Homework 8

Snigdha Peddi

Exercises

Question 1. Apply a median regression analysis on the **clouds** data. Compare this to the linear regression model from Chapter 6. Write up a formal summary of the two analyses and provide a justified recommendation on which analysis the researcher should be using.

INTRODUCTION:

Clouds data from **HSAUR3** package consists of experimental data investigating the use of massive amounts of Silver Iodide (100 to 1000 grams per cloud) in cloud seeding to increase rainfall.This data is collected from an area in Florida ,in the summer of 1975. 24 Days were judged suitable for seeding based on the measured suitability criterion. This data has 24 observations and 7 variables.These variables include *Seeding* which is factor indicating if seeding has occurred,*time* is the number of days after the first day of experiment, *sne* is the suitability criterion, *cloudcover* is percentage of experimental area covered with clouds, *prewetness* is the total rainfall in the target area one hour before seeding, *echomotion* is a factor showing if the radar echo is moving or stationary and *rainfall* is the amount of rain in cubic meters times 10ˆ7.

ANALYSIS AND RESULTS:

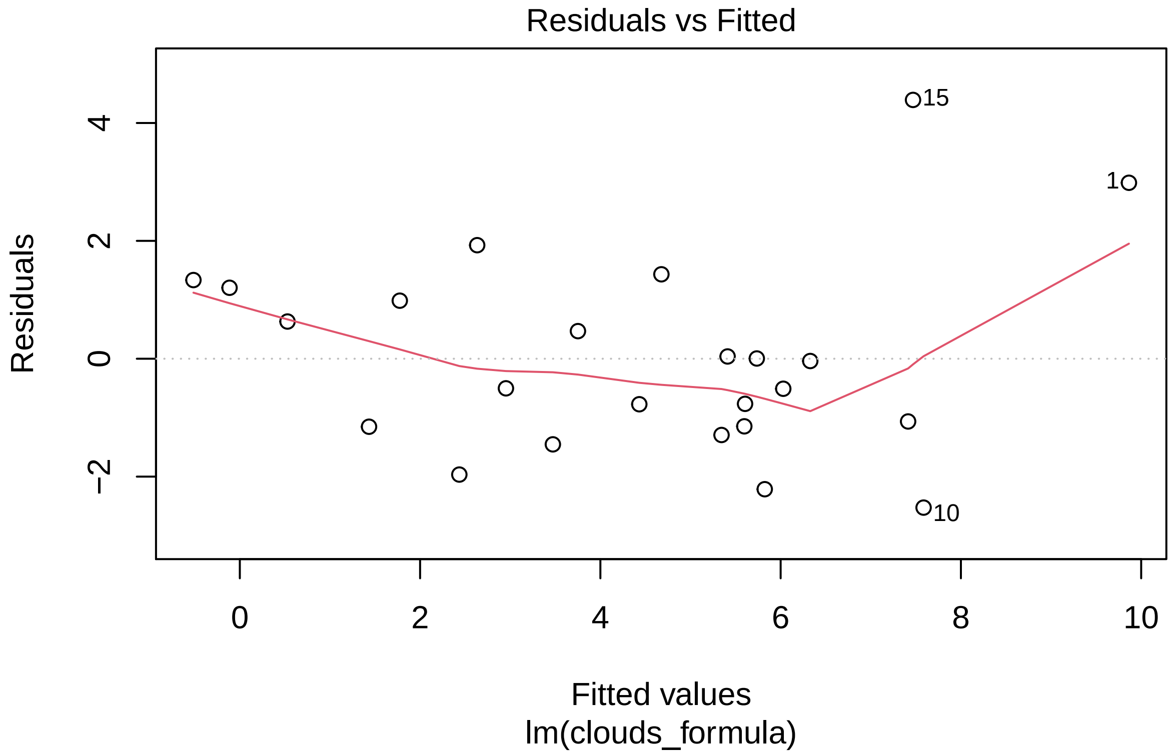
The Clouds data is studied and presented by Brian S. Everitt and Torsten Hothorn in their book, A Handbook of Statistical Analyses using R. A Linear regression model is fit using rainfall as dependent variable. Seeding, time and interaction terms for seeding were used as the covariates in this model.

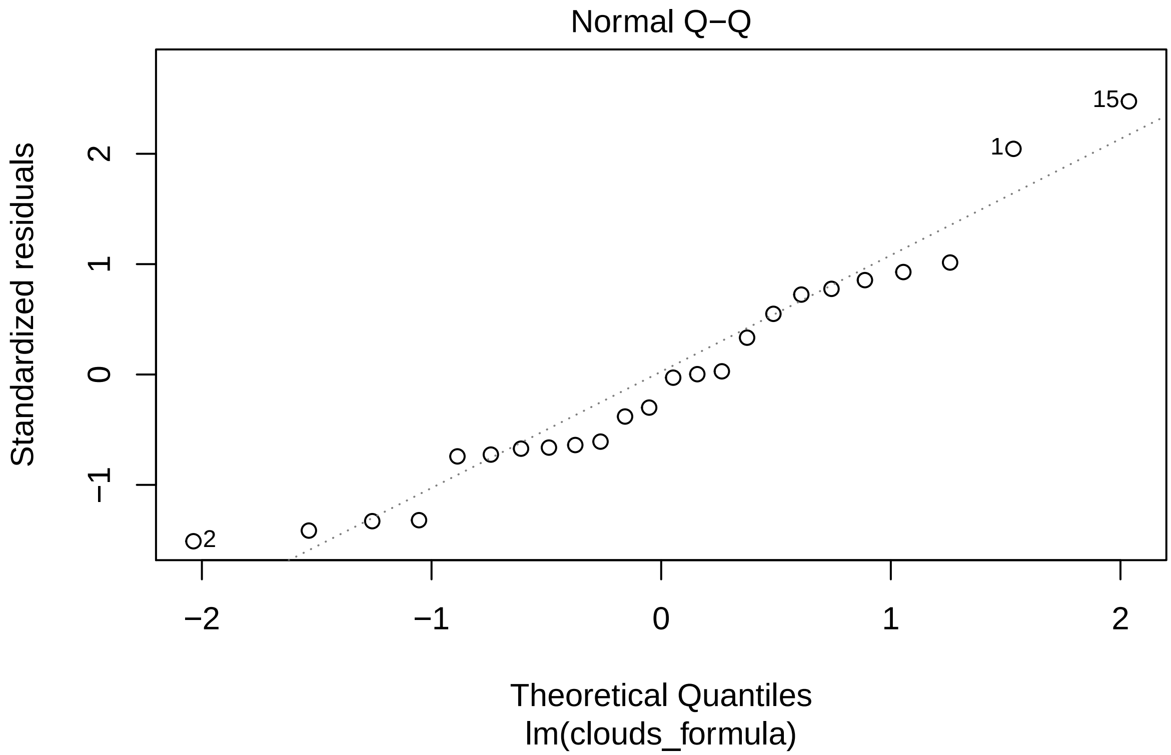
```
clouds_formula<- rainfall~seeding+seeding:(sne+cloudcover+prewetness+echomotion)+time
clouds_lm <-lm(clouds_formula,data=cloud)
```

```
##                                 Coefficients.of.Linear.Model     p.value
## (Intercept)                                      -0.34624093 0.903055585
## seedingyes                                       15.68293481 0.003715091
## time                                             -0.04497427 0.095897358
## seedingno:sne                                     0.41981393 0.627420901
## seedingyes:sne                                   -2.77737613 0.010404243
## seedingno:cloudcover                              0.38786207 0.098385030
## seedingyes:cloudcover                            -0.09839285 0.388538450
## seedingno:prewetness                              4.10834188 0.274499083
## seedingyes:prewetness                             1.55127493 0.574408155
## seedingno:echomotionstationary                    3.15281358 0.126772119
## seedingyes:echomotionstationary                   2.59059513 0.177565483
```

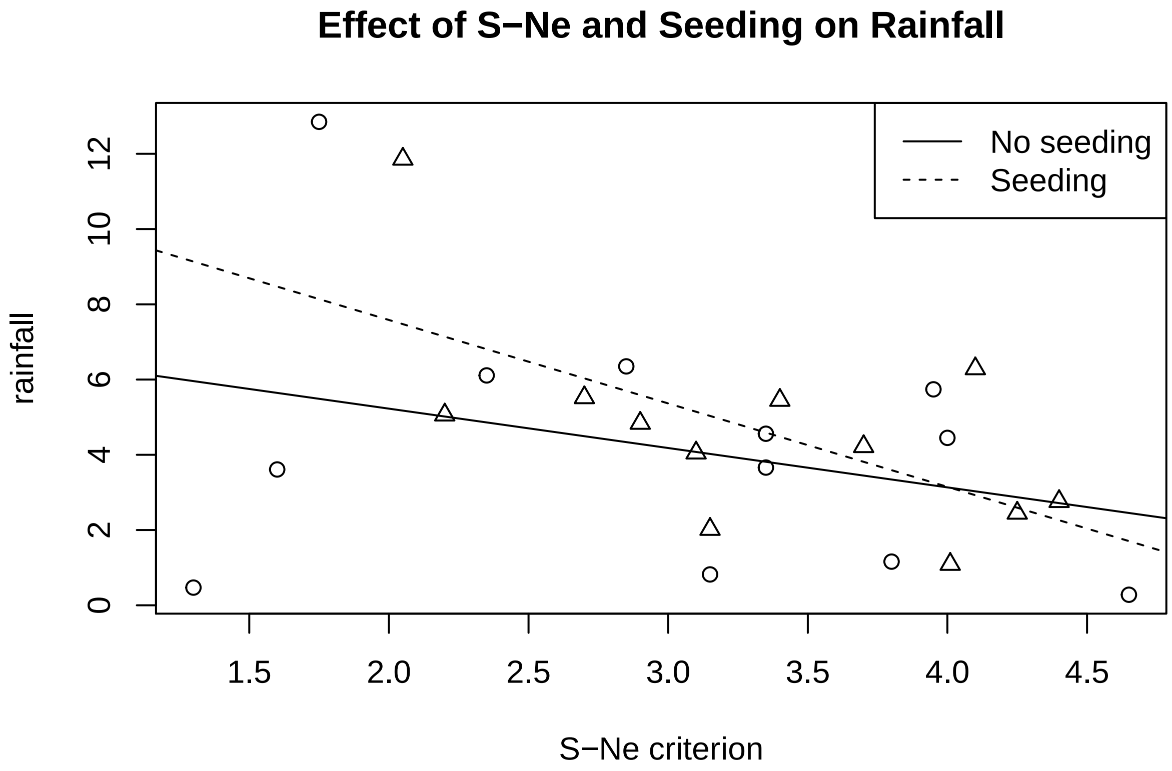
The Linear model of clouds data suggests that seeding the clouds increases the rainfall and higher value of SNe leads to lower rainfall when seeding is done. The interaction of Sne significantly effects rainfall.

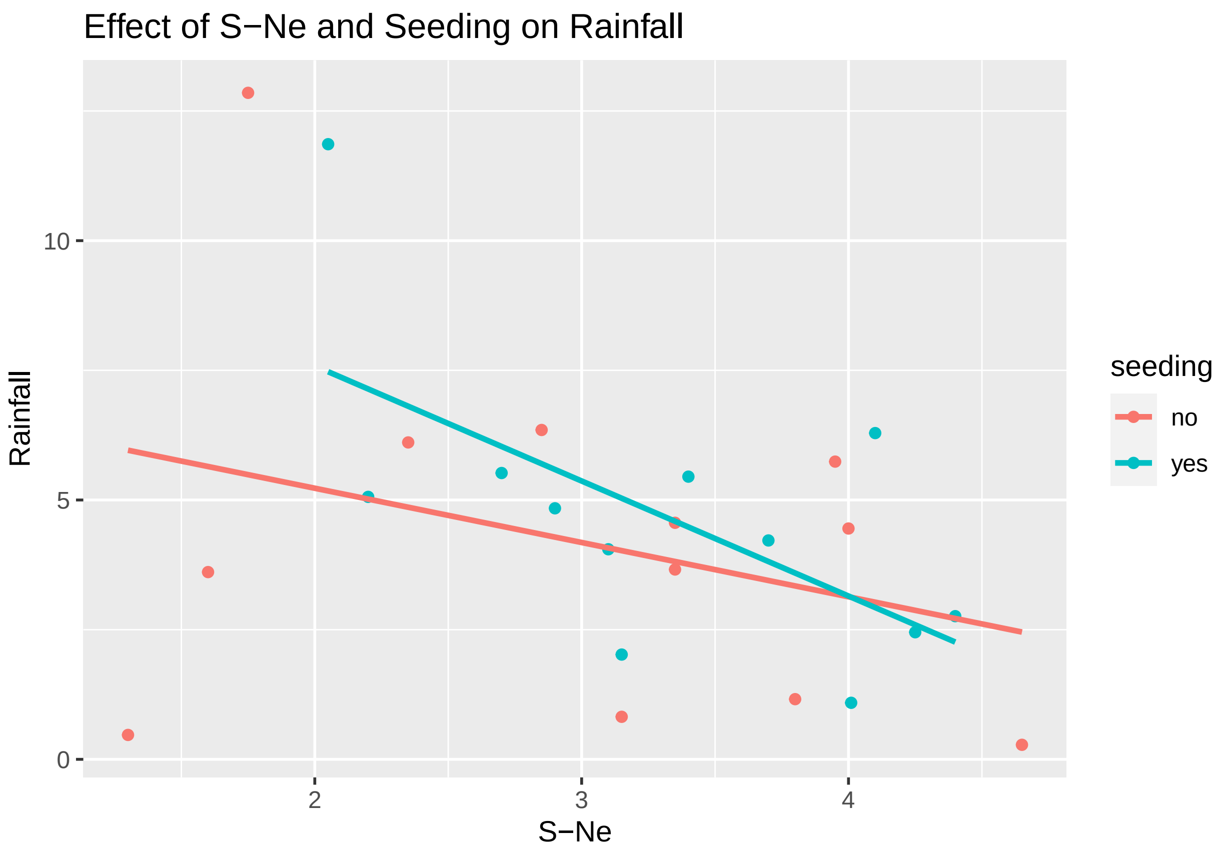
The lower p values of less than 0.05 indicates that the effect is significant at 95% confidence interval. The Residual vs fitted plot shows that the residuals forms a curve and are not normally distributed and Quantile-qunatile plot resulted in an approximately straight line showing that the residuals are approximately normally distributed.Removing the outliers and refitting the model can improve the model.





The relationship between rainfall and S-Ne for seeding and non-seeding days is plotted and a linear regression is fit for both seeding and non seeding data.The plots indicates that for a lower S-Ne value,seeding produces more rainfall than non-seeding.A cross over is observed around a S-Ne value of 4.0 indicating that seeding has to be done at S-Ne values lower than 4.0 to improve the rainfall. Similar plot is created using ggplot2.


Effect of S−Ne and Seeding on Rainfall
No seeding
Seeding
rainfall
0
2
4
6
8
10
12
1.5
2.0
2.5
3.0
3.5
4.0
4.5
S−Ne criterion



To better understand the Clouds data and relationship between S-Ne and rainfall with and without seeding, a median Quantile regression model is fit. A Quantile regression model estimates the conditional median rather than the conditional mean estimates of linear regression model. This model will help us understand the relation between the variables at 50th percentile or 50% Quantile.

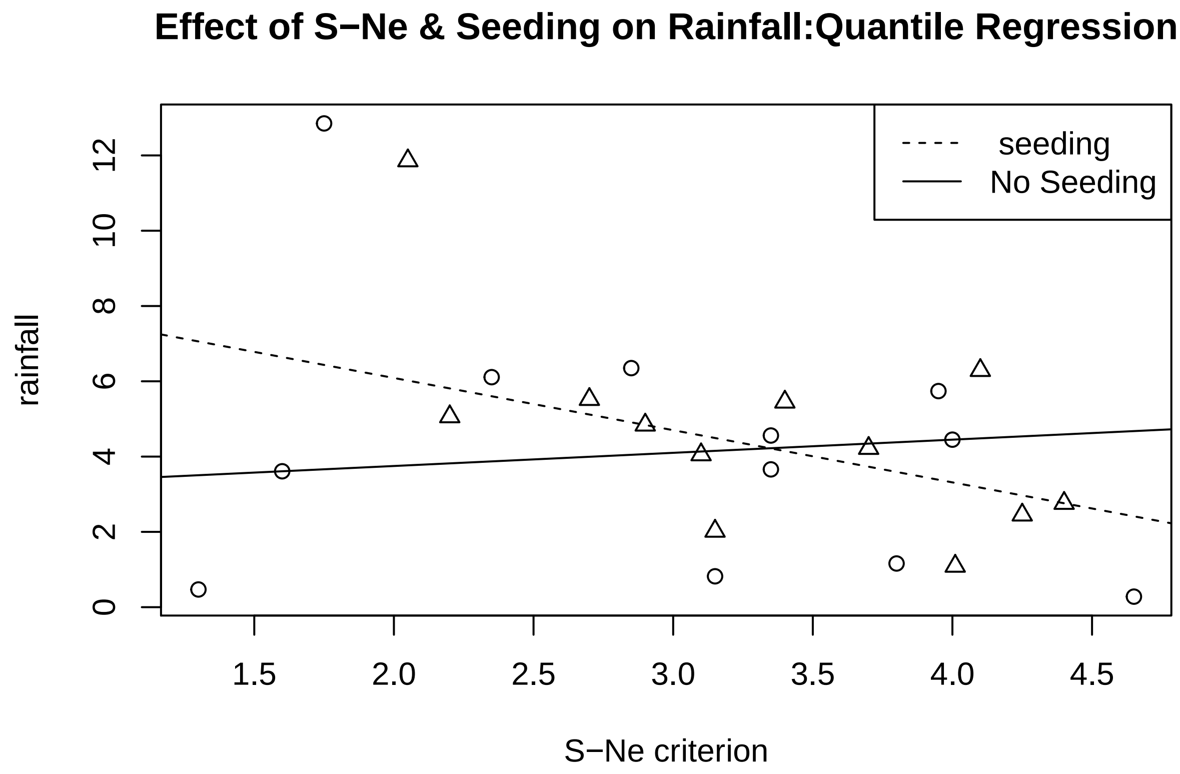
```
  rq1 <- rq(rainfall ~ seeding+seeding:(sne+cloudcover+prewetness+echomotion)+time,
  data = cloud, tau = 0.5)
```

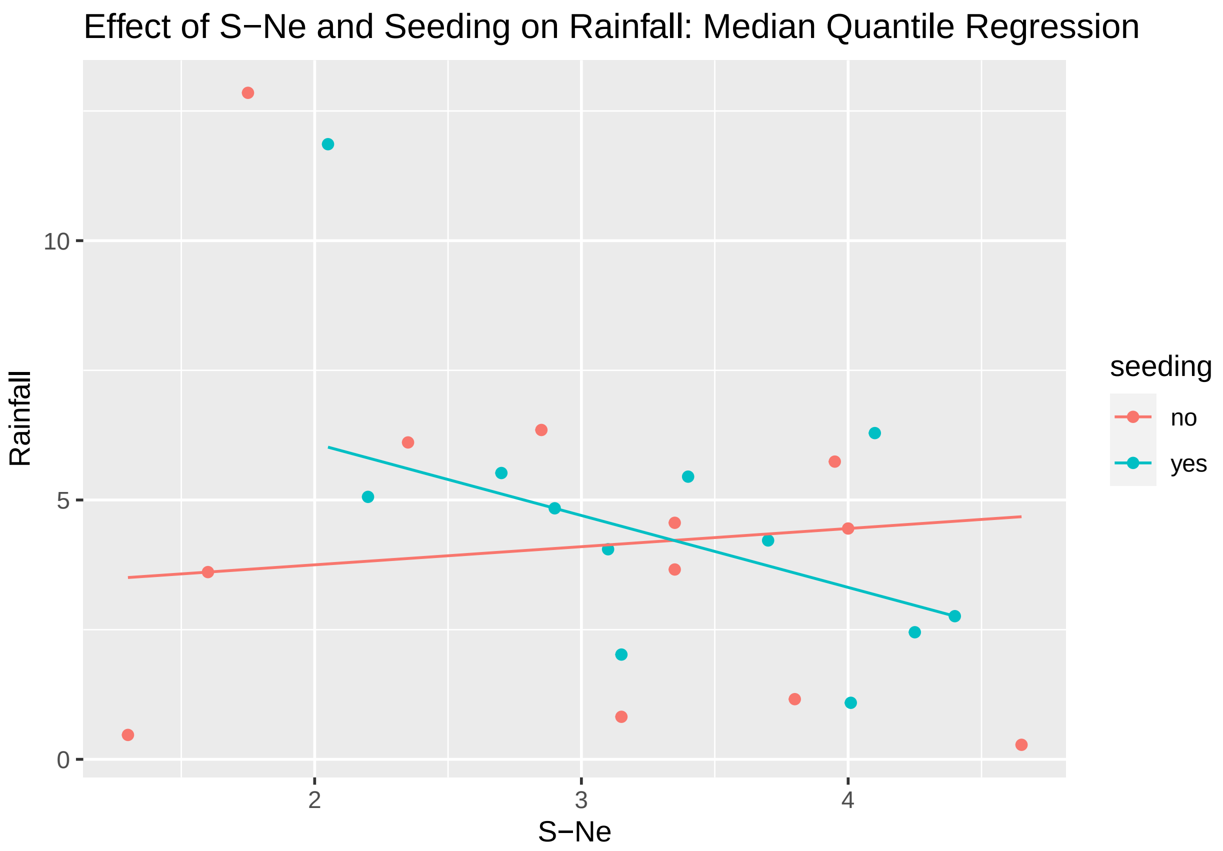
```
##                                  Coefficients.at.50th.Quantile
## (Intercept)                                        -0.39510353
## seedingyes                                          9.28416250
## time                                               -0.02682160
## seedingno:sne                                       0.36860476
## seedingyes:sne                                     -1.33267160
## seedingno:cloudcover                                0.20691306
## seedingyes:cloudcover                              -0.06071068
## seedingno:prewetness                                5.22263667
## seedingyes:prewetness                               2.01808261
## seedingno:echomotionstationary                      2.13502276
## seedingyes:echomotionstationary                     2.78255068
```

```
  rq2 <- rq(rainfall ~sne, data = cloud, tau = 0.5,subset=seeding=="yes")
  rq3 <- rq(rainfall ~sne, data = cloud, tau = 0.5,subset=seeding=="no")
```

```
##                Seeding Non.Seeding
## (Intercept)  8.861333        3.05
## sne         -1.386667        0.35
```

It is clear from the coefficients of the model that, at 50% Quantile, the regression line is less inclined (-1.39) when seeding is performed and has a negative slope (0.35) when no seeding is performed compared to the linear model with coefficients -2.78 and 0.42 respectively. From the plot, one can clearly see a cross over that occurs at around S-Ne of 3.4 indicating that when seeding is done at S-Ne values lower than 3.4 there is a greater potential for rainfall. A similar plot is created using ggplot2.





CONCLUSION:

At the 50th quantile, there is no significant effect on rainfall when seeding is not performed and with seeding, increase in rainfall is by 9.25 X 10^7 cubic meters which is lower than linear regression model (15.68 x 10^7). The regression line is less inclined with a coefficient greater than the linear model (-1.39 > -2.78). I would recommend the use of Median Quantile regression over the linear regression as the effect of outliers is reduced at median level and it helps to understand the relation between S-Ne and Rainfall better, which in turn helps pick suitable days for seeding.

REFERENCES

- **Quantile Regression Documentation** from tidyverse.org (https://ggplot2.tidyverse.org/reference/geom_quantile.html)
- Lecture code
- Ani Katchova, **Quantile Regression**, 2013 (https://www.youtube.com/watch?v=P9lMmEkXuBw)
- Ani Katchova, **Quantile Regression in R**, 2013 (https://www.youtube.com/watch?v=ucURUTVjBRo)

Question 2. Reanalyze the **bodyfat** data from the **TH.data** package.

a) Compare the regression tree approach from chapter 9 of the textbook to median regression and summarize the different findings.
b) Choose one dependent variable. For the relationship between this variable and DEXfat, create linear regression models for the 5%, 10%, 90%, and 95% quantiles. Plot DEXfat vs that dependent variable and plot the lines from the models on the graph.
c) Provide a formal write up of the methodologies and of your results

INTRODUCTION

Bodyfat data from *mboost* package consists of anthropometric measurements obtained from 71 German women along with their body composition measured by Dual Energy X-Ray Absorptiometry (DXA).Body fat is considered as a better predictor of metabiloc syndromes like diabetes mellitus and cardiovascular diseases. Body fat measured by DXA finds little applicability due to its high cost and methodological efforts needed.This data has 6 variables.*DEXfat* is the body fat measured by Dual Energy X-Ray Absortiometry,*age* is the age of women in years,*waistcirc* is the waist circumference,*hipcirc* is the hip circumference,elbowbreadth is the breadth of elbow and *kneebreadth* is the breadth of the knee.

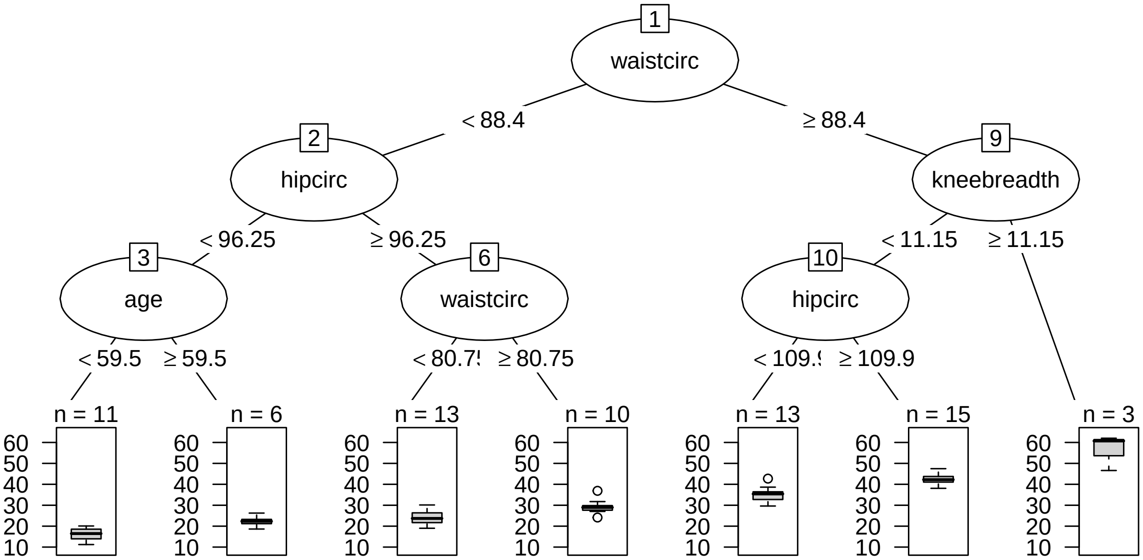
ANALYSIS AND RESULTS

A multivariate analysis is done by fitting a Regression tree model.A decision tree is created by partitioning the observations by univariate splits in a recursive way.*rpart* function is used to grow the decision tree. All the independent variables were used to fit the model and DEXfat is used as dependent variable.A control is placed in the model where, a minimum of 10 observations is required to continue the split.

```
   bodyfat_rpart <-rpart(DEXfat ~ age+waistcirc+hipcirc+elbowbreadth+kneebreadth,
   data=bodyfat,control=rpart.control(minsplit=10))
```

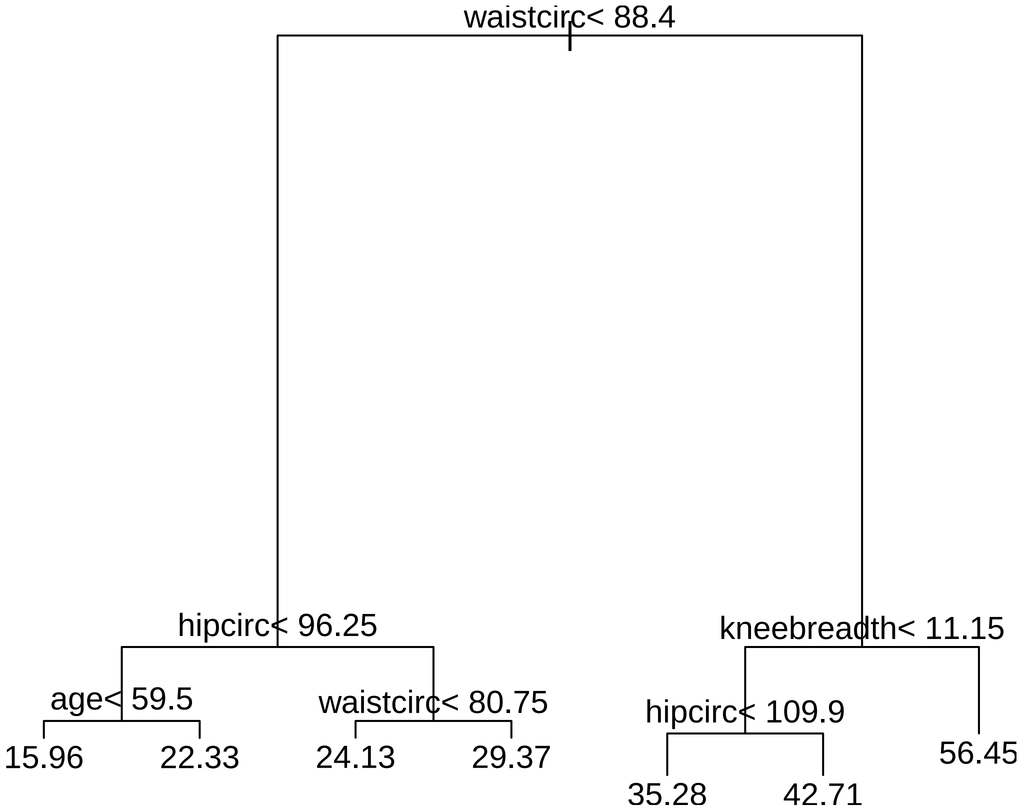
Based on the variable importance the tree is split at several nodes to the leaf nodes.From the plot it is clear that the Bodyfat data is split into 13 leaf nodes, with a total of 6 splits.If a condition is satisfied the observations are branched to the left and those do not satisfy branch towards right.The higher values of the Waist circumference, Hip circumference and knee breadth corresponds to the higher DEXfat. Pruning of the tree can be performed based the cross validation error(Xerror) from the CP table of the model summary. In case of the Bodyfat data the best tree has four splits.

```
##              Variable.Importance
## waistcirc              6412.4407
## hipcirc                5683.2506
## kneebreadth            4617.7360
## elbowbreadth           1391.6317
## age                     806.1666
```



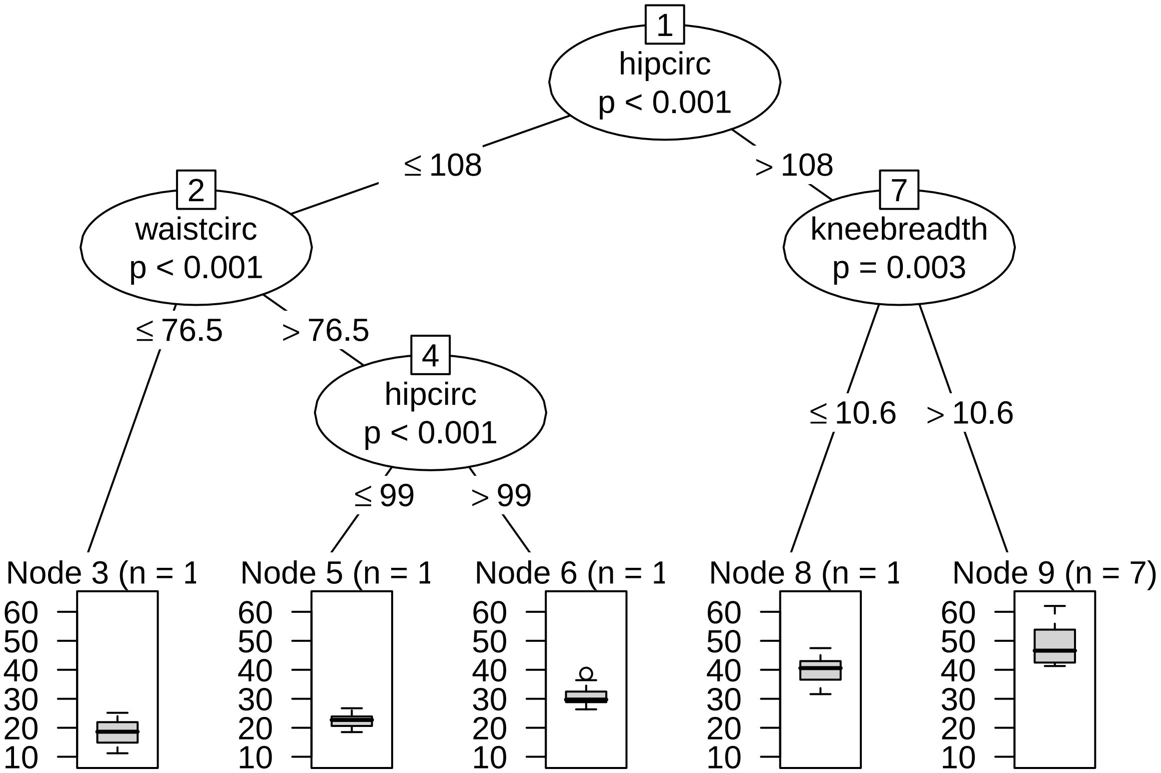
```
##           CP nsplit  rel.error    xerror       xstd
## 1 0.66289544      0 1.00000000 1.0283444 0.16914894
## 2 0.09376252      1 0.33710456 0.4415941 0.09479676
## 3 0.07703606      2 0.24334204 0.4286021 0.08708594
## 4 0.04507506      3 0.16630598 0.3710368 0.07283959
## 5 0.01844561      4 0.12123092 0.2953664 0.06165134
## 6 0.01818982      5 0.10278532 0.2835620 0.06105742
## 7 0.01000000      6 0.08459549 0.2617490 0.06044859
```

Dendrogram of the Bodyfat Regression Tree



A Conditional inference tree can be computed using ctree function which is based on minimum Standard Error at each node. This is more advantageous as we do not need to prune the initial large trees since the tree is spliit based on a p-value(stopping criterion).

```
bodyfat_ctree <- ctree (DEXfat ~age+waistcirc+hipcirc+elbowbreadth+kneebreadth,data=bodyfat )
```



A median regression is performed with DEXfat as dependent variable and all independent variables. Quantile regression methodology is a non parametric method which is used to understand the relationship between the variables outside the mean. A median regression is obtained at 50% Quantile where observations are divided into two equal groups.The Coefficients show that waist circumference (0.28), Hip circumference (0.51) and knee breadth (0.76)are more significant compared to other variables.At 50th percentile, With increase in a unit of hip circumference there is 0.51 units increase in Dexfat and With increase in a unit of knee breadth there is 0.76 units increase in Dexfat.

```
Tree <- DEXfat ~ age+waistcirc+hipcirc+elbowbreadth+kneebreadth
Tree.qr.50 <-rq(Tree ,data=bodyfat,tau = 0.5)
```

```
##              Median.Quantile
## (Intercept)     -57.30031520
## age               0.06839443
## waistcirc         0.28332466
## hipcirc           0.51073243
## elbowbreadth     -0.11982312
## kneebreadth       0.76452936
```

Linear regression lines are fit at different quantiles (5%,10%,90% and 95%) and the coefficients are compared to check if the there is any significant difference at these levels. From the Coefficients it is clear that as the we go from 5th Quantile to 95th Quantile the significance of both hip circumference and knee breath increases indicating heteroscedasticity in the data.

```
Tree.qr.5 <-rq(Tree ,data=bodyfat,tau = 0.05)
```

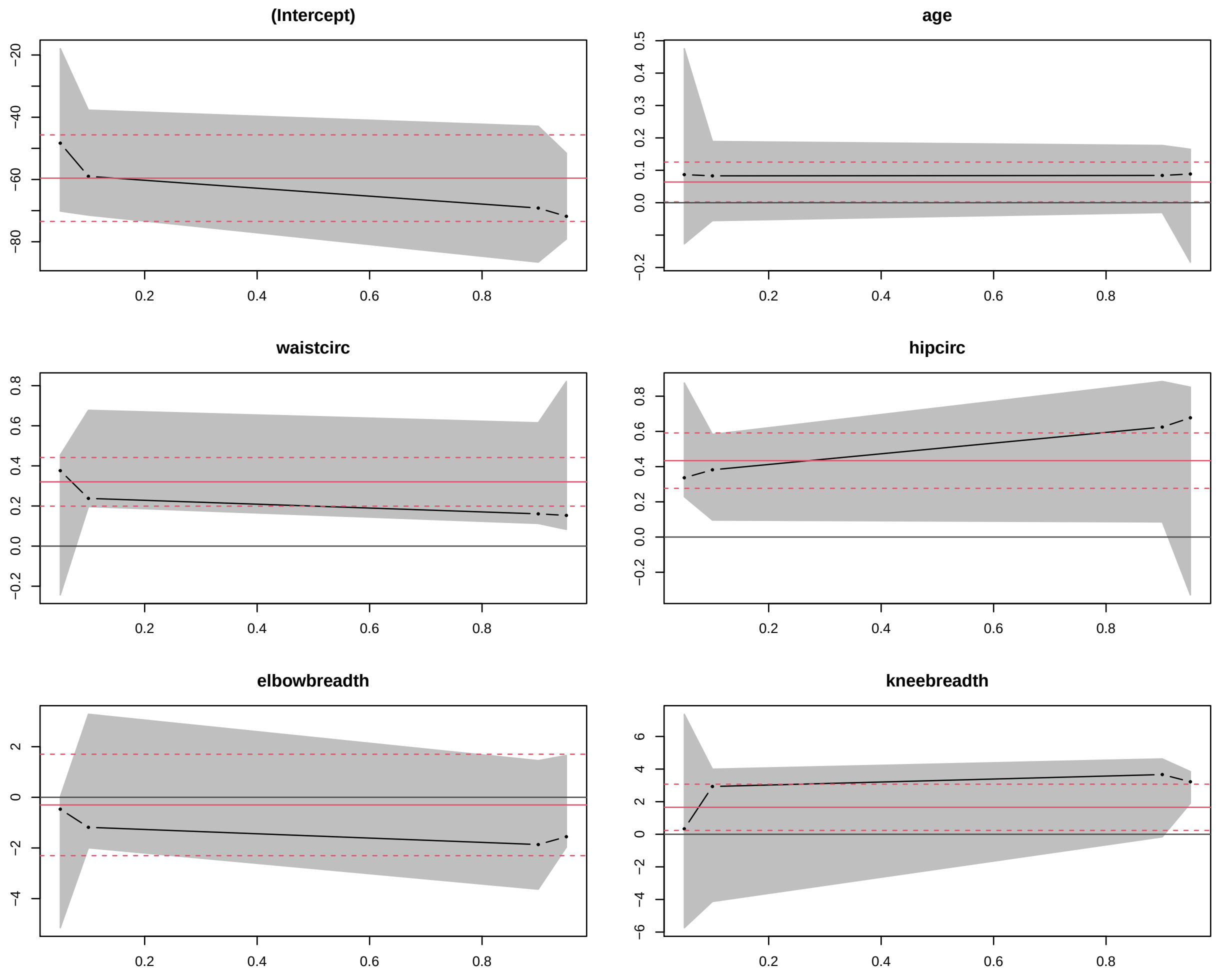
```
   Tree.qr.10 <-rq(Tree ,data=bodyfat,tau = 0.1)
   Tree.qr.90 <-rq(Tree ,data=bodyfat,tau = 0.9)
   Tree.qr.95 <-rq(Tree ,data=bodyfat,tau = 0.95)
```

```
##              X5th.Quantile X10th.Quantile X90th.Quantile X95th.Quantile
## (Intercept)   -48.33253829   -58.97738973   -69.18027856   -71.82885612
## age             0.08663026     0.08285144     0.08411451     0.08860546
## waistcirc       0.37613071     0.23786244     0.16076824     0.15295434
## hipcirc         0.33670446     0.38183871     0.62462417     0.67737884
## elbowbreadth   -0.46897330    -1.18452551    -1.86525827    -1.55603803
## kneebreadth     0.33454217     2.93043990     3.66332552     3.22144122
```

Analysis of variance is performed to verify if there is a significant difference in the models built at different quantiles. A p-value of 0.02316 indicate that the null hypothesis is rejected and the models are significantly different at 95% confidence interval, confirming the heteroscedasticity in the data.

```
## Quantile Regression Analysis of Deviance Table
##
## Model: DEXfat ~ age + waistcirc + hipcirc + elbowbreadth + kneebreadth
## Joint Test of Equality of Slopes: tau in {  0.05 0.1 0.9 0.95  }
##
##   Df Resid Df F value Pr(>F)
## 1 15      269  0.8974 0.5677
```

The plots show the relationship between the variables and the DEXfat at different Quantiles compared to the linear regression line which is at the center of each plot.For the variables, age and elbow width the Quantile regression line falls within the confidence intervel of the linear regression line indicating no significant difference in the observations between quantiles. However, for the variables waist circumference,hip circumference and knee width the quantile regression line is not parallel and are partly outside the confidence intervel of the linear regression line indicating a difference.



A Quantile regression line is fit foe each of the Quantiles, 5%,10%,90% and 95% with Dexfat as dependent variable and hipcirc as independent variable. Hip circumference is choosed to fit the model as it is clear from the initial analysis that this variable is significantly different at different quantiles compared to linear regression line.

```
BF <- DEXfat ~ hipcirc
BF.5<-rq(BF ,data=bodyfat,tau = 0.05)
BF.10<-rq(BF ,data=bodyfat,tau = 0.1)
BF.90<-rq(BF ,data=bodyfat,tau = 0.90)
BF.95<-rq(BF ,data=bodyfat,tau = 0.95)
```

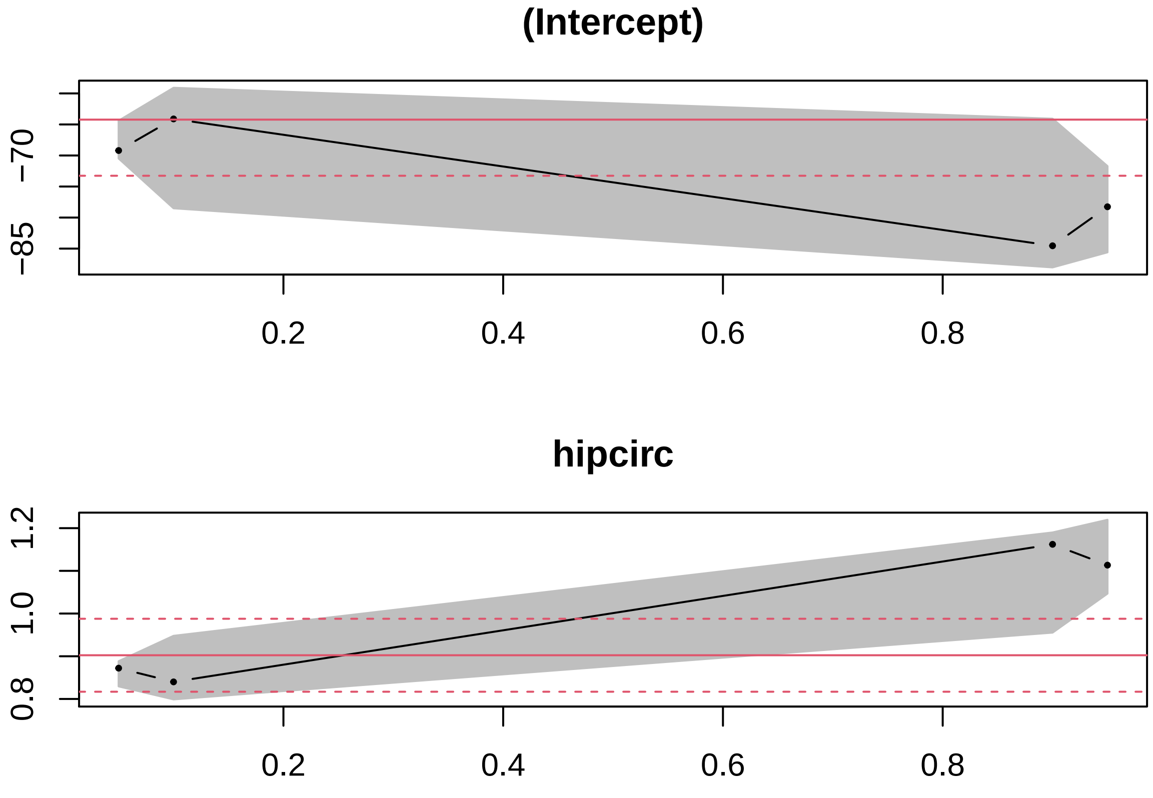
On comparision, there is difference is the effect of hip circumference on DEXfat at different quantiles. The coefficients are listed in the table. From 5th percentile to 95th percentile there is an increase in significance.However, there is no big difference between 5th and 10th percentile ,90th and 95th percentile. At 5th percentile with a unit increase in hip circumference there is 0.87 units increase in DEXfat.At 10th percentile with a unit increase in hip circumference there is 0.84 units increase in DEXfat.At 90th percentile with a unit increase in hip circumference there is 01.16 units increase in DEXfat.At 95th percentile with a unit increase in hip circumference there is 1.11 units increase in DEXfat.

```
##             X5th.Quantile X10th.Quantile X90th.Quantile X95th.Quantile
## (Intercept)   -69.1985930        -64.108     -84.545668     -78.260000
## hipcirc         0.8721106          0.840       1.162125       1.113333
```

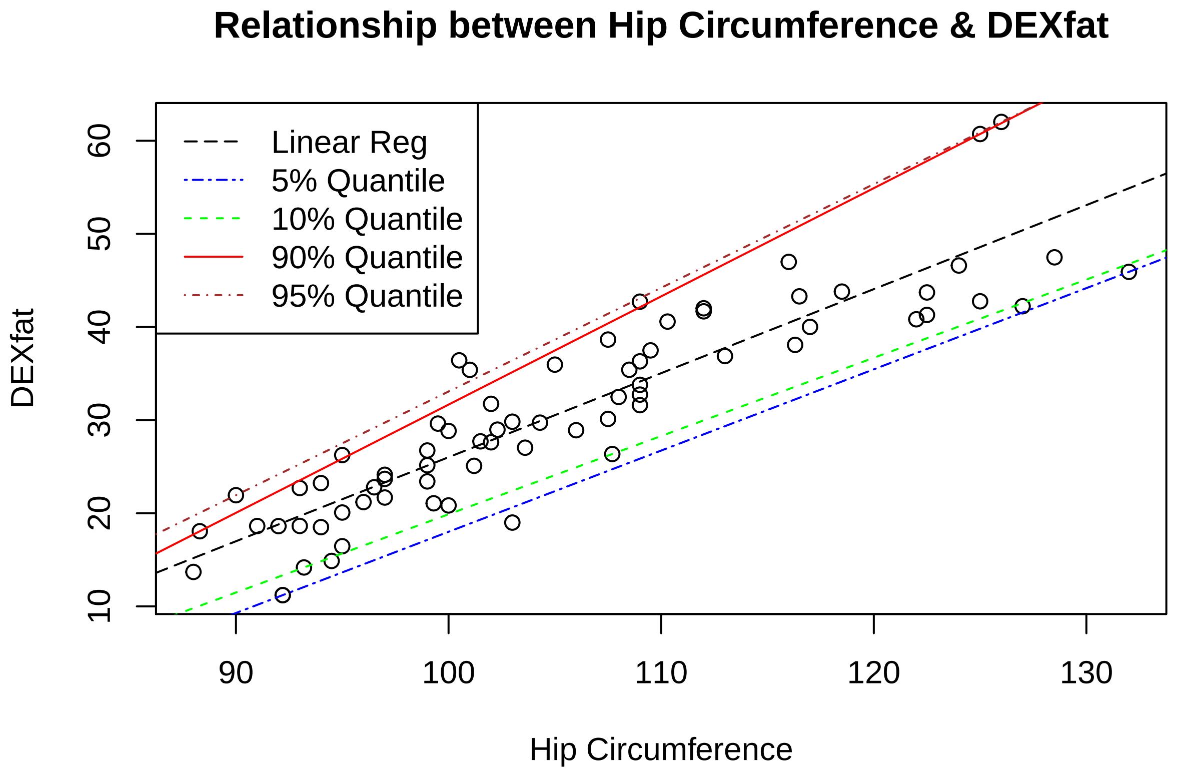
Analysis of variance is performed to verify if there is a significant difference in the models built at different quantiles. A p-value of 0.01169 indicate that the null hypothesis is rejected and the models are significantly different at 95% confidence interval, confirming the heteroscedasticity in the data.

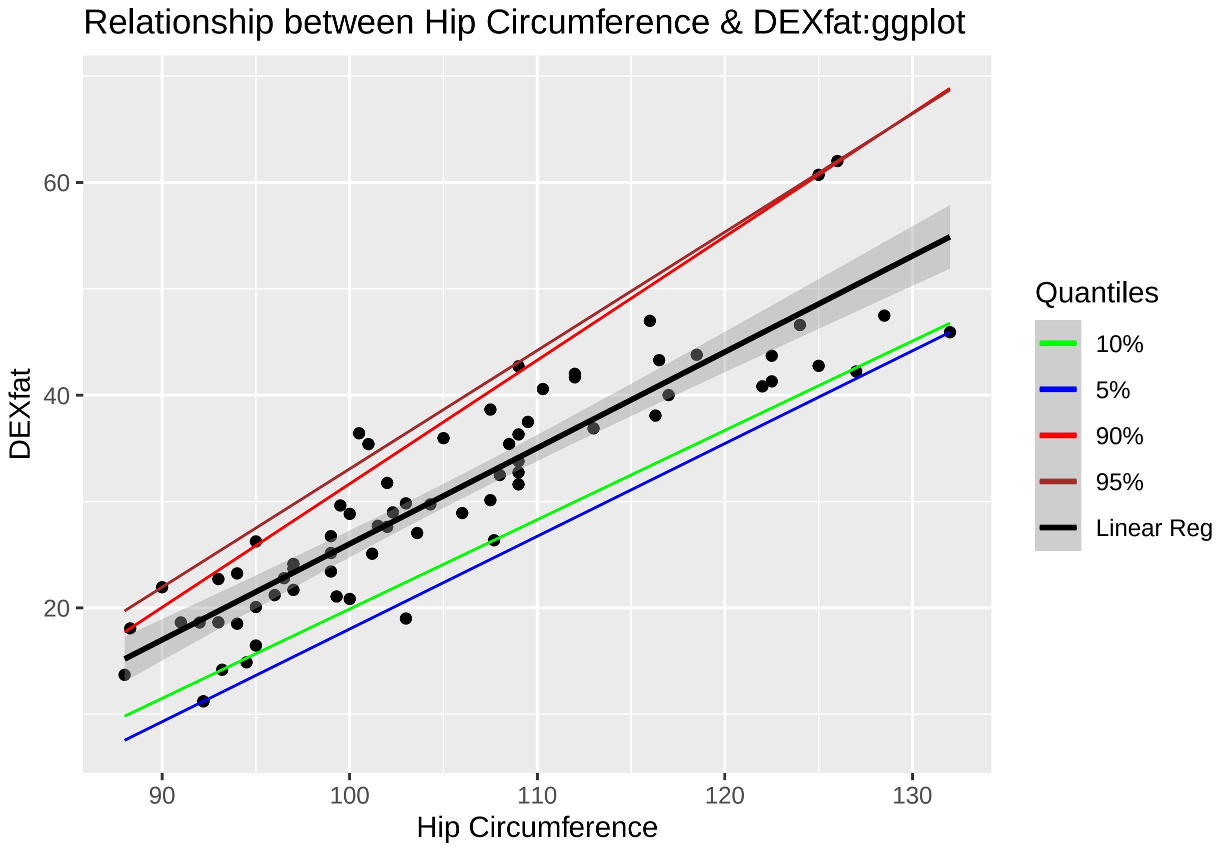
```
## Quantile Regression Analysis of Deviance Table
##
## Model: DEXfat ~ hipcirc
## Joint Test of Equality of Slopes: tau in {  0.05 0.1 0.9 0.95  }
##
##   Df Resid Df F value  Pr(>F)
## 1  3      281  3.7348 0.01169 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The plots show the relationship between the hip circumference and the DEXfat at different Quantiles compared to the linear regression line which is at the center of each plot.The quantile regression line is not parallel and are mostly outside the confidence intervel of the linear regression line indicating a significant difference. At the lower Quantiles the observations are lower than the mean and towards the higher Quantiles the they are higher than mean.



The linear regression lines from all the quantiles are plotted over the scatter plot of the Bodyfat data.From the plot it is clear that these lines are not parallel and have a different slope indicating heteroscedasticity of data. More information can be obtained from the Quantile regression than the linear regression. A similar plot is created using ggplot2.


Relationship between Hip Circumference & DEXfat
Linear Reg
5% Quantile
10% Quantile
90% Quantile
95% Quantile
DEXfat
10
20
30
40
50
60
90
100
110
120
130
Hip Circumference



CONCLUSION

Bodyfat data from the TH.data package is used to analyze the effect of anthropometric measurements on the DEXfat. A decision tree is built using Recursive Partitioning methodology. Waist circumference, Hip circumference and Knee breadth variables are found to have more significance on the DEXfat.A median regression line (at 50th percentile) is fit. The Coefficients indicate that the Waist circumference, Hip Circumference and Knee breadth variables are more significant at 50th percentile as well. With a unit increase of waist circumference,hip circumference and knee breadth there is an increase of 0.28,0.51 and 0.76 units of DEXfat respectively. Relationship between the variables and the DEXfat at 5%,10%,90% and 95% Quantiles is analysed. A difference is observed at different quantiles compared to the linear regression line. Similarly,a Quantile regression model is fit with DEXfat as dependent variable and hip circumference as independent variable at 5%,10%,90% and 95% quantiles. From the coefficients, it is clear that hip circumference has different effect at different quantiles. From 5th percentile to 95th percentile there is an increase in significance.However, there is no big difference between 5th and 10th percentile & 90th and 95th percentile.A plot is created showing the relationship between Hip circumference and DEXfat.The Quantile regression lines are not parallel to the linear regresion line indicating that there is a difference in slope and more information can be obtained from these quantiles.

REFERENCES

- Lecture Code
- **Quantile Regression Documentation** from tidyverse.org (https://ggplot2.tidyverse.org/reference/geom_quantile.html)
- Clay Ford,*Getting Started with Quantile Regression*, September 20,2015 (https://data.library.virginia.edu/getting-started-with-quantile-regression/)

- Lecture by Scott Burk,*R50 Quantile Regresiion in R. Robust,nonparametric,regression*, April 28,2019 (https://www.youtube.com/watch?v=NjfJfpC1PUA)