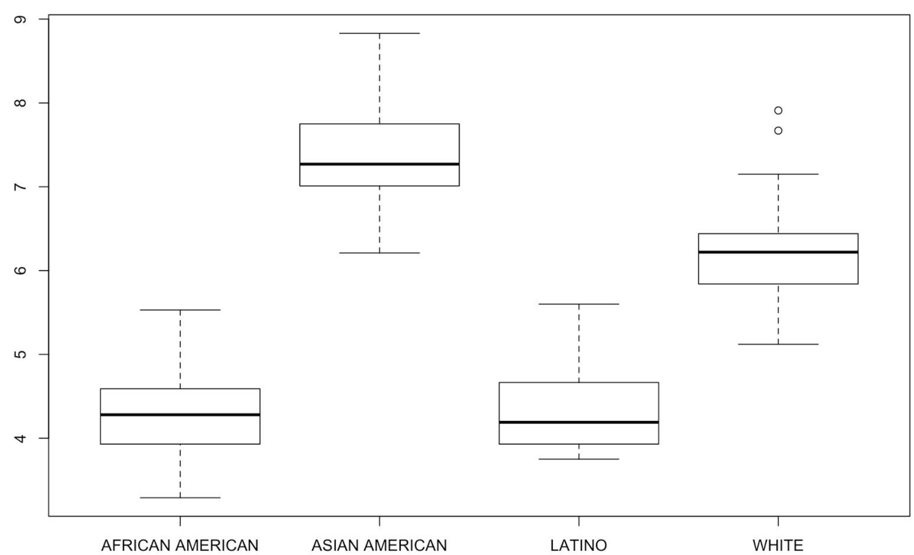
 # Mid Term 2

```
install.packages("readxl")
setwd("/opt/Code/My/R-Tests/regression/mt2")
demoData <- readxl::read_xlsx("MOAData.xlsx")
colnames(demoData) <- c("race","year", "area", "hdIndex",
"lifeExpAtBirth","gradDeg","schEnroll","medianEarn","healthIndex","eduIndex","incomeIndex")
par(mfrow=c(1,1))
```

a) Compare HD index for the other races

```
boxplot(demoData$hdIndex~demoData$race)
```



Comments: By looking at the plot it seems that mean HD index is over all higher for asian american people. It is relatively lower for african american and latino people.
Where mean HD Index for white people is higher than Latino and African american but lower than Asian American people. There are a few outlier in the higher side of HD index for white people.

b) Predictor - Race & median income, additive regression model for response HD Index

```
demoHdLm1<-lm(hdIndex~factor(race)+medianEarn,demoData)
summary(demoHdLm1)
```

```
Call:
lm(formula = hdIndex ~ factor(race) + medianEarn, data = demoData)
Residuals:
    Min      1Q  Median      3Q     Max
-0.7308 -0.2324  0.0073  0.1906  0.8993
Coefficients:
                              Estimate Std. Error t value   Pr(>|t|)
(Intercept)                 1.20537586 0.22447945    5.37 0.00000066 ***
factor(race)ASIAN AMERICAN 2.25357499 0.11743121   19.19    < 2e-16 ***
factor(race)LATINO          0.68021373 0.10995107    6.19 0.00000002 ***
```

```
factor(race)WHITE         0.70494933 0.12802229    5.51 0.00000037 ***
medianEarn              0.00011090 0.00000772   14.36    < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.34 on 86 degrees of freedom
Multiple R-squared:  0.949,        Adjusted R-squared:  0.946
F-statistic:  398 on 4 and 86 DF,  p-value: <2e-16
```

Comments:

The model is Y= B0 + B1 X1 + B2 X2 + B3 X3 + B4 X4

Here

B0 =   1.205 (Intercept)

B1 = 2.254; X1 = 1 (race = Asian American), 0 (Otherwise)

B2 = 0.68; X2 = 1 (Latino), 0 (Otherwise)

B3 = 0.704; X3 = 1 (White), 0 (Otherwise)

B4 = 0.00011; X4 => median earn

R-squared value is large enough to validate the significance of the model. Also p-value being being very small confirms the validity of the coefficients.

F-statistic is slightly high so we can try including interaction terms.

c) Full model with all possible interaction terms

```
demoHdLm2<-lm(hdIndex~factor(race)+medianEarn + factor(race)*medianEarn,demoData)
summary(demoHdLm2)
```

```
Call:
lm(formula = hdIndex ~ factor(race) + medianEarn + factor(race) *
    medianEarn, data = demoData)

Residuals:
    Min      1Q  Median      3Q     Max
-0.7198 -0.2167 -0.0191  0.1733  0.8536

Coefficients:
                                      Estimate  Std. Error t value     Pr(>|t|)
(Intercept)                         1.38075138  0.45994217    3.00       0.0035
factor(race)ASIAN AMERICAN          1.98979746  0.67911040    2.93       0.0044
factor(race)LATINO                 -2.53625134  1.27979430   -1.98       0.0508
factor(race)WHITE                   0.70284335  0.63238786    1.11       0.2696
medianEarn                          0.00010457  0.00001641    6.37 0.0000000099
factor(race)ASIAN AMERICAN:medianEarn  0.00000879  0.00002143    0.41       0.6829
factor(race)LATINO:medianEarn       0.00014448  0.00005659    2.55       0.0125
factor(race)WHITE:medianEarn        0.00000186  0.00001977    0.09       0.9251

(Intercept)                         **
factor(race)ASIAN AMERICAN          **
factor(race)LATINO                  .
factor(race)WHITE
```

```
medianEarn                              ***
factor(race)ASIAN AMERICAN:medianEarn
factor(race)LATINO:medianEarn          *
factor(race)WHITE:medianEarn
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.328 on 83 degrees of freedom
Multiple R-squared:  0.953,	Adjusted R-squared:  0.949
F-statistic:  239 on 7 and 83 DF,  p-value: <0.0000000000000002
```

Comments:
After adding the interaction terms, F-Statistic value has decreased. R-squared is high enough to validate the model.
Full model : Y = B0 + B1 X1 + B2 X2 + B3 X3 + B4 X4 + B5 X1 X4 + B6 X2 X4 + B7 X3 X4
B0 = 1.381 (Intercept)
B1 = 1.99; X1 = 1 (race = Asian American), 0 (Otherwise)
B2 = -2.54; X2 = 1 (Latino), 0 (Otherwise);
B3 =0.70; X3 = 1 (White), 0 (Otherwise);
B4 = 0.00010; X4 => median earn
B5 = 0.00000879 (Interaction of Asian American race with Median Income); **p-value = 0.68**
B6 = 0.00014448 (Interaction of Latino race with Median Income)
B7 = 0.00000186 (Interaction of White race with Median Income); **p-value= 0.9251**

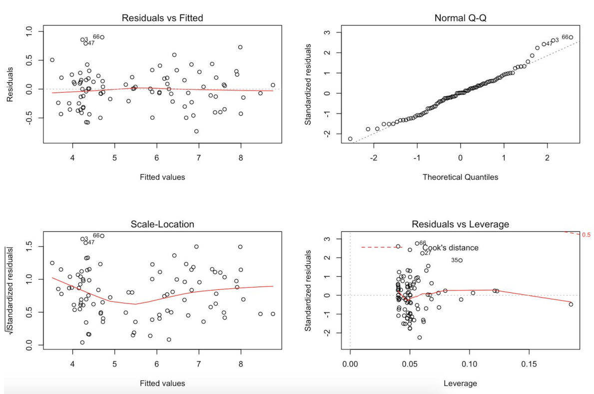
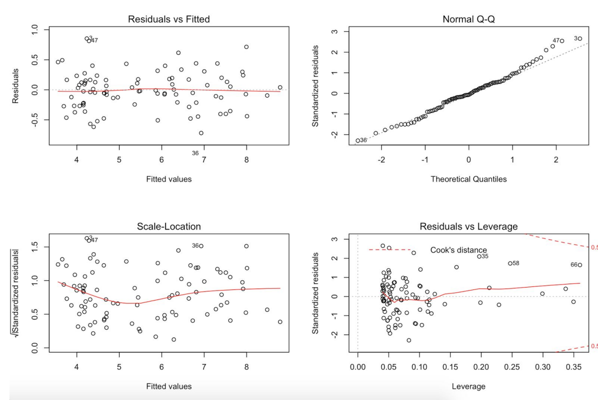
By looking at p values for B5 and B7, we should remove these terms as p-value is higher than 0.5 but since the coefficients are very small, we can ignore these.

d) Visual displays for both additive and full model

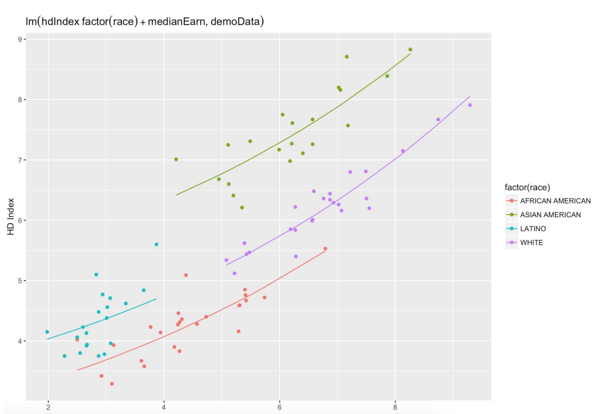
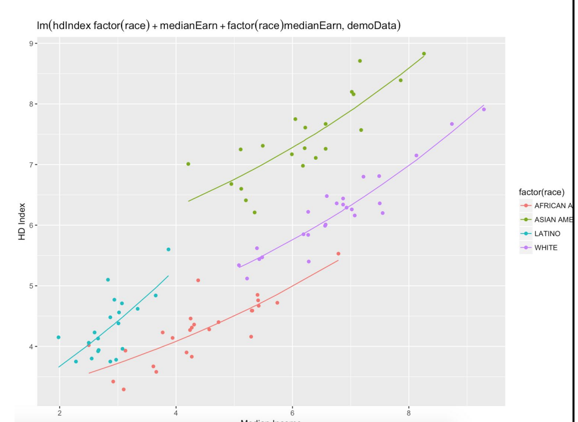
```
> anova(demoHdLm1, demoHdLm2)
Analysis of Variance Table

Model 1: hdIndex ~ factor(race) + medianEarn
Model 2: hdIndex ~ factor(race) + medianEarn + factor(race) * medianEarn
  Res.Df  RSS Df Sum of Sq    F Pr(>F)  
1     86 9.68                           
2     83 8.94  3     0.738 2.28  0.085 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
par(mfrow=c(2,2))
install.packages("ggplot2")
library(ggplot2)
```

Y= B0 + B1 X1 + B2 X2 + B3 X3 + B4 X4	Y = B0 + B1 X1 + B2 X2 + B3 X3 + B4 X4 + B5 X1 X4 + B6 X2 X4 + B7 X3 X4

<table>
<tr><td>plot(demoHdLm1)</td><td>plot(demoHdLm2)</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td></td><td>

- Residual vs fitted plot is better with this model
- Slightly better r-squared

</td></tr>
<tr><td>ggplot(data=demoData, aes(x=incomeIndex, y=hdIndex, colour=factor(race))) + geom_point() + xlab("Median Income") + ylab("HD Index") + geom_line(aes(y=demoHdLm1$fitted.values)) + ggtitle(summary(demoHdLm1)$call)</td><td>ggplot(data=demoData, aes(x=incomeIndex, y=hdIndex, colour=factor(race))) + geom_point() + xlab("Median Income") + ylab("HD Index") + geom_line(aes(y=demoHdLm2$fitted.values)) + ggtitle(summary(demoHdLm2)$call)</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td></td><td>As we can see, this model is improved as the observations (for African american and Latino) are closer to regression line.</td></tr>
</table>

e) Include life expectancy as predictor to build additive model

```
demoHdLm3 <-lm(hdIndex~factor(race)+medianEarn+lifeExpAtBirth,demoData)
summary(demoHdLm3)
```

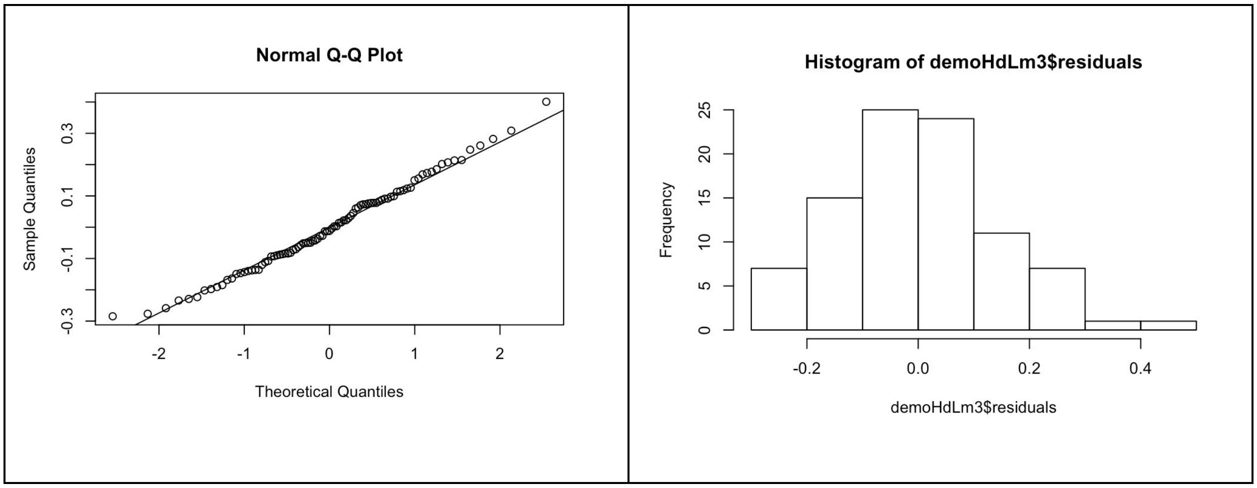
```
Call:
lm(formula = hdIndex ~ factor(race) + medianEarn + lifeExpAtBirth,
    data = demoData)

Residuals:
    Min      1Q  Median      3Q     Max
-0.2847 -0.0931 -0.0117  0.0912  0.4008

Coefficients:
                          Estimate  Std. Error t value            Pr(>|t|)
(Intercept)             -10.2242078   0.6101484  -16.76 < 0.0000000000000002
factor(race)ASIAN AMERICAN   0.5251107   0.1046638    5.02       0.000002840516
factor(race)LATINO          -0.6690835   0.0859496   -7.78       0.000000000015
factor(race)WHITE            0.2202769   0.0617825    3.57               0.0006
medianEarn                 0.0000944   0.0000035   26.95 < 0.0000000000000002
lifeExpAtBirth             0.1574541   0.0082948   18.98 < 0.0000000000000002

(Intercept)                ***
factor(race)ASIAN AMERICAN ***
factor(race)LATINO         ***
factor(race)WHITE          ***
medianEarn                 ***
lifeExpAtBirth             ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.147 on 85 degrees of freedom
Multiple R-squared:  0.99,	Adjusted R-squared:  0.99
F-statistic: 1.72e+03 on 5 and 85 DF,  p-value: <0.0000000000000002
```

Comments:
Y= B0 + B1 X1 + B2 X2 + B3 X3 + B4 X4 + B5 X5
Where
B0 = -10.2242078 (Interccept)
B1 = 0.5251107; X1 = 1 (race = Asian American), 0 (Otherwise)
B2 = -0.6690835; X2 = 1 (Latino), 0 (Otherwise)
B3 = 0.2202769; X3 = 1 (White), 0 (Otherwise)
B4 = 0.0000944; X4 =>median income
B5 = 0.1574541; X5 => lifeExpAtBirth
R-squared value is large enough (and has increased to 0.99 in comparison with other models) to validate the significance of the model. Also p-value being being very small confirms the significance of the coefficients.
F-statistic is high so we can try including interaction terms.

Residual plots for normality and formal tests:

```
plot(demoHdLm3)
qqnorm(demoHdLm3$residuals)
```

```
qqline(demoHdLm3$residuals)
hist(demoHdLm3$residuals)
```



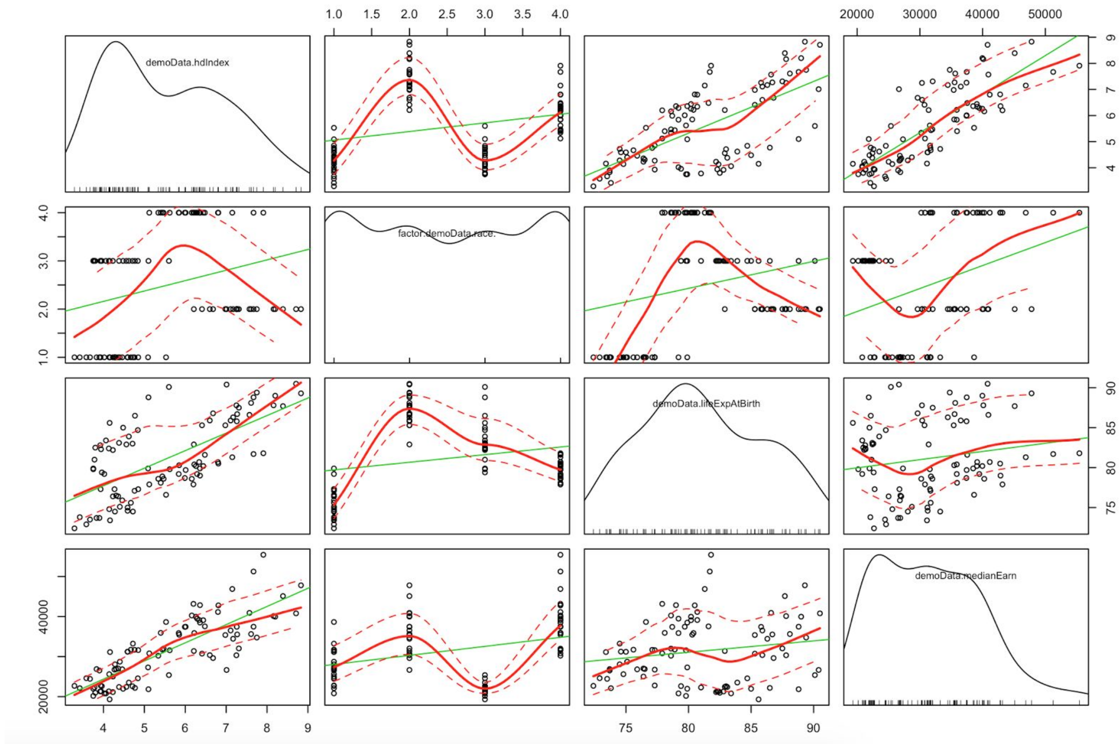
Residual distribution looks Normal.

Shapiro-Wilk normality test

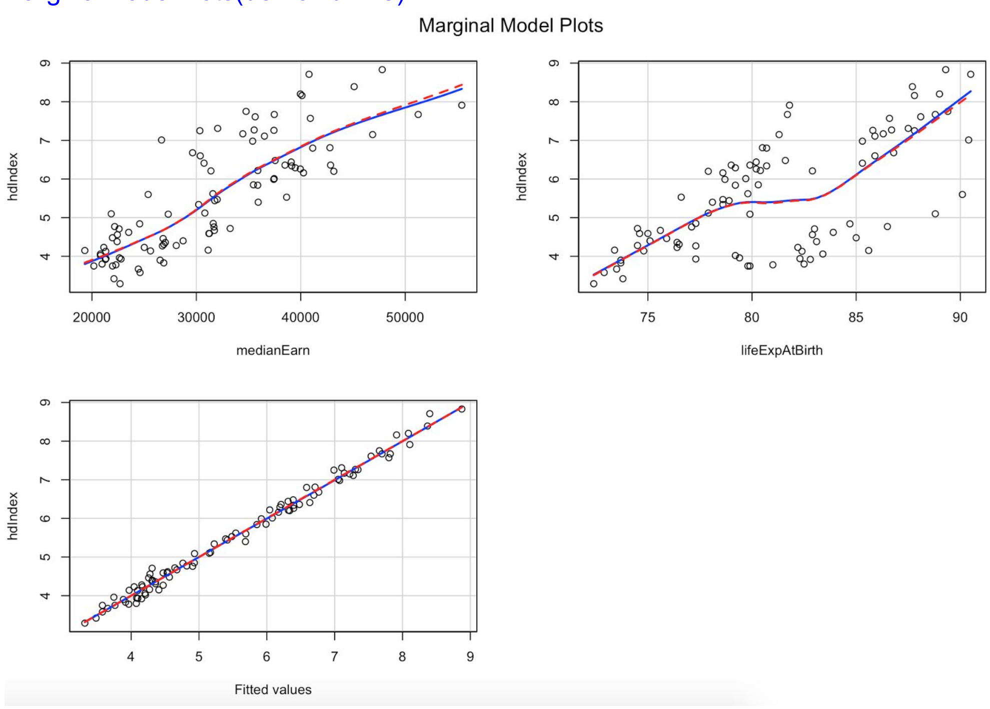
data: demoHdLm3$residuals
W = 1, p-value = 0.7

Scatter and marginal model plots

```
install.packages("car")
library(car)
scatterplotMatrix(~demoData$hdIndex+factor(demoData$race)+demoData$lifeExpAtBirth+demoData$inc
omeIndex)
```



marginalModelPlots(demoHdLm3)



By looking at the fitted vs observed value plot, it seems that the model fits great.

Constancy of Variance:

> ncvTest(demoHdLm3)

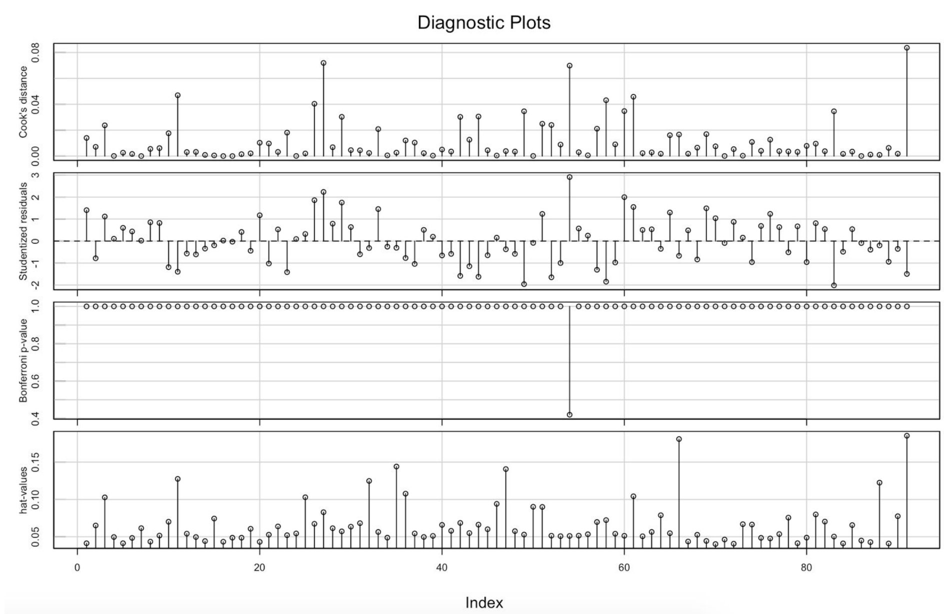
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.0001   Df = 1   p = 0.992

High p value confirms the constancy in variance.

f) Checking outliers in data

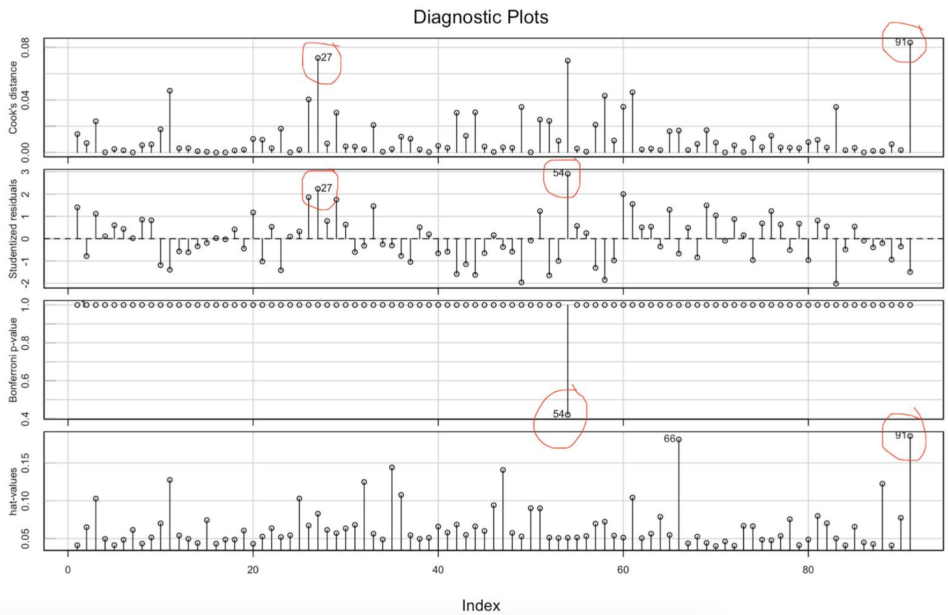
```
outlierTest(demoHdLm3)
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
  rstudent unadjusted p-value Bonferonni p
54     2.91          0.00461       0.42
```

```
influenceIndexPlot(demoHdLm3)
```

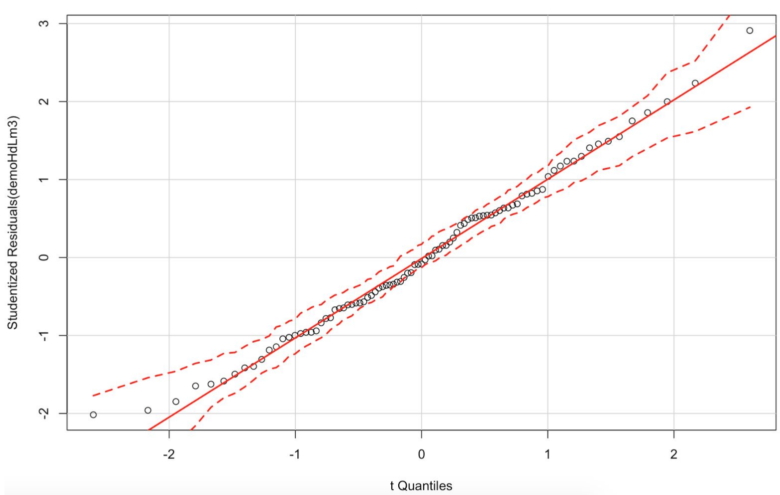


Finding top two in each

```
> influenceIndexPlot(demoHdLm3, id.n=2)
```

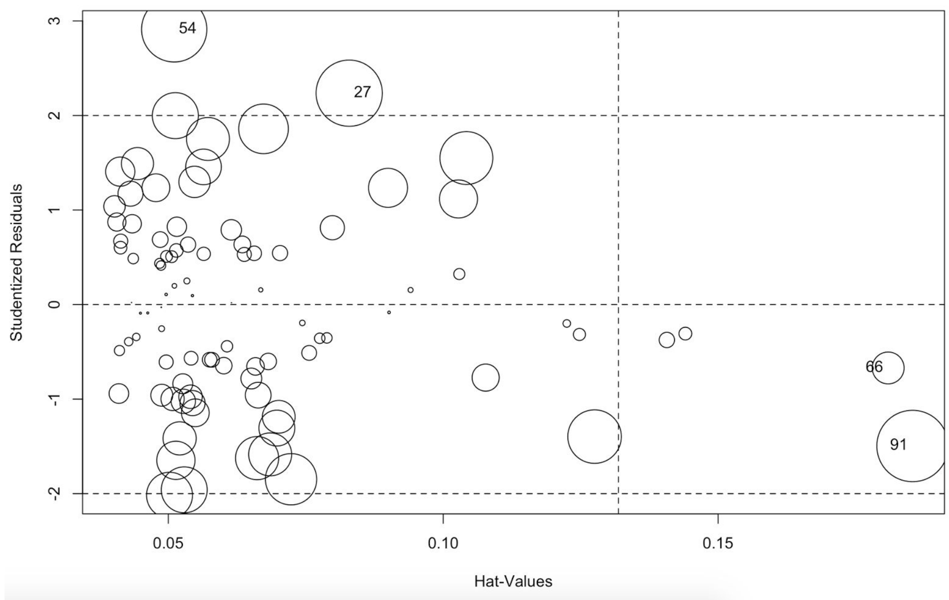


qqPlot(demoHdLm3)



QQ-Plot for studentized residuals.

Influence Plot for Cook's distance confirms the outliers to 27, 54, 66, 91.
Two points with largest influence are 54 and 91.



g) Using AIC to justify the choice of model with 3 predictors

```
# Step function
nullModel <- lm(hdIndex~1,demoData)
summary(nullModel)
Call:
lm(formula = hdIndex ~ 1, data = demoData)

Residuals:
   Min     1Q Median     3Q    Max
-2.262 -1.277 -0.152  1.088  3.278

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)
(Intercept)    5.552      0.152    36.5 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.45 on 90 degrees of freedom

step(nullModel, scope=list(lower=nullModel, upper=demoHdLm3), direction="forward")
Start:  AIC=68.5
hdIndex ~ 1

                Df Sum of Sq   RSS   AIC
+ factor(race)   3     156.1  32.9 -84.6
+ medianEarn     1     126.7  62.3 -30.5
+ lifeExpAtBirth 1      82.3 106.7  18.4
<none>                       189.0  68.5
```

```
Step:  AIC=-84.6
hdIndex ~ factor(race)

               Df Sum of Sq  RSS    AIC
+ medianEarn      1      23.2  9.7 -193.9
+ lifeExpAtBirth  1      15.3 17.6 -139.3
<none>                        32.9  -84.6

Step:  AIC=-194
hdIndex ~ factor(race) + medianEarn

               Df Sum of Sq  RSS  AIC
+ lifeExpAtBirth  1      7.83 1.85 -343
<none>                        9.68 -194
```

```
Step:  AIC=-343
hdIndex ~ factor(race) + medianEarn + lifeExpAtBirth


Call:
lm(formula = hdIndex ~ factor(race) + medianEarn + lifeExpAtBirth,
    data = demoData)

Coefficients:
            (Intercept)  factor(race)ASIAN AMERICAN
            -10.2242078                 0.5251107
      factor(race)LATINO          factor(race)WHITE
             -0.6690835                 0.2202769
             medianEarn             lifeExpAtBirth
              0.0000944                 0.1574541
hdIndex ~ factor(race) + medianEarn + lifeExpAtBirth
```

Above model is the most significant and efficient model as this one has lowest AIC value.