STOR 455 Homework #1

25 points - Due Friday 9/3 at 5:00pm

Directions: This first assignment is meant to be a brief introduction to working with R in RStudio. You may (and should) collaborate with other students. If you do so, you must identify them on the work that you turn in. You should complete the assignment in an R Notebook, including all calculations, plots, and explanations. Make use of the white space outside of the R chunks for your explanations rather than using comments inside of the chunks. For your submission, you should knit the notebook to PDF and submit the file to Gradescope.

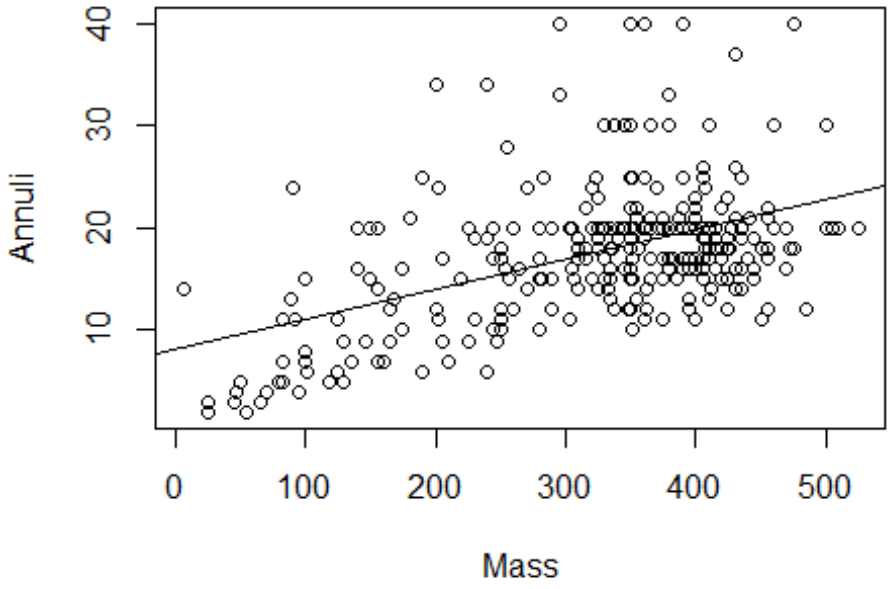
Eastern Box Turtles: The Box Turtle Connection is a long-term study anticipating at least 100 years of data collection on box turtles. Their purpose is to learn more about the status and trends in box turtle populations, identify threats, and develop strategies for long-term conservation of the species. Eastern Box Turtle populations are in decline in North Carolina and while they are recognized as a threatened species by the International Union for Conservation of Nature, the turtles have no protection in North Carolina. There are currently more than 30 active research study sites across the state of North Carolina. Turtles are weighed, measured, photographed, and permanently marked. These data, along with voucher photos (photos that document sightings), are then entered into centralized database managed by the NC Wildlife Resources Commission. The *Turtles* dataset (found at the link below) contains data collected at The Piedmont Wildlife Center in Durham.

https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/Turtles.csv

1) The *Annuli* rings on a turtle represent growth on the scutes of the carapace and plastron. In the past, it was thought that annuli corresponded to age, but recent findings suggest that this is not the case. However, the annuli are still counted since it may yield important life history information. Construct a least squares regression line that predicts turtles' *Annuli* by their *Mass*.

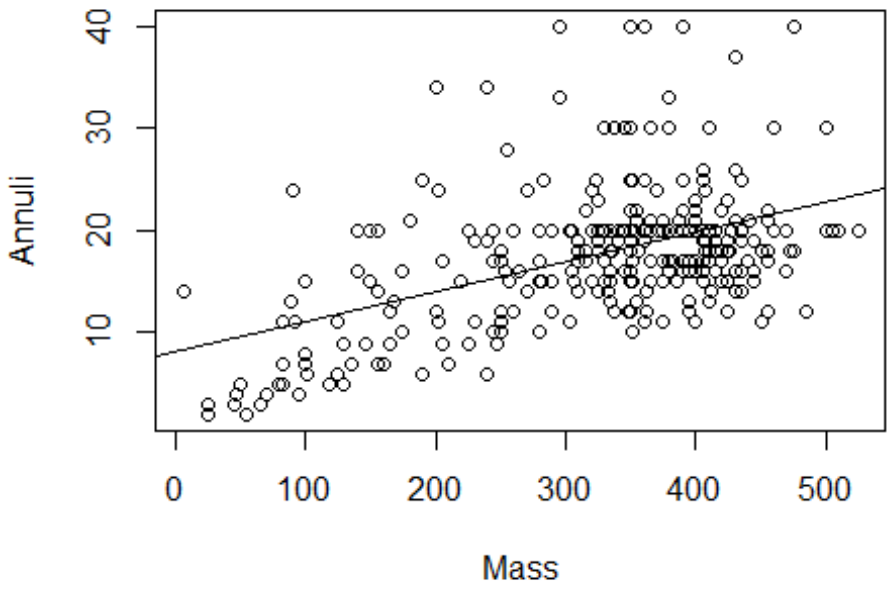
```
turtles <- read.csv("https://raw.githubusercontent.com/JA-
McLean/STOR455/master/data/Turtles.csv
") # Calling in the data
plot(Annuli~Mass, data = turtles) # Checking to see if the data looks linear,
or if there are any patterns.

# The data fans out a little as mass increases, the annuli variability
increases. It is not consistent.

mod1 <- lm(Annuli~Mass, turtles)
mod1

##
## Call:
## lm(formula = Annuli ~ Mass, data = turtles)
```

```
##
## Coefficients:
## (Intercept)         Mass
##     8.08494      0.02957

abline(mod1)
```

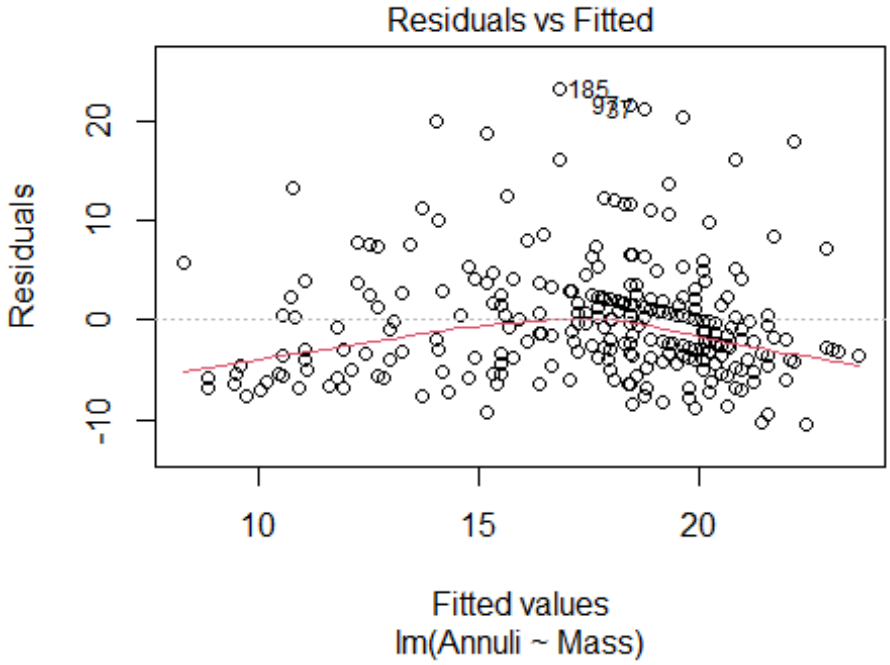


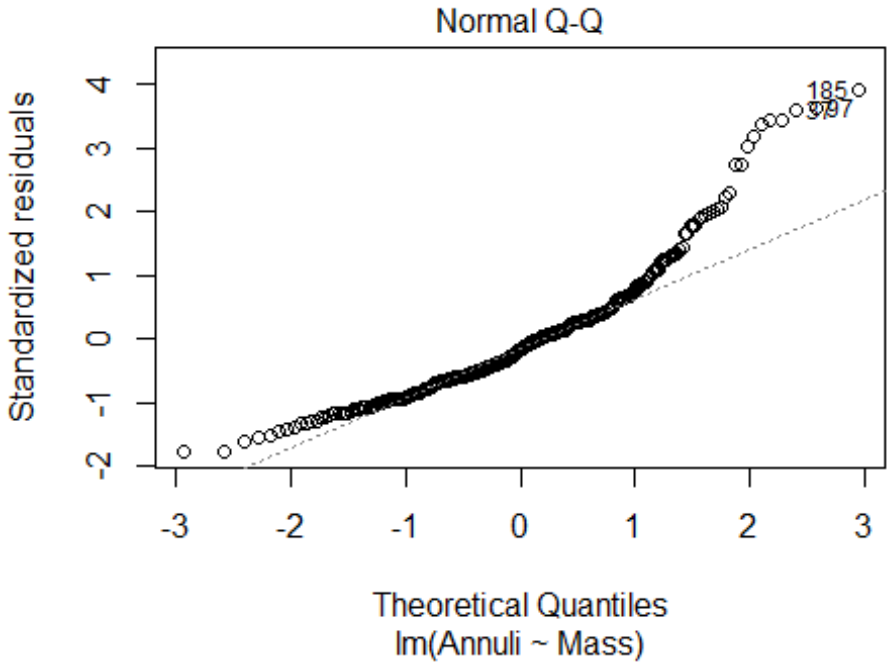
2) Produce a scatterplot of this relationship (and include the least squares line on the plot).

```
plot(Annuli~Mass, data = turtles)
abline(mod1)
```




```r
plot(mod1, 1:2) # This is for me to see what other plots look like
```





3) The turtle in the ninth row of the *Turtles* dataset has a mass of 475 grams. What does your model predict for this turtle's number of *Annuli*? What is the residual for this case?

```
# Linear Regression Line Formula: y = mx + b; y = 0.02957x + 8.08494
mass <- 475
yhat1 <- 0.02957*mass + 8.08494 #This is mod1's regression line
yhat1

## [1] 22.13069

# Based on our model, this turtle should have 18 annuli.
# Actual:
y1 <- turtles[10, "Annuli"]
y1

## [1] 40

# The turtle's true # of annuli is 18.
residual1 <- y1-yhat1
residual1

## [1] 17.86931

turtles[10,]

##    LifeStage  Sex Annuli Mass StraightlineCL MaxCW PL_AnteriortoHinge
## 10     Adult Male     40  475            137   105                 52
##    PL_HingetoPosterior ShellHeightatHinge
## 10                  79                63
```

4) Which turtle (by row number in the dataset) has the largest positive residual? What is the value of that residual?

Turtle number 185 has the largest positive residual of 23.19151.

```
mod1 <- lm(Annuli~Mass, turtles)
turtles$residual <- mod1$residuals
max(turtles$residual)

## [1] 23.19151

x <- which.max(turtles$residual)
turtles[x,]

##     LifeStage    Sex Annuli Mass StraightlineCL MaxCW PL_AnteriortoHinge
## 185     Adult Female     40  295            109    85                 44
##     PL_HingetoPosterior ShellHeightatHinge residual
## 185                  64                61 23.19151
```

5) Which turtle (by row number in the dataset) has the most negative residual? What is the value of that residual?

Turtle number 93 has the smallest residual of -10.42705.

```
mod1 <- lm(Annuli~Mass, turtles)
turtles$residual <- mod1$residuals
min(turtles$residual)
```

```
## [1] -10.42705
```

```
y <- which.min(turtles$residual)
turtles[y,]
```

```
##    LifeStage    Sex Annuli Mass StraightlineCL  MaxCW PL_AnteriortoHinge
## 93     Adult Female     12  485          129.2 104.82              52.59
##    PL_HingetoPosterior ShellHeightatHinge  residual
## 93               79.27              65.09 -10.42705
```

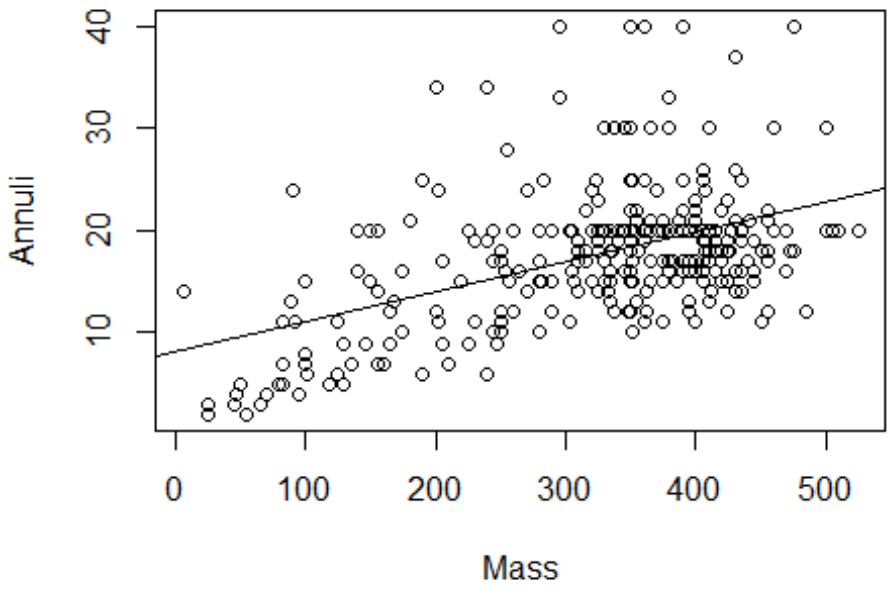
6) Comment how the conditions for a simple linear model are met this model. Include at least two plots (in addition to the plot in question 2) - with commentary on what each plot tells you specifically about the appropriateness of conditions.

The below r chunk contains the scatterplot with the least squares regression line from question 2, a fitted residuals plot, and a normal qq plot.

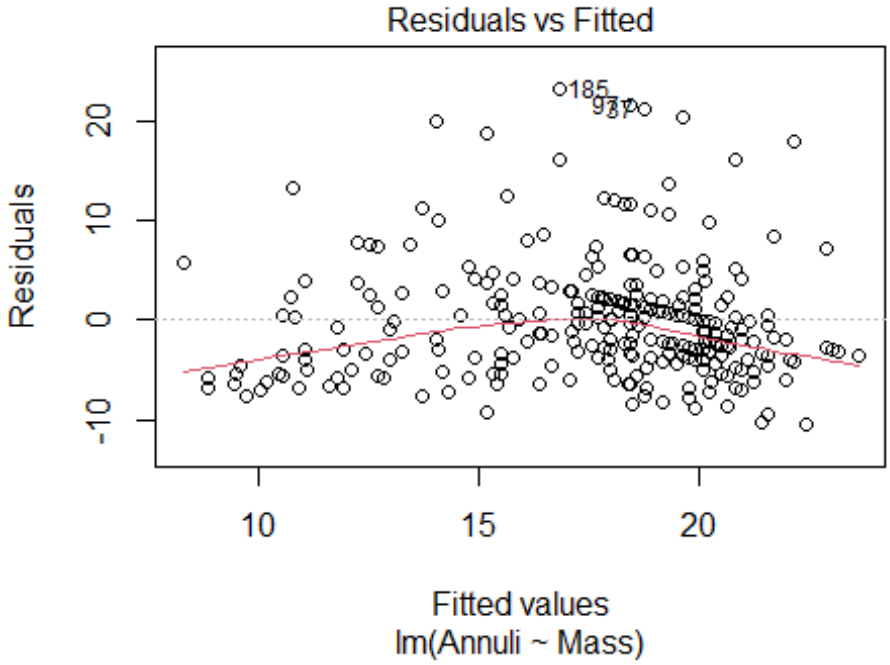
- the first scatterplot with the line shows the least squares regression model
- the fitted residuals plot shows how off our predictions were
- The normal gg plot shows what our data would look like in comparision to a normal distribution of a similar dataset.

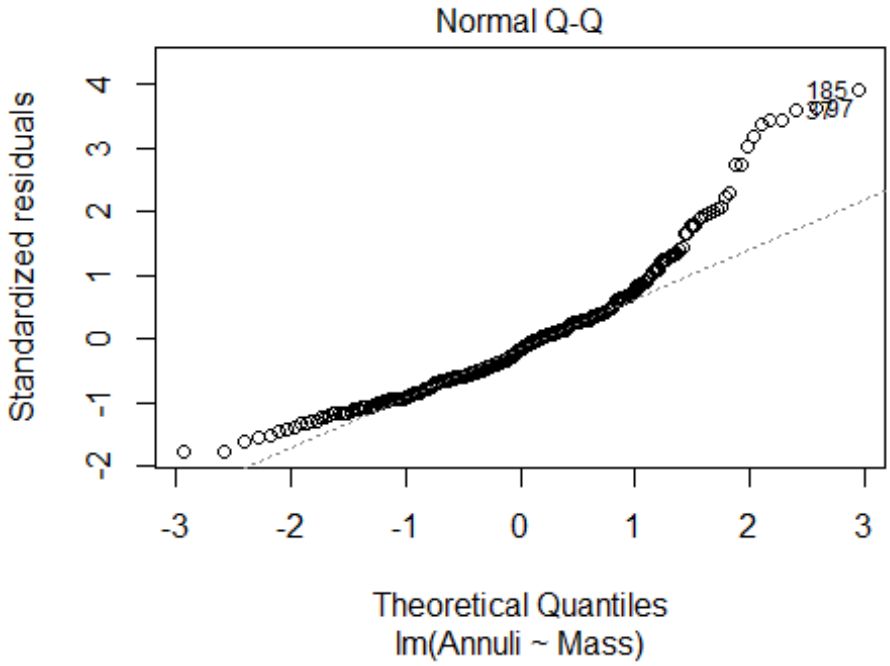
We know that the conditions for a simple linear model are met because all five of the following conditions are met: - Linearity: The data in the first scatterplot has a very loosely positive, linear slope. - Zero mean: We can see that the distribution of errors is roughly centered at zero based on the fitted residuals plot. This is because the dots appear to be randomly placed to the top and bottom of the zero line. I can also see that there may be a pull towards higher residual numbers. - COnstant Variance: The data in the fitted residuals plot and the first scatterplot show random dots with no distinct pattern (other than the slightly positive linear path). - Independence: We can assume independence between these turtles because one turtle's weight does not affect another turtle's annuli or weight. - Normality: The distribution is roughly normal based on the normal qq plot. The normal qq plot tells us how close our data would fit a normally distributed pattern of similar data. Our data tails off at the upper end, which may be something to look into later and to keep in mind. When later looking at the histogram of the residuals, the data is slightly skewed to the right.

```
plot(Annuli~Mass, data = turtles)
abline(mod1)
```

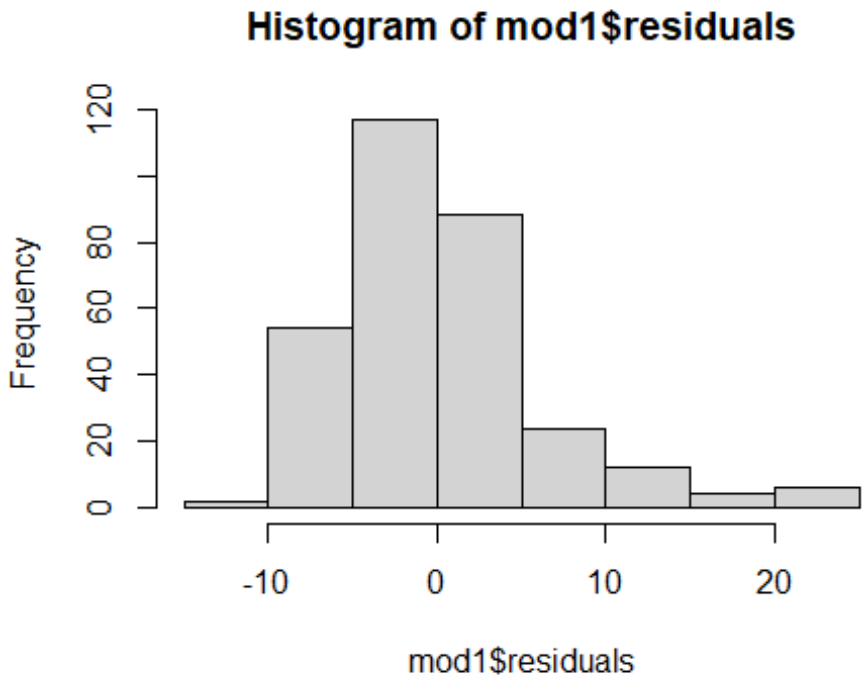



```r
plot(mod1, 1:2)
```



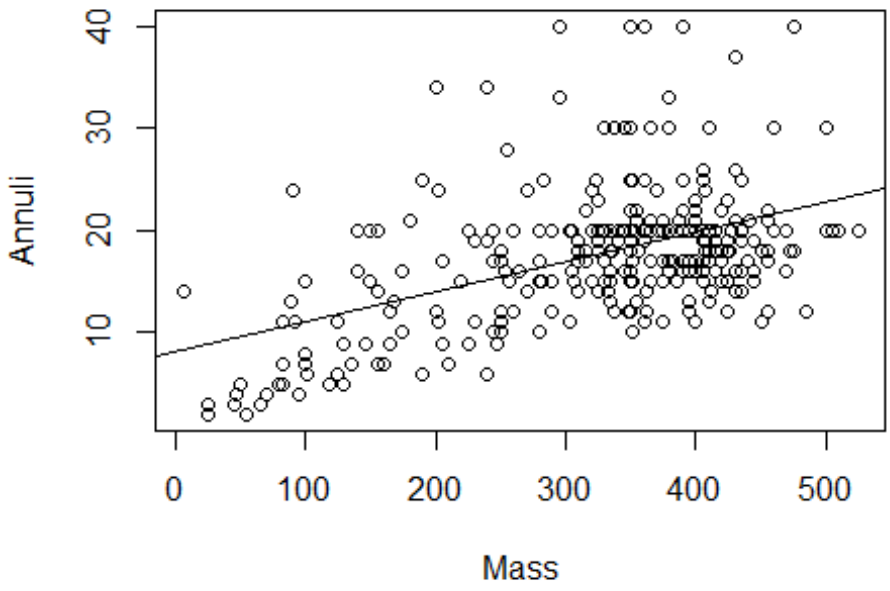



```r
hist(mod1$residuals)
```

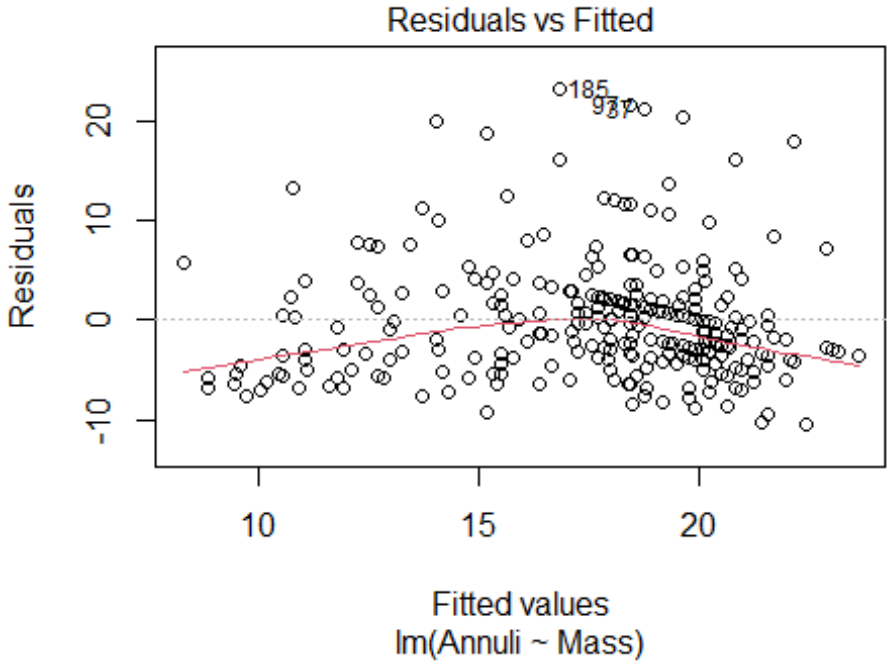


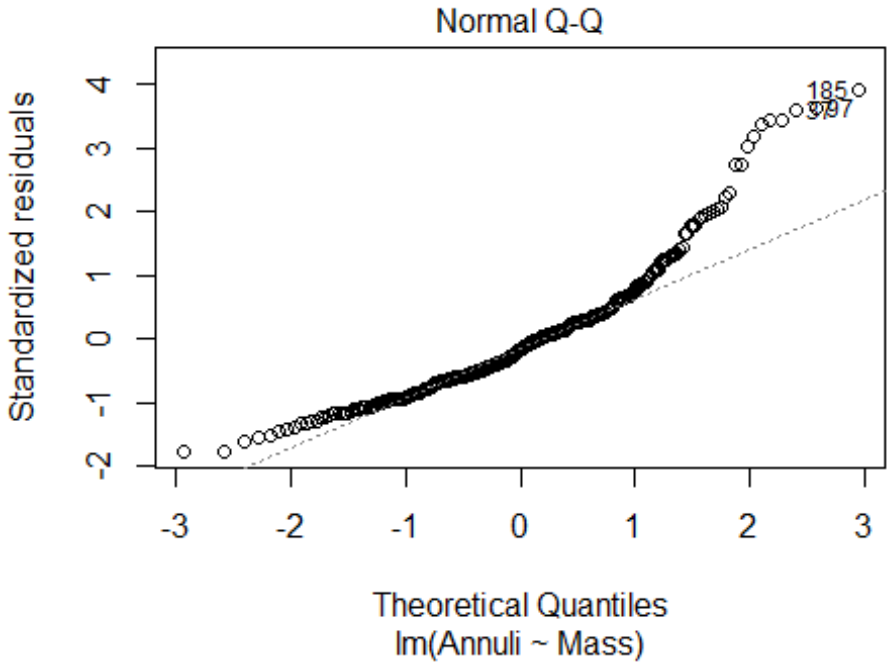
7) Experiment with at least two transformations to determine if models constructed with these transformations appear to do a better job of satisfying the simple linear model conditions. Include the summary outputs for fitting these model and scatterplots of the transformed variable(s) with the least square lines.

```
plot(Annuli~Mass, data = turtles) # This is the original
abline(lm(Annuli~Mass, data = turtles))
```

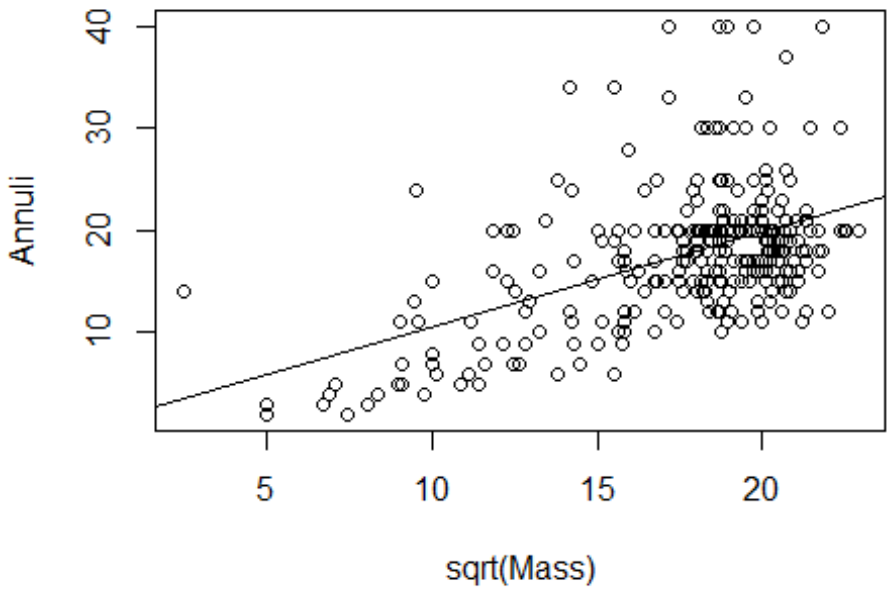



```
plot(lm(Annuli~Mass, data = turtles), 1:2)
```

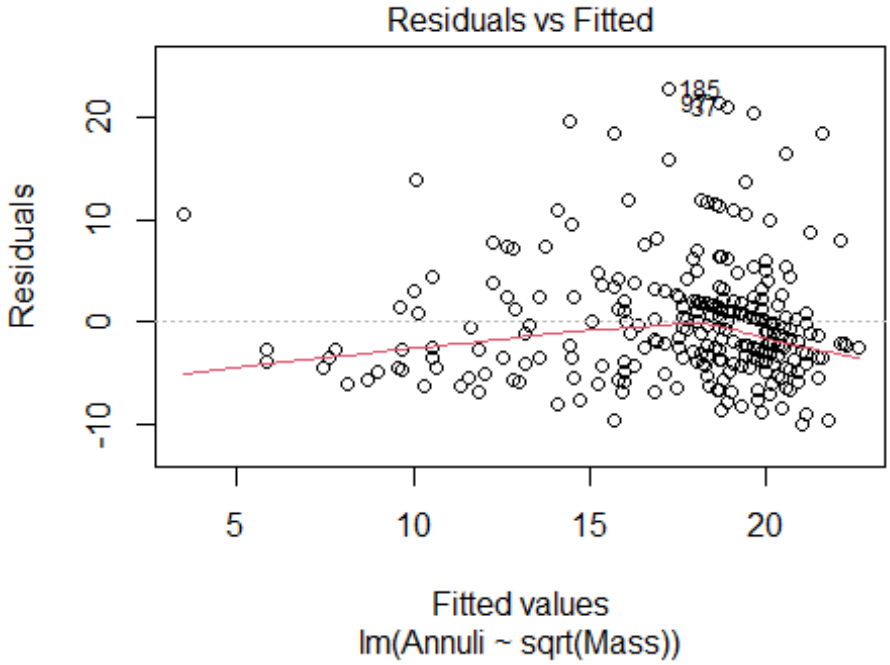


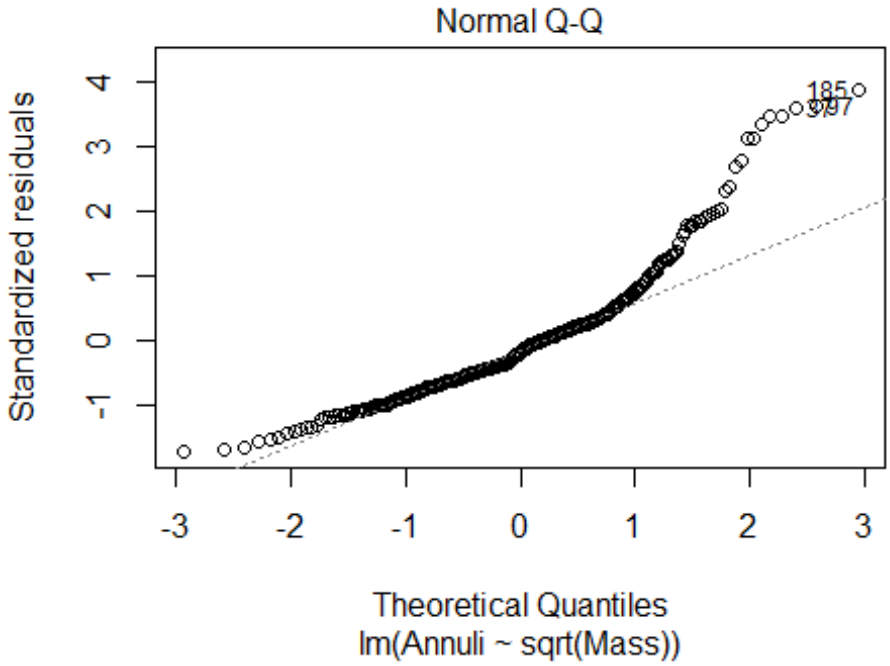



```r
# square root Mass
plot(Annuli~sqrt(Mass), data = turtles)
abline(lm(Annuli~sqrt(Mass), data = turtles))
```

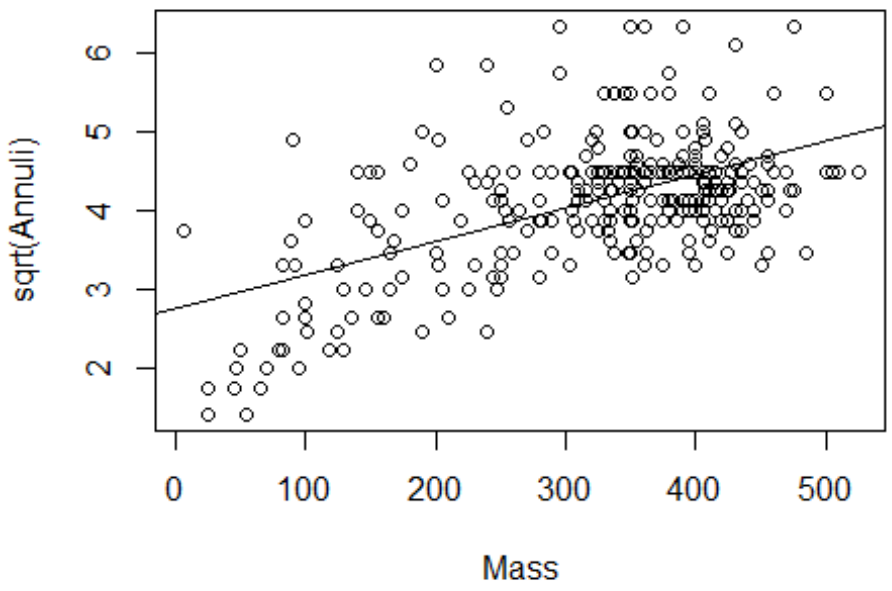



```
plot(lm(Annuli~sqrt(Mass), data = turtles), 1:2)
```

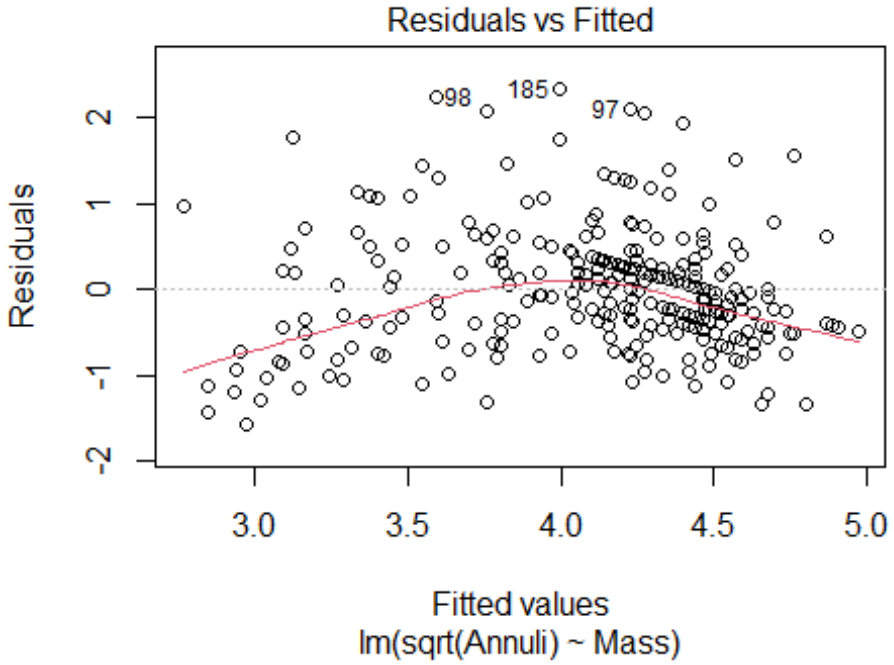


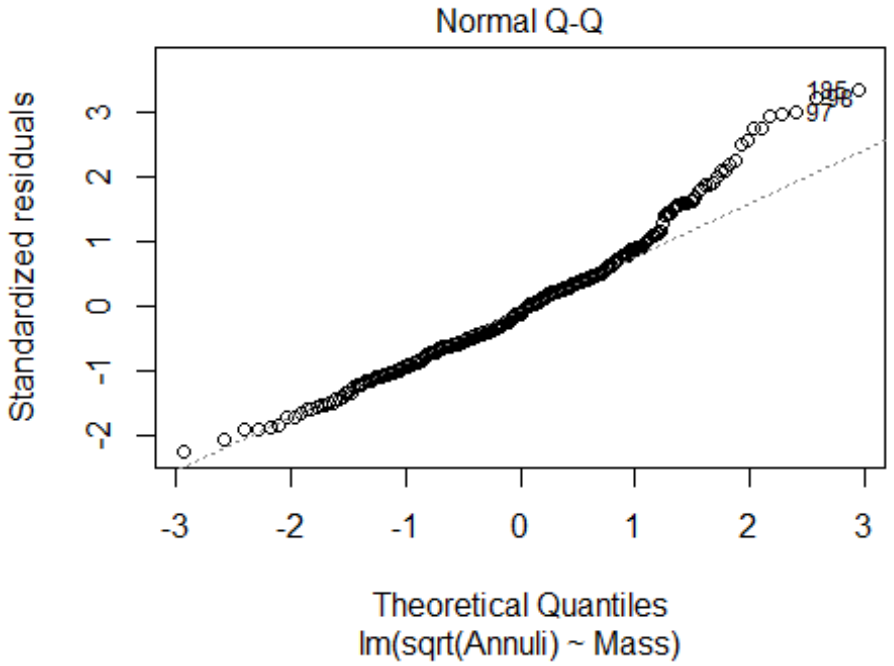



```r
# Square root Annuli
plot(sqrt(Annuli)~Mass, data = turtles)
abline(lm(sqrt(Annuli)~Mass, data = turtles))
```

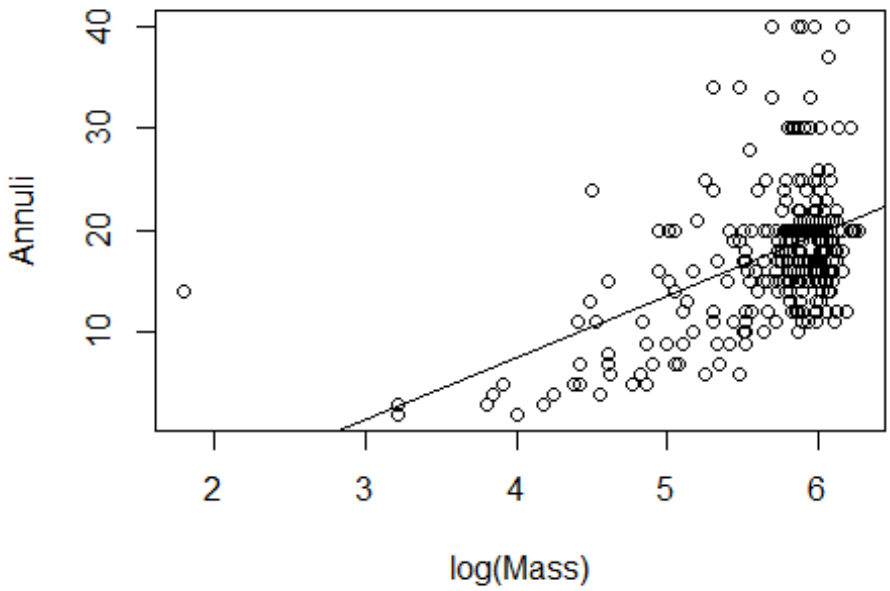



```
plot(lm(sqrt(Annuli)~Mass, data = turtles), 1:2)
```

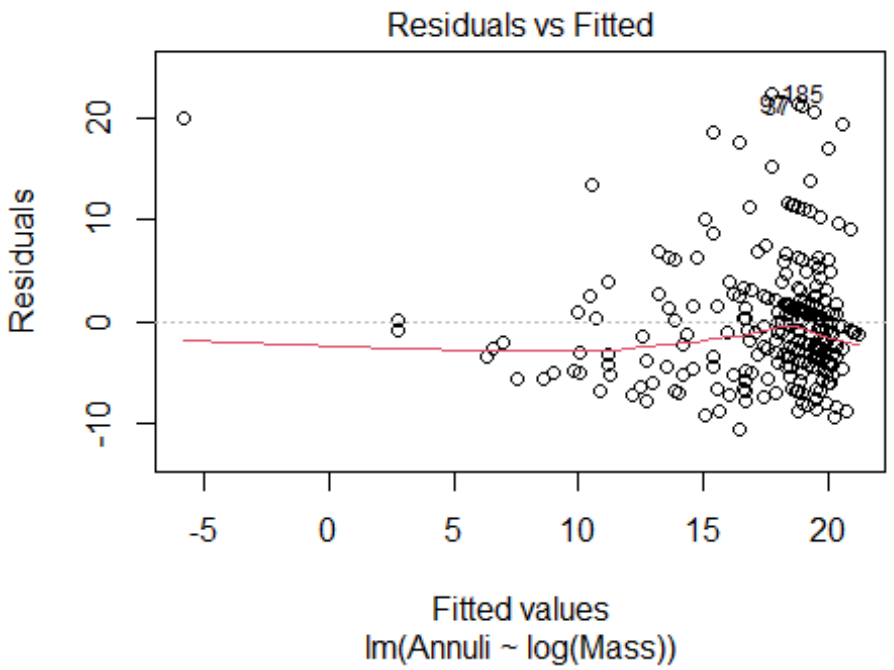


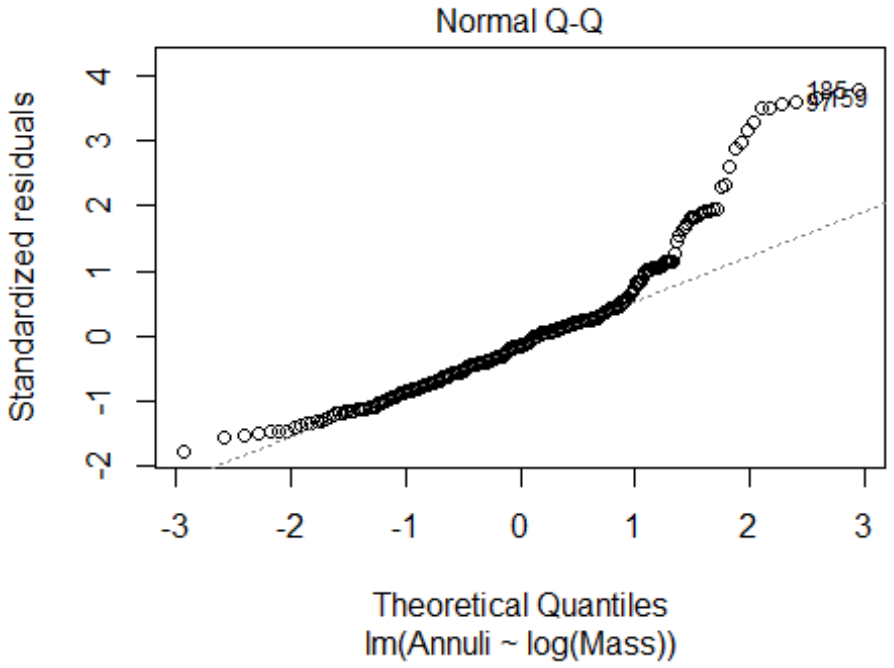



```r
# Log Mass
plot(Annuli~log(Mass), data = turtles)
abline(lm(Annuli~log(Mass), data = turtles))
```

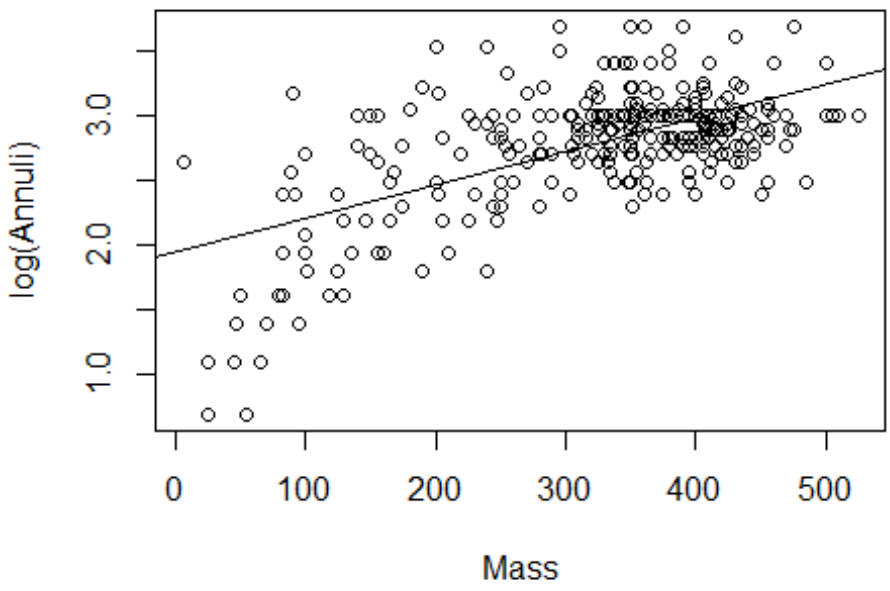



```r
plot(lm(Annuli~log(Mass), data = turtles), 1:2)
```

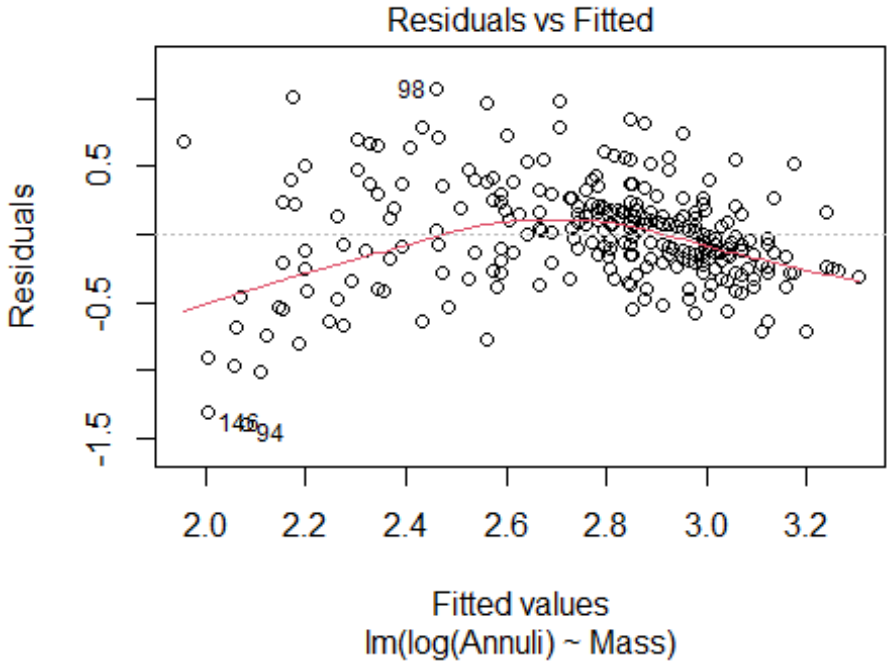


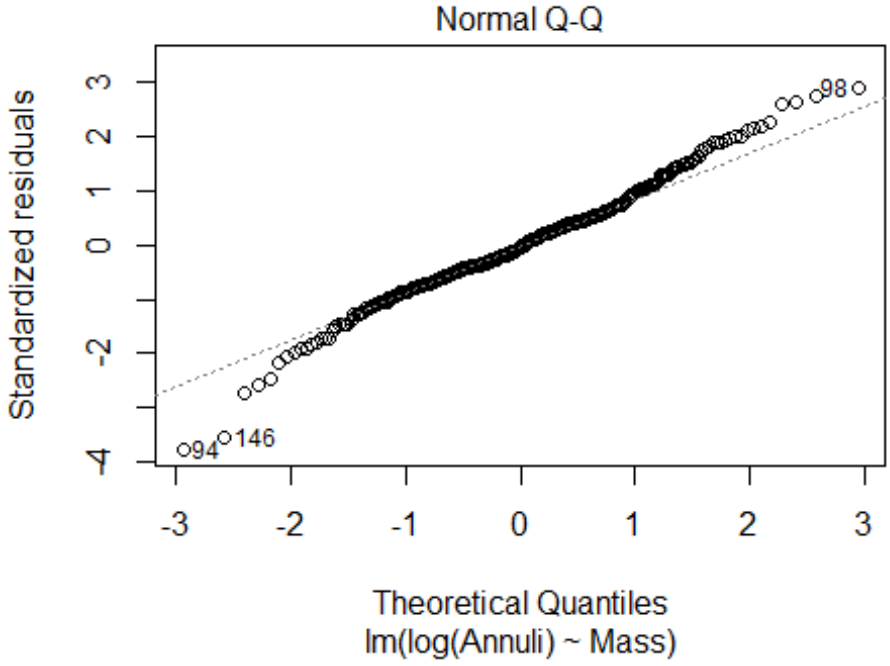



```r
# Log Annuli
plot(log(Annuli)~Mass, data = turtles)
abline(lm(log(Annuli)~Mass, data = turtles))
```

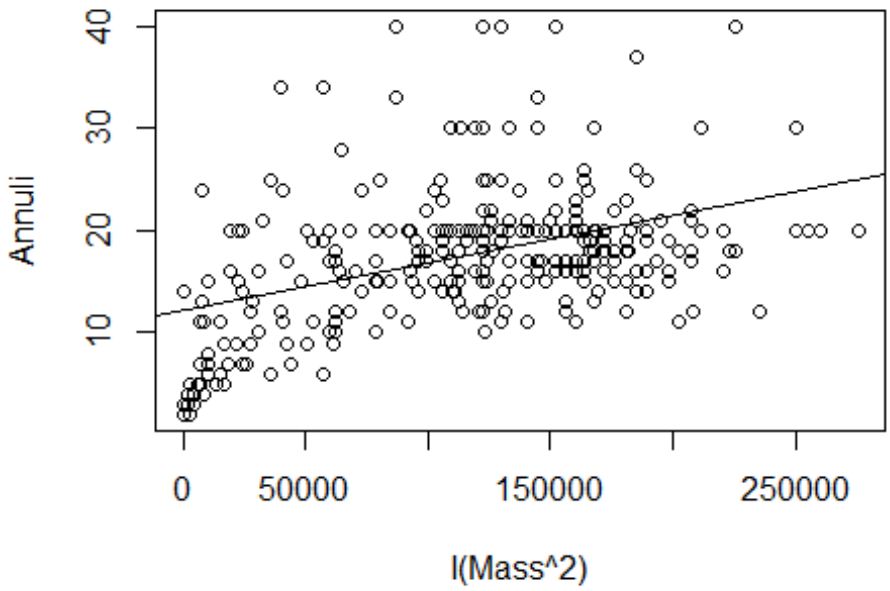



```
plot(lm(log(Annuli)~Mass, data = turtles), 1:2)
```

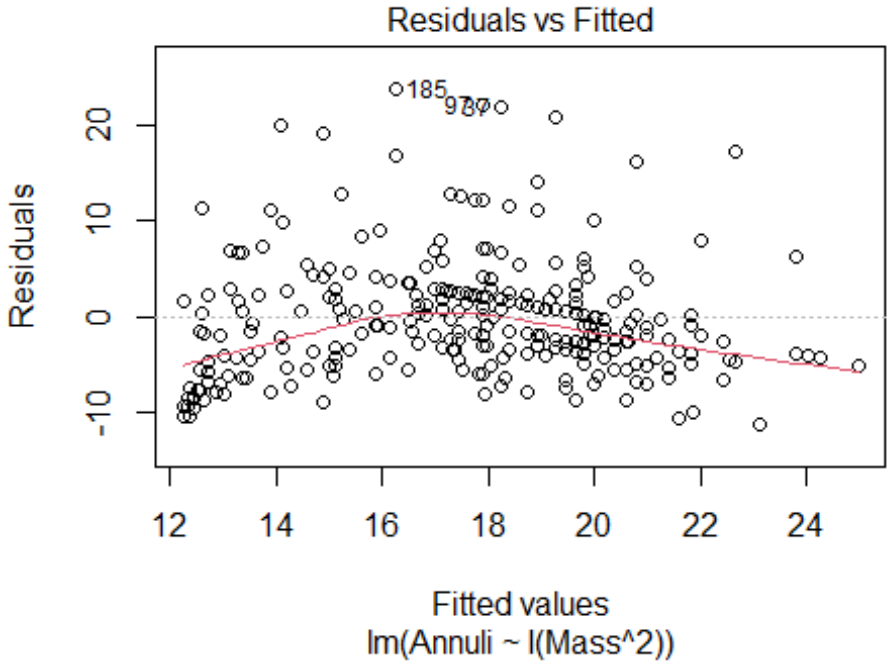


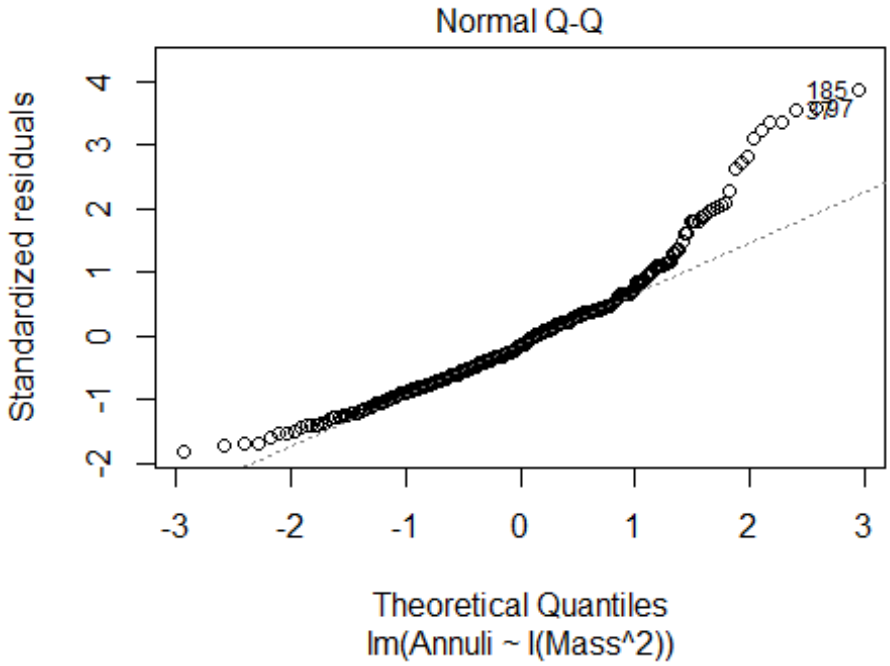



```r
# Add Exponent to Mass
plot(Annuli~I(Mass^2), data = turtles)
abline(lm(Annuli~I(Mass^2), data = turtles))
```

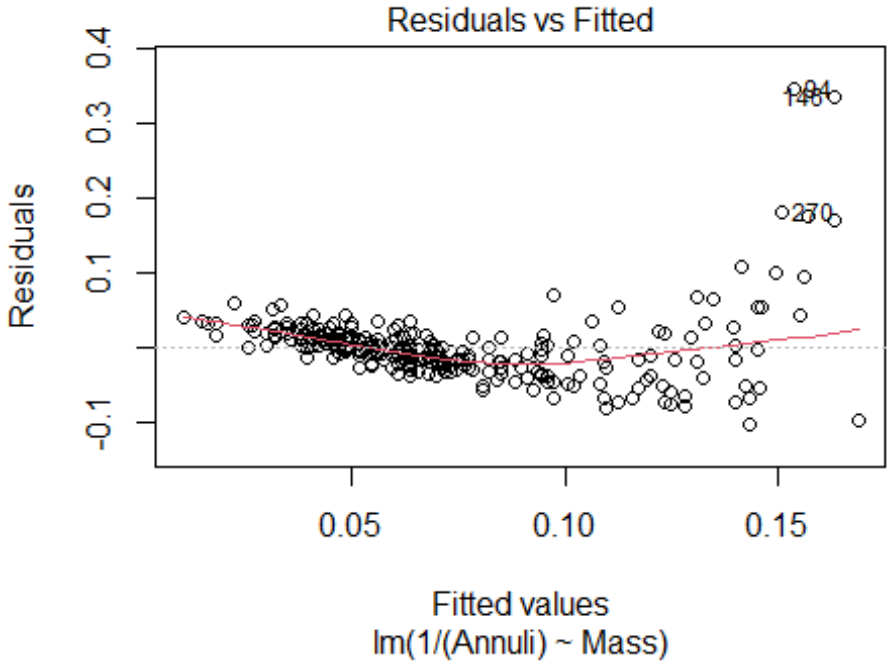


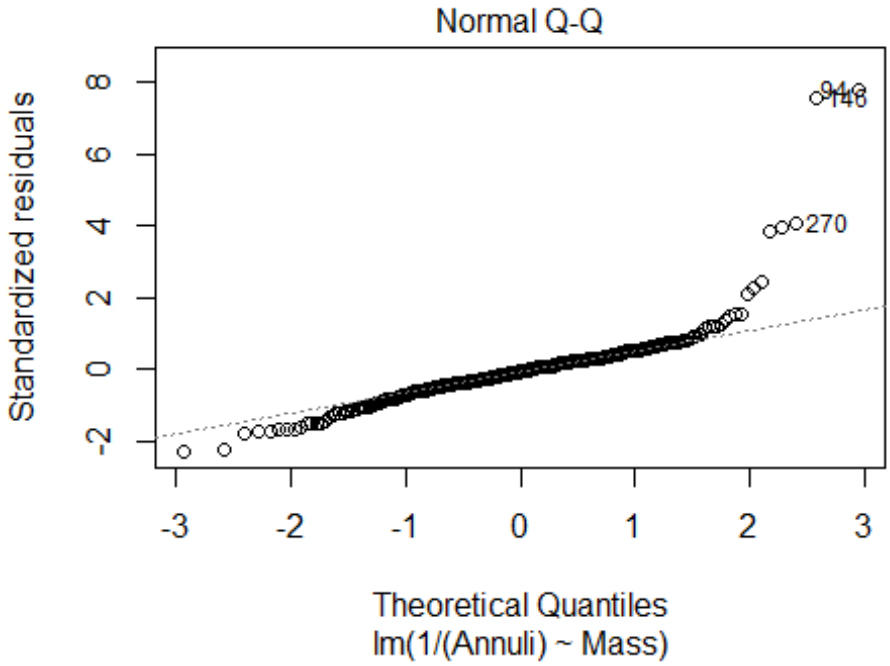

```
plot(lm(Annuli~I(Mass^2), data = turtles), 1:2)
```

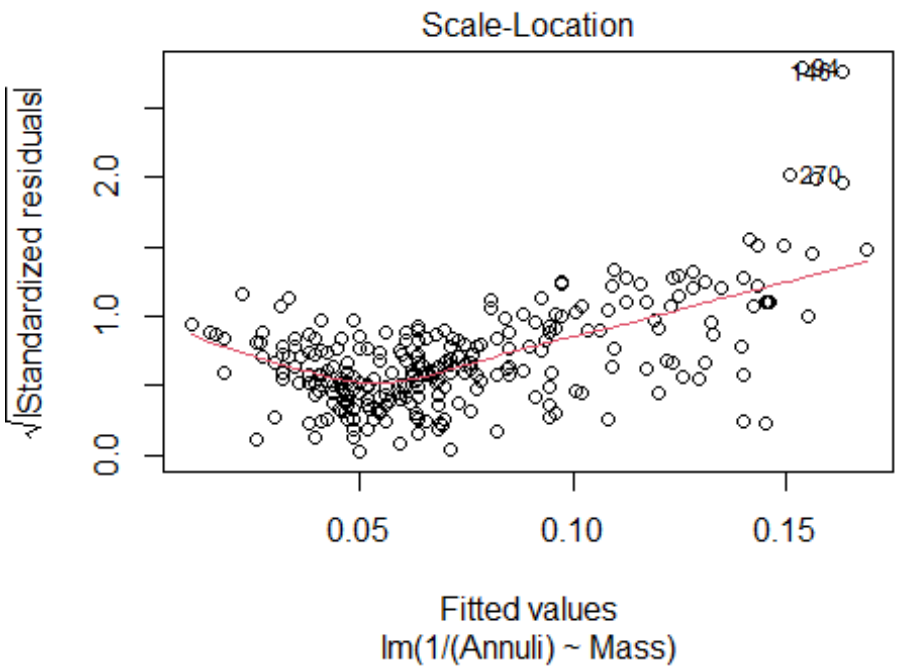




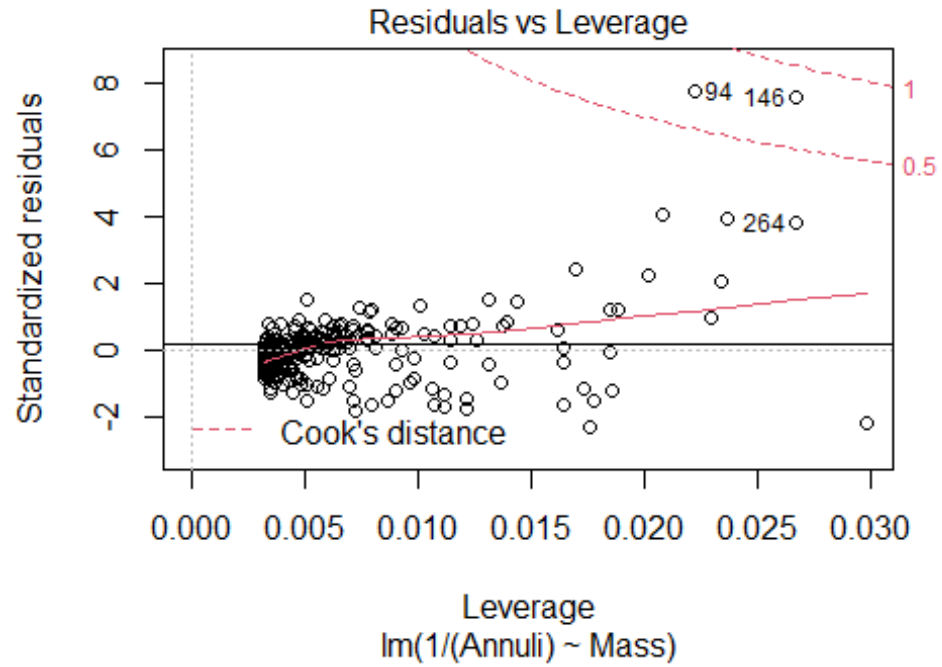

```r
# Division Annuli
plot(lm(1/(Annuli)~Mass, data=turtles))
```


Residuals vs Fitted
Residuals
Fitted values
lm(1/(Annuli) ~ Mass)
0.05
0.10
0.15
-0.1
0.1
0.2
0.3
0.4
270

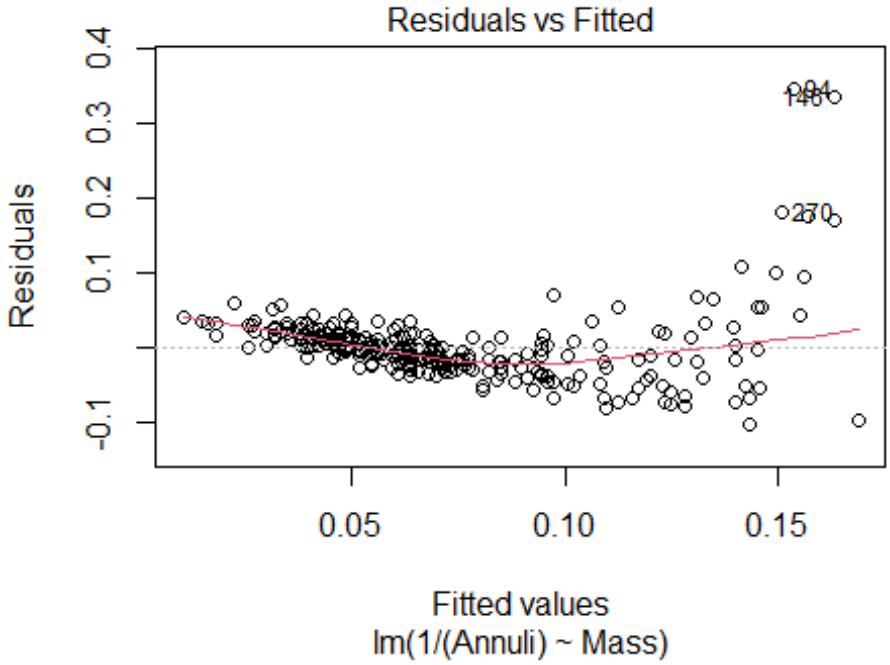

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(1/(Annuli) ~ Mass)
-3
-2
-1
0
1
2
3
-2
0
2
4
6
8
270

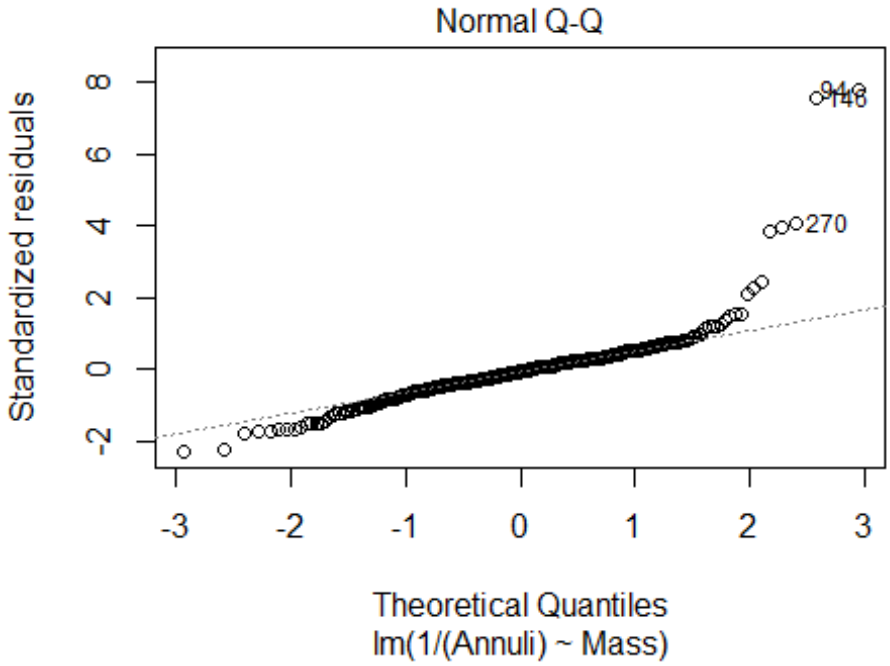



```
abline(lm(1/(Annuli)~Mass, data=turtles))
```




```
plot(lm(1/(Annuli)~Mass, data=turtles), 1:2)
```



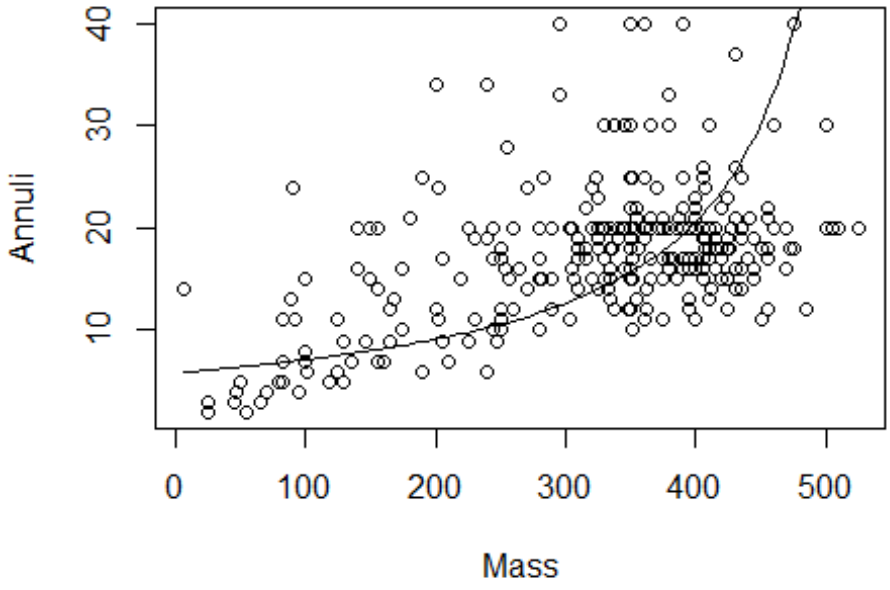


The # Division Annuli is the best transformation that fits the data. This is because I first compared all the transformation's QQ norm plots. I found that the log transformation and the division transformation were the closest fitting qq norm plots. After this, I checked the fitted residual plots. # Division Annuli's fitted residual plot is much closer to the line than # Log

Annuli. This means that # Division Annuli's predictions tend to be closer to the actual values than # Log Annuli. Therefore, I concluded that #Division Annuli is the best transformation that fits the data.

8) For your model using the best transformation from question 7, plot the raw data (not transformed) with the model (likely a curve) on the same axes.

```
B0 = summary(lm(1/(Annuli)~(Mass), data=turtles))$coefficients[1,1]
B1 = summary(lm(1/(Annuli)~(Mass), data=turtles))$coefficients[2,1]


plot(Annuli~Mass, data = turtles)
curve(1/(B0+B1*x), add=TRUE)
```



9) Again, the turtle in the ninth row of the *Turtles* dataset has a mass of 325 grams. For your model using the best transformation from question 7, what does this model predict for this turtle's number of *Annuli*? In terms of *Annuli*, how different is this prediction from the observed value?

Our predicted value is 0.8309 less than the actual number of turtle annuli for the turtle with a mass of 475 grams.

```
B0 = summary(lm(1/(Annuli)~(Mass),
                data=turtles))$coefficients[1,1] # Intercept
B1 = summary(lm(1/(Annuli)~(Mass),
                data=turtles))$coefficients[2,1] # Slope
```

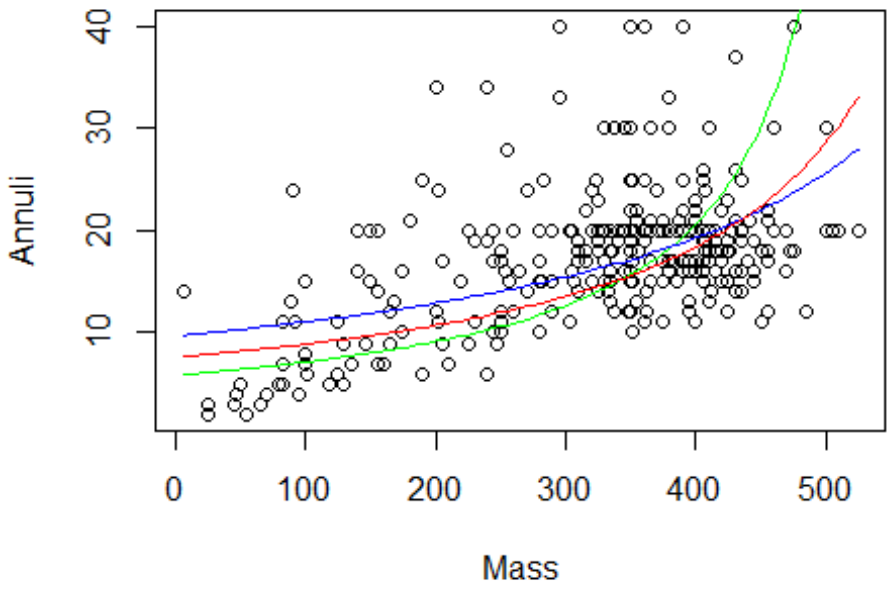
```
predicted <- 1/(B0+B1*mass)
predicted

## [1] 39.1691

turtles$Annuli[10] - predicted

## [1] 0.8309016
```

10) For your model using the best transformation from question 7, could the relationship between *Mass* and *Annuli* be different depending on the *Sex* of the turtle? Construct two new dataframes, one with only male turtles, and one with only female turtles. Using your best transformation from question 7, construct two new models to predict *Annuli* with *Mass* for male and female turtles separately. Plot the raw data for *Anulli* and *Mass* as well as each of these new models on the same plot. You should use different colors for each model (which are likely curves). What does this plot tell you about the relationship between *Mass* and *Annuli* depending on the *Sex* of the turtle?

This linear regression model predicts that as the mass of the turtles increase, so do their number of Annuli. The different curves on the graphs below predicts that males have more Annuli than females from a certain mass range, but once the turtles surpass about 425g of weight, the female regression line predicts females have more Annuli than males. This analysis tells me that it is most likely the female points that are pulling the last bit of the regression line on the far right.

```
males <- turtles[turtles$Sex == "Male",]
females <- turtles[turtles$Sex == "Female",]
head(males)

##    LifeStage  Sex Annuli Mass StraightlineCL  MaxCW PL_AnteriortoHinge
## 1      Adult Male     13  410         127.00 102.00              48.00
## 2      Adult Male     19  340         113.62  93.96              44.87
## 4      Adult Male     16  175         127.70 101.16              54.76
## 9      Adult Male     18  325         115.00  94.00              45.00
## 10     Adult Male     40  475         137.00 105.00              52.00
## 11     Adult Male     15  405         123.00  99.00              49.00
##    PL_HingetoPosterior ShellHeightatHinge   residual
## 1                68.00              61.00 -7.2091983
## 2                67.61              55.88  0.8607977
## 4                84.72              61.97  2.7400737
## 9                68.00              55.00  0.3043682
## 10               79.00              63.00 17.8686627
## 11               72.00              61.00 -5.0613414

head(females)

##    LifeStage    Sex Annuli Mass StraightlineCL  MaxCW PL_AnteriortoHinge
## 3   Juvenile Female      7  160          89.49  73.51              39.60
## 5   Juvenile Female      7  100          81.00  69.00              35.00
## 7      Adult Female     18  472         131.00 104.00              49.00
## 8      Adult Female     20  155         122.85  99.38              51.68
```

```
## 14     Adult Female     30  345         105.00  89.00              40.00
## 15     Adult Female     19  240         124.68 102.30              48.92
##    PL_HingetoPosterior ShellHeightatHinge  residual
## 3                53.65              43.48 -5.816356
## 5                44.00              39.00 -4.042074
## 7                80.00              59.00 -4.042623
## 8                74.73              64.60  7.331501
## 14               66.00              56.00 11.712941
## 15               71.92              58.47  3.817935

B0.trans = summary(lm(1/(Annuli)~(Mass),
                data=turtles))$coefficients[1,1] # Intercept
B1.trans = summary(lm(1/(Annuli)~(Mass),
                data=turtles))$coefficients[2,1] # Slope

B0.trans.males = summary(lm(1/(Annuli)~(Mass),
                data=males))$coefficients[1,1] # Intercept
B1.trans.males = summary(lm(1/(Annuli)~(Mass),
                data=males))$coefficients[2,1] # Slope

B0.trans.females = summary(lm(1/(Annuli)~(Mass),
                data=females))$coefficients[1,1] # Intercept
B1.trans.females = summary(lm(1/(Annuli)~(Mass),
                data=females))$coefficients[2,1] # Slope

plot(Annuli~Mass, turtles)
curve(1/(B0.trans+B1.trans*x), col = "green", add=TRUE)
curve(1/(B0.trans.males+B1.trans.males*x), col = "blue", add=TRUE)
curve(1/(B0.trans.females+B1.trans.females*x), col = "red", add=TRUE)
```


Annuli
10
20
30
40
0
100
200
300
400
500
Mass