Turtles/Extinction Project

2025-02-15

```r
#QUESTION 1:
library(dplyr)
```

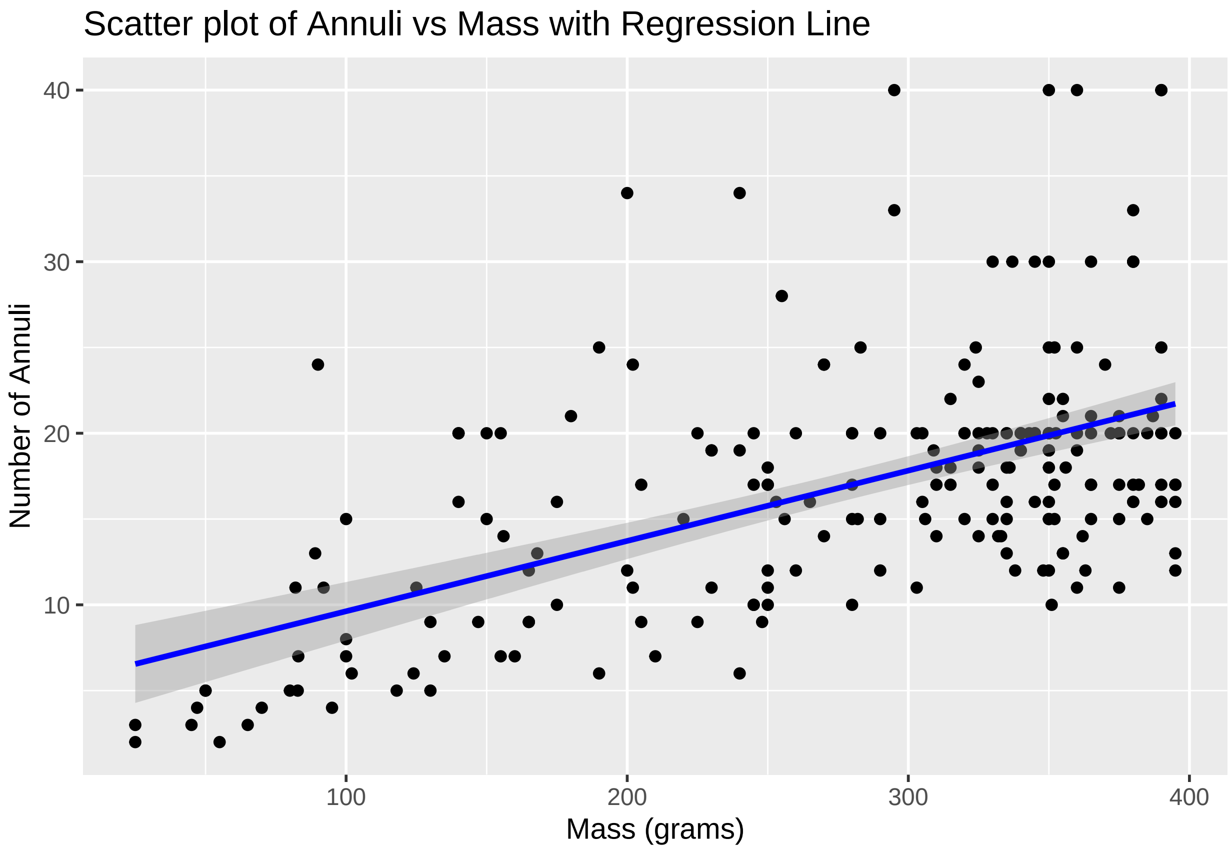
```
##
## Attaching package: ’dplyr’

## The following objects are masked from ’package:stats’:
##
##     filter, lag

## The following objects are masked from ’package:base’:
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)

library(readr)
Turtles <- read.csv("Turtles.csv")

Turtles_under_400g <- Turtles %>%
  filter(Mass < 400 & Mass != 6)

model <- lm(Annuli ~ Mass, data = Turtles_under_400g)
summary(model)
```

```
##
## Call:
## lm(formula = Annuli ~ Mass, data = Turtles_under_400g)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9146 -4.2587 -0.8985  2.1264 22.3811
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.525504   1.249598   4.422 1.55e-05 ***
## Mass        0.040995   0.004206   9.746  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 6.147 on 215 degrees of freedom
## Multiple R-squared:  0.3064, Adjusted R-squared:  0.3032
## F-statistic: 94.98 on 1 and 215 DF,  p-value: < 2.2e-16
```

```
ggplot(Turtles_under_400g, aes(x = Mass, y = Annuli)) +
  geom_point() +
  geom_smooth(method = "lm", col = "blue") +
  labs(title = "Scatter plot of Annuli vs Mass with Regression Line",
       x = "Mass (grams)",
       y = "Number of Annuli")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```
confint(model, level = 0.95)
```

```
##                   2.5 %     97.5 %
## (Intercept) 3.06247307 7.98853454
## Mass        0.03270346 0.04928569
```

```
new_data <- data.frame(Mass = 200)
prediction <- predict(model, newdata = new_data, interval = "prediction", level = 0.90)
prediction
```

```
##        fit      lwr      upr
## 1 13.72442 3.531087 23.91775
```

```
residuals <- Turtles_under_400g %>%
  filter(Mass == 200) %>%
  mutate(Residual = Annuli - predict(model, newdata = .))
residuals
```

```
##   LifeStage    Sex Annuli Mass StraightlineCL MaxCW PL_AnteriortoHinge
## 1  Juvenile   Male     34  200             96    79                 39
## 2     Adult Female     12  200             97    79                 39
##   PL_HingetoPosterior ShellHeightatHinge  Residual
## 1                  59                 45 20.275581
## 2                  55                 47 -1.724419
```

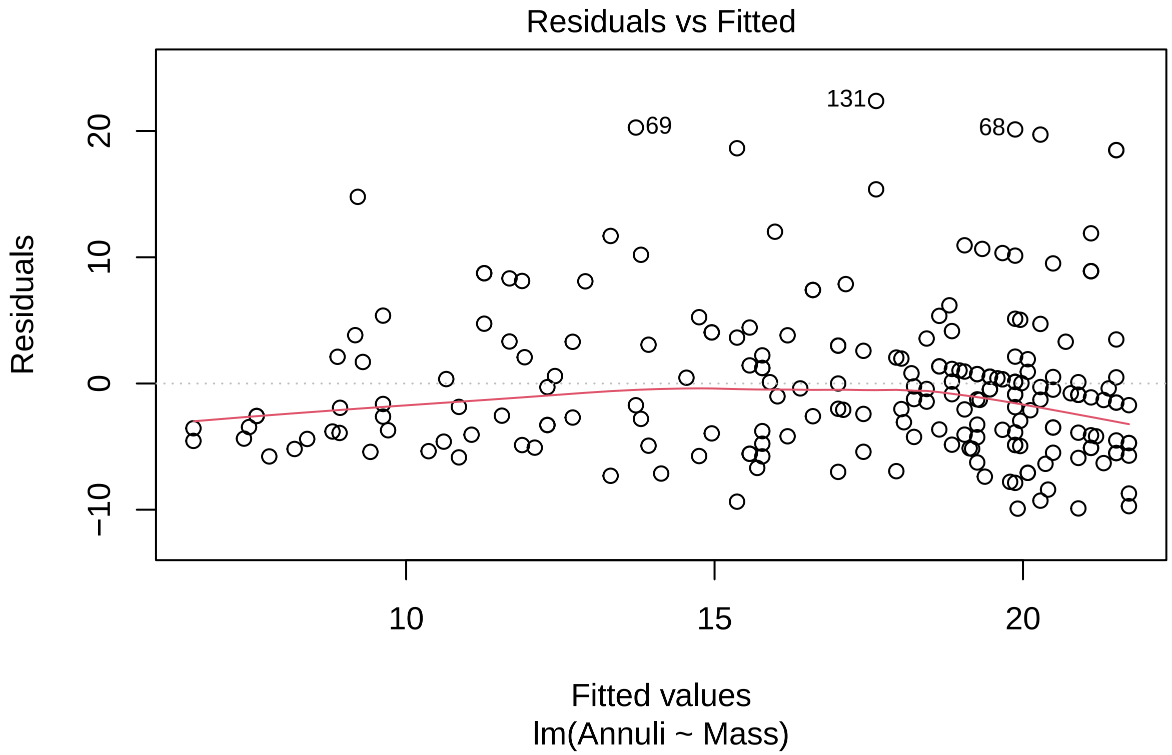
```
largest_positive_residual <- which.max(residuals(model))
Turtles_under_400g[largest_positive_residual, ]
```

```
##     LifeStage    Sex Annuli Mass StraightlineCL MaxCW PL_AnteriortoHinge
## 131     Adult Female     40  295            109    85                 44
##     PL_HingetoPosterior ShellHeightatHinge
## 131                  64                 61
```

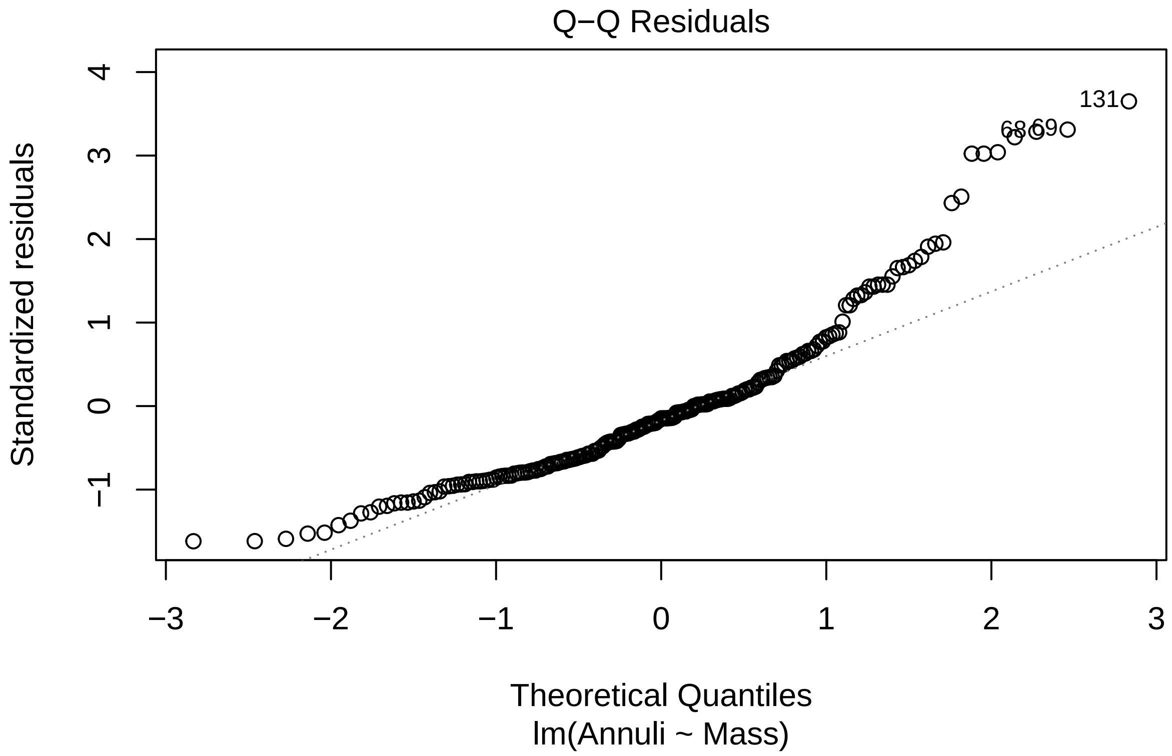
```
most_negative_residual <- which.min(residuals(model))
Turtles_under_400g[most_negative_residual, ]
```

```
##    LifeStage  Sex Annuli Mass StraightlineCL MaxCW PL_AnteriortoHinge
## 64     Adult Male     10  351            123    94                 49
##    PL_HingetoPosterior ShellHeightatHinge
## 64                  67                 57
```

```
plot(model, which = 1)
```



```
plot(model, which = 2)
```



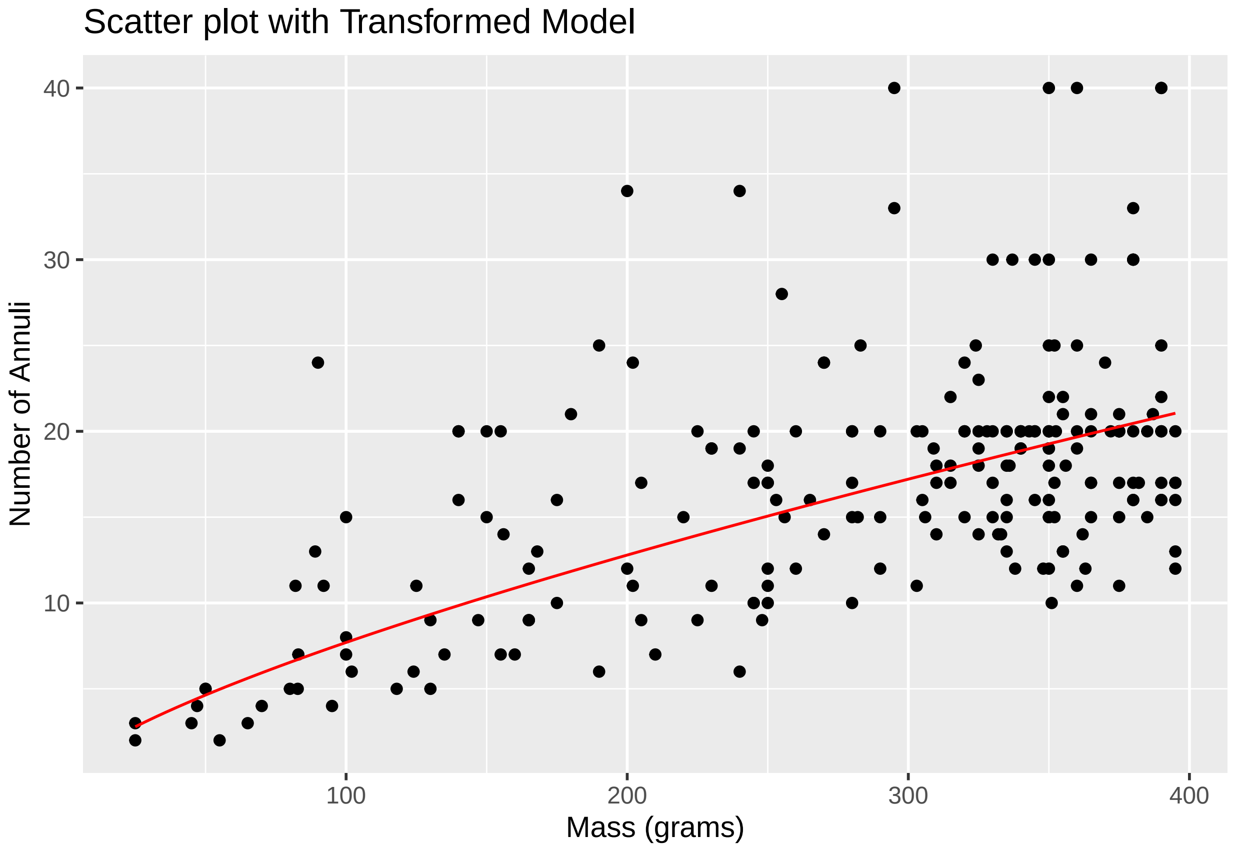
```
model_log <- lm(log(Annuli) ~ log(Mass), data = Turtles_under_400g)
summary(model_log)
```

```
##
## Call:
## lm(formula = log(Annuli) ~ log(Mass), data = Turtles_under_400g)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.91052 -0.25044 -0.01327  0.18116  1.21384
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.33008    0.25233  -5.271 3.29e-07 ***
## log(Mass)    0.73210    0.04542  16.119  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3572 on 215 degrees of freedom
## Multiple R-squared:  0.5472, Adjusted R-squared:  0.5451
## F-statistic: 259.8 on 1 and 215 DF,  p-value: < 2.2e-16
```

```
model_sqrt <- lm(sqrt(Annuli) ~ sqrt(Mass), data = Turtles_under_400g)
summary(model_sqrt)
```

```
##
## Call:
## lm(formula = sqrt(Annuli) ~ sqrt(Mass), data = Turtles_under_400g)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.41614 -0.50245 -0.06504  0.38426  2.19937
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.17940    0.23196   5.085 8.01e-07 ***
## sqrt(Mass)   0.17339    0.01386  12.509  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7105 on 215 degrees of freedom
## Multiple R-squared:  0.4212, Adjusted R-squared:  0.4185
## F-statistic: 156.5 on 1 and 215 DF,  p-value: < 2.2e-16
```

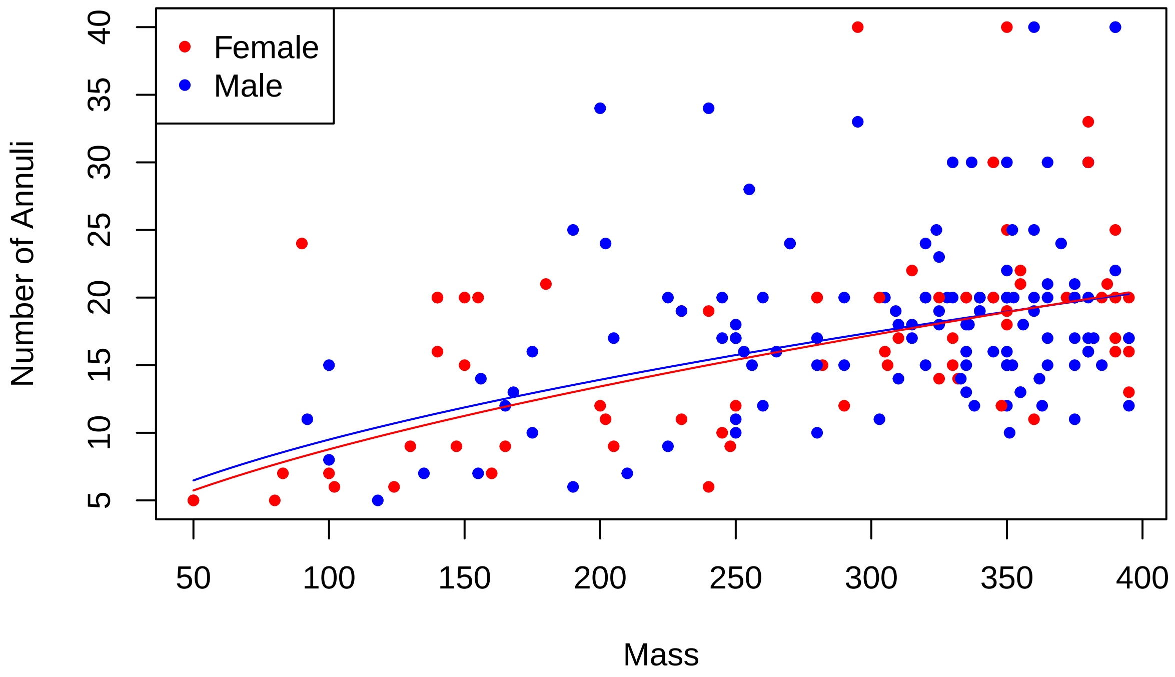
```
ggplot(Turtles_under_400g, aes(x = Mass, y = Annuli)) +
  geom_point() +
  stat_function(fun = function(x) exp(predict(model_log, newdata = data.frame(Mass = x))), col = "red")
  labs(title = "Scatter plot with Transformed Model",
       x = "Mass (grams)",
       y = "Number of Annuli")
```



```r
prediction_log <- exp(predict(model_log, newdata = new_data, interval = "prediction", level = 0.90))
prediction_log
```

```
##        fit      lwr      upr
## 1 12.79171 7.079152 23.11403
```

```r
Turtles_under_400g_sex <- Turtles_under_400g %>%
  filter(Sex != "Unknown")

male_turtles <- Turtles_under_400g_sex %>%
  filter(Sex == "Male")
female_turtles <- Turtles_under_400g_sex %>%
  filter(Sex == "Female")

model_male <- lm(log(Annuli) ~ log(Mass), data = male_turtles)
model_female <- lm(log(Annuli) ~ log(Mass), data = female_turtles)

plot(Turtles_under_400g_sex$Mass, Turtles_under_400g_sex$Annuli,
     xlab = "Mass", ylab = "Number of Annuli",
     col = c("red", "blue")[as.factor(Turtles_under_400g_sex$Sex)], pch = 20)
legend("topleft", legend = c("Female", "Male"), col = c("red", "blue"), pch = c(20, 20))

curve(exp(predict(model_male, newdata = data.frame(Mass = x))), add = TRUE, col = "blue")
curve(exp(predict(model_female, newdata = data.frame(Mass = x))), add = TRUE, col = "red")
```



```r
summary(model_male)
```

```
##
## Call:
## lm(formula = log(Annuli) ~ log(Mass), data = male_turtles)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.81331 -0.21519  0.01527  0.15719  0.89297
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.29202    0.55889  -0.522    0.602
## log(Mass)    0.55214    0.09812   5.627 1.17e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.3332 on 123 degrees of freedom
## Multiple R-squared:  0.2047, Adjusted R-squared:  0.1983
## F-statistic: 31.66 on 1 and 123 DF,  p-value: 1.175e-07
```

```r
summary(model_female)
```

```
##
## Call:
## lm(formula = log(Annuli) ~ log(Mass), data = female_turtles)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.91706 -0.23989 -0.02565  0.20728  1.07085
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.65286    0.47535  -1.373    0.174
## log(Mass)    0.61337    0.08624   7.113 8.88e-10 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.3718 on 68 degrees of freedom
## Multiple R-squared:  0.4266, Adjusted R-squared:  0.4182
## F-statistic: 50.59 on 1 and 68 DF,  p-value: 8.883e-10
```

```r
#COMMENTS:

#c) The slope parameter in the regression model (Annuli ~ Mass) is 0.040995.
#This means that for every 1-gram increase in the mass of a turtle, the number
#of annuli is expected to increase by 0.040995, on average.
#The slope is statistically significant, indicating a strong
#positive relationship between mass and annuli.

#g) The condiitons for a simple linear model are: linearity, independence,
#homoscedasticity, and normality. This means that the relationship between
```

```
#Mass and Annuli should be linear, the residuals should be independent,
#the residuals should have a constant variance, and the residuals should be
#normally distributed. Firstly, the normal Q-Q plot checks if the residuals
#follow a normal distribution.
#The points should lie close to the reference line.
#Here, the Q-Q plot shows some deviation at the tails, indicating that the
#residuals are not perfectly normally distributed.
#Secondly, the residuals vs.fitted plot shows whether the residuals are randomly
#scattered around zero. In this case, the plot suggests some non-linearity and
#potential heteroscedasticity, as the residuals are not perfectly random.

#h) The log transformation results in improved linearity.
#The relationship between log(Mass) and log(Annuli) appears more linear.
#There is also improved homoscedasticity because
#the residuals are more evenly spread.
#Lastly, there is improved normality becasue the Q-Q plot for the
#log-transformed model shows residuals closer to our refrence line.
#However, the square root transformation only resulted in slightly improved
#linearity and homoscedasticity, so it was not as effective as the
#log transformatio. All in all, the log transformation significantly improves
#the model conditions, making it a better fit for the data.

#j) Prediction for Turtles with a Mass of 200 Grams: In the original model,
#the predicted Annuli was 13.72 and the 90% prediction
#interval was (3.53, 23.92). In the log-transformed model, the predicted annuli
#12.79 and the 90% prediction interval was (7.08, 23.11).
#For turtles with a mass of 200 grams, the observed annuli values can be
#compared to the predicted values. The difference between the observed and
#predicted values is the residual. The log-transformed model provides a slightly
#lower prediction than the original model. The residual is the actual difference
#between the observed and predicted values for a specific turtle.
#The prediction interval gives a range of plausible values for future
#predictions, while the residual is specific to the observed data.

#k) Relationship between Mass and Annuli: The male turtles had a slope of
#0.55214 and an R^2 value of 0.2047. The female turtles had a slope of 0.61337
#and an R^2 value of 0.4266. The relationship between mass and annuli does
#differ by sex. Female turtles show a stronger relationship
#(higher slope and R^2) compared to male turtles.

#Comparing goodness-of-fit: The model fits better for females, as indicated by
#the higher R^2 value (0.4266) and a more significant slope. The model fits less
#well for males, with a lower R^2 value (0.2047). The scatterplot with separate
#curves for males and females shows that the relationship between mass
#and annuli is stronger for females.

#QUESTION 2:

#Part A:
set.seed(123)  # For reproducibility
alpha <- 0.05  # Significance level
N <- 100000    # Number of simulations
n_values <- c(20, 50, 100, 200)  # Sample sizes
```

```
beta0 <- 2      # True intercept
beta1 <- 5      # True slope
sigma <- 5      # Standard deviation of errors

coverage_probability <- function(n) {
  coverages <- numeric(N)

  for (i in 1:N) {
    x <- rnorm(n, mean = 0, sd = 1)
    epsilon <- rnorm(n, mean = 0, sd = sigma)
    y <- beta0 + beta1 * x + epsilon

    model <- lm(y ~ x)

    ci <- confint(model, level = 1 - alpha)["x", ]

    coverages[i] <- (ci[1] <= beta1) & (beta1 <= ci[2])
  }

  mean(coverages)
}

coverage_probs_a <- sapply(n_values, coverage_probability)
names(coverage_probs_a) <- n_values

coverage_probs_a
```

```
##      20      50     100     200
## 0.94886 0.95006 0.95041 0.95013
```

```
#Part B:
coverage_probability_hetero <- function(n) {
  coverages <- numeric(N)

  for (i in 1:N) {
    x <- rnorm(n, mean = 0, sd = 1)
    sigma_i <- 5 * sqrt(abs(x))
    epsilon <- rnorm(n, mean = 0, sd = sigma_i)
    y <- beta0 + beta1 * x + epsilon

    model <- lm(y ~ x)

    ci <- confint(model, level = 1 - alpha)["x", ]

    coverages[i] <- (ci[1] <= beta1) & (beta1 <= ci[2])
  }

  mean(coverages)
}

coverage_probs_b <- sapply(n_values, coverage_probability_hetero)
names(coverage_probs_b) <- n_values
```

```
coverage_probs_b
```

```
##      20      50     100     200 
## 0.84449 0.83631 0.83550 0.83634
```

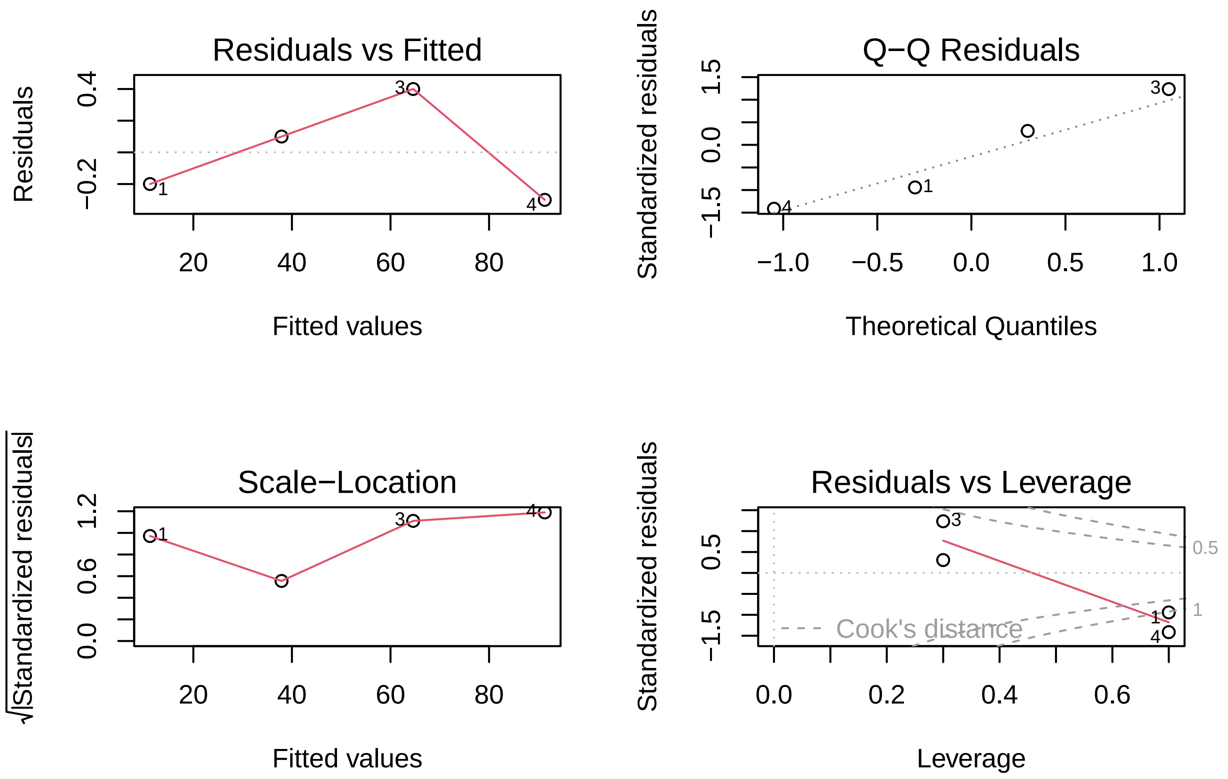
```
results <- data.frame(
  Sample_Size = n_values,
  Homoscedastic_Coverage = coverage_probs_a,
  Heteroscedastic_Coverage = coverage_probs_b
)
results
```

```
##     Sample_Size Homoscedastic_Coverage Heteroscedastic_Coverage
## 20           20                0.94886                  0.84449
## 50           50                0.95006                  0.83631
## 100         100                0.95041                  0.83550
## 200         200                0.95013                  0.83634
```

```
#Part A Comments: The coverage probabilities are 0.94886 for n=20, 0.95006 for
#n=50, 0.95041 for n=100, 0.95013 for n=200. The coverage probabilities for the
#homoscedastic model are close to the theoretical value of 0.95 for all
#sample sizes (n= 20, 50, 100, 200). This confirms that the confidence intervals
#are performing as expected under the homoscedastic assumption.

#Part B Comments: Compared to the coverage probabilities in part a, the coverage
#probabilities are below 0.95, indicating that the confidence intervals are too
#narrow when the homoscedasticity assumption is not true.
#As the sample size increases, the coverage probability
#improves but still does not reach 0.95.

#QUESTION 3:

#Part A:
extinction_data <- data.frame(
  Episode = c(4, 3, 2, 1),
  Millions_of_Years = c(11, 38, 65, 91)
)

model <- lm(Millions_of_Years ~ Episode, data = extinction_data)
summary(model)
```

```
## 
## Call:
## lm(formula = Millions_of_Years ~ Episode, data = extinction_data)
## 
## Residuals:
##    1    2    3    4 
## -0.2  0.1  0.4 -0.3 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 118.0000     0.4743   248.8 1.62e-05 ***
## Episode     -26.7000     0.1732  -154.2 4.21e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.3873 on 2 degrees of freedom
## Multiple R-squared:  0.9999, Adjusted R-squared:  0.9999
## F-statistic: 2.376e+04 on 1 and 2 DF,  p-value: 4.208e-05
```

```r
par(mfrow = c(2, 2))
plot(model)
```



```r
#Part B:
next_episode <- data.frame(Episode = 5)
prediction <- predict(model, newdata = next_episode, interval = "prediction", level = 0.95)
prediction
```

```
##      fit       lwr       upr
## 1 -15.5 -18.13483 -12.86517
```

```r
#Part C:
future_time <- 12
is_concerned <- future_time >= prediction[2] & future_time <= prediction[3]
is_concerned
```

```
## [1] FALSE
```

```
#Comments for Part A: The sample size is very small (n = 4),
#which limits the reliability of the model.
#However, the data shows a strong linear trend,
#and the diagnostics do not reveal significant failure of the assumptions.

#Comments for Part C: The predicted time for the next extinction episode
#(Episode = 5) is -15.5 million years from now, with a 95% prediction interval
#of (-18.13, -12.87) million years. This negative value indicates that the model
#predicts the next extinction episode would have
#occurred 15.5 million years ago, not in the future. The code in part c checks
#whether 12 million years in the future falls within the 95% prediction interval
#for the next extinction episode, which returns as false.
#This means that 12 million years in the future does not fall within the
#predicted interval for the next extinction episode.
```