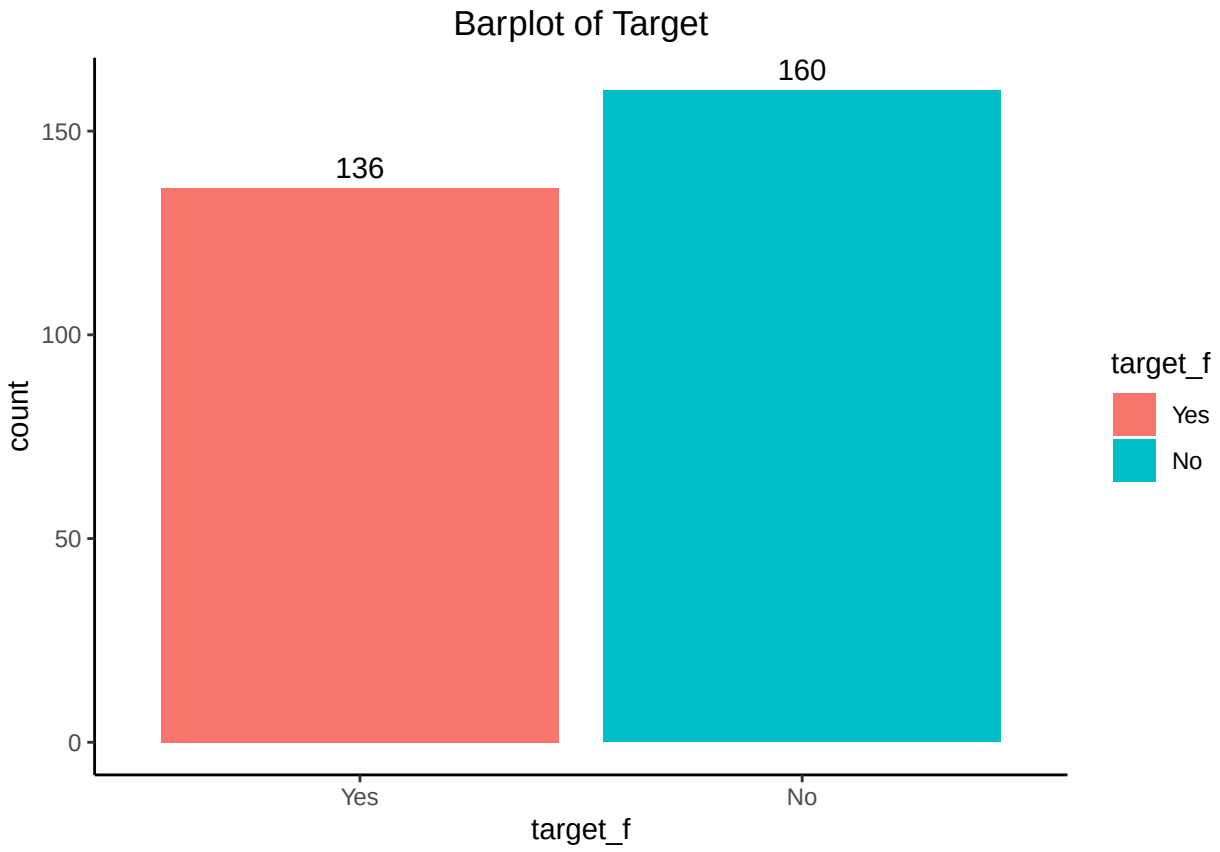

```
grid.arrange(g7,g8)
```




```
grid.arrange(g2,g4)
```



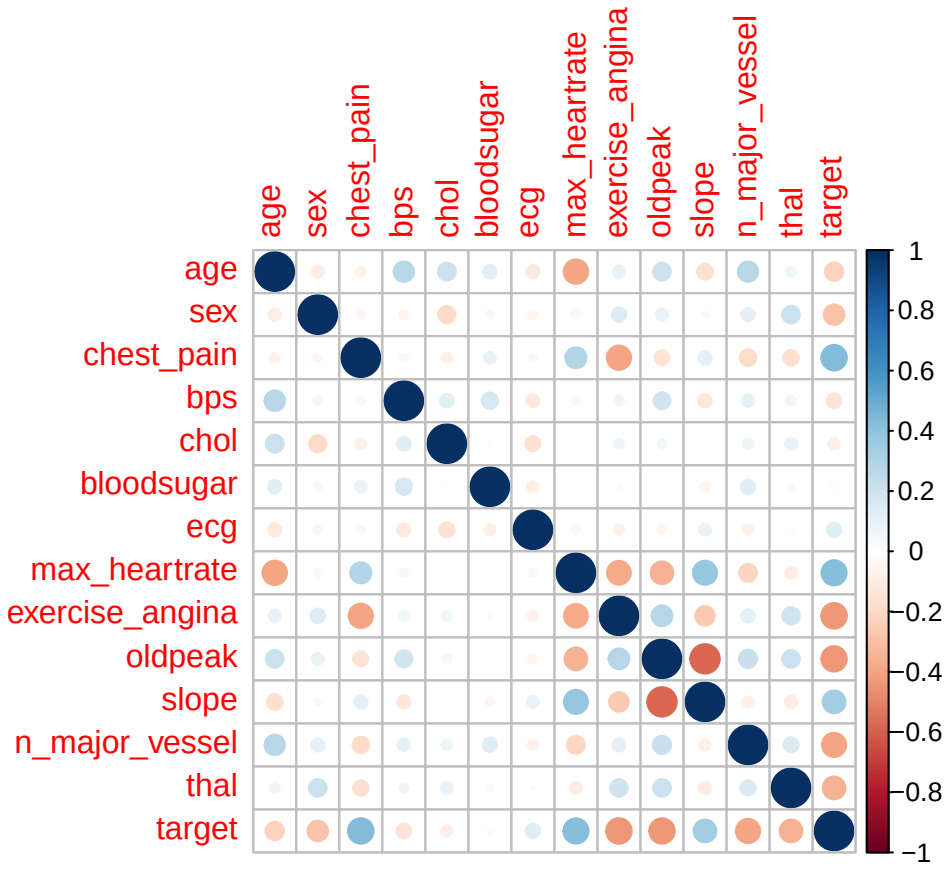

```
grid.arrange(g9)
```



- We can see all the category variables and the composition of them. We can roughly find that males are more likely to get a heart disease. And we can also look other disease indicators, unluckliy I do not understand what they means even after googling. What makes me confused is that the graph shows that a asymptomatic person is more likely to get a heart disease, it looks so weird, but the original dataset also shows this pattern.
- Besides, the composition of whether having a heart disease is well balanced in this dataset.

Correlation Plots

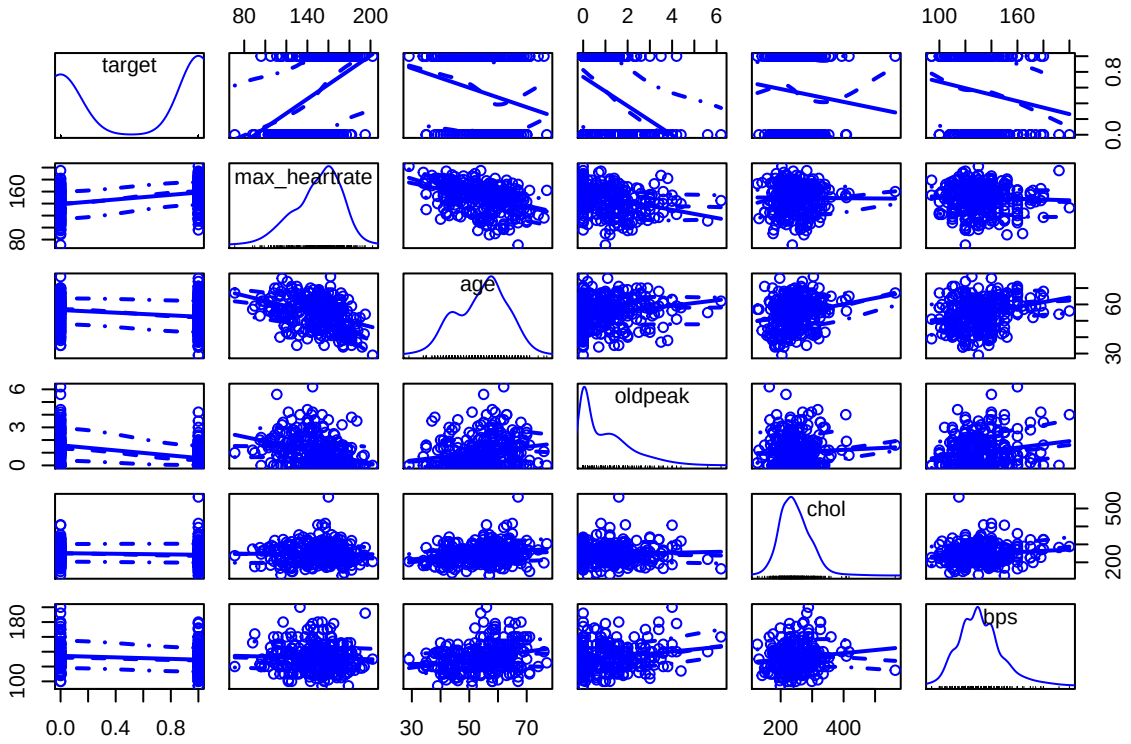
```
# Since correlation plots do not work for factor variables, we used the orginal dataset.
par(mfrow=c(1,1))
corrplot(cor(h))
```



- Mainly we have to see the correlation between target and other variables, and what we find is that chest pain, max heartrate and slope are relatively highly related to target.
- We also find the possibility of multi-coolinearity that can be tested formally in the following content.

Scatter Plot

```
#Since our dependent variables are 0-1 variables, the scatter plot does not make much sense. Anyway, I
scatterplotMatrix(data=h,~target+max_heartrate+age+oldpeak+chol+bps)
```



- As you can see, this really does not make much sense.

(b)

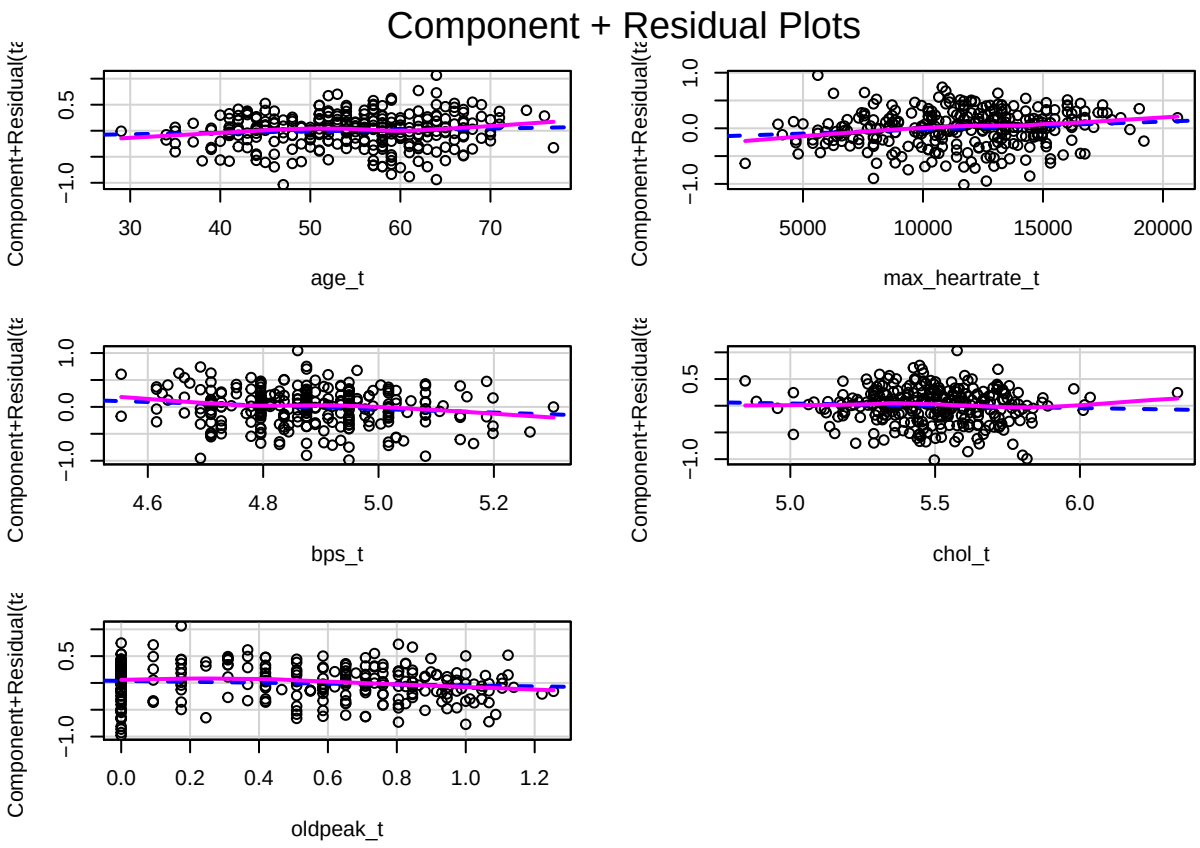
Transformation with yj Power

```
#we can use the yjPower transformation
t_yj = powerTransform(data=h2,cbind(age,max_heartrate,bps,chol,oldpeak)~1.,family="yjPower")
h2_yj = yjPower(with(h2, cbind(age,max_heartrate,bps,chol,oldpeak)), coef(t_yj,round=T))
colnames(h2_yj) = c("age_t","max_heartrate_t","bps_t","chol_t","oldpeak_t")
h2 = h2 %>% cbind(h2_yj)
```

Verification for linearity

- Then we see the component residual plot to verify the linearity.

```
reg_for_linearity = lm(target~age_t+max_heartrate_t+bps_t+chol_t+
                    oldpeak_t+sex_f+chest_pain_f+bloodsugar_f
                    +ecg_f+exercise_angina_f+slope_f+thal_f+
                      n_major_vessel_f,data=h2)
crPlots(reg_for_linearity,terms = ~age_t+max_heartrate_t+bps_t+chol_t+
          oldpeak_t)
```



- We can see that it seems that the linearity work for all continuous variables, but we have to verify them formally.

```
suppressWarnings(boxTidwell(target~age_t+max_heartrate_t+bps_t+chol_t,other.x = ~sex_f + chest_pain_f +
```

```
##                  MLE of lambda Score Statistic (z) Pr(>|z|)
## age_t                  -1.7625             -0.1740   0.8618
## max_heartrate_t        -2.3647             -0.1175   0.9065
## bps_t                  12.2544             -0.1601   0.8728
## chol_t                 -6.3971              0.2597   0.7951
##
## iterations =  26
```

- We can see that the p value is very big for each variable, so it means that we can not reject the null hypothesis that we do not need a further transformation.
- So we can conclude that the transformation suggested with yjPower above is a suitable one.

(c)

Basic Multiple Regression

```r
# we have to determine the baseline for each factor variable.
h2$ecg_f <- relevel(h2$ecg_f, "Normal")

reg_basal = lm(target~age+chol+bps+max_heartrate+oldpeak+
                 sex_f+chest_pain_f+exercise_angina_f+ecg_f+slope_f+
                 bloodsugar_f+n_major_vessel_f+thal_f,data=h2)
summary(reg_basal)
```

```
##
## Call:
## lm(formula = target ~ age + chol + bps + max_heartrate + oldpeak +
##     sex_f + chest_pain_f + exercise_angina_f + ecg_f + slope_f +
##     bloodsugar_f + n_major_vessel_f + thal_f, data = h2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.01153 -0.19126  0.03463  0.22160  1.02731
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  0.8459824  0.2866814   2.951 0.003441 **
## age                          0.0024813  0.0026829   0.925 0.355852
## chol                        -0.0003092  0.0004037  -0.766 0.444387
## bps                         -0.0023069  0.0012270  -1.880 0.061155 .
## max_heartrate                0.0020818  0.0011382   1.829 0.068481 .
## oldpeak                     -0.0416979  0.0231305  -1.803 0.072527 .
## sex_fmale                   -0.1588395  0.0486907  -3.262 0.001245 **
## chest_pain_fNon_Angina       0.2156284  0.0545718   3.951 9.89e-05 ***
## chest_pain_fAtypical_Angina  0.1535696  0.0648824   2.367 0.018631 *
## chest_pain_fTypical_Angina   0.2581515  0.0814873   3.168 0.001708 **
## exercise_angina_fYes        -0.0854813  0.0503526  -1.698 0.090705 .
## ecg_fLeft_Hypertrophy       -0.0490422  0.0414226  -1.184 0.237456
## ecg_fAbnormality            -0.1304531  0.1774053  -0.735 0.462760
## slope_fFlat                 -0.1336206  0.0503790  -2.652 0.008459 **
## slope_fDown                 -0.0731007  0.0954252  -0.766 0.444302
## bloodsugar_f>120             0.0509529  0.0590300   0.863 0.388796
## n_major_vessel_f1           -0.2695257  0.0532974  -5.057 7.79e-07 ***
## n_major_vessel_f2           -0.3460248  0.0686389  -5.041 8.40e-07 ***
## n_major_vessel_f3           -0.2960619  0.0864956  -3.423 0.000714 ***
## thal_fFix                   -0.0679198  0.0919817  -0.738 0.460898
## thal_fReversable            -0.2182611  0.0504307  -4.328 2.11e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.336 on 275 degrees of freedom
## Multiple R-squared:  0.5777, Adjusted R-squared:  0.547
## F-statistic: 18.81 on 20 and 275 DF,  p-value: < 2.2e-16
```

Interpretation of estimates

- I just want to discuss the significant ones.

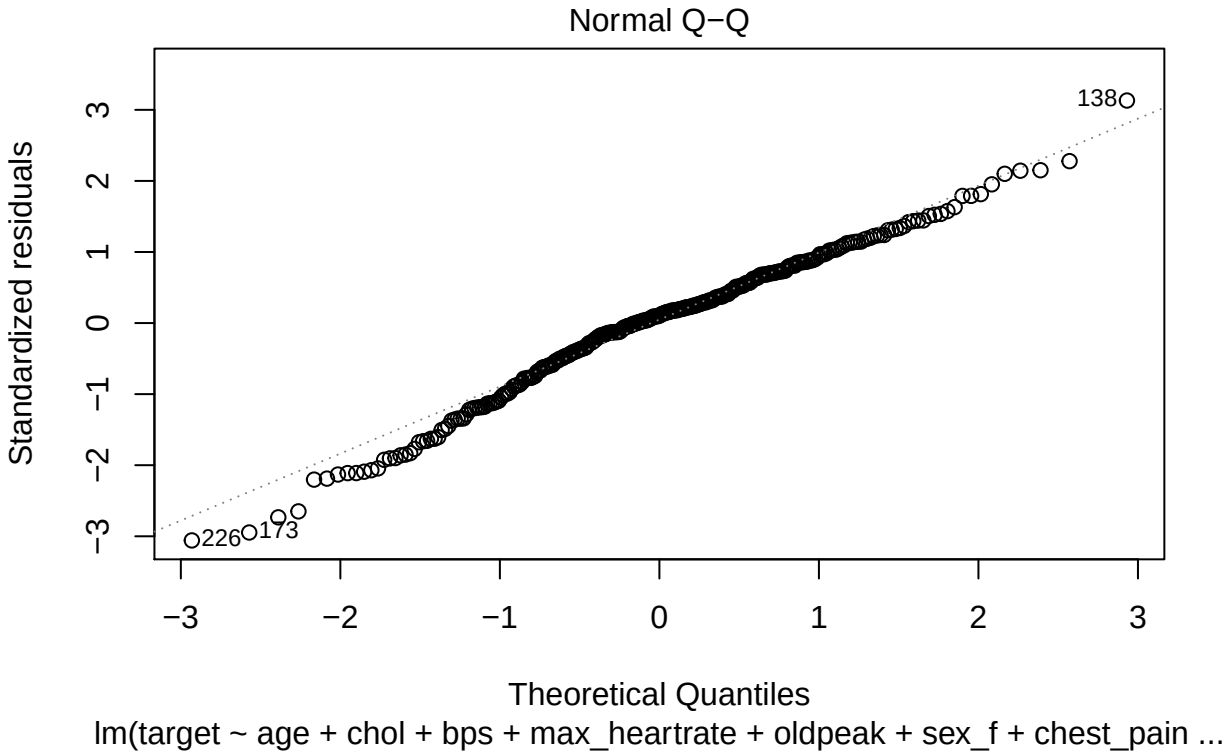
- The coefficient of the sex_male means that averagely men are about 16% more possible to get a heart disease than woman
- The coeffient of the different levels of chest pain seems weird because of the reason I mentioned above, but it means that if you have angina, you will be less likely to get a heartdisease by seperately 21%, 15%, 25%.
- The other statistically significant variables are hard to explain, but after googling at least I know they make sense.
- The coefficient of slope_flat and slope_down means that if this some kinds of slope is nearly 0 or negative, you will be separetely 13% and 7% more likely to get a heart disease.
- The coefficient of n_major vessel shows that if the number of major vessels are 1 or 2 or 3, the possibility of getting a heart disease will separately increase about 26%, 34% and 29%.
- The coefficient of thal_fReversable means that if in a thal test there shows a reversable defect, the probability of getting a heart disease will increase 21%.

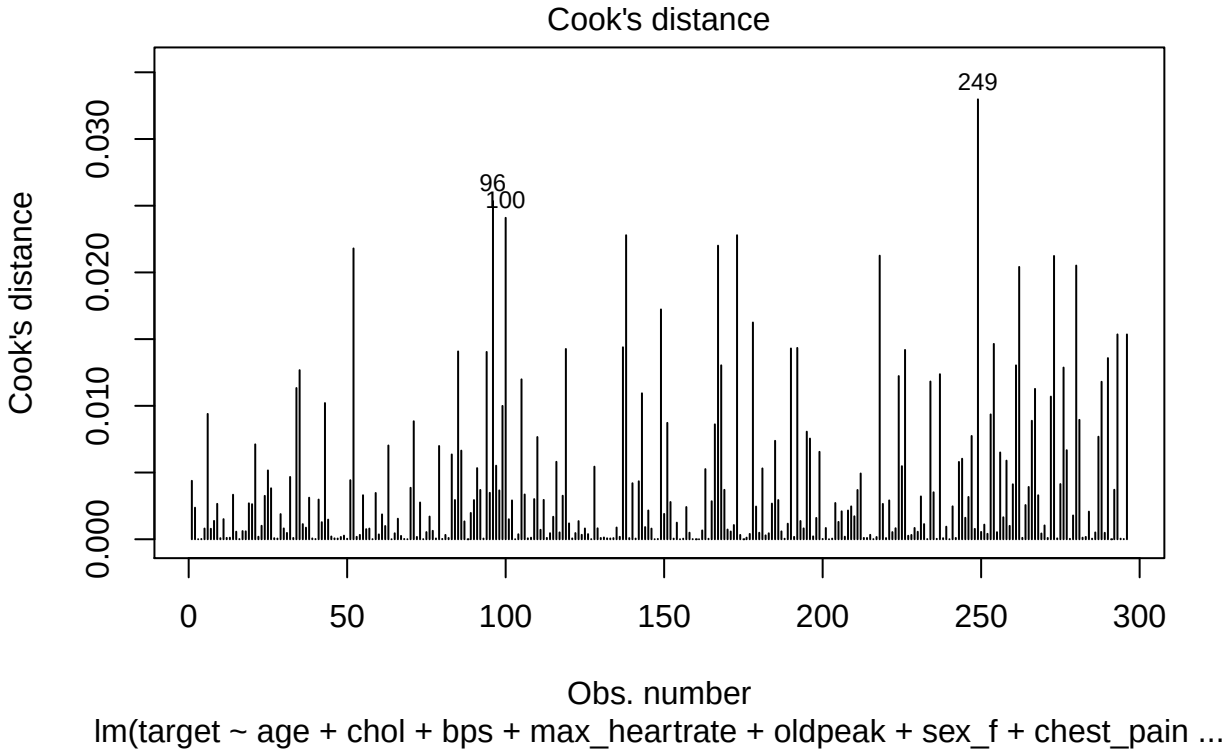
(d)

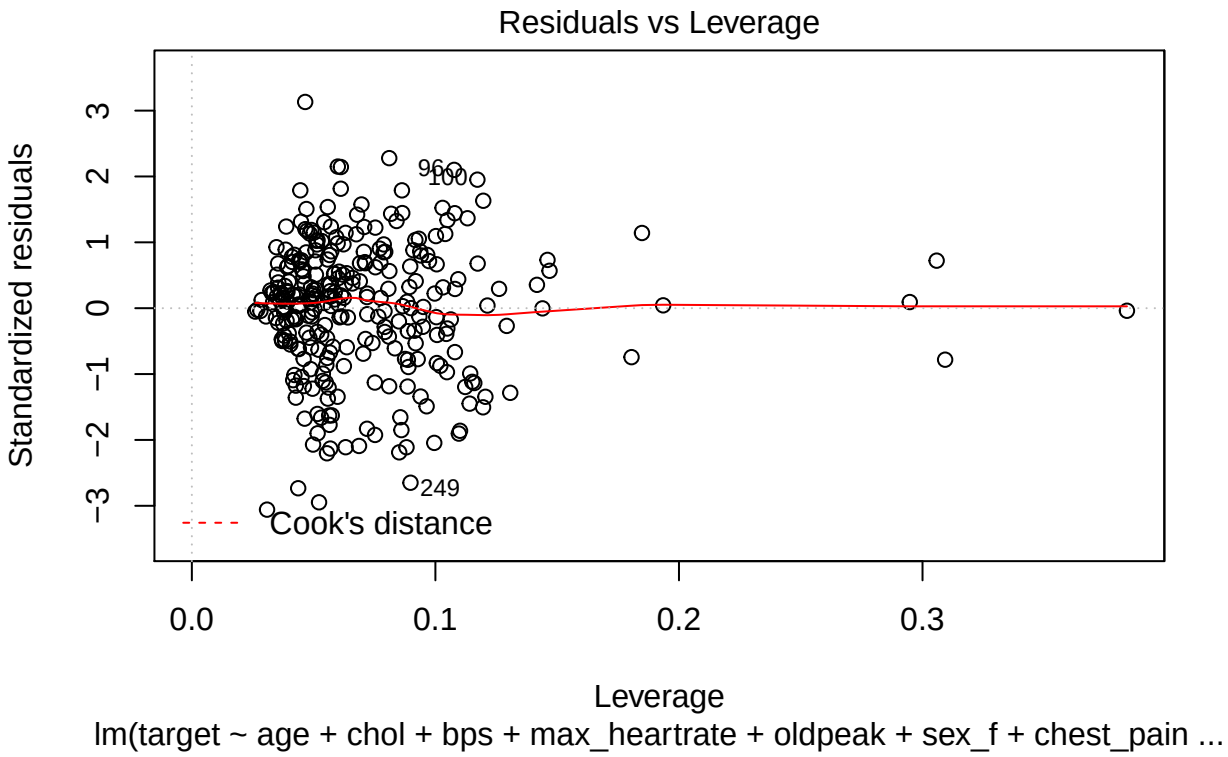
Test for outliers

- We draw the quantile plot, cook distance and residuals ve leverage plots to see vividly whethere there is an outlier.

```
plot(reg_basal,which = c(2,4,5))
```




Cook's distance
Cook's distance
0.000
0.010
0.020
0.030
96
100
249
0
50
100
150
200
250
300
Obs. number
lm(target ~ age + chol + bps + max_heartrate + oldpeak + sex_f + chest_pain ...



- From the graph, we have several potential outliers, then we validate them formally with outlier test.

```
outlierTest(reg_basal)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 138 3.182831          0.0016265      0.48143
```

- As we can see from the results, there are no significant outliers in the model.

(e)

Best Subset Regression with Backward Method

- We run a subset regression to get the candidate models. Since if a factor varible have n levels, it will enter the regression with three indicator varibles, totally we have 20 variables. It takes too long to do the exhaustive subset regression, so we use a backward method.
- And if we want get every candidate models, we should let

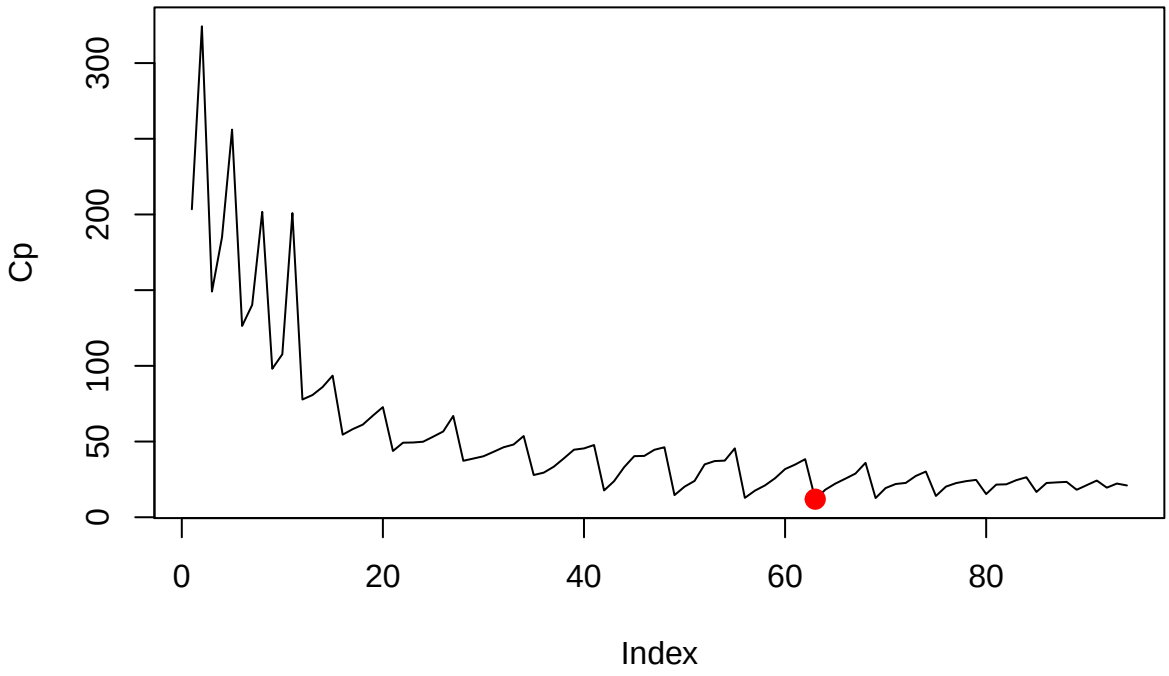
$$nbest = \binom{20}{11} = 184756$$

. While it takes too much time, so we just let nbest = 1000.

```
reg_subset = suppressWarnings(regsubsets(target~n_major_vessel_f+thal_f+slope_f+exercise_angina_f+ecg_f
reg_subset_s = summary(reg_subset)
```

- Then we try to get the regression with the smallest cp and draw the plot of cp.

```
plot(reg_subset_s$cp, ylab = "Cp", type = "l")
cp_min = which.min(reg_subset_s$cp)
points(cp_min, reg_subset_s$cp[cp_min], col = "red", cex = 2, pch = 20)
```




```
cp_min = which.min(reg_subset_s$cp)
print(cp_min)
```

```
## [1] 63
```

- So we extract No.63 regression model.

```
coeff_63 = coef(reg_subset,63)
print(coeff_63)
```

```
##                 (Intercept)          n_major_vessel_f1
##                0.908677439               -0.259936529
##          n_major_vessel_f2          n_major_vessel_f3
```

```
##                -0.315592246                -0.289245220
##            thal_fReversable                 slope_fFlat
##                -0.198942712                -0.129241731
##        exercise_angina_fYes      chest_pain_fNon_Angina
##                -0.091554136                 0.235851585
## chest_pain_fAtypical_Angina  chest_pain_fTypical_Angina
##                 0.163119259                 0.272634043
##                   sex_fmale                         bps
##                -0.163061893                -0.002252095
##               max_heartrate                     oldpeak
##                 0.001787411                -0.055157767
```

```
# Since we have just one level of thal and slope, we have to get the indicator variable by ourself.
h2 = h2 %>%
  mutate(thal_Reversable = ifelse(thal_f=="Reversable",1,0),
         slope_flat = ifelse(slope_f=="Flat",1,0)
                  )
reg_subset_best = lm(target~n_major_vessel_f+thal_Reversable+slope_flat+
                       exercise_angina_f+chest_pain_f+sex_f+bps+
                       max_heartrate+oldpeak,data=h2)
```

Test of Multi-coolinearity

- Then we test the multi-coolinearity of this model.

```
tidy(vif(reg_subset_best))
```

```
## Warning: 'tidy.numeric' is deprecated.
## See help("Deprecated")
```

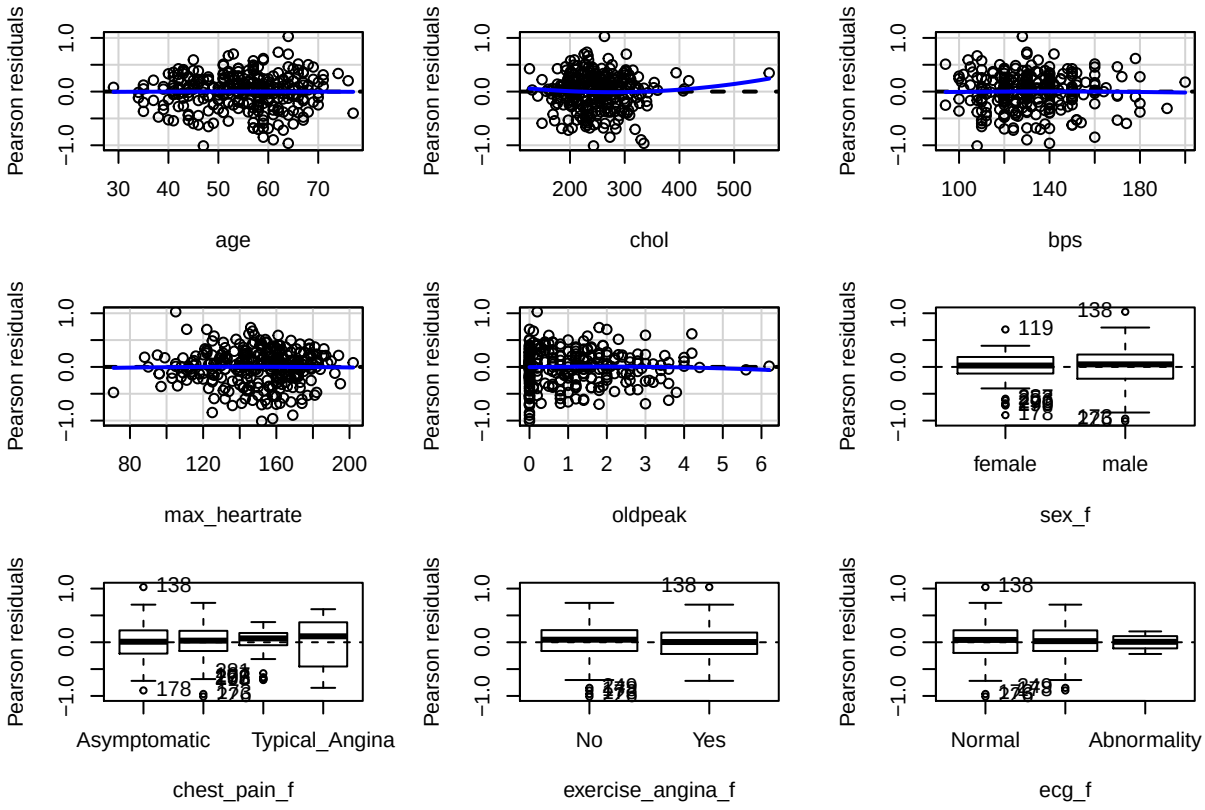
```
## # A tibble: 13 x 2
##    names                          x
##    <chr>                      <dbl>
##  1 n_major_vessel_f1           1.18
##  2 n_major_vessel_f2           1.20
##  3 n_major_vessel_f3           1.16
##  4 thal_Reversable             1.35
##  5 slope_flat                  1.30
##  6 exercise_angina_fYes        1.45
##  7 chest_pain_fNon_Angina      1.50
##  8 chest_pain_fAtypical_Angina 1.51
##  9 chest_pain_fTypical_Angina  1.23
## 10 sex_fmale                   1.18
## 11 bps                         1.09
## 12 max_heartrate               1.54
## 13 oldpeak                     1.42
```

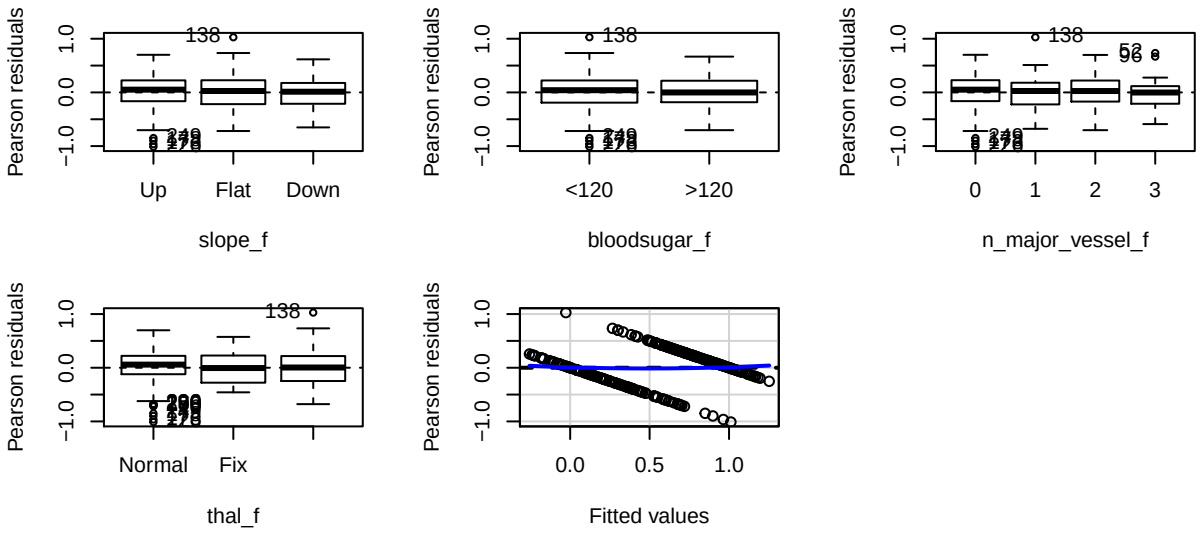
- We can see that none of the independent variables has a VIF over 4, so we can conclude that there does not exist collinearity.

(f)

Respective Residuals vs X

```
residualPlots(reg_basal)
```



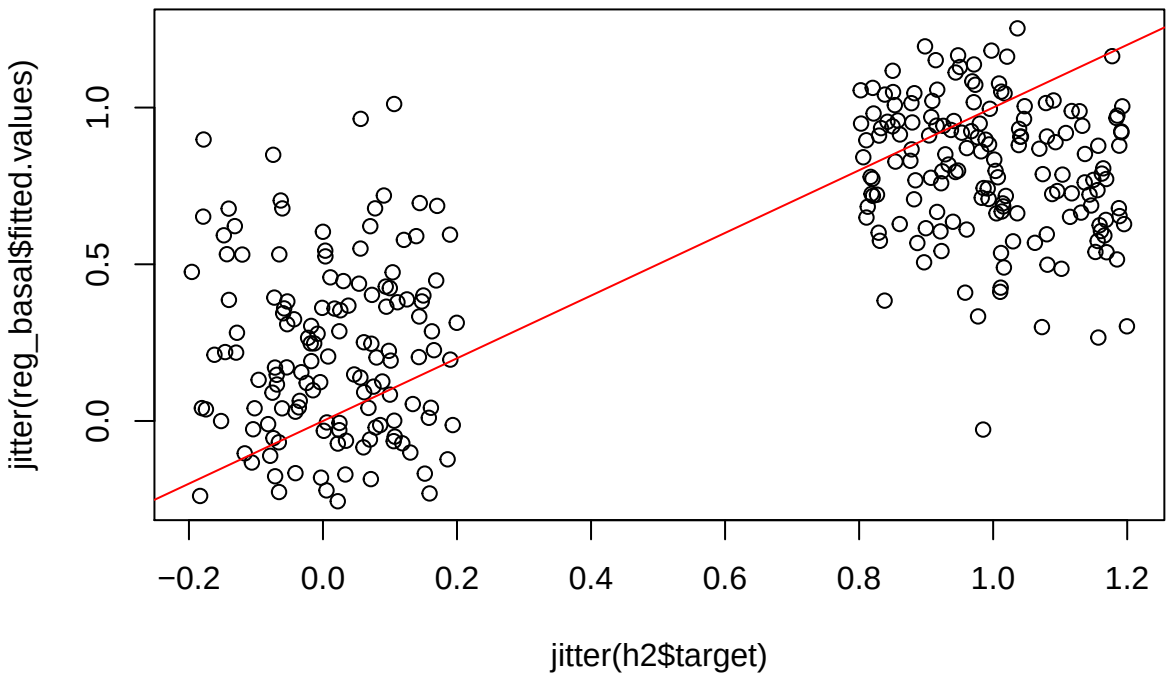



```
##                   Test stat Pr(>|Test stat|)
## age                 -0.0464           0.9630
## chol                 0.9915           0.3223
## bps                 -0.1275           0.8987
## max_heartrate       -0.1352           0.8926
## oldpeak             -0.3489           0.7274
## sex_f
## chest_pain_f
## exercise_angina_f
## ecg_f
## slope_f
## bloodsugar_f
## n_major_vessel_f
## thal_f
## Tukey test           0.7014           0.4831
```

- We can see that the residual plots for continuous variable looks not so bad. The constant variation almost holds, but with chol,bps and oldpeak, the value of x is too concentrated.

Y vs fitted value of Y

```
plot(jitter(h2$target),jitter(reg_basal$fitted.values))
abline(0,1,col="red")
```

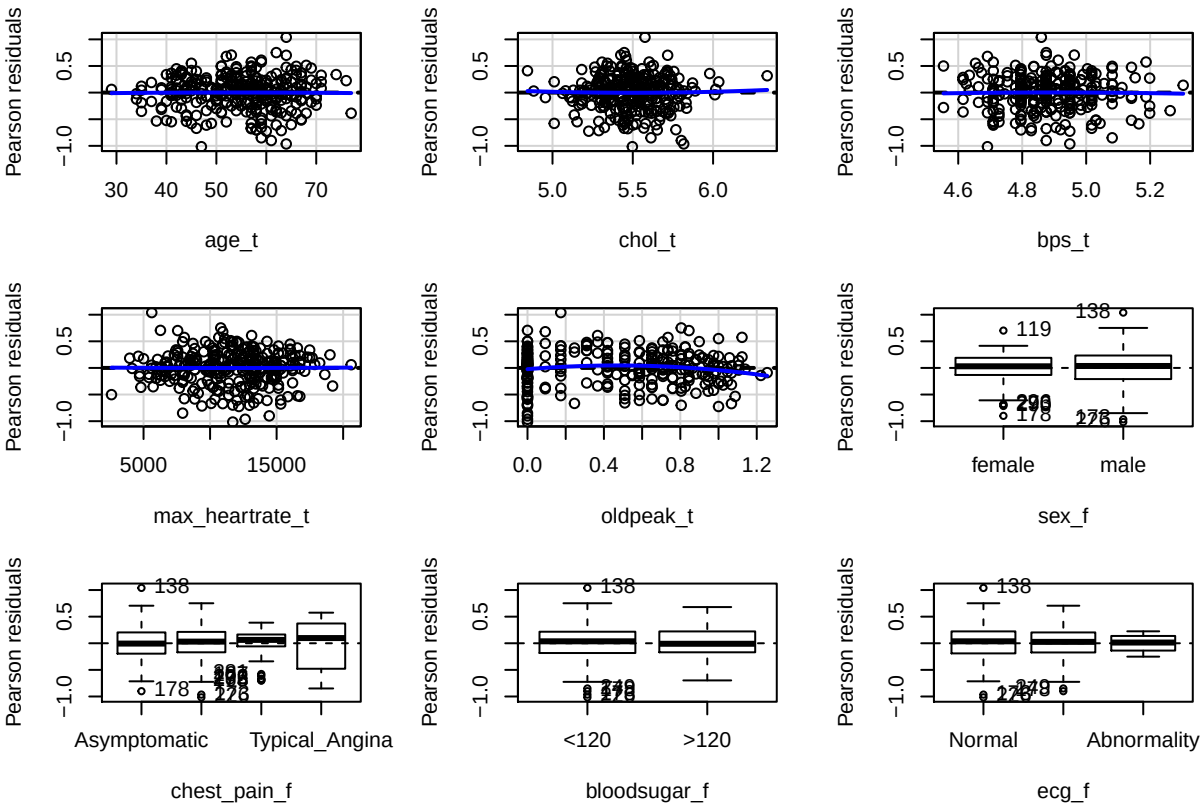


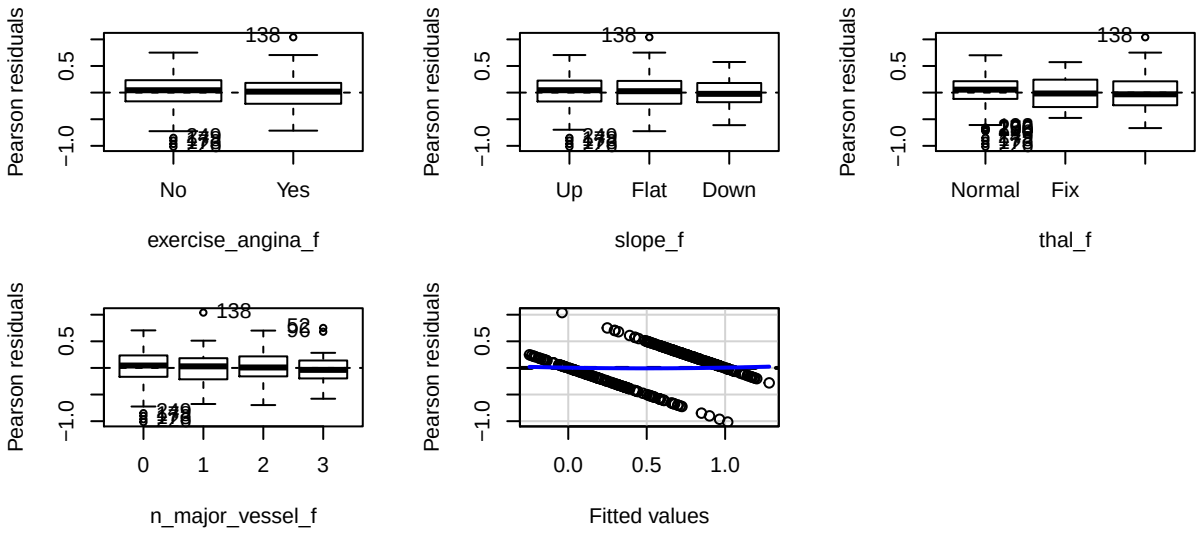
- We draw the picture of y and $\hat{y}$. As we all know, the y is a 0-1 variable, the $\hat{y}$ is a continuous variable, so it looks quite weird and it is an inevitable result.

(g)

Mulitiple Regression with Transformed Independent Variables

```
reg_robust = lm(target~age_t+chol_t+bps_t+max_heartrate_t+oldpeak_t
   +sex_f+chest_pain_f+bloodsugar_f+ecg_f+exercise_angina_f
   +slope_f+thal_f+n_major_vessel_f,data=h2)
residualPlots(reg_robust)
```

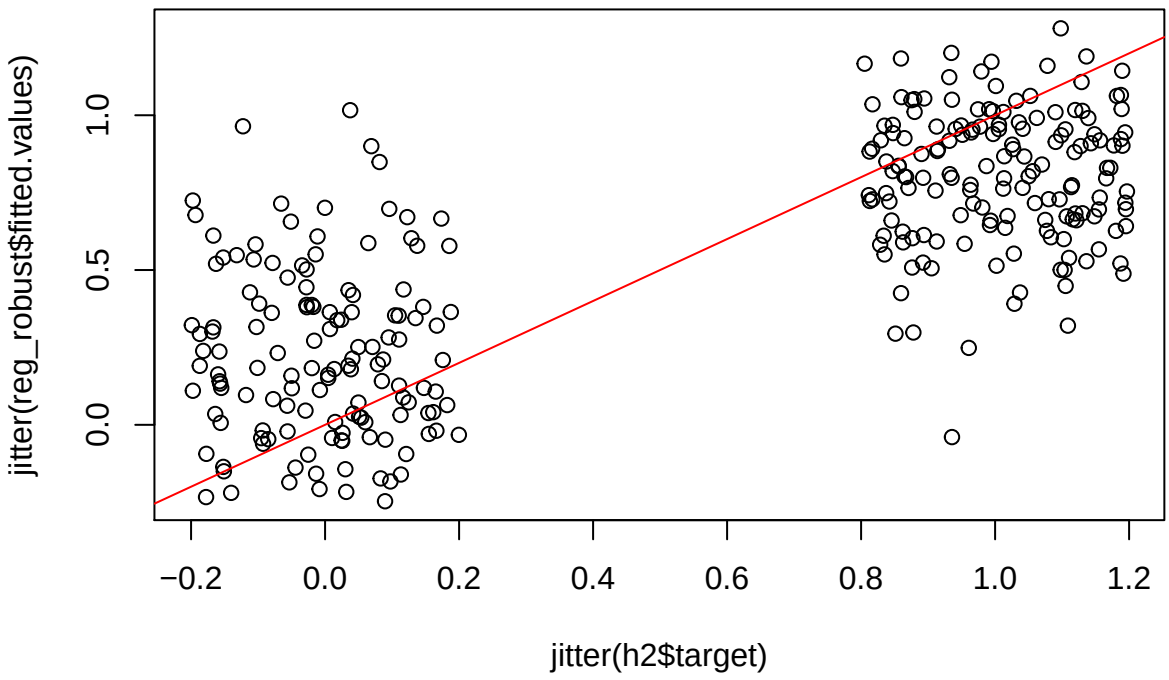

Pearson residuals
0.5
−1.0
30
40
50
60
70
age_t
5.0
5.5
6.0
chol_t
4.6
4.8
5.0
5.2
bps_t
5000
15000
max_heartrate_t
0.0
0.4
0.8
1.2
oldpeak_t
119
138
178
female
male
sex_f
Asymptomatic
Typical_Angina
chest_pain_f
<120
>120
bloodsugar_f
Normal
Abnormality
ecg_f




```
##                   Test stat Pr(>|Test stat|)
## age_t               -0.0944           0.9249
## chol_t               0.2806           0.7792
## bps_t               -0.1778           0.8590
## max_heartrate_t      0.0623           0.9504
## oldpeak_t           -1.9739           0.0494 *
## sex_f
## chest_pain_f
## bloodsugar_f
## ecg_f
## exercise_angina_f
## slope_f
## thal_f
## n_major_vessel_f
## Tukey test           0.3755           0.7073
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- From the residual plots, we can not see there are no significant improvements when using this model, but the problem of over-concentration of x is a little improved.

```
plot(jitter(h2$target),jitter(reg_robust$fitted.values))
abline(0,1,col="red")
```



- The plot is as same as before, but I do not think it makes any sense.

Comparison with AIC and BIC

- Then we compare this two models with AIC and BIC.

```
AIC(reg_basal,reg_robust)
```

```
##            df      AIC
## reg_basal  22 216.5553
## reg_robust 22 218.3901
```

```
BIC(reg_basal,reg_robust)
```

```
##            df      BIC
## reg_basal  22 297.7432
## reg_robust 22 299.5780
```

- These two criterions all suggest that the model in (b) is better.

(h)

Multiple Regression with Interaction Terms

- I think that sex may have a interaction term with other variables.

```
reg_interaction = update(reg_basal,.~.+sex_f:max_heartrate+sex_f:bps+sex_f:chol+sex_f:oldpeak+sex_f:age
summary(reg_interaction)
```

```
##
## Call:
## lm(formula = target ~ age + chol + bps + max_heartrate + oldpeak +
##     sex_f + chest_pain_f + exercise_angina_f + ecg_f + slope_f +
##     bloodsugar_f + n_major_vessel_f + thal_f + max_heartrate:sex_f +
##     bps:sex_f + chol:sex_f + oldpeak:sex_f + age:sex_f, data = h2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.99426 -0.17463  0.02882  0.22435  1.04994
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  1.035e+00  5.000e-01   2.071 0.039320 *
## age                          3.138e-03  4.582e-03   0.685 0.494024
## chol                         1.183e-05  5.847e-04   0.020 0.983877
## bps                         -3.379e-03  2.095e-03  -1.613 0.107866
## max_heartrate                9.482e-04  2.156e-03   0.440 0.660400
## oldpeak                     -4.290e-02  3.941e-02  -1.088 0.277367
## sex_fmale                   -3.852e-01  5.943e-01  -0.648 0.517390
## chest_pain_fNon_Angina       2.120e-01  5.551e-02   3.820 0.000166 ***
## chest_pain_fAtypical_Angina  1.570e-01  6.546e-02   2.398 0.017165 *
## chest_pain_fTypical_Angina   2.533e-01  8.222e-02   3.081 0.002275 **
## exercise_angina_fYes        -7.448e-02  5.187e-02  -1.436 0.152174
## ecg_fLeft_Hypertrophy       -5.269e-02  4.216e-02  -1.250 0.212491
## ecg_fAbnormality            -1.407e-01  1.847e-01  -0.762 0.446769
## slope_fFlat                 -1.261e-01  5.153e-02  -2.448 0.015017 *
## slope_fDown                 -5.706e-02  9.781e-02  -0.583 0.560148
## bloodsugar_f>120             4.291e-02  6.032e-02   0.711 0.477490
## n_major_vessel_f1           -2.624e-01  5.396e-02  -4.864 1.96e-06 ***
## n_major_vessel_f2           -3.333e-01  7.035e-02  -4.737 3.51e-06 ***
## n_major_vessel_f3           -2.803e-01  8.884e-02  -3.155 0.001787 **
## thal_fFix                   -7.170e-02  9.380e-02  -0.764 0.445281
## thal_fReversable            -2.249e-01  5.124e-02  -4.389 1.64e-05 ***
## max_heartrate:sex_fmale      1.589e-03  2.429e-03   0.654 0.513416
## bps:sex_fmale                1.717e-03  2.596e-03   0.661 0.508922
## chol:sex_fmale              -6.544e-04  8.347e-04  -0.784 0.433729
## oldpeak:sex_fmale            1.498e-04  4.266e-02   0.004 0.997200
## age:sex_fmale               -1.393e-03  5.484e-03  -0.254 0.799704
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.338 on 270 degrees of freedom
## Multiple R-squared:  0.5803, Adjusted R-squared:  0.5415
## F-statistic: 14.93 on 25 and 270 DF,  p-value: < 2.2e-16
```

Comparison with AIC and BIC

- From the results, we can see that interaction terms work not so good, then we compare it with our basal model with AIC and BIC.

```
AIC(reg_basal,reg_interaction,reg_robust)
```

```
##                 df      AIC
## reg_basal       22 216.5553
## reg_interaction 27 224.7062
## reg_robust      22 218.3901
```

```
BIC(reg_basal,reg_interaction,reg_robust)
```

```
##                 df      BIC
## reg_basal       22 297.7432
## reg_interaction 27 324.3459
## reg_robust      22 299.5780
```

- The basal model is really good compared to the model with interaction terms and the robust model.

Test for Adding Higher Power Form

- Then we test whether there should be a higher power form.

```
resettest(reg_basal , power=2, type="regressor")
```

```
##
##  RESET test
##
## data:  reg_basal
## RESET = 0.22755, df1 = 5, df2 = 270, p-value = 0.9504
```

```
resettest(reg_basal , power=3, type="regressor")
```

```
##
##  RESET test
##
## data:  reg_basal
## RESET = 0.25286, df1 = 5, df2 = 270, p-value = 0.9382
```

- Normally we consider for quadratic and cubic forms, the results show that we do not need them.

(i)

Model selection

- We do not take our best subset model into consideration. Now we compare it with our basal model.

```r
AIC(reg_basal,reg_subset_best)
```

```
##                 df      AIC
## reg_basal       22 216.5553
## reg_subset_best 15 207.8150
```

```r
BIC(reg_basal,reg_subset_best)
```

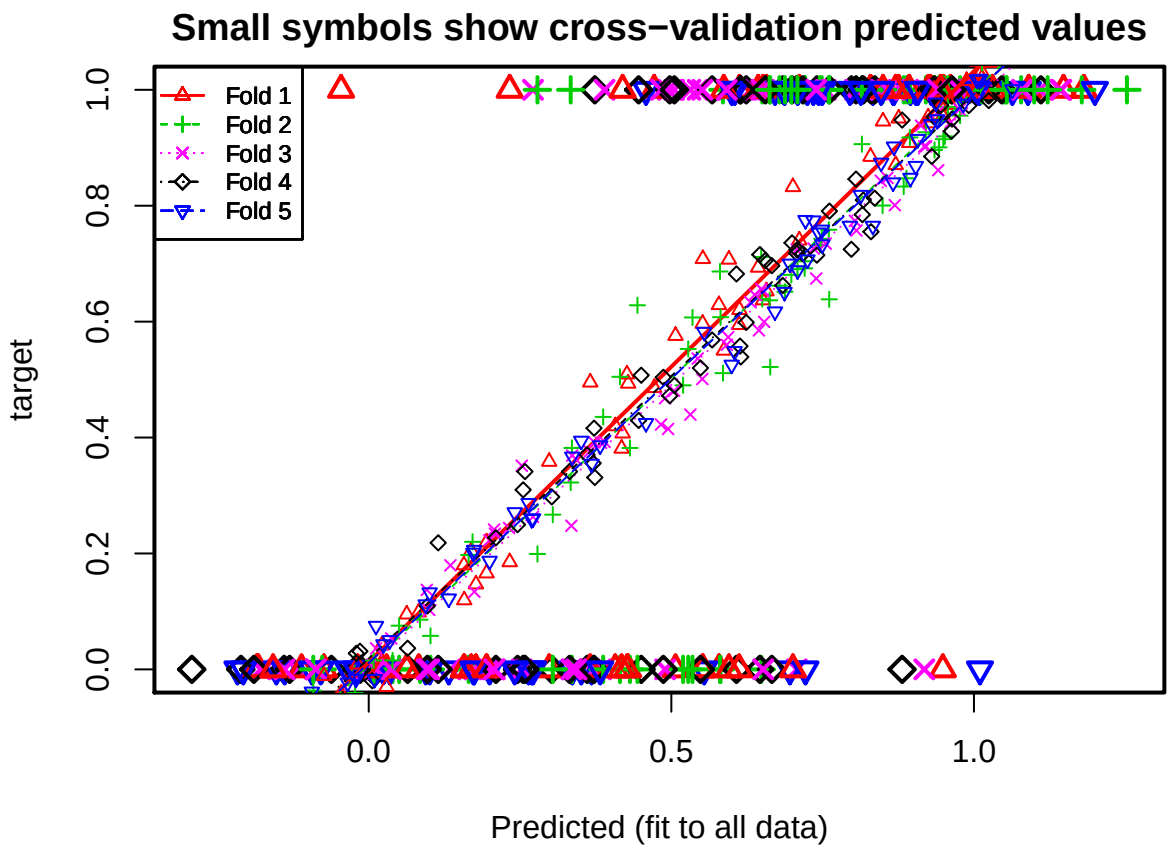
```
##                 df      BIC
## reg_basal       22 297.7432
## reg_subset_best 15 263.1704
```

- So finally I will choose the best subset model to do the cross validation.

5-Fold Cross Validation

```r
cv_results = CVlm(data=h2,reg_subset_best,m=5,plotit=T,printit=T)
```

```
## Analysis of Variance Table
##
## Response: target
##                    Df Sum Sq Mean Sq F value  Pr(>F)
## n_major_vessel_f    3  18.23    6.08   54.22 < 2e-16 ***
## thal_Reversable     1  10.84   10.84   96.77 < 2e-16 ***
## slope_flat          1   3.27    3.27   29.14 1.4e-07 ***
## exercise_angina_f   1   3.18    3.18   28.39 2.0e-07 ***
## chest_pain_f        3   3.07    1.02    9.12 8.9e-06 ***
## sex_f               1   1.30    1.30   11.61 0.00075 ***
## bps                 1   0.64    0.64    5.74 0.01725 *
## max_heartrate       1   0.52    0.52    4.65 0.03194 *
## oldpeak             1   0.86    0.86    7.69 0.00591 **
## Residuals         282  31.60    0.11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Warning in CVlm(data = h2, reg_subset_best, m = 5, plotit = T, printit = T):
##
##  As there is >1 explanatory variable, cross-validation
##  predicted values for a fold are not a linear function
##  of corresponding overall predicted values.  Lines that
##  are shown for the different folds are approximate
```




```
##
## fold 1
## Observations in test set: 59
##                  7     9     17    24     31    38      39      40      47
## Predicted    0.829 0.657  1.182 0.613 0.8757 0.651  1.0159  1.0100  0.9879
## cvpred       0.885 0.653  1.183 0.620 0.9508 0.637  1.0254  1.0263  1.0155
## target       1.000 1.000  1.000 1.000 1.0000 1.000  1.0000  1.0000  1.0000
## CV residual 0.115 0.347 -0.183 0.380 0.0492 0.363 -0.0254 -0.0263 -0.0155
##                 48      50    52     54     63     64     81    86     99
## Predicted    0.9495  1.0905 0.233  1.148 0.711 0.9611 0.9243 0.472 0.643
## cvpred       0.9527  1.0802 0.185  1.161 0.740 0.9744 0.9527 0.486 0.693
## target       1.0000  1.0000 1.000  1.000 1.000 1.0000 1.0000 1.000 1.000
## CV residual 0.0473 -0.0802 0.815 -0.161 0.260 0.0256 0.0473 0.514 0.307
##                 107   112    113    126     127    131    136   137     138
## Predicted    1.0304 0.587 0.8930 0.8497  1.0056 0.9280 0.871 0.420 -0.0457
## cvpred       1.0469 0.550 0.9086 0.9456  1.0486 0.9425 0.871 0.407 -0.0421
## target       1.0000 1.000 1.0000 1.0000  1.0000 1.0000 1.000 1.000  1.0000
## CV residual -0.0469 0.450 0.0914 0.0544 -0.0486 0.0575 0.129 0.593  1.0421
##                 155    156    166    173    179    181     182    192
## Predicted    1.0086 0.9450  0.429  0.949  0.298  0.507  0.0826  0.579
## cvpred       1.0274 0.9724  0.493  0.955  0.358  0.576  0.0985  0.629
## target       1.0000 1.0000  0.000  0.000  0.000  0.000  0.0000  0.000
## CV residual -0.0274 0.0276 -0.493 -0.955 -0.358 -0.576 -0.0985 -0.629
##                 194    201    205    214    217    218    220    222
## Predicted    -0.182  0.195  0.169 -0.0740  0.0293  0.552 -0.1110  0.157
## cvpred       -0.268  0.219  0.204 -0.0112 -0.0300  0.708 -0.0907  0.180
```

```
## target       0.000  0.000  0.000  0.0000  0.0000  0.000  0.0000  0.000
## CV residual  0.268 -0.219 -0.204  0.0112  0.0300 -0.708  0.0907 -0.180
##                  236    237    239    241    244    248    252    267
## Predicted   -0.01831  0.552 -0.185  0.407  0.366  0.158  0.177  0.418
## cvpred       0.00867  0.597 -0.175  0.420  0.495  0.119  0.147  0.381
## target       0.00000  0.000  0.000  0.000  0.000  0.000  0.000  0.000
## CV residual -0.00867 -0.597  0.175 -0.420 -0.495 -0.119 -0.147 -0.381
##                  270    276     278    281     283    293    294    296
## Predicted    0.195  0.427  0.0218  0.595  0.0628  0.612 -0.158  0.701
## cvpred       0.166  0.509  0.0419  0.707  0.0950  0.594 -0.156  0.832
## target       0.000  0.000  0.0000  0.000  0.0000  0.000  0.000  0.000
## CV residual -0.166 -0.509 -0.0419 -0.707 -0.0950 -0.594  0.156 -0.832
##
## Sum of squares = 8.84    Mean square = 0.15    n = 59
##
## fold 2
## Observations in test set: 60
##                 11    13     15    20    23    26     28     30     33
## Predicted   0.650 0.889  1.078 0.721 0.748 0.663 0.9205 0.8942 0.9520
## cvpred      0.639 0.847  1.106 0.704 0.743 0.522 0.9268 0.9179 0.9197
## target      1.000 1.000  1.000 1.000 1.000 1.000 1.0000 1.0000 1.0000
## CV residual 0.361 0.153 -0.106 0.296 0.257 0.478 0.0732 0.0821 0.0803
##                  36    44    46    49      53    55    61     68     72
## Predicted   0.9419 0.720 0.935 0.884  1.0518 0.585 0.745 0.9423 0.9770
## cvpred      0.9484 0.692 0.896 0.833  1.0958 0.511 0.732 0.9007 0.9554
## target      1.0000 1.000 1.000 1.000  1.0000 1.000 1.000 1.0000 1.0000
## CV residual 0.0516 0.308 0.104 0.167 -0.0958 0.489 0.268 0.0993 0.0446
##                 75     78      80    94    95   101    102   114   119
## Predicted   0.849 0.9494  1.0492 0.334 0.679 0.761 0.8149  1.18 0.279
## cvpred      0.800 0.9149  1.0934 0.322 0.662 0.639 0.9063  1.18 0.199
## target      1.000 1.0000  1.0000 1.000 1.000 1.000 1.0000  1.00 1.000
## CV residual 0.200 0.0851 -0.0934 0.678 0.338 0.361 0.0937 -0.18 0.801
##                 120    123    139     141    146     147    148   150   152
## Predicted   0.760  1.253  1.122  1.0543  1.100  1.0133 0.9568 0.662 0.698
## cvpred      0.758  1.273  1.117  1.0871  1.134  1.0462 0.9706 0.637 0.681
## target      1.000  1.000  1.000  1.0000  1.000  1.0000 1.0000 1.000 1.000
## CV residual 0.242 -0.273 -0.117 -0.0871 -0.134 -0.0462 0.0294 0.363 0.319
##                 154   157     164    171     172     177    180    185
## Predicted   0.688 0.708  0.0503  0.172  0.0848  0.1020  0.165  0.581
## cvpred      0.652 0.691  0.0754  0.220  0.0858  0.0578  0.197  0.608
## target      1.000 1.000  0.0000  0.000  0.0000  0.0000  0.000  0.000
## CV residual 0.348 0.309 -0.0754 -0.220 -0.0858 -0.0578 -0.197 -0.608
##                 190    195     200    208      216    224    225     227
## Predicted    0.649  0.432 -0.0226  0.336  0.00806  0.580  0.415 -0.0769
## cvpred       0.712  0.382 -0.0407  0.382  0.01534  0.686  0.505 -0.0881
## target       0.000  0.000  0.0000  0.000  0.00000  0.000  0.000  0.0000
## CV residual -0.712 -0.382  0.0407 -0.382 -0.01534 -0.686 -0.505  0.0881
##                 240    254    256    261    266    277     282    287
## Predicted    0.0395  0.444  0.519  0.304  0.387  0.528 -0.0917  0.535
## cvpred       0.0155  0.628  0.490  0.267  0.436  0.553 -0.0365  0.607
## target       0.0000  0.000  0.000  0.000  0.000  0.000  0.0000  0.000
## CV residual -0.0155 -0.628 -0.490 -0.267 -0.436 -0.553  0.0365 -0.607
##
## Sum of squares = 6.68    Mean square = 0.11    n = 60
```

```
##
## fold 3
## Observations in test set: 59
##                  4     6    19    25    29    32    35     37    43    51
## Predicted   0.9125 0.544 0.631 0.653 0.806 0.505 0.531  1.144 0.390 0.594
## cvpred      0.9381 0.536 0.633 0.600 0.757 0.481 0.440  1.129 0.393 0.573
## target      1.0000 1.000 1.000 1.000 1.000 1.000 1.000  1.000 1.000 1.000
## CV residual 0.0619 0.464 0.367 0.400 0.243 0.519 0.560 -0.129 0.607 0.427
##                 56    59    60    65     69    71    76    82    85     88
## Predicted   0.803 0.645 0.869 0.846 0.9208 0.495 0.735 0.941 0.755 0.9617
## cvpred      0.774 0.585 0.801 0.843 0.9025 0.415 0.726 0.861 0.735 0.9441
## target      1.000 1.000 1.000 1.000 1.0000 1.000 1.000 1.000 1.000 1.0000
## CV residual 0.226 0.415 0.199 0.157 0.0975 0.585 0.274 0.139 0.265 0.0559
##                 96   106   110   111    117   118     121    133   142
## Predicted   0.271 0.638 0.483 0.857  1.143 0.649  1.1082  1.079 0.592
## cvpred      0.263 0.646 0.422 0.848  1.159 0.656  1.0918  1.111 0.560
## target      1.000 1.000 1.000 1.000  1.000 1.000  1.0000  1.000 1.000
## CV residual 0.737 0.354 0.578 0.152 -0.159 0.344 -0.0918 -0.111 0.440
##                145   151    153    158    162    163    167    170    178
## Predicted   0.739 0.551 0.9762 0.9406 -0.173  0.365  0.707 -0.179  0.918
## cvpred      0.674 0.501 0.9661 0.9621 -0.172  0.376  0.724 -0.174  0.902
## target      1.000 1.000 1.0000 1.0000  0.000  0.000  0.000  0.000  0.000
## CV residual 0.326 0.499 0.0339 0.0379  0.172 -0.376 -0.724  0.174 -0.902
##                 186    191    197    199    204    210    221    234
## Predicted    0.384  0.135 -0.118  0.336  0.232  0.175  0.209  0.373
## cvpred       0.391  0.179 -0.152  0.369  0.244  0.134  0.234  0.392
## target       0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000
## CV residual -0.391 -0.179  0.152 -0.369 -0.244 -0.134 -0.234 -0.392
##                 238    243    245     255    265    268    275    285
## Predicted    0.0999  0.253  0.199  0.0122  0.490  0.335  0.343  0.207
## cvpred       0.1026  0.351  0.223  0.0361  0.468  0.248  0.362  0.241
## target       0.0000  0.000  0.000  0.0000  0.000  0.000  0.000  0.000
## CV residual -0.1026 -0.351 -0.223 -0.0361 -0.468 -0.248 -0.362 -0.241
##                 286    289    290     291     295
## Predicted    0.096 -0.139  0.652  0.0371 -0.0875
## cvpred       0.137 -0.118  0.653  0.0534 -0.1246
## target       0.000  0.000  0.000  0.0000  0.0000
## CV residual -0.137  0.118 -0.653 -0.0534  0.1246
##
## Sum of squares = 7.57    Mean square = 0.13    n = 59
##
## fold 4
## Observations in test set: 59
##                  2     5    14    21    27    34    41    42     45    57
## Predicted   0.830 0.805 0.708 0.374 0.836 0.684 0.740 0.797 0.9915 0.817
## cvpred      0.755 0.846 0.724 0.331 0.813 0.662 0.715 0.725 0.9728 0.810
## target      1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.0000 1.000
## CV residual 0.245 0.154 0.276 0.669 0.187 0.338 0.285 0.275 0.0272 0.190
##                 62    70     74    79    87    89    90    92    97   100
## Predicted   0.9629 0.624 1.0244 0.567 0.761 0.717 0.699 0.707 0.614 0.446
## cvpred      0.9284 0.599 0.9813 0.568 0.791 0.717 0.736 0.718 0.558 0.429
## target      1.0000 1.000 1.0000 1.000 1.000 1.000 1.000 1.000 1.000 1.000
## CV residual 0.0716 0.401 0.0187 0.432 0.209 0.283 0.264 0.282 0.442 0.571
##                105    108     124   128   129   130    132    135   140
```

```
## Predicted   0.497 0.9451  1.1106 0.655 0.816 0.930 0.9345  1.046 0.615
## cvpred      0.472 0.9734  1.0841 0.706 0.785 0.885 0.9531  1.129 0.539
## target      1.000 1.0000  1.0000 1.000 1.000 1.000 1.0000  1.000 1.000
## CV residual 0.528 0.0266 -0.0841 0.294 0.215 0.115 0.0469 -0.129 0.461
##                149    160     161    165    169     176    183      202
## Predicted   0.505 0.9638 -0.0145  0.246  0.302 -0.0616 -0.140  0.00583
## cvpred      0.490 0.9496  0.0312  0.250  0.298 -0.0860 -0.154 -0.01966
## target      1.000 1.0000  0.0000  0.000  0.000  0.0000  0.000  0.00000
## CV residual 0.510 0.0504 -0.0312 -0.250 -0.298  0.0860  0.154  0.01966
##                 203    212    215     219     230    231    232    235
## Predicted    0.0643  0.371  0.115 -0.1307  0.0968  0.361  0.255  0.450
## cvpred       0.0364  0.356  0.218 -0.0936  0.1093  0.370  0.310  0.507
## target       0.0000  0.000  0.000  0.0000  0.0000  0.000  0.000  0.000
## CV residual -0.0364 -0.356 -0.218  0.0936 -0.1093 -0.370 -0.310 -0.507
##                246    249     250    251    253    258    259    260
## Predicted   -0.292  0.881 -0.0204 -0.211  0.548  0.332  0.210  0.372
## cvpred      -0.285  0.948  0.0277 -0.187  0.520  0.341  0.227  0.416
## target       0.000  0.000  0.0000  0.000  0.000  0.000  0.000  0.000
## CV residual  0.285 -0.948 -0.0277  0.187 -0.520 -0.341 -0.227 -0.416
##                262    264    272    279    280    288
## Predicted    0.608  0.258  0.666 -0.190  0.646  0.487
## cvpred       0.682  0.341  0.697 -0.142  0.716  0.504
## target       0.000  0.000  0.000  0.000  0.000  0.000
## CV residual -0.682 -0.341 -0.697  0.142 -0.716 -0.504
##
## Sum of squares = 7.37    Mean square = 0.12    n = 59
##
## fold 5
## Observations in test set: 59
##                  1       3     8    10    12     16    18    22        58
## Predicted   0.833  1.0092 0.749 0.866  1.09 0.9392 0.904 0.895  1.063512
## cvpred      0.765  1.0187 0.759 0.840  1.11 0.9489 0.868 0.848  1.000576
## target      1.000  1.0000 1.000 1.000  1.00 1.0000 1.000 1.000  1.000000
## CV residual 0.235 -0.0187 0.241 0.160 -0.11 0.0511 0.132 0.152 -0.000576
##                66     67    73    77    83    84    91     93    98    103
## Predicted   0.725 0.9295 0.671 0.795 0.600 0.735 0.458 0.9370 0.709 0.9823
## cvpred      0.707 0.9406 0.617 0.765 0.525 0.774 0.425 0.9777 0.688 0.9756
## target      1.000 1.0000 1.000 1.000 1.000 1.000 1.000 1.0000 1.000 1.0000
## CV residual 0.293 0.0594 0.383 0.235 0.475 0.226 0.575 0.0223 0.312 0.0244
##                104   109   115   116    122   125    134   143   144
## Predicted   0.8679 0.687 0.749 0.605  1.200 0.743 0.9073 0.847 0.813
## cvpred      0.9019 0.651 0.733 0.549  1.214 0.754 0.9149 0.874 0.818
## target      1.0000 1.000 1.000 1.000  1.000 1.000 1.0000 1.000 1.000
## CV residual 0.0981 0.349 0.267 0.451 -0.214 0.246 0.0851 0.126 0.182
##                159    168    174     175    184    187     188    189
## Predicted    1.0050  0.696  0.132  0.0942  0.336 -0.139  0.0120 -0.217
## cvpred       1.0178  0.700  0.122  0.1119  0.367 -0.165  0.0748 -0.220
## target       1.0000  0.000  0.000  0.0000  0.000  0.000  0.0000  0.000
## CV residual -0.0178 -0.700 -0.122 -0.1119 -0.367  0.165 -0.0748  0.220
##                193    196    198    206    207    209    211      213
## Predicted    0.101  0.554  0.272  0.351 -0.153  0.368  0.269  0.00976
## cvpred       0.133  0.582  0.260  0.394 -0.142  0.355  0.259 -0.01673
## target       0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.00000
## CV residual -0.133 -0.582 -0.260 -0.394  0.142 -0.355 -0.259  0.01673
```

```
##                223   226     228    229     233     242    247    257
## Predicted    0.200  1.01 -0.0793  0.174  0.0242  0.0349  0.264 -0.207
## cvpred       0.187  1.02 -0.0771  0.206  0.0431  0.0503  0.287 -0.173
## target       0.000  0.00  0.0000  0.000  0.0000  0.0000  0.000  0.000
## CV residual -0.187 -1.02  0.0771 -0.206 -0.0431 -0.0503 -0.287  0.173
##                263    269     271    273     274    284    292
## Predicted   -0.0347  0.173 -0.0939  0.722 -0.0210  0.241  0.382
## cvpred      -0.0230  0.201 -0.0386  0.775 -0.0165  0.271  0.386
## target       0.0000  0.000  0.0000  0.000  0.0000  0.000  0.000
## CV residual  0.0230 -0.201  0.0386 -0.775  0.0165 -0.271 -0.386
##
## Sum of squares = 5.39    Mean square = 0.09    n = 59
##
## Overall (Sum over all 59 folds)
##    ms
## 0.121
```

- As we know before, because of our dependent variable is 0-1 variable, the graph will look quite weird. But we can get the MSE for each fold and the cv error for this model. The MSE for each fold is seperately 0.15, 0.11, 0.13 0.12, 0.09. The CV error is as follows.

```
attr(cv_results,"ms")
```

```
## [1] 0.121
```

- And we can see that the 5 fitted lines are near to each other, proving that our results are robust.

Question 2

Import necessary packages

```
library(tidyverse) # for ggplot
library(car) # for EDA plots
library(leaps) # for subset regression
library(lmtest) # for ramsey test
library(MASS) # for stepAIC
library(multcomp) # for glht
library(corrplot)
library(emmeans) # for ANOVA
library(broom)
library(effects)
```

Import data

```
setwd("C:/Users/zyj37/Desktop/MAE/ECON 430")
hc = read.csv("Homework/2/healthcare.csv")
```

Clean data and generate some useful variables

Deal With Missing Values

- We drop "TI" and "HSAT" due to explanation, aggregate the year variables to only one indicating whether the obsevation is after 1987. .

```
hc2 = hc
drops = c("TI","HSAT","YEAR1984","YEAR1985","YEAR1986","YEAR1987","YEAR1988","YEAR1991","YEAR1994")
hc2 = hc2[,!(names(hc2) %in% drops)]
hc2$TIME = ifelse(hc2$YEAR>1987,1,0)
hc2=subset(hc2,select=-YEAR)
hc2_omit = na.omit(hc2)
(nrow(hc2)-nrow(hc2_omit))/nrow(hc2)
```

```
## [1] 0.000951
```

- We can see that the missing values take up a very small part of the dateset,so we can just remove them.

```
hc2= na.omit(hc2)
```

(a)

- First, since we have so many variables, we use VIF to delete some variables. Then we use subset regression that uses adjusted R-square, Cp, AIC and BIC as the criterion to get our candidates.

```
reg_whole = lm(data=hc2,DOCVIS~.)
tidy(vif(reg_whole))
```

```
## Warning: 'tidy.numeric' is deprecated.
## See help("Deprecated")
```

```
## # A tibble: 33 x 2
##    names       x
##    <chr>   <dbl>
##  1 ID       1.02
##  2 FEMALE   1.36
##  3 AGE      1.58
##  4 HANDDUM  1.66
##  5 ALC      1.00
##  6 FAMHIST  1.00
##  7 HANDPER  1.82
##  8 HHKIDS   1.38
##  9 EDUC    10.1
## 10 MARRIED  1.39
## # ... with 23 more rows
```

```
reg_whole = update(reg_whole,.~.-ID-HAUPTS-ABITUR-WORKING,data=hc2) # VIF>10
tidy(vif(reg_whole))
```

```
## Warning: 'tidy.numeric' is deprecated.
## See help("Deprecated")

## # A tibble: 29 x 2
##    names       x
##    <chr>   <dbl>
##  1 FEMALE   1.34
##  2 AGE      1.55
##  3 HANDDUM  1.66
##  4 ALC      1.00
##  5 FAMHIST  1.00
##  6 HANDPER  1.81
##  7 HHKIDS   1.37
##  8 EDUC     3.95
##  9 MARRIED  1.38
## 10 REALS    1.21
## # ... with 19 more rows
```

```
reg_whole = update(reg_whole,.~.-WHITEC,data=hc2) # VIF>4
tidy(vif(reg_whole))
```

```
## Warning: 'tidy.numeric' is deprecated.
## See help("Deprecated")

## # A tibble: 28 x 2
##    names       x
##    <chr>   <dbl>
##  1 FEMALE   1.34
##  2 AGE      1.54
##  3 HANDDUM  1.66
##  4 ALC      1.00
##  5 FAMHIST  1.00
##  6 HANDPER  1.81
##  7 HHKIDS   1.37
##  8 EDUC     3.94
##  9 MARRIED  1.38
## 10 REALS    1.21
## # ... with 18 more rows
```

```
reg_whole = update(reg_whole,.~.-HHNINC,data=hc2) # VIF>4
tidy(vif(reg_whole))
```

```
## Warning: 'tidy.numeric' is deprecated.
## See help("Deprecated")

## # A tibble: 27 x 2
##    names       x
##    <chr>   <dbl>
##  1 FEMALE   1.34
##  2 AGE      1.54
##  3 HANDDUM  1.66
##  4 ALC      1.00
```

```
##  5 FAMHIST  1.00
##  6 HANDPER  1.80
##  7 HHKIDS   1.37
##  8 EDUC     3.94
##  9 MARRIED  1.36
## 10 REALS    1.21
## # ... with 17 more rows
```

```
reg_subset = regsubsets(formula(reg_whole),method="exhaustive",nbest=40,data=hc2)
reg_subset_s = summary(reg_subset)
which.max(reg_subset_s$adjr2)
```

```
## [1] 268
```

```
which.min(reg_subset_s$bic)
```

```
## [1] 268
```

```
which.min(reg_subset_s$cp)
```

```
## [1] 268
```

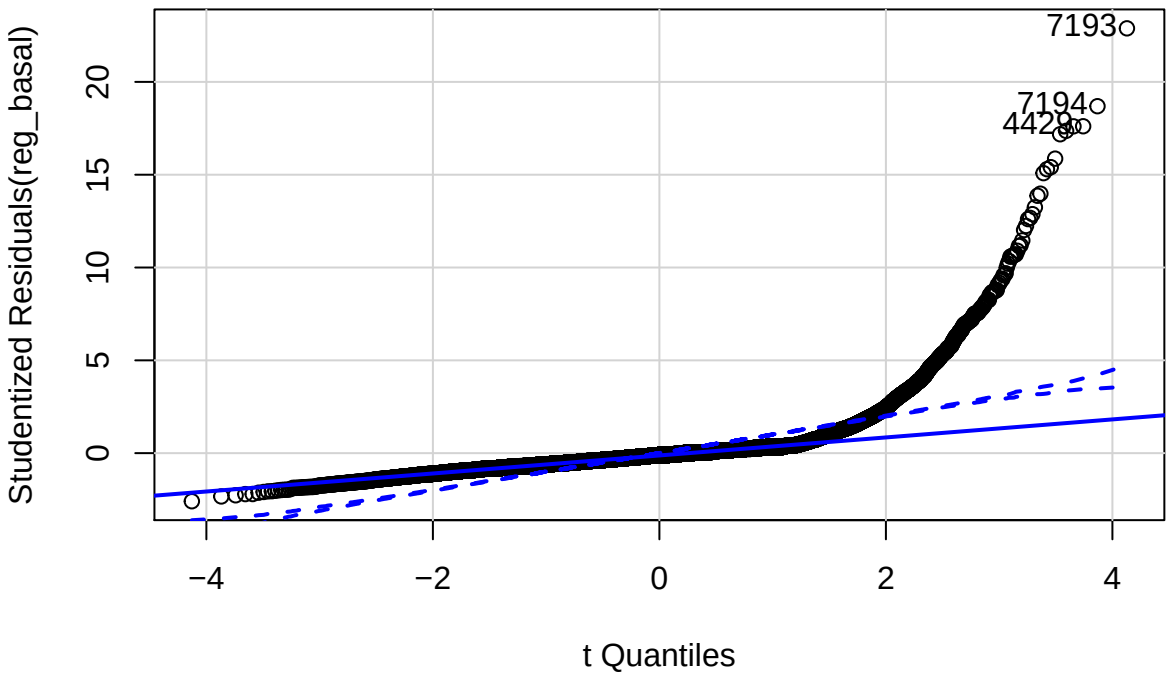
```
coef(reg_subset,268)
```

```
## (Intercept)      FEMALE     HANDPER     HOSPVIS      DOCTOR     HEALTHY
##       4.616       0.456       0.023       0.250       3.925       0.684
##    HOSPITAL     NEWHSAT        TIME
##       1.772      -0.689      -0.483
```

```
reg_basal = lm(DOCVIS~FEMALE+HANDPER+HOSPVIS+DOCTOR+HEALTHY+HOSPITAL+
                NEWHSAT+TIME,data=hc2)
```

- Then we try to test whether there are unusual observations.

```
# outlier
qqPlot(reg_basal,id=list(n=3))
```

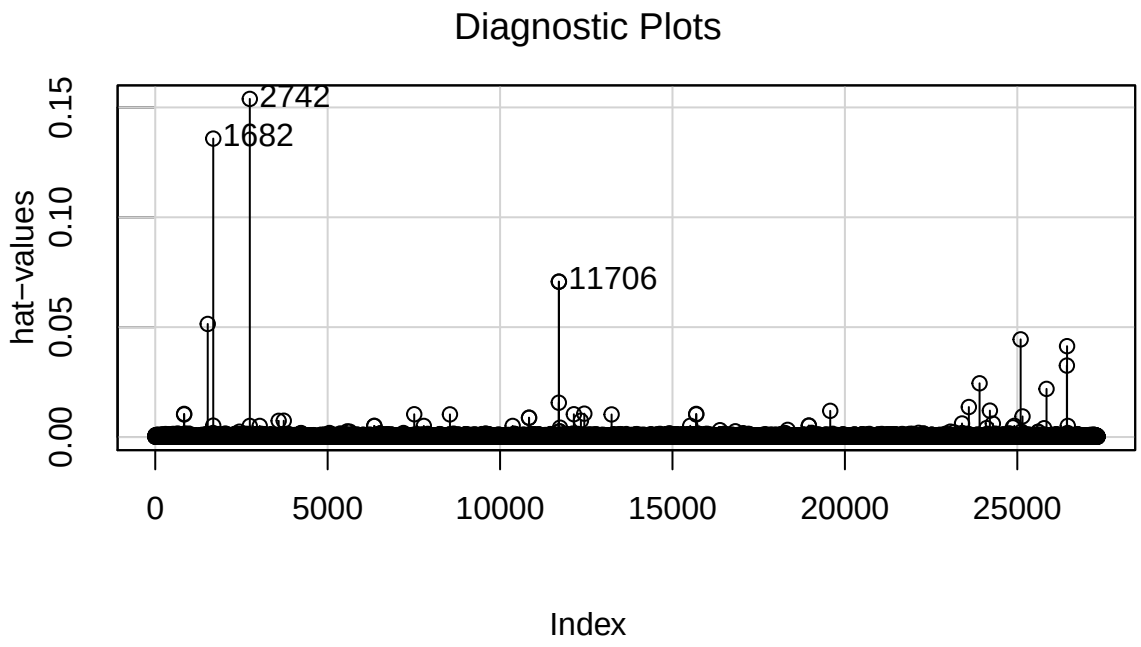



```
## 4429 7193 7194
## 4409 7173 7174
```

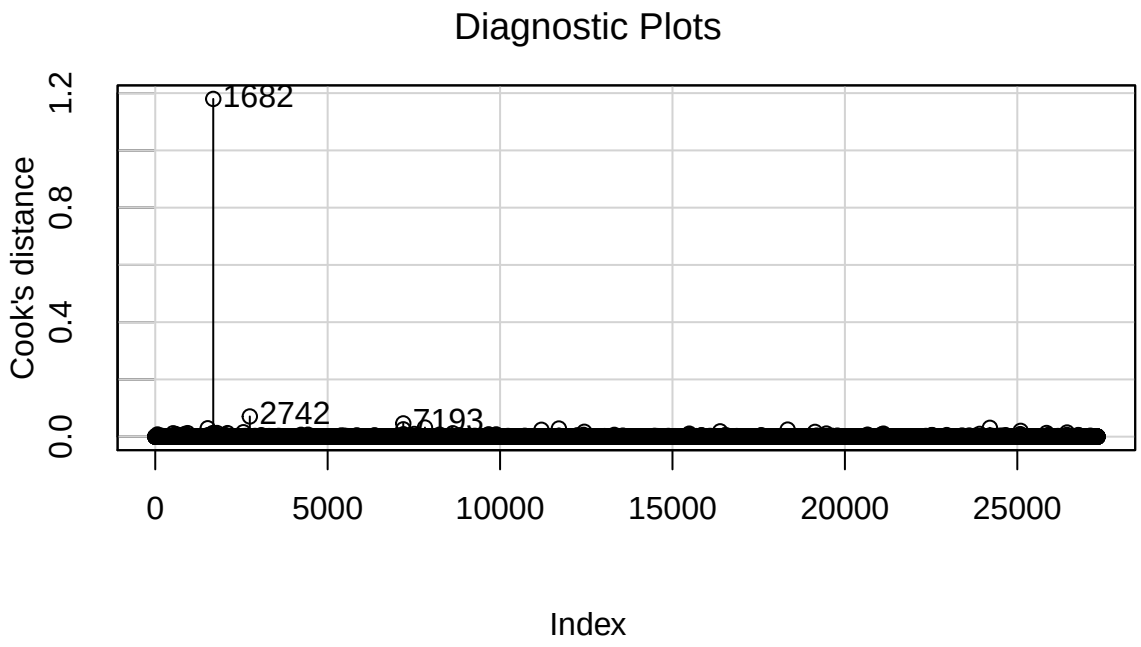
```
## 4429 7193 and 7194 are potential outliers
outlierTest(reg_basal)
```

```
##       rstudent unadjusted p-value Bonferroni p
## 7193      22.9          8.08e-115    2.21e-110
## 7194      18.7           1.84e-77     5.01e-73
## 4429      17.6           4.97e-69     1.36e-64
## 11200     17.6           5.37e-69     1.47e-64
## 7822      17.4           3.63e-67     9.91e-63
## 16382     17.2           8.59e-66     2.34e-61
## 19133     15.9           1.97e-56     5.37e-52
## 15484     15.4           2.48e-53     6.78e-49
## 596       15.3           1.32e-52     3.59e-48
## 22952     15.1           3.33e-51     9.09e-47
```

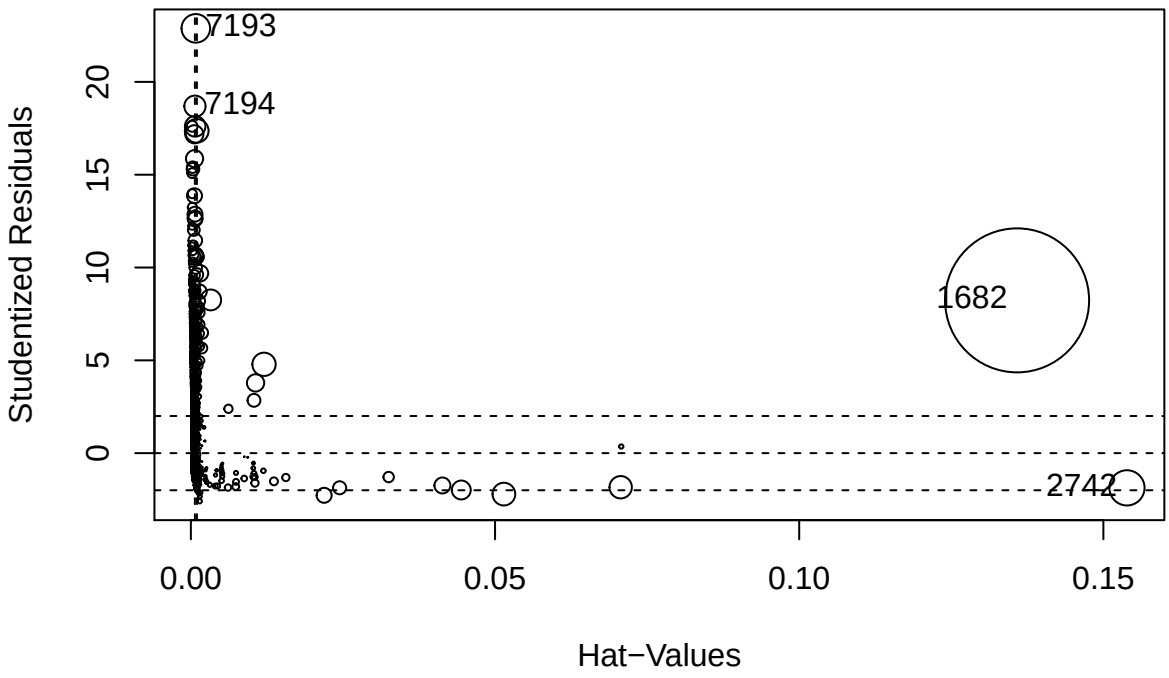
```
## 7193 7194 4429 11200 16382 7822 596 19133 15484 22952 are tested outliers

# leverage
influenceIndexPlot(reg_basal,id=list(n=3),vars="hat")
```




```
## 1682 2742 11706 are leverages

# influence
influenceIndexPlot(reg_basal,id=list(n=3),vars="cook")
```




```
## 1682 2742 7193 are influences.

# Depict the influence
influencePlot(reg_basal)
```




```
##      StudRes      Hat  CookD
## 1682    8.23 0.135821 1.1799
## 2742   -1.87 0.153878 0.0707
## 7193   22.88 0.000812 0.0464
## 7194   18.69 0.000668 0.0256
```

```
AIC(reg_basal)
```

```
## [1] 163744
```

- In all, we will try to delete these observations and compare this model to the basic model.

```
reg_nounusal =update(reg_basal,subset=-c(7193,7194,4429,11200,16382,
                                         7822,19133,15484,596,22952,
                                        25976,25981,1682,2742,11706))
AIC(reg_nounusal,reg_basal)
```

```
## Warning in AIC.default(reg_nounusal, reg_basal): models are not all fitted
## to the same number of observations
```

```
##              df    AIC
## reg_nounusal 10 163666
## reg_basal    10 163744
```

```
BIC(reg_nounusal,reg_basal)
```

```
## Warning in BIC.default(reg_nounusal, reg_basal): models are not all fitted
## to the same number of observations
```

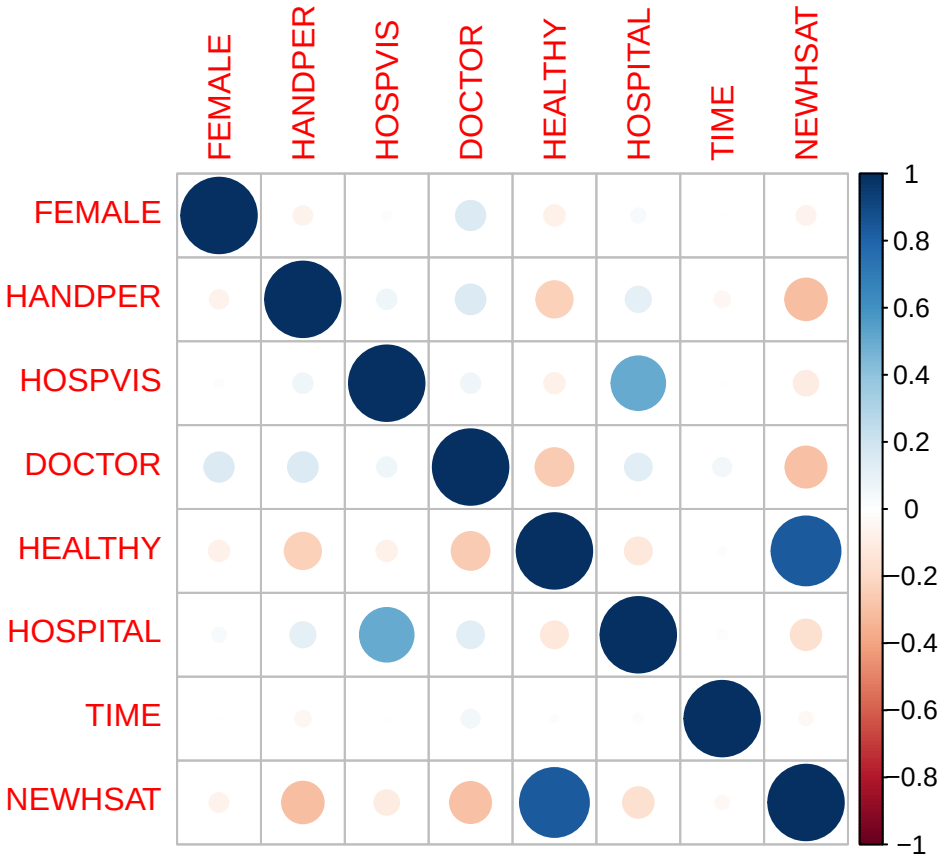
```
##              df    BIC
## reg_nounusal 10 163748
## reg_basal    10 163827
```

- We can see that the model deleting the unusual observations are better. Then we try to test the multi-coolinearity.

```
coef(reg_nounusal)
```

```
## (Intercept)      FEMALE     HANDPER     HOSPVIS      DOCTOR     HEALTHY
##       4.616       0.456       0.023       0.249       3.925       0.684
##    HOSPITAL     NEWHSAT        TIME
##       1.771      -0.689      -0.483
```

```
corrplot(cor(hc2[,c("FEMALE","HANDPER","HOSPVIS","DOCTOR","HEALTHY",
"HOSPITAL","TIME","NEWHSAT")]))
```



```
vif(reg_nounusal)
```

```
##   FEMALE  HANDPER  HOSPVIS   DOCTOR  HEALTHY HOSPITAL  NEWHSAT     TIME
##     1.03     1.13     1.34     1.13     3.26     1.38     3.50     1.01
```

- From the correlation plot and the VIF values we suggest that HEALTHY and NEWHSAT maybe highly correlated. We remove one of it and see whether there is improvement.

```
reg_nomc1 = update(reg_nounusal,.~.-HEALTHY,data=hc2)
reg_nomc2 = update(reg_nounusal,.~.-NEWHSAT,data=hc2)
AIC(reg_nomc1,reg_nomc2,reg_nounusal)
```

```
##              df    AIC
## reg_nomc1     9 163703
## reg_nomc2     9 164478
## reg_nounusal 10 163666
```

```
BIC(reg_nomc1,reg_nomc2,reg_nounusal)
```

```
##              df    BIC
## reg_nomc1     9 163777
## reg_nomc2     9 164552
## reg_nounusal 10 163748
```

-We find that removing these terms do not make the resutls better. Then we try the transformation.

```
t = powerTransform(with(hc2,cbind(HANDPER,HOSPVIS,NEWHSAT))~1,family="yjPower")
t_s = yjPower(with(hc2, cbind(HANDPER,HOSPVIS,NEWHSAT)),coef(t,round=TRUE))
hc3 = cbind(hc2,t_s)
reg_trans = lm(DOCVIS~FEMALE+HANDPER+HOSPVIS+DOCTOR+HEALTHY+HOSPITAL+
                 NEWHSAT+TIME,data=hc3)
AIC(reg_trans,reg_nounusal)
```

```
##              df    AIC
## reg_trans    10 163744
## reg_nounusal 10 163666
```

```
BIC(reg_trans,reg_nounusal)
```

```
##              df    BIC
## reg_trans    10 163827
## reg_nounusal 10 163748
```

- The AIC and BIC show that we do not need the transform.
- The component residual plots also show that there is a linearity relationship between dependent variables and continuous independent variables. Then we use a resettest to decide if we need to add a higher degree polynomial.

```
resettest(reg_nounusal,power=2,type="regressor")
```

```
##
##  RESET test
##
## data:  reg_nounusal
## RESET = 66, df1 = 8, df2 = 27269, p-value <2e-16
```

```
resettest(reg_nounusal,power=3,type="regressor")
```

```
##
##  RESET test
##
## data:  reg_nounusal
## RESET = 65, df1 = 8, df2 = 27269, p-value <2e-16
```

- According to the ramsey RESET test, we need to add a higher degree polynomial.

```
coef(reg_nounusal)
```

```
## (Intercept)      FEMALE     HANDPER     HOSPVIS      DOCTOR     HEALTHY
##       4.616       0.456       0.023       0.249       3.925       0.684
##    HOSPITAL     NEWHSAT        TIME
##       1.771      -0.689      -0.483
```

```
reg_inter2 = lm(DOCVIS~FEMALE+HANDPER+HOSPVIS+DOCTOR+HEALTHY+HOSPITAL+NEWHSAT+TIME+I(HOSPVIS^2),data=hc
summary(reg_inter2)
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + TIME + I(HOSPVIS^2), data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12.51  -2.23  -0.48   0.90 109.94
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.61606    0.14195   32.52  < 2e-16 ***
## FEMALE        0.45641    0.05982    7.63  2.4e-14 ***
## HANDPER       0.02304    0.00162   14.23  < 2e-16 ***
## HOSPVIS       0.14984    0.08146    1.84    0.066 .
## DOCTOR        3.92529    0.06477   60.60  < 2e-16 ***
## HEALTHY       0.68589    0.10878    6.31  2.9e-10 ***
## HOSPITAL      1.89793    0.15195   12.49  < 2e-16 ***
## NEWHSAT      -0.68888    0.02395  -28.77  < 2e-16 ***
## TIME         -0.48256    0.05934   -8.13  4.4e-16 ***
## I(HOSPVIS^2)  0.00332    0.00239    1.39    0.165
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.85 on 27291 degrees of freedom
## Multiple R-squared:  0.273,  Adjusted R-squared:  0.273
## F-statistic: 1.14e+03 on 9 and 27291 DF,  p-value: <2e-16
```

```
reg_inter3 = lm(DOCVIS~FEMALE+HANDPER+HOSPVIS+DOCTOR+HEALTHY+HOSPITAL+NEWHSAT+TIME+I(NEWHSAT^2),data=hc
summary(reg_inter3)
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + TIME + I(NEWHSAT^2), data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14.67  -2.12  -0.33   0.60 108.60
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.25908    0.18632   38.96  < 2e-16 ***
## FEMALE        0.43304    0.05932    7.30  3.0e-13 ***
## HANDPER       0.01865    0.00162   11.53  < 2e-16 ***
## HOSPVIS       0.23064    0.03816    6.04  1.5e-09 ***
## DOCTOR        4.01707    0.06437   62.41  < 2e-16 ***
## HEALTHY       0.21202    0.11004    1.93    0.054 .
## HOSPITAL      1.63602    0.12093   13.53  < 2e-16 ***
## NEWHSAT      -1.77222    0.05540  -31.99  < 2e-16 ***
## TIME         -0.40304    0.05895   -6.84  8.3e-12 ***
## I(NEWHSAT^2)  0.09668    0.00447   21.65  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.81 on 27291 degrees of freedom
## Multiple R-squared:  0.285,  Adjusted R-squared:  0.285
## F-statistic: 1.21e+03 on 9 and 27291 DF,  p-value: <2e-16
```

```
reg_inter4 = lm(DOCVIS~FEMALE+HANDPER+HOSPVIS+DOCTOR+HEALTHY+HOSPITAL+NEWHSAT+TIME+I(HANDPER^2)+I(HOSPV
summary(reg_inter4)
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + TIME + I(HANDPER^2) + I(HOSPVIS^2),
##     data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11.86  -2.20  -0.51   0.92 109.51
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.54e+00   1.42e-01   31.86  < 2e-16 ***
```

```
## FEMALE        4.82e-01   5.99e-02    8.04  9.5e-16 ***
## HANDPER       5.44e-02   5.28e-03   10.30  < 2e-16 ***
## HOSPVIS       1.57e-01   8.14e-02    1.92    0.054 .
## DOCTOR        3.91e+00   6.48e-02   60.41  < 2e-16 ***
## HEALTHY       7.02e-01   1.09e-01    6.46  1.1e-10 ***
## HOSPITAL      1.89e+00   1.52e-01   12.45  < 2e-16 ***
## NEWHSAT      -6.89e-01   2.39e-02  -28.78  < 2e-16 ***
## TIME         -4.43e-01   5.96e-02   -7.42  1.2e-13 ***
## I(HANDPER^2) -4.17e-04   6.68e-05   -6.24  4.5e-10 ***
## I(HOSPVIS^2)  3.16e-03   2.39e-03    1.32    0.186
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.85 on 27290 degrees of freedom
## Multiple R-squared:  0.274,  Adjusted R-squared:  0.274
## F-statistic: 1.03e+03 on 10 and 27290 DF,  p-value: <2e-16
```

```
reg_inter5 = lm(DOCVIS~FEMALE+HANDPER+HOSPVIS+DOCTOR+HEALTHY+HOSPITAL+NEWHSAT+TIME+I(HANDPER^2)+I(NEWHS
summary(reg_inter5)
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + TIME + I(HANDPER^2) + I(NEWHSAT^2),
##     data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13.43  -2.12  -0.36   0.64 108.07
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.21e+00   1.86e-01   38.71  < 2e-16 ***
## FEMALE        4.62e-01   5.94e-02    7.78  7.5e-15 ***
## HANDPER       5.53e-02   5.24e-03   10.55  < 2e-16 ***
## HOSPVIS       2.33e-01   3.81e-02    6.10  1.0e-09 ***
## DOCTOR        4.00e+00   6.43e-02   62.23  < 2e-16 ***
## HEALTHY       2.24e-01   1.10e-01    2.04    0.042 *
## HOSPITAL      1.63e+00   1.21e-01   13.52  < 2e-16 ***
## NEWHSAT      -1.79e+00   5.54e-02  -32.30  < 2e-16 ***
## TIME         -3.55e-01   5.93e-02   -5.99  2.1e-09 ***
## I(HANDPER^2) -4.88e-04   6.63e-05   -7.35  2.0e-13 ***
## I(NEWHSAT^2)  9.82e-02   4.47e-03   22.00  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.81 on 27290 degrees of freedom
## Multiple R-squared:  0.286,  Adjusted R-squared:  0.286
## F-statistic: 1.1e+03 on 10 and 27290 DF,  p-value: <2e-16
```

```
reg_inter6 = lm(DOCVIS~FEMALE+HANDPER+HOSPVIS+DOCTOR+HEALTHY+HOSPITAL+NEWHSAT+TIME+I(HOSPVIS^2)+I(NEWHS
summary(reg_inter6)
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + TIME + I(HOSPVIS^2) + I(NEWHSAT^2),
##     data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14.62  -2.14  -0.33   0.60 108.56
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.26066    0.18632   38.97  < 2e-16 ***
## FEMALE        0.43341    0.05932    7.31  2.8e-13 ***
## HANDPER       0.01873    0.00162   11.58  < 2e-16 ***
## HOSPVIS       0.12044    0.08078    1.49    0.136
## DOCTOR        4.01737    0.06436   62.42  < 2e-16 ***
## HEALTHY       0.21353    0.11004    1.94    0.052 .
## HOSPITAL      1.77541    0.15077   11.78  < 2e-16 ***
## NEWHSAT      -1.77310    0.05540  -32.01  < 2e-16 ***
## TIME         -0.40252    0.05895   -6.83  8.8e-12 ***
## I(HOSPVIS^2)  0.00367    0.00237    1.55    0.122
## I(NEWHSAT^2)  0.09673    0.00447   21.66  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.81 on 27290 degrees of freedom
## Multiple R-squared:  0.285,  Adjusted R-squared:  0.285
## F-statistic: 1.09e+03 on 10 and 27290 DF,  p-value: <2e-16
```

```
reg_inter7 = lm(DOCVIS~FEMALE+HANDPER+HOSPVIS+DOCTOR+HEALTHY+HOSPITAL+NEWHSAT+TIME+I(HANDPER^2)+I(HOSPV
summary(reg_inter7)
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + TIME + I(HANDPER^2) + I(HOSPVIS^2) +
##     I(NEWHSAT^2), data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13.38  -2.12  -0.36   0.64 108.03
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.21e+00   1.86e-01   38.72  < 2e-16 ***
## FEMALE        4.62e-01   5.94e-02    7.79  7.2e-15 ***
## HANDPER       5.53e-02   5.24e-03   10.55  < 2e-16 ***
## HOSPVIS       1.28e-01   8.07e-02    1.59    0.113
## DOCTOR        4.00e+00   6.43e-02   62.24  < 2e-16 ***
## HEALTHY       2.25e-01   1.10e-01    2.05    0.041 *
## HOSPITAL      1.77e+00   1.51e-01   11.72  < 2e-16 ***
## NEWHSAT      -1.79e+00   5.54e-02  -32.32  < 2e-16 ***
## TIME         -3.55e-01   5.93e-02   -5.99  2.2e-09 ***
```

```
## I(HANDPER^2) -4.87e-04   6.63e-05   -7.34  2.3e-13 ***
## I(HOSPVIS^2)  3.49e-03   2.37e-03    1.47    0.141
## I(NEWHSAT^2)  9.83e-02   4.47e-03   22.01  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.81 on 27289 degrees of freedom
## Multiple R-squared:  0.286,  Adjusted R-squared:  0.286
## F-statistic:  996 on 11 and 27289 DF,  p-value: <2e-16
```

```
# all the quadratic terms will either be insignificant or make others insignificant.
```

- Then we do a manual way to choose our 10 variables, since we want to keep the main effect variables, so I use the manual backward way.

```
coef(reg_nounusal)
```

```
## (Intercept)      FEMALE     HANDPER     HOSPVIS      DOCTOR     HEALTHY
##       4.616       0.456       0.023       0.249       3.925       0.684
##    HOSPITAL     NEWHSAT        TIME
##       1.771      -0.689      -0.483
```

```
reg_candidate = lm(DOCVIS~(FEMALE+HANDPER+HOSPVIS+DOCTOR+HEALTHY+HOSPITAL+NEWHSAT+TIME)^2,data=hc2)
summary(reg_candidate)
```

```
##
## Call:
## lm(formula = DOCVIS ~ (FEMALE + HANDPER + HOSPVIS + DOCTOR +
##     HEALTHY + HOSPITAL + NEWHSAT + TIME)^2, data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -16.67  -1.82  -0.21   0.26 107.22
##
## Coefficients: (1 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.98e+00   3.52e-01    5.61  2.0e-08 ***
## FEMALE            5.87e-01   2.78e-01    2.12  0.03435 *
## HANDPER          -2.29e-02   6.29e-03   -3.64  0.00027 ***
## HOSPVIS           1.43e-01   1.49e-01    0.96  0.33852
## DOCTOR            8.45e+00   3.14e-01   26.92  < 2e-16 ***
## HEALTHY          -3.61e+00   3.87e-01   -9.32  < 2e-16 ***
## HOSPITAL          2.45e+00   5.04e-01    4.87  1.1e-06 ***
## NEWHSAT          -4.33e-01   6.58e-02   -6.57  5.0e-11 ***
## TIME             -5.62e-02   2.75e-01   -0.20  0.83828
## FEMALE:HANDPER    7.16e-03   3.24e-03    2.21  0.02707 *
## FEMALE:HOSPVIS   -8.43e-02   8.44e-02   -1.00  0.31745
## FEMALE:DOCTOR     5.76e-01   1.28e-01    4.50  6.8e-06 ***
## FEMALE:HEALTHY   -1.85e-01   2.16e-01   -0.86  0.39229
## FEMALE:HOSPITAL  -1.06e-01   2.46e-01   -0.43  0.66593
## FEMALE:NEWHSAT   -5.23e-02   4.76e-02   -1.10  0.27200
## FEMALE:TIME      -1.09e-01   1.18e-01   -0.92  0.35746
```

```
## HANDPER:HOSPVIS   3.26e-03   1.90e-03    1.71  0.08664 .
## HANDPER:DOCTOR    2.56e-02   4.59e-03    5.57  2.6e-08 ***
## HANDPER:HEALTHY  -7.68e-04   5.41e-03   -0.14  0.88724
## HANDPER:HOSPITAL  1.36e-03   5.37e-03    0.25  0.80006
## HANDPER:NEWHSAT   3.19e-03   9.60e-04    3.32  0.00091 ***
## HANDPER:TIME     -1.72e-03   3.19e-03   -0.54  0.58943
## HOSPVIS:DOCTOR    1.66e-01   1.14e-01    1.46  0.14298
## HOSPVIS:HEALTHY  -3.89e-02   1.56e-01   -0.25  0.80369
## HOSPVIS:HOSPITAL        NA         NA      NA       NA
## HOSPVIS:NEWHSAT   3.46e-05   2.64e-02    0.00  0.99896
## HOSPVIS:TIME     -1.85e-01   9.01e-02   -2.05  0.04002 *
## DOCTOR:HEALTHY    1.11e-01   2.42e-01    0.46  0.64623
## DOCTOR:HOSPITAL   1.52e+00   3.22e-01    4.71  2.5e-06 ***
## DOCTOR:NEWHSAT   -6.45e-01   5.68e-02  -11.37  < 2e-16 ***
## DOCTOR:TIME      -6.19e-01   1.28e-01   -4.83  1.4e-06 ***
## HEALTHY:HOSPITAL  4.30e-01   4.34e-01    0.99  0.32106
## HEALTHY:NEWHSAT   6.22e-01   5.38e-02   11.57  < 2e-16 ***
## HEALTHY:TIME      3.87e-02   2.15e-01    0.18  0.85690
## HOSPITAL:NEWHSAT -3.69e-01   8.16e-02   -4.52  6.3e-06 ***
## HOSPITAL:TIME    -1.02e-01   2.48e-01   -0.41  0.68195
## NEWHSAT:TIME      1.93e-02   4.77e-02    0.41  0.68513
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.75 on 27265 degrees of freedom
## Multiple R-squared:  0.305,  Adjusted R-squared:  0.304
## F-statistic:  342 on 35 and 27265 DF,  p-value: <2e-16
```

```
reg_candidate = update(reg_candidate,.~.-HOSPVIS:HOSPITAL) # remove perfect multicoolinearity
summary(reg_candidate)
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + TIME + FEMALE:HANDPER + FEMALE:HOSPVIS +
##     FEMALE:DOCTOR + FEMALE:HEALTHY + FEMALE:HOSPITAL + FEMALE:NEWHSAT +
##     FEMALE:TIME + HANDPER:HOSPVIS + HANDPER:DOCTOR + HANDPER:HEALTHY +
##     HANDPER:HOSPITAL + HANDPER:NEWHSAT + HANDPER:TIME + HOSPVIS:DOCTOR +
##     HOSPVIS:HEALTHY + HOSPVIS:NEWHSAT + HOSPVIS:TIME + DOCTOR:HEALTHY +
##     DOCTOR:HOSPITAL + DOCTOR:NEWHSAT + DOCTOR:TIME + HEALTHY:HOSPITAL +
##     HEALTHY:NEWHSAT + HEALTHY:TIME + HOSPITAL:NEWHSAT + HOSPITAL:TIME +
##     NEWHSAT:TIME, data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -16.67  -1.82  -0.21   0.26 107.22
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.98e+00   3.52e-01    5.61  2.0e-08 ***
## FEMALE            5.87e-01   2.78e-01    2.12  0.03435 *
## HANDPER          -2.29e-02   6.29e-03   -3.64  0.00027 ***
## HOSPVIS           1.43e-01   1.49e-01    0.96  0.33852
## DOCTOR            8.45e+00   3.14e-01   26.92  < 2e-16 ***
```

```
## HEALTHY          -3.61e+00   3.87e-01   -9.32  < 2e-16 ***
## HOSPITAL          2.45e+00   5.04e-01    4.87  1.1e-06 ***
## NEWHSAT          -4.33e-01   6.58e-02   -6.57  5.0e-11 ***
## TIME             -5.62e-02   2.75e-01   -0.20  0.83828
## FEMALE:HANDPER    7.16e-03   3.24e-03    2.21  0.02707 *
## FEMALE:HOSPVIS   -8.43e-02   8.44e-02   -1.00  0.31745
## FEMALE:DOCTOR     5.76e-01   1.28e-01    4.50  6.8e-06 ***
## FEMALE:HEALTHY   -1.85e-01   2.16e-01   -0.86  0.39229
## FEMALE:HOSPITAL  -1.06e-01   2.46e-01   -0.43  0.66593
## FEMALE:NEWHSAT   -5.23e-02   4.76e-02   -1.10  0.27200
## FEMALE:TIME      -1.09e-01   1.18e-01   -0.92  0.35746
## HANDPER:HOSPVIS   3.26e-03   1.90e-03    1.71  0.08664 .
## HANDPER:DOCTOR    2.56e-02   4.59e-03    5.57  2.6e-08 ***
## HANDPER:HEALTHY  -7.68e-04   5.41e-03   -0.14  0.88724
## HANDPER:HOSPITAL  1.36e-03   5.37e-03    0.25  0.80006
## HANDPER:NEWHSAT   3.19e-03   9.60e-04    3.32  0.00091 ***
## HANDPER:TIME     -1.72e-03   3.19e-03   -0.54  0.58943
## HOSPVIS:DOCTOR    1.66e-01   1.14e-01    1.46  0.14298
## HOSPVIS:HEALTHY  -3.89e-02   1.56e-01   -0.25  0.80369
## HOSPVIS:NEWHSAT   3.46e-05   2.64e-02    0.00  0.99896
## HOSPVIS:TIME     -1.85e-01   9.01e-02   -2.05  0.04002 *
## DOCTOR:HEALTHY    1.11e-01   2.42e-01    0.46  0.64623
## DOCTOR:HOSPITAL   1.52e+00   3.22e-01    4.71  2.5e-06 ***
## DOCTOR:NEWHSAT   -6.45e-01   5.68e-02  -11.37  < 2e-16 ***
## DOCTOR:TIME      -6.19e-01   1.28e-01   -4.83  1.4e-06 ***
## HEALTHY:HOSPITAL  4.30e-01   4.34e-01    0.99  0.32106
## HEALTHY:NEWHSAT   6.22e-01   5.38e-02   11.57  < 2e-16 ***
## HEALTHY:TIME      3.87e-02   2.15e-01    0.18  0.85690
## HOSPITAL:NEWHSAT -3.69e-01   8.16e-02   -4.52  6.3e-06 ***
## HOSPITAL:TIME    -1.02e-01   2.48e-01   -0.41  0.68195
## NEWHSAT:TIME      1.93e-02   4.77e-02    0.41  0.68513
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.75 on 27265 degrees of freedom
## Multiple R-squared:  0.305,  Adjusted R-squared:  0.304
## F-statistic:  342 on 35 and 27265 DF,  p-value: <2e-16
```

```
reg_candidate = update(reg_candidate,.~.-FEMALE:HOSPVIS-FEMALE:HEALTHY-FEMALE:HOSPITAL-FEMALE:NEWHSAT-FE
summary(reg_candidate)
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + TIME + FEMALE:HANDPER + FEMALE:DOCTOR +
##     HANDPER:HOSPVIS + HANDPER:DOCTOR + HANDPER:NEWHSAT + HOSPVIS:TIME +
##     DOCTOR:HOSPITAL + DOCTOR:NEWHSAT + DOCTOR:TIME + HEALTHY:HOSPITAL +
##     HEALTHY:NEWHSAT + HOSPITAL:NEWHSAT, data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -16.52  -1.82  -0.21   0.26 107.23
##
## Coefficients:
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.216775   0.287632    7.71  1.3e-14 ***
## FEMALE           -0.011182   0.097857   -0.11  0.90903
## HANDPER          -0.025144   0.005760   -4.37  1.3e-05 ***
## HOSPVIS           0.252454   0.045166    5.59  2.3e-08 ***
## DOCTOR            8.443654   0.263752   32.01  < 2e-16 ***
## HEALTHY          -3.626376   0.356477  -10.17  < 2e-16 ***
## HOSPITAL          2.220417   0.421471    5.27  1.4e-07 ***
## NEWHSAT          -0.460630   0.047071   -9.79  < 2e-16 ***
## TIME              0.070332   0.096635    0.73  0.46673
## FEMALE:HANDPER    0.009487   0.003096    3.06  0.00218 **
## FEMALE:DOCTOR     0.649335   0.123356    5.26  1.4e-07 ***
## HANDPER:HOSPVIS   0.003904   0.001489    2.62  0.00875 **
## HANDPER:DOCTOR    0.025391   0.004580    5.54  3.0e-08 ***
## HANDPER:NEWHSAT   0.003217   0.000631    5.10  3.4e-07 ***
## HOSPVIS:TIME     -0.258116   0.068210   -3.78  0.00015 ***
## DOCTOR:HOSPITAL   1.742995   0.274393    6.35  2.2e-10 ***
## DOCTOR:NEWHSAT   -0.635444   0.032365  -19.63  < 2e-16 ***
## DOCTOR:TIME      -0.680785   0.121046   -5.62  1.9e-08 ***
## HEALTHY:HOSPITAL  0.399291   0.359856    1.11  0.26719
## HEALTHY:NEWHSAT   0.624597   0.052181   11.97  < 2e-16 ***
## HOSPITAL:NEWHSAT -0.376351   0.069174   -5.44  5.4e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.75 on 27280 degrees of freedom
## Multiple R-squared:  0.305,  Adjusted R-squared:  0.304
## F-statistic:  597 on 20 and 27280 DF,  p-value: <2e-16
```

```
reg_candidate = update(reg_candidate,.~.-FEMALE:DOCTOR-FEMALE:HANDPER) # remove interactions with femal
summary(reg_candidate)
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + TIME + HANDPER:HOSPVIS + HANDPER:DOCTOR +
##     HANDPER:NEWHSAT + HOSPVIS:TIME + DOCTOR:HOSPITAL + DOCTOR:NEWHSAT +
##     DOCTOR:TIME + HEALTHY:HOSPITAL + HEALTHY:NEWHSAT + HOSPITAL:NEWHSAT,
##     data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -16.02  -1.83  -0.22   0.30 106.88
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.99222    0.28541    6.98  3.0e-12 ***
## FEMALE            0.46895    0.05862    8.00  1.3e-15 ***
## HANDPER          -0.02012    0.00562   -3.58  0.00035 ***
## HOSPVIS           0.25006    0.04520    5.53  3.2e-08 ***
## DOCTOR            8.78005    0.25543   34.37  < 2e-16 ***
## HEALTHY          -3.62569    0.35674  -10.16  < 2e-16 ***
## HOSPITAL          2.21646    0.42167    5.26  1.5e-07 ***
## NEWHSAT          -0.45750    0.04710   -9.71  < 2e-16 ***
```

```
## TIME              0.08003    0.09670    0.83  0.40789
## HANDPER:HOSPVIS   0.00427    0.00149    2.87  0.00412 **
## HANDPER:DOCTOR    0.02438    0.00457    5.34  9.4e-08 ***
## HANDPER:NEWHSAT   0.00306    0.00063    4.86  1.2e-06 ***
## HOSPVIS:TIME     -0.26222    0.06826   -3.84  0.00012 ***
## DOCTOR:HOSPITAL   1.77760    0.27453    6.48  9.6e-11 ***
## DOCTOR:NEWHSAT   -0.64245    0.03236  -19.85  < 2e-16 ***
## DOCTOR:TIME      -0.69094    0.12112   -5.70  1.2e-08 ***
## HEALTHY:HOSPITAL  0.41292    0.36010    1.15  0.25152
## HEALTHY:NEWHSAT   0.62627    0.05222   11.99  < 2e-16 ***
## HOSPITAL:NEWHSAT -0.38071    0.06921   -5.50  3.8e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.75 on 27282 degrees of freedom
## Multiple R-squared:  0.303,  Adjusted R-squared:  0.303
## F-statistic:  660 on 18 and 27282 DF,  p-value: <2e-16
```

```
reg_final = regsubsets(formula(reg_candidate),nvmax=15,nbest=1000,method="backward",data=hc2)
```

```
## Warning in obj$ress < obj$rss * (1 - 1e-08):
##
```

```
## Warning in warn.extra(rval): model with initial
## (3,4,5,6,7,8,9,10,11,12,13,6,7,8,9,10,11,12,13,9,10,11,12,13,14,15,12,13,2)
## variables was better, and is reported
```

```
reg_final_s = summary(reg_final)
which.min(reg_final_s$cp)
```

```
## [1] 91
```

```
coef(reg_final,91)
```

```
##      (Intercept)           FEMALE          HANDPER          HOSPVIS
##          2.03061          0.47240         -0.01862          0.27978
##           DOCTOR          HEALTHY         HOSPITAL          NEWHSAT
##          8.71351         -3.57658          2.18191         -0.46200
##   HANDPER:DOCTOR  HANDPER:NEWHSAT     HOSPVIS:TIME  DOCTOR:HOSPITAL
##          0.02542          0.00283         -0.24807          1.79682
##   DOCTOR:NEWHSAT      DOCTOR:TIME  HEALTHY:NEWHSAT HOSPITAL:NEWHSAT
##         -0.63874         -0.61122          0.62465         -0.34004
```

```
reg_final = lm(DOCVIS~FEMALE+HANDPER+HOSPVIS+DOCTOR+HEALTHY+HOSPITAL+NEWHSAT+HANDPER:DOCTOR+HANDPER:NEW
summary(reg_final)
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + HANDPER:DOCTOR + HANDPER:NEWHSAT + HOSPVIS:TIME +
##     DOCTOR:HOSPITAL + DOCTOR:NEWHSAT + DOCTOR:TIME + HEALTHY:NEWHSAT +
```

```
##     HOSPITAL:NEWHSAT, data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -15.43  -1.83  -0.23   0.28 106.96
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.030606   0.279128    7.27  3.6e-13 ***
## FEMALE            0.472405   0.058605    8.06  7.9e-16 ***
## HANDPER          -0.018620   0.005590   -3.33  0.00087 ***
## HOSPVIS           0.279777   0.044041    6.35  2.2e-10 ***
## DOCTOR            8.713507   0.250600   34.77  < 2e-16 ***
## HEALTHY          -3.576577   0.356073  -10.04  < 2e-16 ***
## HOSPITAL          2.181912   0.387615    5.63  1.8e-08 ***
## NEWHSAT          -0.462005   0.046149  -10.01  < 2e-16 ***
## HANDPER:DOCTOR    0.025418   0.004551    5.59  2.4e-08 ***
## HANDPER:NEWHSAT   0.002833   0.000626    4.53  6.0e-06 ***
## HOSPVIS:TIME     -0.248074   0.068097   -3.64  0.00027 ***
## DOCTOR:HOSPITAL   1.796819   0.274106    6.56  5.7e-11 ***
## DOCTOR:NEWHSAT   -0.638735   0.032333  -19.76  < 2e-16 ***
## DOCTOR:TIME      -0.611222   0.073925   -8.27  < 2e-16 ***
## HEALTHY:NEWHSAT   0.624651   0.051835   12.05  < 2e-16 ***
## HOSPITAL:NEWHSAT -0.340042   0.041024   -8.29  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.75 on 27285 degrees of freedom
## Multiple R-squared:  0.303,  Adjusted R-squared:  0.303
## F-statistic:  791 on 15 and 27285 DF,  p-value: <2e-16
```

```
reg_final = update(reg_final,.~.-HOSPVIS:TIME-DOCTOR:TIME,data=hc2)
summary(reg_final)
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + HANDPER:DOCTOR + HANDPER:NEWHSAT + DOCTOR:HOSPITAL +
##     DOCTOR:NEWHSAT + HEALTHY:NEWHSAT + HOSPITAL:NEWHSAT, data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14.90  -1.92  -0.22   0.31 107.39
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.098347   0.279512    7.51  6.2e-14 ***
## FEMALE            0.475399   0.058693    8.10  5.7e-16 ***
## HANDPER          -0.019974   0.005598   -3.57  0.00036 ***
## HOSPVIS           0.199411   0.037831    5.27  1.4e-07 ***
## DOCTOR            8.382533   0.247983   33.80  < 2e-16 ***
## HEALTHY          -3.722429   0.356339  -10.45  < 2e-16 ***
## HOSPITAL          2.093991   0.387976    5.40  6.8e-08 ***
## NEWHSAT          -0.473647   0.046211  -10.25  < 2e-16 ***
```

```
## HANDPER:DOCTOR    0.026452   0.004557    5.80  6.5e-09 ***
## HANDPER:NEWHSAT   0.003063   0.000627    4.89  1.0e-06 ***
## DOCTOR:HOSPITAL   1.799786   0.274568    6.55  5.7e-11 ***
## DOCTOR:NEWHSAT   -0.633504   0.032383  -19.56  < 2e-16 ***
## HEALTHY:NEWHSAT   0.644870   0.051879   12.43  < 2e-16 ***
## HOSPITAL:NEWHSAT -0.335651   0.041090   -8.17  3.3e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.76 on 27287 degrees of freedom
## Multiple R-squared:  0.301,  Adjusted R-squared:   0.3
## F-statistic:  903 on 13 and 27287 DF,  p-value: <2e-16
```

```
reg_final_1 = update(reg_final,.~.-HANDPER:NEWHSAT,data=hc2)
summary(reg_final_1) # insignificant
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + HANDPER:DOCTOR + DOCTOR:HOSPITAL + DOCTOR:NEWHSAT +
##     HEALTHY:NEWHSAT + HOSPITAL:NEWHSAT, data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -15.92  -1.91  -0.22   0.30 107.15
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.74497    0.27012    6.46  1.1e-10 ***
## FEMALE             0.46786    0.05870    7.97  1.6e-15 ***
## HANDPER           -0.00159    0.00415   -0.38      0.7
## HOSPVIS            0.20052    0.03785    5.30  1.2e-07 ***
## DOCTOR             8.38431    0.24809   33.80  < 2e-16 ***
## HEALTHY           -3.35337    0.34840   -9.63  < 2e-16 ***
## HOSPITAL           2.01730    0.38782    5.20  2.0e-07 ***
## NEWHSAT           -0.40631    0.04413   -9.21  < 2e-16 ***
## HANDPER:DOCTOR     0.02240    0.00448    5.00  5.9e-07 ***
## DOCTOR:HOSPITAL    1.82069    0.27465    6.63  3.4e-11 ***
## DOCTOR:NEWHSAT    -0.62989    0.03239  -19.45  < 2e-16 ***
## HEALTHY:NEWHSAT    0.57760    0.05004   11.54  < 2e-16 ***
## HOSPITAL:NEWHSAT -0.32517    0.04105   -7.92  2.4e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.76 on 27288 degrees of freedom
## Multiple R-squared:   0.3,   Adjusted R-squared:   0.3
## F-statistic:  975 on 12 and 27288 DF,  p-value: <2e-16
```

```
reg_final_2 = update(reg_final,.~.-HANDPER:DOCTOR,data=hc2)
summary(reg_final_2) # insignificant
```

```
##
## Call:
```

```
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + HANDPER:NEWHSAT + DOCTOR:HOSPITAL +
##     DOCTOR:NEWHSAT + HEALTHY:NEWHSAT + HOSPITAL:NEWHSAT, data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14.86  -1.89  -0.17   0.39 107.42
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.767122   0.273790    6.45  1.1e-10 ***
## FEMALE            0.471763   0.058725    8.03  9.9e-16 ***
## HANDPER           0.005837   0.003403    1.72    0.086 .
## HOSPVIS           0.200245   0.037853    5.29  1.2e-07 ***
## DOCTOR            8.736425   0.240516   36.32  < 2e-16 ***
## HEALTHY          -3.602214   0.355950  -10.12  < 2e-16 ***
## HOSPITAL          2.047016   0.388124    5.27  1.3e-07 ***
## NEWHSAT          -0.434901   0.045754   -9.51  < 2e-16 ***
## HANDPER:NEWHSAT   0.002401   0.000616    3.90  9.8e-05 ***
## DOCTOR:HOSPITAL   1.852734   0.274581    6.75  1.5e-11 ***
## DOCTOR:NEWHSAT   -0.667699   0.031862  -20.96  < 2e-16 ***
## HEALTHY:NEWHSAT   0.626068   0.051809   12.08  < 2e-16 ***
## HOSPITAL:NEWHSAT -0.333981   0.041114   -8.12  4.7e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.76 on 27288 degrees of freedom
## Multiple R-squared:   0.3,   Adjusted R-squared:   0.3
## F-statistic:  974 on 12 and 27288 DF,  p-value: <2e-16
```

```
reg_final_3 = update(reg_final,.~.-DOCTOR:HOSPITAL,data=hc2)
summary(reg_final_3) # 0.2993
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + HANDPER:DOCTOR + HANDPER:NEWHSAT + DOCTOR:NEWHSAT +
##     HEALTHY:NEWHSAT + HOSPITAL:NEWHSAT, data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14.96  -1.90  -0.17   0.35 107.38
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.949863   0.278807    6.99  2.7e-12 ***
## FEMALE            0.473710   0.058737    8.06  7.6e-16 ***
## HANDPER          -0.021033   0.005600   -3.76  0.00017 ***
## HOSPVIS           0.198225   0.037859    5.24  1.7e-07 ***
## DOCTOR            8.527034   0.247191   34.50  < 2e-16 ***
## HEALTHY          -3.693126   0.356585  -10.36  < 2e-16 ***
## HOSPITAL          3.972018   0.261814   15.17  < 2e-16 ***
## NEWHSAT          -0.459688   0.046198   -9.95  < 2e-16 ***
## HANDPER:DOCTOR    0.027444   0.004558    6.02  1.8e-09 ***
```

```
## HANDPER:NEWHSAT   0.003127   0.000627    4.99  6.2e-07 ***
## DOCTOR:NEWHSAT   -0.640322   0.032391  -19.77  < 2e-16 ***
## HEALTHY:NEWHSAT   0.639613   0.051913   12.32  < 2e-16 ***
## HOSPITAL:NEWHSAT -0.408545   0.039587  -10.32  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.76 on 27288 degrees of freedom
## Multiple R-squared:   0.3,   Adjusted R-squared:  0.299
## F-statistic:  973 on 12 and 27288 DF,  p-value: <2e-16
```

```
reg_final_4= update(reg_final,.~.-DOCTOR:NEWHSAT,data=hc2)
summary(reg_final_4) # 0.2906
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + HANDPER:DOCTOR + HANDPER:NEWHSAT + DOCTOR:HOSPITAL +
##     HEALTHY:NEWHSAT + HOSPITAL:NEWHSAT, data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -15.19  -2.11  -0.14   0.51 107.36
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       6.122207   0.190579   32.12  < 2e-16 ***
## FEMALE            0.463899   0.059099    7.85  4.3e-15 ***
## HANDPER          -0.031033   0.005608   -5.53  3.2e-08 ***
## HOSPVIS           0.202884   0.038094    5.33  1.0e-07 ***
## DOCTOR            3.716138   0.068276   54.43  < 2e-16 ***
## HEALTHY          -5.787385   0.342718  -16.89  < 2e-16 ***
## HOSPITAL          2.348810   0.390460    6.02  1.8e-09 ***
## NEWHSAT          -1.050937   0.035810  -29.35  < 2e-16 ***
## HANDPER:DOCTOR    0.042668   0.004512    9.46  < 2e-16 ***
## HANDPER:NEWHSAT   0.002783   0.000631    4.41  1.0e-05 ***
## DOCTOR:HOSPITAL   1.972324   0.276339    7.14  9.8e-13 ***
## HEALTHY:NEWHSAT   0.957064   0.049708   19.25  < 2e-16 ***
## HOSPITAL:NEWHSAT -0.394954   0.041264   -9.57  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.79 on 27288 degrees of freedom
## Multiple R-squared:  0.291,  Adjusted R-squared:  0.291
## F-statistic:  933 on 12 and 27288 DF,  p-value: <2e-16
```

```
reg_final_5= update(reg_final,.~.-HEALTHY:NEWHSAT,data=hc2)
summary(reg_final_5) # insignificant
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + HANDPER:DOCTOR + HANDPER:NEWHSAT + DOCTOR:HOSPITAL +
```

```
##     DOCTOR:NEWHSAT + HOSPITAL:NEWHSAT, data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -15.20  -1.93  -0.25   0.23 107.56
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.044703   0.220626   -0.20    0.839
## FEMALE            0.485361   0.058852    8.25  < 2e-16 ***
## HANDPER          -0.005046   0.005483   -0.92    0.357
## HOSPVIS           0.202642   0.037936    5.34  9.3e-08 ***
## DOCTOR            9.262149   0.238339   38.86  < 2e-16 ***
## HEALTHY           0.497980   0.108467    4.59  4.4e-06 ***
## HOSPITAL          2.553462   0.387296    6.59  4.4e-11 ***
## NEWHSAT          -0.068705   0.032867   -2.09    0.037 *
## HANDPER:DOCTOR    0.022915   0.004561    5.02  5.1e-07 ***
## HANDPER:NEWHSAT   0.000997   0.000606    1.65    0.100 .
## DOCTOR:HOSPITAL   1.747019   0.275306    6.35  2.2e-10 ***
## DOCTOR:NEWHSAT   -0.757326   0.030899  -24.51  < 2e-16 ***
## HOSPITAL:NEWHSAT -0.396999   0.040907   -9.70  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.77 on 27288 degrees of freedom
## Multiple R-squared:  0.297,  Adjusted R-squared:  0.296
## F-statistic:  960 on 12 and 27288 DF,  p-value: <2e-16
```

```
reg_final_6= update(reg_final,.~.-HOSPITAL:NEWHSAT,data=hc2)
summary(reg_final_6) # insignificant
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + HANDPER:DOCTOR + HANDPER:NEWHSAT + DOCTOR:HOSPITAL +
##     DOCTOR:NEWHSAT + HEALTHY:NEWHSAT, data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13.64  -1.96  -0.21   0.30 108.49
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.428687   0.276904    8.77  < 2e-16 ***
## FEMALE            0.470183   0.058760    8.00  1.3e-15 ***
## HANDPER          -0.018351   0.005601   -3.28   0.0011 **
## HOSPVIS           0.219853   0.037793    5.82  6.0e-09 ***
## DOCTOR            8.479992   0.247994   34.19  < 2e-16 ***
## HEALTHY          -3.989571   0.355262  -11.23  < 2e-16 ***
## HOSPITAL         -0.369090   0.244437   -1.51   0.1311
## NEWHSAT          -0.533008   0.045691  -11.67  < 2e-16 ***
## HANDPER:DOCTOR    0.026191   0.004562    5.74  9.5e-09 ***
## HANDPER:NEWHSAT   0.002796   0.000626    4.46  8.1e-06 ***
## DOCTOR:HOSPITAL   2.406781   0.264639    9.09  < 2e-16 ***
```

```
## DOCTOR:NEWHSAT  -0.653019   0.032334  -20.20  < 2e-16 ***
## HEALTHY:NEWHSAT  0.695771   0.051565   13.49  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.77 on 27288 degrees of freedom
## Multiple R-squared:  0.299,  Adjusted R-squared:  0.299
## F-statistic:  970 on 12 and 27288 DF,  p-value: <2e-16
```

```
AIC(reg_final_3,reg_final_4)
```

```
##             df    AIC
## reg_final_3 14 162723
## reg_final_4 14 163061
```

```
BIC(reg_final_3,reg_final_4)
```

```
##             df    BIC
## reg_final_3 14 162838
## reg_final_4 14 163176
```

- We successfully to remove 1 variable, then we try to remove another.

```
reg_final_3_1 = update(reg_final_3,.~.-HANDPER:DOCTOR,data=hc2)
summary(reg_final_3_1) #insignificant
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + HANDPER:NEWHSAT + DOCTOR:NEWHSAT + HEALTHY:NEWHSAT +
##     HOSPITAL:NEWHSAT, data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14.92  -1.93  -0.20   0.39 107.41
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.601295   0.272907    5.87  4.5e-09 ***
## FEMALE            0.469882   0.058772    8.00  1.3e-15 ***
## HANDPER           0.005743   0.003406    1.69    0.092 .
## HOSPVIS           0.199055   0.037883    5.25  1.5e-07 ***
## DOCTOR            8.899022   0.239501   37.16  < 2e-16 ***
## HEALTHY          -3.567368   0.356202  -10.02  < 2e-16 ***
## HOSPITAL          3.980613   0.261979   15.19  < 2e-16 ***
## NEWHSAT          -0.419018   0.045731   -9.16  < 2e-16 ***
## HANDPER:NEWHSAT   0.002442   0.000617    3.96  7.6e-05 ***
## DOCTOR:NEWHSAT   -0.676048   0.031864  -21.22  < 2e-16 ***
## HEALTHY:NEWHSAT   0.619923   0.051843   11.96  < 2e-16 ***
## HOSPITAL:NEWHSAT -0.409038   0.039613  -10.33  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.77 on 27289 degrees of freedom
## Multiple R-squared:  0.299,  Adjusted R-squared:  0.298
## F-statistic: 1.06e+03 on 11 and 27289 DF,  p-value: <2e-16
```

```
reg_final_3_2 = update(reg_final_3,.~.-HANDPER:NEWHSAT,data=hc2)
summary(reg_final_3_2) #insignificant
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + HANDPER:DOCTOR + DOCTOR:NEWHSAT + HEALTHY:NEWHSAT +
##     HOSPITAL:NEWHSAT, data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -16.00  -1.95  -0.18   0.37 107.14
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.58725    0.26928    5.89  3.8e-09 ***
## FEMALE             0.46600    0.05874    7.93  2.2e-15 ***
## HANDPER           -0.00227    0.00415   -0.55     0.58
## HOSPVIS            0.19934    0.03788    5.26  1.4e-07 ***
## DOCTOR             8.53057    0.24730   34.50  < 2e-16 ***
## HEALTHY           -3.31592    0.34862   -9.51  < 2e-16 ***
## HOSPITAL           3.91599    0.26169   14.96  < 2e-16 ***
## NEWHSAT           -0.39076    0.04410   -8.86  < 2e-16 ***
## HANDPER:DOCTOR     0.02332    0.00448    5.20  2.0e-07 ***
## DOCTOR:NEWHSAT    -0.63671    0.03240  -19.65  < 2e-16 ***
## HEALTHY:NEWHSAT    0.57086    0.05007   11.40  < 2e-16 ***
## HOSPITAL:NEWHSAT -0.39871    0.03956  -10.08  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.77 on 27289 degrees of freedom
## Multiple R-squared:  0.299,  Adjusted R-squared:  0.299
## F-statistic: 1.06e+03 on 11 and 27289 DF,  p-value: <2e-16
```

```
reg_final_3_3 = update(reg_final_3,.~.-DOCTOR:NEWHSAT,data=hc2)
summary(reg_final_3_3)
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + HANDPER:DOCTOR + HANDPER:NEWHSAT + HEALTHY:NEWHSAT +
##     HOSPITAL:NEWHSAT, data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -15.27  -2.12  -0.15   0.56 107.35
##
```

```
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        6.006831   0.190066   31.60  < 2e-16 ***
## FEMALE             0.461911   0.059152    7.81  6.0e-15 ***
## HANDPER           -0.032326   0.005610   -5.76  8.4e-09 ***
## HOSPVIS            0.201624   0.038128    5.29  1.2e-07 ***
## DOCTOR             3.819558   0.066782   57.19  < 2e-16 ***
## HEALTHY           -5.779621   0.343030  -16.85  < 2e-16 ***
## HOSPITAL           4.412011   0.262722   16.79  < 2e-16 ***
## NEWHSAT           -1.042441   0.035823  -29.10  < 2e-16 ***
## HANDPER:DOCTOR     0.043948   0.004513    9.74  < 2e-16 ***
## HANDPER:NEWHSAT    0.002849   0.000631    4.51  6.4e-06 ***
## HEALTHY:NEWHSAT    0.954983   0.049752   19.19  < 2e-16 ***
## HOSPITAL:NEWHSAT -0.475619   0.039722  -11.97  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.8 on 27289 degrees of freedom
## Multiple R-squared:  0.29,   Adjusted R-squared:  0.289
## F-statistic: 1.01e+03 on 11 and 27289 DF,  p-value: <2e-16
```

```
reg_final_3_4 = update(reg_final_3,.~.-HEALTHY:NEWHSAT,data=hc2)
summary(reg_final_3_4) #insignificant
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + HANDPER:DOCTOR + HANDPER:NEWHSAT + DOCTOR:NEWHSAT +
##     HOSPITAL:NEWHSAT, data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -15.25  -1.97  -0.22   0.26 107.55
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -0.171904   0.219872   -0.78    0.434
## FEMALE             0.483643   0.058894    8.21  2.3e-16 ***
## HANDPER           -0.006193   0.005484   -1.13    0.259
## HOSPVIS            0.201465   0.037963    5.31  1.1e-07 ***
## DOCTOR             9.395484   0.237582   39.55  < 2e-16 ***
## HEALTHY            0.493022   0.108543    4.54  5.6e-06 ***
## HOSPITAL           4.373227   0.260498   16.79  < 2e-16 ***
## NEWHSAT           -0.058358   0.032850   -1.78    0.076 .
## HANDPER:DOCTOR     0.023907   0.004561    5.24  1.6e-07 ***
## HANDPER:NEWHSAT    0.001076   0.000606    1.77    0.076 .
## DOCTOR:NEWHSAT    -0.762966   0.030909  -24.68  < 2e-16 ***
## HOSPITAL:NEWHSAT -0.467287   0.039407  -11.86  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.78 on 27289 degrees of freedom
## Multiple R-squared:  0.296,  Adjusted R-squared:  0.295
## F-statistic: 1.04e+03 on 11 and 27289 DF,  p-value: <2e-16
```

```
reg_final_3_5 = update(reg_final_3,.~.-HOSPITAL:NEWHSAT,data=hc2)
summary(reg_final_3_5)
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + HANDPER:DOCTOR + HANDPER:NEWHSAT + DOCTOR:NEWHSAT +
##     HEALTHY:NEWHSAT, data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13.34  -2.00  -0.13   0.38 108.82
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.317951   0.277050    8.37  < 2e-16 ***
## FEMALE            0.466112   0.058846    7.92  2.4e-15 ***
## HANDPER          -0.019371   0.005608   -3.45  0.00055 ***
## HOSPVIS           0.224547   0.037846    5.93  3.0e-09 ***
## DOCTOR            8.719040   0.246966   35.30  < 2e-16 ***
## HEALTHY          -4.031003   0.355764  -11.33  < 2e-16 ***
## HOSPITAL          1.568946   0.119923   13.08  < 2e-16 ***
## NEWHSAT          -0.531468   0.045759  -11.61  < 2e-16 ***
## HANDPER:DOCTOR    0.027541   0.004567    6.03  1.7e-09 ***
## HANDPER:NEWHSAT   0.002804   0.000627    4.47  7.8e-06 ***
## DOCTOR:NEWHSAT   -0.668973   0.032334  -20.69  < 2e-16 ***
## HEALTHY:NEWHSAT   0.704135   0.051634   13.64  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.77 on 27289 degrees of freedom
## Multiple R-squared:  0.297,  Adjusted R-squared:  0.297
## F-statistic: 1.05e+03 on 11 and 27289 DF,  p-value: <2e-16
```

```
AIC(reg_final_3_3,reg_final_3_5)
```

```
##                df    AIC
## reg_final_3_3 13 163110
## reg_final_3_5 13 162828
```

```
BIC(reg_final_3_3,reg_final_3_5)
```

```
##                df    BIC
## reg_final_3_3 13 163216
## reg_final_3_5 13 162935
```

- We successfully to remove 1 variable again, then we try to remove another variable.

```
reg_final_3_5_1 = update(reg_final_3_5,.~.-HANDPER:DOCTOR,data=hc2)
summary(reg_final_3_5_1) # insignificant
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + HANDPER:NEWHSAT + DOCTOR:NEWHSAT + HEALTHY:NEWHSAT,
##     data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13.30  -2.02  -0.22   0.44 108.85
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.968592   0.271102    7.26  3.9e-13 ***
## FEMALE           0.462261   0.058881    7.85  4.3e-15 ***
## HANDPER          0.007502   0.003408    2.20  0.02773 *
## HOSPVIS          0.225411   0.037870    5.95  2.7e-09 ***
## DOCTOR           9.092580   0.239228   38.01  < 2e-16 ***
## HEALTHY         -3.905207   0.355382  -10.99  < 2e-16 ***
## HOSPITAL         1.574662   0.119997   13.12  < 2e-16 ***
## NEWHSAT         -0.490741   0.045287  -10.84  < 2e-16 ***
## HANDPER:NEWHSAT  0.002116   0.000617    3.43  0.00061 ***
## DOCTOR:NEWHSAT  -0.704860   0.031803  -22.16  < 2e-16 ***
## HEALTHY:NEWHSAT  0.684453   0.051564   13.27  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.78 on 27290 degrees of freedom
## Multiple R-squared:  0.296,  Adjusted R-squared:  0.296
## F-statistic: 1.15e+03 on 10 and 27290 DF,  p-value: <2e-16
```

```
reg_final_3_5_2 = update(reg_final_3_5,.~.-HANDPER:NEWHSAT,data=hc2)
summary(reg_final_3_5_2) # insignificant
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + HANDPER:DOCTOR + DOCTOR:NEWHSAT + HEALTHY:NEWHSAT,
##     data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14.31  -1.98  -0.14   0.36 108.57
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.98394    0.26688    7.43  1.1e-13 ***
## FEMALE           0.45934    0.05885    7.81  6.1e-15 ***
## HANDPER         -0.00254    0.00416   -0.61     0.54
## HOSPVIS          0.22498    0.03786    5.94  2.8e-09 ***
## DOCTOR           8.71806    0.24705   35.29  < 2e-16 ***
## HEALTHY         -3.68452    0.34734  -10.61  < 2e-16 ***
## HOSPITAL         1.57058    0.11996   13.09  < 2e-16 ***
## NEWHSAT         -0.46794    0.04351  -10.75  < 2e-16 ***
## HANDPER:DOCTOR   0.02383    0.00449    5.30  1.1e-07 ***
```

```
## DOCTOR:NEWHSAT  -0.66510    0.03233  -20.57  < 2e-16 ***
## HEALTHY:NEWHSAT  0.64092    0.04968   12.90  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.77 on 27290 degrees of freedom
## Multiple R-squared:  0.296,  Adjusted R-squared:  0.296
## F-statistic: 1.15e+03 on 10 and 27290 DF,  p-value: <2e-16
```

```
reg_final_3_5_3 = update(reg_final_3_5,.~.-DOCTOR:NEWHSAT,data=hc2)
summary(reg_final_3_5_3)
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + HANDPER:DOCTOR + HANDPER:NEWHSAT + HEALTHY:NEWHSAT,
##     data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13.38  -2.12  -0.16   0.61 109.04
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.651415   0.182758   36.39  < 2e-16 ***
## FEMALE           0.452380   0.059301    7.63  2.5e-14 ***
## HANDPER         -0.030969   0.005624   -5.51  3.7e-08 ***
## HOSPVIS          0.232672   0.038139    6.10  1.1e-09 ***
## DOCTOR           3.797711   0.066931   56.74  < 2e-16 ***
## HEALTHY         -6.285371   0.341307  -18.42  < 2e-16 ***
## HOSPITAL         1.616794   0.120835   13.38  < 2e-16 ***
## NEWHSAT         -1.157204   0.034607  -33.44  < 2e-16 ***
## HANDPER:DOCTOR   0.044929   0.004524    9.93  < 2e-16 ***
## HANDPER:NEWHSAT  0.002457   0.000632    3.89    1e-04 ***
## HEALTHY:NEWHSAT  1.047203   0.049280   21.25  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.81 on 27290 degrees of freedom
## Multiple R-squared:  0.286,  Adjusted R-squared:  0.286
## F-statistic: 1.09e+03 on 10 and 27290 DF,  p-value: <2e-16
```

```
reg_final_3_5_4 = update(reg_final_3_5,.~.-HEALTHY:NEWHSAT,data=hc2)
summary(reg_final_3_5_4) # insignificant
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + HANDPER:DOCTOR + HANDPER:NEWHSAT + DOCTOR:NEWHSAT,
##     data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
```

```
## -13.41  -1.99  -0.17   0.35 109.24
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.006891   0.219914    0.03   0.9750
## FEMALE           0.475987   0.059041    8.06  7.8e-16 ***
## HANDPER         -0.002526   0.005489   -0.46   0.6454
## HOSPVIS          0.232393   0.037970    6.12  9.5e-10 ***
## DOCTOR           9.720007   0.236603   41.08  < 2e-16 ***
## HEALTHY          0.590989   0.108504    5.45  5.2e-08 ***
## HOSPITAL         1.631153   0.120241   13.57  < 2e-16 ***
## NEWHSAT         -0.094680   0.032791   -2.89   0.0039 **
## HANDPER:DOCTOR   0.023605   0.004573    5.16  2.5e-07 ***
## HANDPER:NEWHSAT  0.000462   0.000605    0.76   0.4459
## DOCTOR:NEWHSAT  -0.810578   0.030725  -26.38  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.79 on 27290 degrees of freedom
## Multiple R-squared:  0.292,  Adjusted R-squared:  0.292
## F-statistic: 1.13e+03 on 10 and 27290 DF,  p-value: <2e-16
```

```
reg_final_3_5_5 = update(reg_final_3_5,.~.-HOSPITAL:NEWHSAT,data=hc2)
summary(reg_final_3_5_5)
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + HANDPER:DOCTOR + HANDPER:NEWHSAT + DOCTOR:NEWHSAT +
##     HEALTHY:NEWHSAT, data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13.34  -2.00  -0.13   0.38 108.82
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.317951   0.277050    8.37  < 2e-16 ***
## FEMALE           0.466112   0.058846    7.92  2.4e-15 ***
## HANDPER         -0.019371   0.005608   -3.45  0.00055 ***
## HOSPVIS          0.224547   0.037846    5.93  3.0e-09 ***
## DOCTOR           8.719040   0.246966   35.30  < 2e-16 ***
## HEALTHY         -4.031003   0.355764  -11.33  < 2e-16 ***
## HOSPITAL         1.568946   0.119923   13.08  < 2e-16 ***
## NEWHSAT         -0.531468   0.045759  -11.61  < 2e-16 ***
## HANDPER:DOCTOR   0.027541   0.004567    6.03  1.7e-09 ***
## HANDPER:NEWHSAT  0.002804   0.000627    4.47  7.8e-06 ***
## DOCTOR:NEWHSAT  -0.668973   0.032334  -20.69  < 2e-16 ***
## HEALTHY:NEWHSAT  0.704135   0.051634   13.64  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.77 on 27289 degrees of freedom
## Multiple R-squared:  0.297,  Adjusted R-squared:  0.297
```

```
## F-statistic: 1.05e+03 on 11 and 27289 DF,  p-value: <2e-16
```

```
AIC(reg_final_3_5_3,reg_final_3_5_5)
```

```
##                 df    AIC
## reg_final_3_5_3 12 163251
## reg_final_3_5_5 13 162828
```

```
BIC(reg_final_3_5_3,reg_final_3_5_5)
```

```
##                 df    BIC
## reg_final_3_5_3 12 163349
## reg_final_3_5_5 13 162935
```

- We successfully to remove 1 variable again, then we try to remove the last one to keep 10 variables in.

```
reg_final_3_5_5_1 = update(reg_final_3_5_5,.~.-HANDPER:DOCTOR,data=hc2)
summary(reg_final_3_5_5_1) # insignificant
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + HANDPER:NEWHSAT + DOCTOR:NEWHSAT + HEALTHY:NEWHSAT,
##     data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13.30  -2.02  -0.22   0.44 108.85
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.968592   0.271102    7.26  3.9e-13 ***
## FEMALE           0.462261   0.058881    7.85  4.3e-15 ***
## HANDPER          0.007502   0.003408    2.20  0.02773 *
## HOSPVIS          0.225411   0.037870    5.95  2.7e-09 ***
## DOCTOR           9.092580   0.239228   38.01  < 2e-16 ***
## HEALTHY         -3.905207   0.355382  -10.99  < 2e-16 ***
## HOSPITAL         1.574662   0.119997   13.12  < 2e-16 ***
## NEWHSAT         -0.490741   0.045287  -10.84  < 2e-16 ***
## HANDPER:NEWHSAT  0.002116   0.000617    3.43  0.00061 ***
## DOCTOR:NEWHSAT  -0.704860   0.031803  -22.16  < 2e-16 ***
## HEALTHY:NEWHSAT  0.684453   0.051564   13.27  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.78 on 27290 degrees of freedom
## Multiple R-squared:  0.296,  Adjusted R-squared:  0.296
## F-statistic: 1.15e+03 on 10 and 27290 DF,  p-value: <2e-16
```

```r
reg_final_3_5_5_2 = update(reg_final_3_5_5,.~.-HANDPER:NEWHSAT,data=hc2)
summary(reg_final_3_5_5_2) # insignificant
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + HANDPER:DOCTOR + DOCTOR:NEWHSAT + HEALTHY:NEWHSAT,
##     data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14.31  -1.98  -0.14   0.36 108.57
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.98394    0.26688    7.43  1.1e-13 ***
## FEMALE            0.45934    0.05885    7.81  6.1e-15 ***
## HANDPER          -0.00254    0.00416   -0.61     0.54
## HOSPVIS           0.22498    0.03786    5.94  2.8e-09 ***
## DOCTOR            8.71806    0.24705   35.29  < 2e-16 ***
## HEALTHY          -3.68452    0.34734  -10.61  < 2e-16 ***
## HOSPITAL          1.57058    0.11996   13.09  < 2e-16 ***
## NEWHSAT          -0.46794    0.04351  -10.75  < 2e-16 ***
## HANDPER:DOCTOR    0.02383    0.00449    5.30  1.1e-07 ***
## DOCTOR:NEWHSAT   -0.66510    0.03233  -20.57  < 2e-16 ***
## HEALTHY:NEWHSAT   0.64092    0.04968   12.90  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.77 on 27290 degrees of freedom
## Multiple R-squared:  0.296,  Adjusted R-squared:  0.296
## F-statistic: 1.15e+03 on 10 and 27290 DF,  p-value: <2e-16
```

```r
reg_final_3_5_5_3 = update(reg_final_3_5_5,.~.-DOCTOR:NEWHSAT,data=hc2)
summary(reg_final_3_5_5_3)
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + HANDPER:DOCTOR + HANDPER:NEWHSAT + HEALTHY:NEWHSAT,
##     data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13.38  -2.12  -0.16   0.61 109.04
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.651415   0.182758   36.39  < 2e-16 ***
## FEMALE           0.452380   0.059301    7.63  2.5e-14 ***
## HANDPER         -0.030969   0.005624   -5.51  3.7e-08 ***
## HOSPVIS          0.232672   0.038139    6.10  1.1e-09 ***
```

```
## DOCTOR           3.797711   0.066931   56.74  < 2e-16 ***
## HEALTHY         -6.285371   0.341307  -18.42  < 2e-16 ***
## HOSPITAL         1.616794   0.120835   13.38  < 2e-16 ***
## NEWHSAT         -1.157204   0.034607  -33.44  < 2e-16 ***
## HANDPER:DOCTOR   0.044929   0.004524    9.93  < 2e-16 ***
## HANDPER:NEWHSAT  0.002457   0.000632    3.89    1e-04 ***
## HEALTHY:NEWHSAT  1.047203   0.049280   21.25  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.81 on 27290 degrees of freedom
## Multiple R-squared:  0.286,  Adjusted R-squared:  0.286
## F-statistic: 1.09e+03 on 10 and 27290 DF,  p-value: <2e-16
```

```
reg_final_3_5_5_4 = update(reg_final_3_5_5,.~.-HEALTHY:NEWHSAT,data=hc2)
summary(reg_final_3_5_5_4) # insignificant
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + HANDPER:DOCTOR + HANDPER:NEWHSAT + DOCTOR:NEWHSAT,
##     data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13.41  -1.99  -0.17   0.35 109.24
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.006891   0.219914    0.03   0.9750
## FEMALE           0.475987   0.059041    8.06  7.8e-16 ***
## HANDPER         -0.002526   0.005489   -0.46   0.6454
## HOSPVIS          0.232393   0.037970    6.12  9.5e-10 ***
## DOCTOR           9.720007   0.236603   41.08  < 2e-16 ***
## HEALTHY          0.590989   0.108504    5.45  5.2e-08 ***
## HOSPITAL         1.631153   0.120241   13.57  < 2e-16 ***
## NEWHSAT         -0.094680   0.032791   -2.89   0.0039 **
## HANDPER:DOCTOR   0.023605   0.004573    5.16  2.5e-07 ***
## HANDPER:NEWHSAT  0.000462   0.000605    0.76   0.4459
## DOCTOR:NEWHSAT  -0.810578   0.030725  -26.38  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.79 on 27290 degrees of freedom
## Multiple R-squared:  0.292,  Adjusted R-squared:  0.292
## F-statistic: 1.13e+03 on 10 and 27290 DF,  p-value: <2e-16
```

```
reg_final = reg_final_3_5_5_3
summary(reg_final)
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
```

```
##     HOSPITAL + NEWHSAT + HANDPER:DOCTOR + HANDPER:NEWHSAT + HEALTHY:NEWHSAT,
##     data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13.38  -2.12  -0.16   0.61 109.04
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       6.651415   0.182758   36.39  < 2e-16 ***
## FEMALE            0.452380   0.059301    7.63  2.5e-14 ***
## HANDPER          -0.030969   0.005624   -5.51  3.7e-08 ***
## HOSPVIS           0.232672   0.038139    6.10  1.1e-09 ***
## DOCTOR            3.797711   0.066931   56.74  < 2e-16 ***
## HEALTHY          -6.285371   0.341307  -18.42  < 2e-16 ***
## HOSPITAL          1.616794   0.120835   13.38  < 2e-16 ***
## NEWHSAT          -1.157204   0.034607  -33.44  < 2e-16 ***
## HANDPER:DOCTOR    0.044929   0.004524    9.93  < 2e-16 ***
## HANDPER:NEWHSAT   0.002457   0.000632    3.89    1e-04 ***
## HEALTHY:NEWHSAT   1.047203   0.049280   21.25  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.81 on 27290 degrees of freedom
## Multiple R-squared:  0.286,  Adjusted R-squared:  0.286
## F-statistic: 1.09e+03 on 10 and 27290 DF,  p-value: <2e-16
```

- Since we at most can have 10 predictors, this model is our final model.
- Our final model have a adjusted R-square 28.56%, I just interpret some of these.
- The number of visiting doctors will increase if a person is a female or if she has seen a doctor or gone to the hospital before.
- The number of visting doctors will decrease if he or she is healthy.

(b)

```
lm_did_female = update(reg_final,.~.-FEMALE+FEMALE*TIME,data=hc2)
summary(lm_did_female)
```

```
##
## Call:
## lm(formula = DOCVIS ~ HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + FEMALE + TIME + HANDPER:DOCTOR + HANDPER:NEWHSAT +
##     HEALTHY:NEWHSAT + FEMALE:TIME, data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13.63  -2.12  -0.25   0.63 108.88
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       6.760002   0.185886   36.37  < 2e-16 ***
```

```
## HANDPER         -0.030868   0.005619   -5.49  4.0e-08 ***
## HOSPVIS          0.231844   0.038107    6.08  1.2e-09 ***
## DOCTOR           3.820569   0.066951   57.07  < 2e-16 ***
## HEALTHY         -6.107692   0.341925  -17.86  < 2e-16 ***
## HOSPITAL         1.631852   0.120745   13.51  < 2e-16 ***
## NEWHSAT         -1.151321   0.034585  -33.29  < 2e-16 ***
## FEMALE           0.527401   0.079142    6.66  2.7e-11 ***
## TIME            -0.324266   0.081396   -3.98  6.8e-05 ***
## HANDPER:DOCTOR   0.045154   0.004520    9.99  < 2e-16 ***
## HANDPER:NEWHSAT  0.002282   0.000632    3.61   0.0003 ***
## HEALTHY:NEWHSAT  1.023365   0.049352   20.74  < 2e-16 ***
## FEMALE:TIME     -0.182037   0.117179   -1.55   0.1203
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.81 on 27288 degrees of freedom
## Multiple R-squared:  0.287,  Adjusted R-squared:  0.287
## F-statistic:  916 on 12 and 27288 DF,  p-value: <2e-16
```

```
lm_did_unemployed = update(reg_final,.~.+UNEMPLOY*TIME,data=hc2)
summary(lm_did_unemployed)
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + UNEMPLOY + TIME + HANDPER:DOCTOR + HANDPER:NEWHSAT +
##     HEALTHY:NEWHSAT + UNEMPLOY:TIME, data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13.61  -2.09  -0.25   0.63 108.70
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.735188   0.185768   36.26  < 2e-16 ***
## FEMALE           0.375052   0.064325    5.83  5.6e-09 ***
## HANDPER         -0.031440   0.005624   -5.59  2.3e-08 ***
## HOSPVIS          0.230727   0.038102    6.06  1.4e-09 ***
## DOCTOR           3.819048   0.066947   57.05  < 2e-16 ***
## HEALTHY         -6.087547   0.341977  -17.80  < 2e-16 ***
## HOSPITAL         1.633089   0.120764   13.52  < 2e-16 ***
## NEWHSAT         -1.148038   0.034605  -33.18  < 2e-16 ***
## UNEMPLOY         0.247839   0.088331    2.81  0.00502 **
## TIME            -0.357157   0.071285   -5.01  5.5e-07 ***
## HANDPER:DOCTOR   0.044851   0.004521    9.92  < 2e-16 ***
## HANDPER:NEWHSAT  0.002275   0.000632    3.60  0.00032 ***
## HEALTHY:NEWHSAT  1.020299   0.049360   20.67  < 2e-16 ***
## UNEMPLOY:TIME   -0.141803   0.126480   -1.12  0.26223
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.81 on 27287 degrees of freedom
## Multiple R-squared:  0.287,  Adjusted R-squared:  0.287
## F-statistic:  846 on 13 and 27287 DF,  p-value: <2e-16
```

- From the results above, we can see at a 1% significant level, neither does the policy work for women, nor does it work for the unemployed.

(c)

- We do the ANOVA, and our null hypothesis is $\beta_{FEMALE} \leq 0$

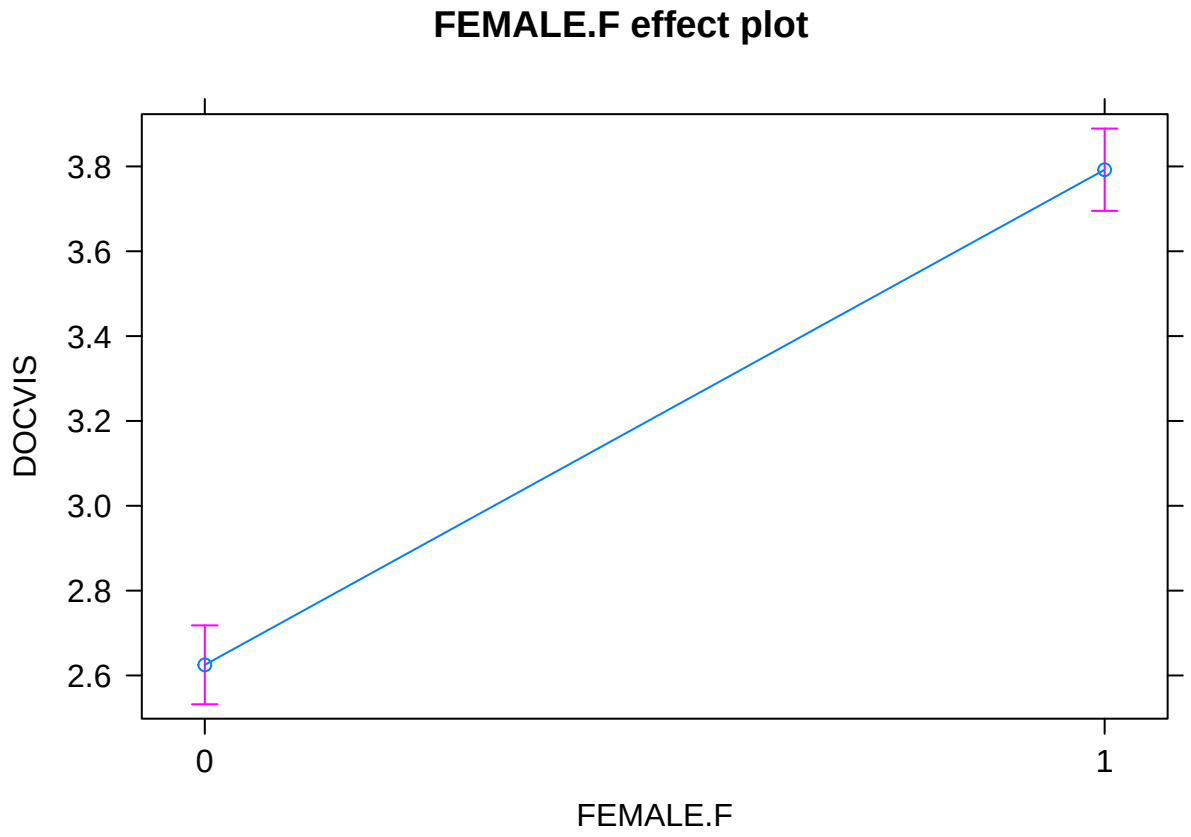
```
hc2$FEMALE.F = as.factor(hc2$FEMALE)
anv = lm(DOCVIS~FEMALE.F,data=hc2)
summary(anv)
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE.F, data = hc2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -3.79  -2.63  -1.63   0.37 118.37
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.6251     0.0475    55.3   <2e-16 ***
## FEMALE.F1     1.1672     0.0686    17.0   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.66 on 27299 degrees of freedom
## Multiple R-squared:  0.0105, Adjusted R-squared:  0.0105
## F-statistic:  290 on 1 and 27299 DF,  p-value: <2e-16
```

```
emmeans(anv,trt.vs.ctrl~FEMALE.F)
```

```
## $emmeans
##  FEMALE.F emmean     SE    df lower.CL upper.CL
##  0          2.63 0.0475 27299     2.53     2.72
##  1          3.79 0.0495 27299     3.70     3.89
##
## Confidence level used: 0.95
##
## $contrasts
##  contrast estimate     SE    df t.ratio p.value
##  1 - 0        1.17 0.0686 27299 17.020  <.0001
```

```
plot(effect("FEMALE.F",anv))
```



- From the linear regression form of anova, the estiamted marginal means and the effect plot, we all can know that the number of doctor visits a patient has over a 3 month period is greater for women than for men.

(d)

- I want to test that if a woman is healthy, the number of visiting doctors will decrease 3. That's to say our null hypothesis is $\lambda = 1 * \beta_{HEALTHY} + 1 * \beta_{FEMALE} = -6$

```
glht = glht(reg_final,linfct =c("1*FEMALE+1*HEALTHY=-6"))
summary(glht)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + HANDPER:DOCTOR + HANDPER:NEWHSAT + HEALTHY:NEWHSAT,
##     data = hc2)
##
## Linear Hypotheses:
##                                Estimate Std. Error t value Pr(>|t|)
## 1 * FEMALE + 1 * HEALTHY == -6   -5.833      0.348    0.48     0.63
## (Adjusted p values reported -- single-step method)
```

```r
confint(glht)
```

```
##
##   Simultaneous Confidence Intervals
##
## Fit: lm(formula = DOCVIS ~ FEMALE + HANDPER + HOSPVIS + DOCTOR + HEALTHY +
##     HOSPITAL + NEWHSAT + HANDPER:DOCTOR + HANDPER:NEWHSAT + HEALTHY:NEWHSAT,
##     data = hc2)
##
## Quantile = 1.96
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##                                Estimate lwr    upr
## 1 * FEMALE + 1 * HEALTHY == -6 -5.833   -6.514 -5.152
```

- I can not reject the null hypothesis, and it means that we can not deny that if a woman is healthy, the number of visiting doctors will decrease 6. And the interval estimate shows that -6 is in the confidence interval.