1) Introduction

- The main source of water for Northern California comes from the Sierra Nevada Mountains. However, because rain and snowfall are different yearly, the Forest Service of the United States Department of Agriculture (USDA) operates a gamma transmission snow gauge in the Central Sierra Nevada, which helps determine a death profile of snow density to monitor the water supply generated from the mountains. The snow gauge does not disturb the snow in the measurement process, which means the same snow-pack can be measured repeatedly. When rain falls on snow the snow absorbs the water up to a certain point, after which flooding occurs. The denser the snowpack, the less water it can absorb. Analyzing the snowpack profile may help with monitoring the water supply and flood management. The data used in the case study are from a calibration run of the USDA's Forest Service, and polyethylene blocks are used to simulate snow. For each polyethylene block, 30 measurements are taken and only the middle 10 datasets are reported. The dataset consists of 10 measurements of each of the 9 densities in gram per cubic centimeter of polyethylene. The snowpack density measured typically ranges between 0.1 and 0.6 g/cm^3. This case study will provide a simple procedure for converting measured gain into predicted snow density when the gauge is in operation.

2) Analysis

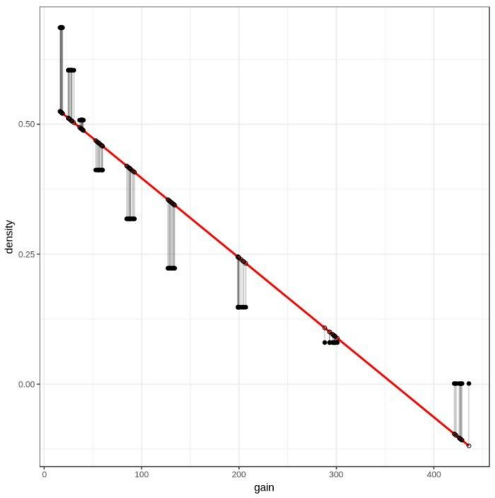
1. **[Fitting]**

   Methods:

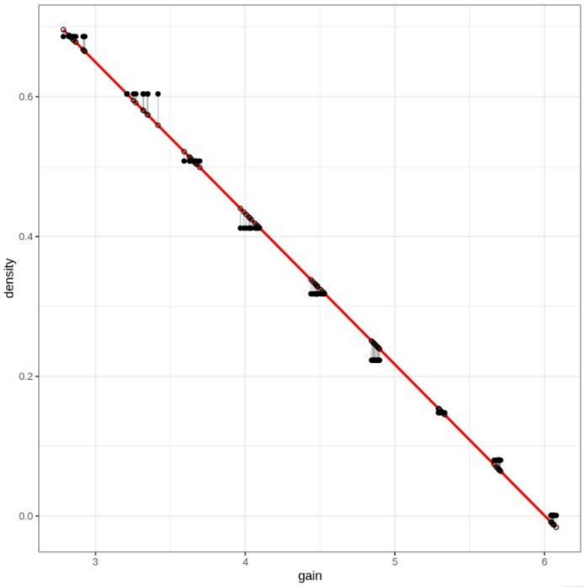
- First, we plotted a histogram for the gain column of our dataset, which is our explanatory variable. Then we used the built-in function "summary()" to receive statistical information about the residuals and correlations of our data. We then fit our data setting our x-axis as the "gain" and response variable y-axis as "density". We plotted the residuals of each data point with respect to the regression line, and also plotted a scatter plot that displays a straight horizontal line (using abline()) showing how much above or below the residuals are from 0.00. Histogram of the residuals as well as the normal Q-Q plot gave us information about the distribution and how far deviated the residuals are from our predicted regression line. After observing that our data is severely right-skewed, we applied the logarithm function to “gain” in order to alleviate the level of skewness. We then repeated the same process mentioned above by plotting the ggplot with our regression line and residuals, along with the scatter plot containing abline(), and the histogram showing the distribution of the residuals.

Analysis:

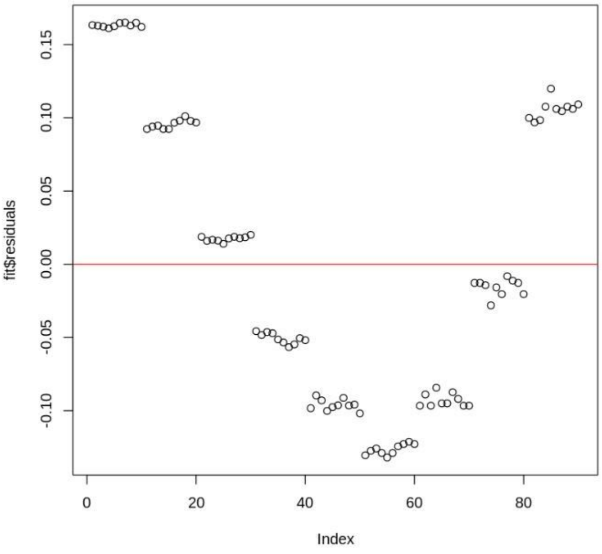
- The data of “gain” was generally right-skewed, so we applied the logarithmic function to fit “gain” in its form of a less skewed data. The histogram of the observed data when we applied the logarithmic function clearly showed that the data became less skewed, as the distribution was not perfect but close to normal distribution form. The residuals of the fitted “gain” was generally bigger in size (greater distance from the regression line) when we were dealing with skewed data than the less skewed data. The size of the residuals dropped significantly for the less skewed data, and the scatter plot showed how many of the residuals were very close to the line that stretched horizontally from 0.00. In other words, for our original data, the variability of the points around our least squares line is not quite constant, implying that the variability of residuals around the 0 line (horizontal) is not so constant. However, our log-fit data is close to achieving homoscedasticity because the variability of points around the least squares line is more constant than that of our original data.



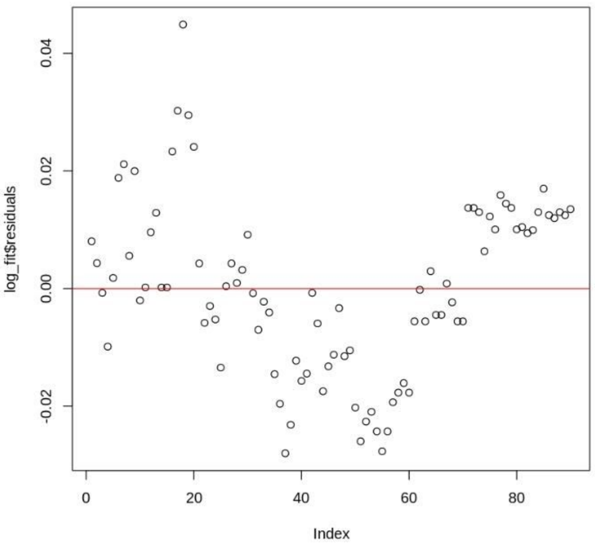
- ggplot of the regression line and residuals plotted (original, skewed data)



- ggplot with regression line and residuals plotted (log-fit, less skewed data)



- scatter plot with residuals displayed with respect to the horizontal abline (original, skewed data)



- scatter plot with residuals displayed with respect to the horizontal abline (log-fit, less skewed data)

- In terms of the histograms of the residuals (as we can see in *Appendix A* and *Appendix C*), we were able to notice that the residuals were close to being evenly distributed (except for only one bar) for the skewed data "gain", but having more spikes on logarithm-fit residual values from -0.02 to 0.02. This, again, visually showed that the size of the residuals was reduced. For the assessment of how well the data points have been fit to the regression line, there is a significant improvement from *Appendix B* to *Appendix D*. *Appendix B* shows how residuals are packed in groups but not quite aligned with the direction of the least squares line, but in contrast *Appendix D* shows how residuals are tightly aligned with the least squares line, although some exceptions do exist.

- A problem that might occur during fitting would be if the densities of the polyethylene blocks are not reported exactly. This problem might occur due to the extreme wear of equipment, or a recording error that might be produced when the polyethylene blocks are observed. Because the sample size of the data set given is very small, a wrongly reported value might significantly alter the predicted values because we may not be able to notice which value is an outlier, and which value is an accurate measurement. For the nine different densities, only 10 measurements were given, and they were all selected from the middle 10 measurements. Because the measurements were not selected at random, the linear regression model might produce inaccurate and invalid density prediction values. If different invalid measurements are recorded from the polyethylene blocks, these problems will have a significant effect on the linear regression model that will predict density given different gain values.

Conclusion:

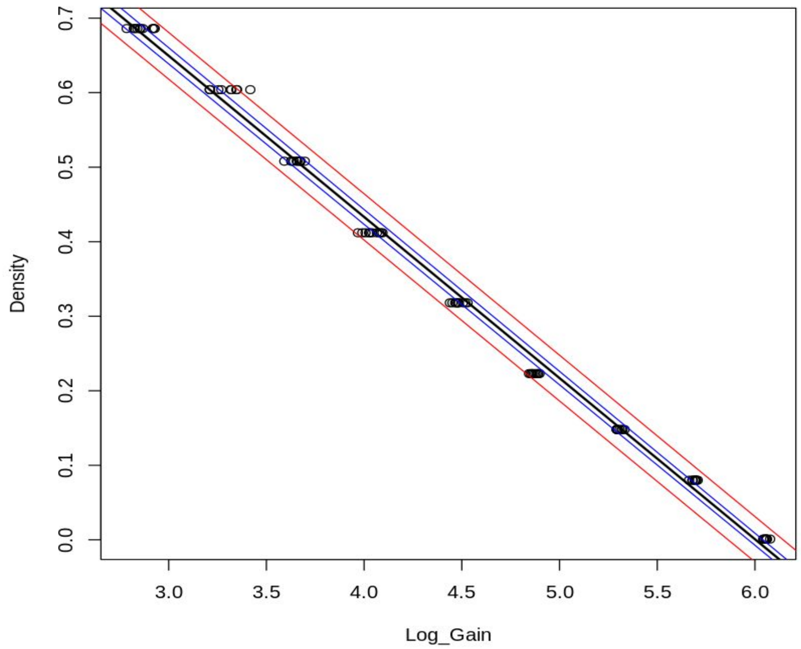
- In conclusion, we were able to achieve our goal of fitting the explanatory variable "gain" to response variable "density" on a linear regression line while minimizing the size of the residuals. Initially, we saw how the residuals raised some problems with the fit because the "gain" data was right-skewed (so much "gain" less than 100 as compared to greater than 100). However, we were able to resolve this problem by applying the logarithmic function on "gain" by reducing the residual standard error from 0.09769 to 0.01471 and increasing the R^2 value from 0.8157 to 0.9958 (proving that the data points are much closer to our fitted line).

2. **[Predicting]**

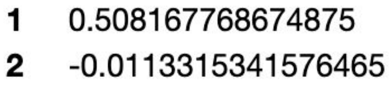
Methods:

- In order to predict density given specific gains of 38.6 and 426.7, we must use the linear regression model produced from the fitting step above. The specific gains were chosen, because they are the average gains for the 0.508 and 0.001 g/cm^3 densities. Using our linear regression model, our main objective is to make sure that the gains listed above produce densities of 0.508 and 0.001 g/cm^3. A fit plot was created, where the feature variable (x-axis) is the log values of gain, and prediction variable (y-axis) is the density values. Also, using the confidence and prediction intervals, and the linear regression model, we estimated the density of 38.6 and 426.7 respectively.

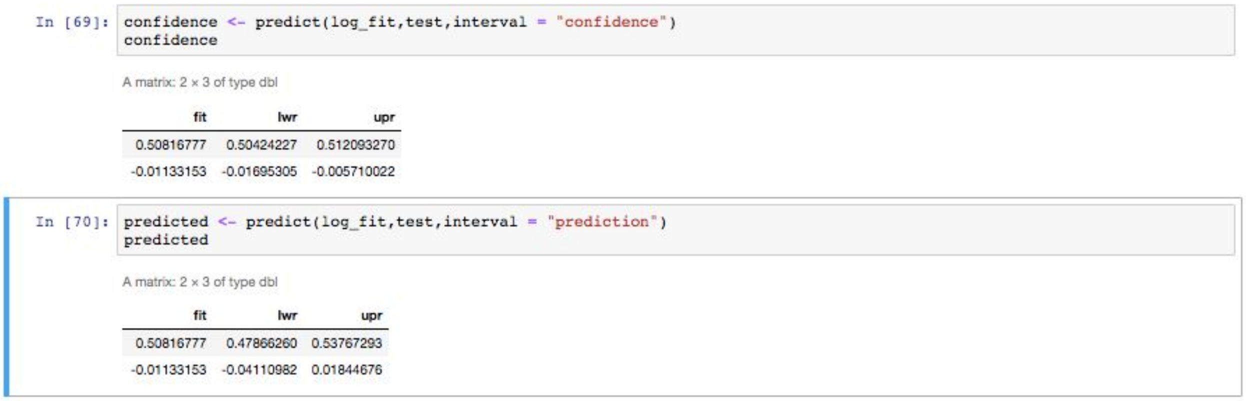
Analysis:



- The fit plot above has three distinctly colored lines, as well as the log density values plotted in between the three lines. The black line that the red and blue lines are centered around represent the linear regression line. The red line represents the 95% prediction interval and the blue line represents the 95% confidence interval. The prediction interval is important because it represents the prediction range of future individual density values given a logarithmic gain. The confidence interval is important because it represents the prediction range of the mean density values given a logarithmic gain. We must take into account that the prediction interval is different from the confidence interval, and is much wider than the confidence interval due to the uncertainties involved in predicting an individual density value, rather than the mean density value. We can observe this exact phenomenon in the fit graph shown above, where the confidence interval is very close to the linear regression line, whereas the prediction interval is much further away from the linear regression line.



```
1    0.508167768674875
2    -0.0113315341576465
```



```
In [69]: confidence <- predict(log_fit,test,interval = "confidence")
         confidence
```

A matrix: 2 × 3 of type dbl

fit	lwr	upr
0.50816777	0.50424227	0.512093270
-0.01133153	-0.01695305	-0.005710022

```
In [70]: predicted <- predict(log_fit,test,interval = "prediction")
         predicted
```

A matrix: 2 × 3 of type dbl

fit	lwr	upr
0.50816777	0.47866260	0.53767293
-0.01133153	-0.04110982	0.01844676

- By using the linear regression model that fit the logarithmic gain, we were able to calculate the predicted density values for the gains of 38.6 and 426.7. For the gain value of 38.6, the linear regression model predicted 0.50817 g/cm^3. This value is almost identical to the actual value of 0.508 g/cm^3, and we found that the 95% confidence interval for the predicted value was (0.50424227, 0.512093270). The 95% prediction interval for the value was (0.47866260, 0.53767293), which was a lot broader than the confidence interval. Our predicted value falls between the 95% confidence interval and the 95% prediction interval. For the gain value of 426.7, the linear regression model predicted -0.01133 g/cm^3, which is almost identical to the actual value of 0.001 g/cm^3. The 95% confidence interval for this gain value was (-0.01695305, -0.005710022). The 95% prediction interval for the value was (-0.04110982, 0.001844676), which was again a lot more broader than the confidence interval values. Through the calculated intervals, we can see that the predicted value also falls between the 95% confidence interval and the 95% prediction interval. However, the linear regression model predicted a negative value, and we believe that this occurred because the actual density value of 0.001 g/cm^3 is very close to zero, therefore the model may have predicted a negative value.

Conclusion:

- From the fit graph and actual calculations of the predicted densities given the gain values of 38.6 and 426.7, we have found that our linear regression model predicted values very similar to the actual density values, and that the predicted values were in between the confidence and prediction intervals. This shows that the linear regression model fitted with the logarithmic gain values estimated the densities pretty accurately, and that the randomness could also be predicted by the linear regression model, because its values came in between the confidence and prediction intervals. Also stated in the introduction, we said that the prediction interval would be broader than the confidence interval, because the predicting individual future density values would contain more uncertainty than predicting the mean density values.

3. **[Random Forest] (Advanced Analysis)**

Method:

- Our data number is only 90, and it could be considered as a small dataset**.** Thus, we used the Random Forest, which is based on a bagging algorithm, to increase accuracy of our model. The reason why we choose this model is that Random Forest creates many trees and makes a result by finding the average of all results from the trees. It will overcome the disadvantages of our dataset, which is a small number of the dataset. Furthermore, our data after transforming to the logarithmic gain does not contain a significant outlier. Thus, this is a good dataset to apply Random Forest. We use RMSE, R^2 and predicted values to compare each model.

Analysis:

- First, our R^2 increases from 0.9958 to 0.9999. It means that our dataset is almost the same with our regression line. The Random Forest RMSE remains the same with linear regression. These R^2 and RMSE are based on predictions from our original dataset, so we worried that our new model is overfitted even though R^2 increases. Therefore, we test by the given test set(38.6,426.7) to compare with the above linear regression. The Random

Forest predicts 0.508000000000004 and 0.001. The density predictions are the same as the given average gains. When we compare with the original linear regression results(0.50816, -0.01133), our Random Forest model is a perfect model to predict densities.

Conclusion:

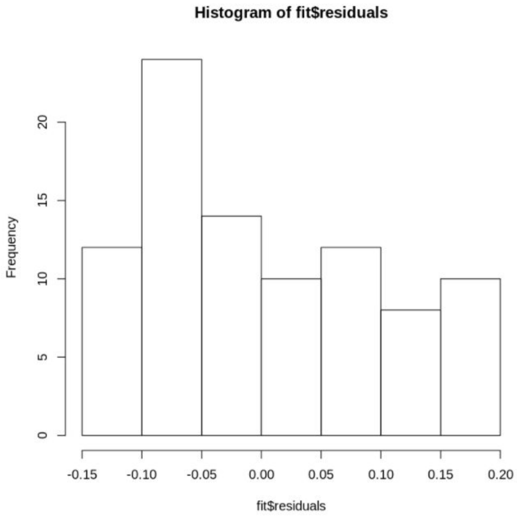
- In conclusion, our dataset is quite small data to predict density, so we have improved our model by using the Random Forest Algorithm. As a result, we got higher R^2. In addition, our predicted values are the same with given average densities. Therefore, we conclude that our new model is a better prediction model.

3) Report Conclusion

- In order to convert gain into a measure of density, we needed to fit the given gain dataset into the linear regression model. However, through different graphical analysis, we found that the gain dataset had right skewness shown in a histogram formed above. To remove this apparent skewness as much as possible, we converted the gain column to a logarithmic dataset to normalize the data. This rescaling proved to be efficient, because it normalized the gain column shown in the histogram above. This rescaling converted the residual scatter plot drawn from heteroscedastic to homoscedastic, and also when fitting this logarithmic gain dataset into the linear regression model, the R^2 value increased. After fitting the linear regression model with the logarithmic dataset, we created a fit plot that contained the linear regression line, confidence interval, and prediction interval. The confidence and prediction interval displayed the 95% interval, which is important because the prediction interval represents the prediction range of future individual density values given the gain and the confidence interval represents the prediction range of the mean density values given the gain. The prediction interval is much broader than the confidence interval, because it contains more uncertainties to predict the range of an individual density value, rather than the mean density value. Our linear regression model predicted the density values closed to the actual density values, and the predicted values came in between the confidence and prediction intervals shown in the fit plot. Also, because the given dataset was quite small, we wanted to improve our linear regression model, so we used a random forest regressor to predict the density values. By using the random forest regressor, we received a higher R² value, and our predicted values were identical with the given average densities. Through these different graphical and predicted analyses, we can conclude that our linear regression model can predict density from logarithmic gain accurately, and therefore predict the water supply from the Sierra Nevada Mountains in Northern California.

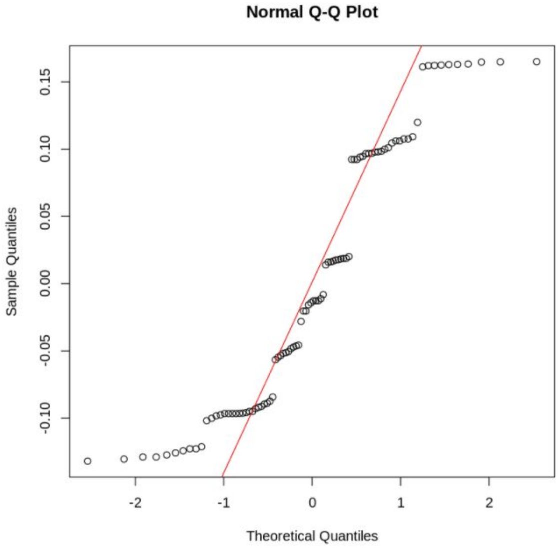
4) Appendix

- **Appendix A**



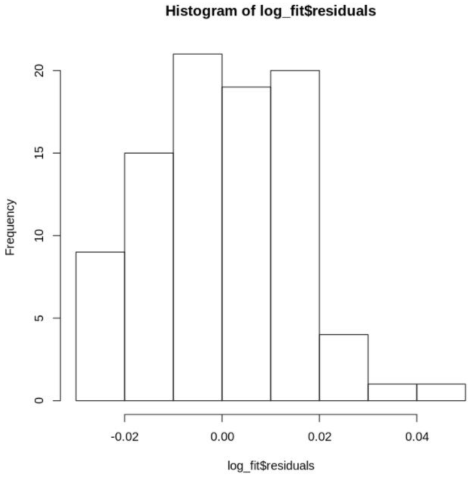
- Histogram of Residuals with original gain.

- **Appendix B**



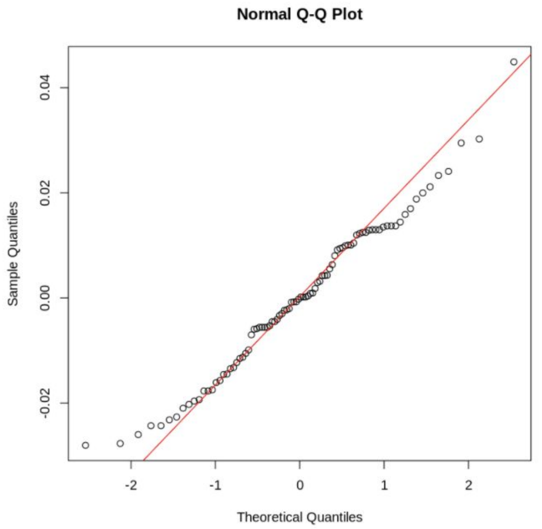
- Q-Q plot of the Residuals with original gain.

- **Appendix C**



- Histogram of residual with log_gain

- **Appendix D**



- Q-Q plot of residual with log_gain

5) Author Contribution Statement

- Youngseo Do developed the analytical reasoning for the “fitting” part, indicating the problem with the initial fit and explaining how log-fit data improved the fitness of the data, as well as comparing between visualized data (ggplot, scatter-plot) of original and log-fit “gain”.
- Seokmin Hong developed the analytical reasoning for the “predicting” part, as well as the fit plot containing the confidence interval, prediction interval, and linear regression line. Also, he calculated the confidence and prediction intervals that were used to check if the predicted density values were in between the intervals.
- Yeongjae Kim developed the code overall, and designed Random Forest Algorithms to improve our linear model . Also, he wrote an analysis of why our team chose Random Forest and explanations of the results of the Random Forest model.
- Youngseo Do edited grammar and code throughout the code part of the homework as well as the written report.
- Seokmin Hong and Yeongjae Kim added to the appendix where statistical information was useful, and also wrote the introduction and final conclusions of the paper.