Homework 8

James Young

Please do the following problems from the text book R Handbook and stated.

Resources: Stack Overflow

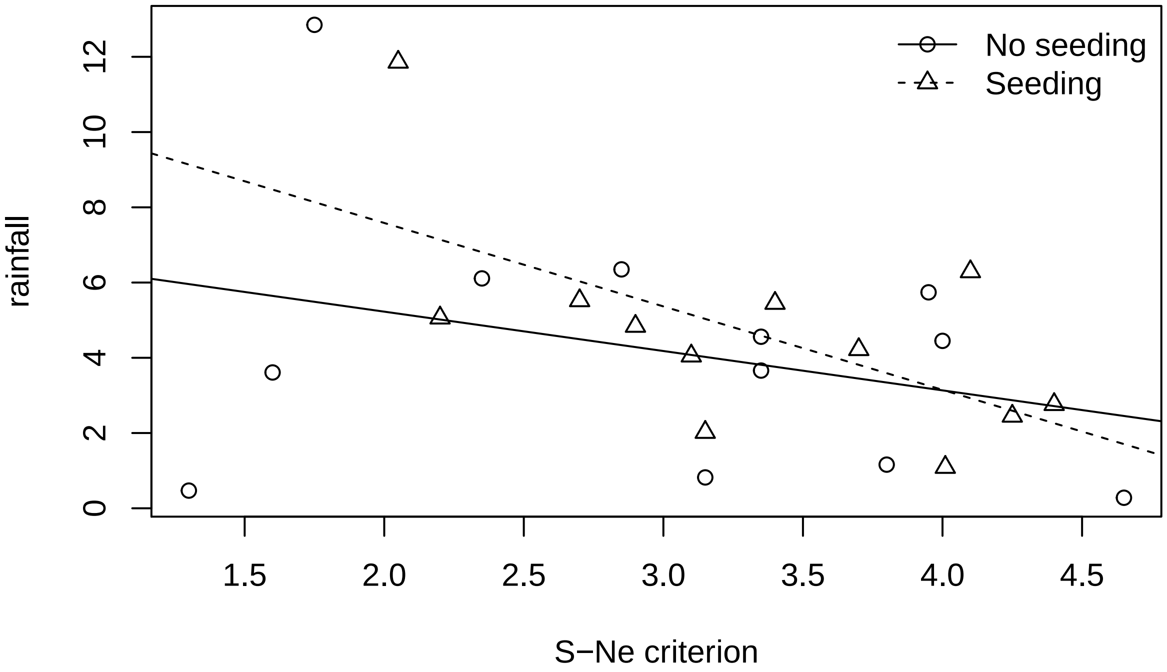
The following libraries were used: library(HSAUR3) library(ggplot2) library(quantreg) library(gamlss.data) library(lattice) library(TH.data) library(rpart) library(partykit) library(gridExtra)

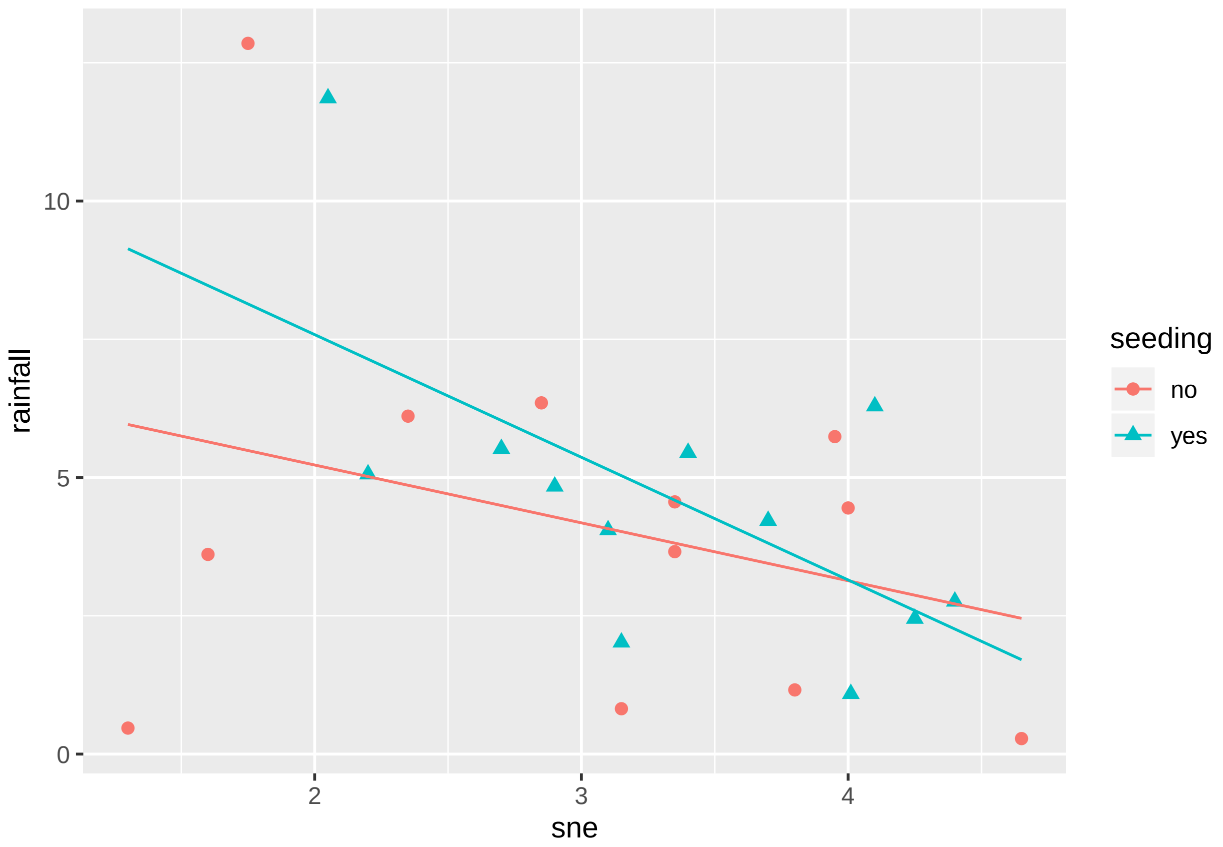
1. Consider the {**clouds**} data from the {**HSAUR3**} package

   a) Review the linear model fitted to this data in Chapter 6 of the text book and report the model and findings.

Replicating the model from chapter 6, we see that rainfall is effected by the treatment of cloud seeding with suitability criterion. We see that rainfall has a significant increase when seeding takes place.

```
##
## 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
```

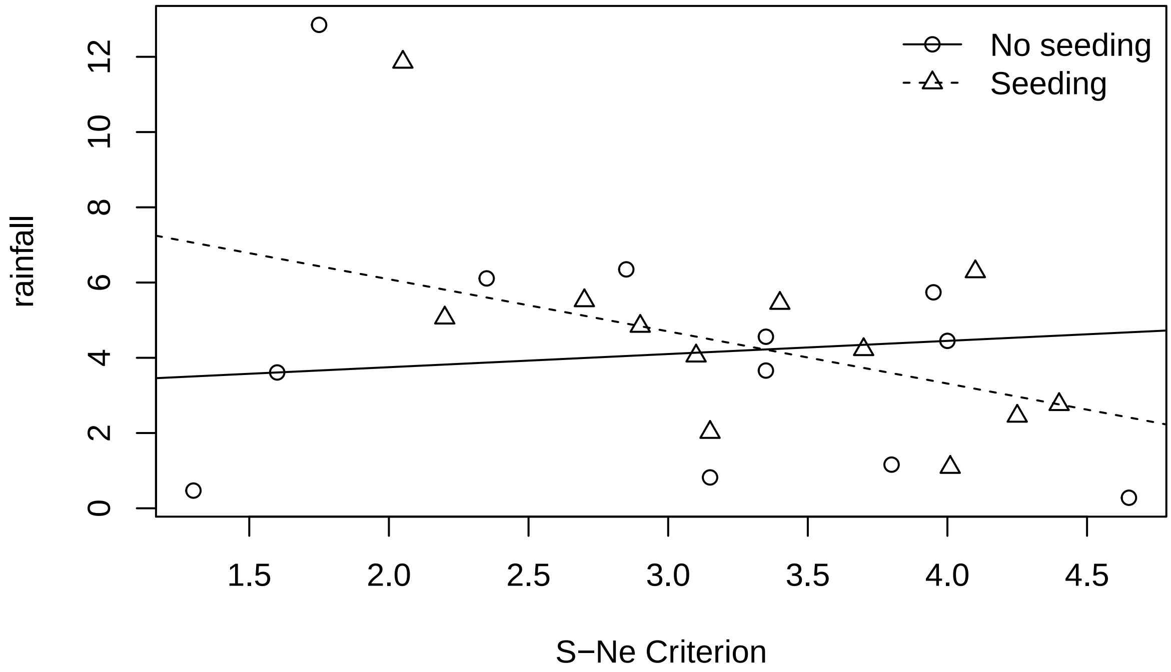

No seeding
Seeding
rainfall
0
2
4
6
8
10
12
1.5
2.0
2.5
3.0
3.5
4.0
4.5
S−Ne criterion

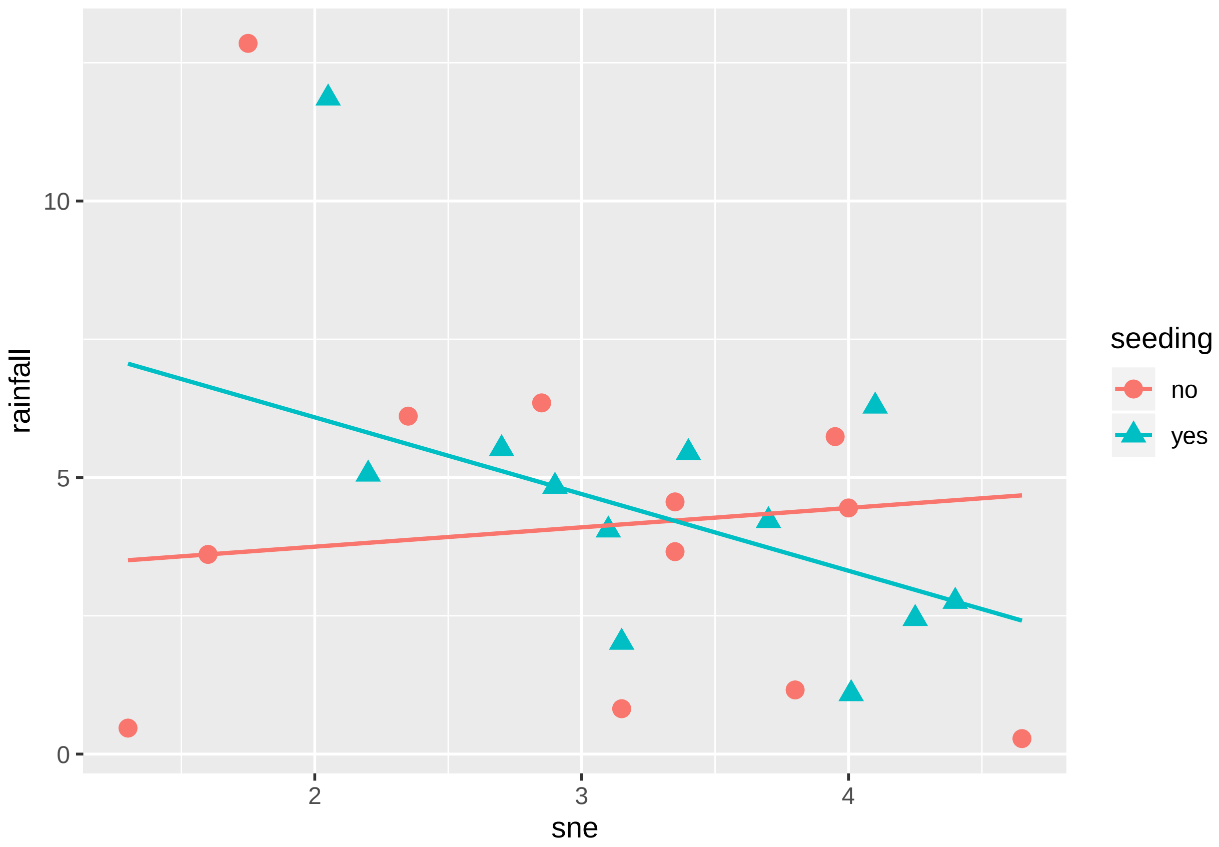


```
b) Fit a median regression model.
```

To fit a median regression model we fit a line with tau = 0.5 (the median) with the same formula from part A looking for difference between seeding treatments.

```
##                                  coefficients      lower bd      upper bd
## (Intercept)                       -0.39510353  -2.032259e+00  1.234196e+01
## seedingyes                         9.28416250   4.632247e+00  2.478669e+01
## time                              -0.02682160  -7.150623e-02 -2.068740e-02
## seedingno:sne                      0.36860476  -1.090559e+00  1.196003e+00
## seedingyes:sne                    -1.33267160  -6.025488e+00 -1.177594e+00
## seedingno:cloudcover               0.20691306   1.818597e-02  1.043587e+00
## seedingyes:cloudcover             -0.06071068  -3.426312e-01  2.468352e-01
## seedingno:prewetness               5.22263667  -9.255066e+00  1.156672e+01
## seedingyes:prewetness              2.01808261 -1.797693e+308  1.797693e+308
## seedingno:echomotionstationary     2.13502276  -4.986951e-01  1.103820e+01
## seedingyes:echomotionstationary    2.78255068 -1.797693e+308  1.797693e+308
```


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



```
c) Compare the two results.
```

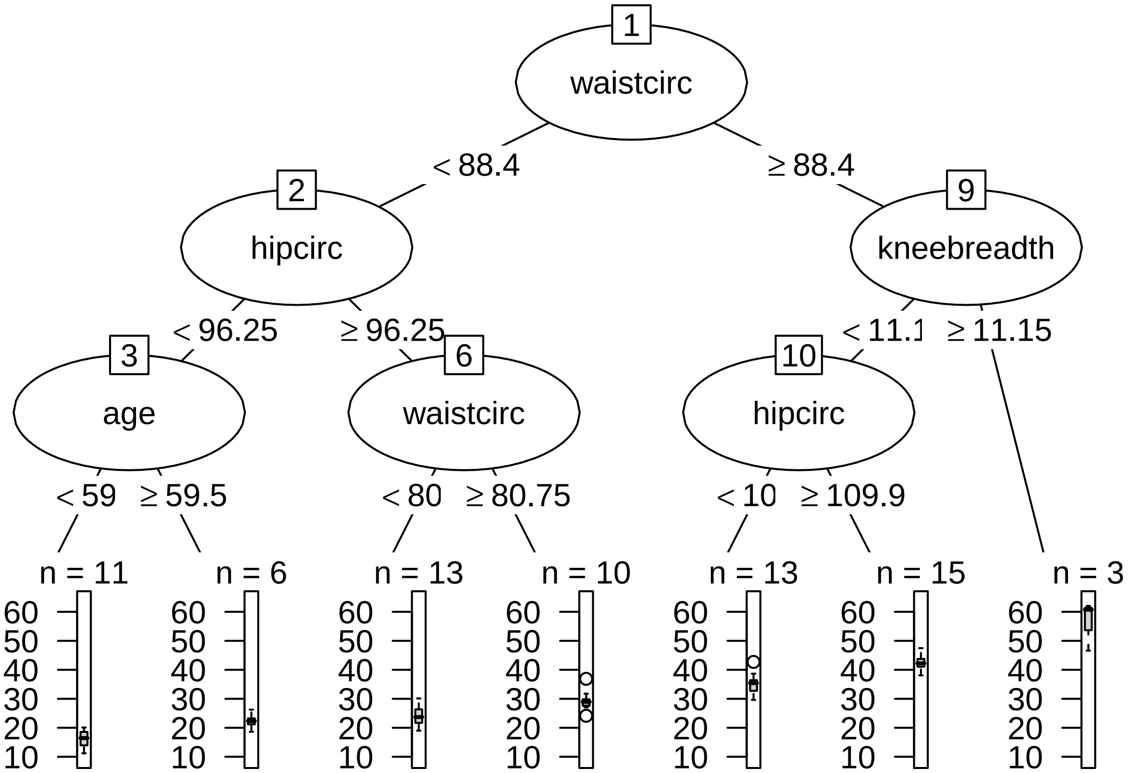
Looking at the graphs from part A and part B we can see a difference in the slop of the line for the "no-seeding" treatment between the linear and median regression, showing there is variability as to the level of covariance between these two variables. We see a more consistent slope between the two models for the "seeded" treatment which makes this data seem more robust, possibly adding evidence to the effect of the "seeding" treatment. Comparing performance metrics, we see the linear regression out performs the median regression in MSE but is outperformed by the median regression in AIC and MAE, which makes sense considering the median regression focus on optimizing absolute residual error instead of least square error like a linear regression.

```
##     Chapt.6.Linear Chapt.6.Median
## MSE       2.632871      4.2491485
## AIC     115.342843    102.4357642
## MAE       1.283675      0.9827484
```

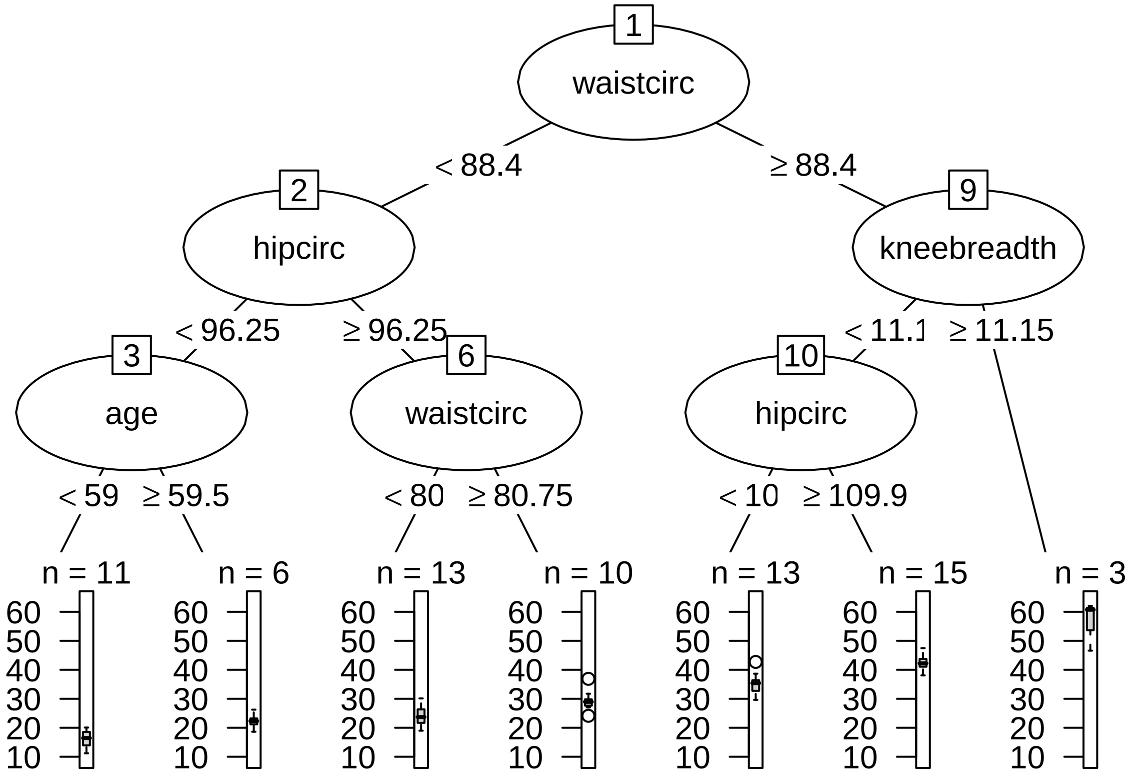
2. Reanalyze the {**bodyfat**} data from the {**TH.data**} package.

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

First I'll make the unpruned tree.



Now I'll make the pruned tree optimizing for error. We can see there isn't a change in this case.



Finally, we will compare the models. We can see below that the tree models outperform median regression in the MSE metric by a considerable amount but the median regression is closer to tree performance with MAE metric which makes sense considering median regression focus on optimizing absolute residuals. However it looks like the tree still outperforms median regression in mean absolute error in this case.

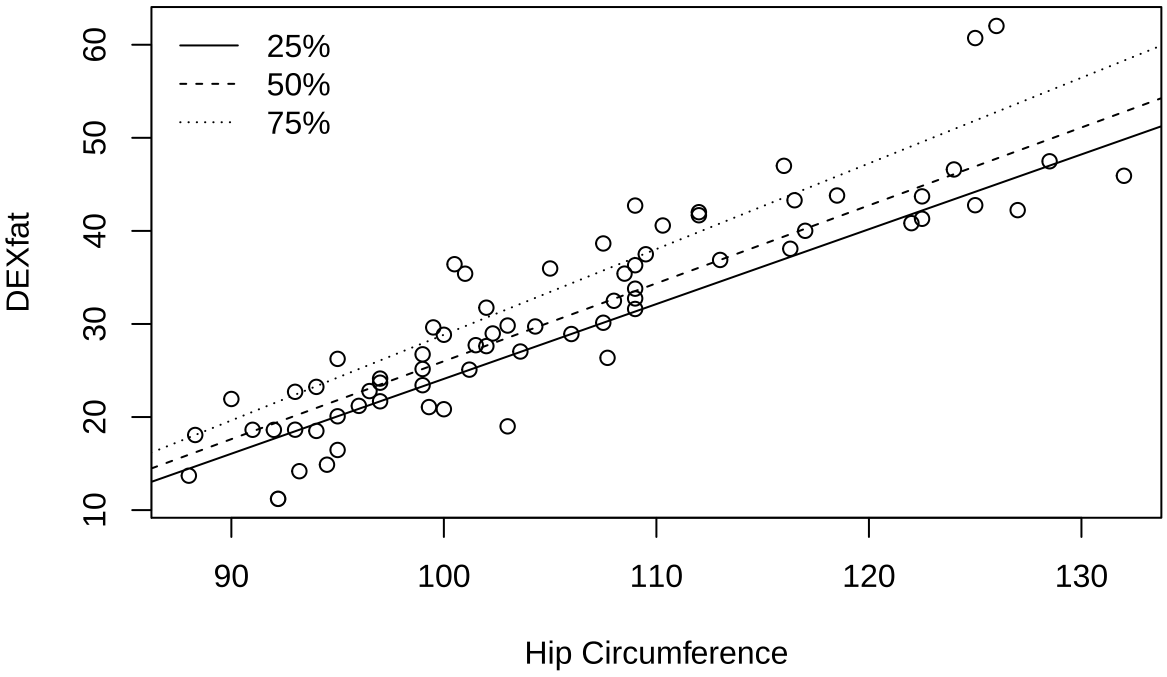
```
##     Unpruned.Tree Pruned.Tree Median.Regression
## MSE     10.170503   10.170503         15.024504
## MAE      2.476113    2.476113          2.947714
```

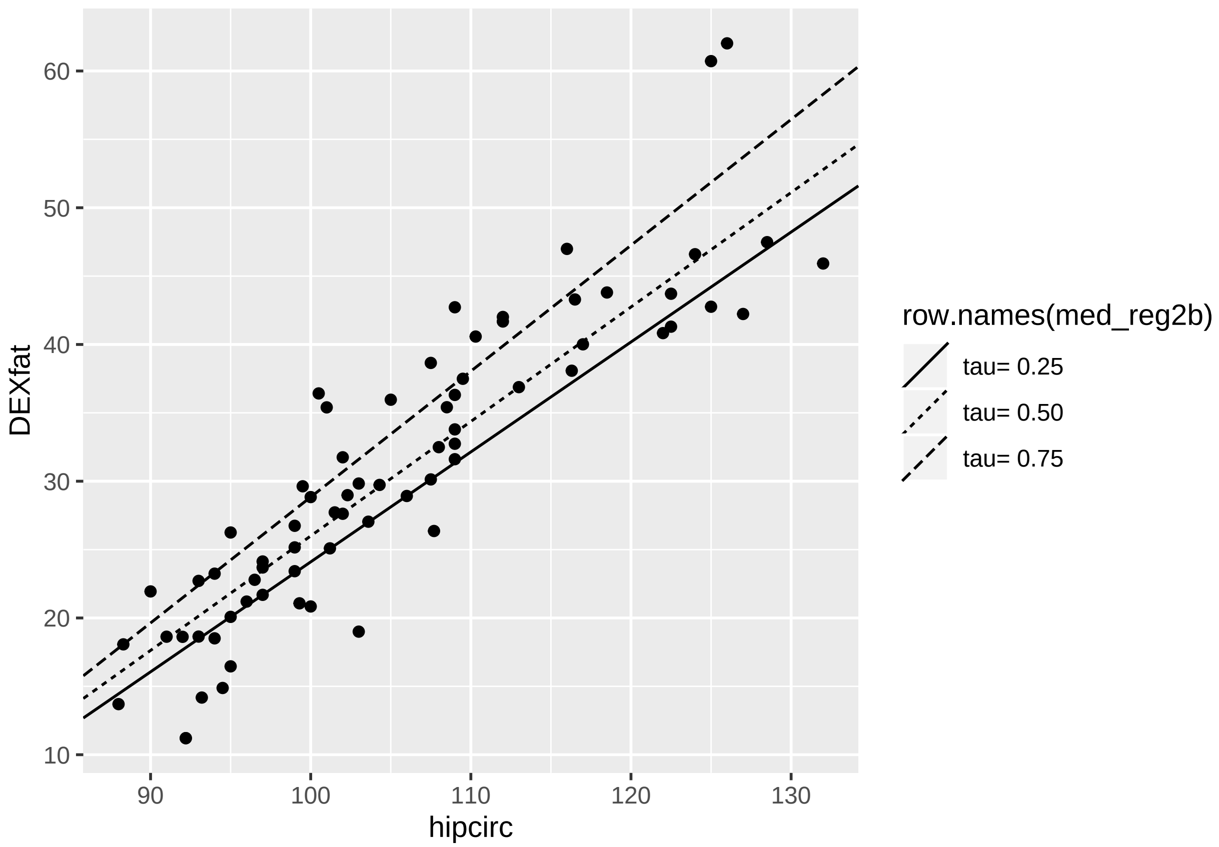
```
b) Choose one independent variable. For the relationship between this variable and DEXfat, create linea
```

Following the above instructions I chose to work with hipcirc which can be seen in the plot below. We see that higher percentiles have a steeper slope than lower percentiles when looking at hipcirc's covariation with DEXfat.

```
##
## Call: rq(formula = DEXfat ~ hipcirc, tau = c(0.25, 0.5, 0.75), data = bodyfat)
##
## tau: [1] 0.25
##
## Coefficients:
##             coefficients lower bd  upper bd
## (Intercept) -56.30000    -65.26907 -51.50345
## hipcirc       0.80400      0.75633   0.88699
```

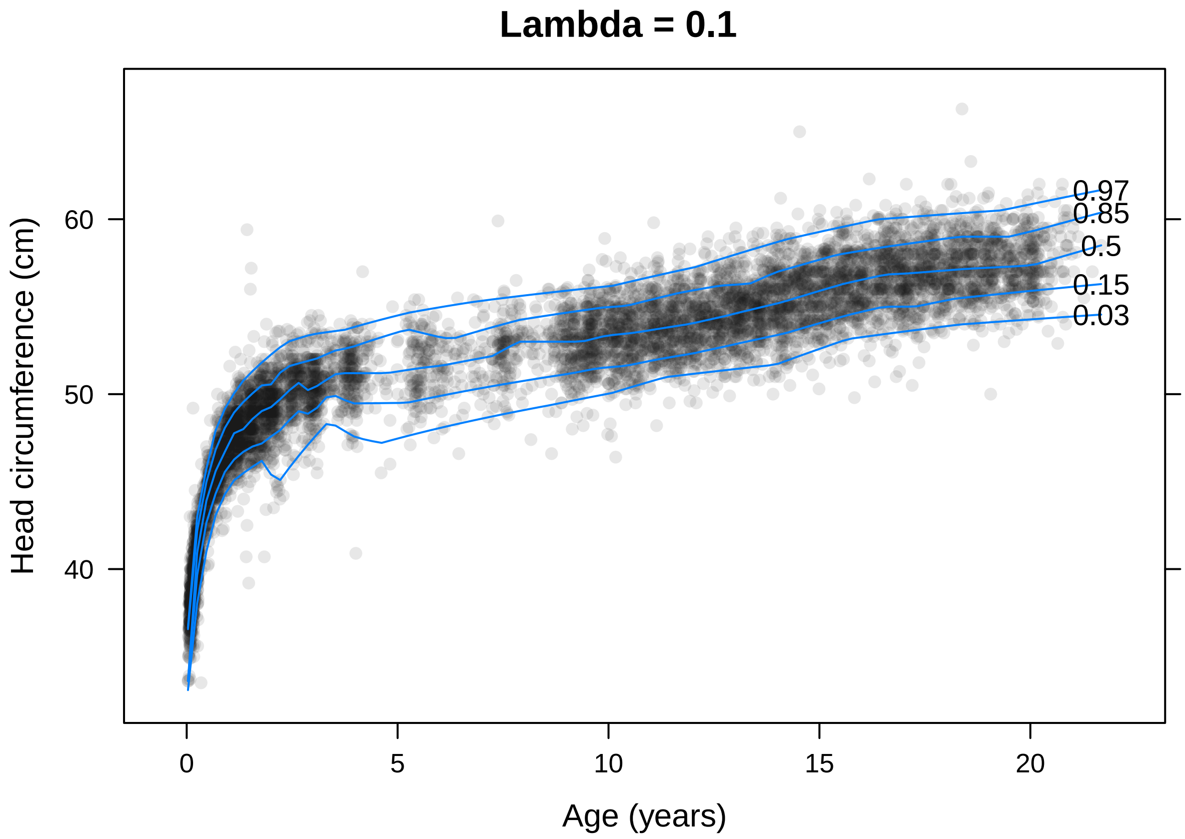
```
##
## Call: rq(formula = DEXfat ~ hipcirc, tau = c(0.25, 0.5, 0.75), data = bodyfat)
##
## tau: [1] 0.5
##
## Coefficients:
##             coefficients lower bd  upper bd
## (Intercept) -57.71714    -72.90737 -47.34519
## hipcirc       0.83714      0.74305   0.98669
##
## Call: rq(formula = DEXfat ~ hipcirc, tau = c(0.25, 0.5, 0.75), data = bodyfat)
##
## tau: [1] 0.75
##
## Coefficients:
##             coefficients lower bd  upper bd
## (Intercept) -63.21128    -90.41231 -47.98983
## hipcirc       0.92051      0.76506   1.19799
```

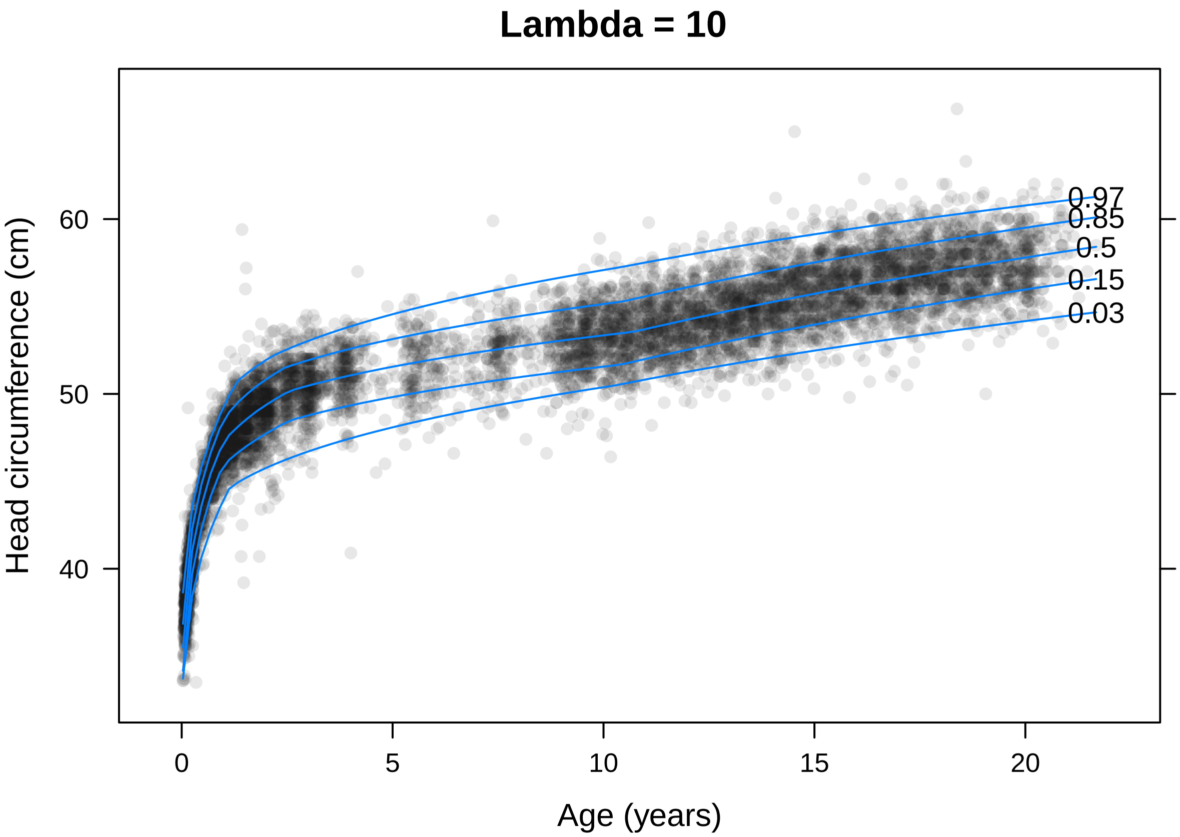


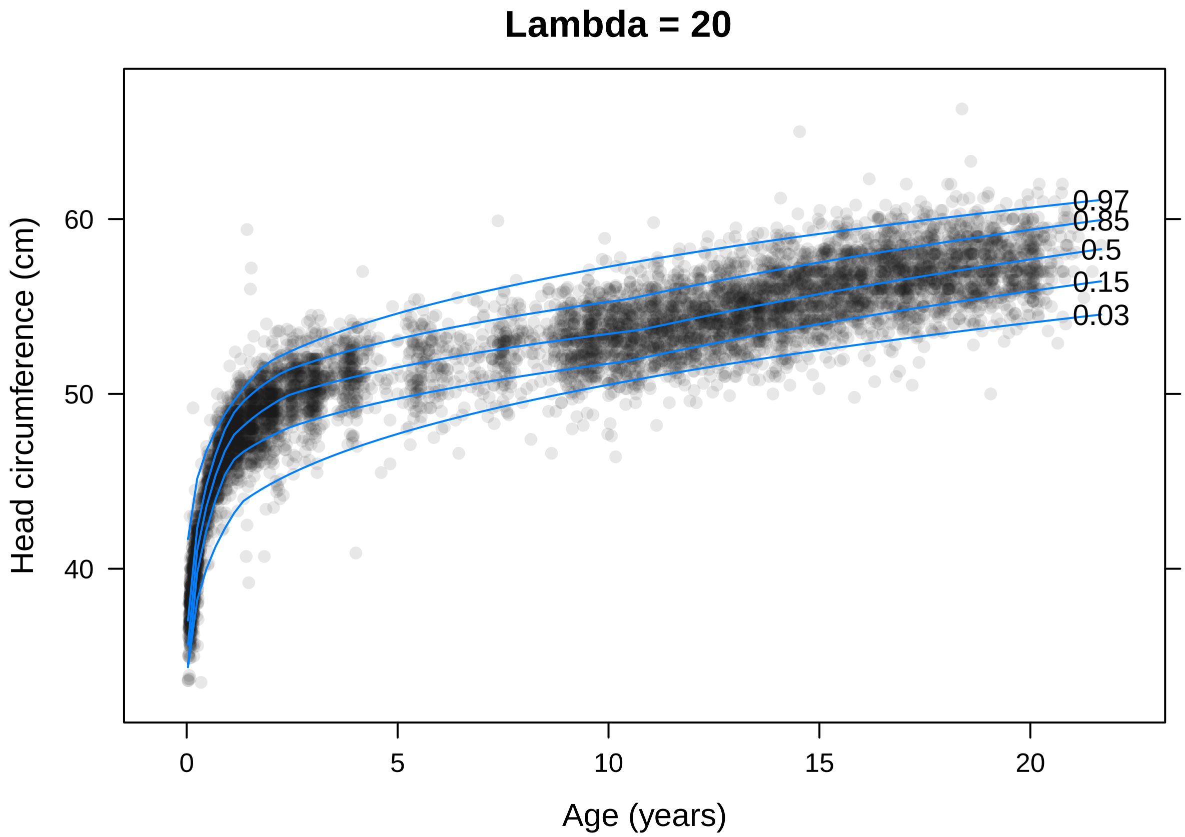


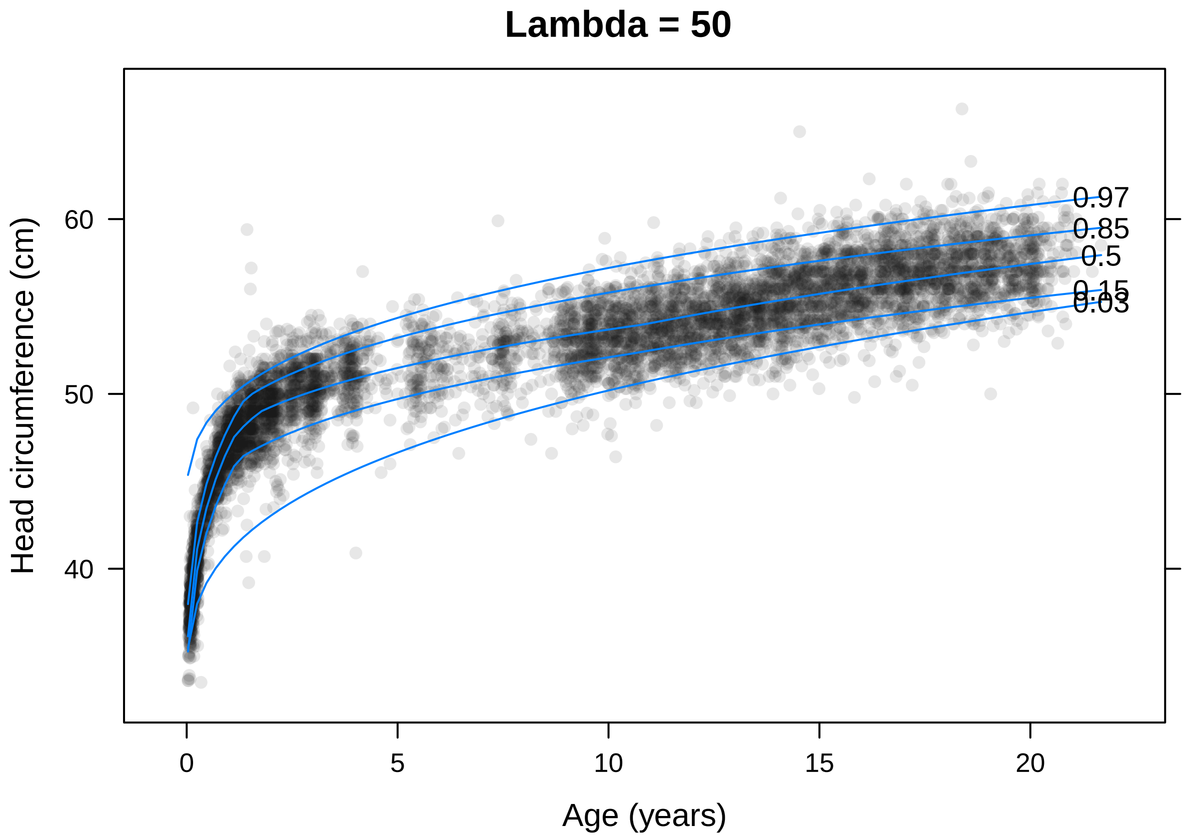
3. Consider {**db**} data from the lecture notes (package {**gamlss.data**}). Refit the additive quantile regression models presented ({**rqssmod**}) with varying values of λ (lambda) in {**qss**}. How do the estimated quantile curves change?

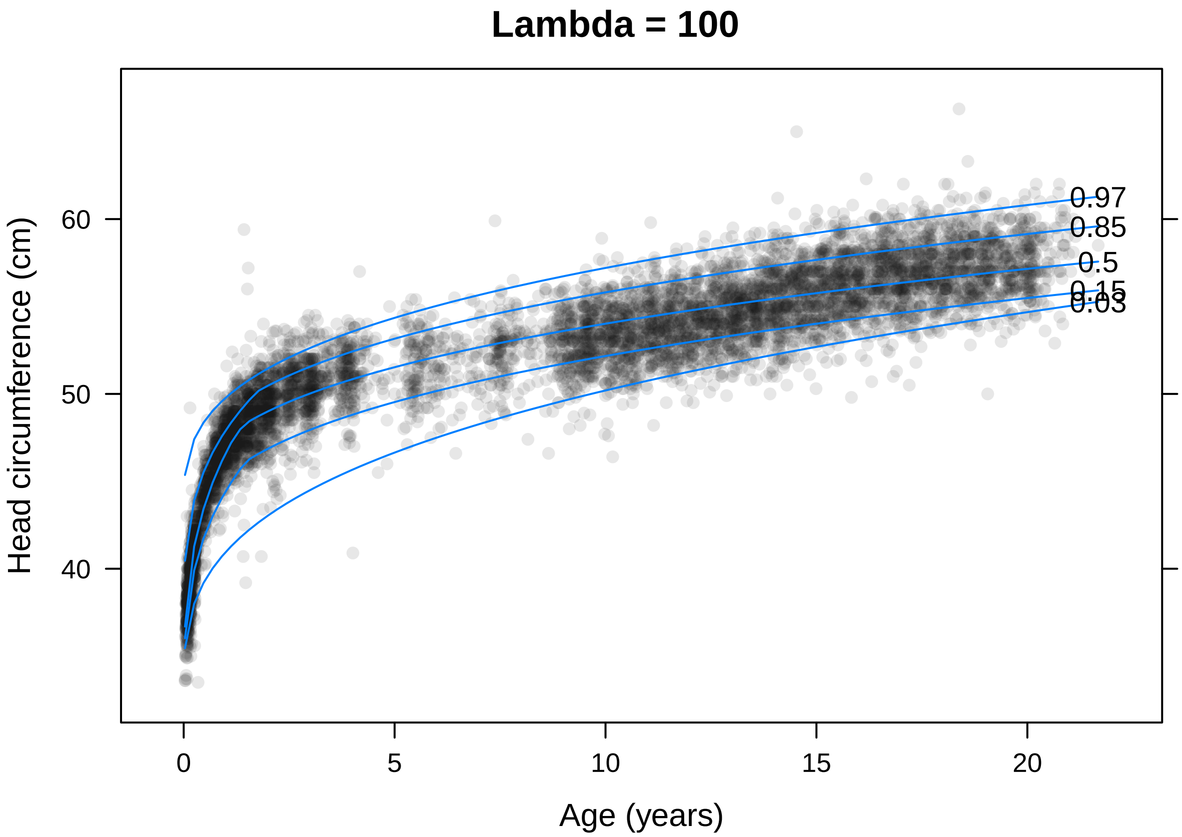
As lambda increases the smoothness of the quantile lines increases. At Lambda of 0.1 we can see some undulation in the quantile lines but by lambda=10 it is pretty much smoothed out and we don't see any visually significant change after lambda = 10. Lamda acts as a penalty term in this case to smooth the lines of fit. I tried to make the base r plots into a grouped frame using par(mfrow) but struggled in this case. However, the ggplots could be easily organized using grid.arrange(). Also, the ggplot y-axis label is abbreviated to alleviate cluster.

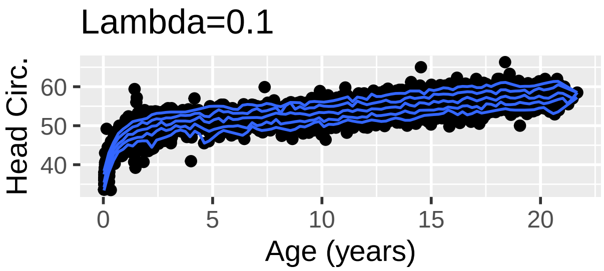

Lambda = 0.1
Head circumference (cm)
60
50
40
0
5
10
15
20
Age (years)
0.97
0.85
0.5
0.15
0.03

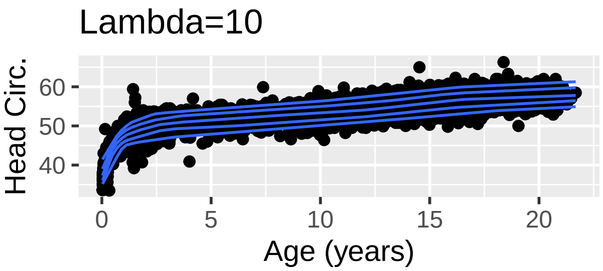

Lambda = 10
Head circumference (cm)
60
50
40
0
5
10
15
20
Age (years)
0.97
0.85
0.5
0.15
0.03

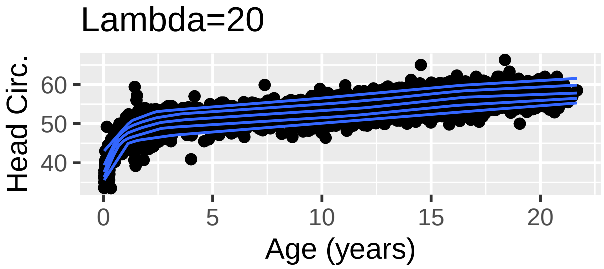

Lambda = 20
Head circumference (cm)
40
50
60
Age (years)
0
5
10
15
20
0.97
0.85
0.5
0.15
0.03

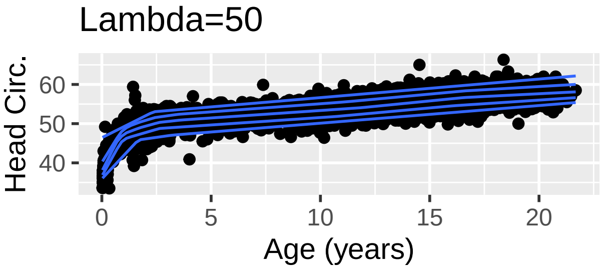

Lambda = 50
Head circumference (cm)
60
50
40
Age (years)
0
5
10
15
20
0.97
0.85
0.5
0.15
0.03

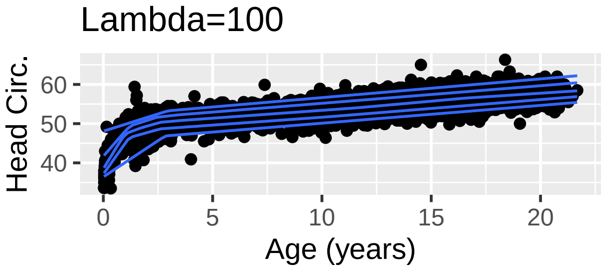

Lambda = 100
Head circumference (cm)
40
50
60
0
5
10
15
20
Age (years)
0.97
0.85
0.5
0.15
0.03











4. Read the paper by Koenker and Hallock (2001), posted on D2L. Write a one page summary of the paper. This should include but not be limited to introduction, motivation, case study considered and findings.

Introduction The paper "Quantile Regression", written by Koenker and Hallock starts by describing the process of breaking populations into equal sized groups along a response variable. A quartile is four groups, a quintile is five groups, a decile is ten groups, and quantiles work in 1/100th increments. Quantiles use the Greek "tau" to represent the quantile an observation is in with the observation being of higher values than "tau" but lesser value than 1-"tau". The purpose of using quantiles in quantile regression is to minimize the absolute residual sum above and below a given "tau" while linear regression tries to minimize mean square error regardless of balance above or below a quantile.

Motivation The motivation for quantile regression is demonstrated in the paper through the example of the association of income with food expenditure. Using a normal regression approach was biased by observations that were outliers. A quantile regression in this case could eliminate the bias introduced by the outliers by focusing on absolute residual errors.

Case Study After introducting the previous example of the advantages in using median regression for certain examples, the paper got into what I consider their case study. The case was examining the association between infant birth weight and many factors. Having a low birth weight was defined in this case as a weight below 5 pounds and 9 ounces at time of birth. The distribution of weights created a long tail to the lower side. A mean regression would be skewed by these much lower weights even though they make up a relatively smaller portion of the population. Quantile regression, with its focus on minimizing absolute residual would more accurately reflect the median observation.

Results/Findings The least square errors based regression found boys were born larger than girls on averag by ~100 grams. At the lowest quantiles the difference between boys and girls was almost half that difference, with boys being 45 grams larger. And at the highest quantiles boys were about 130 grams larger than girls, a larger difference than was seen in the least square errors estimate. The variability in birth weight distribution was explained better by quantile regression in this case than least square errors. Another major result was the difference in birth weights between mothers ethnicity with the difference between the infants of black and white mothers being about 300 grams at the lowest quantiles. Other factors found to be contributors were smoking, marital status, education level, and prenatal care. Most associations were better characterized by quantile regression than linear regression in this case due to variability in associations at different quantiles.

Conclusion This paper demonstrated how data with long tails can skew the mean away from the median and cause problems with the reliability in regular linear regression. Many metrics follow these distributions with long tails, for example income. You may be told the average income in a given field is x amount of money and be lead to believe you have a good chance of making that much money. However, it is possible that the top performers in that field make incredibly more money than the median income and thus skew your expectations based on the mean. Median regression gives a more balanced view in cases where the distribution has outliers or long tails.