Multiple Linear Regression from HSAUR

Yamuna Dhungana

Exercises

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.

```
##
## 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
```

```
## The seedingyes variable is the most significant variable in the model followed by seedingyes:sne
```

```
## Now we choose continous variable sne to fit our linear model
```

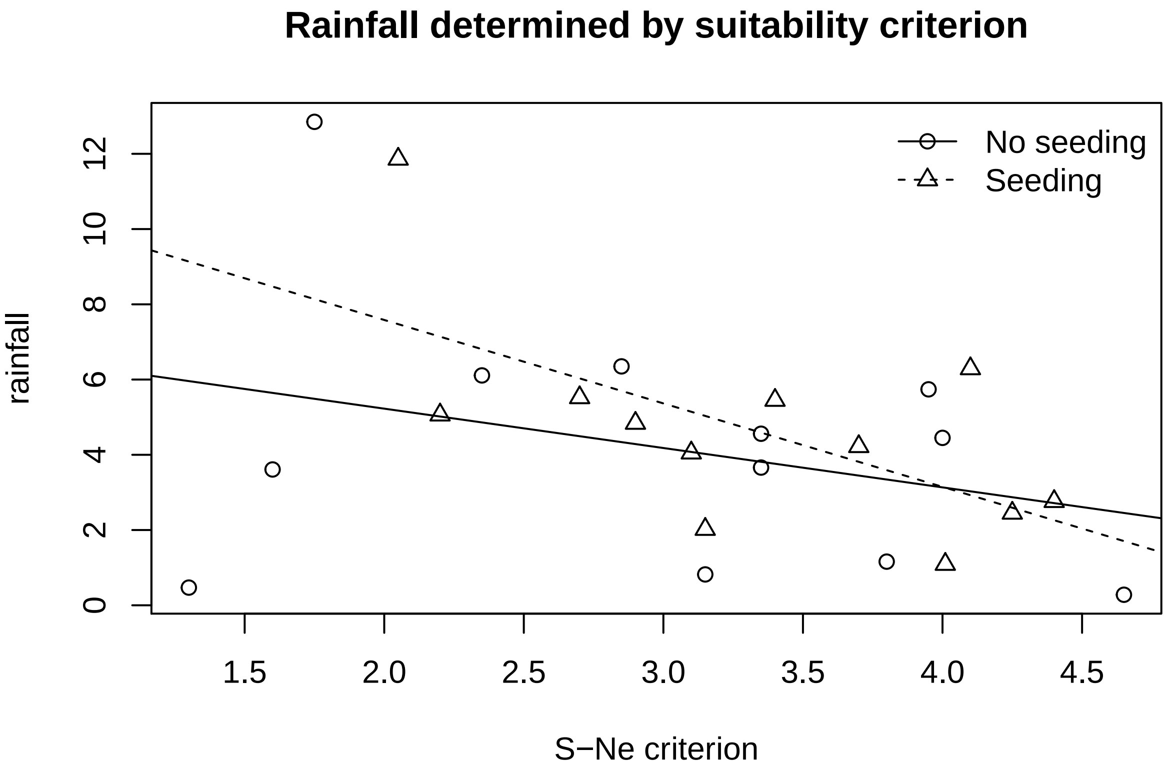
```
##
## Call: rq(formula = rainfall ~ sne, tau = 0.5, data = clouds)
##
## tau: [1] 0.5
##
## Coefficients:
##             Value    Std. Error t value  Pr(>|t|)
## (Intercept)  8.86133  3.64876    2.42859  0.02378
## sne         -1.38667  1.09225   -1.26955  0.21751
```

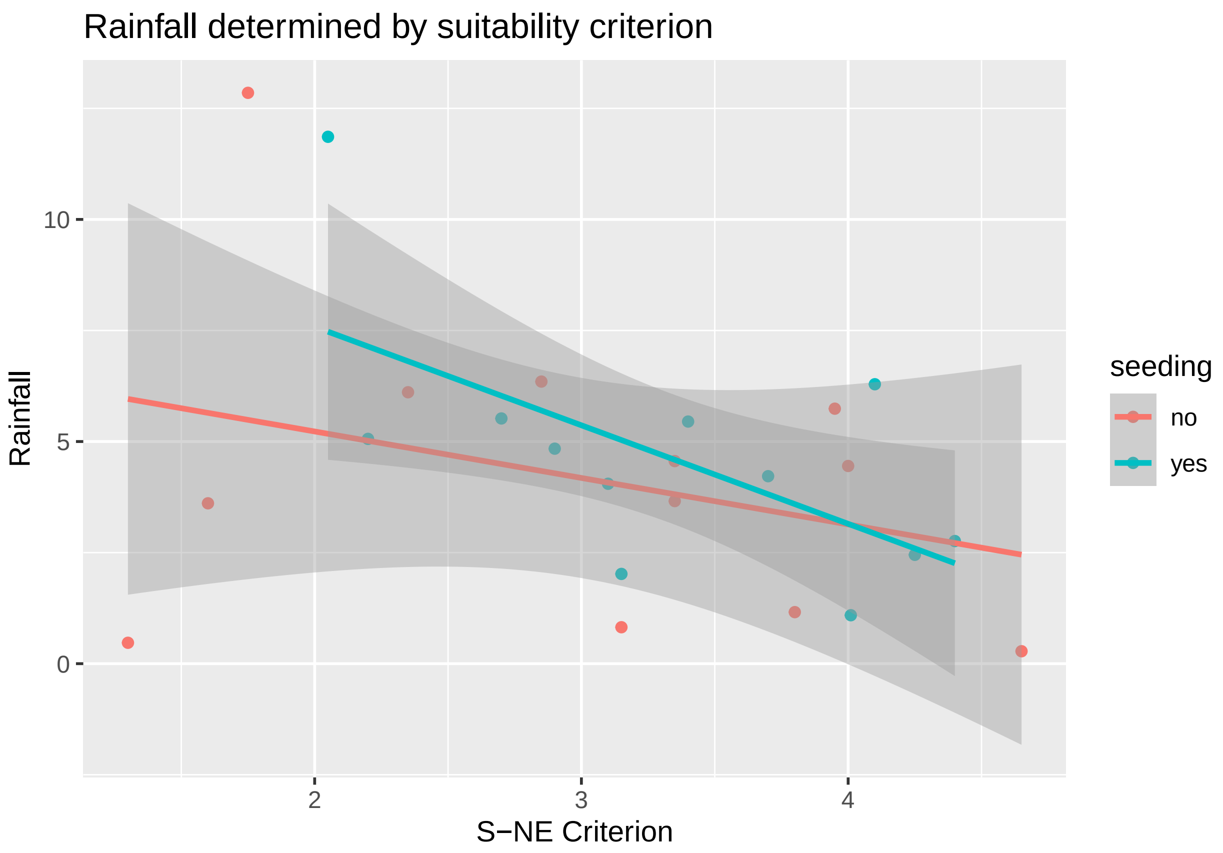
```
## Now we choose continous variable sne to fit our linear model
```

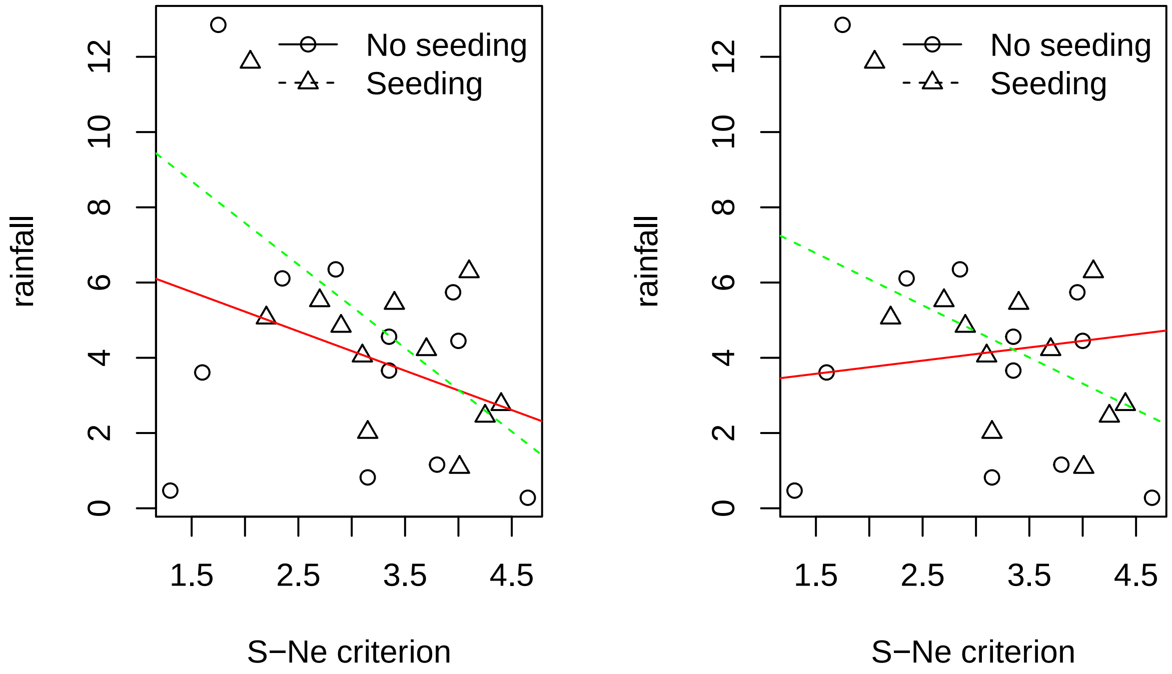
```
##
## Call:
## lm(formula = rainfall ~ sne, data = clouds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4927 -2.1116  0.0556  1.2295  6.5036
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.7430     2.1508   4.065 0.000515 ***
## sne          -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
```

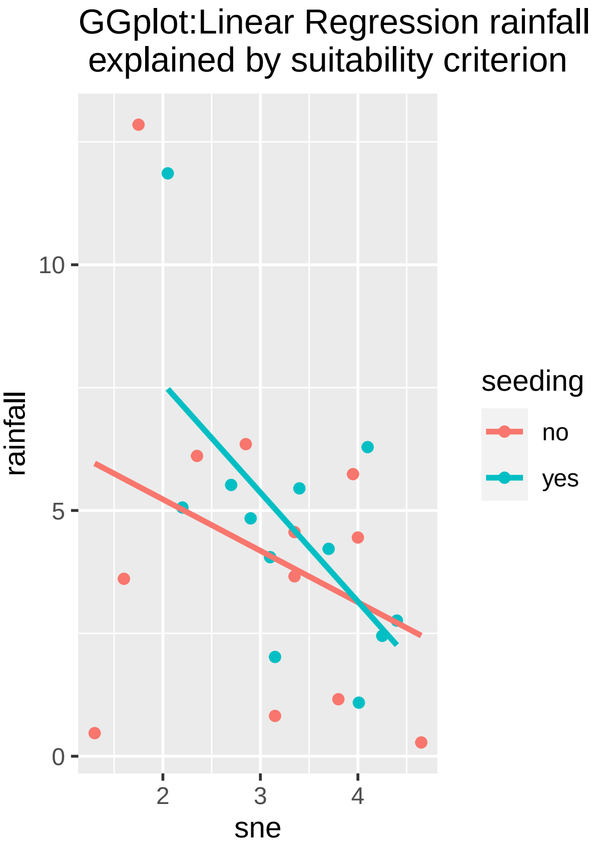
Table 1: MSE for the Median and linear Regression

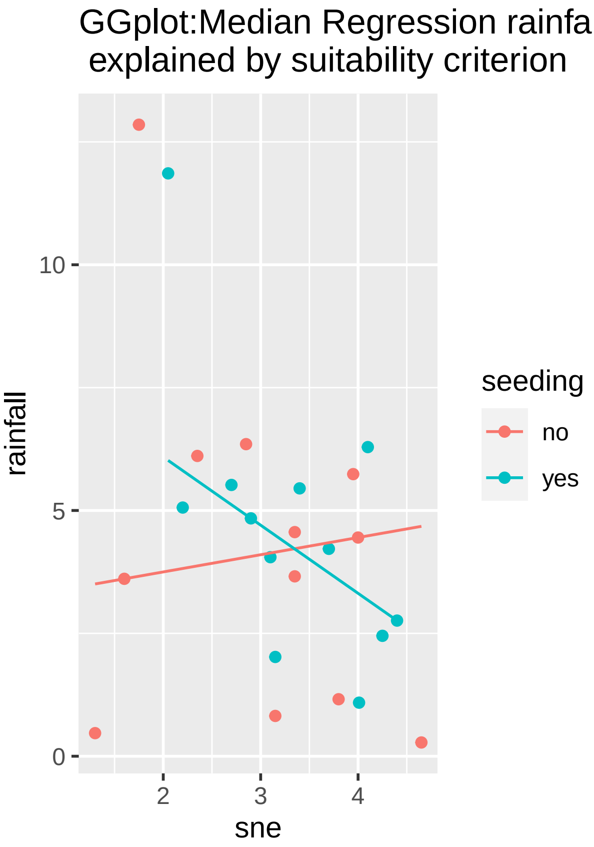
MSE of Median Regression	MSE of Linear regression
7.723	7.718


Rainfall determined by suitability criterion
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


Rainfall determined by suitability criterion
Rainfall
10
5
0
2
3
4
S−NE Criterion
seeding
no
yes


No seeding
Seeding
rainfall
0
2
4
6
8
10
12
1.5
2.5
3.5
4.5
S−Ne criterion
No seeding
Seeding
rainfall
0
2
4
6
8
10
12
1.5
2.5
3.5
4.5
S−Ne criterion





Comparison between median regression analysis with linear regression

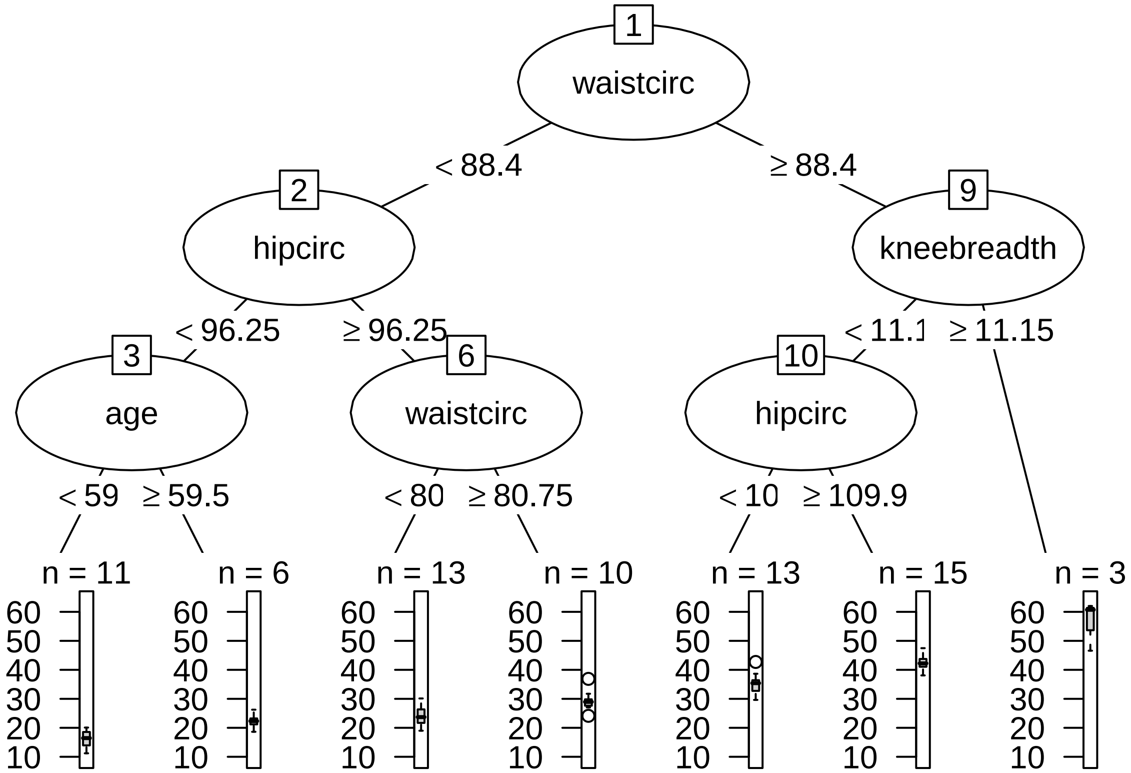
The data ‘Clouds’ from the HSAUR3 library was collected in the summer of 1975 from an experiment to investigate the use of massive amounts of silver iodide 100 to 1000 grams per cloud in the cloud, seeding to increase rainfall. The experiment was conducted in an area of Florida for 24 days. It was judged suitable for seedings on the basis that a measured suitability criterion (SNE). The data has a total variable of 7 with a sample of 24. To compare the model, both median regression and linear regression was fitted, calculated the mean squared errors and visual evidence. Quantreg library was used for performing median regression, and ggplot for the visual evidence.

The model was fitted, with all the variables as in chapter-6 of the textbook. It was done to find out which variable is more significant. From the summary, it appears seedingyes appear significant at 0.1 %. It looks like the seedingyes variable is the most significant variable in the model followed by seedingyes:sne. The results of the linear model fit, suggests that rainfall can be increased by cloud seeding. Moreover, the model indicates that higher values of the S-Ne criterion lead to less rainfall, but only on days when cloud seeding happened, i.e., the interaction of seeding with S-Ne significantly affects rainfall. Now we choose continuous variable sne to fit our linear and median model. Fitting the median regression with rq() and wanted to find out the significant variable in the model so, se = boot is used in the summary. The result showed no significance. I further find the MSE for the two models, here MSE for the linear regression as shone in table:1 is 7.718, and MSE for median regression is 7.723. The difference between the MSE is 0.01. I additionally plotted the graphs to compare the two models. From the plot of the median regression model on the absence of seeding is in such a way that the median regression line has a positive slope. Whereas, in the linear regression, it appears that it has a negative slope. This indicates that there is high variability in the rainfall data when cloud seeding is absent. Also, the median regression line is not weighted by the outliers. Therefore, median regression seems better at explaining the overall variability of the rainfall variable.

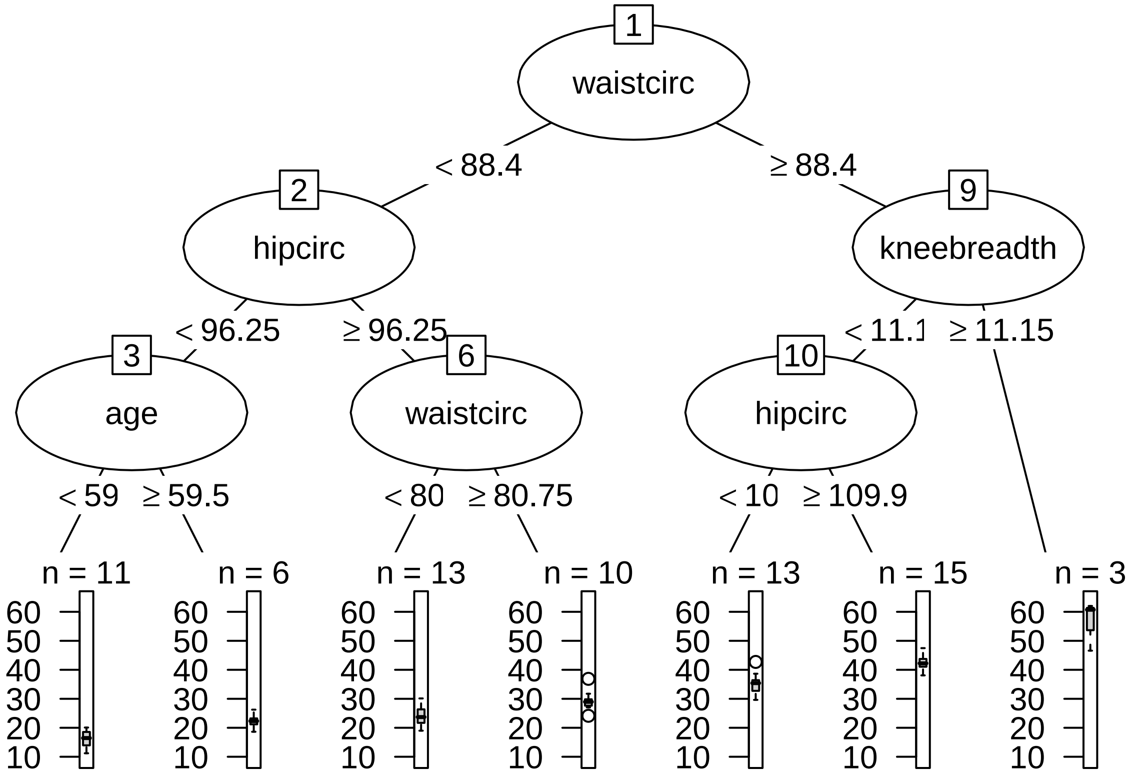
Hence, I recommend median regression for the analysis of cloud data.

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.



```
## CP table for the bodyfat

##           CP nsplit  rel error    xerror       xstd
## 1 0.66289544      0 1.00000000 1.0418641 0.17060953
## 2 0.09376252      1 0.33710456 0.4256876 0.09379196
## 3 0.07703606      2 0.24334204 0.4340129 0.08951396
## 4 0.04507506      3 0.16630598 0.3495848 0.07208586
## 5 0.01844561      4 0.12123092 0.2803559 0.06189382
## 6 0.01818982      5 0.10278532 0.2779785 0.06335328
## 7 0.01000000      6 0.08459549 0.2640637 0.06325110

## [1] "CP value with the lowest xerror is:  7"

## extracting the variable with the lowest xerror

## Predicting the pruned regression tree of bodyfat data

## Based on the pruning, plotting the regression tree
```



```
##
## 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
```

Table 2: MSE for the regression tree and median Regression

MSE of Regression tree	MSE of Median regression
10.171	15.025

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.

```
## Although we can choose waistcirc or hipcirc from the above pruned regression tree because it explain
```

```
##
```

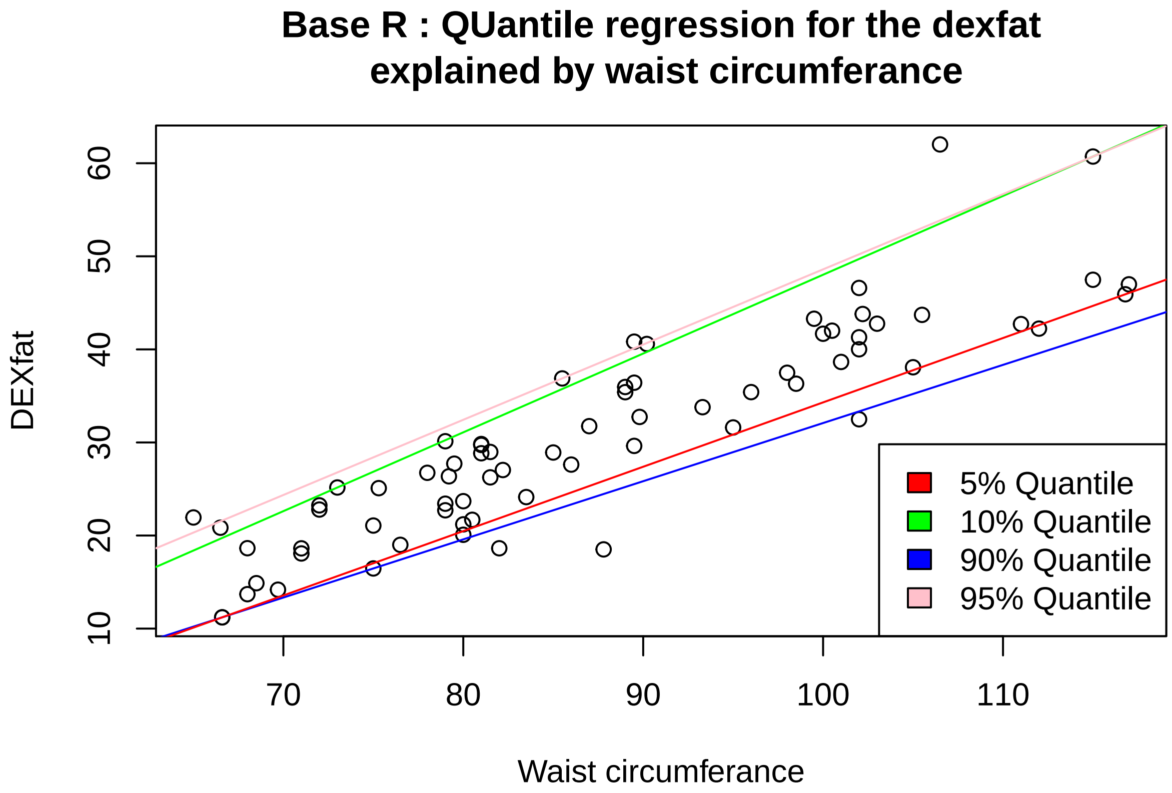
```
## Call:
## lm(formula = DEXfat ~ age + waistcirc + hipcirc + elbowbreadth +
##     kneebreadth, data = bodyfat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1782 -2.4973  0.2089  2.5496 11.6504
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -59.57320    8.45359  -7.047 1.43e-09 ***
## age            0.06381    0.03740   1.706   0.0928 .
## waistcirc      0.32044    0.07372   4.347 4.96e-05 ***
## hipcirc        0.43395    0.09566   4.536 2.53e-05 ***
## elbowbreadth  -0.30117    1.21731  -0.247   0.8054
## kneebreadth    1.65381    0.86235   1.918   0.0595 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.988 on 65 degrees of freedom
## Multiple R-squared:  0.8789, Adjusted R-squared:  0.8696
## F-statistic: 94.34 on 5 and 65 DF,  p-value: < 2.2e-16
```

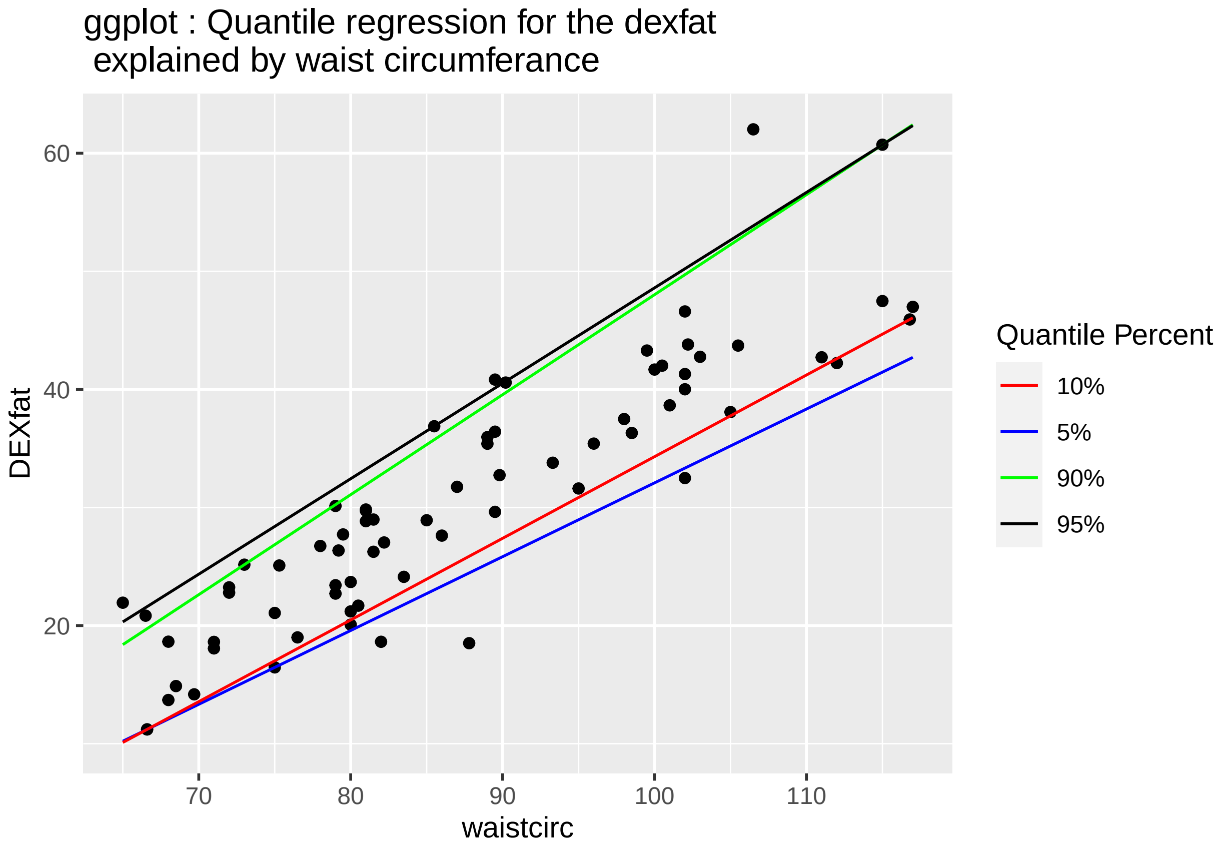
```
##
## Call: rq(formula = DEXfat ~ age + waistcirc, tau = 0.05, data = bodyfat)
##
## tau: [1] 0.05
##
## Coefficients:
##             coefficients lower bd  upper bd
## (Intercept) -42.63981    -54.18271 -28.63610
## age           0.20663     -0.05965   0.24345
## waistcirc     0.63528      0.51034   0.79141
```

```
##
## Call: rq(formula = DEXfat ~ age + waistcirc, tau = 0.1, data = bodyfat)
##
## tau: [1] 0.1
##
## Coefficients:
##             coefficients lower bd  upper bd
## (Intercept) -35.15246    -46.05029 -33.78070
## age           0.04148     -0.02522   0.20622
## waistcirc     0.67174      0.53838   0.70130
```

```
##
## Call: rq(formula = DEXfat ~ age + waistcirc, tau = 0.9, data = bodyfat)
##
## tau: [1] 0.9
##
## Coefficients:
##             coefficients lower bd  upper bd
## (Intercept) -33.34508    -44.32183 -10.91385
## age          -0.06352     -0.23022   0.09110
## waistcirc     0.84060      0.63991   0.93407
```

```
##
## Call: rq(formula = DEXfat ~ age + waistcirc, tau = 0.95, data = bodyfat)
##
## tau: [1] 0.95
##
## Coefficients:
##             coefficients lower bd  upper bd
## (Intercept) -29.80070    -49.79144  -3.93733
## age          -0.02634     -0.30269   0.09892
## waistcirc     0.79653      0.73097   1.05179
```





c) Provide a formal write up of the methodologies and of your results

Analyzing Body fat with the regression tree and median regression

Bodyfat data was from the Garcia et al. (2005) report on the development of predictive regression equations for body fat content through common anthropometric measurements. The data obtained from the healthy German women have ten variables and 71 number of samples. In R, the dataset bodyfat can be available inside the TH.data package. Two models, regression tree and median regression were fitted with the bodyfat data.

Firstly, I loaded all the libraries that will be used while performing the task. Rpart library helps to fit the regression tree in R. The dataset is then loaded with function data (“dataset name”, package = “package name”). With the help of the rpart function, I fit the regression model with the ten minsplit, which means that the data must have a minimum no of 10 observation in a node for a split to draw a regression tree. Then the summary is viewed. Additionally, the weakest link in the model is also found out and removed with the help of prune. The graphical representation of the regression tree before and after pruning is then plotted with the help of the partykit library. Mean square error (MSE) for the model is calculated, which helps in finding a better model.

Secondly, Median regression is fitted with the help of the quantreg library. Since we are fitting for the median, the value of tau is a 0.5 summary of the model is viewed. The mean square of the median regression is then calculated and combined in a table with the regression tree MSE.

Thirdly, the result for both models is viewed. In the regression tree model, it appears waistcirc and hipcirc explain the maximum data of the bodyfat data. Since the model chooses waistcirc as the root of the model. The same information is provided by the plot. The accuracy of the model is calculated with the help of the mean square error method. The MSE of the model is 10.171.The median regression model is then fitted based on the above-pruned tree, the variables waist circumference and hip circumference splits explain the majority of the data, and I chose waist circumference for quantile regression. Based on this analysis, the pruned regression tree has a lower MSE than the median regression model and has an MSE of 15.025. Therefore regression tree seems more favorable.

For the visual evidence, the relationship of Dexfat to Age by Waist Circumference, all four quantiles regression lines have a positive slope. It also appears the slopes of 5 and 10 percentage converges at a point at the end. And the slopes of 90 and 95 percentage converge at a point at the beginning. Within the given percentage, it looks like most of the point have been covered within the interval.