Finding the Relationship between Density and Gain

Author 1 and Author 2

2024-11-24

Header

Author Contributions

Author 1: Contributed to questions 1, 4, 5, and did the formatting for the pdf

Author 2: Contributed to questions 2, 3, and 6, along with doing the advanced analysis.

Use of GPT

ChatGPT was used as a substitute for documentation for R. Since we were unfamiliar with R, we asked ChatGPT how to use R in certain methods in order to find and filter out conditions in the dataset. We additionally used GPT to analyze reasoning and to confirm what we thought was correct about the dataset, as well as to identify extra questions that could be answered for our advanced analysis.

Introduction

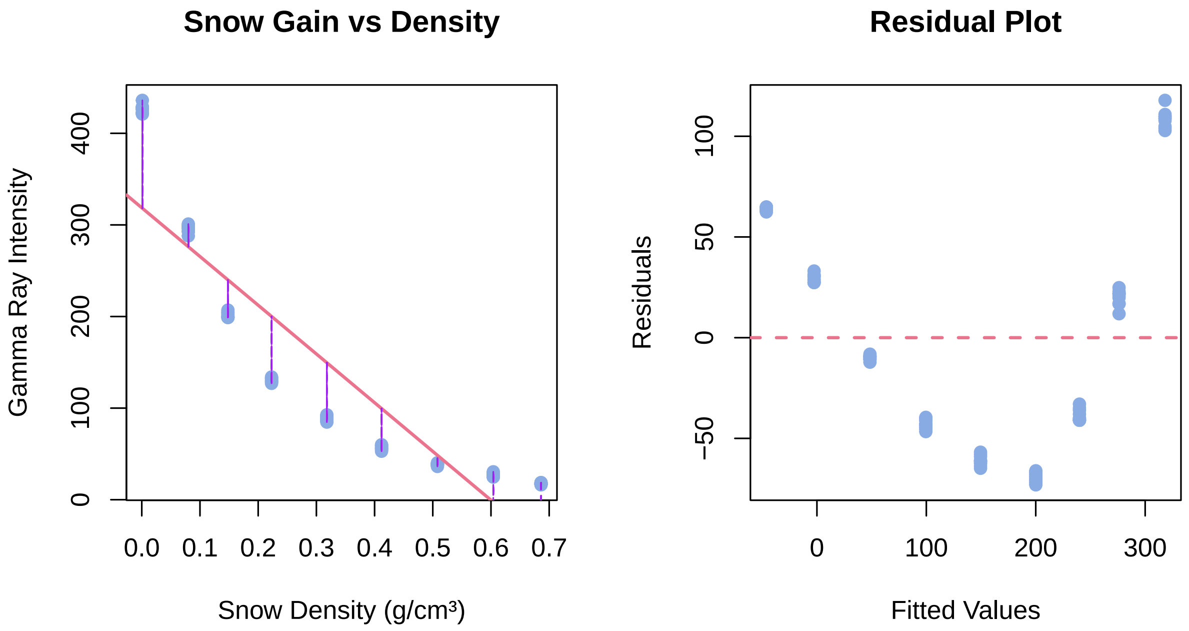
The data used in this analysis comes from a gamma-transmission snow gauge calibration process in the Sierra Nevada mountains of Northern California. The dataset consists of two columns - density measurements of polyethylene blocks (in grams per cubic centimeter) and their corresponding gamma ray intensity readings (gain). For each of the 9 different densities tested, 10 repeated measurements were taken, resulting in 90 total observations. Our objective in this analysis is to establish the relationship between gamma ray intensity and material density, and determine if this relationship follows the expected exponential decay pattern. Essentially, we are testing if the calibration data can provide a reliable inverse function to convert field measurements of gamma intensity into accurate snow density estimates, which is crucial for monitoring water supply in the region. ## Main Research Questions

1. How well does a regression line fit to the data. Examine the residual plot and explain why a transformation is necessary.
2. Determine an appropriate transformation and fit the model to the transformed data, plotting the new fit and examining residuals.
3. Suppose the densities of the polyethylene blocks are not reported exactly. How might this affect the fit, using a simulation to answer the question.
4. Produce point estimates and uncertainty bands for predicting the gain as a function of the measured density. Can some gains be predicted more accurately than others?
5. Invert the forward prediction line and uncertainty bands to produce point estimates and prediction intervals for the density that correspond to the gain measurements 38.6 and 426.7.

6. The reverse prediction for density values 0.508 and 0.001 may be influenced by the fact that the measurement corresponding to the densities 0.508 and 0.001 were included in the fitting. To avoid this, omit the set of measurements corresponding to the block of density 0.508, apply your estimation/calibration procedure (forward fit and reverse prediction) to the remaining data, an provide an interval estimate for the density of a block with an average reading of 38.6. Where does the actual density fall in the interval? Try the same test, for the set of measurements at the 0.001 density.

Question 1: Fit a Regression Line

Methods



```
## [1] "The R² value is 0.816"
```

Analysis

We implemented a regression line along with a residual graph for the snow gain and density measured to determine what type of relationship could be found when analyzing graphically. The regression graph does not look like it follows a line of best fit, potentially showing that the relationship between these two variables is not linear. Additionally, when plotting the residual plot, the same conclusion can be reached. Since the residual plot does not look randomly distributed along the zero line, as a linear relationship would typically be, it shows that our line of best fit doesn't capture the relationship between our variables correctly. Since the line follows an almost parabolic curve, this indicates that there's a good chance that the initial relationship between the two variables could be quadratic, meaning the data needs a transformation or a different model to test something that can accurately capture their relationship.

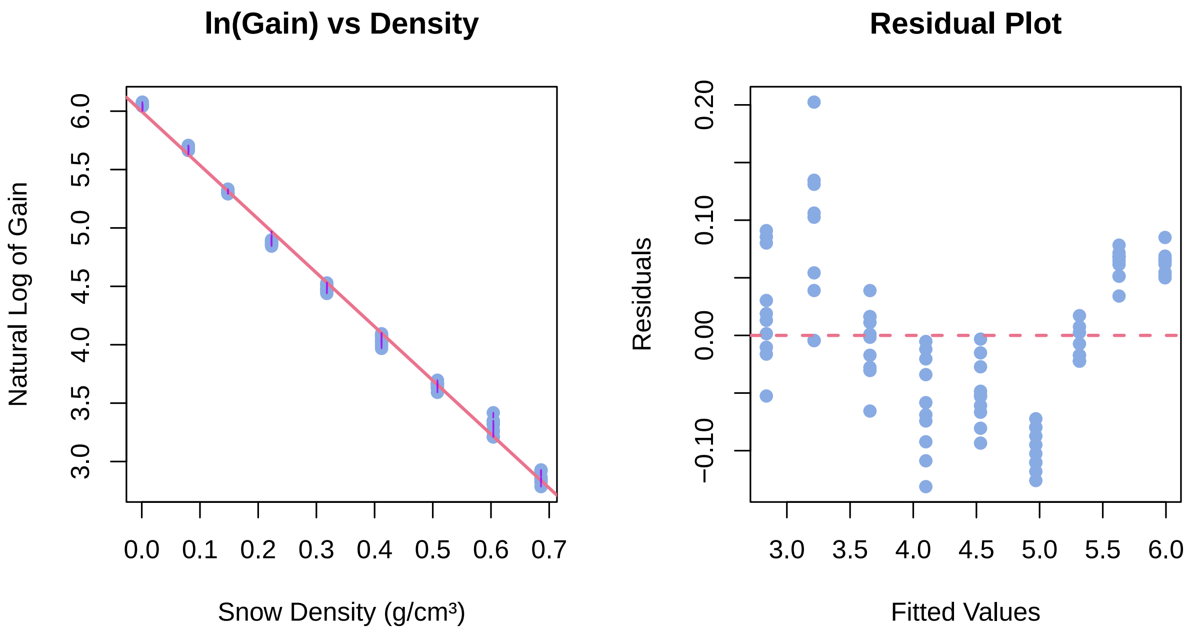
Conclusions

With our R^2 of the model turns out to be 0.816, which means that only 81.6% of the variance of the gain can be explained by the variance in the snow density, meaning there is a likely chance that this model is not one that is extremely predictive of the data, which means it is likely that a transformation is necessary for the data to fit better.

Question 2: Transform Data and Fit

Methods

Based on the theoretical exponential relationship between density and gain, we applied a natural logarithm transformation to gain and looked for a linear relationship. We use a linear regression model to predict the logarithm of gain from density.



```
## Logarithm Regression Results
```

```
## [1] "The R² value for the log_model is 0.996"
```

We compare the above regression results to the regression results before applying the transformation. The

$$R^2$$

metric is useful for comparing the performance of the two models. If our transformation is successful, the transformed model will have a higher

$$R^2$$

, meaning that more of the variance in the response variable (log_gain) can be explained by our independent variable.

```
## Regression Results Before Transformation
```

```
## [1] "The R² value for our regreesion model is 0.816"
```

Analysis

According to physics, the relationship between density and gain is exponential, not linear. Transforming the data with a natural logarithm reveals a linear relationship because the theoretical relationship is

$$g = Ae^{Bd}$$

, which transforms into

$$ln(g) = ln(A) + Bd$$

, establishing a theoretical linear relationship between density and the logarithm of gain.

According to the results of the regression, the logarithm transformation greatly improves the strength of correlation detected. After performing the logarithm, the R^2 increased from **0.816** to **0.996**. This means 99.6% of the variance in the logarithm of gain can be explained by variance in the snow density. Looking at the actual values of residuals, they are in the range of **-0.15 and 0.2**, compared to the pre-transformation values of **-80 and 130**. This also shows a massive improvement in the regression after applying the logarithm transformation.

The shape of the residual plot still has a very slight parabolic shape but it is much less clear than the shape of the pre-transformation residual plot. Additionally, the dramatic decrease in residual values resulting from the transformation is evidence for its validity.

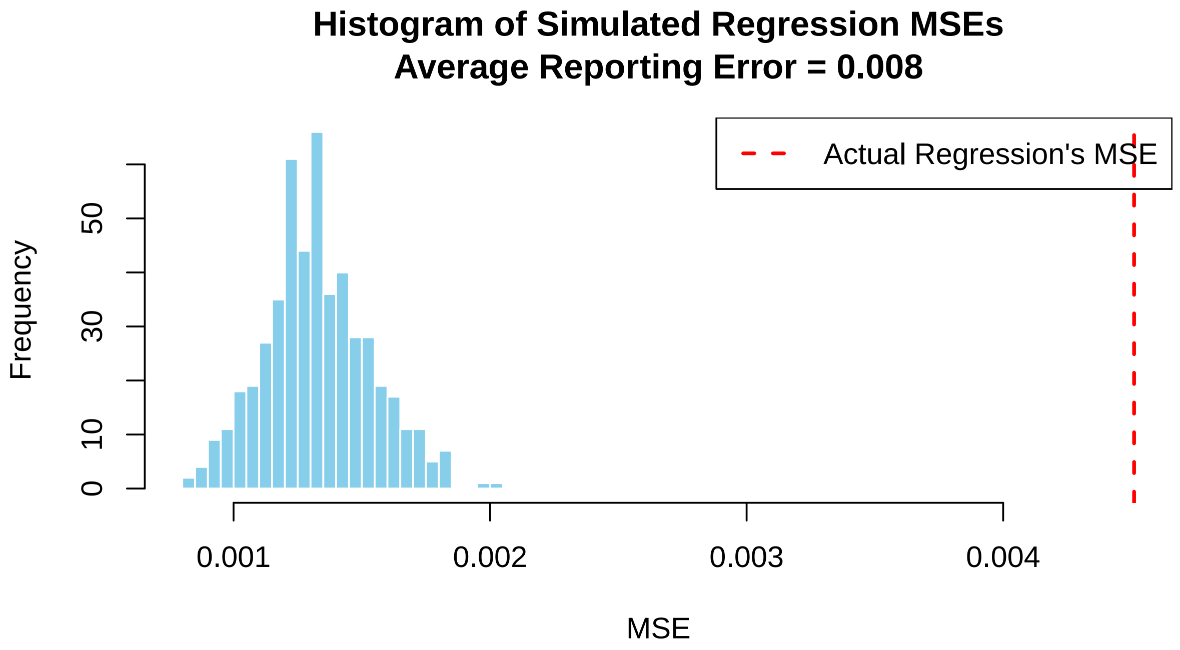
Conclusions

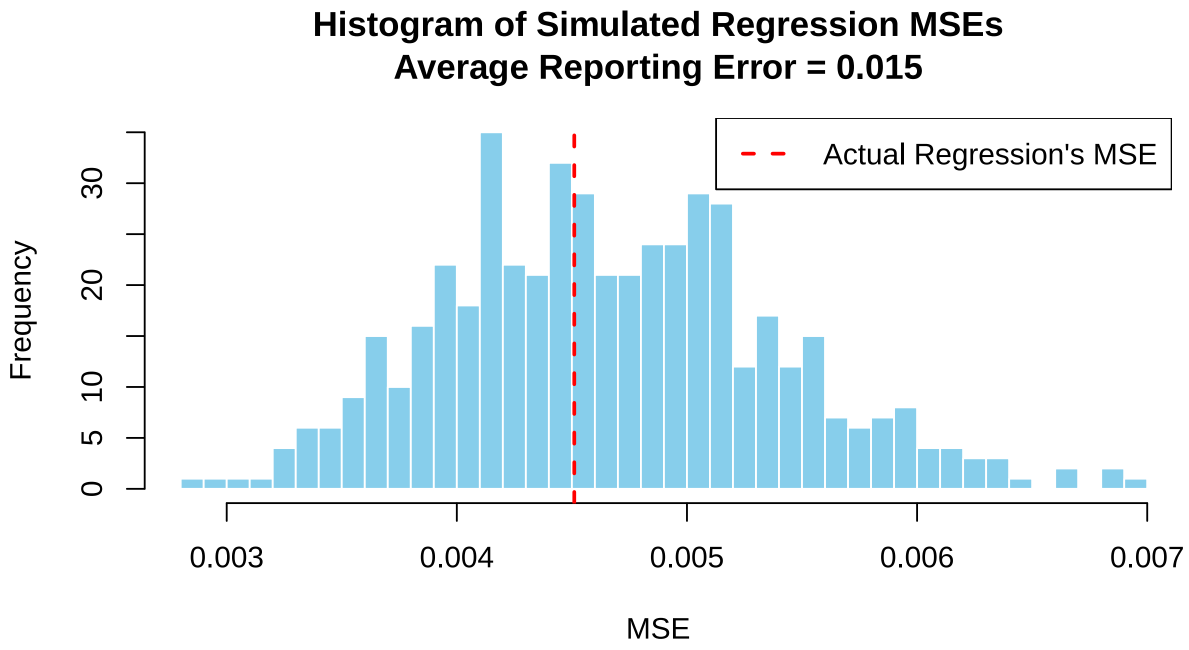
Applying a logarithmic transformation to the gain before fitting the linear model significantly improves the fit of the model. The R^2's increase from 0.816 to 0.996 shows that the model, after the transformation, is able to explain a greater proportion in the variance of the responding variable, log(gain).

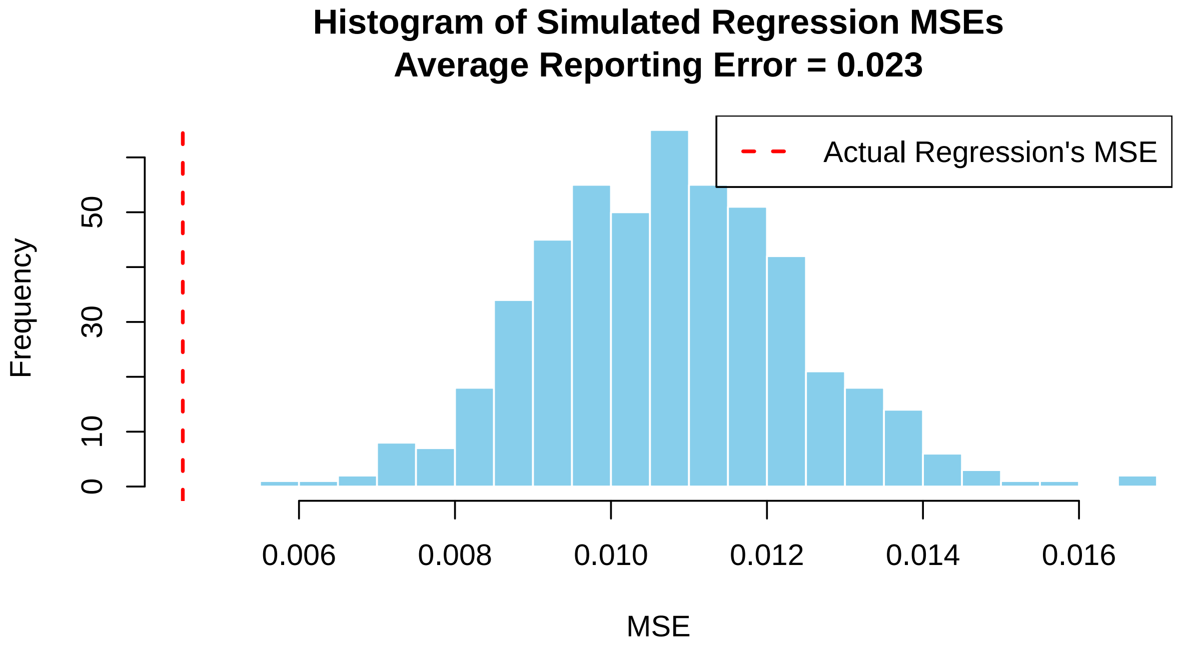
Question 3: Effect of Random Error

Methods

In order to study the effects of rounding errors, we draw simulated errors from a normal distribution and test different standard deviations for the error distribution to see how it affects the MSE of regression on the simulated erroneous data. Below, the histograms can be seen for when the distribution of density reporting error has a standard deviation of **0.008**, **0.015**, and **0.023**.


Histogram of Simulated Regression MSEs
Average Reporting Error = 0.008
Actual Regression's MSE
Frequency
0
10
30
50
0.001
0.002
0.003
0.004
MSE


Histogram of Simulated Regression MSEs
Average Reporting Error = 0.015
Actual Regression's MSE
Frequency
0
5
10
20
30
0.003
0.004
0.005
0.006
0.007
MSE



Analysis

The above visualizations show how the MSE of regression responds to different levels of density reporting error. In context, an average reporting error of **0.015** means that the standard deviation of the distribution of error between measured and true density is **0.015**. The mean of this distribution would be 0 because on average, the measurements should approximate the true density.

As we increase the amount of reporting error from **0.008** to **0.023**, it is clear that the mean squared error (MSE) of the resulting regression increases, meaning the model gets less accurate with increasing reporting error, unsurprisingly. The median MSE for regressions done on simulated reporting error of **0.023** is around **0.011**, which is much higher than the median MSEs of the simulated regressions under smaller reporting errors.

Under the assumption that density reporting errors come from a normal distribution centered around 0 with a standard deviation of **0.015**, the visualization makes it clear that our observed MSE from the actual regression is not unusual. However, under the assumption that the errors come from a normal distribution with a standard deviation of **0.008** or **0.023**, it would **not** be likely to observe the MSE of our regression on the actual data. This could imply that the true amount of reporting error is closer to **0.015** than 0.008 or 0.023. However, it is worth noting that no conclusive statements can be made about the actual reporting error because other factors could be causing the variation in predicted and actual gains.

We can see from the three visualizations above that as the average reporting error increases, the resulting MSEs increase meaning the regression fit is weaker. This makes sense because more noise/error in measurements would lead to a worse regression result.

Conclusions

After conducting the above simulations, it is clear that increasing the amount of random reporting error in densities would increase the Mean Squared Errors (MSEs) of the resulting regressions, meaning the regressions have a weaker fit and more error. The actual observed MSE of our regression most closely matched the simulated regressions under average reporting error of **0.015**.

Question 4: Forward Prediction

Methods

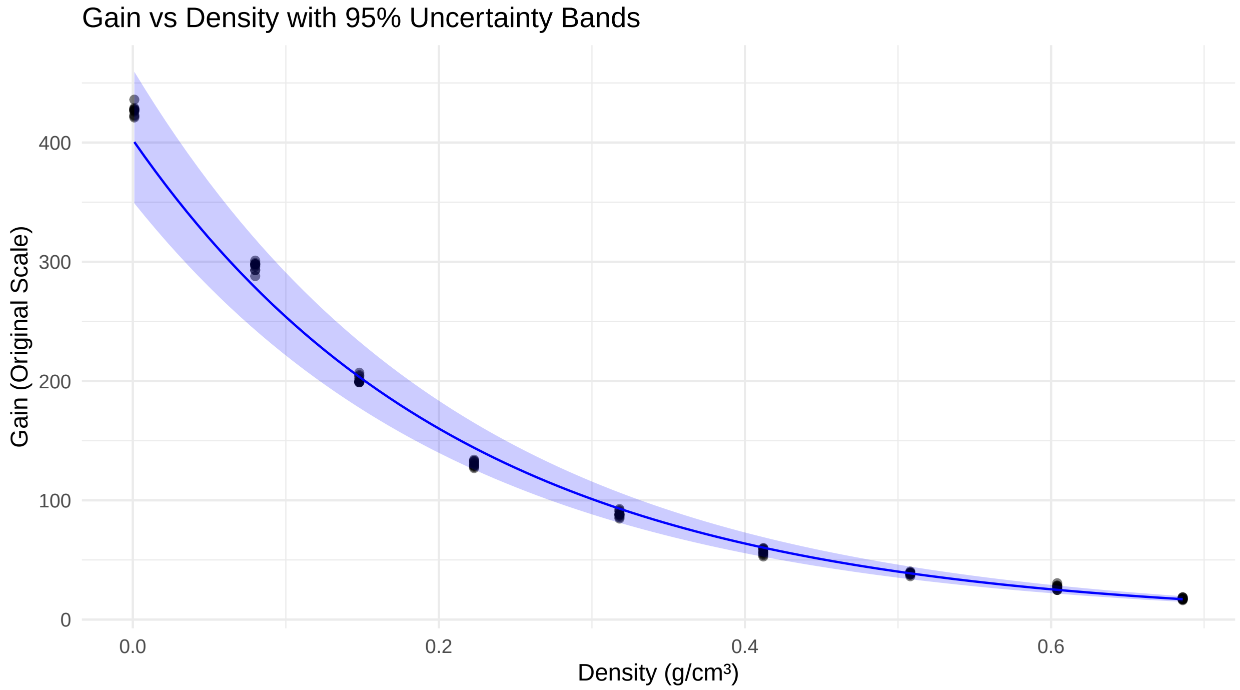


Table 1: Prediction Intervals (Original Scale)

Density	Predicted Value	95% Uncertainty Bands	Actual Range
0.001	400.48	349.09 to 459.43	421 to 436
0.080	278.33	242.78 to 319.08	288 to 301
0.148	203.48	177.57 to 233.18	199 to 207
0.223	144.05	125.74 to 165.01	127 to 134
0.318	93.00	81.2 to 106.52	84.7 to 92.7
0.412	60.32	52.66 to 69.09	52.9 to 60
0.508	38.76	33.83 to 44.42	36.3 to 40.3
0.686	17.07	14.88 to 19.59	16.2 to 18.7

Analysis

In order to do a forward prediction with the logarithmic model we created earlier, we have to untransform it's predictions after predicting gain from density. In order to do that, we exponentiated with the natural base in order to get predicted values in our original untransformed scale. Second, we produced point estimates and uncertainty bands (with a 95% confidence interval) for predicting the gain, where the the returned output of our function returns the interval as well as the point estimate of the predicted function.

Conclusions

Looking through the graph that we created, it can be seen that the 95% confidence interval uncertainty bands that we created worked well, encompassing almost all of the observed values. However, it can be see that some gain values are more easily predicted than others. For instance, if we look at the given densities of 0.508 g/cm^3 and 0.001 g/cm^3, it's seen that the predicted value and uncertainty band of the gain at 0.508 is very much in line with the actual range of the observed values, as the predicted value of 38.76 gain is well within the actual range of 36.3 to 40.3 gain in the observed data. This gives us a sense that at higher values of density within the data like 0.508 g/cm^3, our model is more accurate. On the other hand, at density 0.001 g/cm^3, while our uncertainty band does capture the actual range of the observed data, with the actual range of gain of 421 to 436 gain within our uncertainty band of 349 - 459 gain, our predicted value does not. Our predicted value of 400.48 is not in the actual range of 421 - 436 gain present in the observed data, implying that while our model does a decent job of generalization and prediction of the model, it struggles with the very small value of 0.001 g/cm^3 at the beginning of the observed densities, which can also be noticed with the decreasingly small size of the density bands as the density increases. Some important parts to notice: the model predicts best at densities 0.148, 0.508, and at 0.686, where the predicted value falls within the actual range of the data. On the other hand, densities of 0.001, 0.080, 0.223, 0.318 and 0.412 have uncertainty bands that capture the actual range but fail to predict a value that falls within the actual range. Overall, an observation to notice is that as the density increases within our predicted model, the uncertainty band decreases in size, potentially exhibiting that gain converges to smaller and more tightly observed values easily as density increases.

Question 5: Reverse Prediction

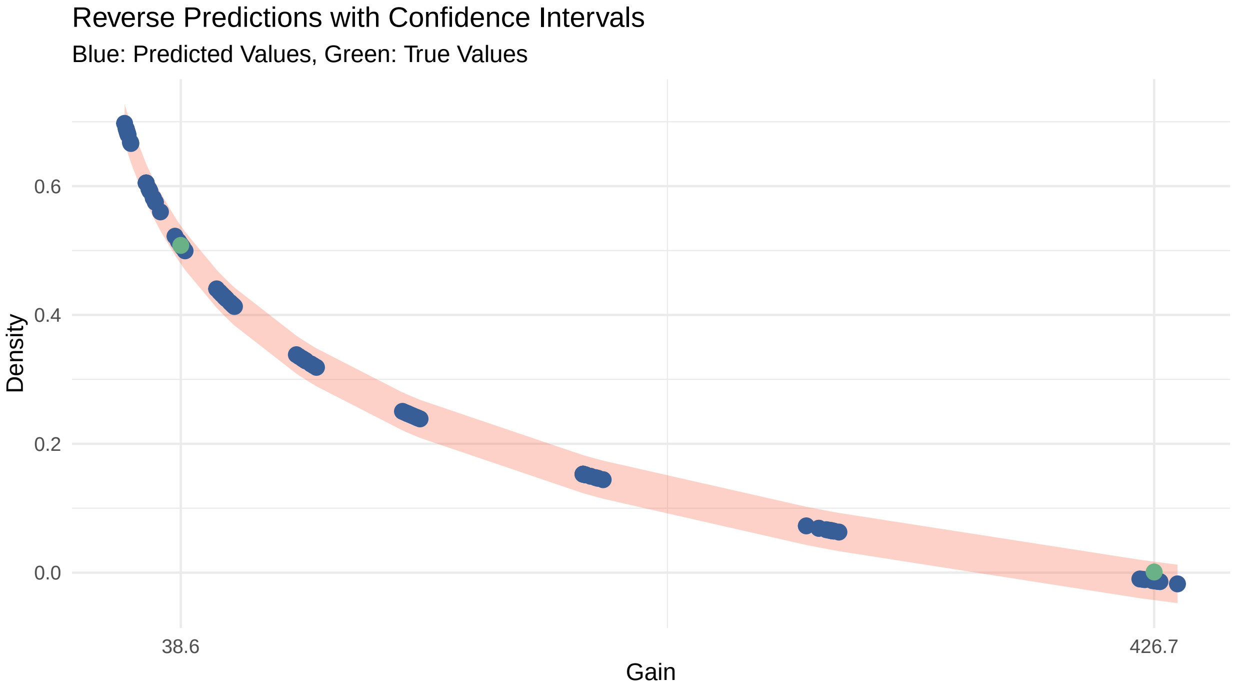


Table 2: Predictions vs Actual Densities

Gain	Predicted Density	Lower CI	Upper CI	Actual Density	Prediction Error
38.6	0.5089	0.5385	0.4793	0.508	0.0009
426.7	-0.0128	0.0171	-0.0426	0.001	0.0138

Analysis

In order to create a prediction for the averaged measured gains, we had to reverse the prediction from our models using the coefficients from our fitted logarithmic model. After doing this, in order to extrapolate our logarithmic model through all of the data, we created a graphical representation to illustrate the affect of gains and what a predicted density would entail. Doing this, we noticed some values that were easier to predict than others. For instance, at a gain level of ~40 and ~200, the predicted and true densities are nearly identical. On the other hand, there are gains where the predicted and true densities are different, as seen in the graph above. For instance, at gains ~130 and ~425, the densities are noticably different from each other. In fact, in the table above, it can be seen that the difference between the predicted and true densities in a gain like 38.6 is very minimal, with a 0.0009 difference, meaning our model predicts these gains very well. On the other hand, at a gain of 426.7, the difference in predicted and true density is 0.0138, which is significant, since the density scale only goes up to 0.686. Additionally, looking at our model, it can be observed that typically, the model generalizes better for predictions where the gain is smaller, because when the gain is large, the model can predict negative density values, as it did for the gain of 426.7. This implies that the smaller gains are easier to predict the density of, while the larger gains have densities that are harder to predict.

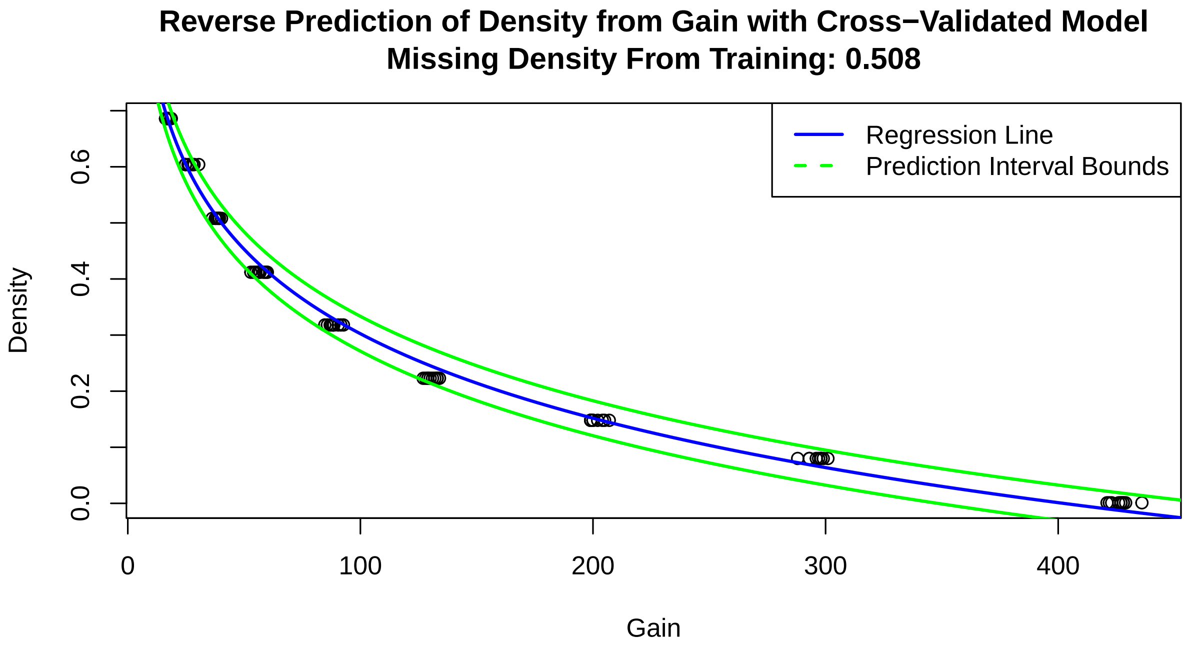
Conclusion

Using the reverse prediction method detailed above, we conclude that it is easier to predict densities where the gain is on it's lower range, while it's more difficult to predict densities where gain is higher, because the relationship present in our model creates a potential for negative predictions of density, which is not possible.

Question 6: Cross Validation

Methods

In order to evaluate the effectiveness with which our model generalizes to unseen data, we will use cross-validation, removing certain data points from the training set, re-training the model, and making predictions on the unseen data. Below, we remove all rows from the data corresponding to the block with density **0.508** and re-train the model. The reverse prediction along with prediction intervals can be seen below.



Using the cross-validated model, we make a prediction on a block producing a gain of **38.6**. If our model generalizes well, this prediction should give a density near **0.508**.

We repeat the same process, removing blocks of density **0.001** from the training set, reversing the prediction function, and predicting density from a gain of **426.7**.

```
##   Average_Gain Point_Estimate Lower_Bound Upper_Bound
## 1         38.6         0.5092      0.4780      0.5404
## 2        426.7        -0.0128     -0.0446      0.0190
```

Analysis

An initial visual inspection of the first cross-validated model (with density 0.508 missing from training) reveals that the model still generalizes well to unseen data. This is particularly evident in the regression plot where density is **0.508** because all of the data points at that density level seem to fall within the prediction interval and are relatively close to the regression estimate.

Our point estimate for the density of a block with a measured gain of **38.6** is **0.5092** with a 95% prediction interval of **(0.478, 0.54)**. The prediction interval means that there is a high (95%) likelihood that a block with an average measured gain of 38.6 has a density between 0.478 and 0.54. Since the actual density falls inside of the interval, our regression seems to generalize well. Not only is the true density inside the interval, but it is close to the middle of the interval (0.5092).

After running cross-validation again with the block of density 0.001 missing from the training set and using this model for reverse prediction of density, we test the model on a block with average gain of **426.7**, and get a density point estimate of **-0.013** with a prediction interval of **(-0.045, 0.019)**. Our true density, 0.001, falls within this interval, but it is further from the middle of the interval. Instead, the true density is closer to the upper bound of the interval.

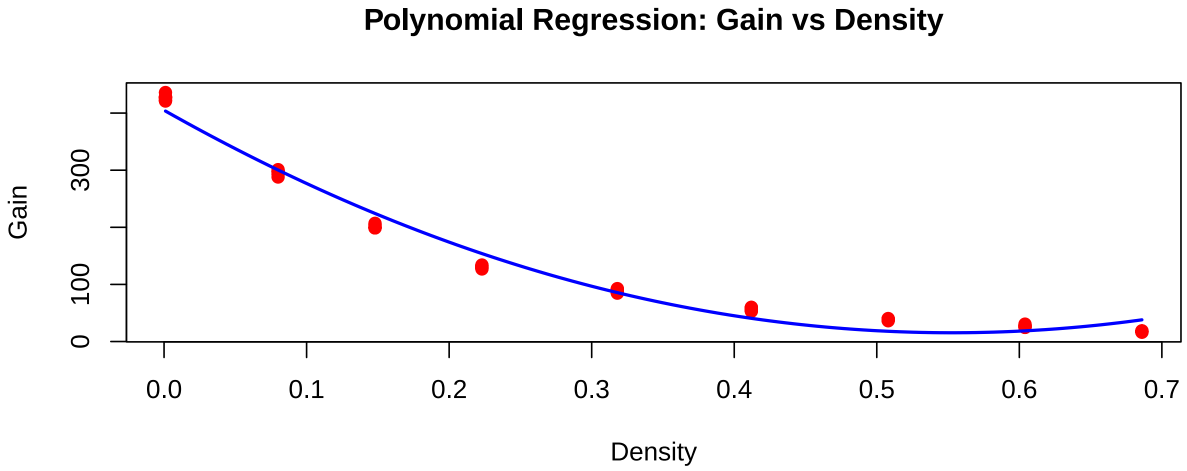
Conclusions

After performing cross-validation, we conclude that the regression model, after using the logarithm-transformed feature, generalizes well to unseen data, meaning it is modeling the underlying data-generating process. For both cross-validations where we removed densities 0.508 and 0.001 from the training sets, the model still made close predictions on the unseen data, with the true density being inside the interval estimates in both cases.

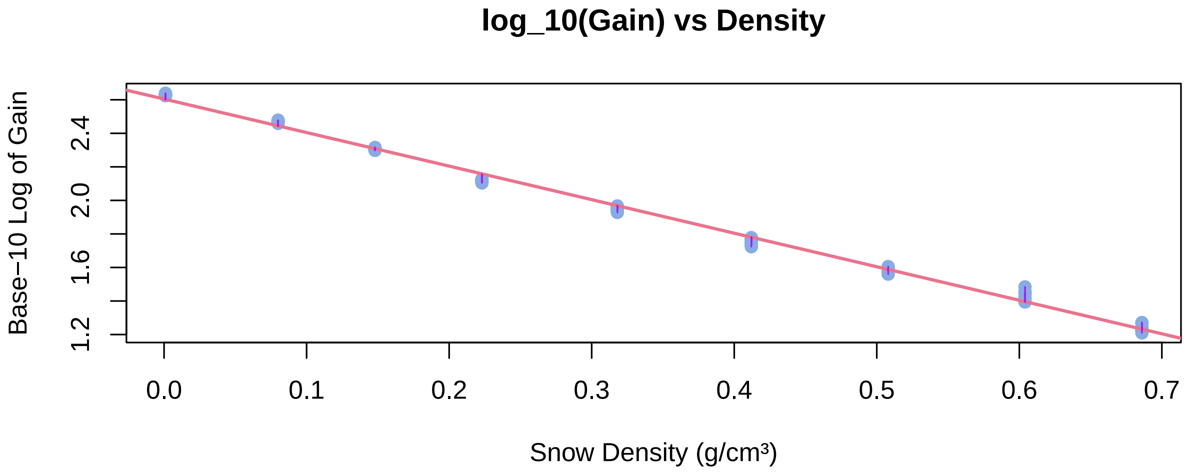
Advanced Analysis: Exploring Transformation Choices

Methods

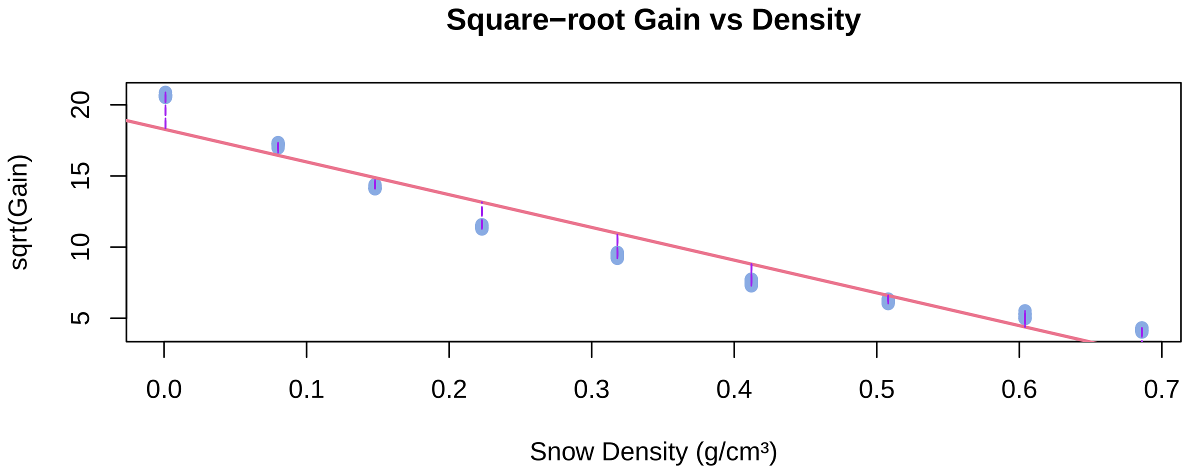
To compare our logarithm transformation to other possible transformations, we engineer polynomial features, another logarithm transformation but with a different base, and a square root feature. We train each of these 3 models and visualize the regression results.



Polynomial Regression: Gain vs Density



log_10(Gain) vs Density



Square−root Gain vs Density

Analysis

After fitting 3 models using polynomial regression, base-10 logarithm transformation, and square-root transformation, the shapes of the regressions in comparison to the actual data reveal that logarithm is still the best and most appropriate transformation.

The polynomial regression graph approximates the data well, but the quadratic starts to turn upwards at the higher density values even though the actual data doesn't do this, implying that this model will not extrapolate well to unseen density values because of an inappropriate choice of quadratic. The square-root transformation results in a regression with a clear pattern in the residuals. Data points at both the low and high ends of the density range have higher values than what is predicted by the regression. This parabolic shape in residuals shows that the square-root transformation is not appropriate. Finally, just based on visual inspection and mathematical knowledge, the base-10 logarithm transformation fits the data's shape better than the other two models. This makes sense because it is mathematically very similar to the transformation done in Question 2 (natural logarithm).

Conclusions

In conclusion, both polynomial and square-root transformations are not appropriate for modeling this data. When either of these transformations are applied before modeling, the residuals have an easily identifiable pattern, meaning they are not appropriate transformations. The logarithmic transformations, however, result in a pattern-less residual plot and are the most appropriate.

Conclusions and Discussion

Summary of Findings

Our findings suggest that there is a solid and reliable relationship between gamma measurements and the actual density of the materials presentw ithin our data, which were a gamma transmission snow gauge which was used to measure snow density. In our analysis, we find that there is a solid chance that the relationship between these two measurements can be represented as an exponential relationship, similar the the equation of exponential decay.

While our original data does not seem to fit a regression line well, we found that an exponential function does (or the equivalent since we transformed the data). Additionally, we tested the robustness of the density of the polyethylene blocks, while also using our original model to test how well the forward and reverse predictions of our gain and densities were, citing areas where the predictions were not as accurate compared to other areas. Additionally, in the analysis of those forward and reverse predictions, we theorized reasons why certain densities or gains were easier to predict, such as fundamental flaws with our equation, or patterns that we found interesting. Finally, we performed a cross-validation of this reverse prediction, providing an interval estimate for the density of our blocks.