Homework #8

Justin Robinette

October 16, 2018

No collaborators for any problem

Problem #1, Part A: Consider the **clouds** data from the **HSAUR3** package. Review the linear model fitted to this data in Chapter 6 of the text book and report model findings.

Results: Here I reported the results of the linear fitted model from Chapter 6. *Figure 1.1* shows the p-value of the model (**0.024**) is significant at an alpha of 0.05.

Figure 1.2 shows the influence the predictors have on predicting rainfall. As we would expect based on the formula provided, **seeding** (whether seeding has occurred) has the most statistically significant impact followed by **sne** (suitability criterion). Lastly, *Figure 1.3* and *Figure 1.4* show the betastar and standard error values, respectively, by variable.

Next, *Figure 1.5* examines the relationship between **rainfall** and **sne** by **seeding**, as the text did. Here we use both ggplot2 and base R to compare. Lastly, *Figure 1.6* recreates the text plot of residual values using ggplot and base R. We see the values are spread pretty evenly about 0.

Figure 1.1: Model P-Value

P-Value
0.0243093

Figure 1.2: Variable P-Values in Model

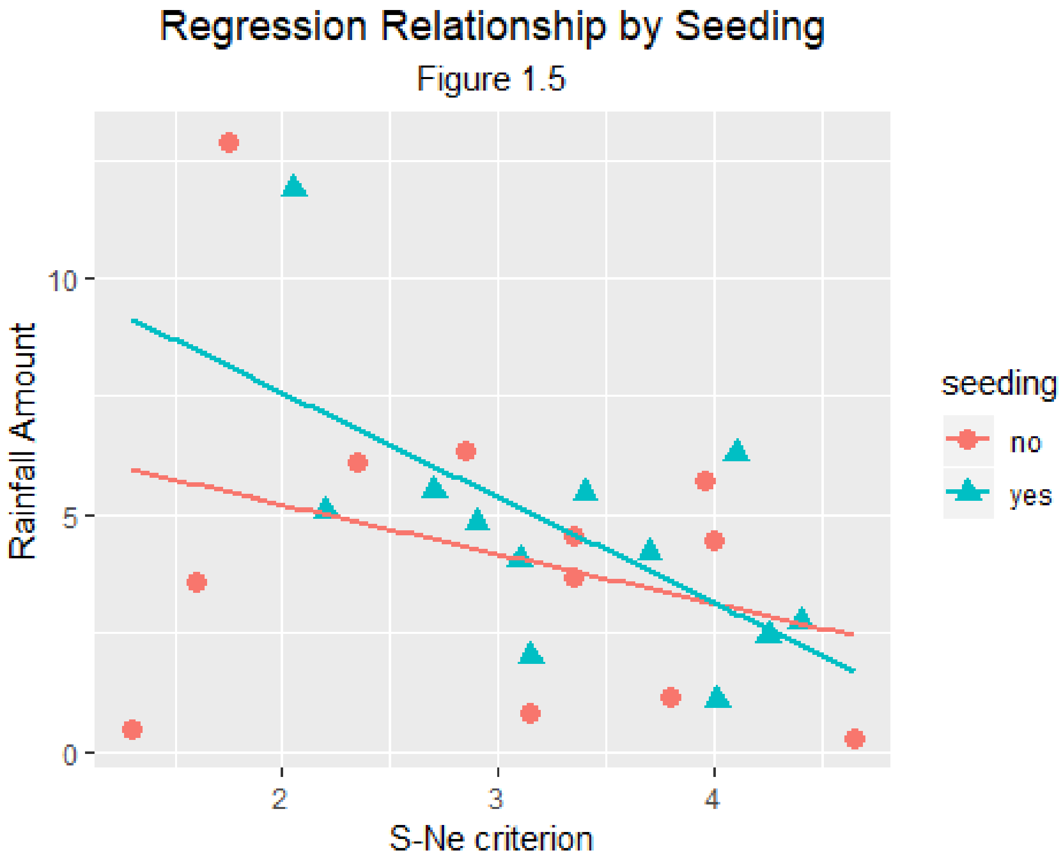
	P-Values
(Intercept)	**0.9030556**
seedingyes	**0.0037151**
time	**0.0958974**
seedingno:sne	**0.6274209**
seedingyes:sne	**0.0104042**
seedingno:cloudcover	**0.0983850**
seedingyes:cloudcover	**0.3885384**
seedingno:prewetness	**0.2744991**
seedingyes:prewetness	**0.5744082**
seedingno:echomotionstationary	**0.1267721**
seedingyes:echomotionstationary	**0.1775655**

Figure 1.3: Variable Beta_star Estimates

	Beta_star
(Intercept)	-0.3462409
seedingyes	15.6829348
time	-0.0449743
seedingno:sne	0.4198139
seedingyes:sne	-2.7773761
seedingno:cloudcover	0.3878621
seedingyes:cloudcover	-0.0983928
seedingno:prewetness	4.1083419
seedingyes:prewetness	1.5512749
seedingno:echomotionstationary	3.1528136
seedingyes:echomotionstationary	2.5905951

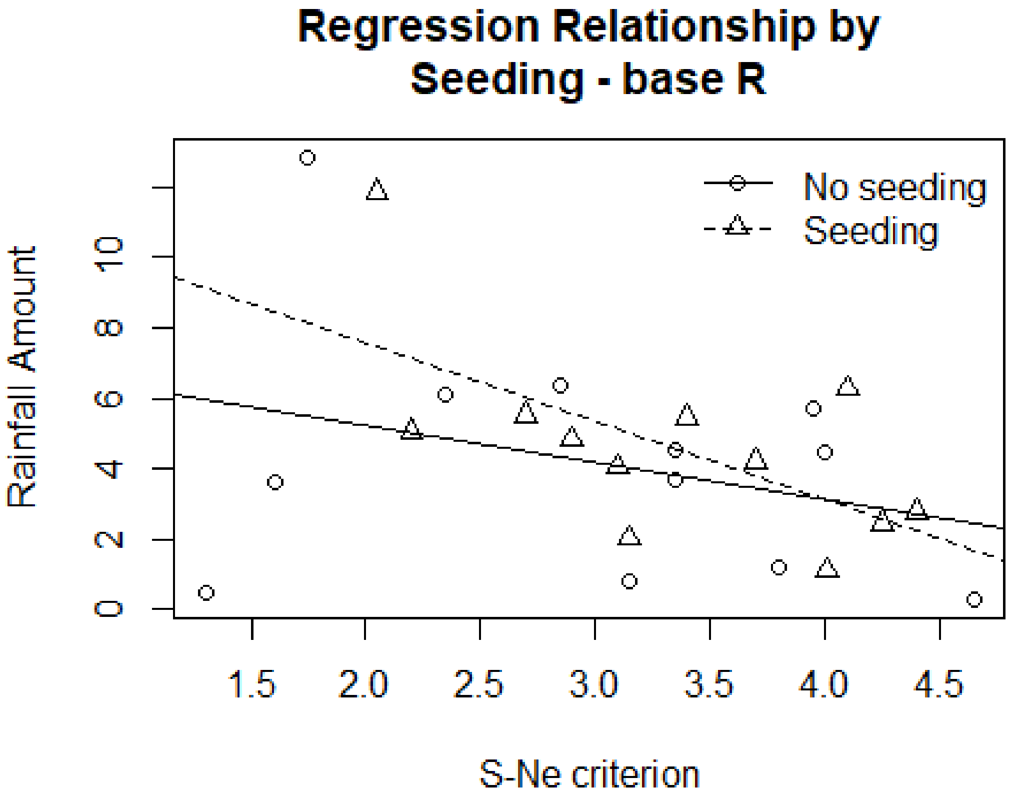
Figure 1.4: Variable Standard Errors

	Standard Errors
(Intercept)	2.7877340
seedingyes	4.4462661
time	0.0250529
seedingno:sne	0.8445299
seedingyes:sne	0.9283701
seedingno:cloudcover	0.2178550
seedingyes:cloudcover	0.1102898
seedingno:prewetness	3.6010069
seedingyes:prewetness	2.6928731
seedingno:echomotionstationary	1.9325259
seedingyes:echomotionstationary	1.8172597

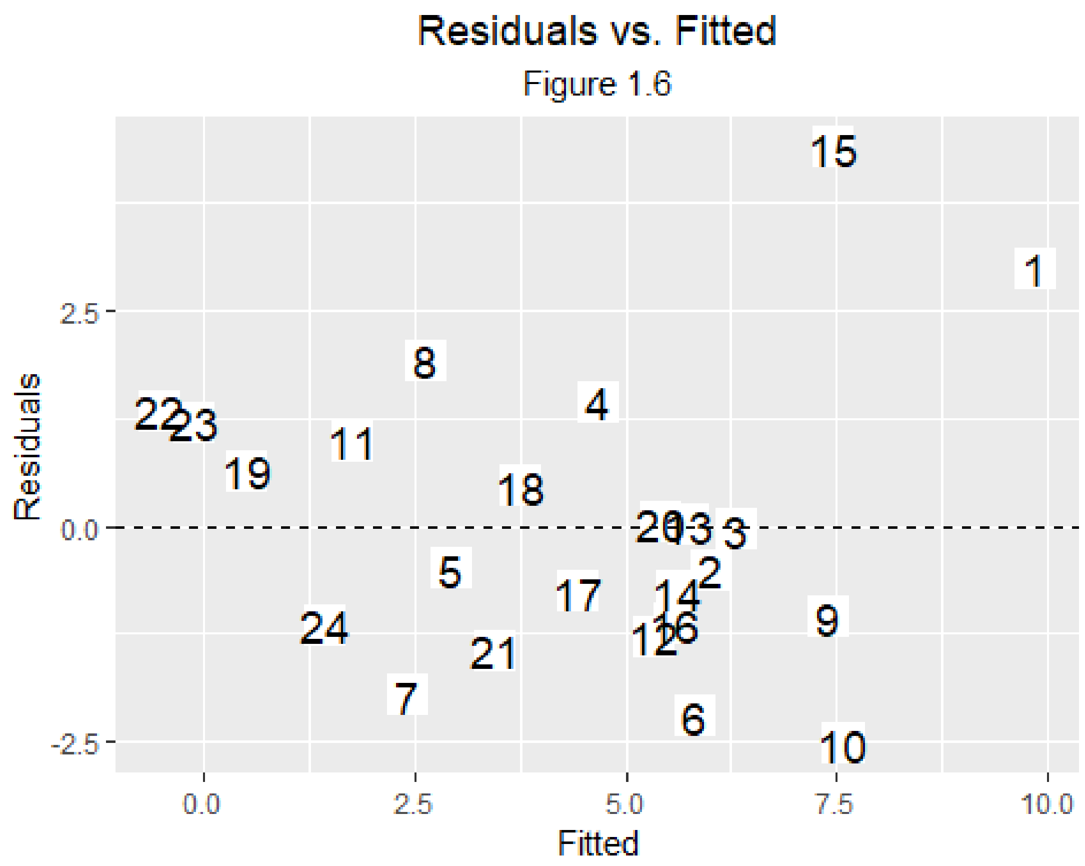


Regression Relationship by Seeding

Figure 1.5

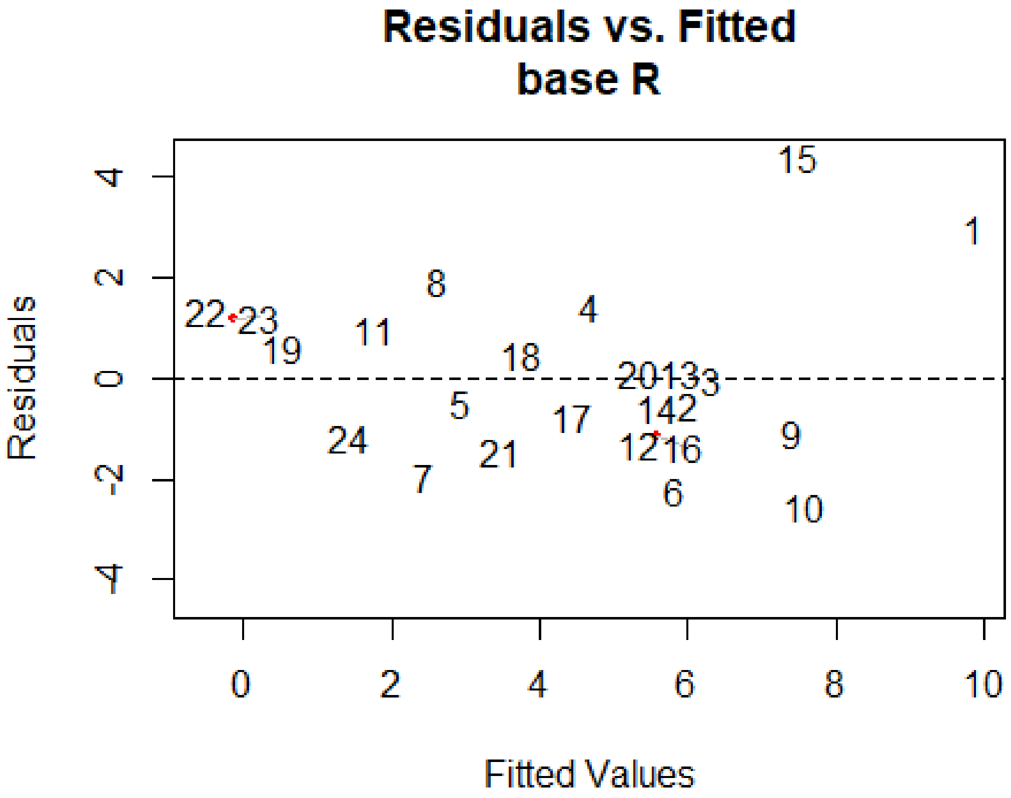


Regression Relationship by Seeding - base R



Residuals vs. Fitted

Figure 1.6



Residuals vs. Fitted
base R

Problem #1, Part B: Fit a median regression model.

Results: I fit a median regression model and printed the call.

```
## rq(formula = clouds_formula, tau = 0.5, data = clouds)
```

Problem #1, Part C: Compare the two results.

Results: *Figure 1.7* and *Figure 1.8* compare coefficient values from the two models. The first thing I notice is that the p-values increased across the board for each variable when going to the Median Regression method.

Figure 1.9 is included for comparison and recreates *Figure 1.5* above showing the relationship between **rainfall** and **sne** by **seeding**. *Figure 1.10* shows the relationships of the same variables using the median regression method. The most interesting part, to me, is that the slope of the line when **seeding** is absent has went from negative to positive. It appears that the median regression lines ignore the outliers more than the linear approach.

Figure 1.10 shows the mean square error values by seeding factor between each method. The linear method does a much better job when there is no seeding. The errors are closer between the two models when seeding does occur.

Figure 1.7: Comparison of Variable P-Values

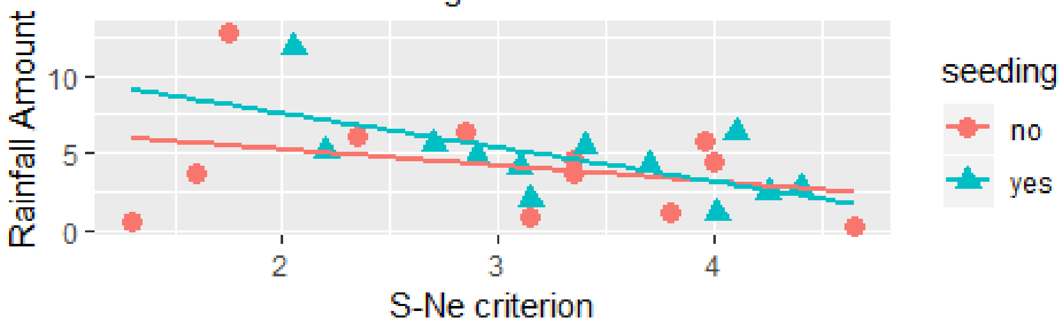
	Linear P-Values	Median Regression P-Values
(Intercept)	**0.9030556**	**0.9117543**
seedingyes	**0.0037151**	**0.6799547**
time	**0.0958974**	**0.9499419**
seedingno:sne	**0.6274209**	**0.9660231**
seedingyes:sne	**0.0104042**	**0.9497113**
seedingno:cloudcover	**0.0983850**	**0.6970934**
seedingyes:cloudcover	**0.3885384**	**0.9844601**
seedingno:prewetness	**0.2744991**	**0.7771197**
seedingyes:prewetness	**0.5744082**	**0.9956736**
seedingno:echomotionstationary	**0.1267721**	**0.1869136**
seedingyes:echomotionstationary	**0.1775655**	**0.5743232**

Figure 1.8: Comparison of Variable Beta_star

	Linear Beta_star	Median Regression Beta_star
(Intercept)	-0.3462409	-0.3951035
seedingyes	15.6829348	9.2841625
time	-0.0449743	-0.0268216
seedingno:sne	0.4198139	0.3686048
seedingyes:sne	-2.7773761	-1.3326716
seedingno:cloudcover	0.3878621	0.2069131
seedingyes:cloudcover	-0.0983928	-0.0607107
seedingno:prewetness	4.1083419	5.2226367
seedingyes:prewetness	1.5512749	2.0180826
seedingno:echomotionstationary	3.1528136	2.1350228
seedingyes:echomotionstationary	2.5905951	2.7825507

Linear Regression Relationship by Seeding

Figure 1.9



Median Regression Relationship by Seeding

Figure 1.10

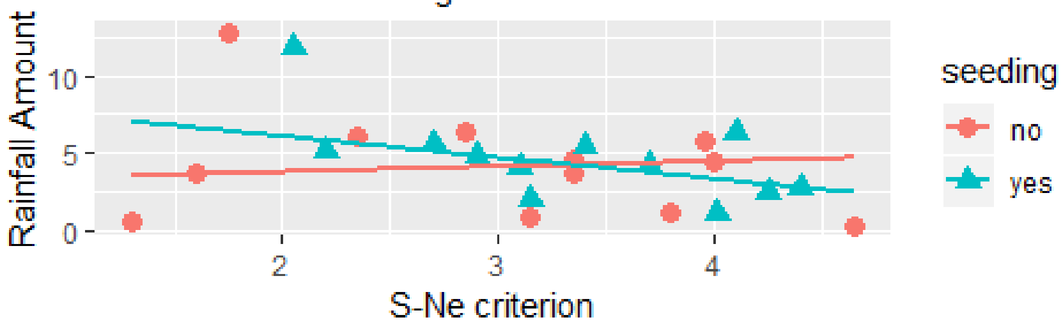


Figure 1.11: MSE Comparison

Linear Model	MSE	Median Regression Model	MSE
Seeding	**11.726147**	**Seeding**	**10.1447**
No Seeding	**6.407436**	**No Seeding**	**10.8319**

Problem #2, Part A: Reanalyze the **bodyfat** data from the **TH.data** package.

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

Results: In this exercise, I compare the regression tree approach to the median regression method to see which is a better predictor of **bodyfat** based on the data. I used the model from page 175 of the text book to create my decision tree and then utilized the same predictors in the median regression model for comparison.

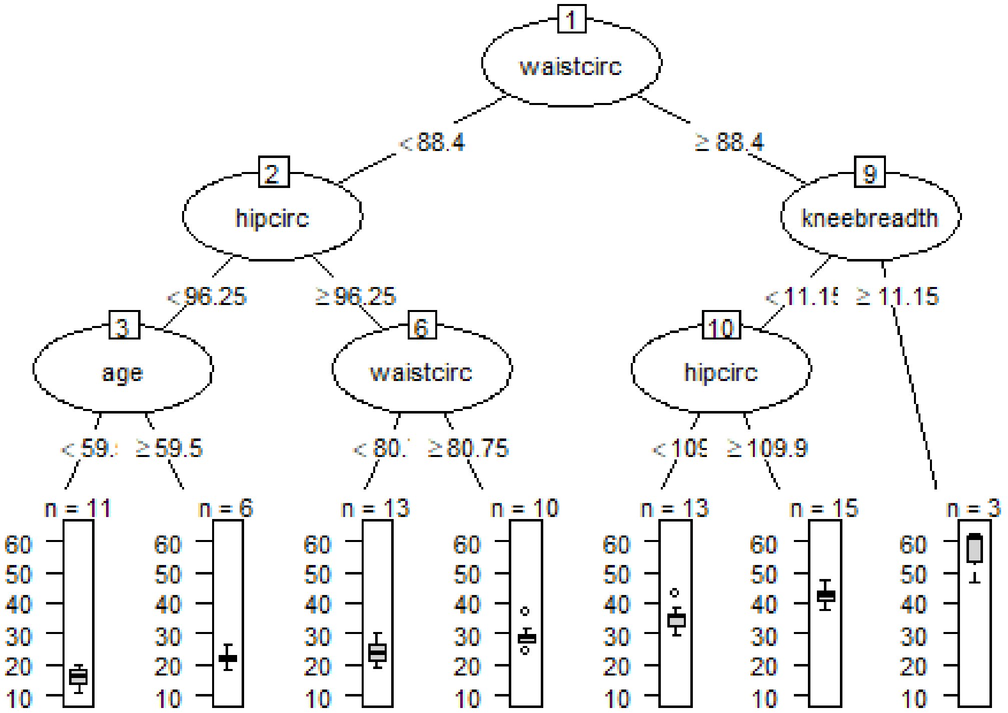
Figure 2.1 shows the tree prior to pruning. We see that **waistcirc** is the root node splitting at 88.4. Pruning technique was used but, as I show in *Figure 2.2*, pruning was not needed for this model.

Figure 2.3 shows the median regression model that was fit for comparing with the decision tree method. I used *Figure 2.4* to show the p-values of the various predictors within the model. Hip and Waist size are the best predictors, according to this chart.

I then used ggplot to show the relationship between bodyfat and hip size (*Figure 2.5*) and bodyfat and waist size (*Figure 2.6*). I used the opposite predictor variable to denote the color of the plot point. If the observation contains a value above the split point of the decision tree, it is a blue triangle. If the observation is below the split point, it is a red circle.

Lastly, *Figure 2.7* shows a comparison of the two methods' Mean Square Error values. As we see, with this dataset, the decision tree method has a better error rate than the median regression method.

Figure 2.1: Original Tree



Original Tree - ggplot

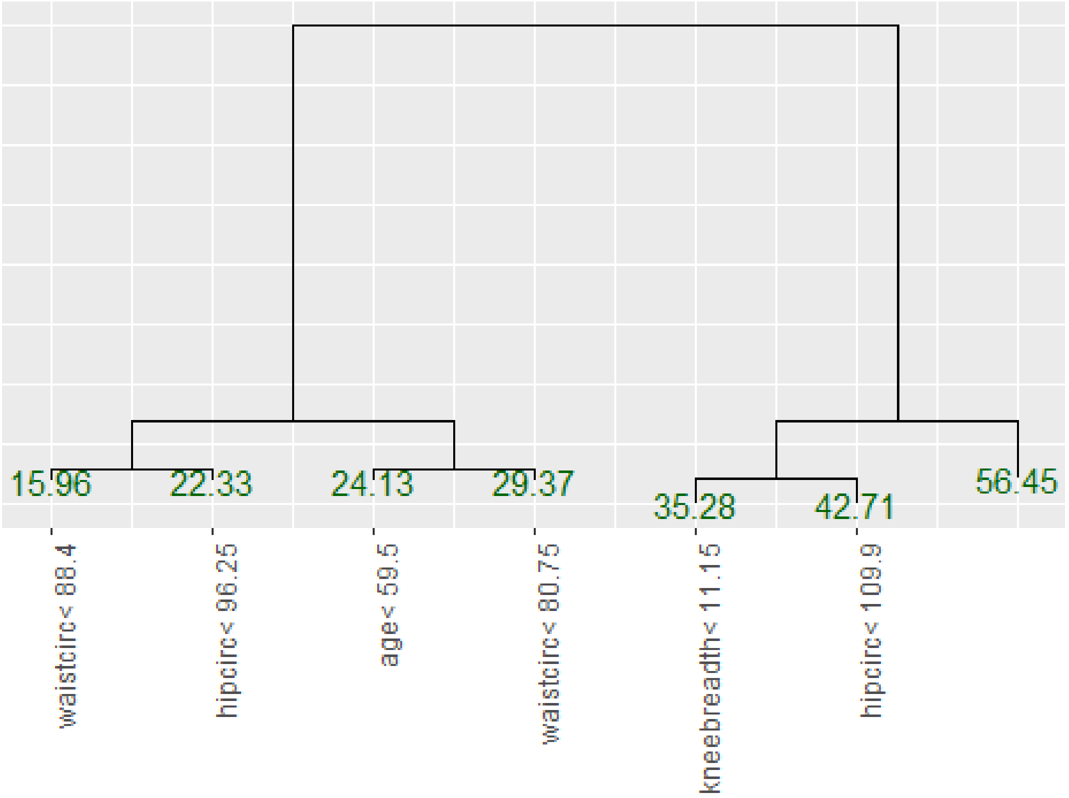
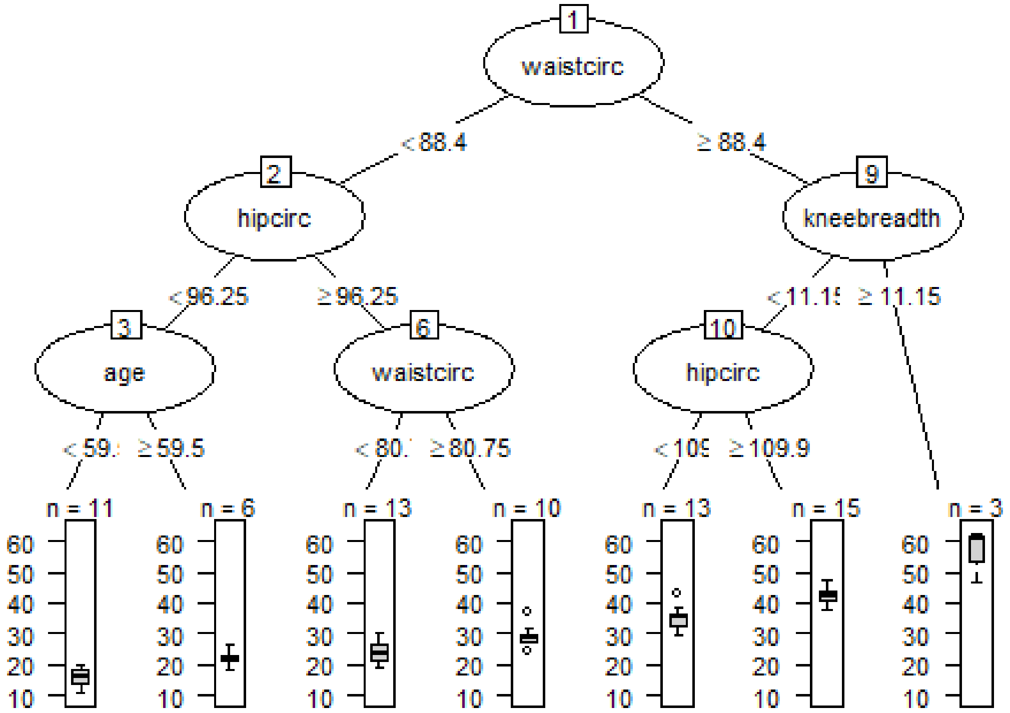
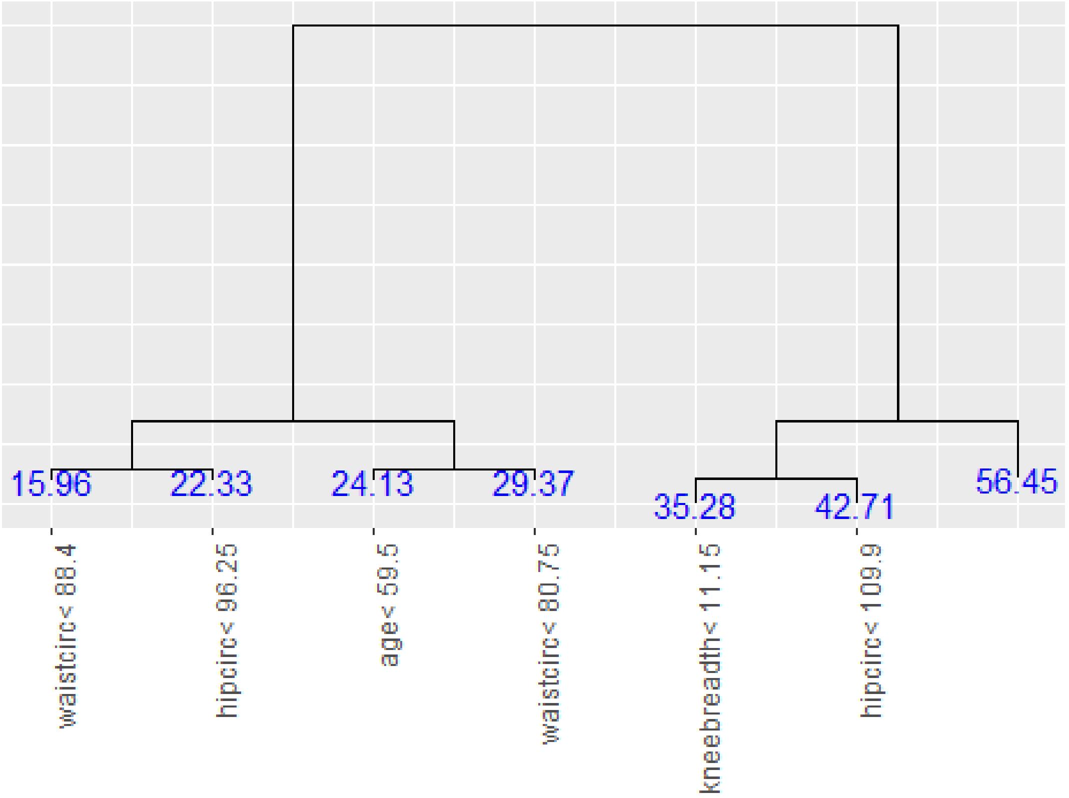


Figure 2.2: Pruned Tree



Pruned Tree - ggplot



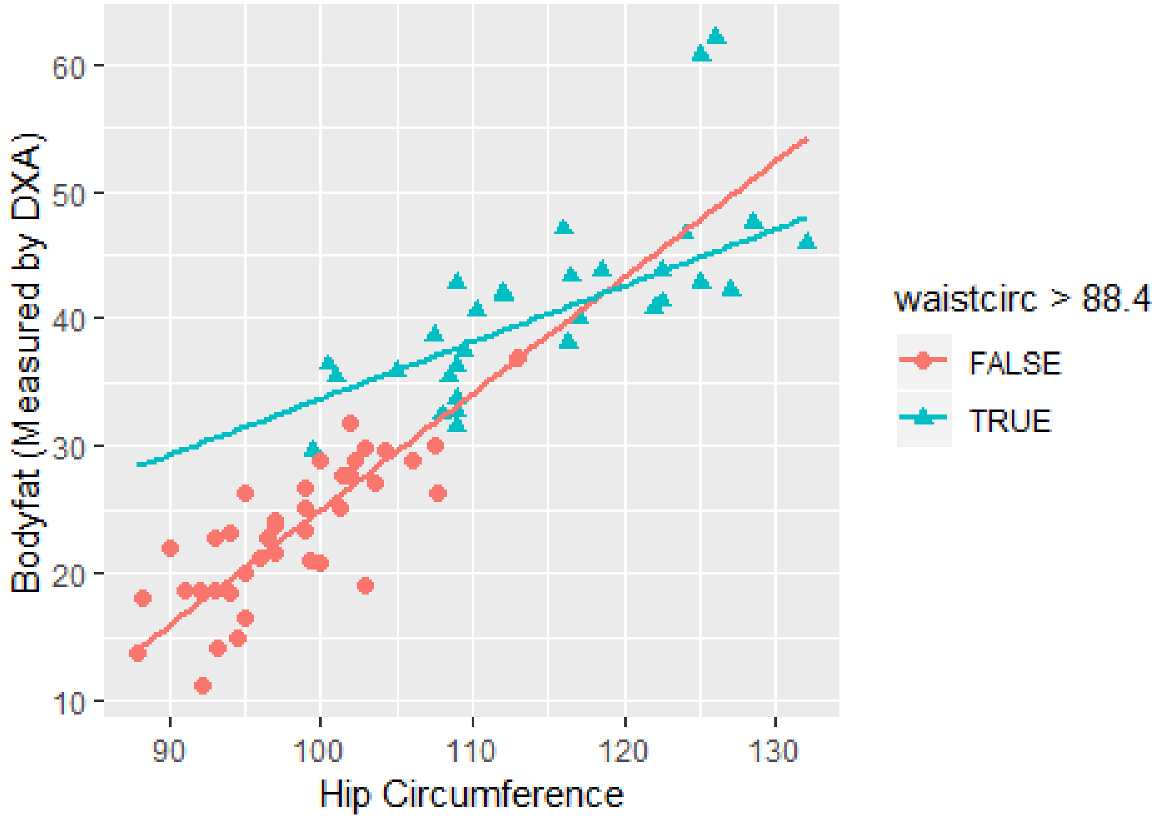
```
## [1] "Figure 2.3: Model Call"

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

Figure 2.4: P-Values by Variable

	P-Values
(Intercept)	**0.0000002**
age	**0.1300021**
waistcirc	**0.0035498**
hipcirc	**0.0000342**
elbowbreadth	**0.9244546**
kneebreadth	**0.4139203**

Bodyfat Relationship with Hip Size by Waist Size

Figure 2.5



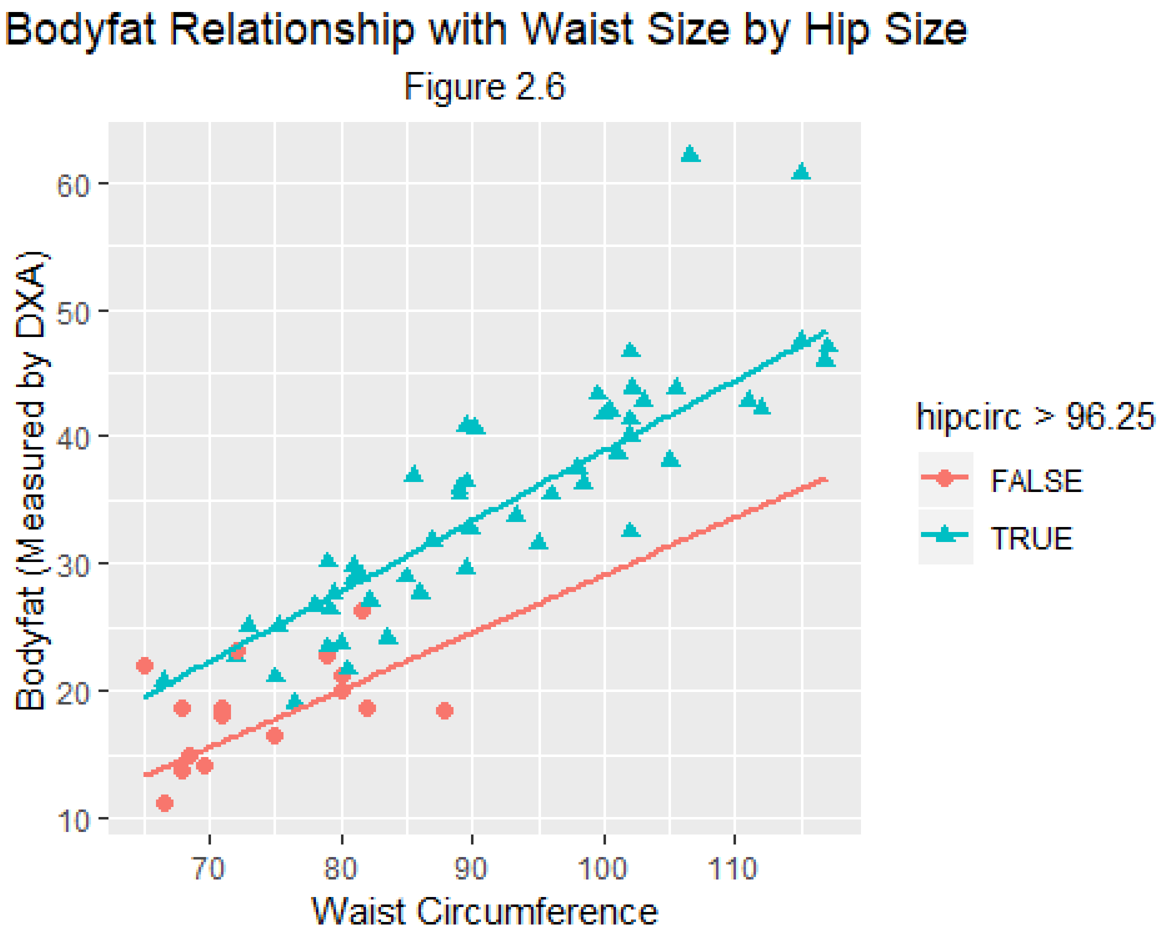


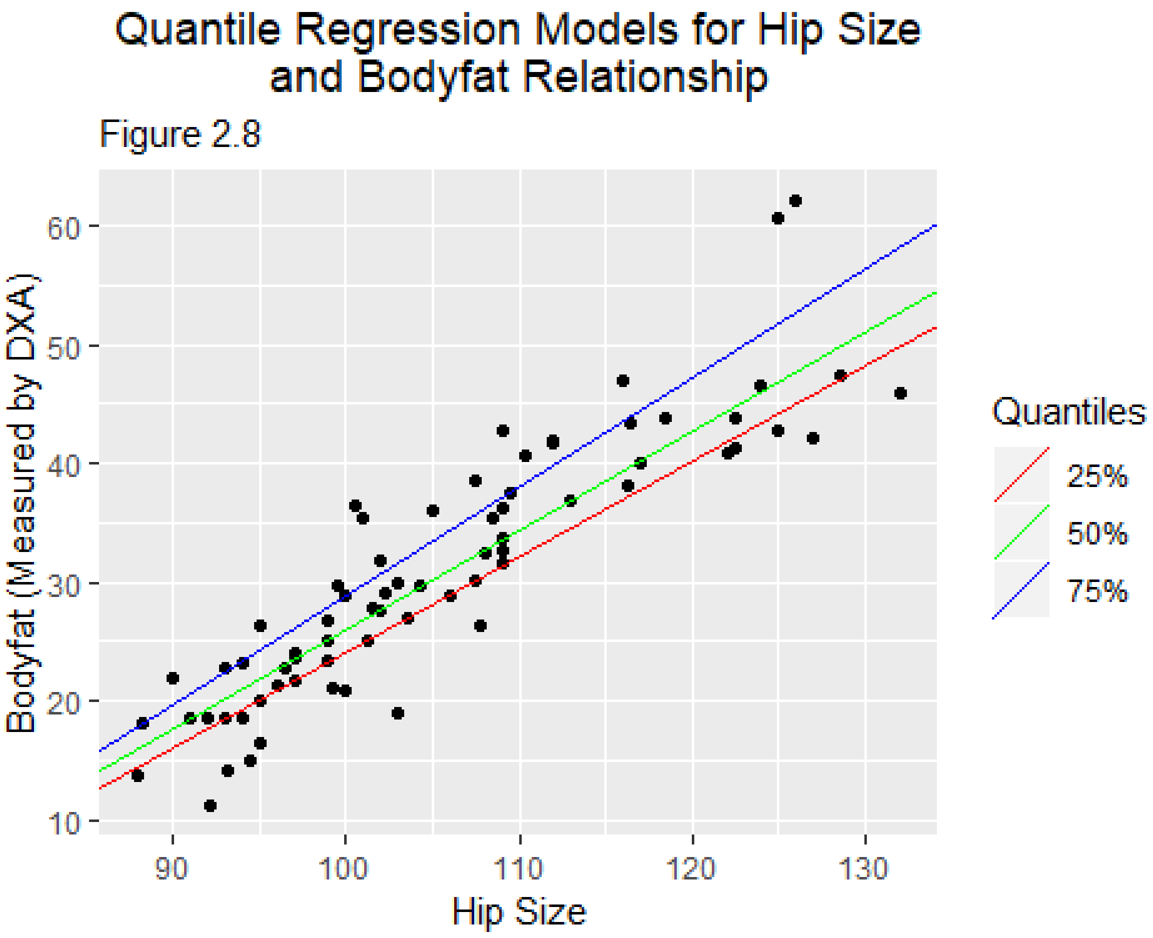
Figure 2.7: Error Rate of Bodyfat Predictors

Regression Tree MSE	Median Regression MSE
10.1705	**15.0245**

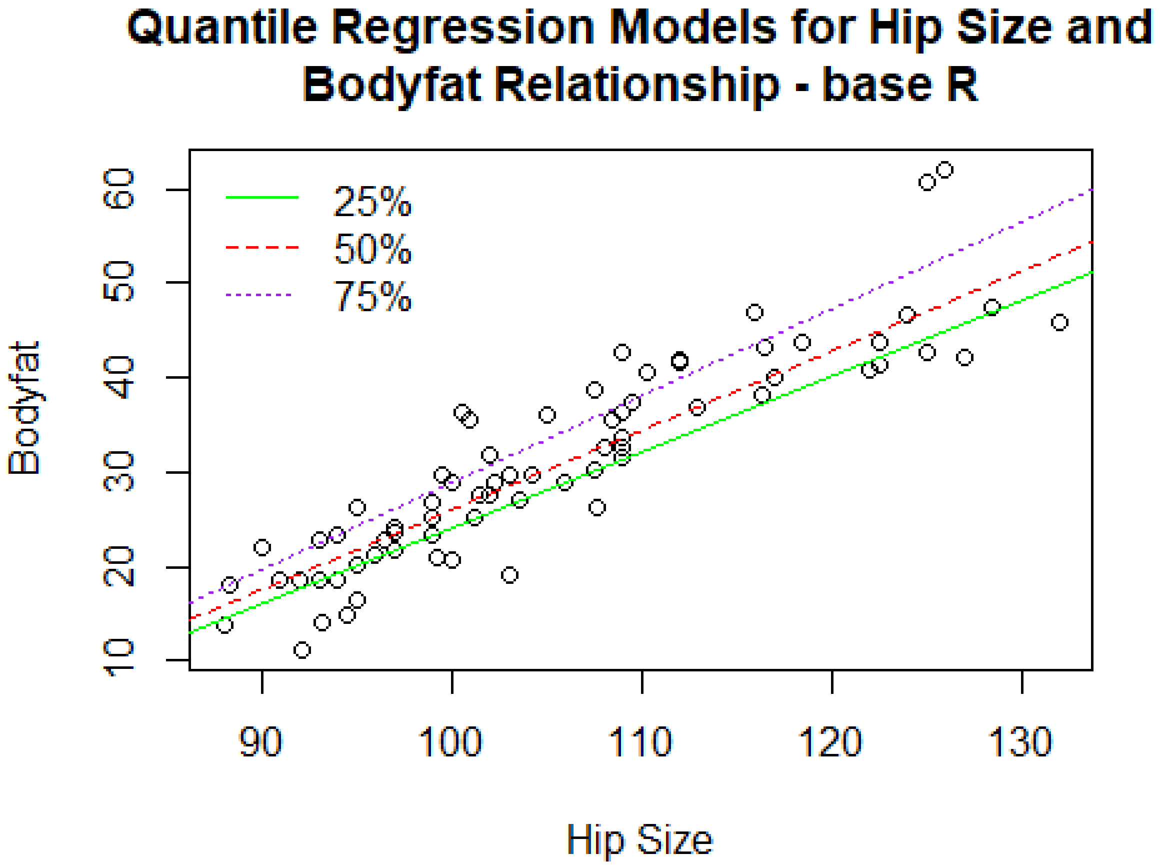
Problem #2, Part B: Choose one independent variable. For the relationship between this variable and DEXfat, create linear regression quantile models for the 25%, 50%, and 75% quantiles. Plot DEXfat vs that independent variable and plot the lines from the models on the graph.

Results: For this exercise, I chose **hipcirc** as my independent variable because, as I showed in *Figure 2.4*, this variable has the lowest p-value inside of the model that was created.

I used the quantile values of 25%, 50% and 75% as lines on the plotted relationship between **bodyfat** and **hipcirc** in *Figure 2.8*. We see the 50% quantile is centered best among the observations. *An analogous base R plot is included per assignment instructions.*



Quantile Regression Models for Hip Size
and Bodyfat Relationship
Figure 2.8
Bodyfat (Measured by DXA)
60
50
40
30
20
10
90
100
110
120
130
Hip Size
Quantiles
25%
50%
75%

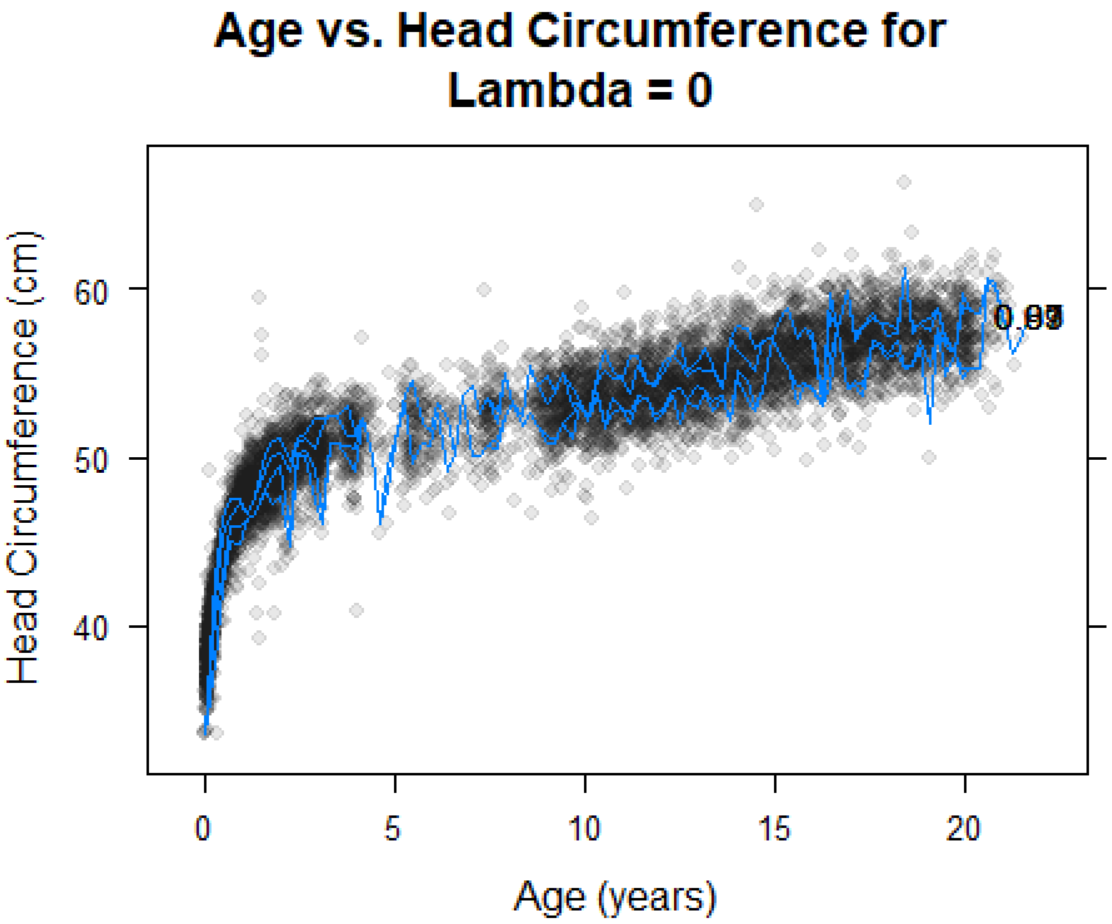


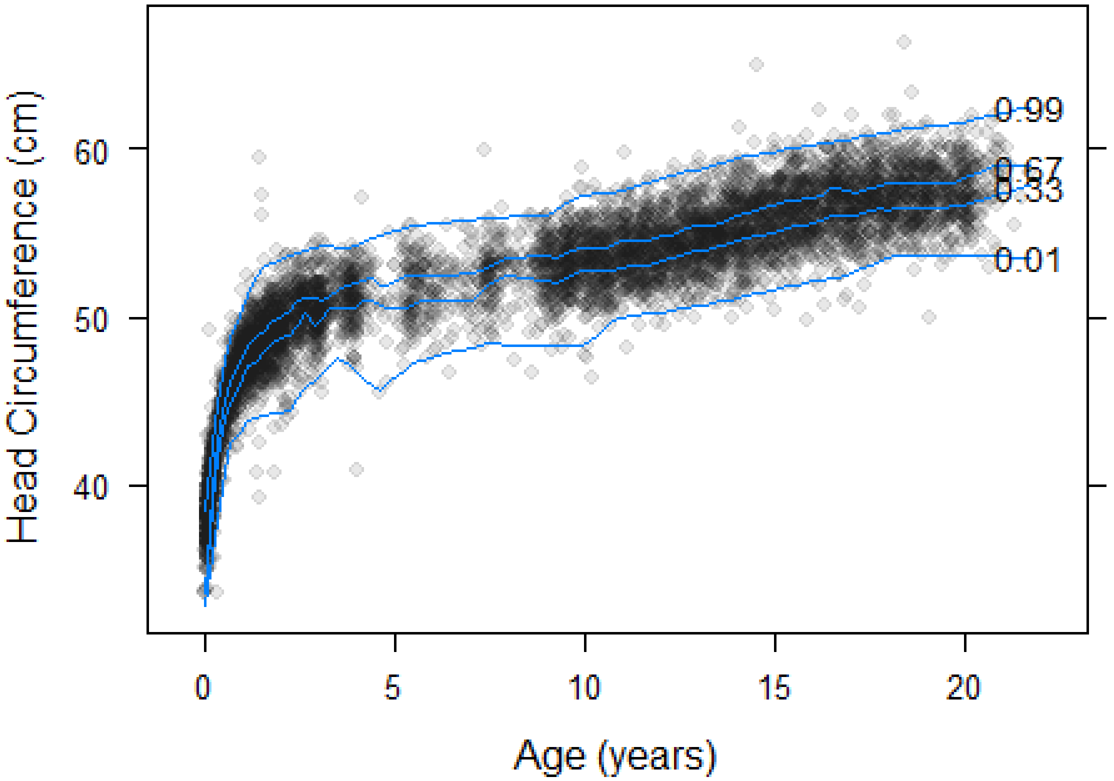
Quantile Regression Models for Hip Size and
Bodyfat Relationship - base R
25%
50%
75%
Bodyfat
60
50
40
30
20
10
90
100
110
120
130
Hip Size

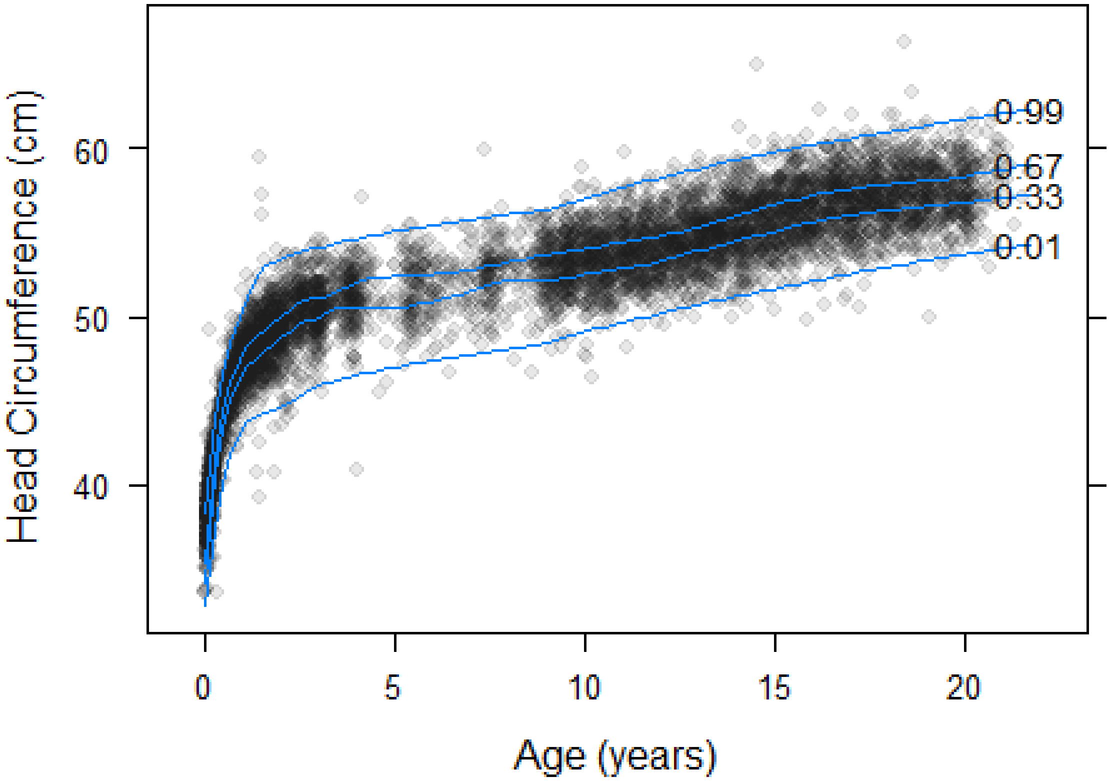
Problem #3: Consider **db** data from the lecture notes (package **gamlss.data**). Refit the additive quantile regression models presented (**rqssmod**) with varying values of lambda in **qss**. How do the estimated quantile curves change?

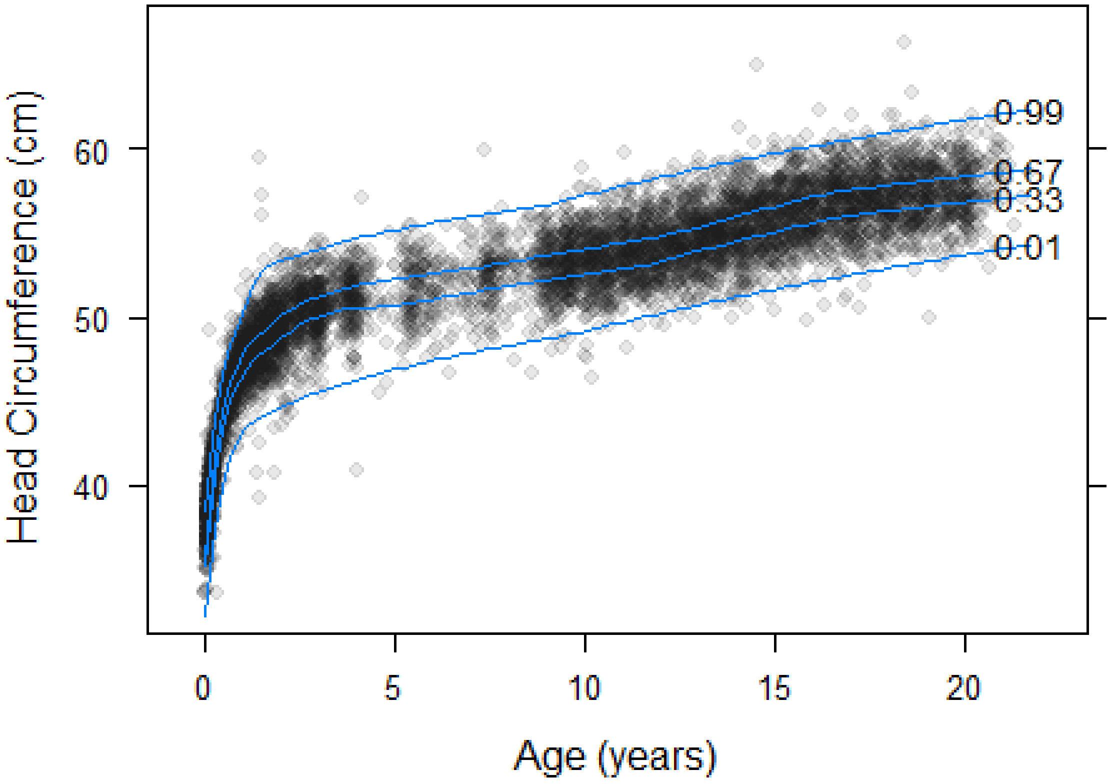
Results: Here I took the lecture notes provided, used my own tau values, and created a function that accepts **lambda** and returns the xyplots.

The shrinkage factor allows the quantile lines to smooth out the larger the **lambda** value. At lambda = 0, the plot is very unsmooth. By lambda = 0.2, the lines have smoothed considerable at the various quantile values. By lambda = 1, we see very little smoothing as the lambda is increased.

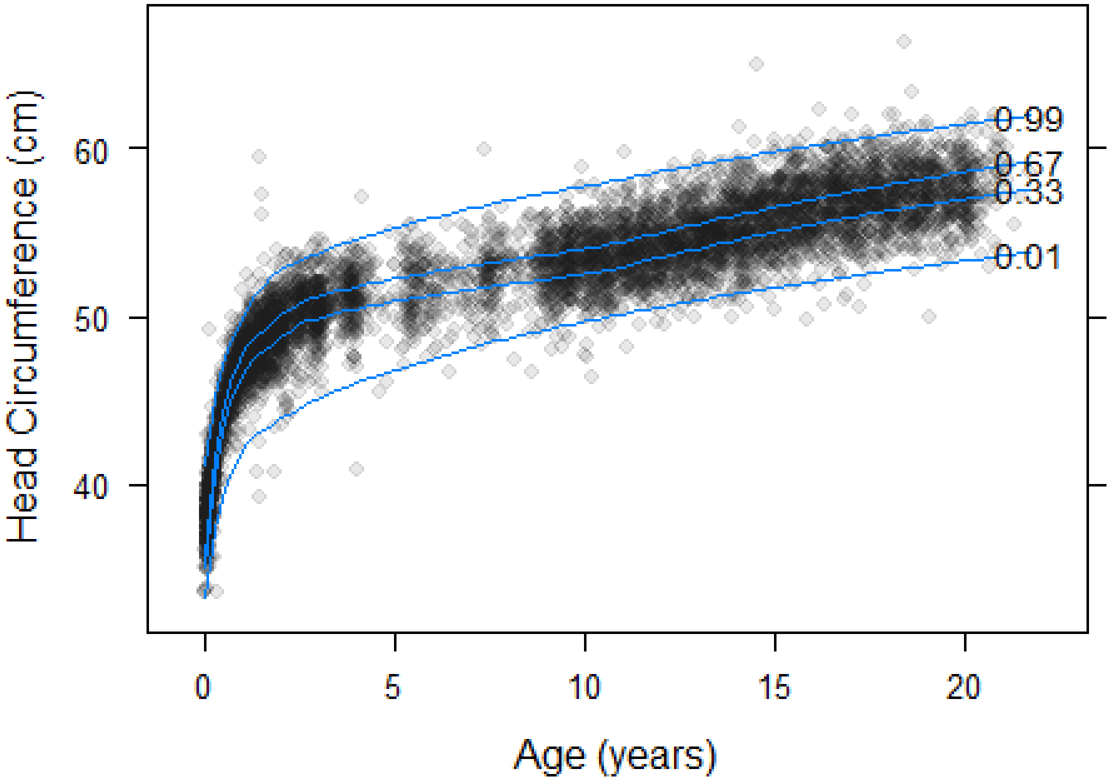



Age vs. Head Circumference for Lambda = 0.05


Age vs. Head Circumference for Lambda = 0.2



Age vs. Head Circumference for Lambda = 1



Age vs. Head Circumference for Lambda = 5

Problem #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 limited to, an introduction, motivation, case study considered and findings.

Results:

Introduction: As modern computing has evolved, so has the availability and usage of statistical computing software for purposes of quantile regression. This paper describes the quantile regression method and how it can be used with a variety of datasets.

Koenker provides a brief overview of some R functionality including how to gain more information through the use of various help features. An excellent explanation of the *summary()* function is provided as well.

Motivation: Quantile Regression provides potentially more successful fits to data due to its estimation of the median. This allows for better fitting model especially when outliers are present that can have a more significant affect on the mean than on the median.

As compared to the Least Squares Method, which is examined through the Engel example, this method can produce a better fit. In the Engel example, there are two extreme outliers that have a heavy impact on the fit when using a Least Squares Method. By utilitizing Quantile Regression, we are able to achieve better fit.

Case Study & Findings: One example used to discuss the functionality of quantile regression, as discussed above, is the Engel dataset regarding Food Expenditures vs. Household Income. As Koenker shows, the use of a mean instead of a median, which is what is used in quantile regression, can significantly skew the fit.

In this example, the Least Squares Method fit an intercept well above that of the quantile regression method due to two outliers where income is very high and food expenditures are low. In fact, the intercept has been skewed so far upward that the intercept has become a “centercept” as Koenker notes is Tukey terminology. This skews the model and makes a Least Squares model less accurate.

Conclusion: The primary thing that I took away from this paper is that the quantile regression method is incredibly useful in many different types of situations. When working with data containing larger outliers, the mean provides misleading interpretations of the data whereas the median (quantile) method can provide a better, more accurate fit.