STOR 455 Homework #1

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

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

```
library(readr)

Turtles <- read_csv("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.
```

```
head(Turtles)
```

```
## # A tibble: 6 x 9
##   LifeStage Sex     Annuli  Mass StraightlineCL MaxCW PL_Anter~1 PL_Hi~2 Shell~3
##   <chr>     <chr>    <dbl> <dbl>          <dbl> <dbl>      <dbl>   <dbl>   <dbl>
## 1 Adult     Male        13   410          127   102         48      68      61
## 2 Adult     Male        19   340          114.   94.0       44.9    67.6    55.9
## 3 Juvenile  Female       7   160           89.5  73.5       39.6    53.6    43.5
## 4 Adult     Male        16   175          128.  101.        54.8    84.7    62.0
## 5 Juvenile  Female       7   100           81    69         35      44      39
## 6 Adult     Unknown     17   410          127.  101.        56.7    81.4    64.2
## # ... with abbreviated variable names 1: PL_AnteriortoHinge,
## #   2: PL_HingetoPosterior, 3: ShellHeightatHinge
```

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.

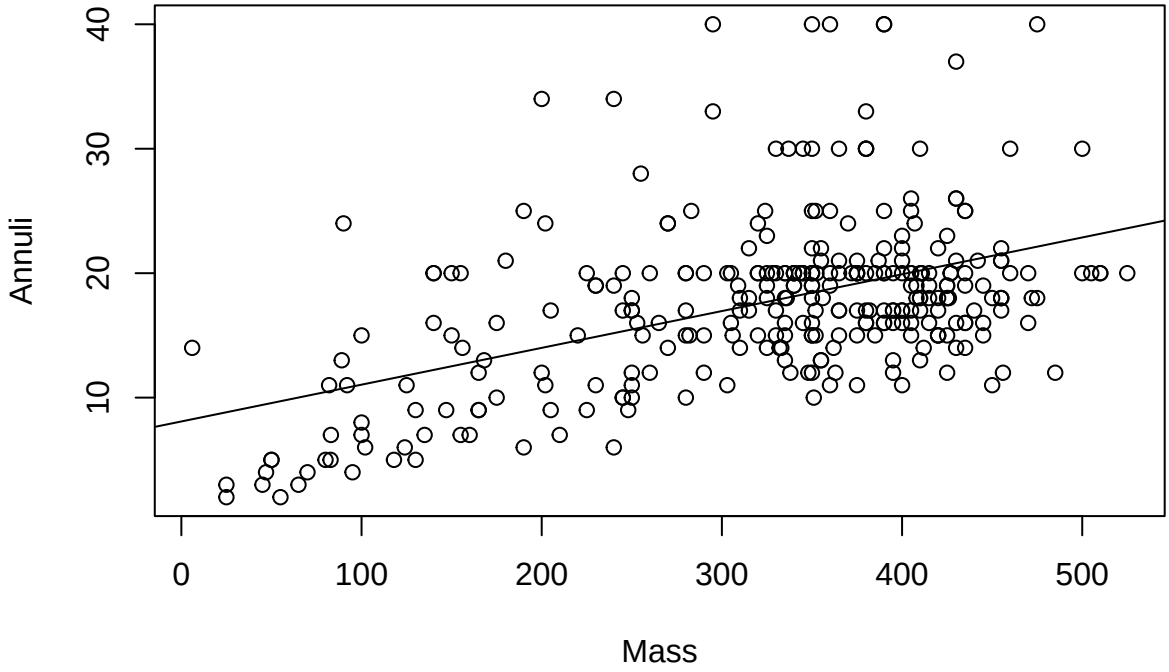
However, the annuli are still counted since it may yield important life history information. Construct a least squares regression line that predicts turtles' *Annuli* by their *Mass*.

```
lsrl = lm(Annuli~Mass, data = Turtles)
lsrl
```

```
##
## Call:
## lm(formula = Annuli ~ Mass, data = Turtles)
##
## Coefficients:
## (Intercept)         Mass
##     8.08494      0.02957
```

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

```
plot(Annuli~Mass, data = Turtles)
abline(lsrl)
```



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?

```
Predicted = Turtles$Mass * lsrl$coefficients[2] + lsrl$coefficients[1]
Residual = Turtles$Annuli - Predicted

Predicted[40]
```

```
## [1] 19.61777
```

```
Residual[40]
```

```
## [1] 20.38223
```

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

```
largest_negative = min(Residual)
largest_negative
```

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

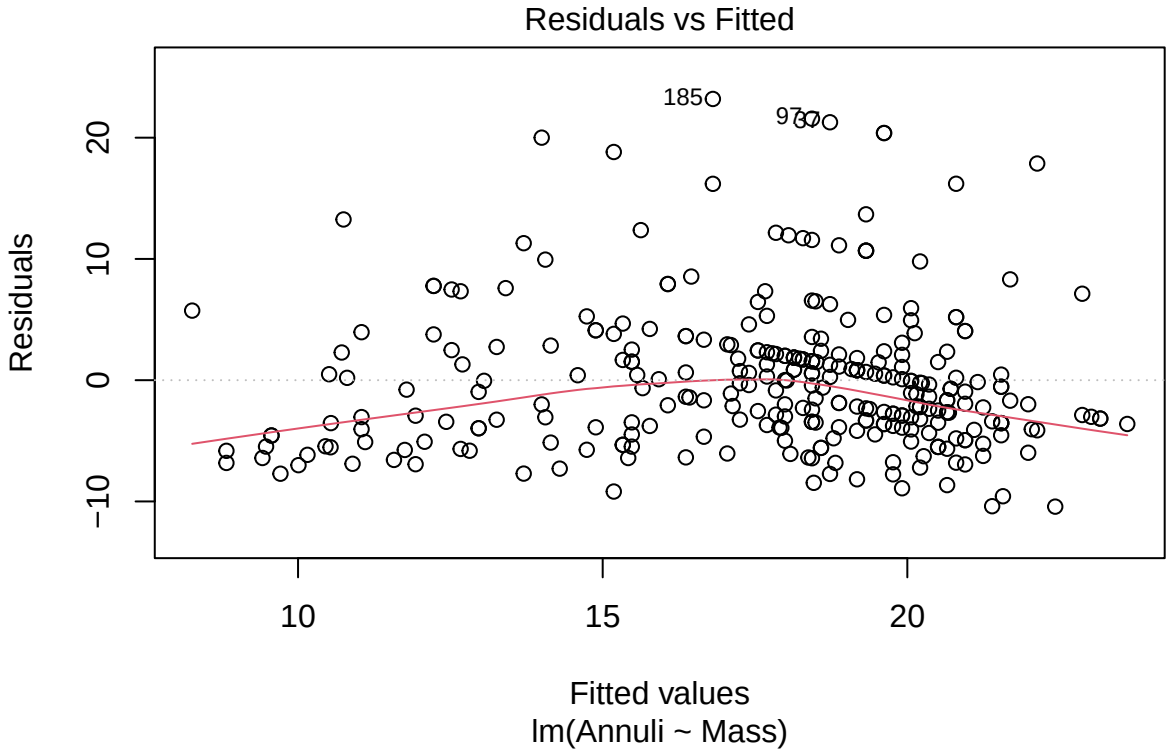
5) Which turtle (by row number in the dataset) has the most negative residual? What is the value of that residual?

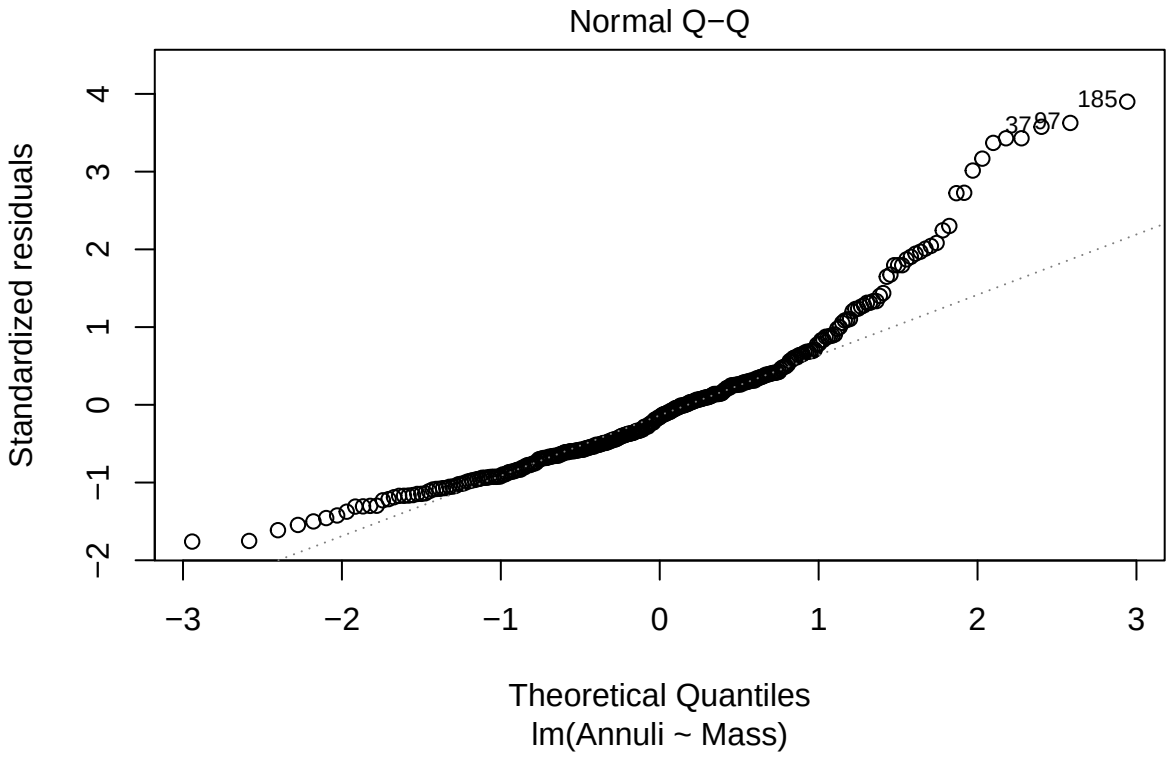
```
largest_positive = max(Residual)
largest_positive
```

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

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(lsrl, 1:2)
```






```
# The red line is around 0 and is relatively flat. The conditions for the linear model are met.
# The QQ plot shows the same thing. There are a lot of points on the line or close to the line.
```

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.

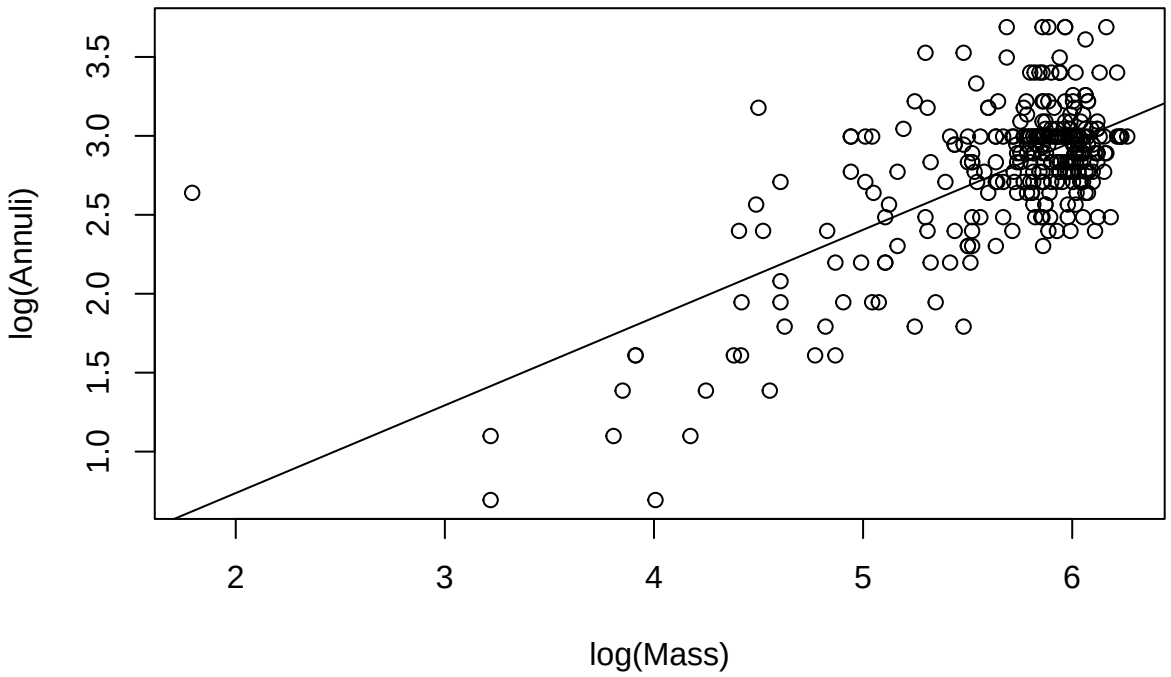
```
log_model = lm(log(Annuli)~log(Mass), data = Turtles)
summary(log_model)
```

```
##
## 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
```

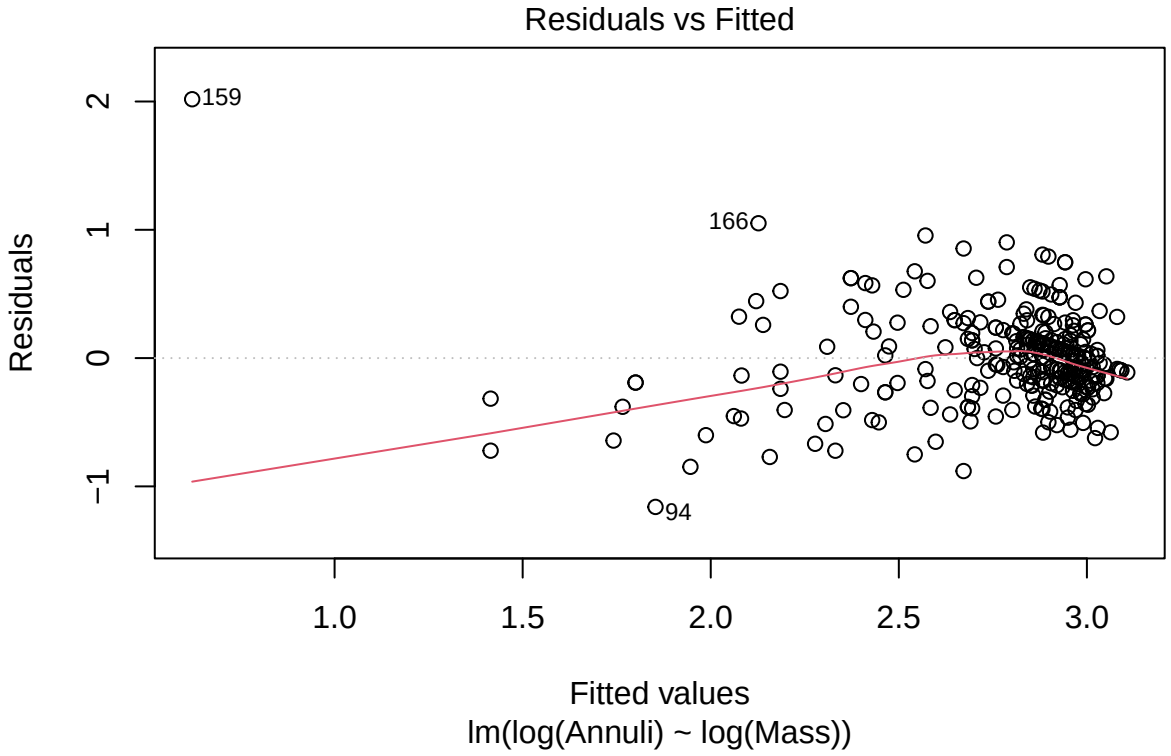
```
exp(-0.37469)
```

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

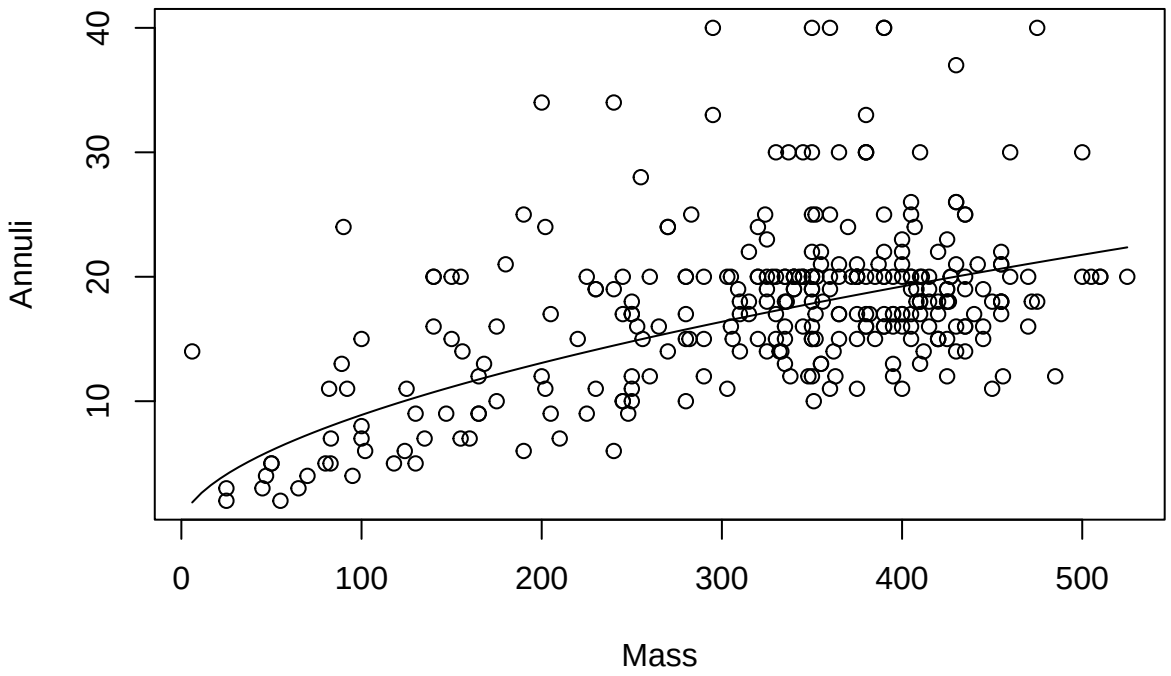
```
plot(log(Annuli)~log(Mass), data = Turtles)
abline(log_model)
```




```
plot(log_model, 1)
```

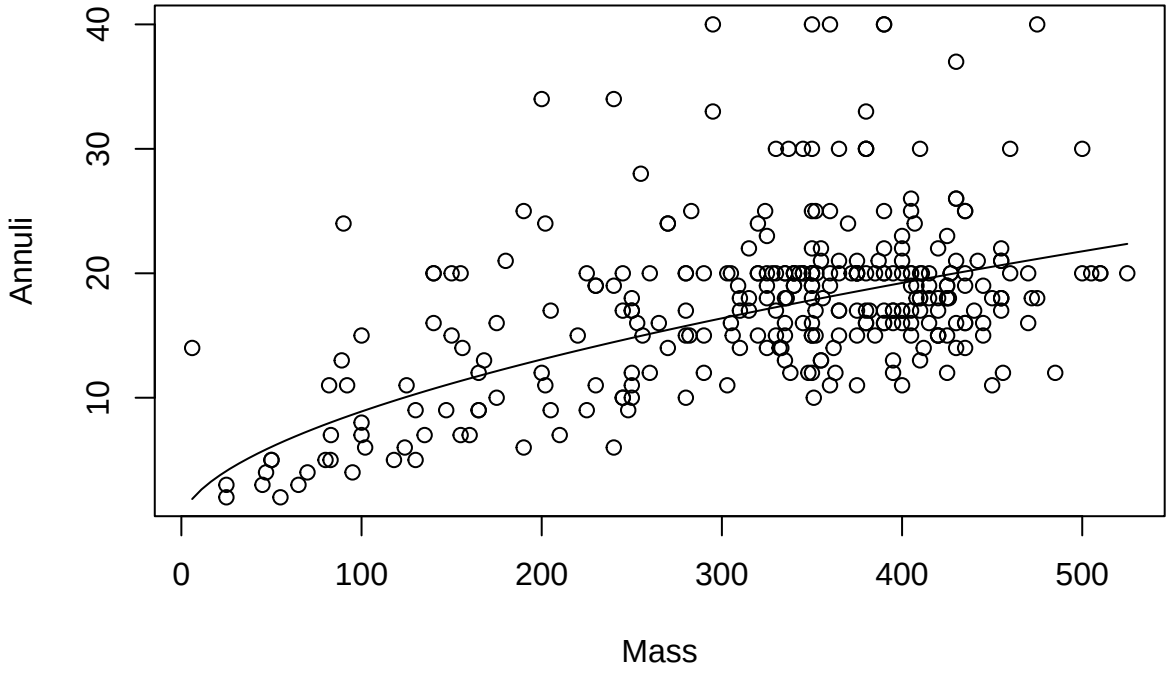


```
plot(Annuli~Mass, data = Turtles)
curve(0.6875024*(x^0.55594), add=TRUE)
```



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(0.6875024*(x^0.55594), add=TRUE)
```



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?

```
log_annuli = 0.6875024*(390^0.55594)
difference = Turtles$Annuli[40] - log_annuli

log_annuli
```

```
## [1] 18.95617
```

```
difference
```

```
## [1] 21.04383
```

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?

```
Male_turtles = Turtles[Turtles$Sex == "Male",]
Female_turtles = Turtles[Turtles$Sex == "Female",]

Male_turtles
```

```
## # A tibble: 170 x 9
##    LifeStage Sex   Annuli  Mass StraightlineCL MaxCW PL_Anteri~1 PL_Hi~2 Shell~3
##    <chr>     <chr>  <dbl> <dbl>          <dbl> <dbl>       <dbl>   <dbl>   <dbl>
##  1 Adult     Male      13   410          127   102          48      68      61
##  2 Adult     Male      19   340          114.   94.0        44.9    67.6    55.9
##  3 Adult     Male      16   175          128.  101.         54.8    84.7    62.0
##  4 Adult     Male      18   325          115    94          45      68      55
##  5 Adult     Male      40   475          137   105          52      79      63
##  6 Adult     Male      15   405          123    99          49      72      61
##  7 Adult     Male      10   175           98.6  71.6        42.3    53.3    46.7
##  8 Adult     Male      28   255          111    85          45      64      53
##  9 Adult     Male      18   336          119    95          47      68      59
## 10 Adult     Male      18   315          122    94          49      70      56
## # ... with 160 more rows, and abbreviated variable names 1: PL_AnteriortoHinge,
## #   2: PL_HingetoPosterior, 3: ShellHeightatHinge
```

```
Female_turtles
```

```
## # A tibble: 106 x 9
##    LifeStage Sex    Annuli  Mass StraightlineCL MaxCW PL_Anter~1 PL_Hi~2 Shell~3
##    <chr>     <chr>   <dbl> <dbl>          <dbl> <dbl>      <dbl>   <dbl>   <dbl>
##  1 Juvenile  Female      7   160           89.5  73.5       39.6    53.6    43.5
##  2 Juvenile  Female      7   100           81    69         35      44      39
##  3 Adult     Female     18   472          131   104         49      80      59
##  4 Adult     Female     20   155          123.   99.4       51.7    74.7    64.6
##  5 Adult     Female     30   345          105    89         40      66      56
##  6 Adult     Female     19   240          125.  102.        48.9    71.9    58.5
##  7 Adult     Female     12   425          120    88         42      99      61
##  8 Adult     Female     20   525          128.  106.        52.4    79.3    63.1
##  9 Adult     Female     11   202          102    78         39      59      49
## 10 Adult     Female     13   395          117    92         46      71      60
## # ... with 96 more rows, and abbreviated variable names 1: PL_AnteriortoHinge,
```

```
## #   2: PL_HingetoPosterior, 3: ShellHeightatHinge
```

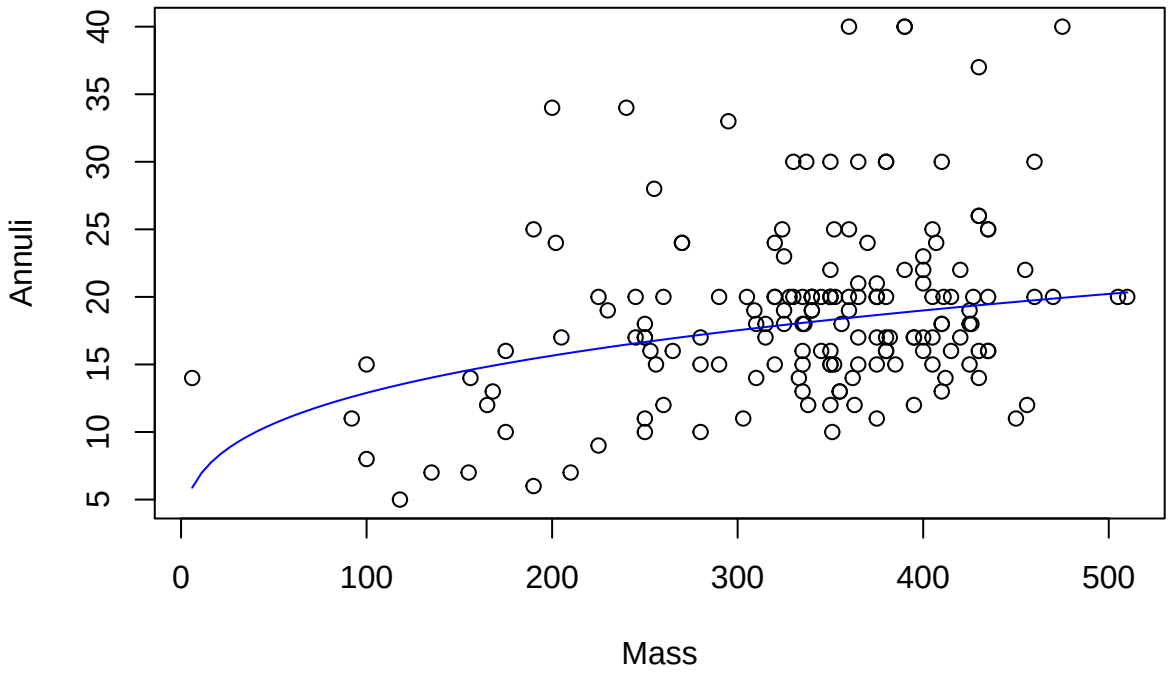
```
male_log_model = lm(log(Annuli)~log(Mass), data = Male_turtles)
summary(male_log_model)
```

```
##
## Call:
## lm(formula = log(Annuli) ~ log(Mass), data = Male_turtles)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.99367 -0.18547  0.01883  0.14543  0.86840
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.26995    0.33782   3.759 0.000235 ***
## log(Mass)    0.27945    0.05848   4.779 3.83e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3308 on 168 degrees of freedom
## Multiple R-squared:  0.1197, Adjusted R-squared:  0.1144
## F-statistic: 22.84 on 1 and 168 DF,  p-value: 3.827e-06
```

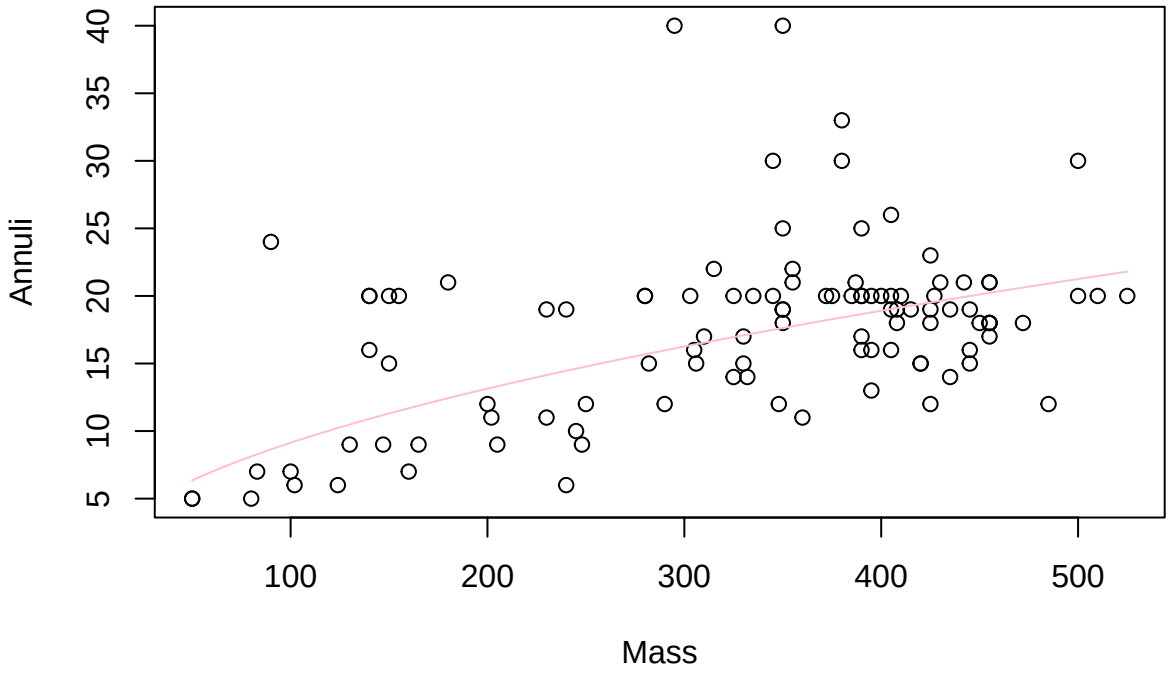
```
female_log_model = lm(log(Annuli)~log(Mass), data = Female_turtles)
summary(female_log_model)
```

```
##
## Call:
## lm(formula = log(Annuli) ~ log(Mass), data = Female_turtles)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8797 -0.1998 -0.0257  0.1605  1.0212
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.20412    0.35579  -0.574    0.567
## log(Mass)    0.52467    0.06227   8.426 2.14e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3252 on 104 degrees of freedom
## Multiple R-squared:  0.4057, Adjusted R-squared:    0.4
## F-statistic:    71 on 1 and 104 DF,  p-value: 2.138e-13
```

```
plot(Annuli~Mass, data = Male_turtles)
curve(exp(1.26995)*(x^0.27945), add=TRUE, col=c("blue", "blue"))
```




```
plot(Annuli~Mass, data = Female_turtles)
curve(exp(-0.20412)*(x^0.52467), add=TRUE, col=c("pink", "pink"))
```




```
# Annuli is more correlated with mass for females compared to males.
# This is proven on the graphs where the females have less variance compared to the males.
```