Homework 8

Alex Soupir

November 7, 2019

Packages: HSAUR3, quantreg, TH.data, gamlss.data, lattice, GGplot2

Collaborators:

Chapter 12

Answer all questions specified on the problem and include a discussion on how your results answered/addressed the question.

Submit your **.rmd** file with the knitted **PDF** (or knitted Word Document saved as a PDF). If you are having trouble with .rmd, let us know and we will help you, but both the .rmd and the PDF are required.

This file can be used as a skeleton document for your code/write up. Please follow the instructions found under Content for Formatting and Guidelines. No code should be in your PDF write-up unless stated otherwise.

For any question asking for plots/graphs, please do as the question asks as well as do the same but using the respective commands in the GGPLOT2 library. (So if the question asks for one plot, your results should have two plots. One produced using the given R-function and one produced from the GGPLOT2 equivalent). This doesn't apply to questions that don't specifically ask for a plot, however I still would encourage you to produce both.

You do not need to include the above statements.

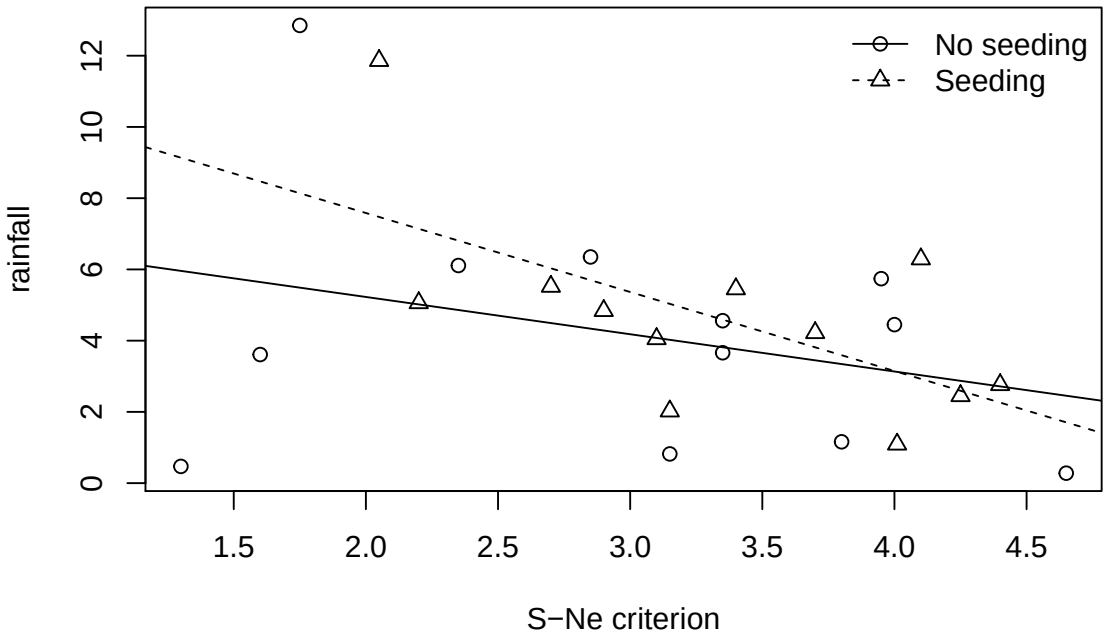
Please do the following problems from the text book R Handbook and stated.

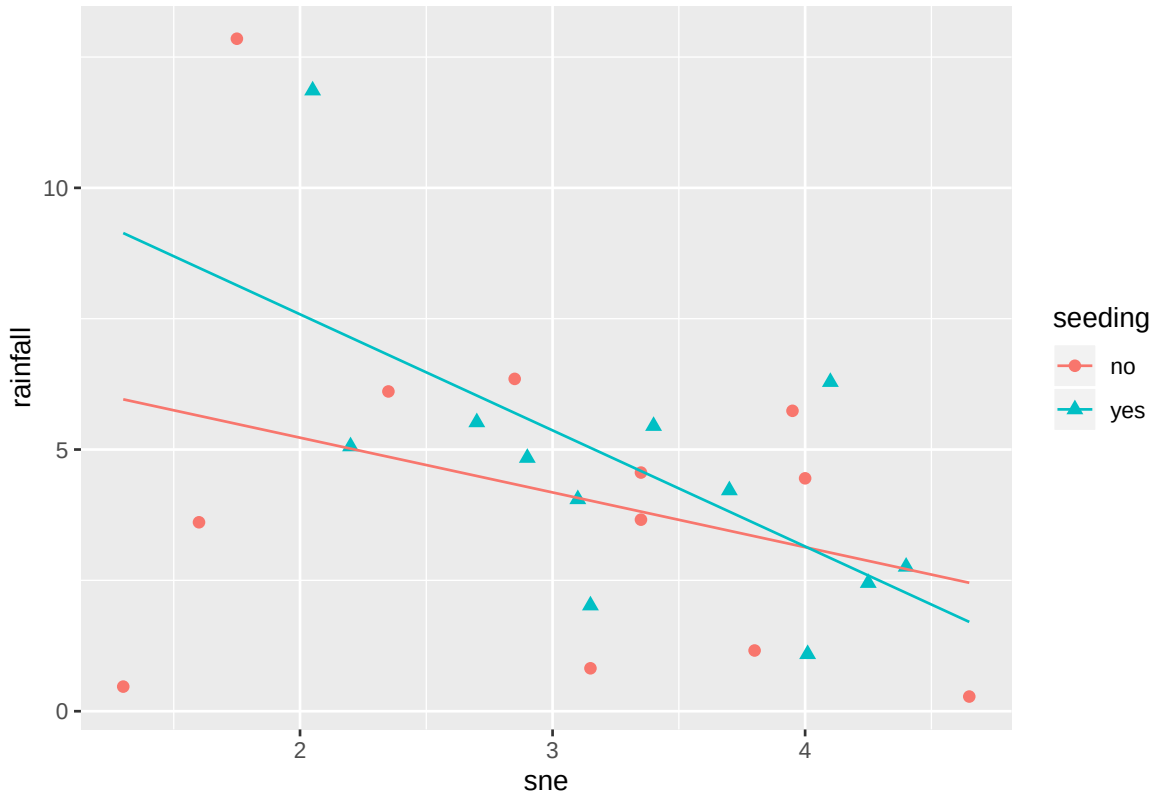
1. Consider the {**clouds**} data from the {**HSAUR3**} package

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

```
##
## 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
```

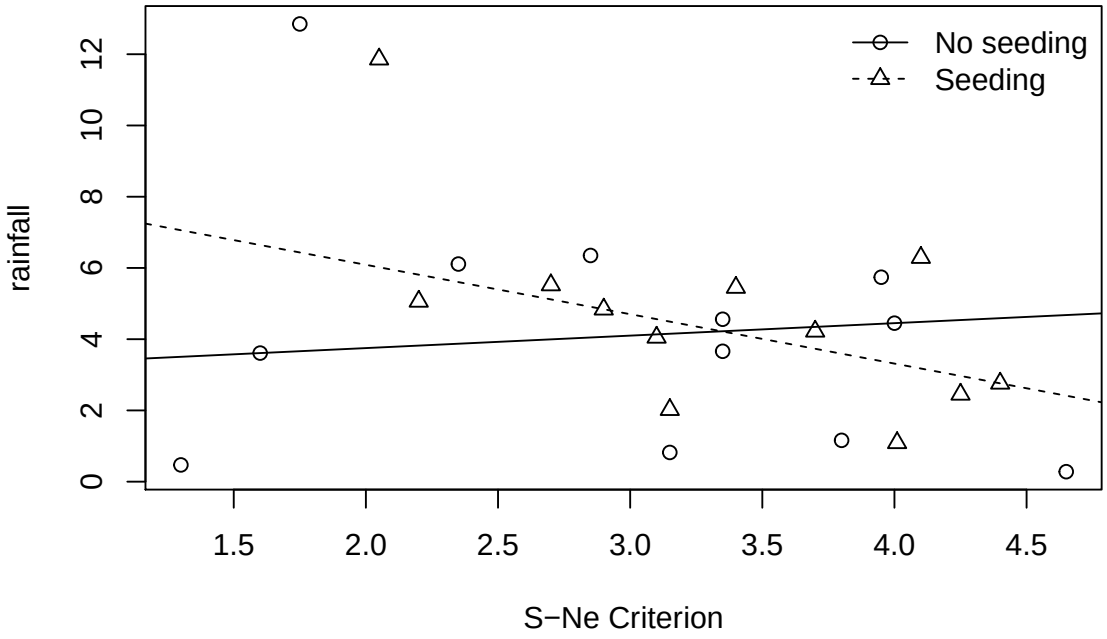


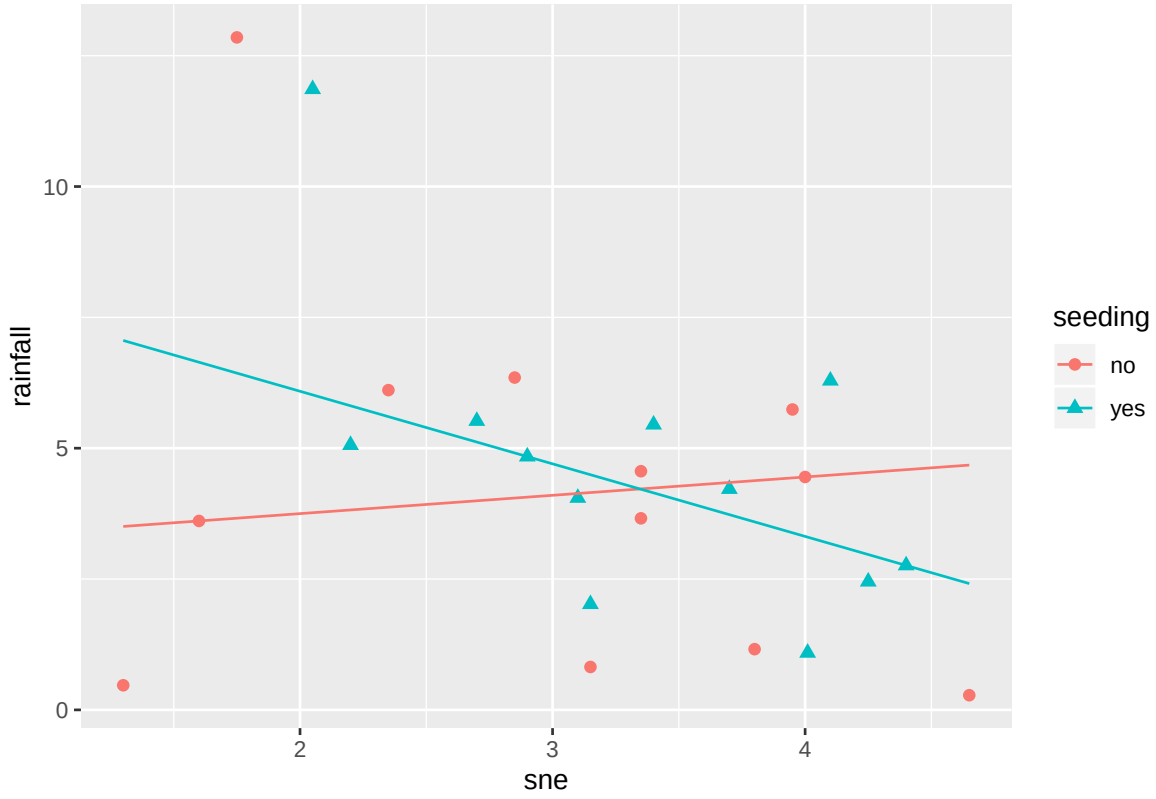


Above is the model summary as well as the plot of the relationship between the suitability criterion and rainfall with and without seeding. The summary shows that a model with seeding is better able to predict the rainfall than a model with no cloud seeding. Seeding clouds and suitability criterion also shows significance to the model meaning that the effect of seeding clouds is not standalone but also depends on the suitability criterion for determining the rainfall.

b) Fit a median regression model.

```
##                                  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
```





A median regression model was fit using *tau = 0.5* on the same formula from the book that was used in *Part A*. The plot was created the same was as *Part A* as well but using

***rq* instead of *lm* and adding *tau*.**

c) Compare the two results.

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

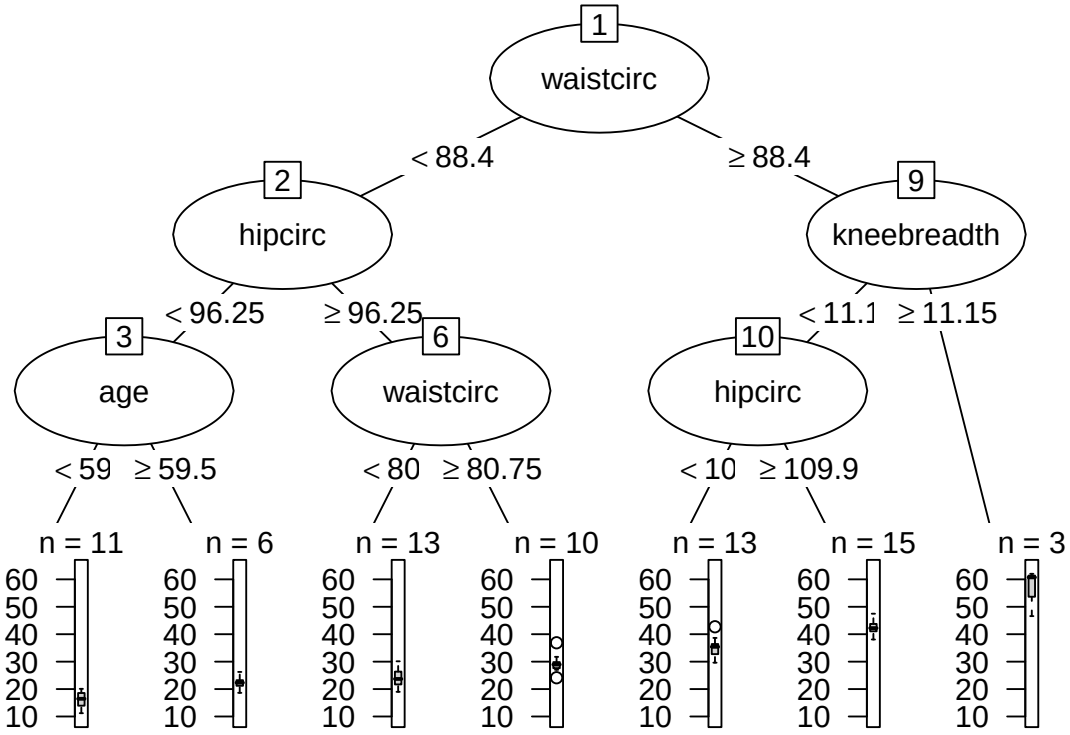
```
##
## Coefficients of linear and median models using clouds_formula from book:
```

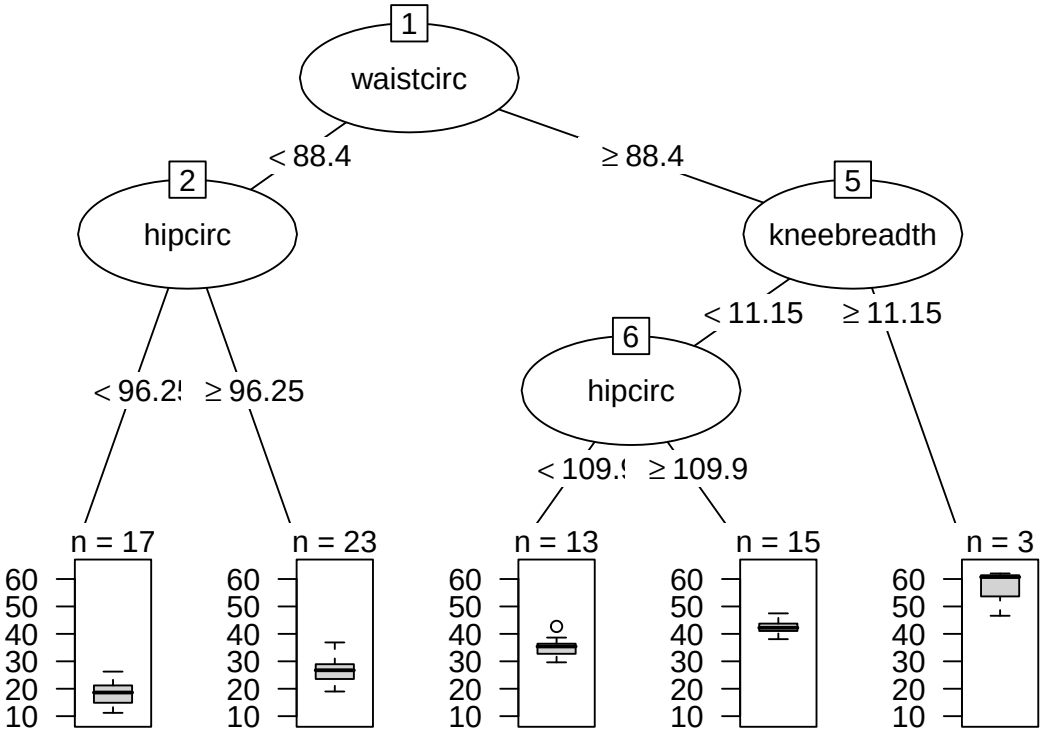
```
##                                       Linear      Median
## (Intercept)                     -0.34624093 -0.39510353
## seedingyes                      15.68293481  9.28416250
## time                            -0.04497427 -0.02682160
## seedingno:sne                    0.41981393  0.36860476
## seedingyes:sne                  -2.77737613 -1.33267160
## seedingno:cloudcover             0.38786207  0.20691306
## seedingyes:cloudcover           -0.09839285 -0.06071068
## seedingno:prewetness             4.10834188  5.22263667
## seedingyes:prewetness            1.55127493  2.01808261
## seedingno:echomotionstationary   3.15281358  2.13502276
## seedingyes:echomotionstationary  2.59059513  2.78255068
```

The linear model of the equation from chapter 6 shows a lower mean squared error than the median model. However, the median model does produce a lower mean absolute error and lower AIC than the linear model. This could be because the median model isn't taking the outliers into account as much as a mean weighted calculation would, but this could produce a higher error if the model shifts too far from the outliers to fit other values better.

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.





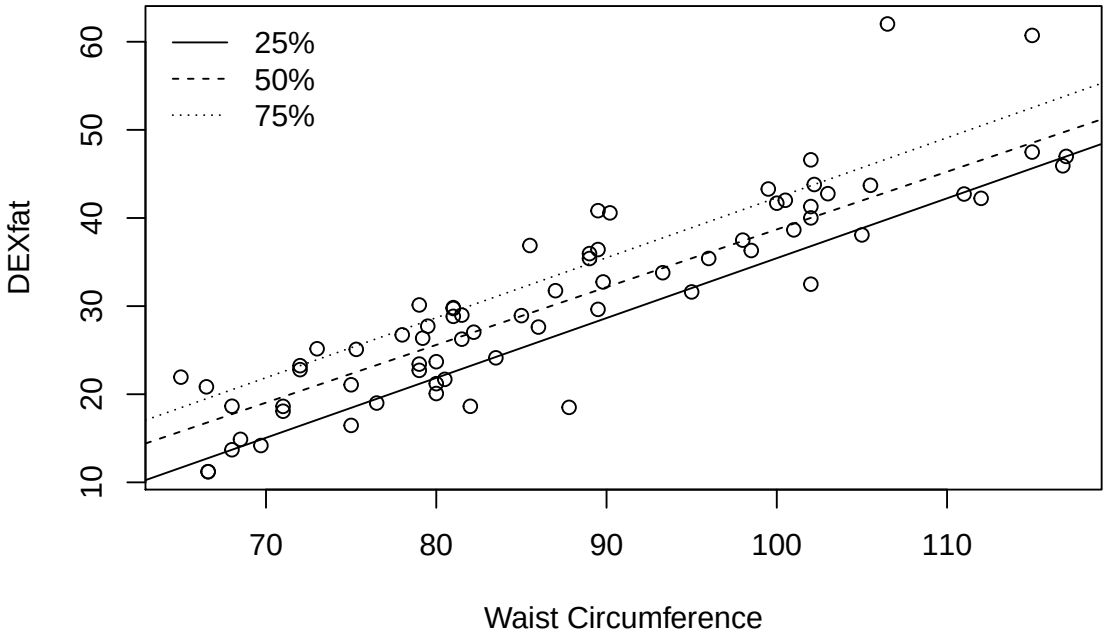

```
##     Unpruned.Tree Pruned.Tree Median.Reg
## MSE     10.170503   14.575003  15.024504
```

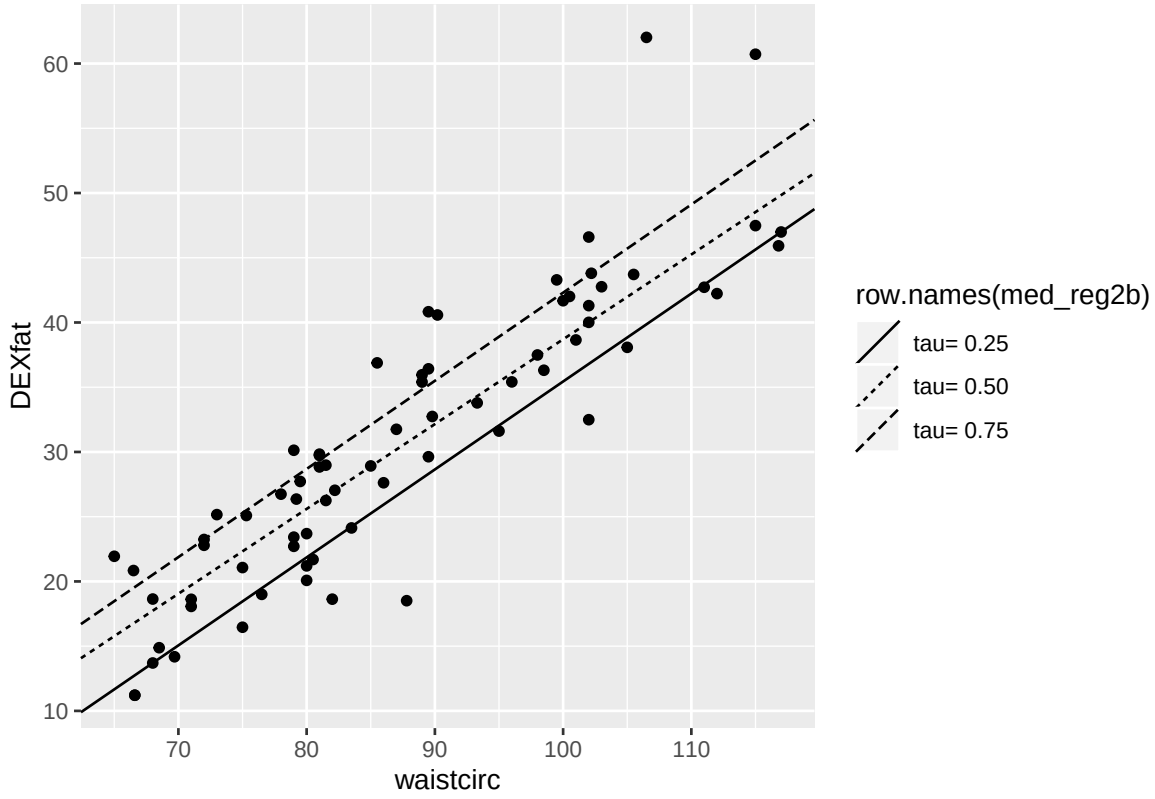
```
## MAE      2.476113    2.984743   2.947714
```

Comparing the regression tree's, both unpruned and pruned, shows that the unpruned tree has lower mean squared error in the predicted DEXfat. A median regression was also created and the mean squared error was calculated. The mean squared error of the median regression was greater than the pruned tree. The mean absolute error of the 3 models shows that the unpruned tree has the least amount of error followed by the median regression. The pruned tree's error is very close to the median regression error but higher.

b) Choose one independent variable. For the relationship between this variable and DEXfat, create linear regression quantile models for the 25%, 50% and 75% quantiles. Plot DEXfat vs that independent variable and plot the lines from the models on the graph.

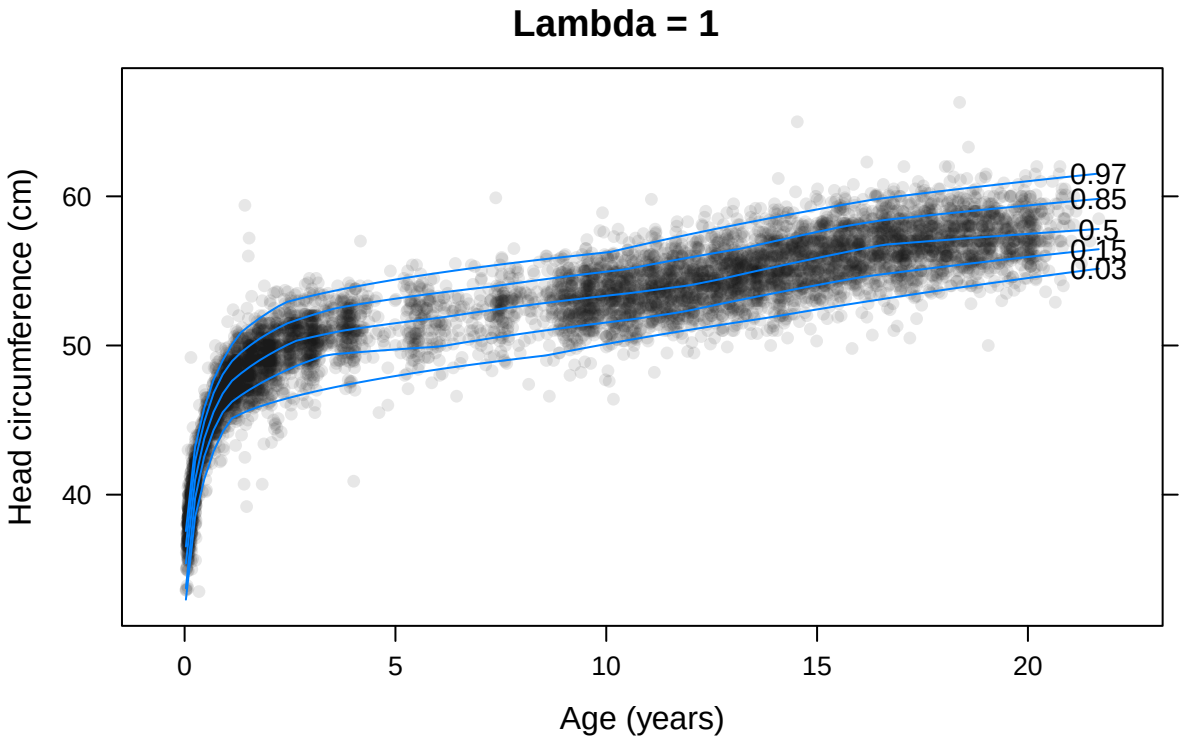
```
##
## Call: rq(formula = DEXfat ~ waistcirc, tau = c(0.25, 0.5, 0.75), data = bodyfat)
##
## tau: [1] 0.25
##
## Coefficients:
##             coefficients lower bd  upper bd
## (Intercept) -32.49837    -37.57797 -25.65679
## waistcirc     0.67939      0.62247   0.74718
##
## Call: rq(formula = DEXfat ~ waistcirc, tau = c(0.25, 0.5, 0.75), data = bodyfat)
##
## tau: [1] 0.5
##
## Coefficients:
##             coefficients lower bd  upper bd
## (Intercept) -26.80515    -36.03078 -16.69178
## waistcirc     0.65505      0.55058   0.77269
##
## Call: rq(formula = DEXfat ~ waistcirc, tau = c(0.25, 0.5, 0.75), data = bodyfat)
##
## tau: [1] 0.75
##
## Coefficients:
##             coefficients lower bd  upper bd
## (Intercept) -25.77722    -37.33159 -18.53524
## waistcirc     0.68079      0.59051   0.82469
```

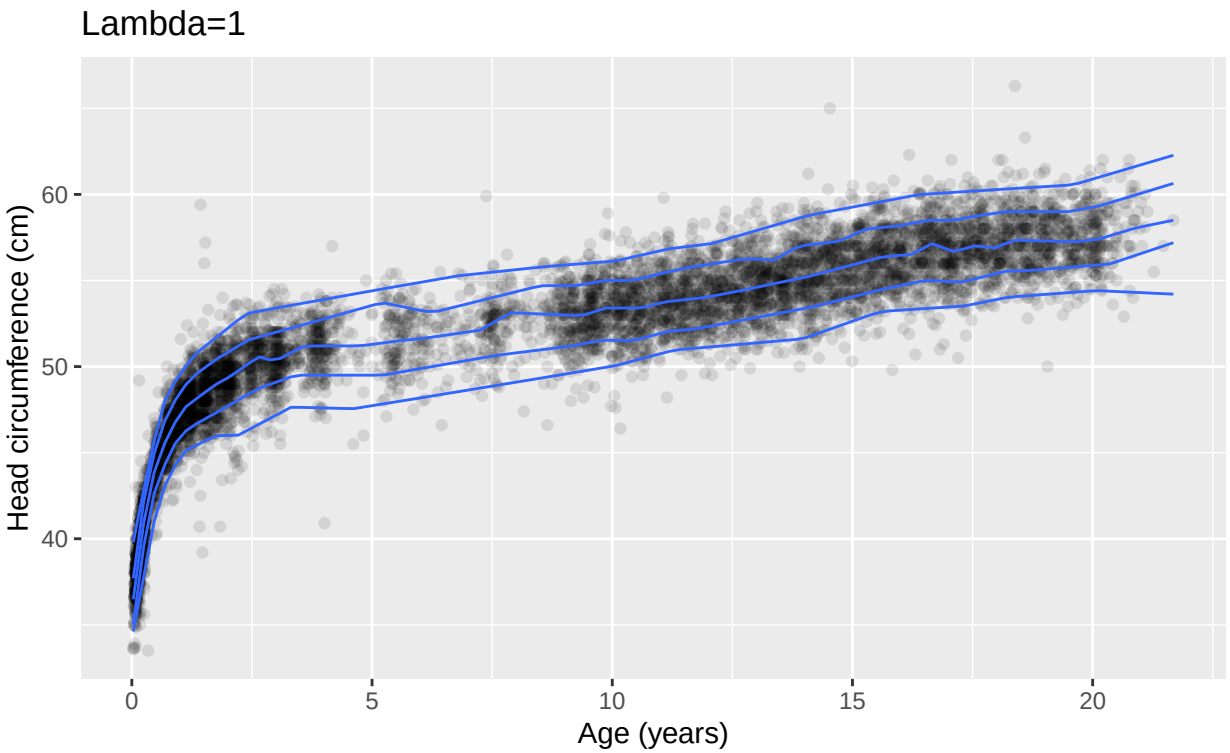


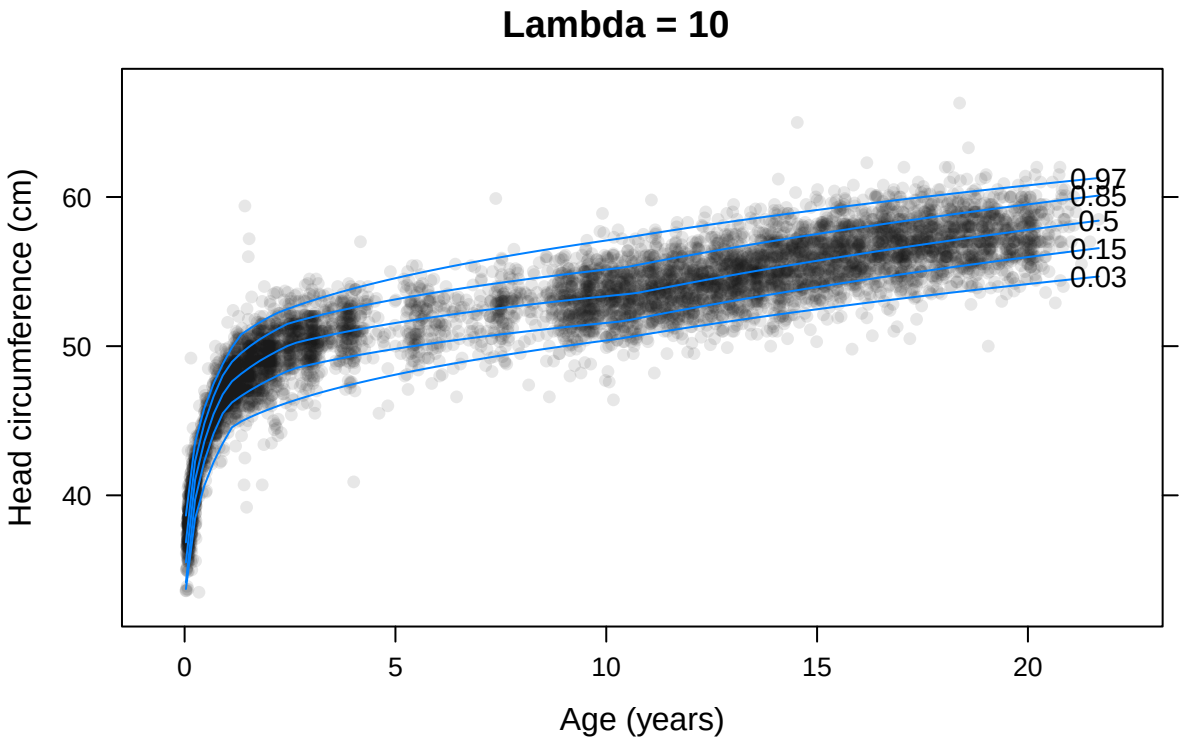


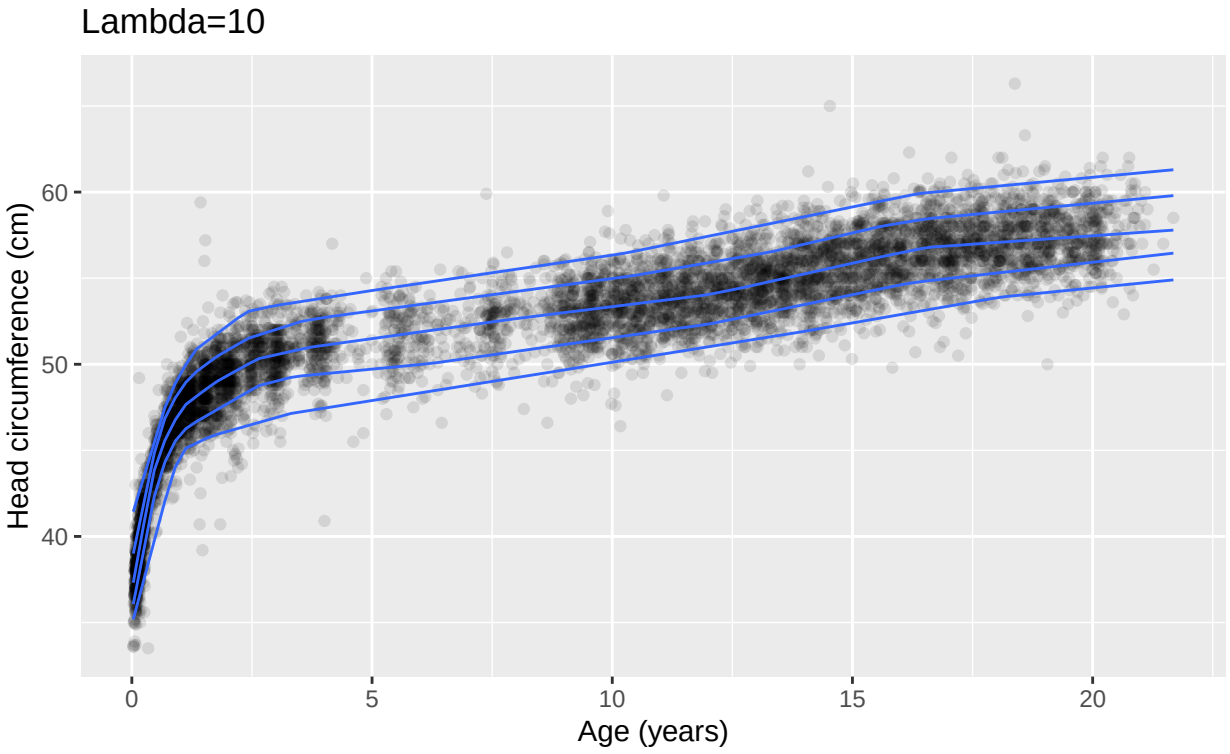
3. Consider {**db**} data from the lecture notes (package {**gamlss.data**}). Refit the additive quantile regression models presented ({**rqssmod**}) with varying values of λ (lambda) in {**qss**}. How do the

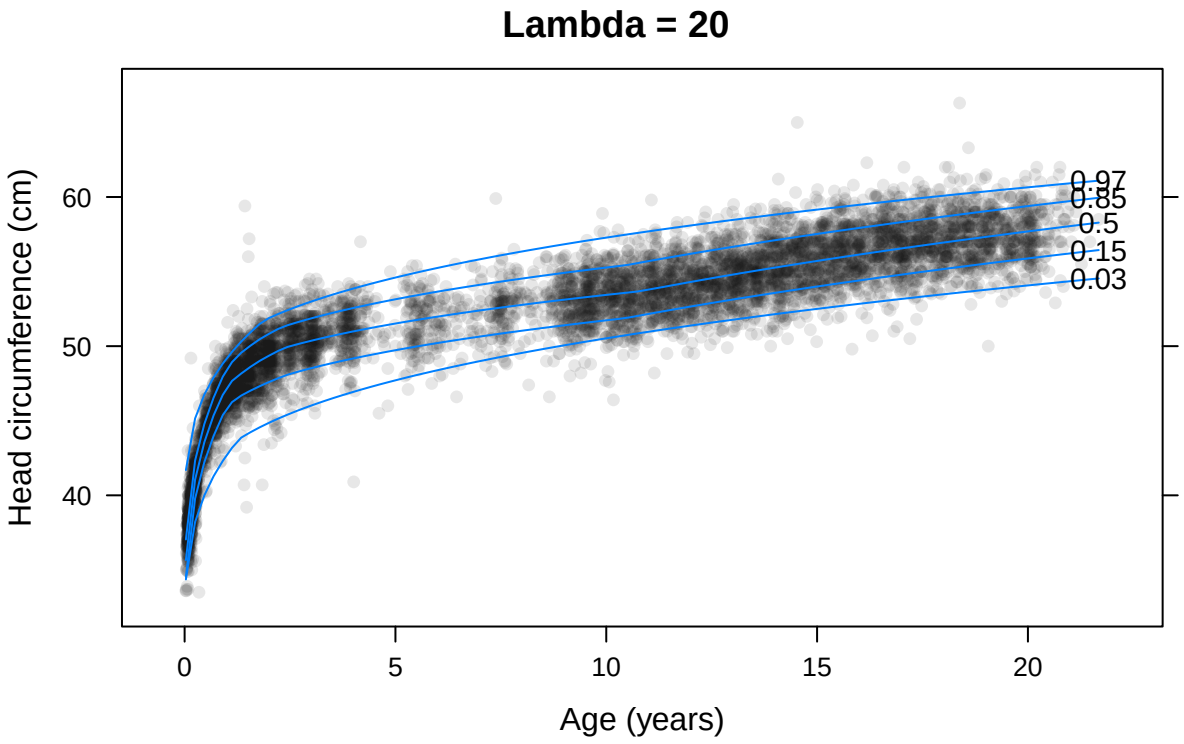
estimated quantile curves change?

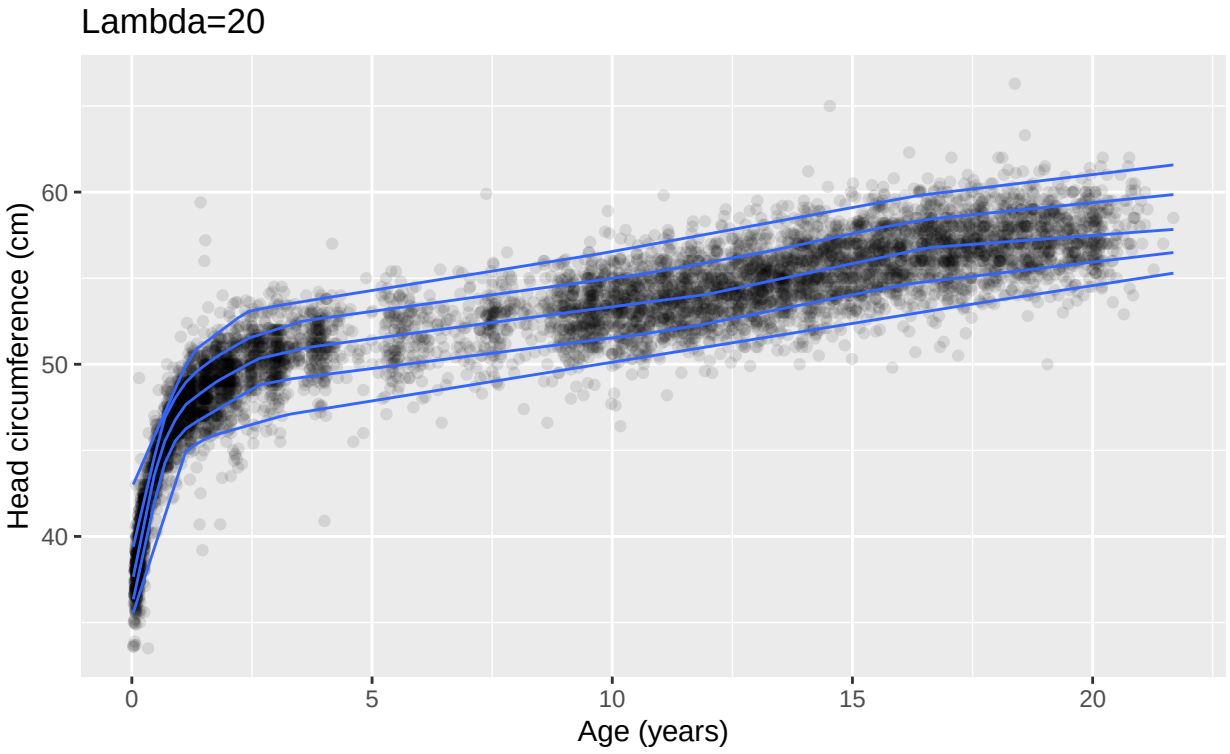


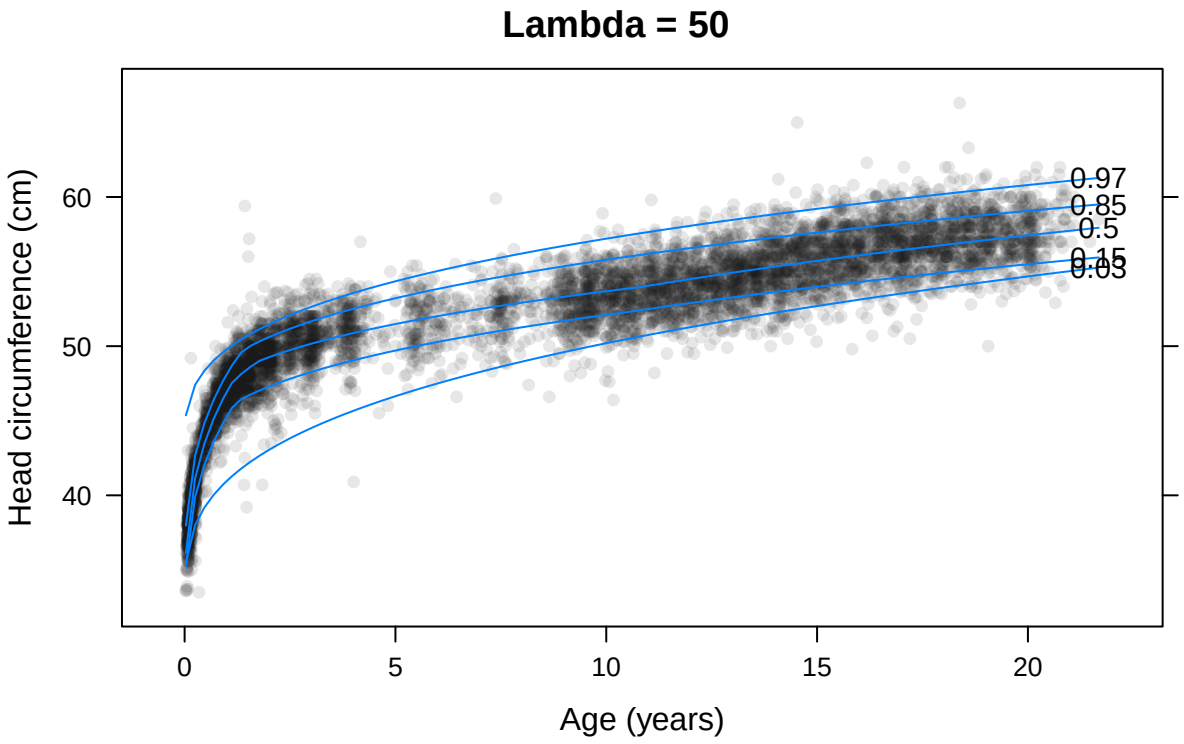


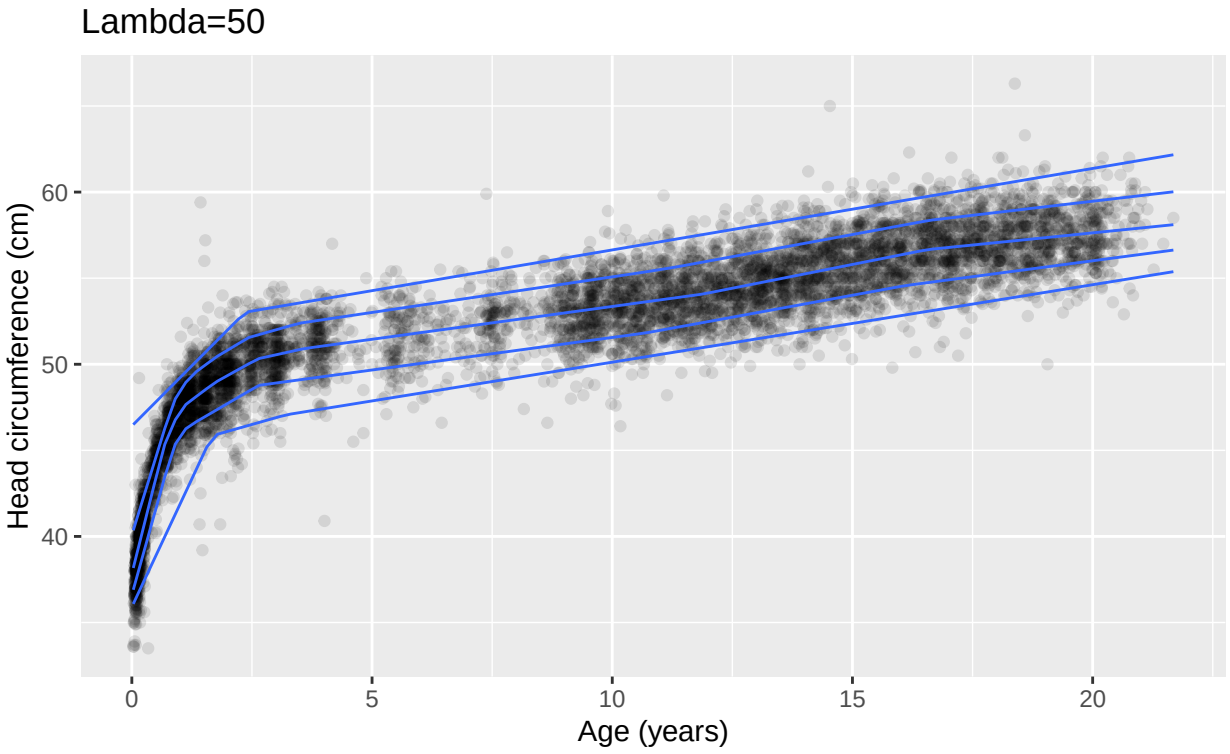

Lambda = 10
Head circumference (cm)
60
50
40
0
5
10
15
20
Age (years)
0.97
0.85
0.5
0.15
0.03

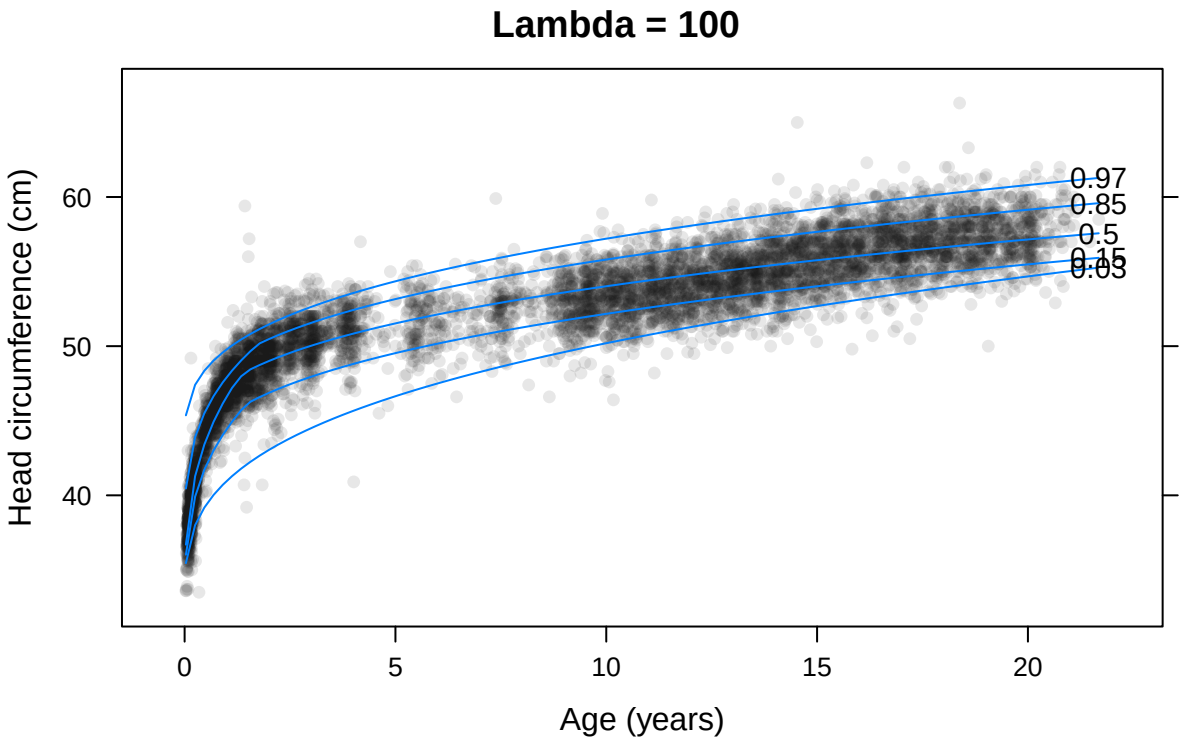

Lambda=10
Head circumference (cm)
60
50
40
0
5
10
15
20
Age (years)

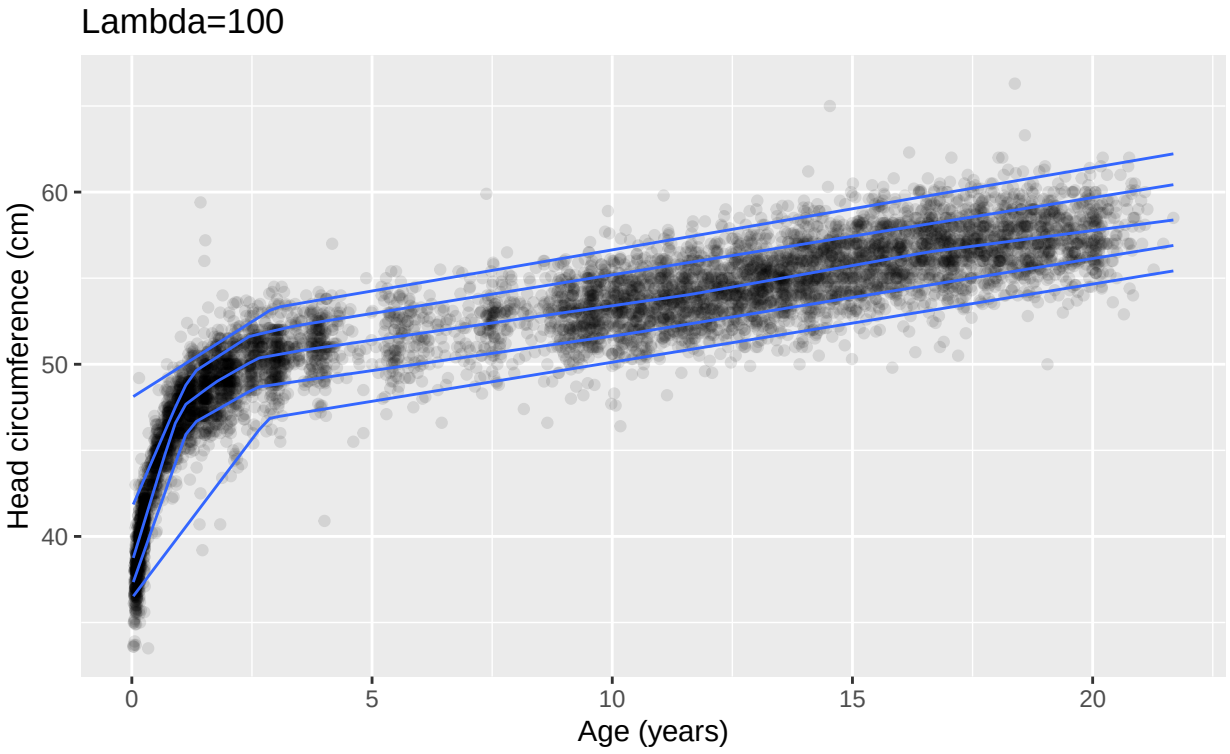

Lambda = 20
Head circumference (cm)
60
50
40
0
5
10
15
20
Age (years)
0.97
0.85
0.5
0.15
0.03


Lambda=20
Head circumference (cm)
60
50
40
0
5
10
15
20
Age (years)


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


Lambda=50
Head circumference (cm)
60
50
40
0
5
10
15
20
Age (years)





An increase in lambda, which is the penalty for the smoothness as said in lecture, smooths the quantile curves out more and more as it increases to 100 from 1. The 0.03 line also starts getting further away from the body of the data between 0 and 5 years of age as the function tries to smooth it out. A lambda of 1 which was generated in lecture keeps the lines tightly around the data where there are sharp curves between 0 and 5, and also has more 'wiggling'

from about 5 to the end. An increase in lambda causes smoother curves.

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

Using quantiles while displaying the data can provide the viewer with a much better understanding the data as a whole rather than only using the mean. The use of quantiles (percentiles or fractiles) allows the sampling groups to be split into groups based on the sample values. For example, a percentile of 0.25 will split the data into 2 groups, one of 25% and another with 75% of the data. Figures like boxplots and scatter plots with growth curves offer more transparent observations of data dispersion than do that of bar plots. There are other plots that utilize quantiles for displaying data which were talked about throughout the article like Quantile Engel Curves and quantile regressions.

Motivation

Quanltile regressions are more appropriate for data with dispersion and outliers because outlying data may pull averages away form the 50th percentile. Quantiles via optimizaiton shows how the minimizaiton of the sum of absolute residuals is better in some cases than is the minimizing a sum of squared residuals for means. Koenker and Hallock show how the least squares estimates can change the regression line over a median regression by only a few data points that may potentially be outliers and density of low household income and high food expenditures (mean line falls below a large portion of these points).

Case Study

The case study that was considered in this article was about using quantile regression and dertminants of infant birthweights conducted by Detailed Natality Data in 1997, which had investigated by Abrevaya in 2001. This was because most of the birthweight analyses are done using the least squares approach and not the least absolute residual values which, as we saw in the previous section, causes issues when working with data that have a high concentration towards one side. Along with the weight of the babies (g), several other metrics were recorded like race, mothers education, prenatal medical care, how much the mother smoked during pregnancy, mothers weight gain during pregnancy, and the age of the baby. If there were any data point metrics missing that baby was removed from the analysis.

Findings

The researchers found that boys are about 100 grams heavier than girls on average, but looking at the quantiles boys are much heavier than girls closer to the upper end of the distribution (about 130 - 140 grams at the 0.95 quantile). They also mention that the difference between black and white mothers is also relatively high. Looking at the graph the average difference shows that babies of black mothers are about 200 grams less than babies from white mothers. However, the babies on the lower side of the distribution close to 350 grams lighter than babies from white mothers while those babies from black mothers on the upper side of the distribution are about 160 grams lighter than babies from white mothers. They also note that there is little difference between sum of squared residuals and sum of absolute residuals for different education levels, and with the exception of education, all of the other plots show that the quantile regression falls outside of the 90% confidence interval at some point.

Resources Used:

- rdocumentation.org
- theanalysisfactor.com
- r-bloggers.com
- Quantile Regression - Roger Koenker and Kevin Hallock