STOR 455 Homework #2

20 points - Due Thursday 09/14 at 12:30pm

Directions: You may (and should) collaborate with other students. However, you must complete the assignment by yourself. You should complete this 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 (it is usually smoother first knit to Word then save the file as pdf) and submit the file to Gradescope. The submitted PDF should not be longer than 20 pages.

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 under "Resources" on Sakai) contains data collected at The Piedmont Wildlife Center in Durham.

Questions

Q1

Consider regressing **Annuli** on **Mass** with the full dataset. 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 - with commentary on what each plot tells you specifically about the appropriateness of conditions.

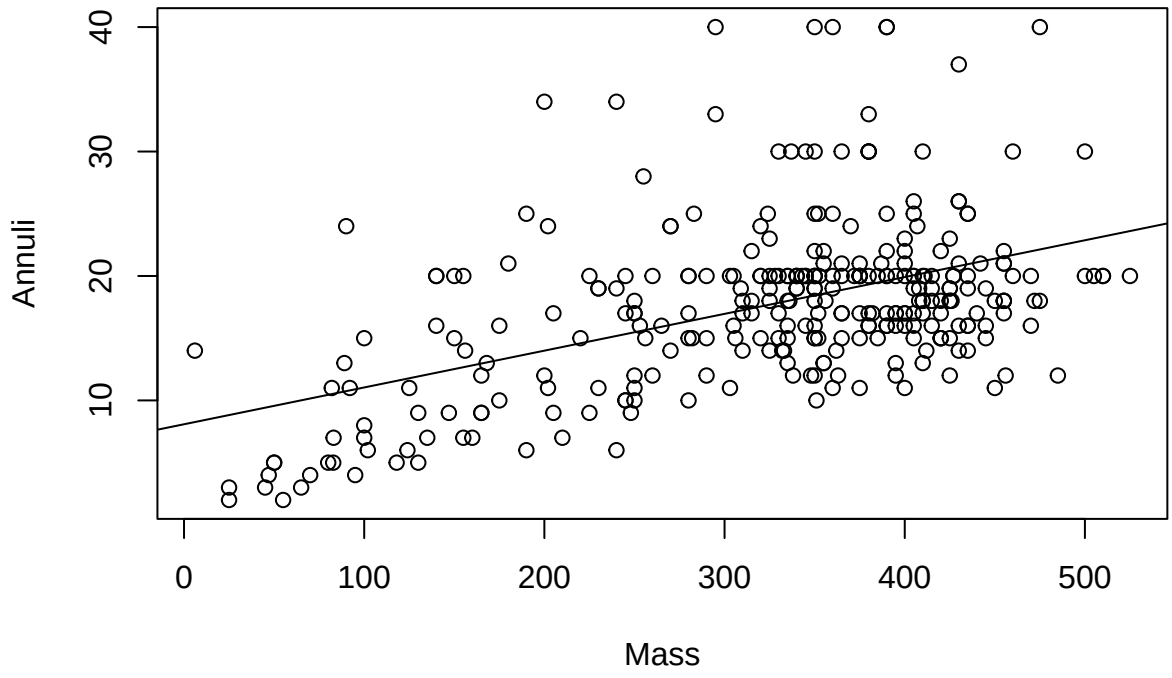
The following graphs show the linearity test of Annuil on Mass through three distinct plots. The first plot is the standard least square regression test, which shows many thing such as the R^2 indicating the goodness of fit. In the example with a .2348 R^2, the model do not explain any of the variability in the dependent variable, and the model provides no predictive value. Also, the Residual vs Fitted values indicated linearity holding, which is represented by the qq plot of the data following closely to the path of the dashed line showing normality. Normality of residuals shows the assumption of linearity holding.

```
library(readr)
#
setwd("C:/Users/Jabbir Ahmed/Desktop/Data")
Turtles <- read.csv("Turtles.csv")
plot(Annuli~Mass, data = Turtles)
mod2 <- lm(Annuli ~ Mass, data = Turtles)
summary(mod2)
```

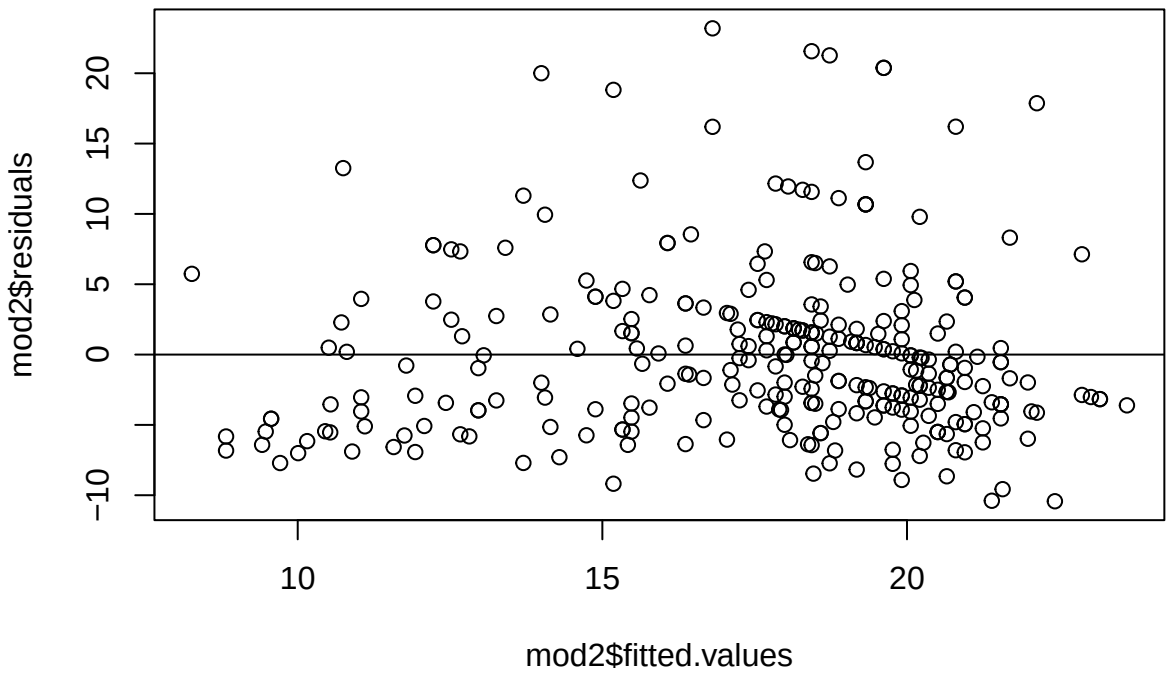
```
##
## Call:
```

```
## 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
```

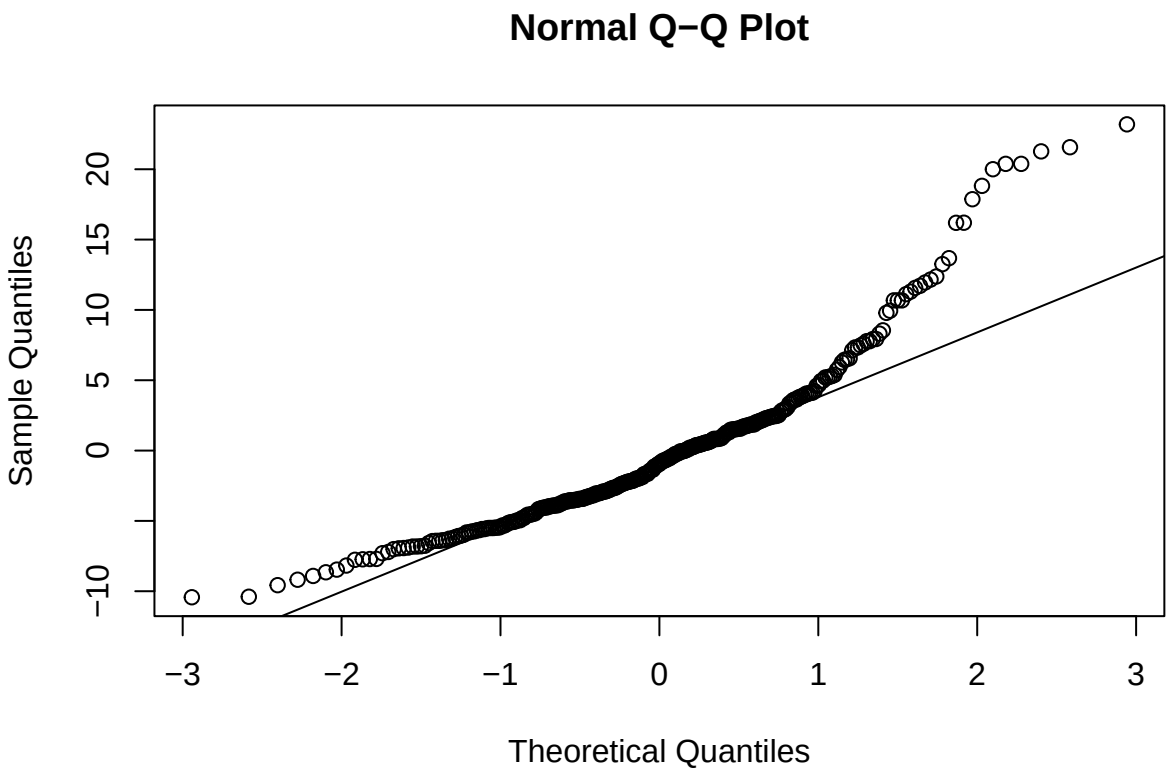
```
abline(mod2)
```




```
plot(mod2$residuals~mod2$fitted.values)
abline(0,0)
```




```
qqnorm(mod2$residuals)
qqline(mod2$residuals)
```

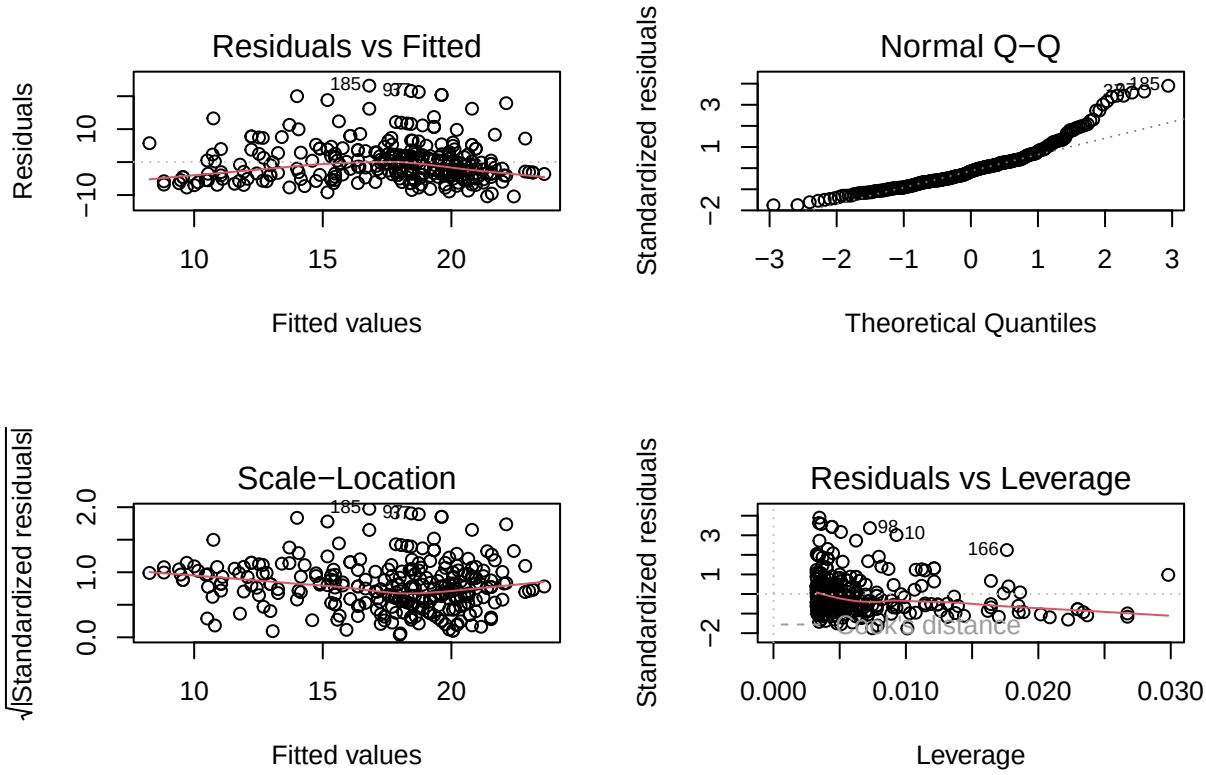


Q2

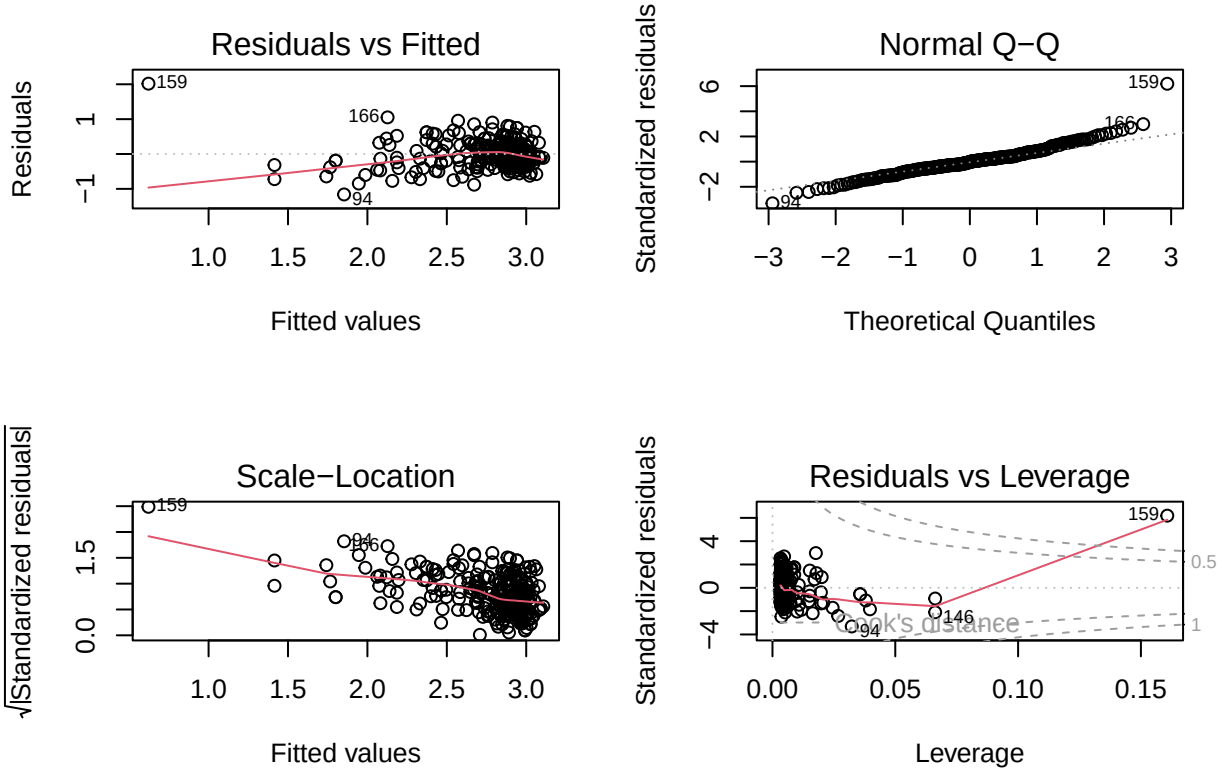
Experiment with at least two transformations to determine if models constructed with these transformations appear to do a better job of satisfying each of the simple linear model conditions. Include the summary outputs for fitting these models and scatterplots of the transformed variable(s) with the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

Using the transformations of square root and logarithms, The logarithms does a better job of satisfying each of the simple linear model conditions because they have similar normality and residual vs fitted take, but the R-squared being closer to 1, indicates a stronger relationship between the variability in the dependent variable, and the model is an better fit to the data.

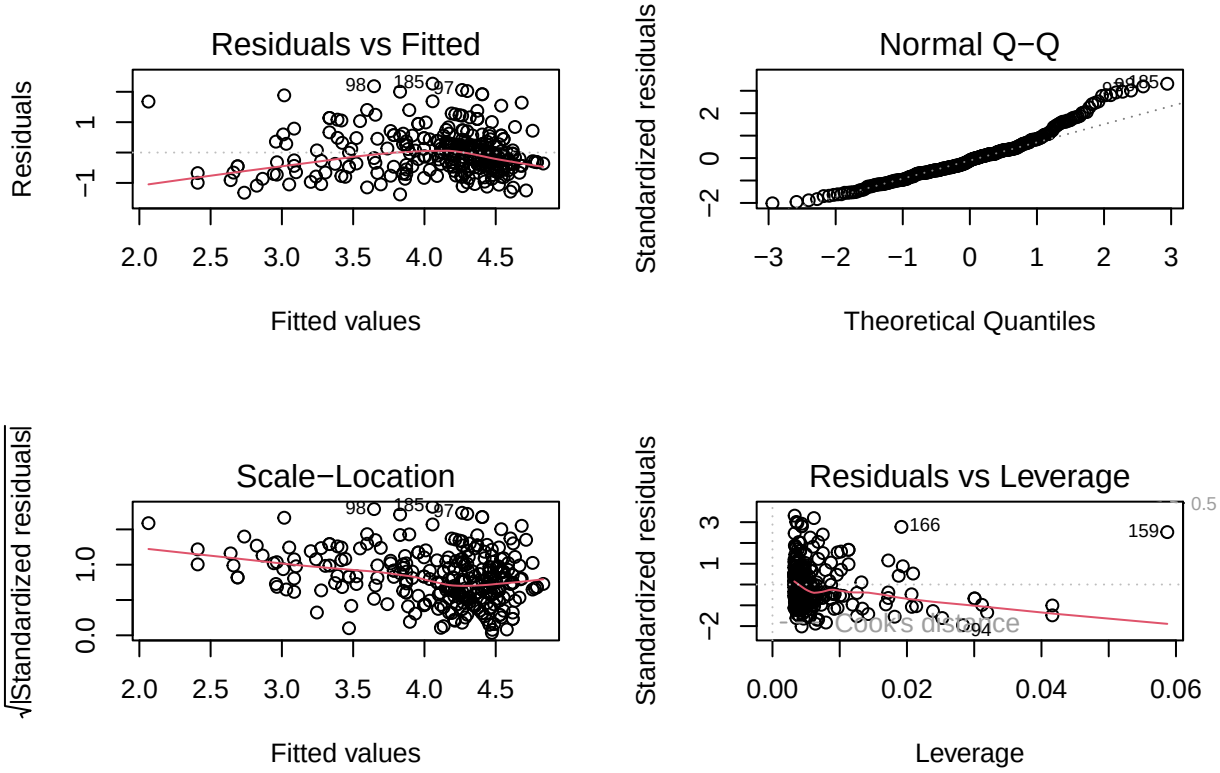
```
#
par(mfrow=c(2,2))
plot(mod2)
```




```
mod3<- lm(log(Annuli)~log(Mass), data=Turtles)
plot(mod3)
```




```
mod4<- lm(sqrt(Annuli)~sqrt(Mass), data=Turtles)
plot(mod4)
```




```
summary(mod2)
```

```
##
## 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
```

```
summary(mod3)
```

```
##
## 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
```

```
summary(mod4)
```

```
##
## Call:
## lm(formula = sqrt(Annuli) ~ sqrt(Mass), data = Turtles)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.37970 -0.44677 -0.06799  0.29990  2.26750
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.73246    0.19158   9.043   <2e-16 ***
## sqrt(Mass)   0.13534    0.01065  12.709   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6856 on 305 degrees of freedom
## Multiple R-squared:  0.3462, Adjusted R-squared:  0.3441
## F-statistic: 161.5 on 1 and 305 DF,  p-value: < 2.2e-16
```

Q3

For your model with the best transformation from question 2 (It still may not be an ideal model), plot the raw data (not transformed) with the model (likely a curve) on the same axes.

```
#
coef(mod3)[1]
```

```
## (Intercept)
##  -0.3746944
```

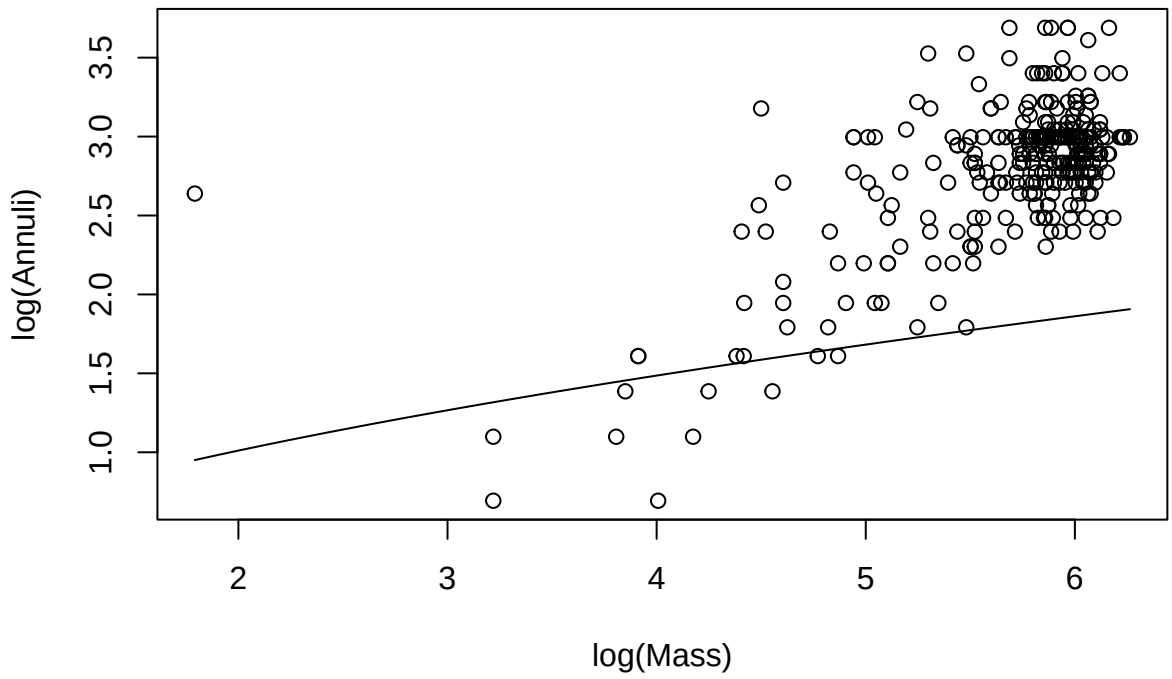
```
coef(mod3)[2]
```

```
## log(Mass)
## 0.5559381
```

```r
exp(-0.37469)
```

```
## [1] 0.6875024
```

```r
plot(log(Annuli)~log(Mass), data=Turtles)
curve( 0.6875024*(x^0.5559381), add = TRUE)
```



Q4

The turtle in the first row of the *Turtles* dataset has a mass of 410 grams. For your model using the best transformation from question 2, 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?

The predicted Annuli from the mass of 410 grams from the transformed model is 19.49038, but the observed value from the data is 13. Meaning that there is a -6.49038 residual, which is better than the original non-transformed model, which was -7.209198. It shows that the transformed model is 0.718818 better than the orginal model.

```r
#
Turtles[1,3]
```

```
## [1] 13
```

```
 0.6875024*(410^0.5559381)
```

```
## [1] 19.49038
```

```
 13-19.49038
```

```
## [1] -6.49038
```

```
Turtles[1,10]
```

```
## NULL
```

```
 -6.49038- -7.209198
```

```
## [1] 0.718818
```

Q5

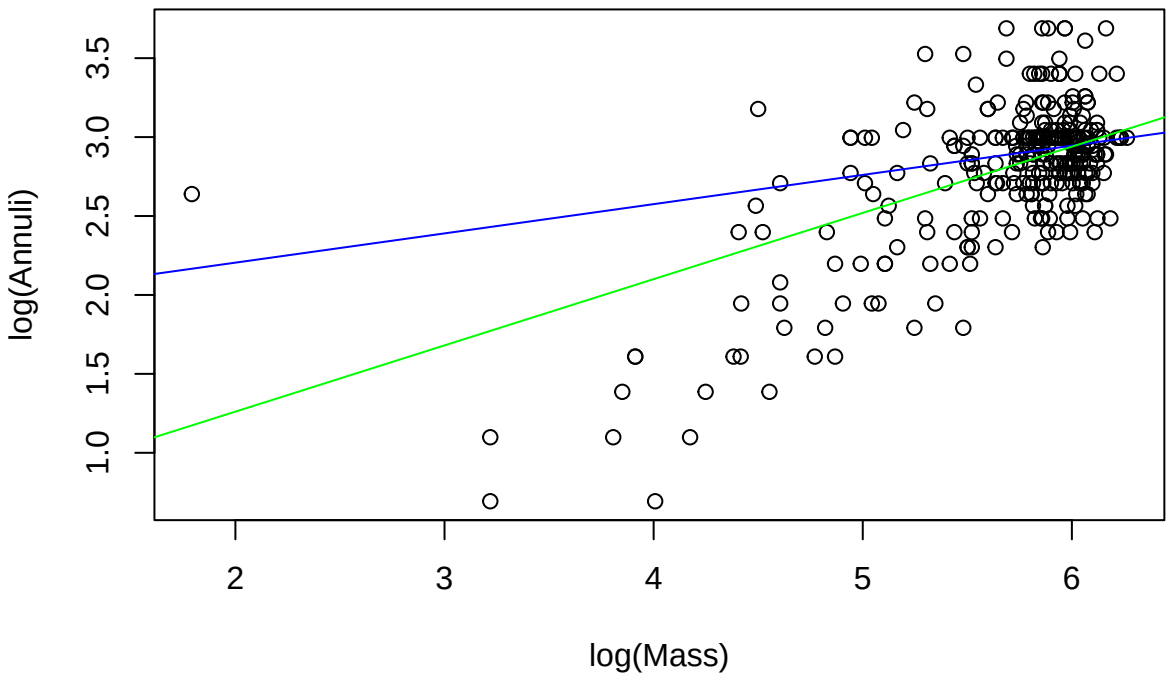
For your model using the best transformation from question 2, could the relationship between **Mass** and **Annuli** be different depending on the **LifeStage** and **Sex** of the turtle? Construct two new dataframes, one with only adult male turtles, and one with only adult female turtles. Using your best transformation from question 2, 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?

The relationship between Mass and Annuli depending on the sex of male is higher than females. It means that the greater the mass of males and higher annuli that are greater than the females with the same mass.

```
#
adult_males_turtles<-(subset(Turtles, LifeStage == "Adult" & Sex == "Male"))
adult_females_turtles<-(subset(Turtles, LifeStage == "Adult"& Sex == "Female"))

mod5<-lm(log(Annuli)~log(Mass), data=adult_males_turtles)
mod6<-lm(log(Annuli)~log(Mass), data=adult_females_turtles)

plot(log(Annuli)~log(Mass), data= Turtles)
abline(mod5, col = "blue")
abline(mod6, col = "green")
```


log(Annuli)
1.0
1.5
2.0
2.5
3.0
3.5
2
3
4
5
6
log(Mass)