STOR 455 Homework #1

20 points - Due Thursday 1/26 at 12:00pm

```
library(readr)
library(mosaic)
turtles <- read_csv("Turtles.csv", show_col_types = FALSE)
```

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

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*.

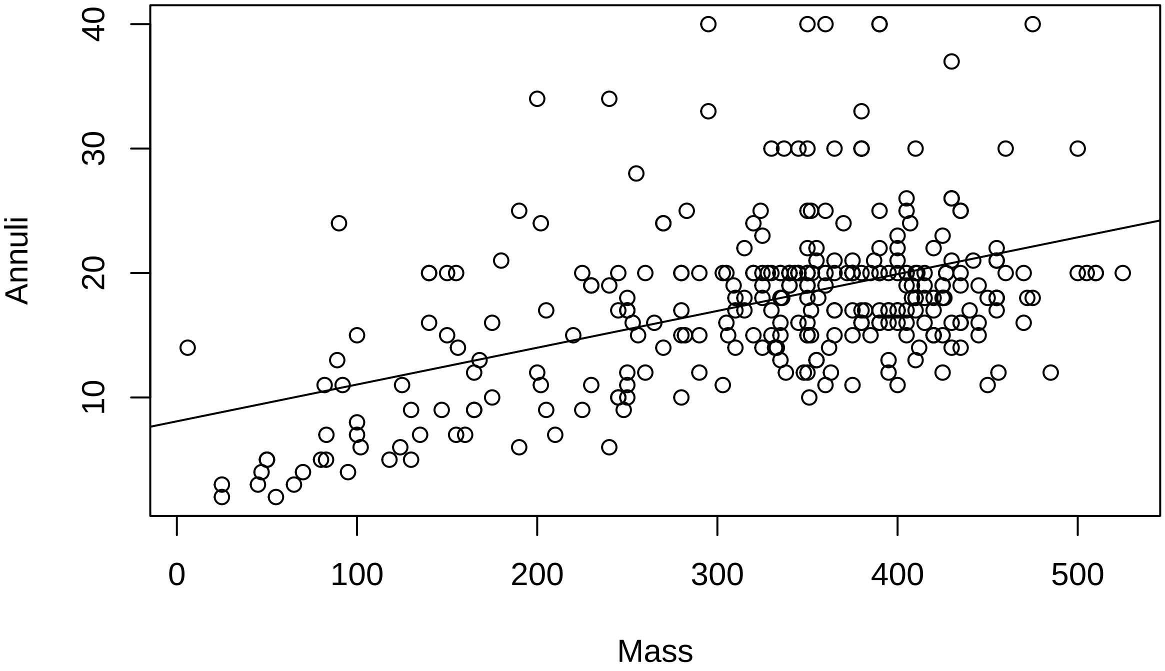
```
annuli_mass_model <- lm(Annuli~Mass, data = turtles)
summary(annuli_mass_model)
```

```
##
## 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
```

Fitted Regression Model: (Predicted Annuli) = 8.085 + 0.030(Mass)

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

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



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?

```
# Fitted Value
annuli_mass_model$fitted.values[40]
```

```
##       40
## 19.61777
```

```
# Residual
annuli_mass_model$residuals[40]
```

```
##       40
## 20.38223
```

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

```
turtle_residuals <- annuli_mass_model$residuals

which.max(turtle_residuals)
```

```
## 185
## 185
```

```
max(turtle_residuals)
```

```
## [1] 23.19151
```

Turtle number 185 and the value of that residual is 23.19151.

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

```
which.min(turtle_residuals)
```

```
## 93
## 93
```

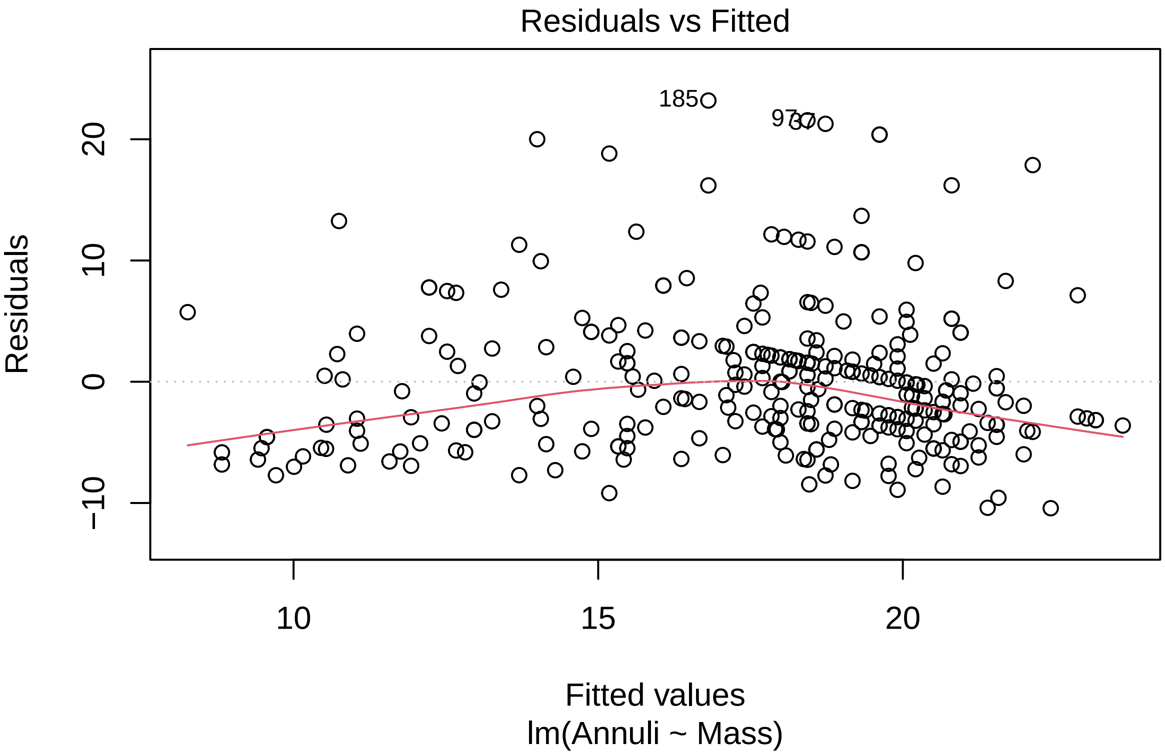
```
min(turtle_residuals)
```

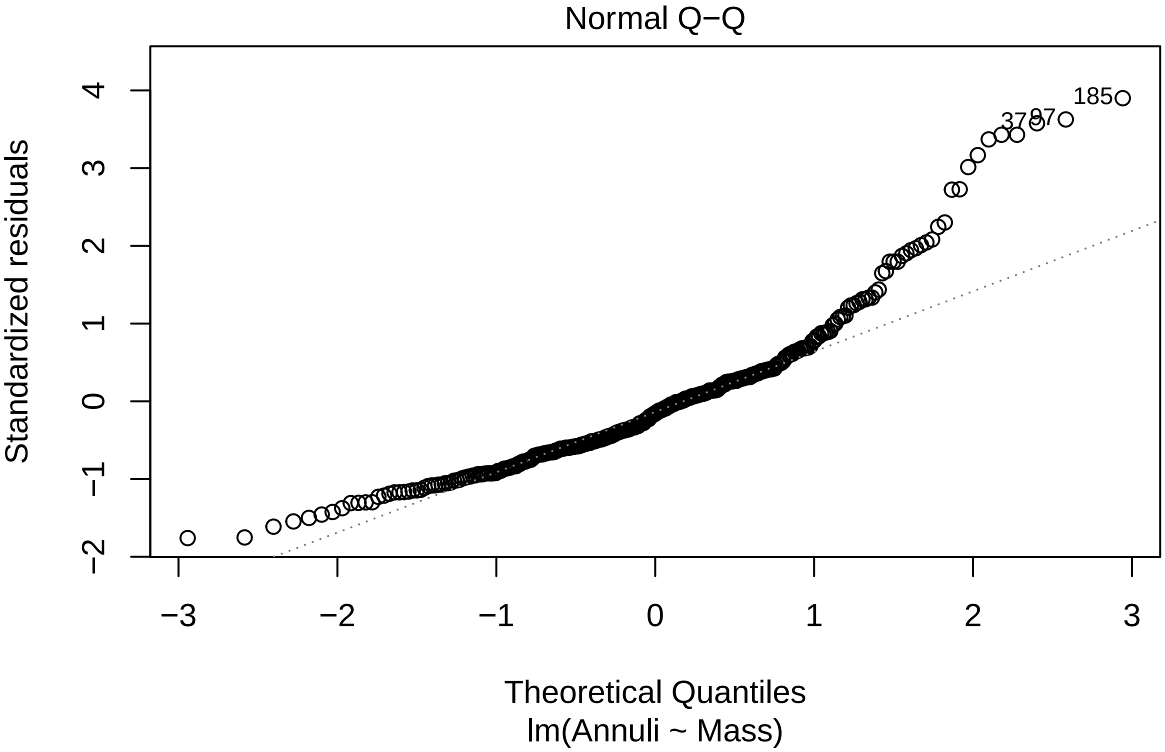
```
## [1] -10.42705
```

Turtle number 93 and the value of that residual is -10.42705

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.

```
plot(annuli_mass_model, 1:2)
```

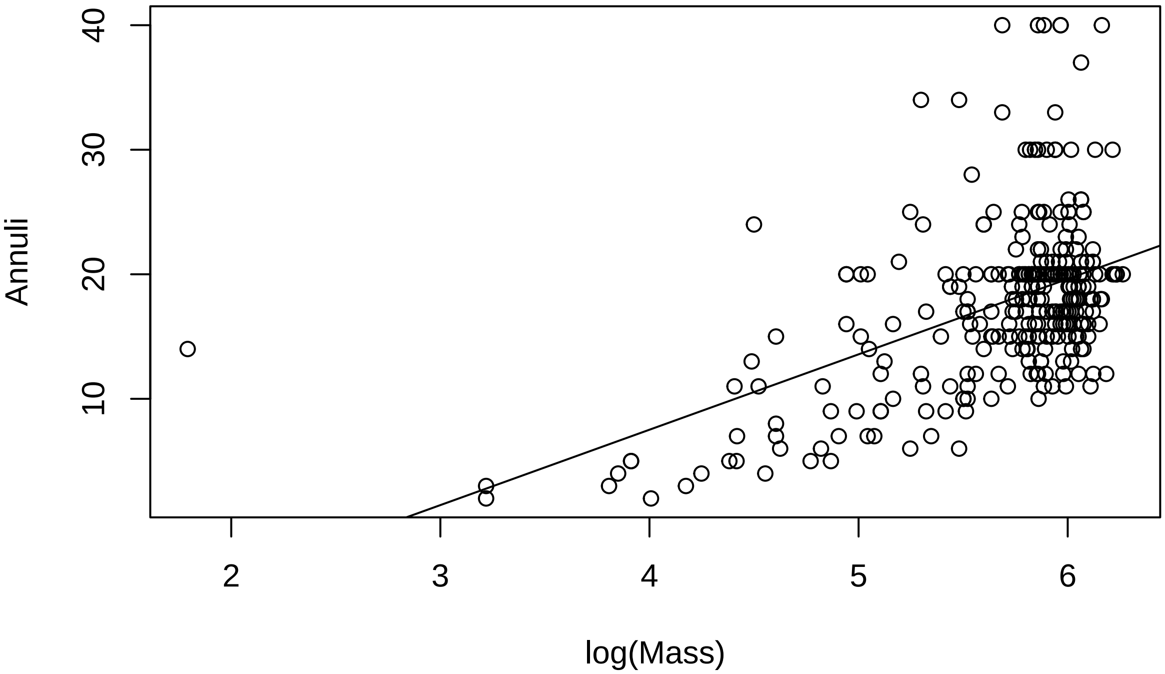




Through the residuals vs fitted values plot we can see that the error distribution is centered around zero and that there is no noticeable pattern in the residuals. Although there is no noticeable pattern to the eyes, the residual line shows a curvature in the residuals, so a transformation in the data could be helpful to minimize this pattern. In addition to the curvature, there is a high variation, so that could be minimized through a transformation. Also through the qqplot we can see that the distribution deviates away from the linear pattern, indicating some form of skew. So it shows an overall linear pattern and uniform spread but making inferences on this data would be difficult as it does not follow a normal distribution and has skewedness.

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.

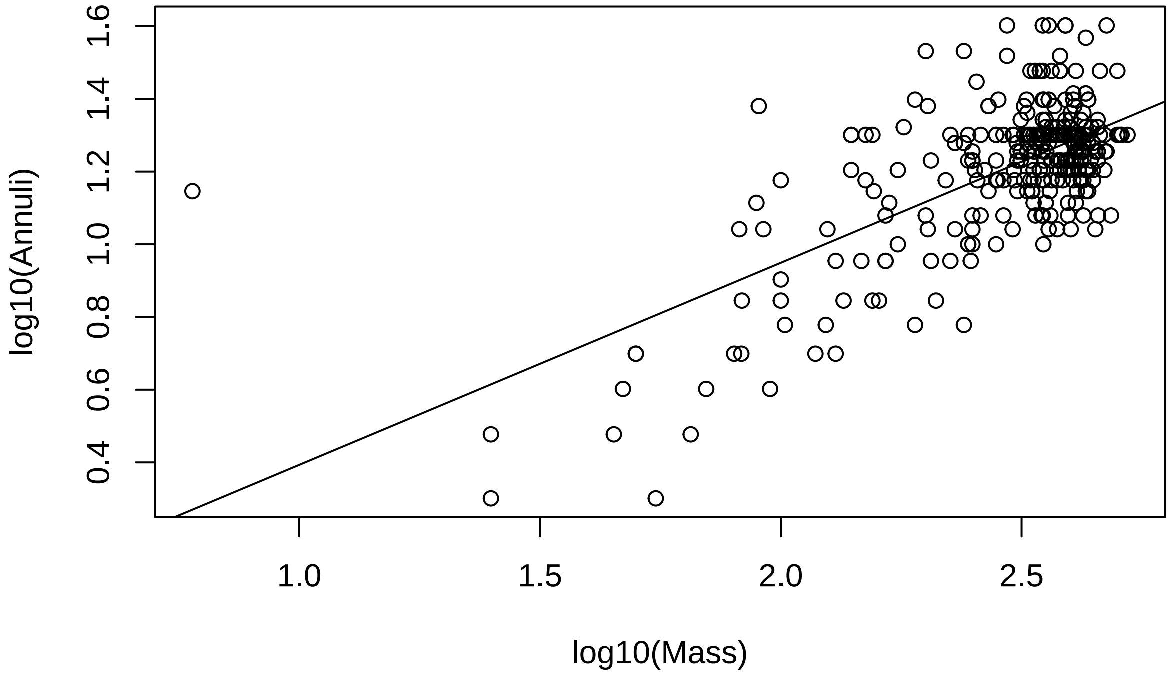
```
# Logged the mass with base e
plot(Annuli~log(Mass), data = turtles)
model1 <- lm(Annuli~log(Mass), data = turtles)
abline(model1)
```



```
summary(model1)
```

```
##
## Call:
## lm(formula = Annuli ~ log(Mass), data = turtles)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.4823  -3.7841  -0.9255   1.7293  22.2686
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -16.6959     3.4416  -4.851 1.96e-06 ***
## log(Mass)     6.0537     0.6036  10.029  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.906 on 305 degrees of freedom
## Multiple R-squared:  0.248,  Adjusted R-squared:  0.2455
## F-statistic: 100.6 on 1 and 305 DF,  p-value: < 2.2e-16
```

```
# Logged both variables with base 10
plot(log10(Annuli)~log10(Mass), data = turtles)
model2 <- lm(log10(Annuli)~log10(Mass), data = turtles)
abline(model2)
```

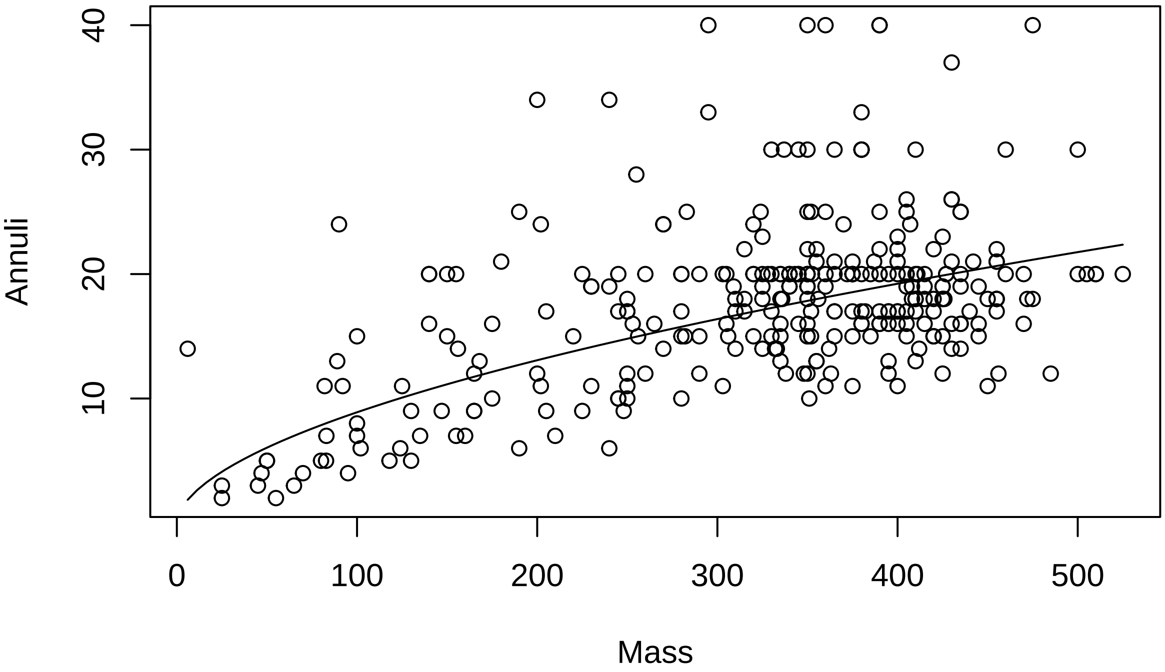


```
summary(model2)
```

```
##
## Call:
## lm(formula = log10(Annuli) ~ log10(Mass), data = turtles)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50378 -0.08509 -0.00308  0.06918  0.87625
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.16273    0.09008  -1.807   0.0718 .
## log10(Mass)  0.55594    0.03638  15.283   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1546 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
```

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.

```
plot(Annuli~Mass, data = turtles)
curve((10^model2$coefficients[1]) * (x^model2$coefficients[2]), add = TRUE)
```



log10(Predicted Annuli) = -0.16273 + 0.55594(log10(Mass)) (Predicted Annuli) = (10^-0.16273) * (Mass^0.55594)

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?

```
# Does this question want the Annuli in the transformed version, or back to normal
(10^-0.16273) * (390^0.55594)
```

```
## [1] 18.95599
```

```
40 - (10^-0.16273) * (390^0.55594)
```

```
## [1] 21.04401
```

We know that the 40th row of the Turtle's dataset has a mass of 390 grams and 40 Annuli. Using our model from our transformations, it predicts an Annuli value of 18.956. The difference from the prediction, or the residual is 21.04401.

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?

```
adult_m_turtles <- subset(turtles, Sex == "Male" & LifeStage == "Adult")
adult_f_turtles <- subset(turtles, Sex == "Female" & LifeStage == "Adult")

adult_m_model <- lm(log10(Annuli)~log10(Mass), adult_m_turtles)
adult_f_model <- lm(log10(Annuli)~log10(Mass), adult_f_turtles)
summary(adult_m_model)
```

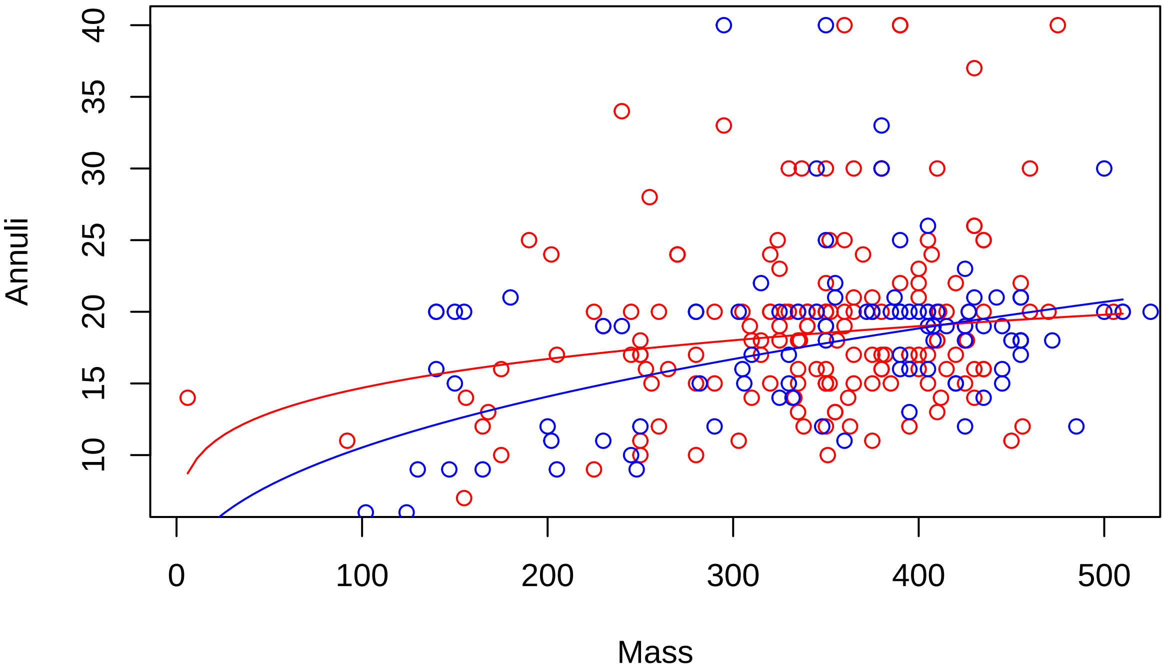
```
##
```

```
## Call:
## lm(formula = log10(Annuli) ~ log10(Mass), data = adult_m_turtles)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35724 -0.07890  0.00365  0.05525  0.33191
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.7965     0.1448   5.500 1.46e-07 ***
## log10(Mass)   0.1853     0.0574   3.228  0.00151 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.129 on 161 degrees of freedom
## Multiple R-squared:  0.06078,    Adjusted R-squared:  0.05495
## F-statistic: 10.42 on 1 and 161 DF,  p-value: 0.001511
```

```
summary(adult_f_model)
```

```
##
## Call:
## lm(formula = log10(Annuli) ~ log10(Mass), data = adult_f_turtles)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28344 -0.07227 -0.00144  0.06766  0.38251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.18310    0.19974   0.917    0.362
## log10(Mass)  0.41965    0.07933   5.290 7.69e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1276 on 96 degrees of freedom
## Multiple R-squared:  0.2257, Adjusted R-squared:  0.2176
## F-statistic: 27.99 on 1 and 96 DF,  p-value: 7.693e-07
```

```
plot(Annuli~Mass, adult_m_turtles, col = "red")
points(Annuli~Mass, adult_f_turtles, col = "blue")
curve(10^adult_m_model$coefficients[1] * x^adult_m_model$coefficients[2], col = "red", add = TRUE)
curve(10^adult_f_model$coefficients[1] * x^adult_f_model$coefficients[2], col = "blue", add = TRUE)
```



This plot tells us that sex does not seem to have an impact on the relationship between Mass and Annuli as the lines intersect at some point. To better understand if there is some impact, we could conduct a t-test for the slopes to see if there is any difference, outside of chance, between the variables.