Modeling and Transformations

Eastern Box Turtles

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*.

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

3) The turtle in the 40th row of the *Turtles* dataset has a mass of 390 grams. What does your model predict for this turtle's number of *Annuli*? What is the residual for this case?

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

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

6) Comment on how each of the conditions for a simple linear model are (or are not) met in 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.

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

8) For your model with the best transformation from question 7 (It still may not be an ideal model), plot the raw data (not transformed) with the model (likely a curve) on the same axes.

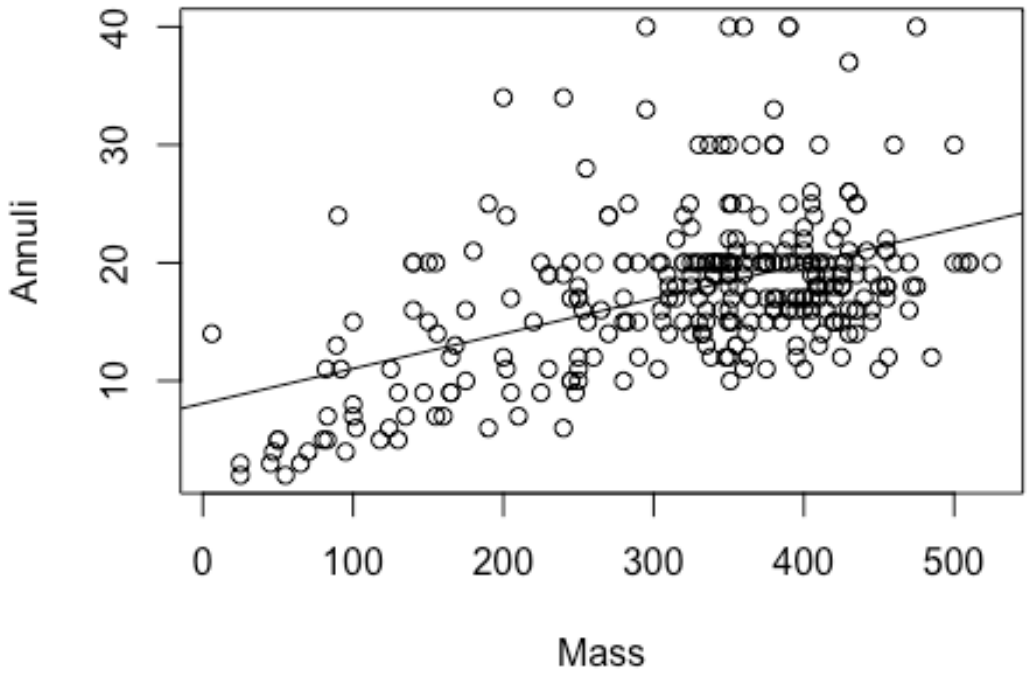
9) Again, the turtle in the 40th row of the *Turtles* dataset has a mass of 390 grams. For your model using the best transformation from question 7, what does this model predict for this turtle's number of *Annuli*? In terms of *Annuli*, how different is this prediction from the observed value?

10) For your model using the best transformation from question 7, could the relationship between *Mass* and *Annuli* be different depending on the *LifeStage* and *Sex* of the turtle? Construct two new dataframes, one with only adult male turtles, and one with only adult female turtles. Using your best transformation from question 7, construct two new models to predict *Annuli* with *Mass* for adult male and adult female turtles separately. Plot the raw data for *Anulli* and *Mass* for all adult turtles 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 adult turtles?

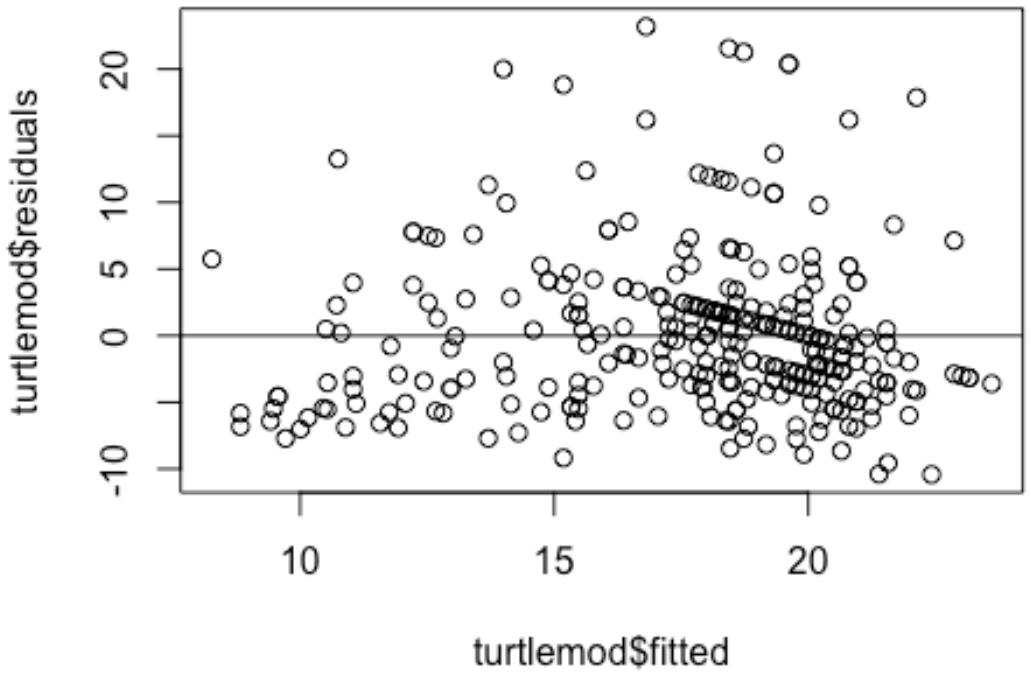
```
#Using Library Readr and importing code
library(readr)
Turtles <- read_csv("https://raw.githubusercontent.com/JA-
McLean/STOR455/master/data/Turtles.csv")

## Rows: 307 Columns: 9
## ── Column specification
────────────────────────────────────────────────────────
## Delimiter: ","
## chr (2): LifeStage, Sex
## dbl (7): Annuli, Mass, StraightlineCL, MaxCW, PL_AnteriortoHinge,
PL_Hingeto...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this
message.

#Looking at top 40 lines of the Turtle Data set
head(`Turtles`, 40)

## # A tibble: 40 × 9
##    LifeStage Sex     Annuli  Mass StraightlineCL MaxCW PL_AnteriortoHinge
##    <chr>     <chr>    <dbl> <dbl>          <dbl> <dbl>               <dbl>
##  1 Adult     Male        13   410          127   102                  48
##  2 Adult     Male        19   340          114.   94.0                44.9
##  3 Juvenile  Female       7   160           89.5  73.5                39.6
##  4 Adult     Male        16   175          128.  101.                 54.8
##  5 Juvenile  Female       7   100           81    69                  35
##  6 Adult     Unknown     17   410          127.  101.                 56.7
```

```
##  7 Adult     Female      18   472          131   104                  49
##  8 Adult     Female      20   155          123.   99.4                51.7
##  9 Adult     Male        18   325          115    94                  45
## 10 Adult     Male        40   475          137   105                  52
## # ℹ 30 more rows
## # ℹ 2 more variables: PL_HingetoPosterior <dbl>, ShellHeightatHinge <dbl>
```

```
#Constructing a least squares regression line that predicts turtles' Annuli
by their Mass
turtlemod = lm(Annuli~Mass, data=`Turtles`)

#Producing a scatterplot of this relationship, while including the least
squares line on the plot
plot(Annuli~Mass, data=`Turtles`)
abline(turtlemod)
```




```
#Creating a residuals vs fitted plots of the annuli
plot(turtlemod$residuals~turtlemod$fitted)
abline(0,0)
```




```
#Looking at a summary of the model we just created
summary(turtlemod)

##
## Call:
## lm(formula = Annuli ~ Mass, data = Turtles)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.4271  -3.9228  -0.9485   2.2938  23.1915
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.084936   1.045886   7.730 1.57e-13 ***
## Mass        0.029571   0.003056   9.675  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.957 on 305 degrees of freedom
## Multiple R-squared:  0.2348, Adjusted R-squared:  0.2323
## F-statistic: 93.61 on 1 and 305 DF,  p-value: < 2.2e-16
```

390 grams = Mass of 40th turtle

Prediction for Annuli = Annuli = 0.029571(Mass) + 8.084936 forty_annuli = 0.029571(390) + 8.084936 forty_annuli = 19.62

```
#Extracting coefficients from the model and calculating annuli prediction for the 40th turtle (which has a mass of 390 grams)
B0_original = summary(turtlemod)$coefficients[1,1]
B1_original = summary(turtlemod)$coefficients[2,1]

fa_P= B1_original * 390 + B0_original
```

40th Turtle Annuli Prediction is = 19.62

```
#Looking at the residual and actual amount of annuli for the 40th observation
turtlemod$resid[40]

##       40
## 20.38223

Turtles$Annuli[40]

## [1] 40
```

The residual for the 40th Turtle Annuli is 20.38 as show above. This means that the actual value of the 40th Annuli is 40, which can also be seen using the Turtles$Annuli[40] function.

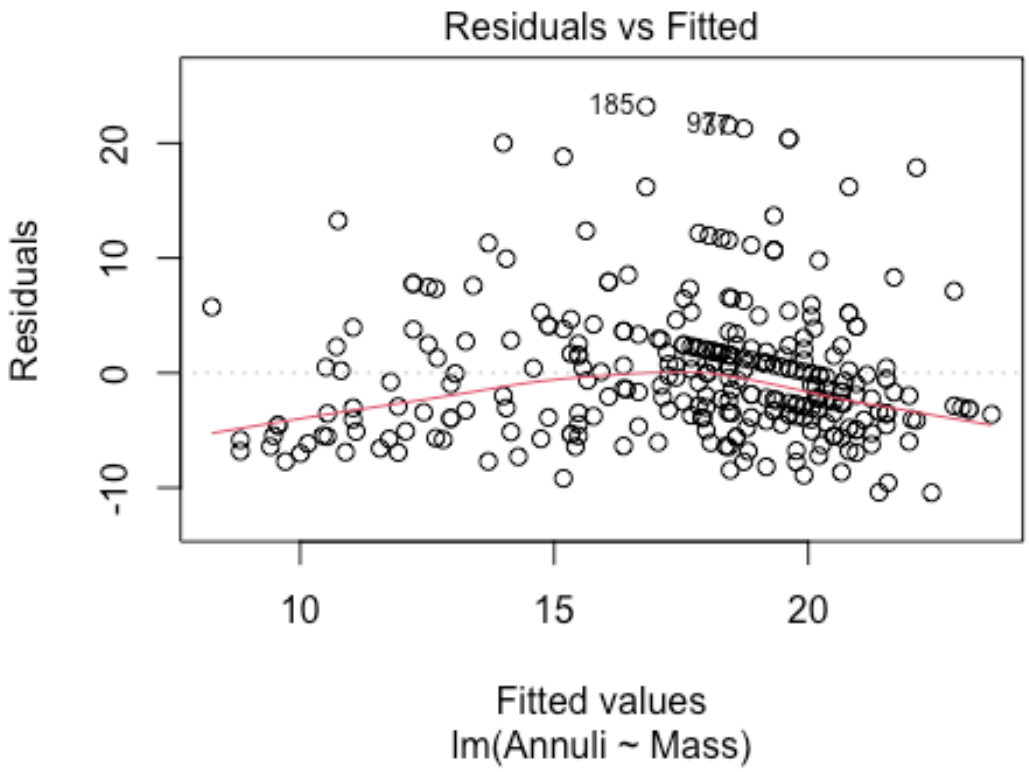
```
#Viewing the max residual for the turtle model and checking which turtle this corresponds to
max(turtlemod$residuals)

## [1] 23.19151

which.max(turtlemod$residuals)

## 185
## 185
```

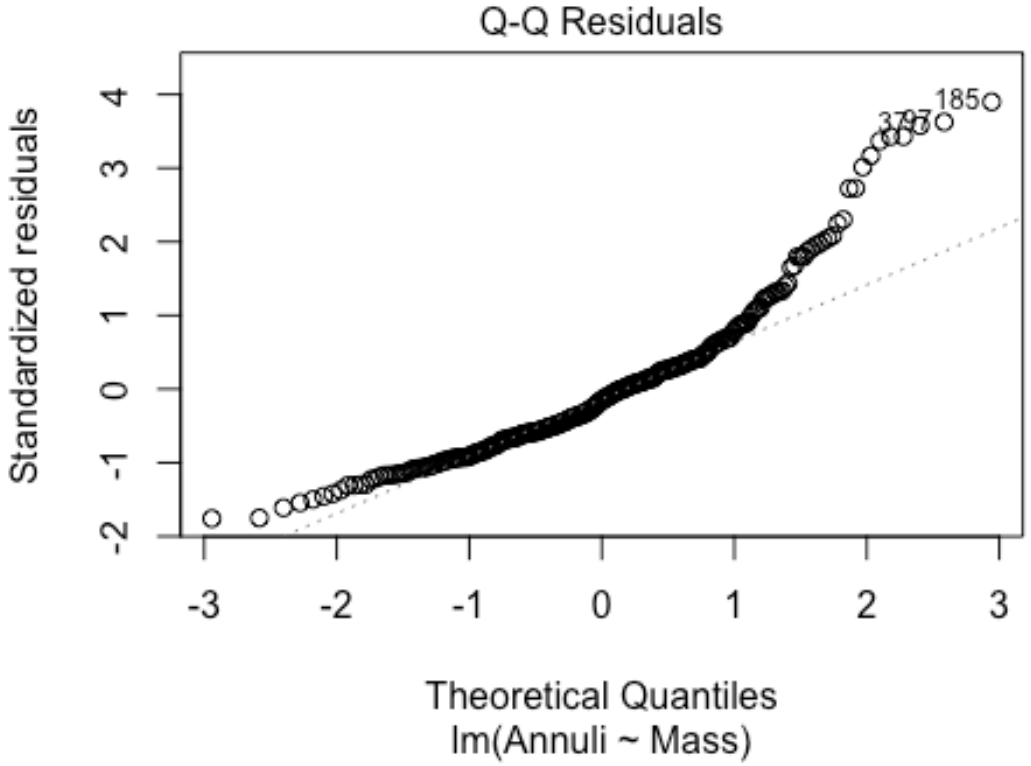
The turtle on 185th row has the largest positive residual. The value is 23.19 as shown above.

```
#Looking at the minimum residual for the turtle model and checking which turtle this corresponds to
min(turtlemod$residuals)

## [1] -10.42705

which.min(turtlemod$residuals)

## 93
## 93
```

The turtle on the 93rd row has the most negative residual. The value is -10.43 as shown above.

```r
#Plotting the turtle model: analyzing residual vs. fitted and QQnorm plots
plot(turtlemod, 1:2)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Annuli ~ Mass)
185
10
15
20
-10
0
10
20



Linearity: Based off of the scatter plot made in question 2, the linearity is very questionable. There seems to be no real trend of the data by the looks of just the scatter plot. There may be a possible positive trend, but nothing definite. The residual plot shown above supports this claim. There is a slight bend.

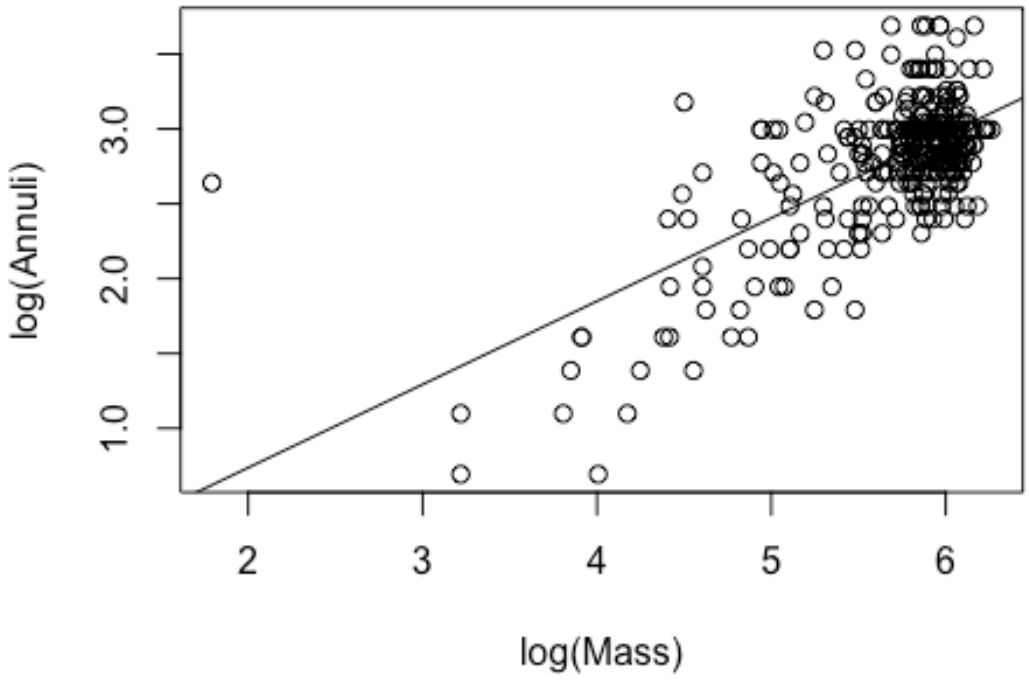
Zero Mean: Because of the linear model formula used, the data is centered around zero.

Constant Variance: Constant variance is definitely questionable for this data. We can tell by looking at both the scatter plot and the residuals vs. fitted plot. The red line in the residuals vs. fitted plot shows a trend of the data. If there was constant variance then this red line would run horizontal across the residuals. There is a greater amount of points below the line, while a smaller amount of points above the line. These points above the line have much larger residuals. This data does not satisfy this condition.

Independence: It is hard to tell independence because we did not gather this data ourselves.

Normality: The normality of the data can be shown in the second graph above. It has 307 observations, which is a fairly large sample. This means that the data should fit the line of normality (dotted line) very well if it is normal. As we can see, the data on the right third of the plot shows very abnormal values and is not close to the line at all. This shows that the plot is not normal and does not satisfy the condition.

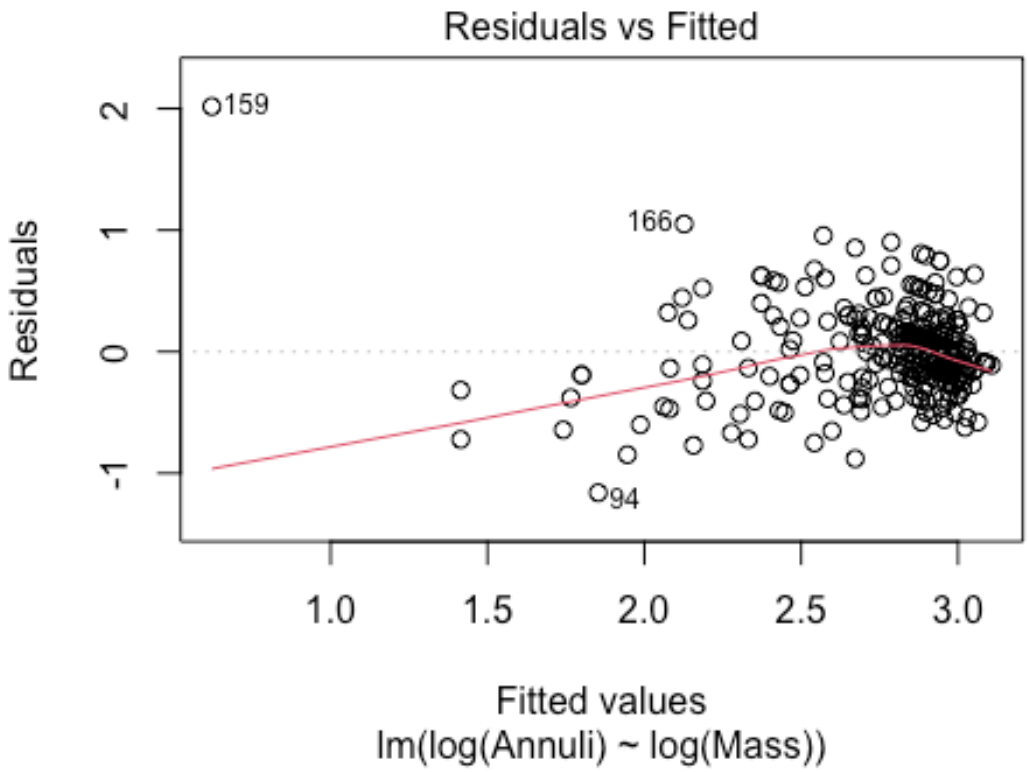
```r
#Plotting the log of both annuli and mass with a new log model line plotted
on this data
plot(log(Annuli)~log(Mass), data=Turtles)
logmod = lm(log(Annuli)~log(Mass), data=Turtles)
abline(logmod)
```

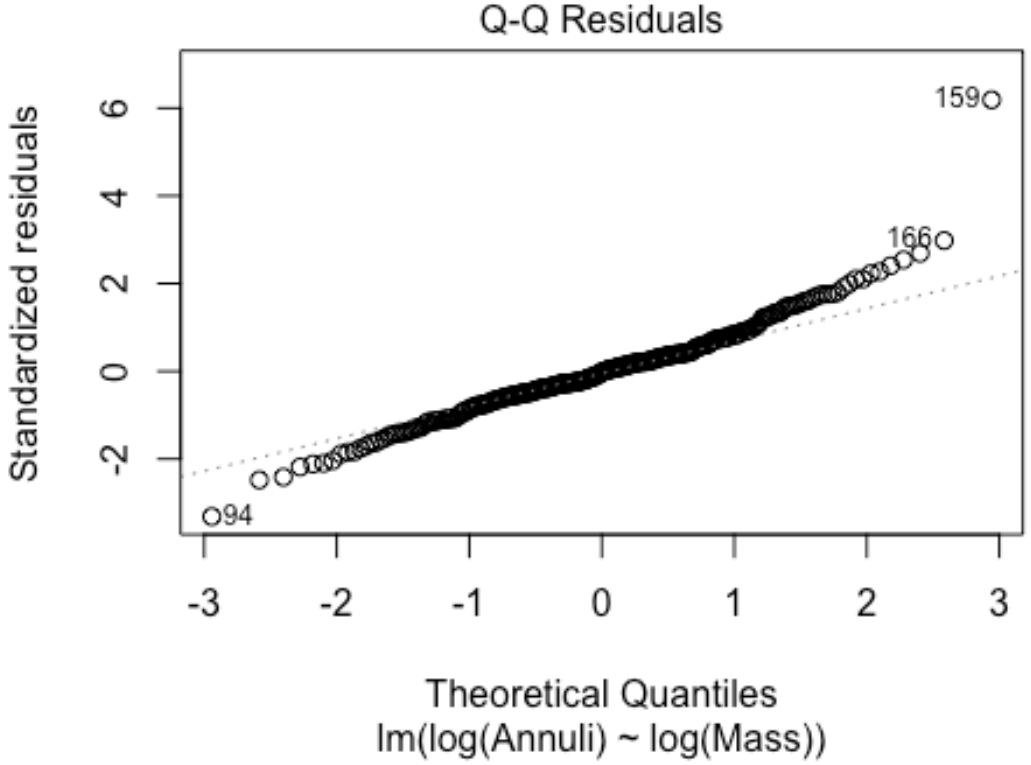


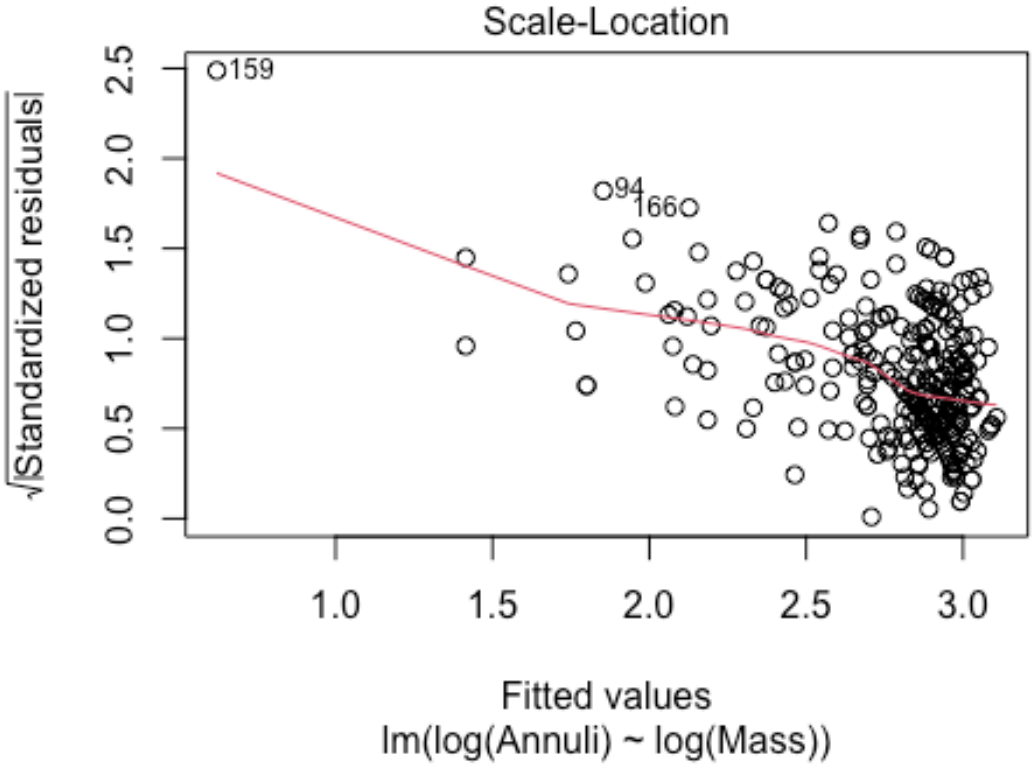

```r
#Looking at a summary of the new model
summary(logmod)
```

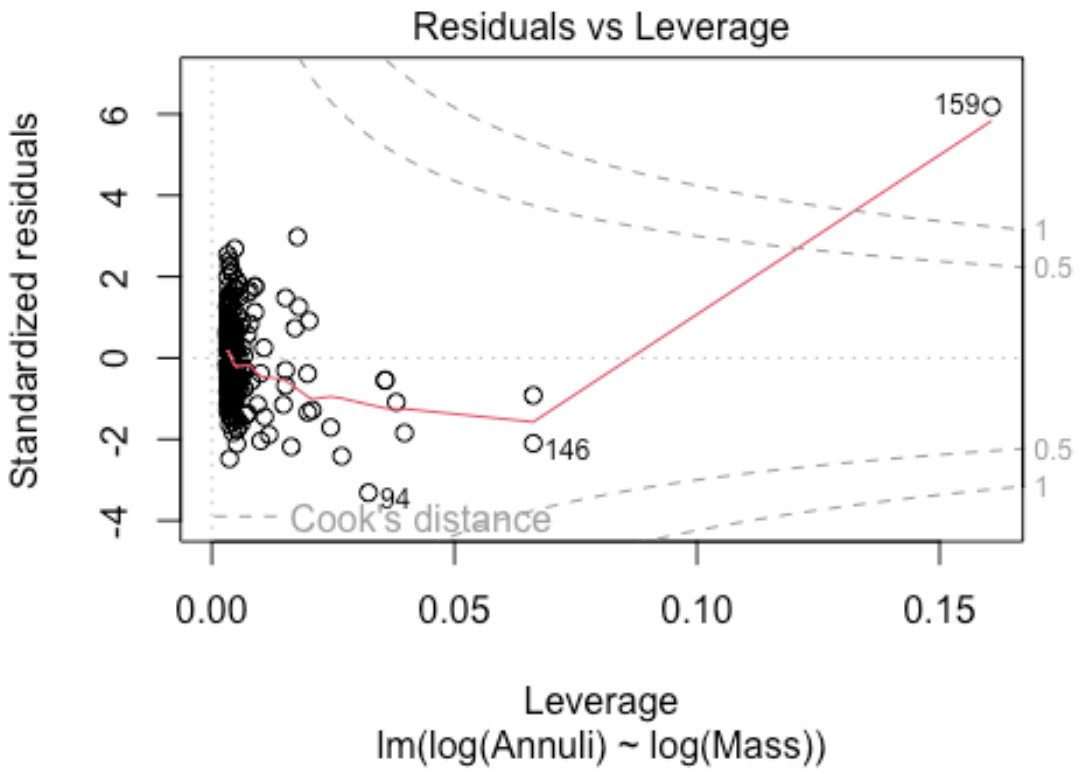
```
##
## Call:
## lm(formula = log(Annuli) ~ log(Mass), data = Turtles)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.15999 -0.19592 -0.00709  0.15929  2.01764
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.37469    0.20741  -1.807   0.0718 .
## log(Mass)    0.55594    0.03638  15.283   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.3559 on 305 degrees of freedom
## Multiple R-squared:  0.4337, Adjusted R-squared:  0.4318
## F-statistic: 233.6 on 1 and 305 DF,  p-value: < 2.2e-16

#Plotting log model to view diagnostics such as residuals vs. fitted and
qqnorm
plot(logmod)
```

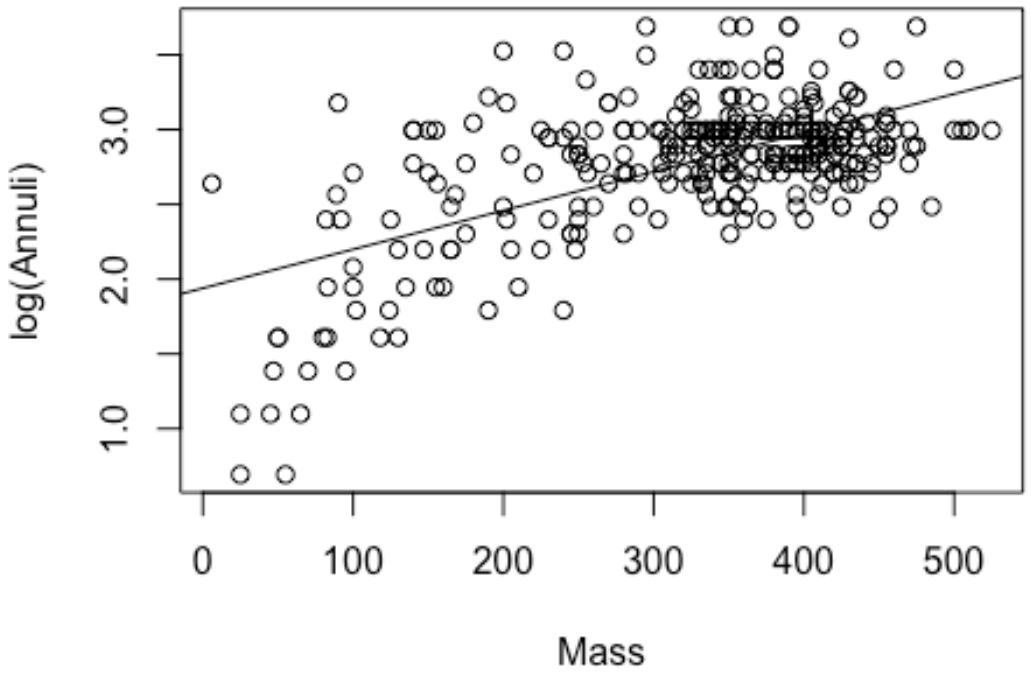

Residuals vs Fitted
159
166
94
Residuals
2
1
0
-1
1.0
1.5
2.0
2.5
3.0
Fitted values
lm(log(Annuli) ~ log(Mass))


Q-Q Residuals
Standardized residuals
-2
0
2
4
6
159
166
94
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(log(Annuli) ~ log(Mass))


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
159
94
166
1.0
1.5
2.0
2.5
3.0
Fitted values
lm(log(Annuli) ~ log(Mass))

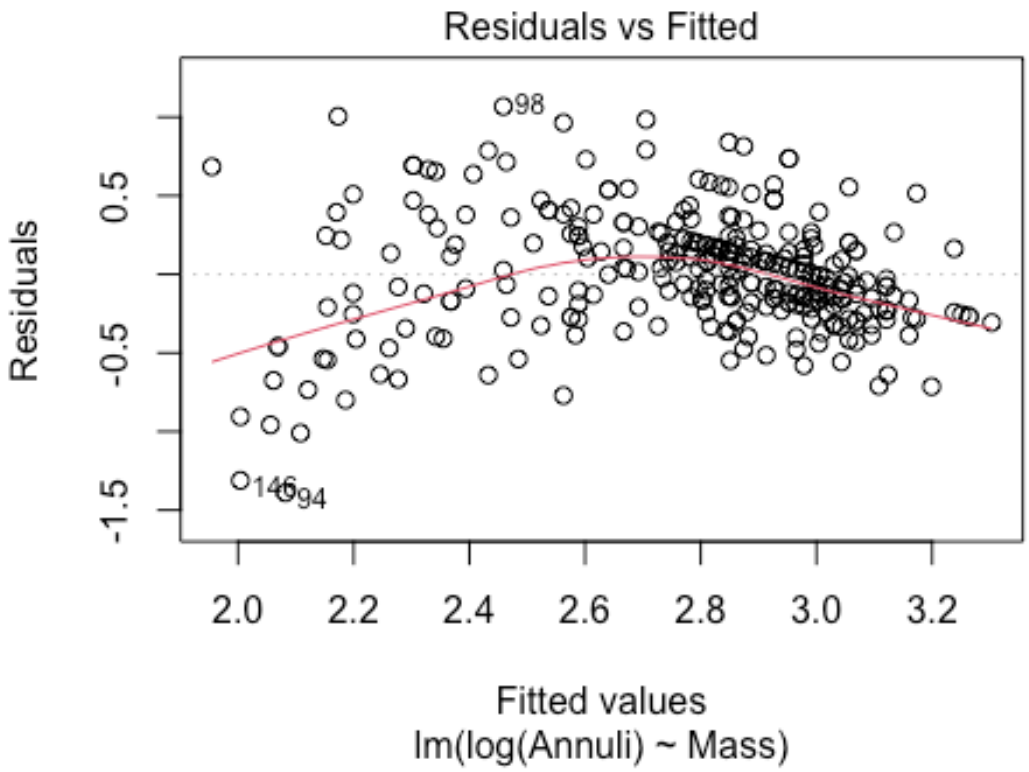


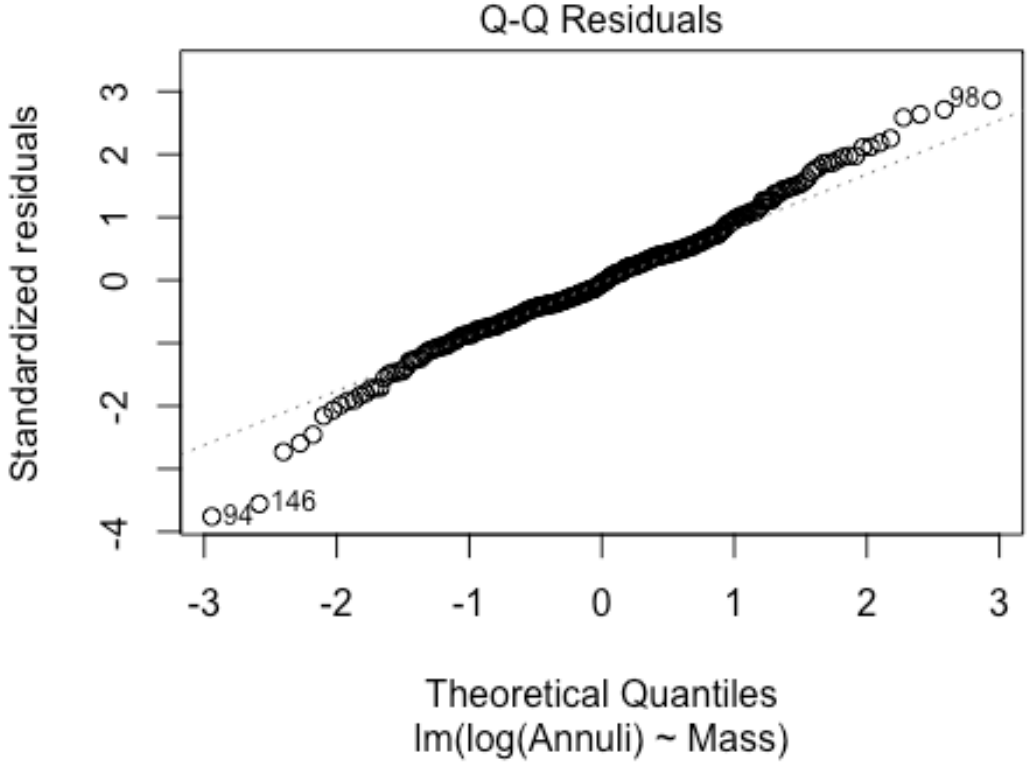

```r
#Plotting a model with log annuli against mass (no log included for mass)
along with the new log annuli model line plotted with it
plot(log(Annuli)~Mass, data=Turtles)
alogmod = lm(log(Annuli)~Mass, data=Turtles)
abline(alogmod)
```

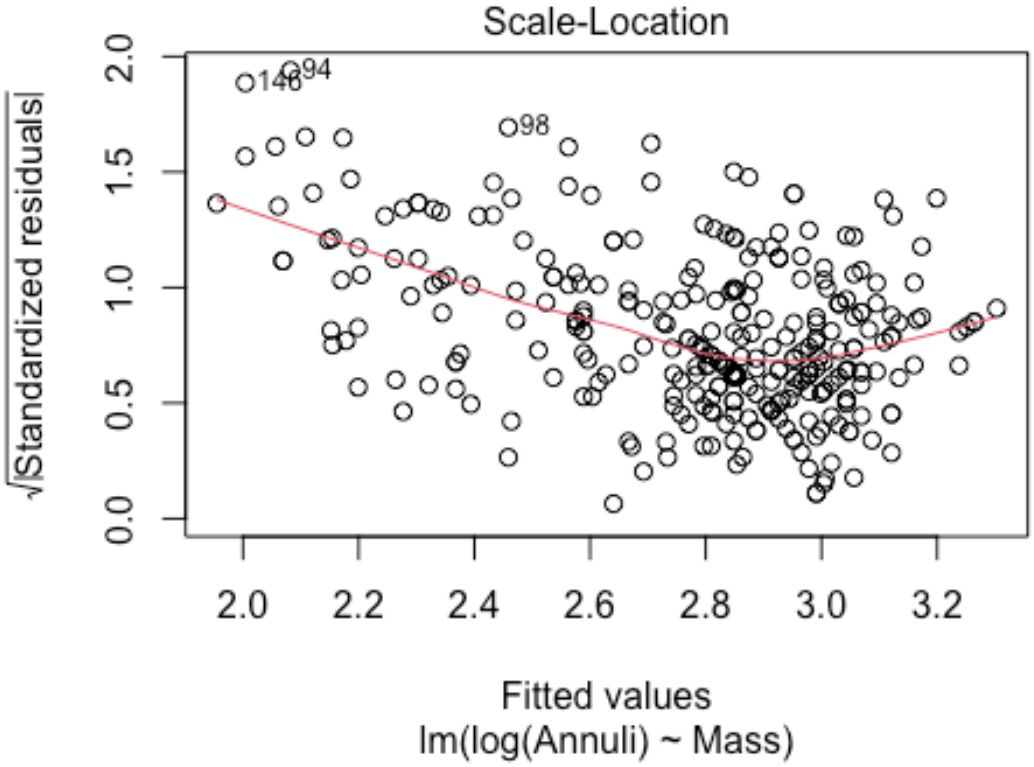


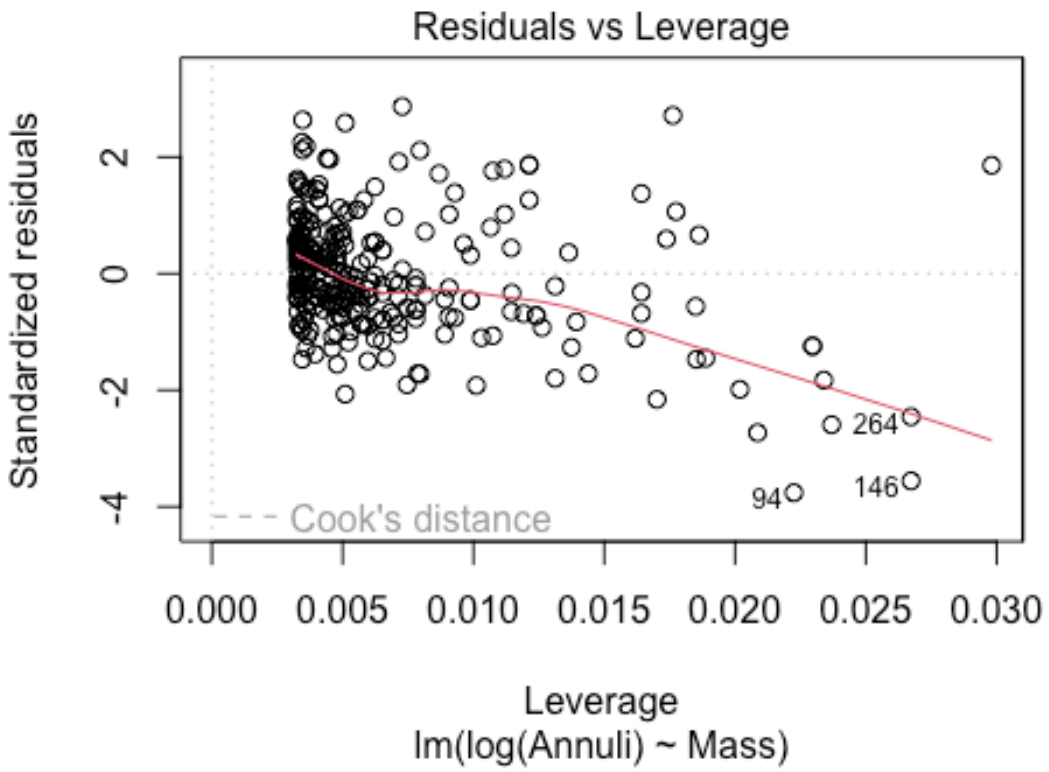

```
#Looking at a summary of the new model
summary(alogmod)

##
## Call:
## lm(formula = log(Annuli) ~ Mass, data = Turtles)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.38902 -0.23076 -0.02038  0.20190  1.06757
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.9393076  0.0656079   29.56   <2e-16 ***
## Mass        0.0025974  0.0001917   13.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3737 on 305 degrees of freedom
## Multiple R-squared:  0.3757, Adjusted R-squared:  0.3736
## F-statistic: 183.5 on 1 and 305 DF,  p-value: < 2.2e-16
```

```r
#Plotting log annuli model to check diagnostics such as residuals vs. fitted and qqnorm plots. This is helping to check if both of these models fit the conditions better
plot(alogmod)
```


Residuals vs Fitted
Residuals
0.5
-0.5
-1.5
98
146
94
2.0
2.2
2.4
2.6
2.8
3.0
3.2
Fitted values
lm(log(Annuli) ~ Mass)

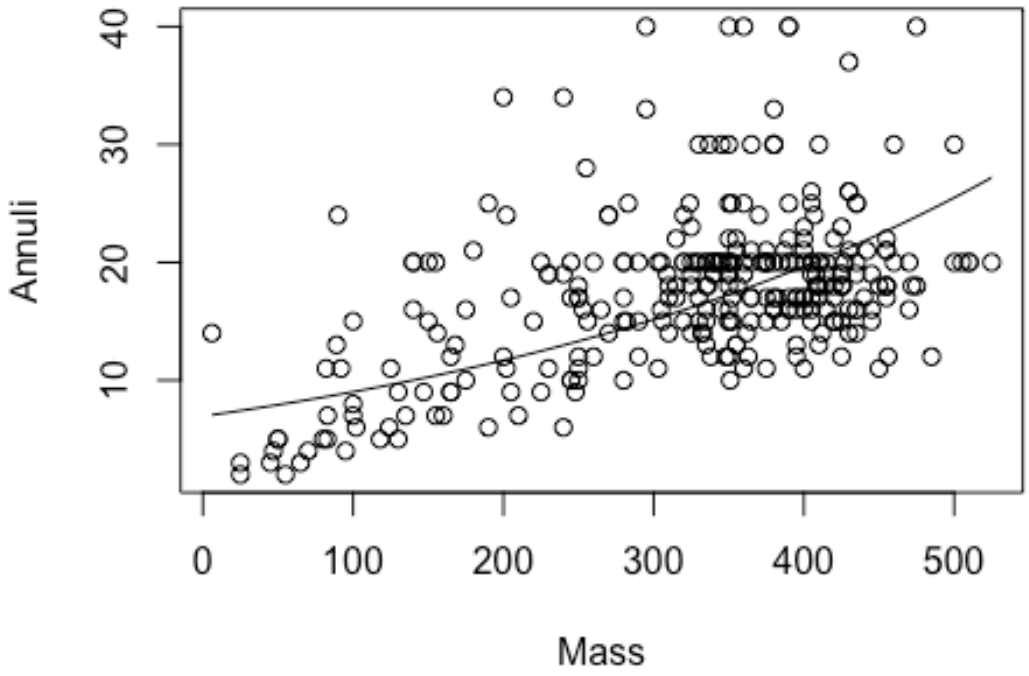

Q-Q Residuals
Standardized residuals
-4
-2
0
1
2
3
94
146
98
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(log(Annuli) ~ Mass)


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
146
94
98
2.0
2.2
2.4
2.6
2.8
3.0
3.2
Fitted values
lm(log(Annuli) ~ Mass)



I chose to use my second transformation to analyze further (log(Annuli) as y, Mass as x). While the residual plot looks very similar to the original, the normality seems to be closer to what we are looking for in the qqnorm plot. The scatter plot also looks more linear than the original.

```
#Plotting the base data along with the new curve we have created from the
transformations in the last model
B0 = summary(alogmod)$coefficients[1,1]
B1 = summary(alogmod)$coefficients[2,1]

plot(Annuli~Mass, data=Turtles)
curve(exp(B1*x+B0), add=TRUE)
```



390 grams = Mass of 40th turtle

Prediction for Annuli with new curve= annuli = e^(B1*mass + B0) forty_annuli = e^(0.0026*mass+1.94)

```
#Calculating the annuli prediction with the new curve
fa_NP = exp(B1*390+B0)
```

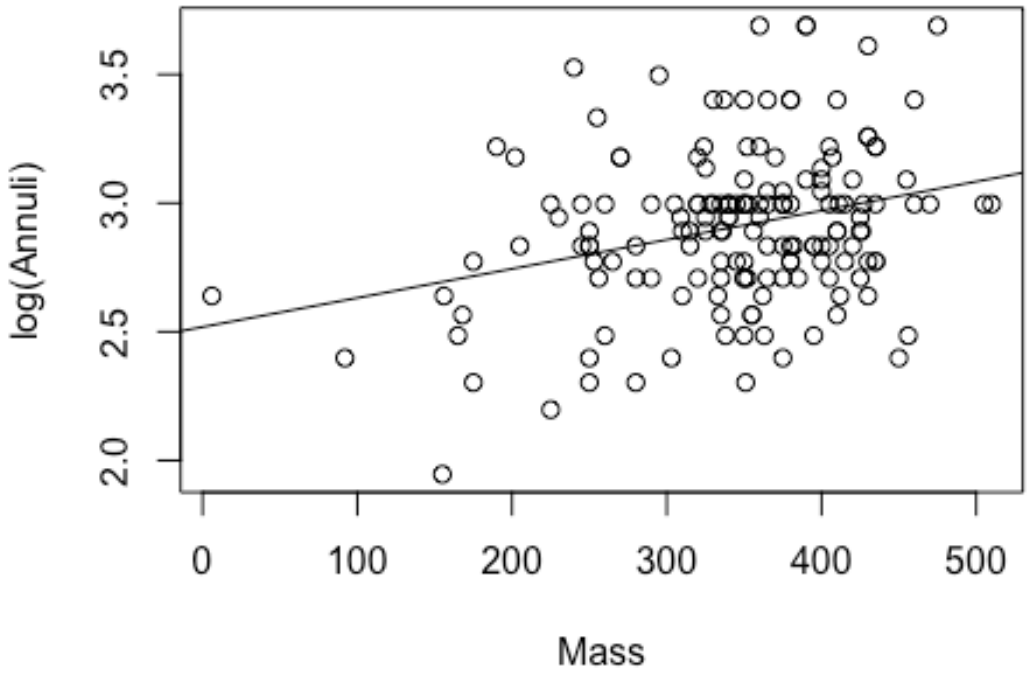
fa_NP = 19.15

The new prediction for the annuli is 19.15. The observed value of the annuli was 40. 40-19.15 = 20.85

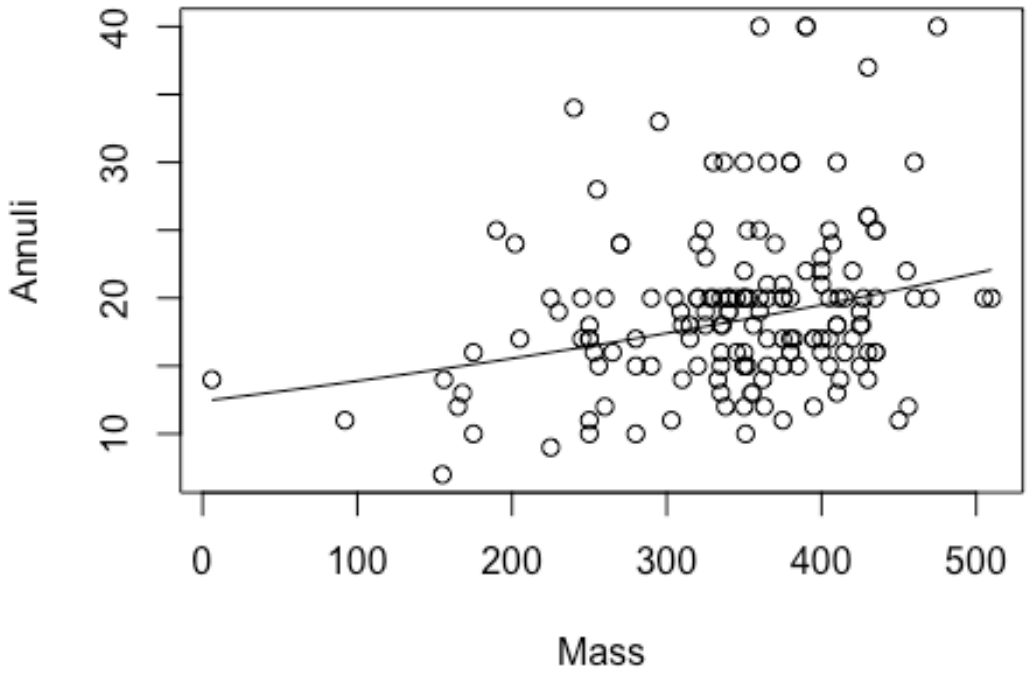
The new prediction is 20.85 away from the observed value, compared to the previous residual of 20.38. These are very similar numbers.

```
#Constructing two new dataframes, one with adult males and another with adult females. A beginning "Adult" subset was created, but this was solely used to make the creation of the 'Male' and 'Female' data sets easier.
Turtles_adult=subset(Turtles, LifeStage=='Adult')
Turtles_adult_male=subset(Turtles_adult, Sex=='Male')
Turtles_adult_female=subset(Turtles_adult, Sex=='Female')
```

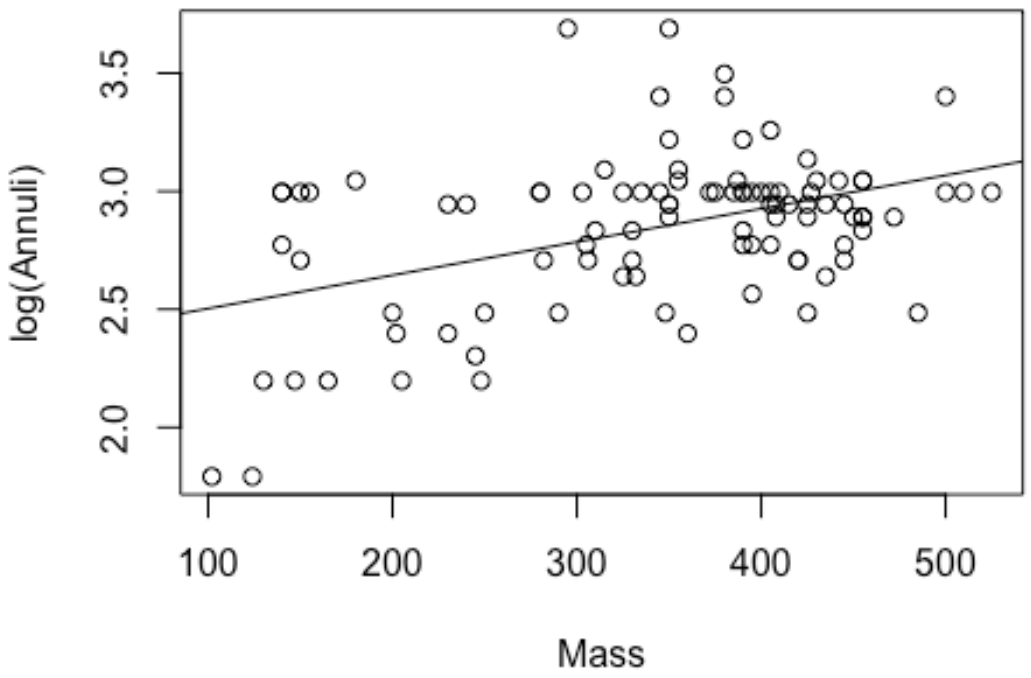
```r
#Using the best transformation, plotting the adult male turtle data frame in
the same manner, extracting coefficients and plotting the new curve
plot(log(Annuli)~Mass, data=Turtles_adult_male)
malemod = lm(log(Annuli)~Mass, data=Turtles_adult_male)
abline(malemod)
```



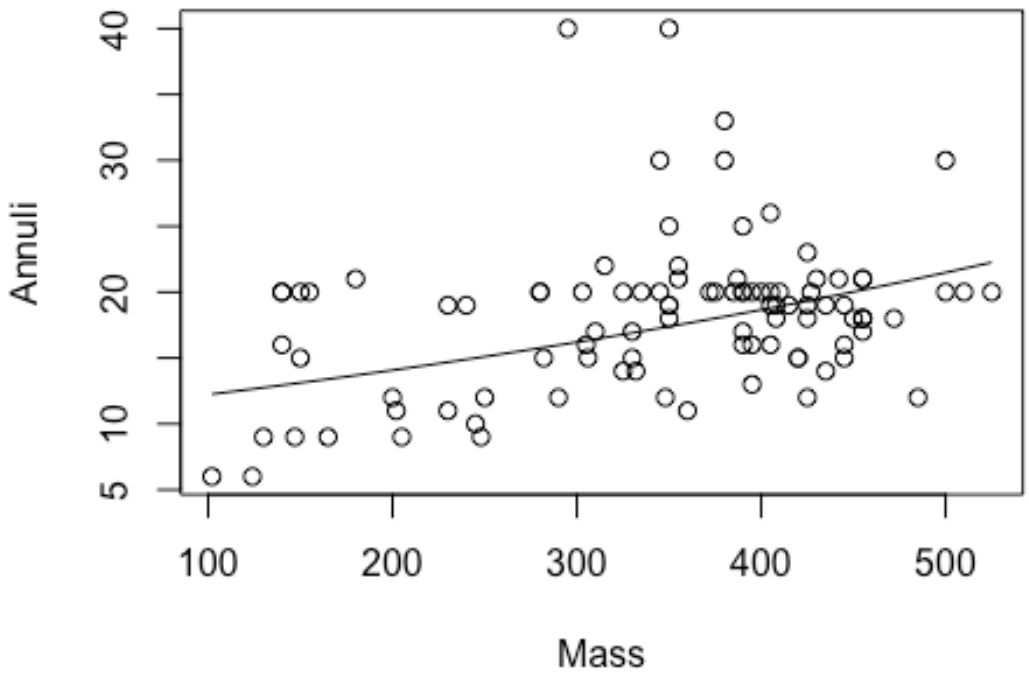

```r
B0_male = summary(malemod)$coefficients[1,1]
B1_male = summary(malemod)$coefficients[2,1]

plot(Annuli~Mass, data=Turtles_adult_male)
curve(exp(B1_male*x+B0_male), add=TRUE)
```



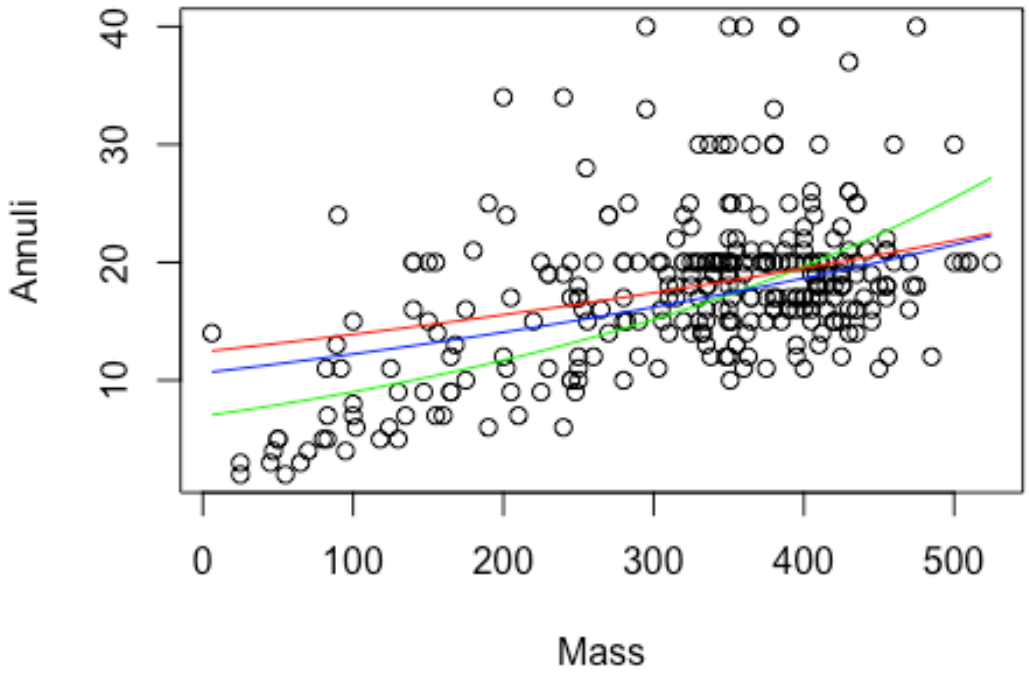

```
#Using the best transformation, plotting the adult female turtle data frame
in the same manner, extracting coefficients and plotting the new curve
plot(log(Annuli)~Mass, data=Turtles_adult_female)
mod5 = lm(log(Annuli)~Mass, data=Turtles_adult_female)
abline(mod5)
```




```
B0_female = summary(mod5)$coefficients[1,1]
B1_female = summary(mod5)$coefficients[2,1]

plot(Annuli~Mass, data=Turtles_adult_female)
curve(exp(B1_female*x+B0_female), add=TRUE)
```




```
#Plotting a summary of the key information and models that we have found so
far. Using this to come up with conclusions about the data. We can see the
original (full data set) plotted in green, the adult male data set plotted in
red, and the adult female data set plotted in blue.
plot(Annuli~Mass, data=Turtles)
curve(exp(B1*x+B0), add=TRUE, col='green')
curve(exp(B1_male*x+B0_male), add=TRUE, col='red')
curve(exp(B1_female*x+B0_female), add=TRUE, col='blue')
```



This visual tells us that male turtles have a larger annuli to mass ratio, on average, than females. We can see the red curve stays slightly above the blue curve in the whole graph. This would cause, on average, an overestimation of female annuli compared to males if we were to use the original dataset. As the turtles increase in mass, the difference in annuli decreases and they seem to cross near the 500 mass mark.