Quantile Regression

Mohamed Ahmed

Note: Amin Baabol and I collaborated in producing this report. We worked together and verified each other's work on every part of this assignment.

Introduction:

The purpose of this analyses is to apply median and linear regression analysis to clouds data.Our goal is to to fit the models and compare them. The linear regression model is from chapter 6.

Data and Model

The data that was used for this analysis is clouds data from the HSAUR3 package. The data contains information such as each seeding, time, and sne.These variables will be used to conduct our analysis. The table below lists all the variables, form the HSAUR3 data set, along with their description.

Symbol	Description
$seeding$	a factor indicating whether seeding action occurred (no or yes).
$Time$	number of days after the first day of the experiment.
sne	suitability criterion.
$cloudcover$	the percentage cloud cover in the experimental area, measured using radar.
$prewetness$	the total rainfall in the target area one hour before seeding (in cubic meters times 1e+8).
$echomotion$	a factor showing whether the radar echo was moving or stationary.
$rainfall$	the amount of rain in cubic meters times 1e+8.

A median regression model will be built using the rq function and another linear regression model will be built, for the same data, then results will be compared to choose which analysis is more suitable for the Clouds data.

Results

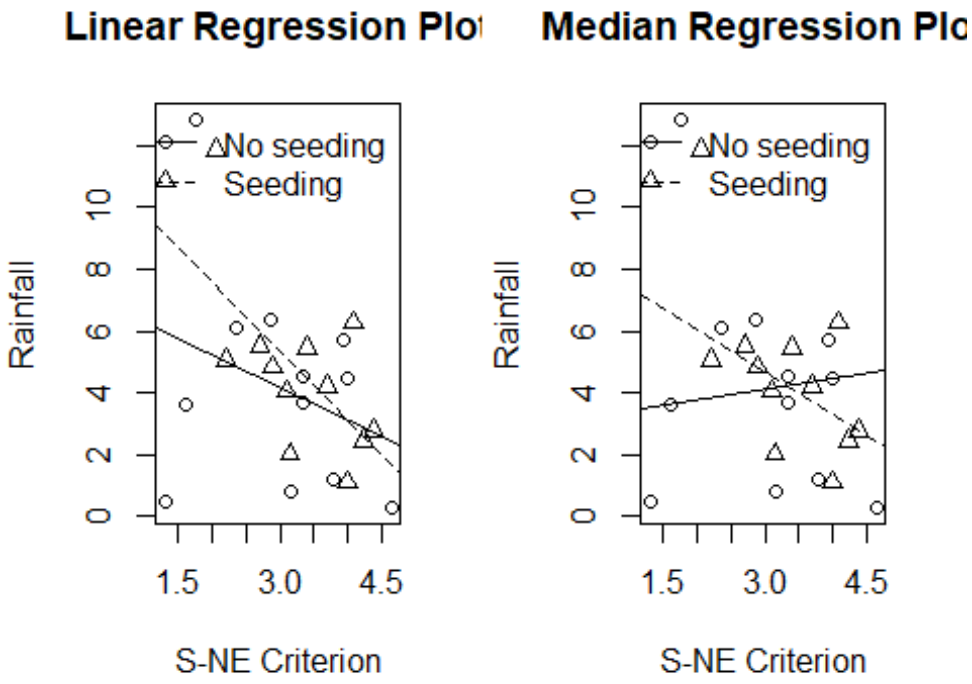
Apply a median regression analysis on the **clouds** data. Compare this to the linear regression model from Chapter 6. Write up a formal summary of the two analyses and provide a justified recommendation on which analysis the researcher should be using.

First, we built a linear a regression model using all explanatory variables. We can see that there is few predictors that are significant with P-Value < 0.5. Those significant predictors

are seeding yes & seeding yes:sne. This also means that seeding along with high or low S-Ne values affects rainfall.The linear model shows that seeding and and S-Ne criterion are the most influential explanatory variables on rainfall. We built a median regression model using the explanatory variables that were found to be significant in the linear regression model. The model's intercept is 8.86 and the intercept's confidence intervals values are (3.14768 14.86666). Also, S-Ne coefficient value is -1.38667 and its confidence intervals values are (-2.46926 0.13118).

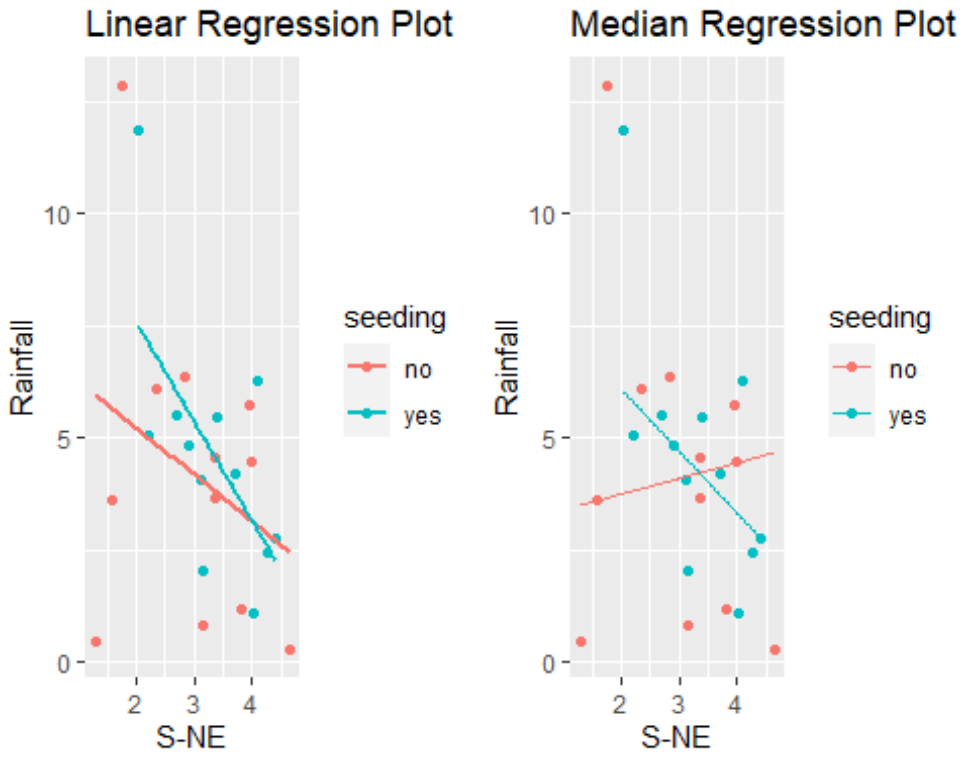
```
##   seeding time  sne cloudcover prewetness echomotion rainfall
## 1      no    0 1.75       13.4      0.274 stationary    12.85
## 2     yes    1 2.70       37.9      1.267     moving     5.52
## 3     yes    3 4.10        3.9      0.198 stationary     6.29
## 4      no    4 2.35        5.3      0.526     moving     6.11
## 5     yes    6 4.25        7.1      0.250     moving     2.45
## 6      no    9 1.60        6.9      0.018 stationary     3.61

## Linear Regression Model with all explantory variables

##
## Call:
## lm(formula = clouds_formula, data = clouds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5259 -1.1486 -0.2704  1.0401  4.3913
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -0.34624    2.78773  -0.124  0.90306
## seedingyes                     15.68293    4.44627   3.527  0.00372 **
## time                           -0.04497    0.02505  -1.795  0.09590 .
## seedingno:sne                   0.41981    0.84453   0.497  0.62742
## seedingyes:sne                 -2.77738    0.92837  -2.992  0.01040 *
## seedingno:cloudcover            0.38786    0.21786   1.780  0.09839 .
## seedingyes:cloudcover          -0.09839    0.11029  -0.892  0.38854
## seedingno:prewetness            4.10834    3.60101   1.141  0.27450
## seedingyes:prewetness           1.55127    2.69287   0.576  0.57441
## seedingno:echomotionstationary  3.15281    1.93253   1.631  0.12677
## seedingyes:echomotionstationary 2.59060    1.81726   1.426  0.17757
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.205 on 13 degrees of freedom
## Multiple R-squared:  0.7158, Adjusted R-squared:  0.4972
## F-statistic: 3.274 on 10 and 13 DF,  p-value: 0.02431

## Linear Regression Model with S-Ne as a predictor

##
## Call:
```

```
## lm(formula = rainfall ~ sne, data = clouds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4927 -2.1116  0.0556  1.2295  6.5036
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.7430     2.1508   4.065 0.000515 ***
## sne          -1.3695     0.6524  -2.099 0.047512 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.902 on 22 degrees of freedom
## Multiple R-squared:  0.1669, Adjusted R-squared:  0.129
## F-statistic: 4.406 on 1 and 22 DF,  p-value: 0.04751

## Median Regression Model

##
## Call: rq(formula = rainfall ~ sne, tau = 0.5, data = clouds)
##
## tau: [1] 0.5
##
## Coefficients:
##             coefficients lower bd upper bd
## (Intercept)  8.86133      3.14768 14.86666
## sne         -1.38667     -2.46926  0.13118
```



The linear regression plot shows that smaller S-Ne values along with seeding causes more rainfall than when seeding is not implemented. However, When we have high S-Ne values, the plot shows that we get less rainfall. The point where where both lines intersect is S-Ne of Four which indicates that rainfall can be maximized by applying seeding with S-Ne values lower than four. We can see that seeding is a better fit than when seeding is not used. Overall, rainfall with seeding and without seeding has a negative slope indicating that the overall trend is lower rainfall amount with higher SN-e values.

The median regression Plot shows that that lower S-Ne when seeding is implemented cause more rainfall than when seeding is implemented. Even with high S-Ne values and no seeding, rainfall levels are lower than rainfall levels when seeding is used. Unlike the linear regression, the median regression shows a positive slope when seeding is not implemented. When seeding is not implemented and negative slope when seeding is implemented.



Conclusion

Median regression is sometimes performs better than linear regression because it is robust to outliers. linear regression estimates the conditional mean function as a linear combination of the predictors where median regression estimates these conditional median function as a linear combination of the predictors. In the linear regression model, It looks like that no seeding line is affected by outliers and, in general, there is high variability in the data, therefore we conclude that median regression is more suitable for the data.

Title

"Understanding Bodyfat Using Predictive Regressions"

Introduction

The goal of this assignment is to understand the various factors that influence the bodyfat.In order to accomplish this task we will reanalyze and compare the **bodyfat** data from the **TH.data** package using two different predictive regression methods. Throughout the duration of this analysis we will: a) Compare the regression tree approach from chapter 9 of the textbook to median regression and summarize the different findings. b) Choose one independent 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 independent variable and plot the lines from the models on the graph.

Methodology

Given the complexity of the explanatory terms, it is clear there is no direct linear relationship between the dependent variable DEXfat and the other independent variables. Hence, our approach to better understand what factors influence response variable is to utilize one of the non-parametric methods. Specifically, we would like to employ decision tree regression method to reanalyze the response variable. Normally we would split the original dataset into two subsets. One for training the model and the other for testing the accuracy of the model. However, in order to save time we will not be testing the accuracy of the model, instead we will prune the decision tree model using the original data and this should fairly be enough to prevent the model from over-fitting.Furthermore, it is fair to point out that this decision tree model is built using binary recursive partitioning using **rpart()** in the **rpart**. This iterative process will iteratively split the data into branches while minimizing the sum of squared deviation from the mean.This process will continue until it each node finally gets to the minimum split and the node becomes a terminal.This method should systematically weed out the unnecessary explanatory variables.

Next, we set up a median regression model to reanalyze the same response variables from the same datset. Median regression is a quantile regression where tau is set to 0.50. We will build this median regression model using the **rq()** function in in the **quantreg** package. This particular method will attempt to estimate the quantile function of the response variable "DEXfat" as a linear combination of the explanatory variables.The conditional quantile we will impose on this quantile regression is 0.50. It is worth noting that this median regression is exceptionally robust to outliers since it is the 50th quantile regression. We will not be making assumptions for this median quantile regression because we are treating this model like the non-linear, non-parametric it is. Making linear quantile regression assumptions defeats the purpose of the median regression. We are not interested in finding the mean, we are interested in estimating how much the explanatory variables explain the dependent variable "DEXfat" on a specified quantile which is the 50 percentile.

Furthermore,upon successfully constructing and running the models we will compare the decision tree regression model and the median quantile regression model. There are two things we will ultimately compare, the fitted versus the observed plots and the mean squared erros of each model. The implication here is that the fitter model will have more accurate plots and lower mean squared error value. Having done this part, we will move one to the last section. This section we will construct a linear quantile regression models for the 5%, 10%, 90% and 95% and then compare there plots with interpretations.

Results

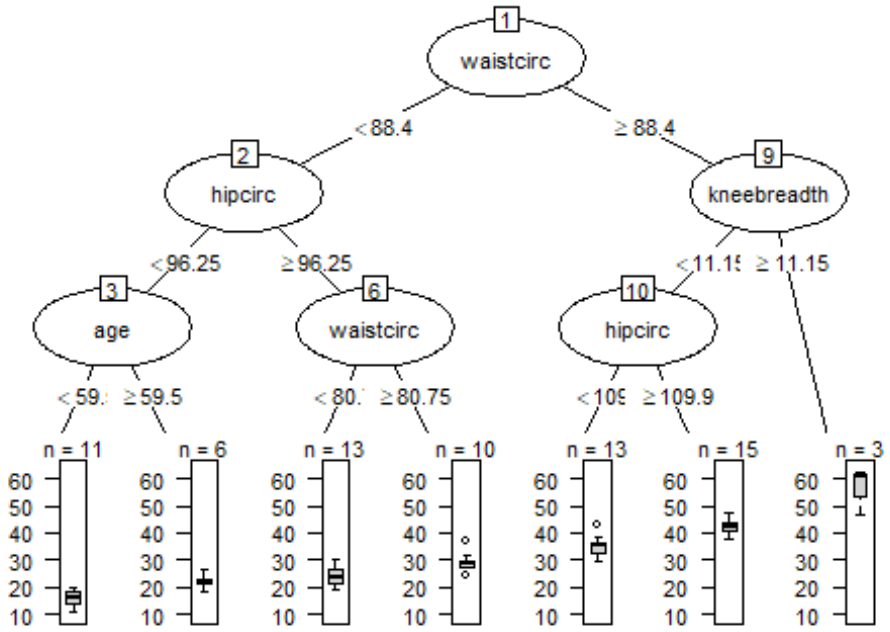
2. Reanalyze the **bodyfat** data from the **TH.data** package.

a) Compare the regression tree approach from chapter 9 of the textbook to median regression and summarize the different findings.

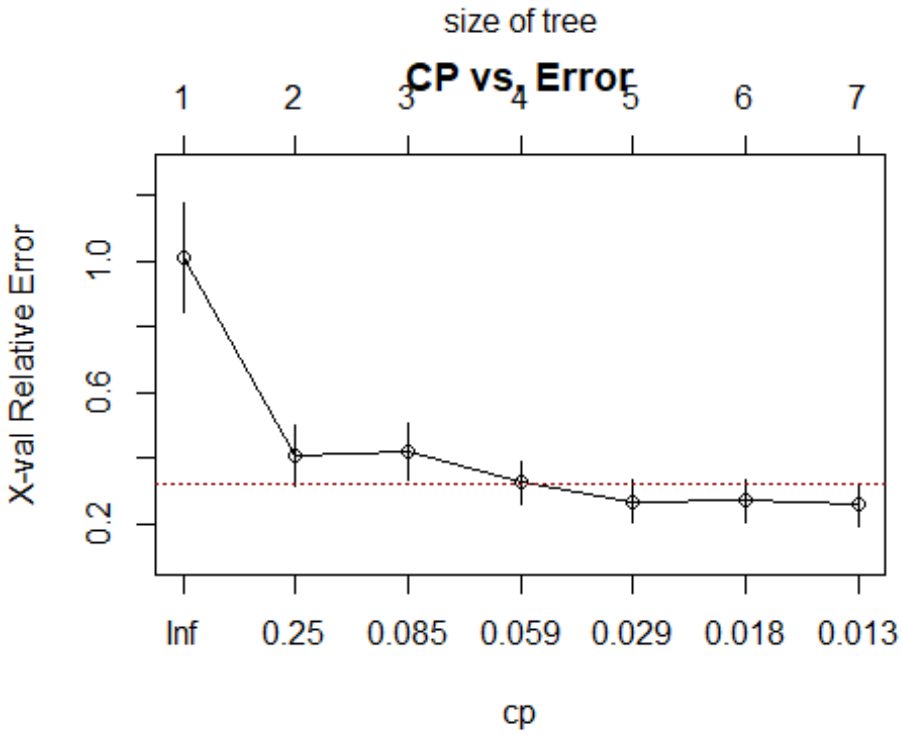
We began the analysis by fitting a decision tree model, we then printed the cp table to check whether it pruning might be necessary or not. According to the cp table an nplit of 7 will yield the lowest error of 0.2574097. We then built a pruned decision tree model using

the **prune()** function. As indicated by figure 2.1a and figure 2.1b there is no need for pruning because both the pruned model and the original decision tree model have the same splits. According to figure 2.2a and 2.2b the pruned decision tree model shows a stepwise plot where as the median regression model shows a fairly linear plot which is not is not entirely unexpected. It went further and plotted the residuals of both models to check their distributions. Figure 2.3a and figure 2.3b both have residuals that are randomly distributed in and around 0 with a margin of error though the median regression model's residuals have wider margin.Finally, we computed the mean squared error of both model, it turns out the pruned decision tree model has mse of 10.1705 whereas the median regression model has mse of 15.0245. Taking all those outputs mentioned above the pruned decision tree will yield the better results in the context of mean squared error.

Figure 2.1a: Non-Pruned Decision Tree Model




```
## 7
## 7

##           CP nsplit  rel error    xerror       xstd
## 1 0.66289544      0 1.00000000 1.0102608 0.16685325
## 2 0.09376252      1 0.33710456 0.4078290 0.09383664
## 3 0.07703606      2 0.24334204 0.4206093 0.08483119
## 4 0.04507506      3 0.16630598 0.3274332 0.06437893
## 5 0.01844561      4 0.12123092 0.2691145 0.06244580
## 6 0.01818982      5 0.10278532 0.2699932 0.06254900
## 7 0.01000000      6 0.08459549 0.2574097 0.06268795
```



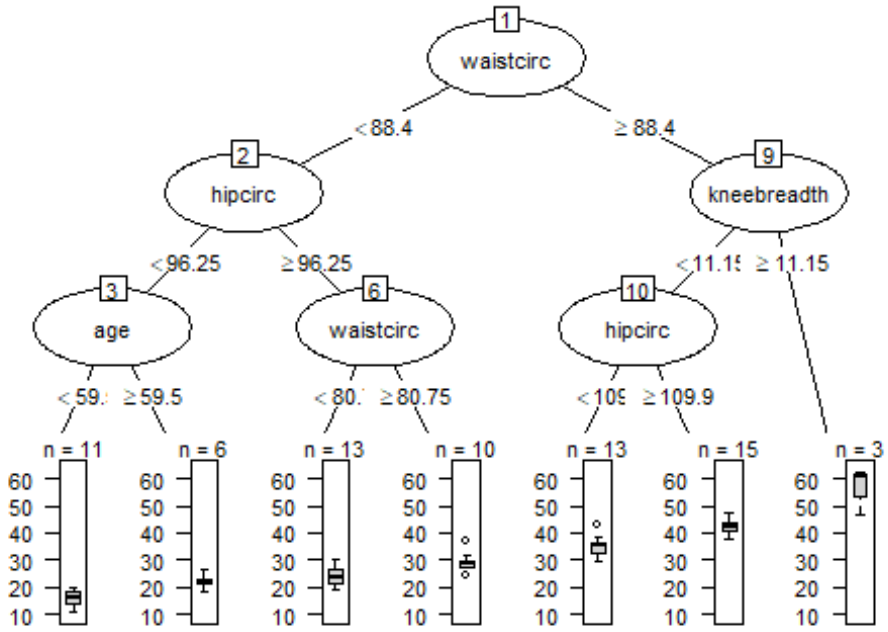

```
## Call:
## rpart(formula = DEXfat ~ age + waistcirc + hipcirc + elbowbreadth +
##     kneebreadth, data = bodyfat, control = rpart.control(minsplit = 10))
##   n= 71
##
##           CP nsplit  rel error    xerror       xstd
## 1 0.66289544      0 1.00000000 1.0102608 0.16685325
## 2 0.09376252      1 0.33710456 0.4078290 0.09383664
## 3 0.07703606      2 0.24334204 0.4206093 0.08483119
## 4 0.04507506      3 0.16630598 0.3274332 0.06437893
## 5 0.01844561      4 0.12123092 0.2691145 0.06244580
## 6 0.01818982      5 0.10278532 0.2699932 0.06254900
## 7 0.01000000      6 0.08459549 0.2574097 0.06268795
##
## Variable importance
##    waistcirc      hipcirc  kneebreadth elbowbreadth          age
##           34           30           24            7            4
##
## Node number 1: 71 observations,    complexity param=0.6628954
##   mean=30.78282, MSE=120.2251
##   left son=2 (40 obs) right son=3 (31 obs)
##   Primary splits:
##       waistcirc    < 88.4   to the left,  improve=0.6628954, (0 missing)
##       hipcirc      < 108.25 to the left,  improve=0.6254333, (0 missing)
##       kneebreadth  < 9.35   to the left,  improve=0.5142133, (0 missing)
##       age          < 40.5   to the left,  improve=0.1570344, (0 missing)
##       elbowbreadth < 6.55   to the left,  improve=0.1169918, (0 missing)
```

```
##   Surrogate splits:
##       hipcirc      < 107.85 to the left,  agree=0.915, adj=0.806, (0
split)
##       kneebreadth  < 9.35   to the left,  agree=0.831, adj=0.613, (0
split)
##       elbowbreadth < 6.55   to the left,  agree=0.648, adj=0.194, (0
split)
##       age          < 47     to the left,  agree=0.592, adj=0.065, (0
split)
##
## Node number 2: 40 observations,    complexity param=0.07703606
##   mean=22.92375, MSE=32.88394
##   left son=4 (17 obs) right son=5 (23 obs)
##   Primary splits:
##       hipcirc      < 96.25  to the left,  improve=0.4999238, (0 missing)
##       waistcirc    < 71.5   to the left,  improve=0.4408508, (0 missing)
##       kneebreadth  < 9.15   to the left,  improve=0.3123752, (0 missing)
##       age          < 41     to the left,  improve=0.2212005, (0 missing)
##       elbowbreadth < 6.65   to the left,  improve=0.0757275, (0 missing)
##   Surrogate splits:
##       waistcirc    < 71.5   to the left,  agree=0.775, adj=0.471, (0
split)
##       age          < 41     to the left,  agree=0.700, adj=0.294, (0
split)
##       kneebreadth  < 8.25   to the left,  agree=0.675, adj=0.235, (0
split)
##       elbowbreadth < 5.75   to the left,  agree=0.600, adj=0.059, (0
split)
##
## Node number 3: 31 observations,    complexity param=0.09376252
##   mean=40.92355, MSE=50.39231
##   left son=6 (28 obs) right son=7 (3 obs)
##   Primary splits:
##       kneebreadth  < 11.15  to the left,  improve=0.51233840, (0 missing)
##       hipcirc      < 109.9  to the left,  improve=0.45671770, (0 missing)
##       waistcirc    < 106    to the left,  improve=0.44843720, (0 missing)
##       elbowbreadth < 6.35   to the left,  improve=0.16017880, (0 missing)
##       age          < 45.5   to the right, improve=0.06131694, (0 missing)
##
## Node number 4: 17 observations,    complexity param=0.01844561
##   mean=18.20765, MSE=16.81845
##   left son=8 (11 obs) right son=9 (6 obs)
##   Primary splits:
##       age          < 59.5   to the left,  improve=0.55069560, (0 missing)
##       waistcirc    < 70.35  to the left,  improve=0.39973880, (0 missing)
##       elbowbreadth < 6.65   to the left,  improve=0.22215850, (0 missing)
##       hipcirc      < 92.6   to the left,  improve=0.16823720, (0 missing)
##       kneebreadth  < 8.55   to the left,  improve=0.08112073, (0 missing)
##   Surrogate splits:
##       elbowbreadth < 6.55   to the left,  agree=0.824, adj=0.500, (0
```

```
split)
##       waistcirc    < 71.5   to the left,  agree=0.765, adj=0.333, (0
split)
##
## Node number 5: 23 observations,    complexity param=0.01818982
##   mean=26.40957, MSE=16.16806
##   left son=10 (13 obs) right son=11 (10 obs)
##   Primary splits:
##       waistcirc    < 80.75  to the left,  improve=0.41753840, (0 missing)
##       hipcirc      < 101.35 to the left,  improve=0.34272770, (0 missing)
##       kneebreadth  < 9.5    to the left,  improve=0.30544320, (0 missing)
##       elbowbreadth < 7.1    to the right, improve=0.06644785, (0 missing)
##       age          < 57     to the right, improve=0.03572739, (0 missing)
##   Surrogate splits:
##       hipcirc      < 101.75 to the left,  agree=0.783, adj=0.5, (0 split)
##       kneebreadth  < 9.5    to the left,  agree=0.696, adj=0.3, (0 split)
##       age          < 66     to the left,  agree=0.652, adj=0.2, (0 split)
##       elbowbreadth < 6.25   to the left,  agree=0.652, adj=0.2, (0 split)
##
## Node number 6: 28 observations,    complexity param=0.04507506
##   mean=39.26036, MSE=21.98307
##   left son=12 (13 obs) right son=13 (15 obs)
##   Primary splits:
##       hipcirc      < 109.9  to the left,  improve=0.62509140, (0 missing)
##       waistcirc    < 99     to the left,  improve=0.47879840, (0 missing)
##       kneebreadth  < 9.85   to the left,  improve=0.28389460, (0 missing)
##       elbowbreadth < 6.35   to the left,  improve=0.18101920, (0 missing)
##       age          < 49.5   to the right, improve=0.04758482, (0 missing)
##   Surrogate splits:
##       waistcirc    < 99     to the left,  agree=0.821, adj=0.615, (0
split)
##       elbowbreadth < 6.45   to the left,  agree=0.714, adj=0.385, (0
split)
##       kneebreadth  < 9.95   to the left,  agree=0.714, adj=0.385, (0
split)
##       age          < 49.5   to the right, agree=0.607, adj=0.154, (0
split)
##
## Node number 7: 3 observations
##   mean=56.44667, MSE=48.76009
##
## Node number 8: 11 observations
##   mean=15.96, MSE=8.818582
##
## Node number 9: 6 observations
##   mean=22.32833, MSE=5.242981
##
## Node number 10: 13 observations
##   mean=24.13077, MSE=9.046699
##
```

```
## Node number 11: 10 observations
##   mean=29.372, MSE=9.899016
##
## Node number 12: 13 observations
##   mean=35.27846, MSE=10.48431
##
## Node number 13: 15 observations
##   mean=42.71133, MSE=6.297998
```

Figure 2.1b:Pruned Decison Tree Model




```
## Call:
## rq(formula = DEXfat ~ age + waistcirc + hipcirc + elbowbreadth +
##     kneebreadth, tau = 0.5, data = bodyfat)
##
## Coefficients:
##  (Intercept)          age    waistcirc      hipcirc elbowbreadth
kneebreadth
## -57.30031520   0.06839443   0.28332466   0.51073243  -0.11982312
0.76452936
##
## Degrees of freedom: 71 total; 65 residual

##
## Call: rq(formula = DEXfat ~ age + waistcirc + hipcirc + elbowbreadth +
##     kneebreadth, tau = 0.5, data = bodyfat)
##
## tau: [1] 0.5
##
```

```
## Coefficients:
##              coefficients lower bd  upper bd
## (Intercept)  -57.30032    -87.22119 -36.39320
## age            0.06839     -0.04338   0.14943
## waistcirc      0.28332      0.07991   0.48638
## hipcirc        0.51073      0.21307   0.75030
## elbowbreadth  -0.11982     -3.62882   2.18220
## kneebreadth    0.76453     -2.30145   2.33329
```

gure 2.2a:Pruned Regression Tree Predicted vs. Obs

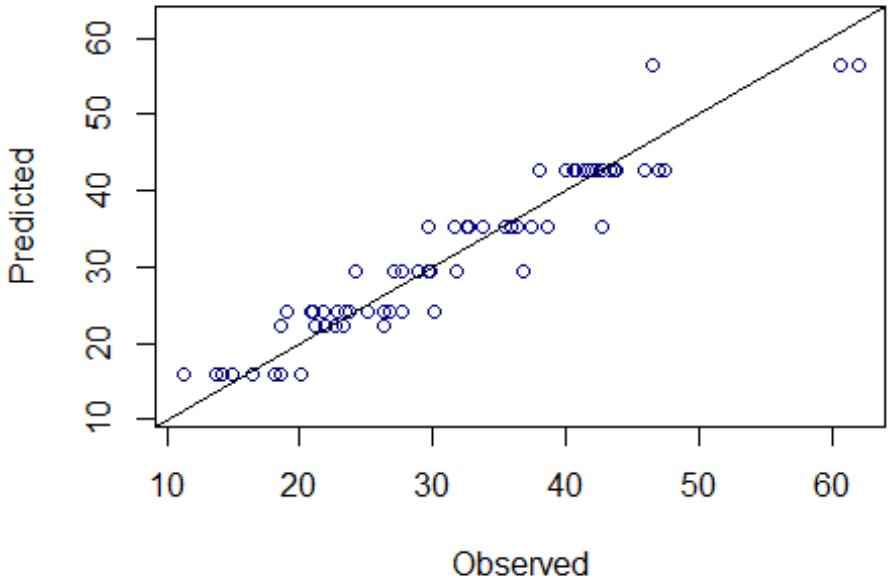


Figure 2.2b:Median Regression Predicted vs. Obser

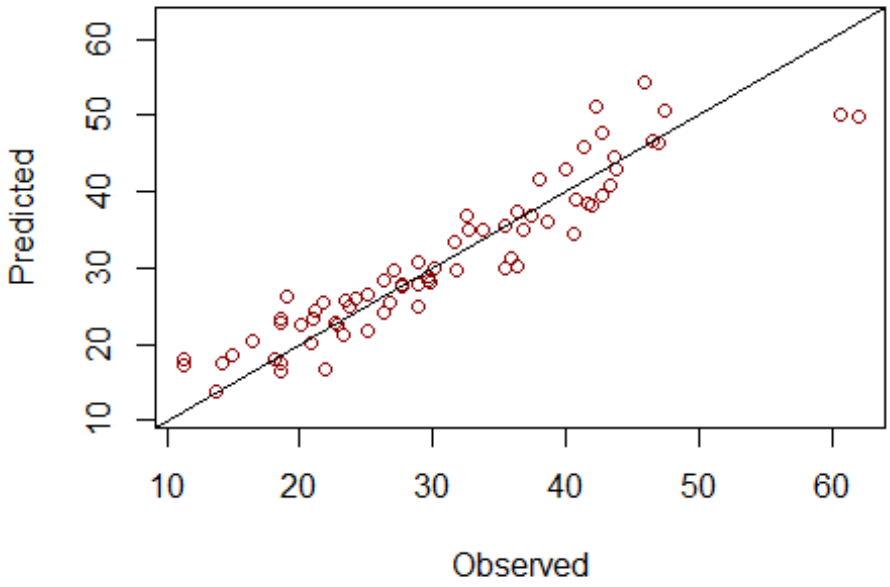


Figure 2.3a:Pruned Regression Tree Model Residu

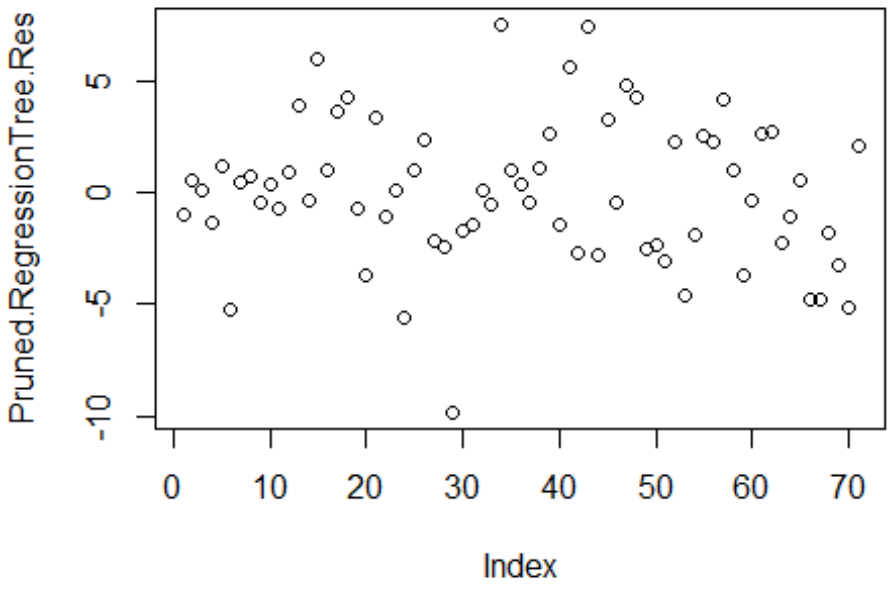
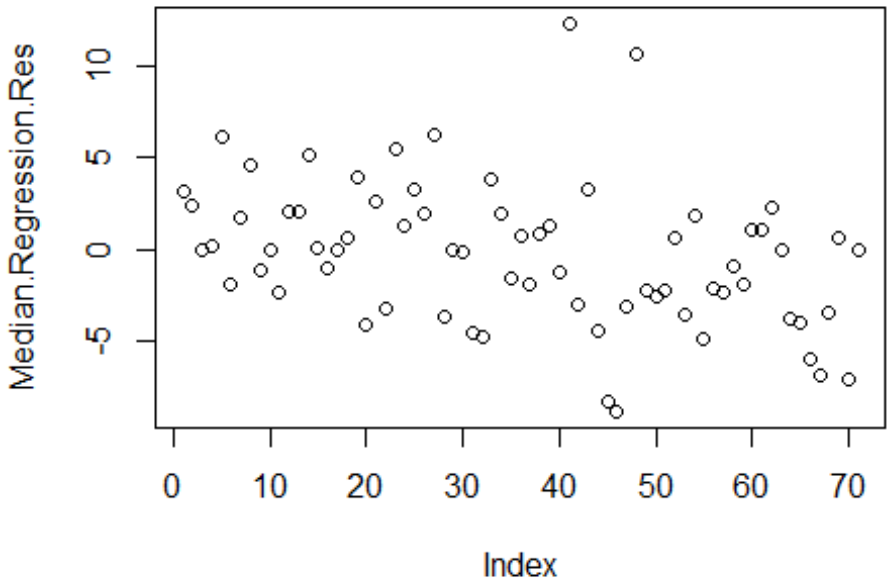


Figure 2.3b:Median Regression Model Residuals




```
##    Actual Predicted.Prune   SSE.Prune
## 47  41.68        42.71133  1.06364844
## 48  43.29        42.71133  0.33485511
## 49  35.41        35.27846  0.01730237
```

```
## 50  22.79        24.13077  1.79766213
## 51  36.42        35.27846  1.30311006
## 52  24.13        29.37200 27.47856400

##    Actual Predicted.Median   SSE.Median
## 47  41.68         38.46850 1.031375e+01
## 48  43.29         40.86192 5.895595e+00
## 49  35.41         35.41000 2.019484e-28
## 50  22.79         22.65437 1.839667e-02
## 51  36.42         30.28407 3.764968e+01
## 52  24.13         26.01941 3.569862e+00

##   Regression Tree Model MSE Median Regression Model MSE
## 1                   10.1705                      15.0245
```

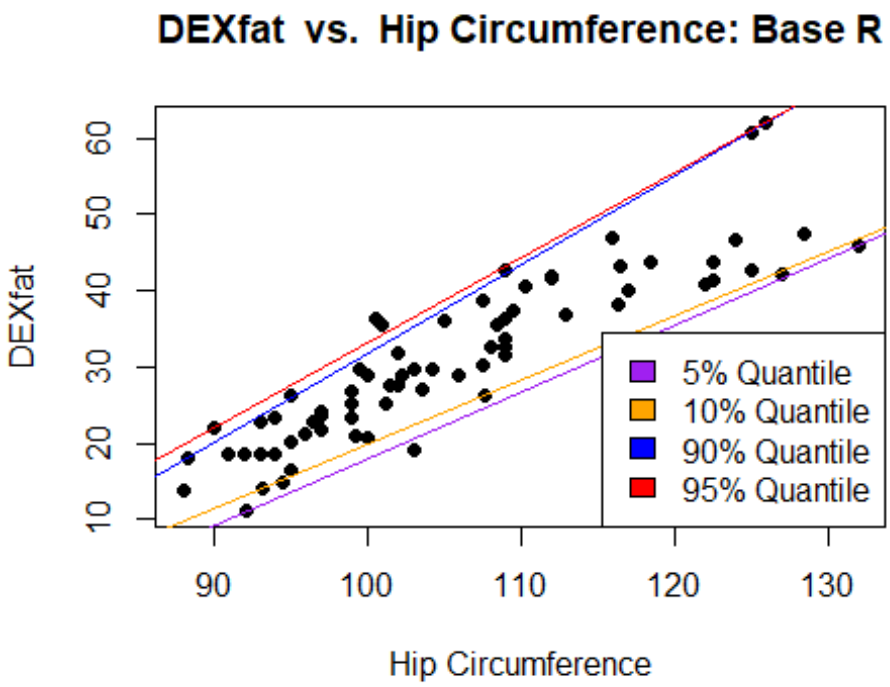
b) Choose one independent variable. For the relationship between this variable and DEXfat, create linear regression models for the 5%, 10%, 90%, and 95% quantiles. Plot DEXfat vs that independent variable and plot the lines from the models on the graph.

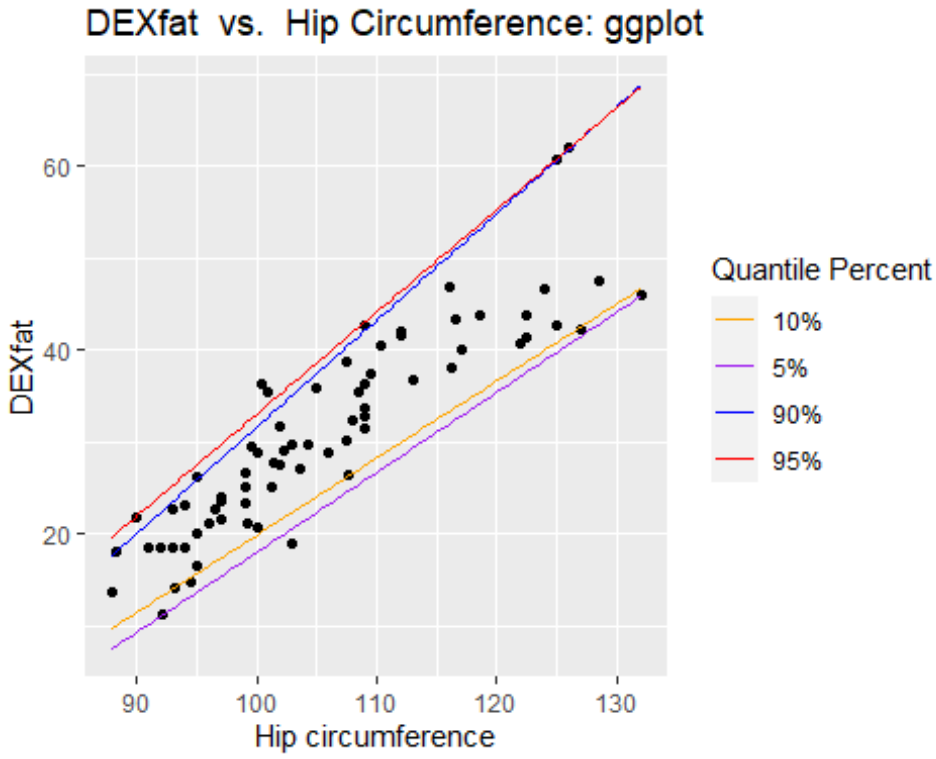
We printed the summary of the median regression model along with the p-values of each independent variable to see which ones are important. According to the model summary there are only three parameters that are significant at the 0.05 level. Those parameters include the intercept, waist circumference and hip circumference. ignoring the intercept we see that hip circumference has the lowest p-value of 0.00003 and the largest t-value of 4.45177 which we presumed to be highly important independent variable.We could have just easily have chose waist circumference too but we decided to with hip circumference the reasons mentioned above.

We plotted the linear regression model to understand the relationship between our independent variable hip circumference and the response variable DEXfat. Additionally, we added the quantile regression lines for the 5%, 10%,90% and 95%. After carefully studying the final plots, it is interesting to note that the intensity of the coefficient of the explanatory variable changes with different quantile values. At the 5% the slope of the independent variable is 0.8721106 compared to when tau is 0.10 the slope drops to 0.840, then picks back up to 1.162125 at the 90% and then slightly drops down again to 1.113333 at the 95%. The plots of the regression lines also seem to be supporting this. The 95% indicated by the red line seems to be much steeper than the 5% indicated by the purple line. The variance of the residual does not stay constant due to the heteroscedasticity.

```
##
## Call: rq(formula = DEXfat ~ age + waistcirc + hipcirc + elbowbreadth +
##     kneebreadth, tau = 0.5, data = bodyfat)
##
## tau: [1] 0.5
##
## Coefficients:
##              Value     Std. Error t value   Pr(>|t|)
## (Intercept)  -57.30032   9.88496   -5.79672   0.00000
## age            0.06839   0.04460    1.53352   0.13000
```

```
## waistcirc      0.28332   0.09363    3.02586   0.00355
## hipcirc        0.51073   0.11473    4.45177   0.00003
## elbowbreadth  -0.11982   1.25874   -0.09519   0.92445
## kneebreadth    0.76453   0.92976    0.82229   0.41392

##               tau= 0.05 tau= 0.10  tau= 0.90  tau= 0.95
## (Intercept) -69.1985930   -64.108 -84.545668 -78.260000
## hipcirc       0.8721106     0.840   1.162125   1.113333
```





Conclusion

The purpose of our report was to properly comprehend in a meaningful way the relationship between bodyfat and a host of other variables contained the in the **bodyfat** in

TH.data package. In effort to accurately accomplish this goal we also compared decision tree model and a median quantile regression model to understand how these two models perform in establishing a relationship between the response variable “DEXfat” and the independent variables. Although in the end decision tree model had the lower mse of 10.1705 while the median regression has mse of 15.0245,however, each model has it is strengths and weaknesses.

Works Cited

1.Michael, Semhar, and Christopher P. Saunders. “Survival Analysis Introduction” Chapter 12. 30 Oct. 2020, South Dakota State University, South Dakota State University. 2.Everitt, Brian, and Torsten Hothorn. A Handbook of Statistical Analyses Using n SECOND EDITION. Taylor and Francis Group LLC, 2010. 3.Neupane, Achal.“Survival Analysis” Achal Neupane,30 Oct.2019,achalneupane.github.io/achalneupane.github.io/post/quantile_regression/