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

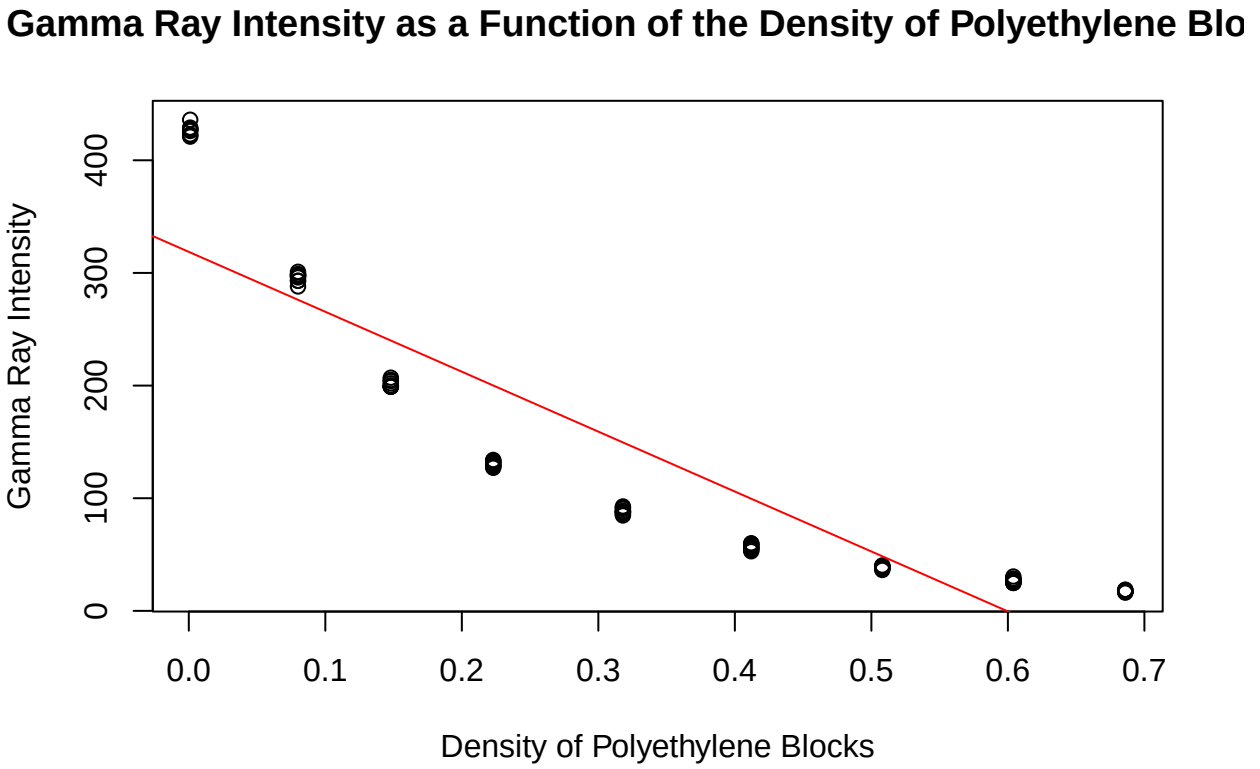
Methods

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

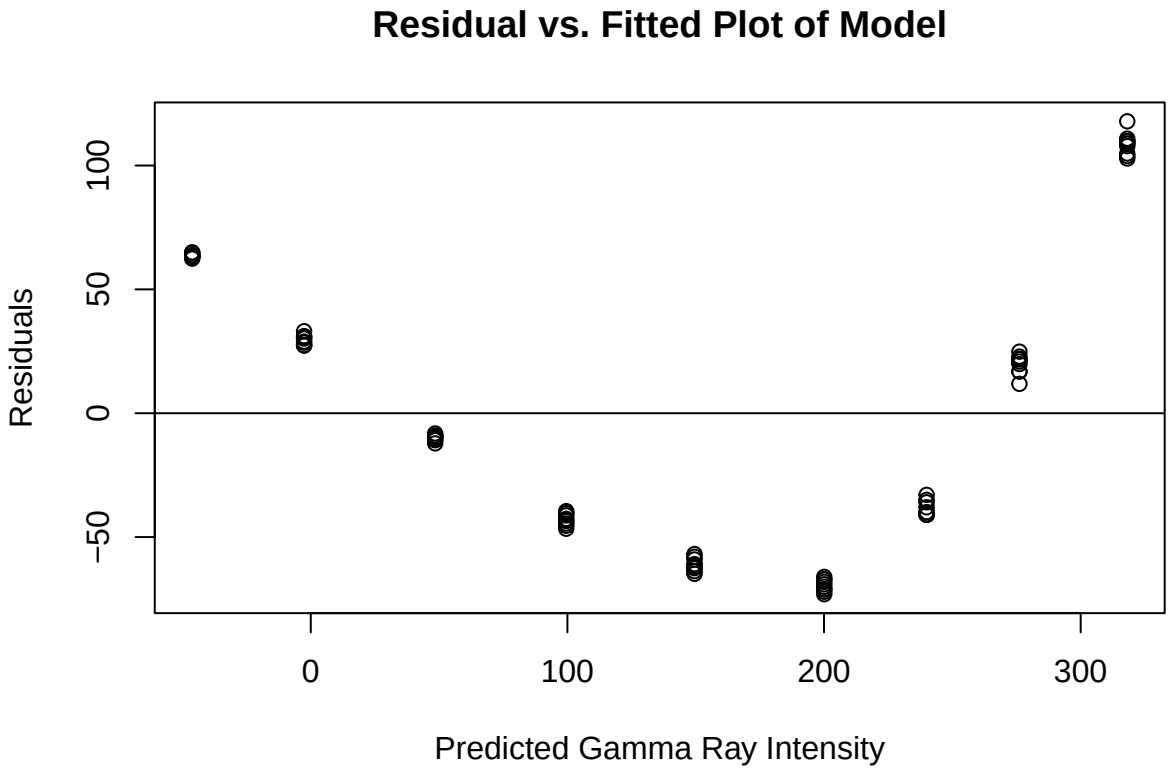
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	318.7	10.79	29.53	1.906e-47
density	-532	26.95	-19.74	4.519e-34

Table 2: Fitting linear model: gain ~ density

Observations	Residual Std. Error	R^2	Adjusted R^2
90	57.54	0.8157	0.8136

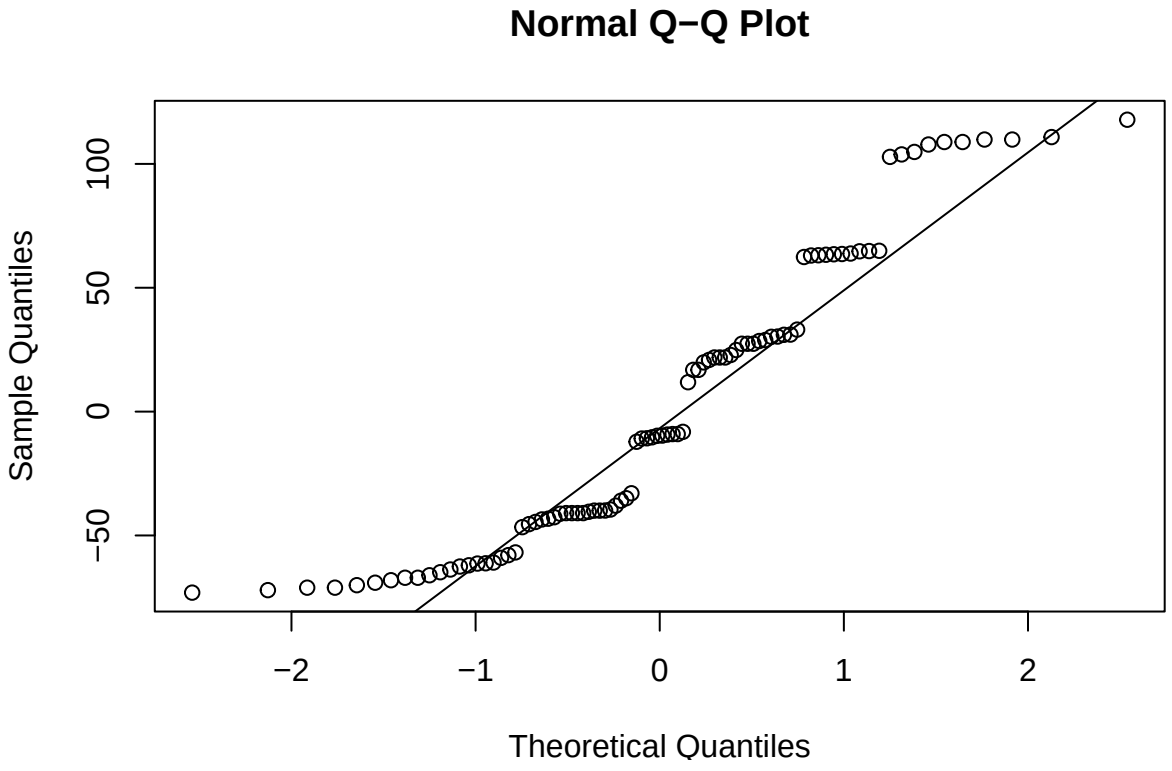


We will extract residuals from the model and observe the residual vs. fitted plot (the predicted values) to understand whether our model has a good fit.

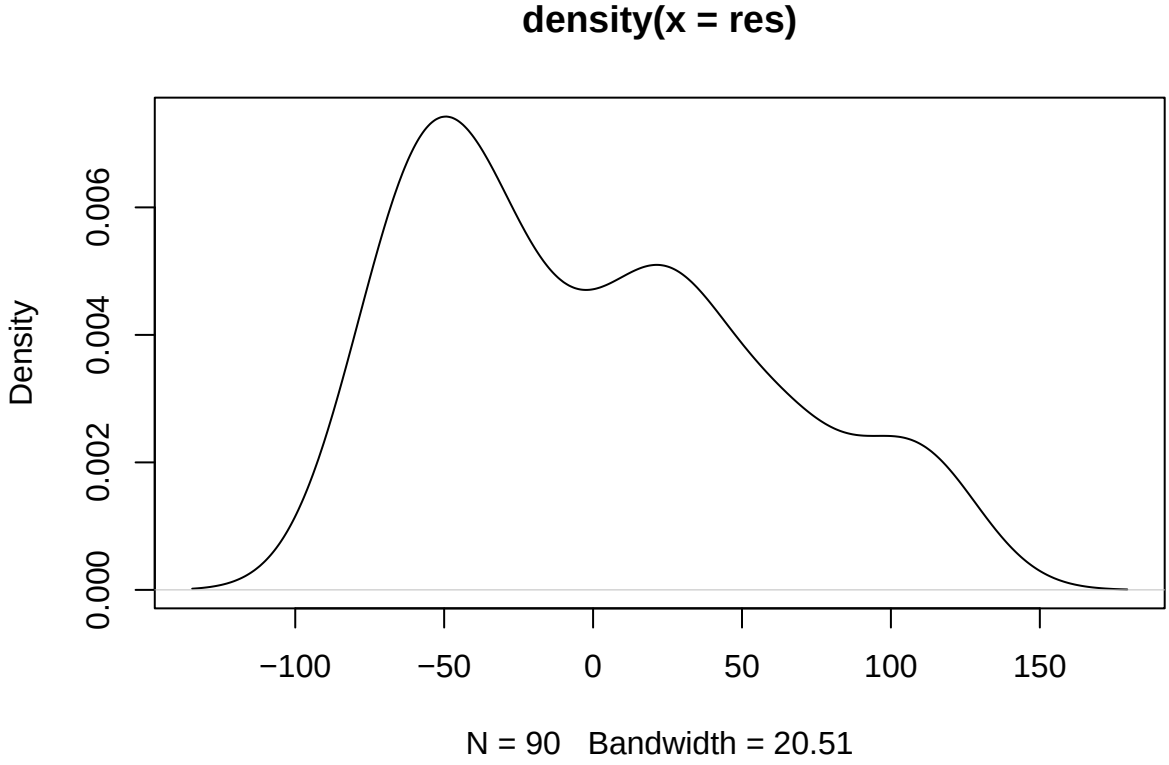


There is a very clear nonlinear pattern between the predicted values from the model and the residuals, which suggests the current model is inappropriate for the data since the relationship between the response variable and the residuals are not linear. There is also homoscedasticity present since the 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.

Table 3: Shapiro-Wilk normality test: `data$gain` With a p-value of 4.599 * 10^(-9) < 0.05, we reject the null hypothesis. This suggests our data is *not* normally distributed.

Test statistic	P value
0.8212	4.599e-09 * * *

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, heteroscedasticity, 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

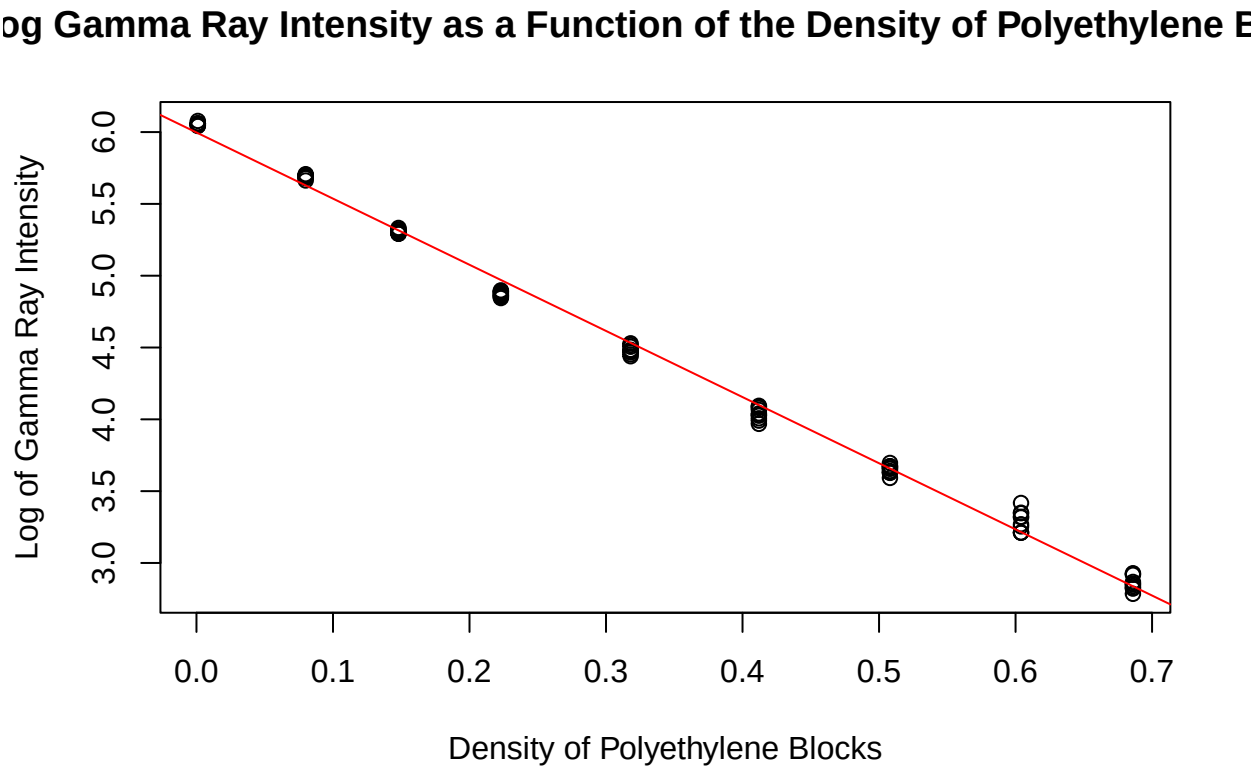
Methods

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.997	0.01274	470.8	1.843e-151
density	-4.606	0.03182	-144.8	1.857e-106

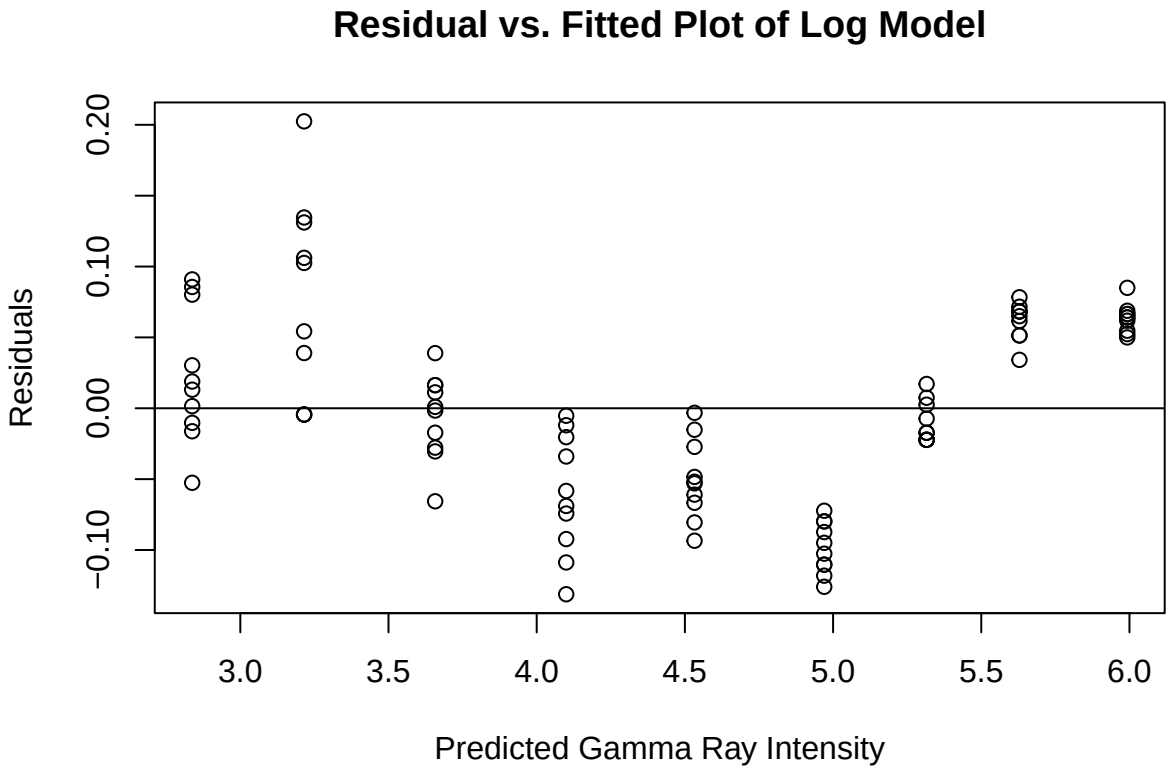
Table 5: Fitting linear model: lggain ~ density

Observations	Residual Std. Error	R^2	Adjusted R^2
90	0.06792	0.9958	0.9958

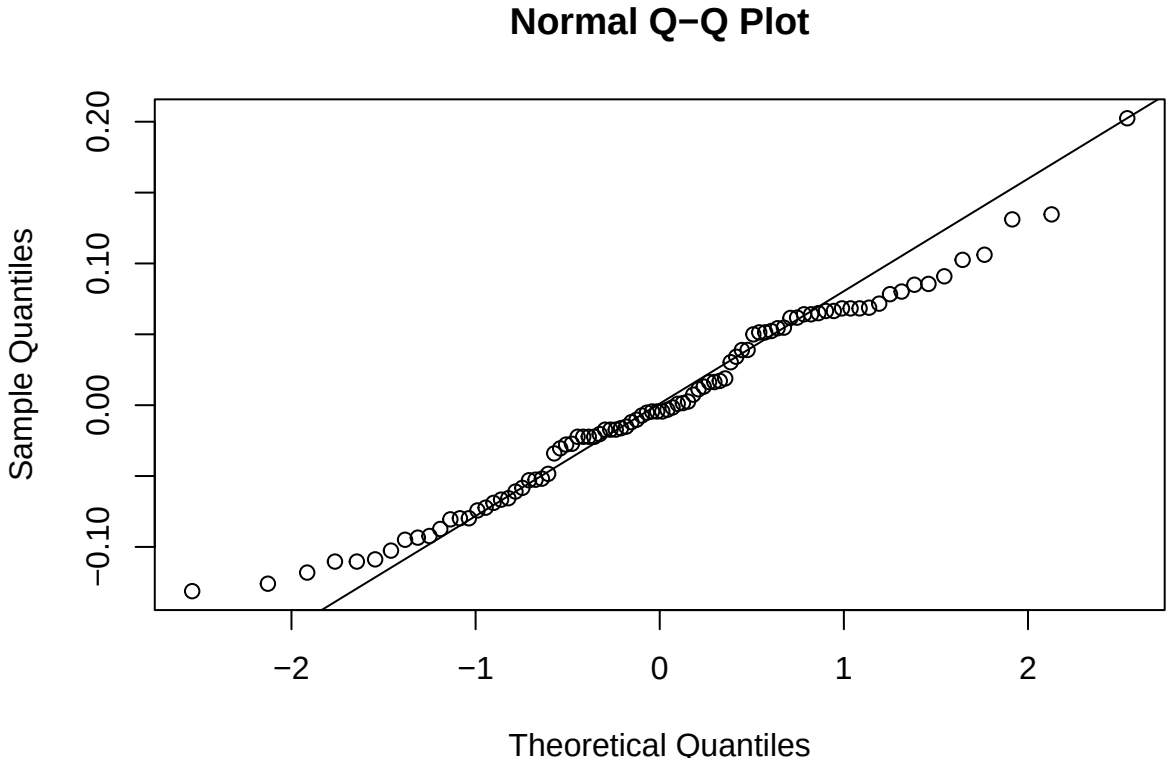


After performing a log transformation on the measured gamma ray intensity, the data becomes linear and a linear model is able to fit the data very closely with a negative slope of -4.605.

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 transformation, 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.

```
##
##  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 the residuals 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 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 transformation failed to completely transform the distribution of predicted gamma ray density into a normal distribution, it significantly reduced the non-linearity of its distribution.

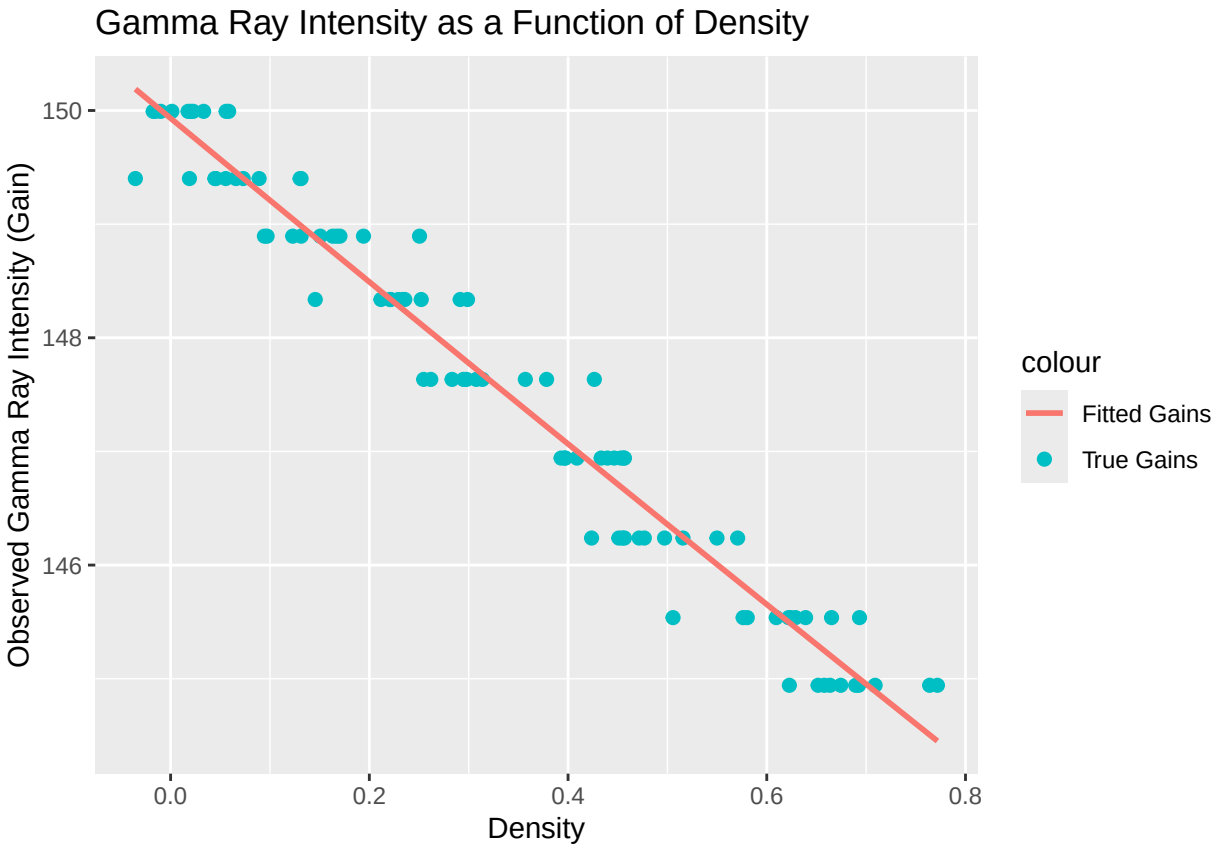
Question 3: ROBUSTNESS

Methods

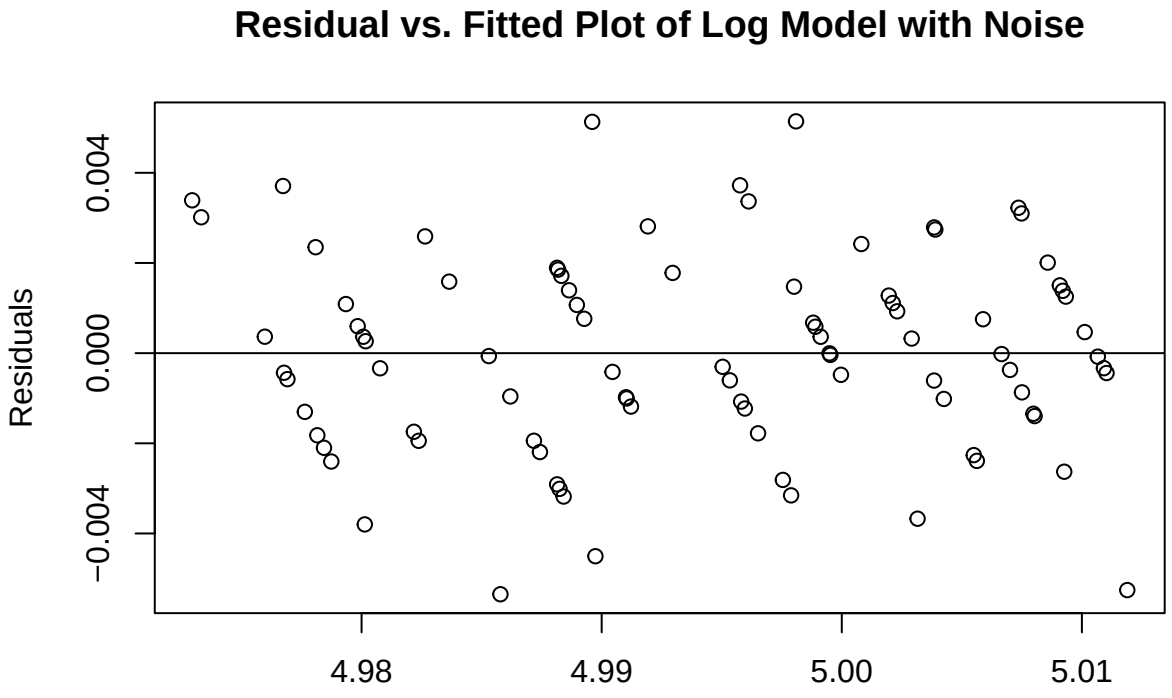
If the densities of the polyethylene blocks are not reported exactly, we can use the error term in the prediction interval to take into account the variability in the gamma ray intensity (the gain).

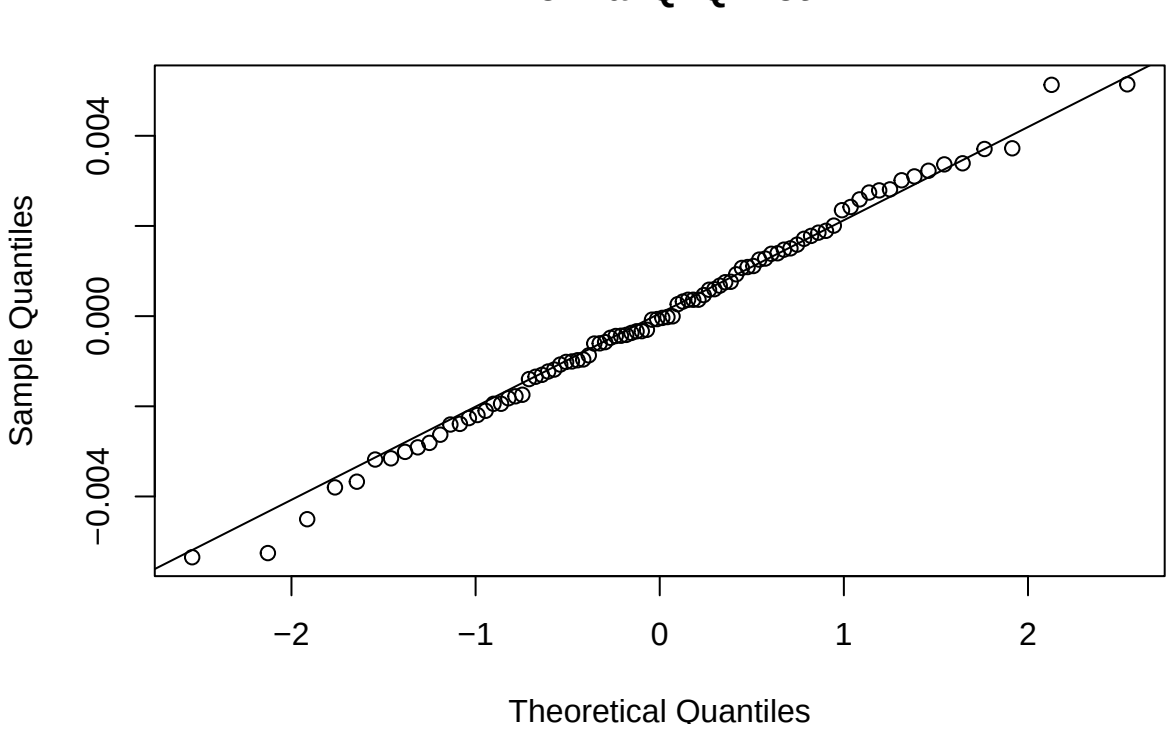
We have generated synthetic data that includes noise for the densities, fit a model to predict gains based on the densities with noise, and created predictions of the gains using the newly fitted model.

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Based on the graph above, our new model that accounts for noise in density closely predicts the true gains from our dataset! Below, we will plot the residuals of the model.





After accounting for noise, it seems as though our residuals follow the line y = x more closely; our residuals are closer to being normally distributed than in our previous model.

Analysis

We will extract the residuals from our model that accounts for noise in density and use the `{r} shapiro.test()` function on the residuals. Our null hypothesis is that the residuals are normally distributed.

Table 6: Shapiro-Wilk normality test: `res_noise`

Test statistic	P value
0.9944	0.9707

Since the p-value = 0.9707 > 0.05, we fail to reject the null hypothesis. The residuals are normally distributed, so accounting for variations in the densities of the polyethylene blocks *does* result in a more accurate fit.

Conclusion

We created a new model under the assumption that densities of the polyethylene blocks were not reported exactly. Since the residuals closely align the 45-degree line in the QQ plot and a Shapiro-Wilk test confirms the residuals are normally distributed, our new fitted model that accounts for variation in densities performs *better* than without accounting for variations.

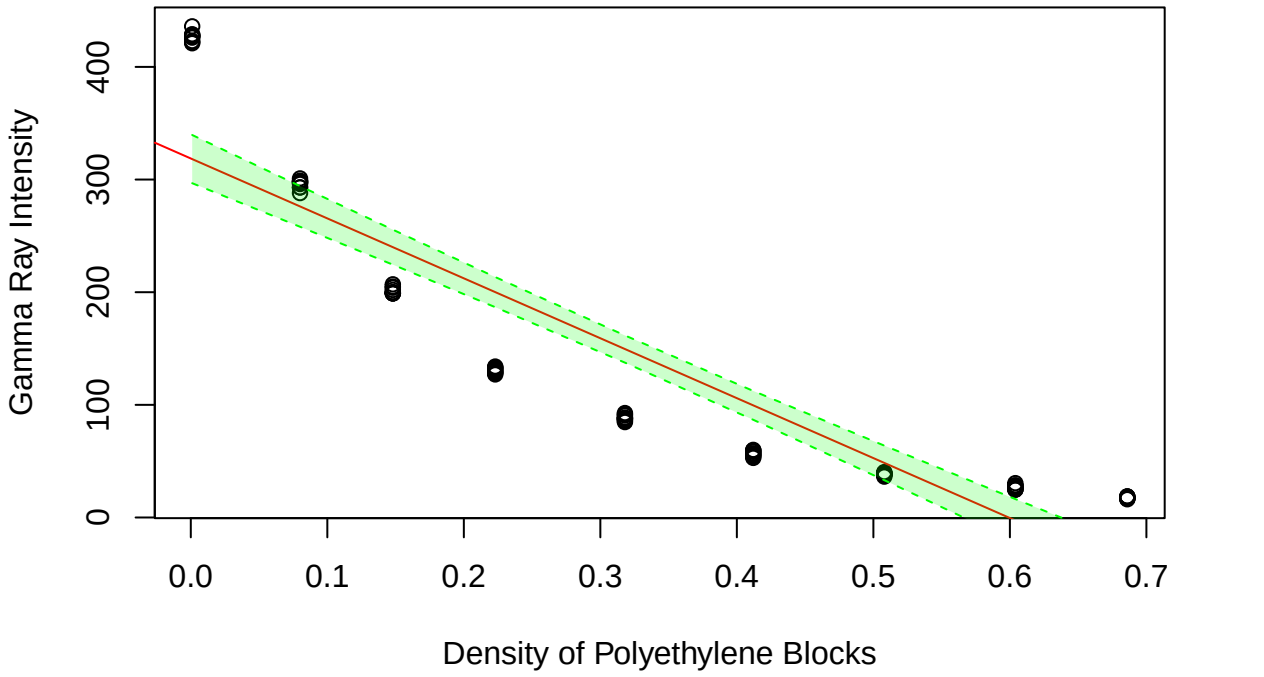
Question 4: FORWARD PREDICTION

Methods

Below is the basic summary of our data.

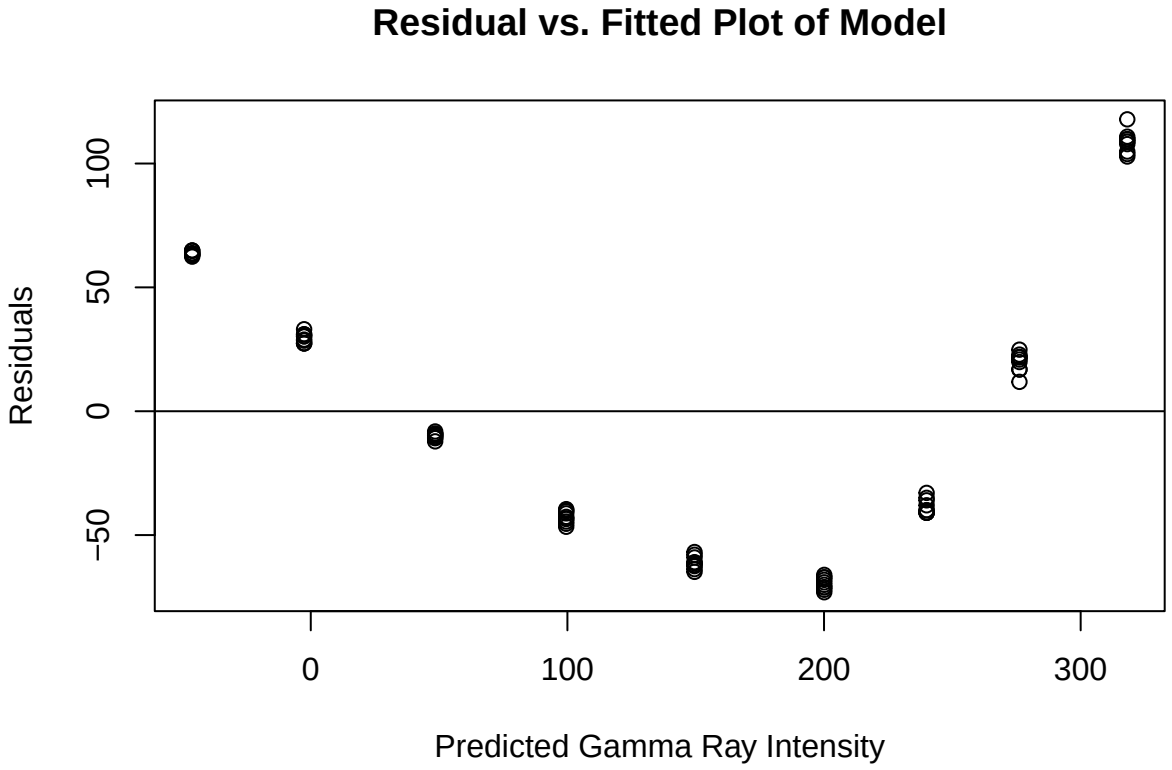
gain	density
Min. : 16.20	Min. :0.0010
1st Qu.: 37.80	1st Qu.:0.1480
Median : 88.25	Median :0.3180
Mean :142.57	Mean :0.3311
3rd Qu.:203.50	3rd Qu.:0.5080
Max. :436.00	Max. :0.6860

We will reuse our model from Question #1 and now include confidence intervals in our visualization to represent the uncertainty bands.



Analysis

From our visualization with the confidence intervals included, it is clear that there are two specific densities whos gain measurements are accurately predicted by the model: 0.1 and 0.5. The predicted gain measurements include the actual gain measurements of those values. Additionally, the recorded densities between these measures fall relatively close to to the linear model in comparison to densities that fall outside of this interval. This can also be visualized in the residual plot of this model, where the predicted gains of densities in this interval are [0, 300]. The residuals within this interval are much closer to 0 than residuals outside this interval.



Conclusion

Based on both the linear model as well as the residual plot, it is reasonable to conclude that some gains can be predicted more accurately than others. This is best represented by the density interval [0.1, 0.5]. On the linear model plot, the data points within this interval fall within the confidence interval or very close to it compared to data points that are outside the interval. The residual plot reinforces this idea, as the predicted gains of densities within this interval are closer to 0 than the predicted gains of densities outside of this interval.

Question 5: REVERSE PREDICTION

Question 6: CROSS-VALIDATION

Advanced Analysis

Conclusion & Discussion