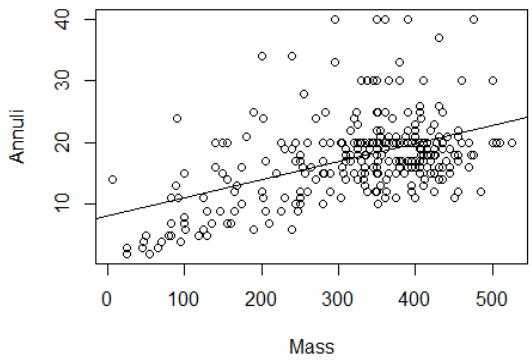
# STOR 455 Homework 1 Turtles R Notebook

*I changed the size of pictures  and color of questions' heading in word for easy viewing. All things are directly exported from RStudio. I did this assignment independently. Signed Yunwei Cao.*

```
library(readr)
library(Stat2Data)

Turtles <- read_csv(
  "https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/Turtles.csv")
```

## Question1&2: a least squares regression line & a scatterplot

```
plot(Annuli~Mass, data=Turtles)
Turtlesmod1 =lm(Annuli~Mass, data=Turtles)
abline(Turtlesmod1)
```




## Question3: the residual for 40th row of the Turtles is 20.38223.

## the model predict for this turtle's number of Annuli is 369.6178.

```
Turtlesmod1$resid[40]

##       40
## 20.38223

390-Turtlesmod1$resid[40]

##       40
## 369.6178
```

Q4: row185 has the largest positive residual. The value is 23.19151.

Q5: row93 has the most negative residual. The value is -10.42705.

```
max(Turtlesmod1$residuals)

## [1] 23.19151

which.max(Turtlesmod1$residuals)

## 185
## 185

Turtles[185,]

## # A tibble: 1 x 9
##   LifeStage Sex    Annuli  Mass StraightlineCL MaxCW PL_AnteriortoHi
nge
##   <chr>     <chr>   <dbl> <dbl>          <dbl> <dbl>              <d
bl>
## 1 Adult     Female     40   295            109    85
 44
## # ... with 2 more variables: PL_HingetoPosterior <dbl>,
## #   ShellHeightatHinge <dbl>

min(Turtlesmod1$residuals)

## [1] -10.42705

which.min(Turtlesmod1$residuals)

## 93
## 93

Turtles[93,]

## # A tibble: 1 x 9
##   LifeStage Sex    Annuli  Mass StraightlineCL MaxCW PL_AnteriortoHi
nge
##   <chr>     <chr>   <dbl> <dbl>          <dbl> <dbl>              <d
bl>
## 1 Adult     Female     12   485           129.  105.               5
2.6
## # ... with 2 more variables: PL_HingetoPosterior <dbl>,
## #   ShellHeightatHinge <dbl>
```

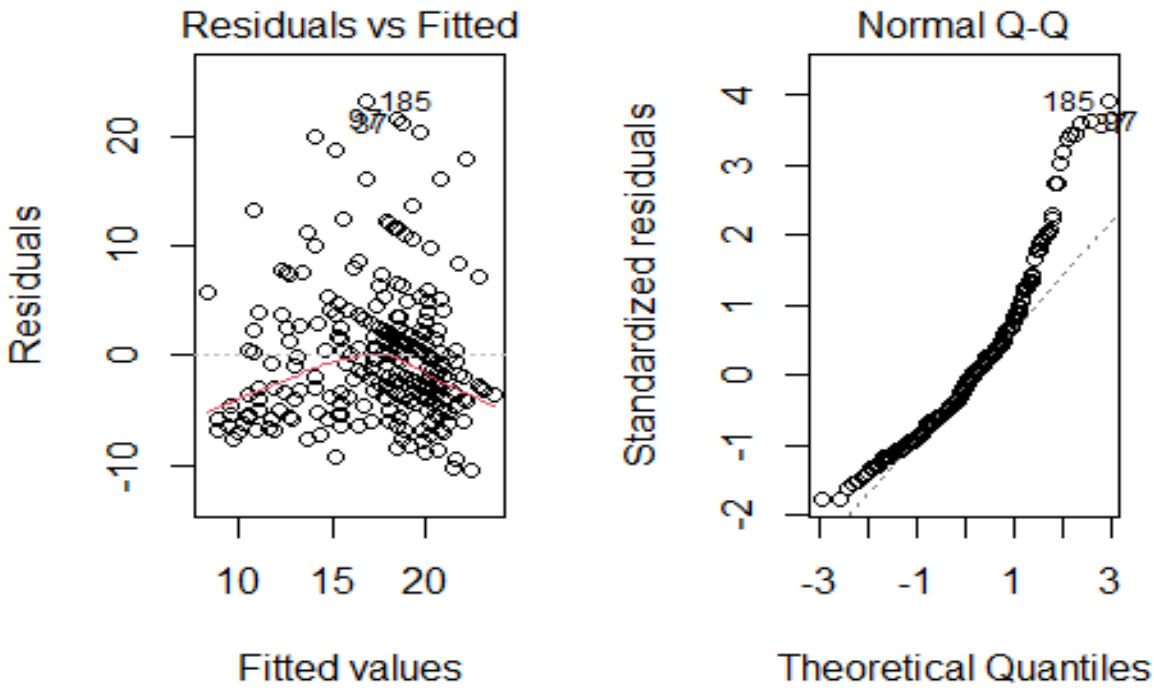
Q6: Simple Linear Model-Conditions

(1)Linearity: Not good. From the residuals vs Fitted plot, we can see that the data does not follow the horizontal line centered at 0, and the red line shows another pattern.

(2)Zero Mean: No problem. This is inherent using the least squares method.
(3)Constant variance(Homoscedasticity): Not good. From the lm picture(Q2), we know that there is a fanning pattern in most of the back place.
(4)Independence:Uncertain. Turtles' annulis may not have internal influence. But we do not know the time and process of data collection.

(5)Normality:not good nor bad. From the NormalQQ, data at both tail fly out, althouth the sample size is not large enough(307 observations).

```
par(mfrow=c(1,2))
plot(Turtlesmod1,1:2)
```



Q7.1: Logarithmic linear transformation

By comparing 4 log-linear plots, I think double-log is the best.

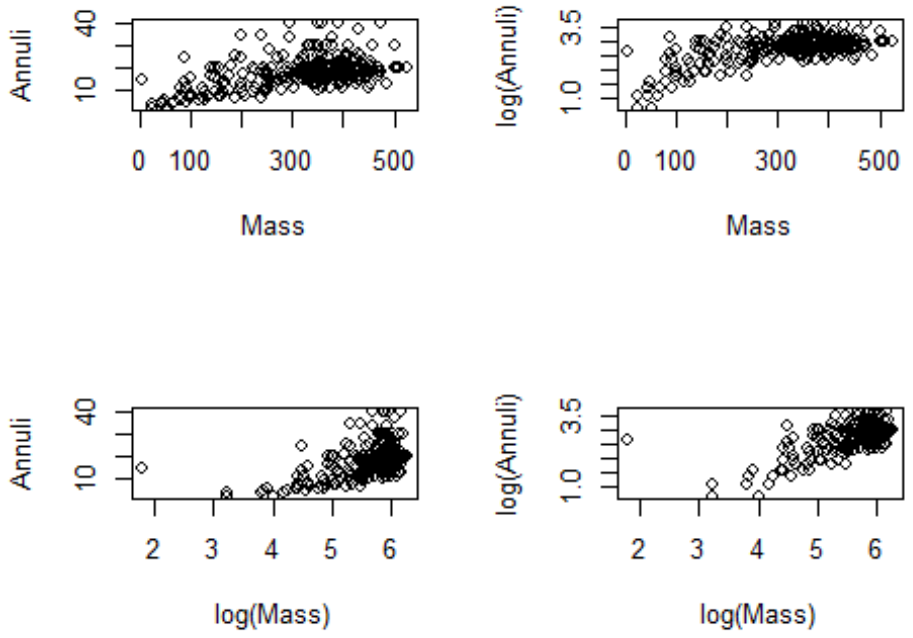
Due to the piled up data is not clear, I made a subset for easy viewing.

Overall the new model is better. Linearity improves(residuals vs Fitted better).

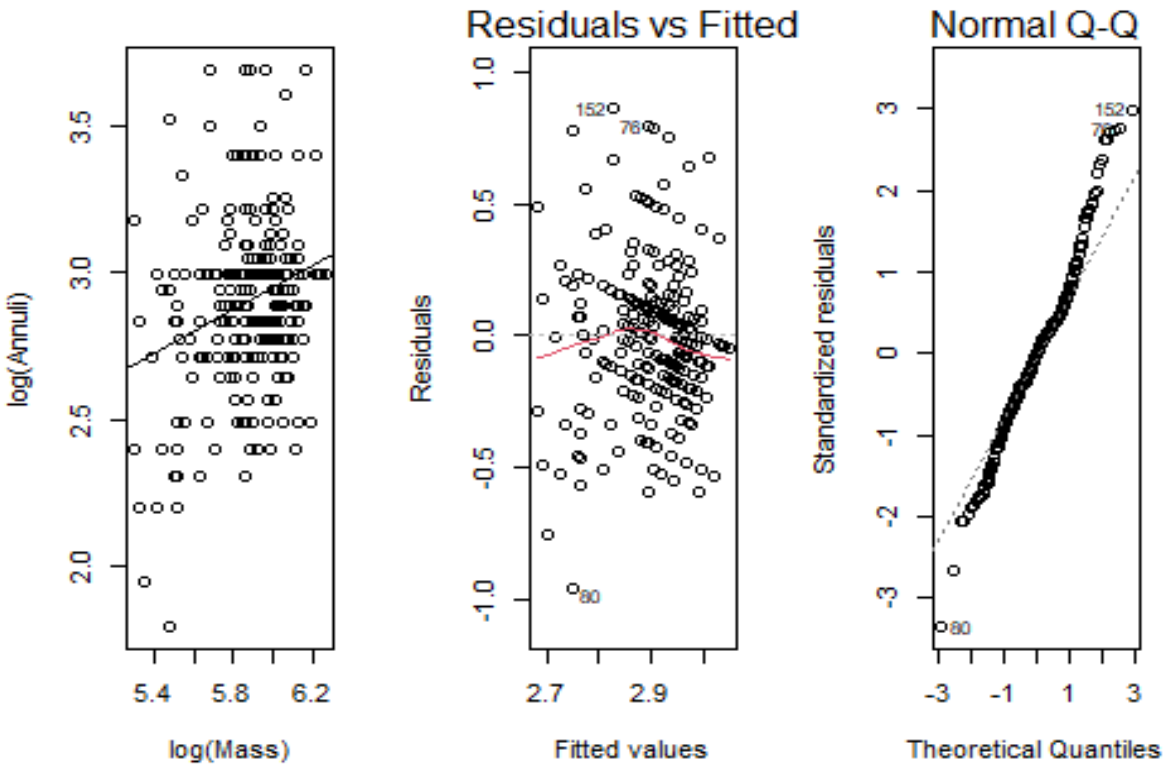
The constant variance improves(no fanning pattern).

Normality improves(NormalQQ better).

```
par(mfrow=c(2,2))

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




```
par(mfrow=c(1,3))

Turtle_above_200mass = subset(Turtles, Mass > 200)
Turtlesmod2 =lm(log(Annuli)~log(Mass), data=Turtle_above_200mass)

plot(log(Annuli)~log(Mass), Turtle_above_200mass)
abline(Turtlesmod2)
plot(Turtlesmod2,1:2)
```




```
summary(Turtlesmod2)

##
## Call:
## lm(formula = log(Annuli) ~ log(Mass), data = Turtle_above_200mass)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95872 -0.16297 -0.00392  0.12972  0.86011
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.67103    0.52349   1.282    0.201
## log(Mass)    0.37942    0.08906   4.260 2.87e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2888 on 256 degrees of freedom
## Multiple R-squared:  0.0662, Adjusted R-squared:  0.06255
## F-statistic: 18.15 on 1 and 256 DF,  p-value: 2.872e-05
```

Q7.2: Power function transformation

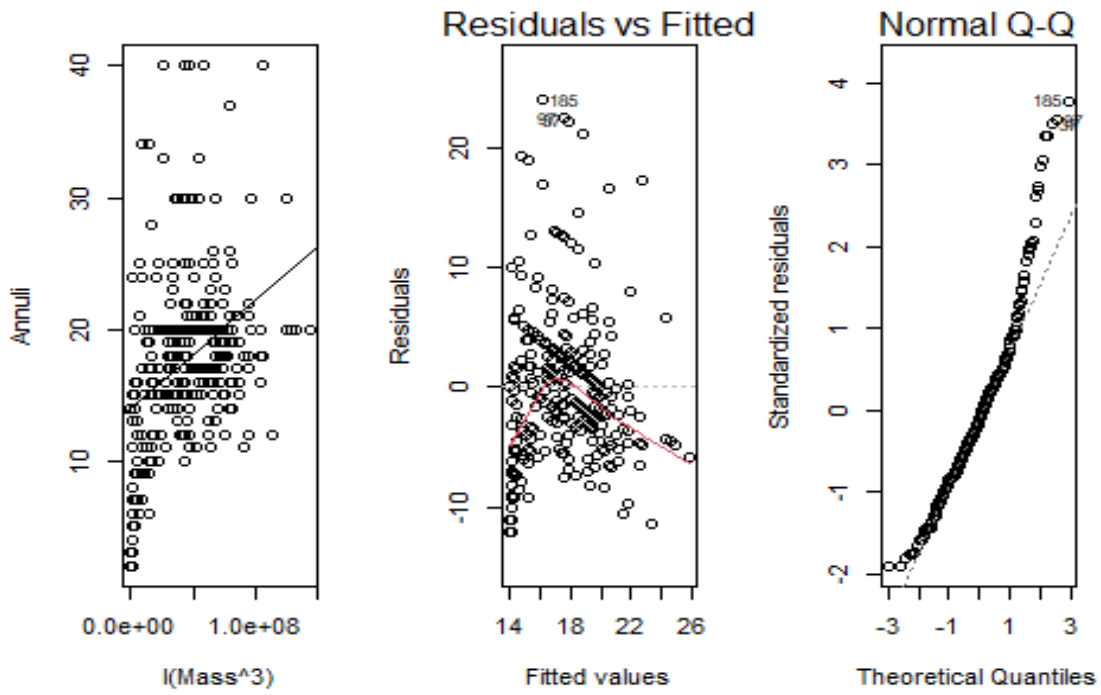
Overall the new model is not good.

Linearity worse(residuals vs Fitted do not improve).

The constant variance does not improve.

Normality not good nor bad still.

```
par(mfrow=c(1,3))

Turtlesmod3=lm(Annuli~I(Mass^3), data=Turtles)

plot(Annuli~I(Mass^3), data=Turtles)
abline(Turtlesmod3)
plot(Turtlesmod3,1:2)
```

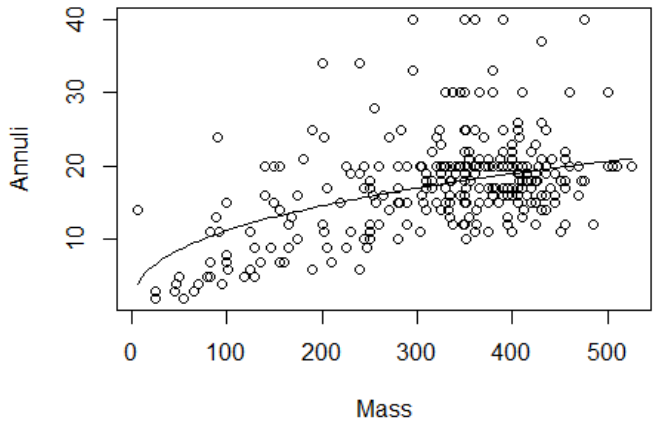



```
summary(Turtlesmod3)

##
## Call:
## lm(formula = Annuli ~ I(Mass^3), data = Turtles)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.0165  -4.2487  -0.7968   2.7762  23.8992
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.400e+01  6.352e-01  22.043  < 2e-16 ***
## I(Mass^3)   8.172e-08  1.170e-08   6.985 1.79e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.323 on 305 degrees of freedom
## Multiple R-squared:  0.1379, Adjusted R-squared:  0.1351
## F-statistic:  48.8 on 1 and 305 DF,  p-value: 1.787e-11
```

Q8: the raw data with the double-log model

```
B0 = summary(Turtlesmod2)$coefficients[1,1]
B1 = summary(Turtlesmod2)$coefficients[2,1]

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




Q9: the residual for 40th row of the Turtles is 0.2101058.

the model predict for this turtle's number of Annuli is 389.7899.

```
Turtlesmod2$resid[40]

##        40
## 0.2101058

390-Turtlesmod2$resid[40]

##       40
## 389.7899
```

Q10: new exploration

the relationship between Mass and Annuli is probably positive correlation.

The increasing speed of female turtles' annulis is faster than the males'.

```
Turtles_adult =subset(Turtles,LifeStage=='Adult')
Turtles_adult_male =subset(Turtles_adult,Sex=='Male')
Turtles_adult_female =subset(Turtles_adult,Sex=='Female')

Turtlesmod_male =lm(log(Annuli)~log(Mass), data=Turtles_adult_male)
Turtlesmod_female =lm(log(Annuli)~log(Mass), data=Turtles_adult_female)

B0 = summary(Turtlesmod_male)$coefficients[1,1]
B1 = summary(Turtlesmod_male)$coefficients[2,1]
B2 = summary(Turtlesmod_female)$coefficients[1,1]
B3 = summary(Turtlesmod_female)$coefficients[2,1]

plot(Annuli~Mass, data=Turtles_adult)
curve(exp(B0)*x^B1, col='blue',add=TRUE)
curve(exp(B2)*x^B3, col='red',add=TRUE)
```