hw4

Student 1 and Student 2

2024-11-21

0. Contribution Statement

Student 1

Student 1 mainly worked on questions...

Student 2

Student 2 mainly worked on questions ...

Introduction

Data

Objective

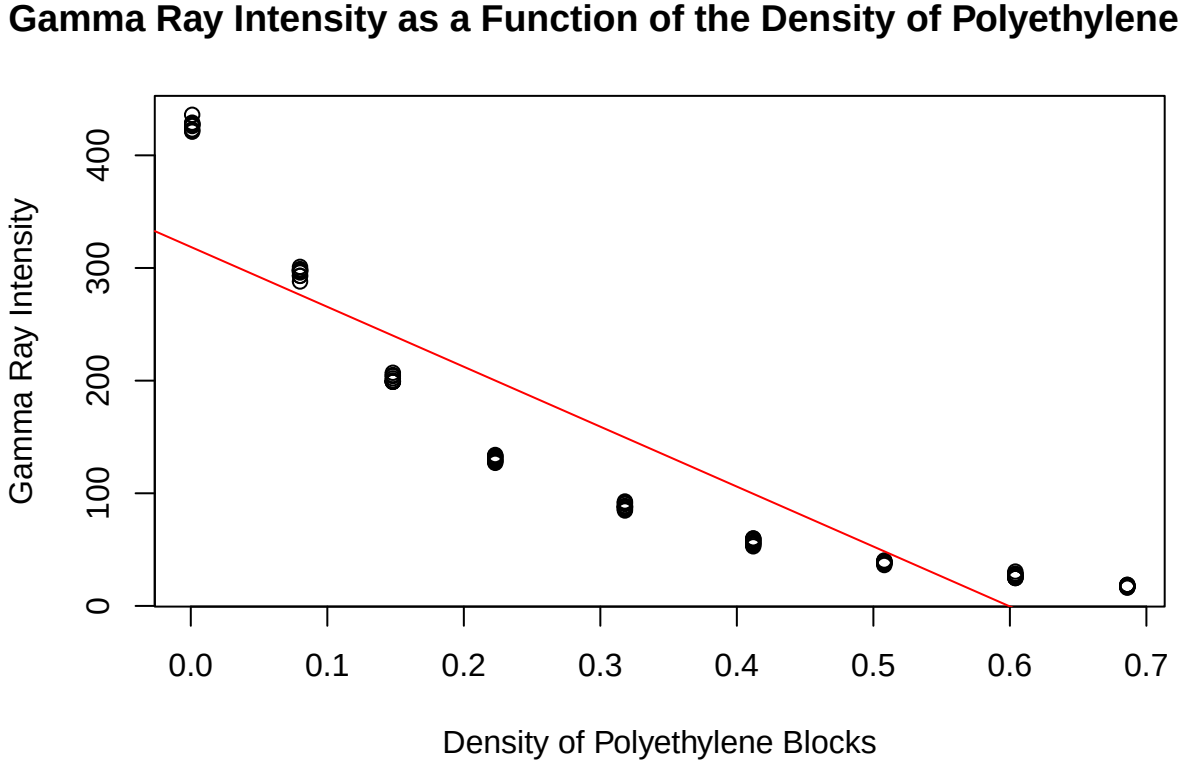
Basic Analysis

Question 1: RAW DATA

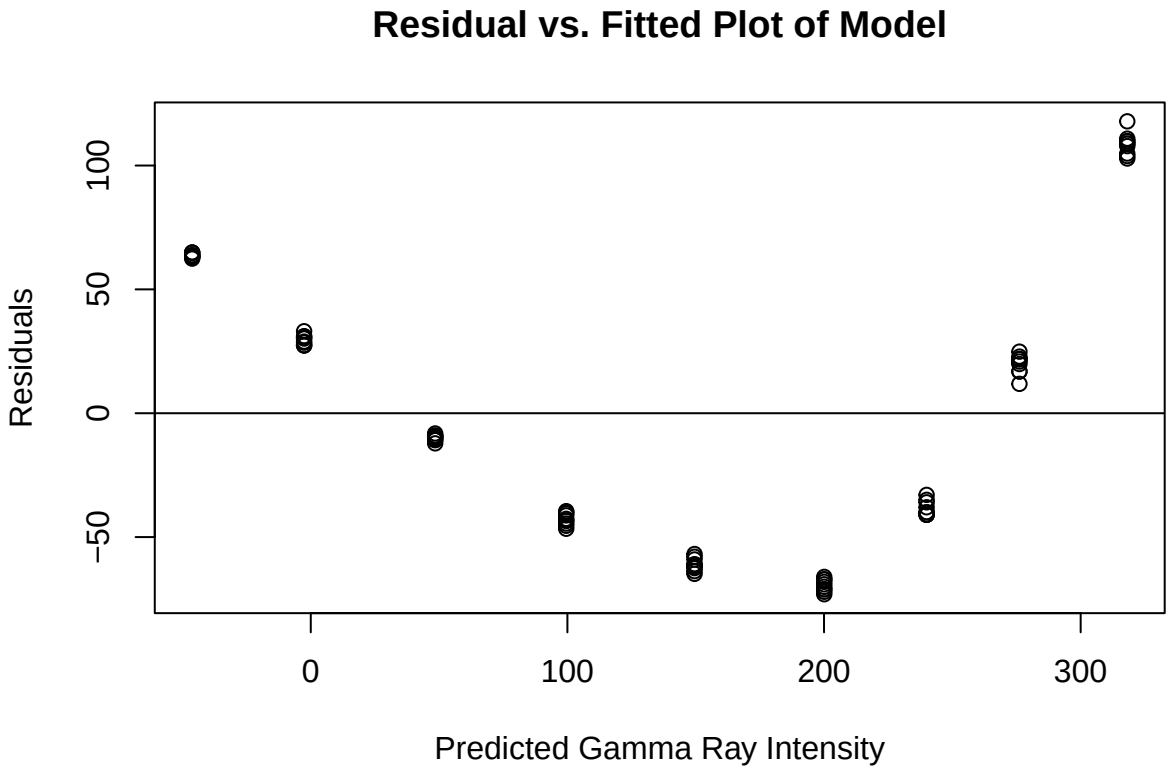
Methods

We will plot the data points and fit the regression line.

```
##
## Call:
## lm(formula = gain ~ density, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -73.08 -44.29  -9.72  30.82 117.83
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   318.70      10.79   29.54   <2e-16 ***
## density      -531.95      26.95  -19.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 57.54 on 88 degrees of freedom
## Multiple R-squared:  0.8157, Adjusted R-squared:  0.8136
## F-statistic: 389.5 on 1 and 88 DF,  p-value: < 2.2e-16
```

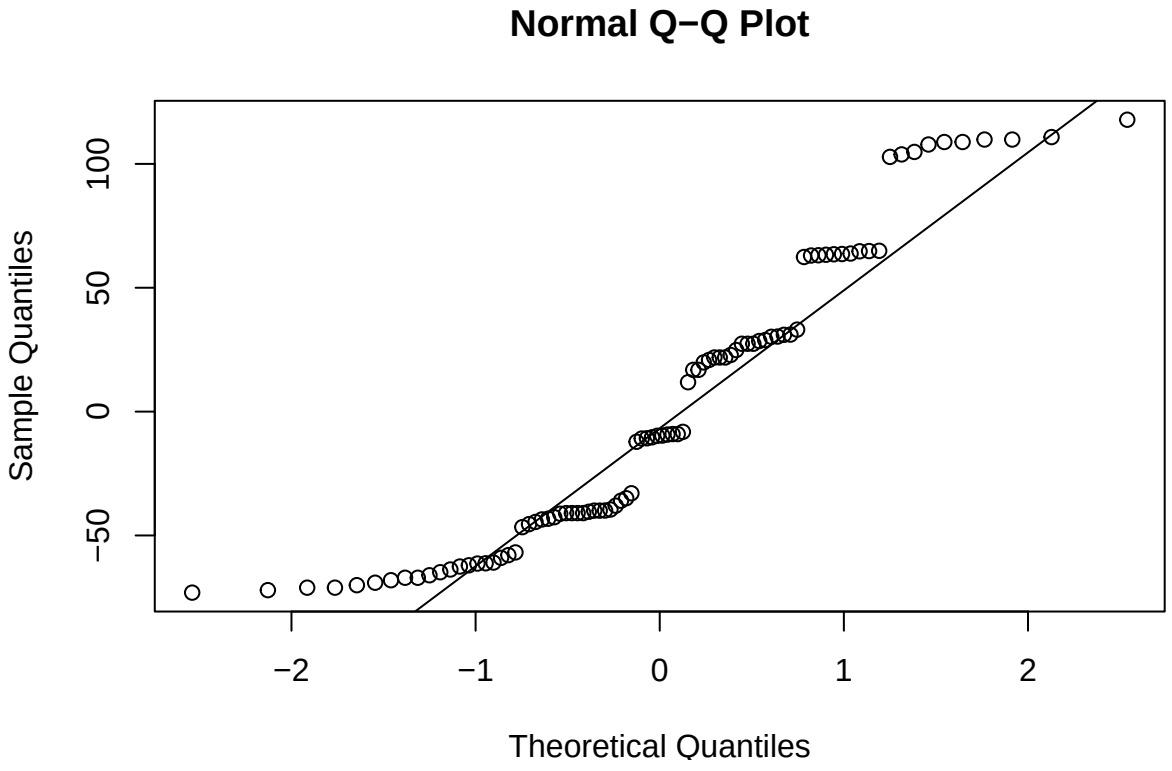


We will extract residuals from the model and observe the residual vs. fitted plot (the predicted values) to understand whether our model has a good fit.

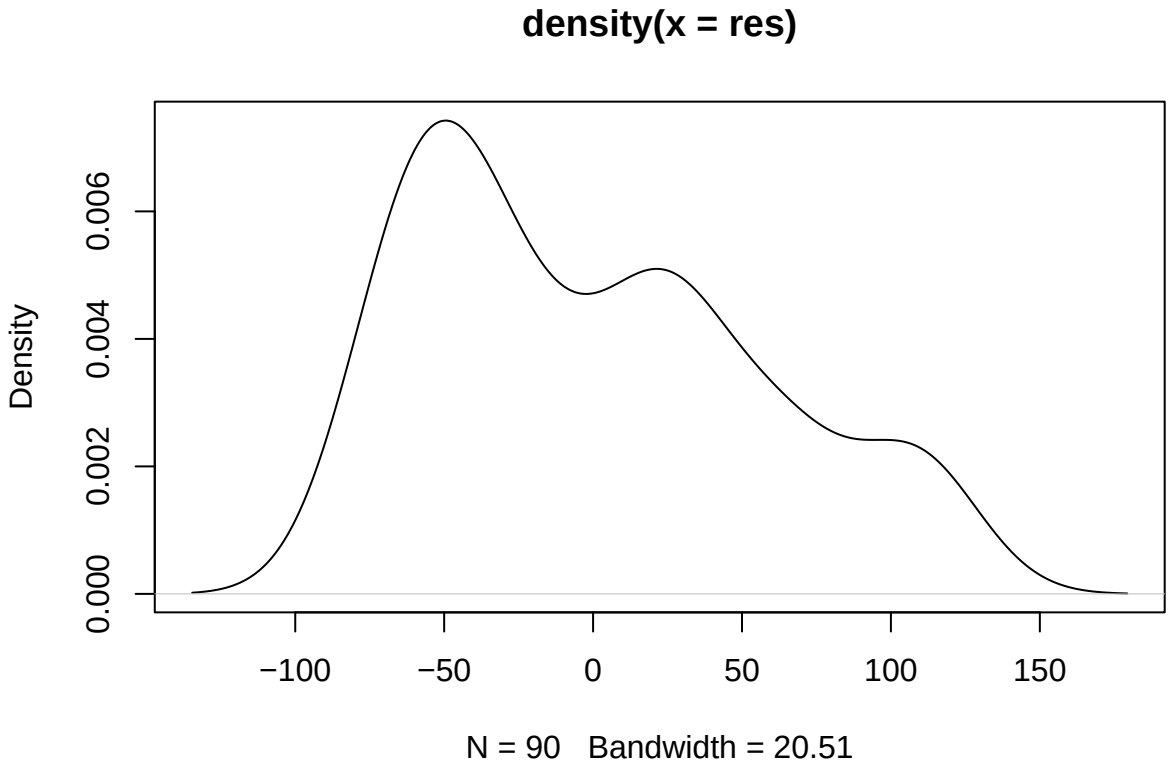


There is a very clear nonlinear pattern between the predicted values from the model and the residuals, which suggests the current model is inappropriate for the data since the relationship between the response variable and the residuals are not linear. There is also homoscedasticity present since the residuals do not vary constantly; some are further or closer away than others, as shown in the graph.

Additionally, we can plot a QQ plot to check if the residuals are normally distributed.



In general, the points on the plot do not fall closely to the line. The points form an 'S-shape', are staggered, and clearly deviate from the 45-degree reference line, which indicates the residuals not normally distributed.



A density plot of our residuals shows that the density of our residuals is skewed right, meaning most of our observed results were below the predicted value (overestimate). We can confirm this by checking our regression line plotted with the observed values. Since the data was concaved up, our regression line was higher than most of the points in the center of the plot.

Analysis

We will need to transform our data if our data is skewed and does not resemble a bell curve. We can test whether our data comes from a normally distributed population with a Shapiro-Wilk test. Our null hypothesis is that our data does come from a normally distributed population.

```
set.seed(123)
shapiro_result <- shapiro.test(data$gain)
print(shapiro_result)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$gain
## W = 0.82115, p-value = 4.599e-09
```

With a p-value of 4.599 * 10^(-9) < 0.05, we reject the null hypothesis. This suggests our data is *not* normally distributed.

Conclusion

A transformation may be necessary because a visual graph shows our fitted model *overestimates* many of our data points since our data is nonlinear and concaved up. A plot of the residuals with the predicted gamma ray intensity also shows that our plot is not homoscedastic, and a QQ plot highlights that our residuals are

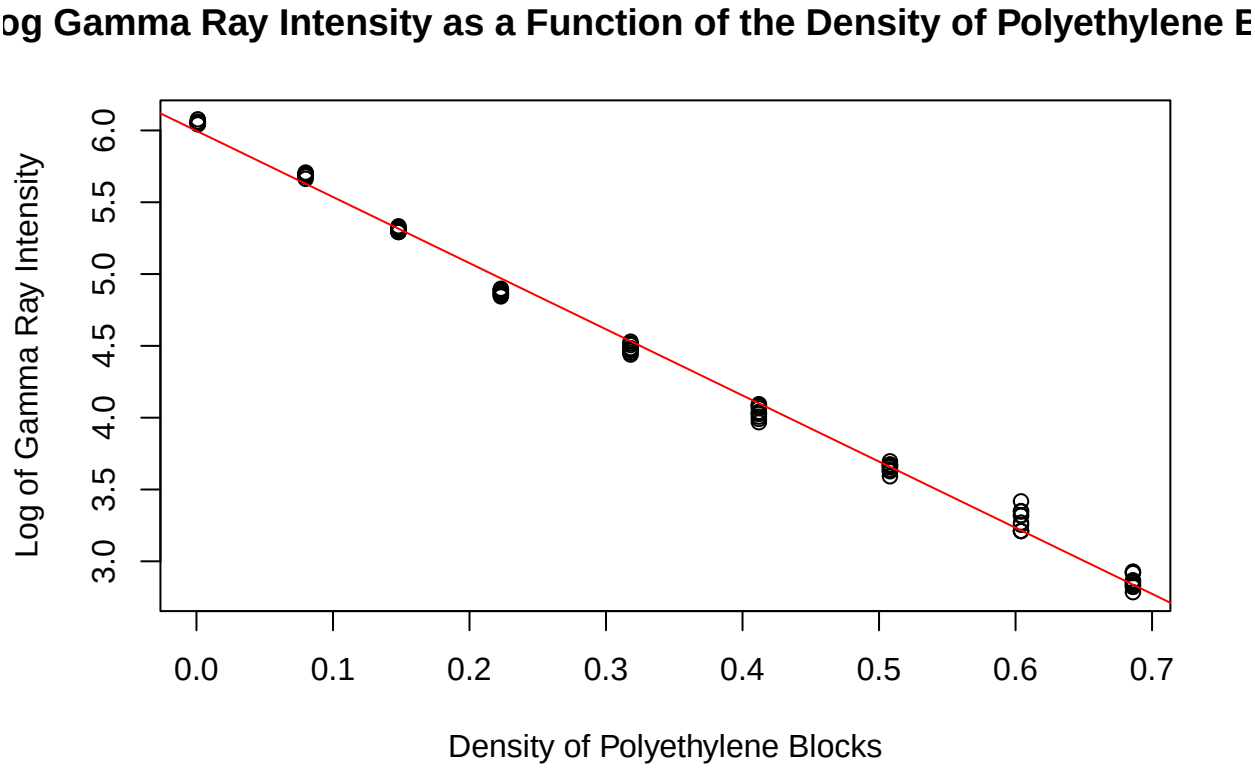
also not normally distributed. This means our data failed the linearity, heteroscedisticity, and normality conditions. Since a Shapiro-Wilk test confirmed that our data is not normal, we must transform our data to help normalize it.

Question 2: TRANSFORMED DATA

Methods

To find a fitting transformation for our data, we experimented with a log transformation on gain and graphed lines of best fit.

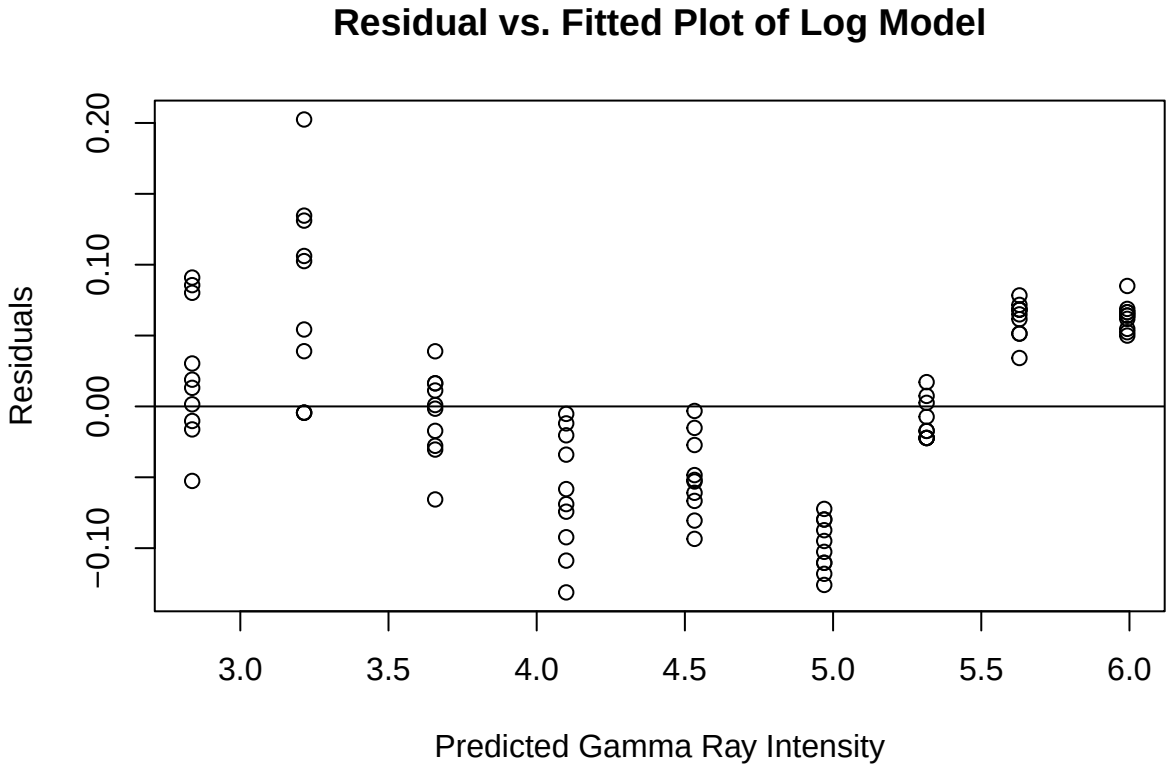
```
##
## Call:
## lm(formula = lggain ~ density, data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.131216 -0.052396 -0.004436  0.054607  0.202447
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.99727    0.01274   470.8   <2e-16 ***
## density     -4.60594    0.03182  -144.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06792 on 88 degrees of freedom
## Multiple R-squared:  0.9958, Adjusted R-squared:  0.9958
## F-statistic: 2.096e+04 on 1 and 88 DF,  p-value: < 2.2e-16
```



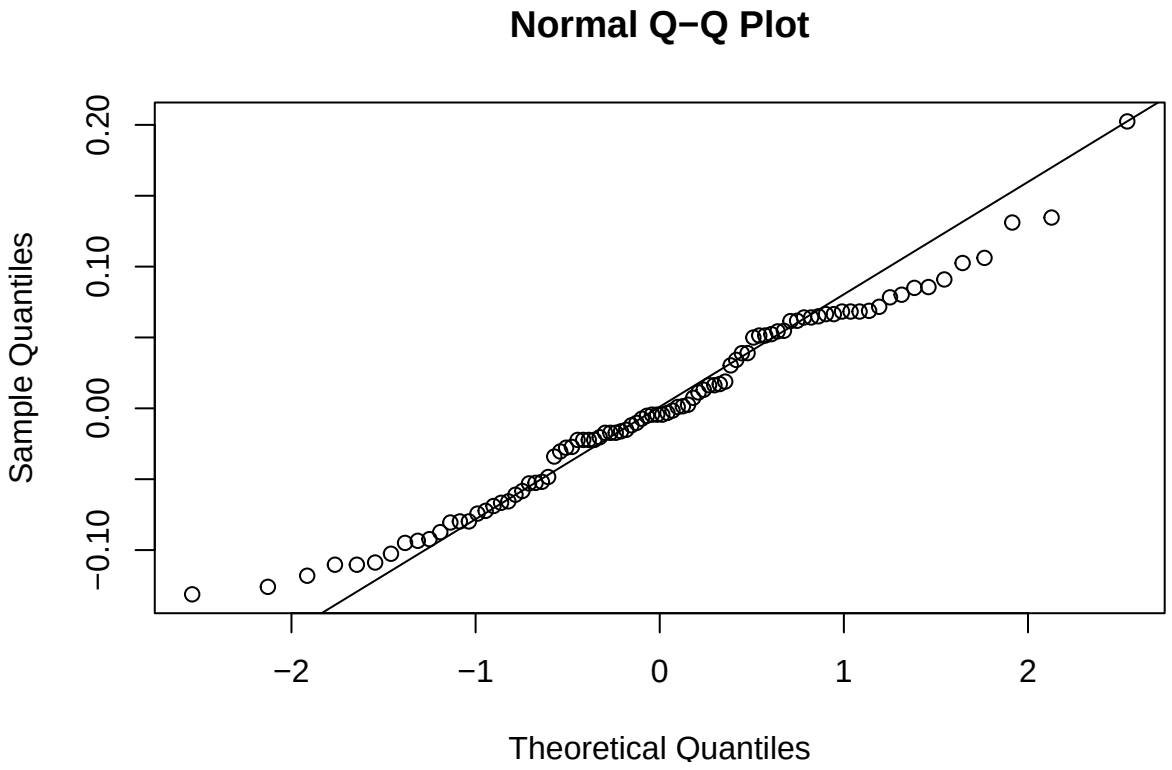
After performing a log transformation on the measured gamma ray intensity, the data becomes linear and a

linear model is able to fit the data very closely with a negative slope of -4.605.

Next, we plotted the residuals of the model to visualize the distribution of each data point.



Addtionally, we plotted a QQ plot of these residuals to ensure that they are normally distributed.



The plot of the residuals and the QQ plot also support the idea that the data is now linear after the log

transfomation, as the resisuals are more randomly scattered and they fall close to the normal line on the QQ plot.

Analysis

Based on our visualizations, the log transformation is a fitting transformation for our data to fit a linear model. To reinforce this idea, we can test whether our data comes from a normally distributed population with a Shapiro-Wilk test. Our null hypothesis is that our data does come from a normally distributed population.

```
set.seed(123)
shapiro_result <- shapiro.test(data$lggain)
print(shapiro_result)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$lggain
## W = 0.93581, p-value = 0.0002461
```

With a p-value of 0.0002461< 0.05, we reject the null hypothesis. Although this suggests that our data is *not* normally distributed, it is a vast improvement from the previous p-value of 4.599 * 10^(-9).

Conclusion

The log transformation on predicted gamma ray density is an appropriate for fitting the data to a linear model. The visualizations of the model itself, its residuals, as well as a QQ plot show us that the log transformation allowed the data to fit a linear model very well. Although the log transfomation failed to completely transform the distribution of predicted gamma ray density into a normal distribution, it significantly reduced the non-linearity of its distribution.

Question 3: ROBUSTNESS

Question 4: FORWARD PREDICTION

Question 5: REVERSE PREDICTION

Question 6: CROSS-VALIDATION

Advanced Analysis

Conclusion & Discussion