Calibrating a Snow Gauge

Math 189 - Investigation #4

Professor Schwartzman

Spring 2020

Arthur Chang - A14410373

Raya Kavosh - A14826756

Siddharth Saha - A15572442

Contributions

Raya - Formulated research questions and methodology based on background information and data. Aided in analysis and interpretation of results. Structured introduction and its sub-discussions as well as conclusion.
Arthur - Performed initial fit of the data and determined if transformation of the dataset was needed. Transformed the dataset with log and provided reasoning behind the transformation.
Siddharth - Took fitted results and carried out entire prediction analysis which includes the calculation of prediction intervals, interpretation of the fit, study into the distribution of the gain values at specific densities, researched methods to tackle said distribution and executed BCA confidence interval for giving confidence to the estimates we elected to use

Introduction

The main source of water for Northern California comes from the Sierra Nevada mountains. To help monitor the water supply, the Forest Service of the United States Department of Agriculture (USDA) operates a gamma transmission snow gauge in the Central Sierra Nevada near Soda Springs, CA to determine a death profile of snow density. The snow gauge does not disturb the snow in the measurement process, which means the same snow-pack can be measured over and over again. With replicate measurements on the same volume of snow, researchers can study snow-pack settlement over the course of the winter season and the dynamics of rain on snow. When rain falls on snow, it can absorb it up to a certain point before flooding. This point depends on the density of the snow- the more densely it is packed, the less water it can absorb. Analysis of the snowpack profile may help with monitoring the water supply and flood management.

The gauge measures the snow density through a conversion of measurements of gamma ray emissions. Instrument wear and radioactive source decay may affect the functions used to convert the measured values into density readings over time. To adjust the conversion method, a calibration run is made each year at the beginning of the winter season. In this investigation, we will develop a procedure to calibrate the snow gauge.

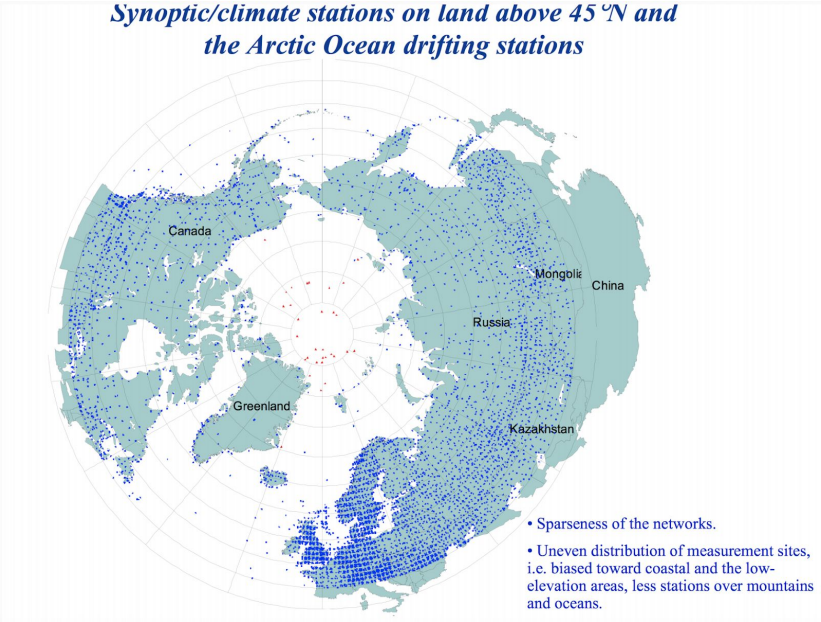
Data

The data we will use are from a calibration run of the USDA Forest Service's snow gauge located in the Central Sierra Nevada mountain range near Soda Springs. The run consists of placing polyethylene blocks- used to simulate snow- of known densities between the two poles of the snow gauge and taking readings on the blocks. 30 measurements are taken per polyethylene block for each of 9 densities in grams per cubic centimeter of polyethylene. Of those measurements, the middle 10 are reported in our data. The measurements reported are amplified versions of the gamma photon count made by the detector. We call the gauge measurement the "gain".

Data Limitations

Some challenges related to the data generating process include the decline of operational networks in the northern regions, including Siberia, Alaska, and Northern Canada, with few stations in the mountain regions. This presents the problem of sustaining and improving data quality and

compatibility across national boundaries. The inconsistencies of operational networks can be observed in the figure below.



Other inconsistencies include large biases in gauge measurements of solid precipitation and the incompatibility of precipitation data due to differences in instruments and methods of data processing. It is difficult to determine precipitation changes in the arctic regions, and to validate precipitation data, including satellite and reanalysis products and fused products at high latitudes.

Background

The snow gauge is a complex and expensive instrument. It is not feasible to establish a broad network of gauges in the watershed area in order to monitor the water supply. Instead, the gauge is primarily used as a research tool to help study snow-pack settling, snow-melt runoff, avalanches and rain-on-snow dynamics.

The gauge in California is located in the center of a forest opening that is roughly 62 meters in diameter. The laboratory site is at 2099 meters elevation and is subject to major high altitude storms, which regularly deposit 5-20 centimeters of wet snow. The snowpack reaches an average depth of 4 meters each winter.

The snow gauge consists of a cesium-137 radioactive source and an energy detector mounted on separate vertical poles approximately 70 centimeters apart. At the top of the poles, the lift mechanism raises and lowers the source and detector together. The radioactive source emits gamma photons, also called gamma rays, at 662 kilo-electron-volts (keV) in all directions. The detector contains a scintillation crystal which counts those photons, eating through the 70-cm gap from the source to the detector crystal. The pulses generated by the photons that reach the deter crystal are transmitted by a cable to a preamplifier and then further amplified and transmitted via a buried coaxial cable to the lab. There, the signal is stabilized, corrected for temperature drift, and converted to a measurement of "gain" that should be directly proportional to the emission rate. The snowpack density typically ranges between 0.1 and 0.6 g/cm3. The gamma rays that are sent in the direction of the detector may be scattered or absorbed by the polyethylene molecules between the source and the detector. With denser polyethylene, fewer gamma rays will reach the detector.

There are complex physical models for the relationship between the polyethylene density and the detector readings. A simplified version of the model that may be workable for the calibration problem of interest is described here. A gamma ray on route to the detector passes a number of polyethylene molecules, depending on the density. A molecule may either absorb the gamma photon, bounce it out of the path to the detector, or allow it to pass. If each molecule acts independently, then the chance that a gamma ray will successfully arrive at the detector is (pm), where p is the chance that a single molecule will neither absorb nor bounce the gamma ray, and m is the number of molecules in a straight line from the source to the detector. This probability can be re-expressed as $e^{m\,log\,(p)} = e^{bx}$, where x (the density), is proportional to m (the number of molecules).

Research Questions

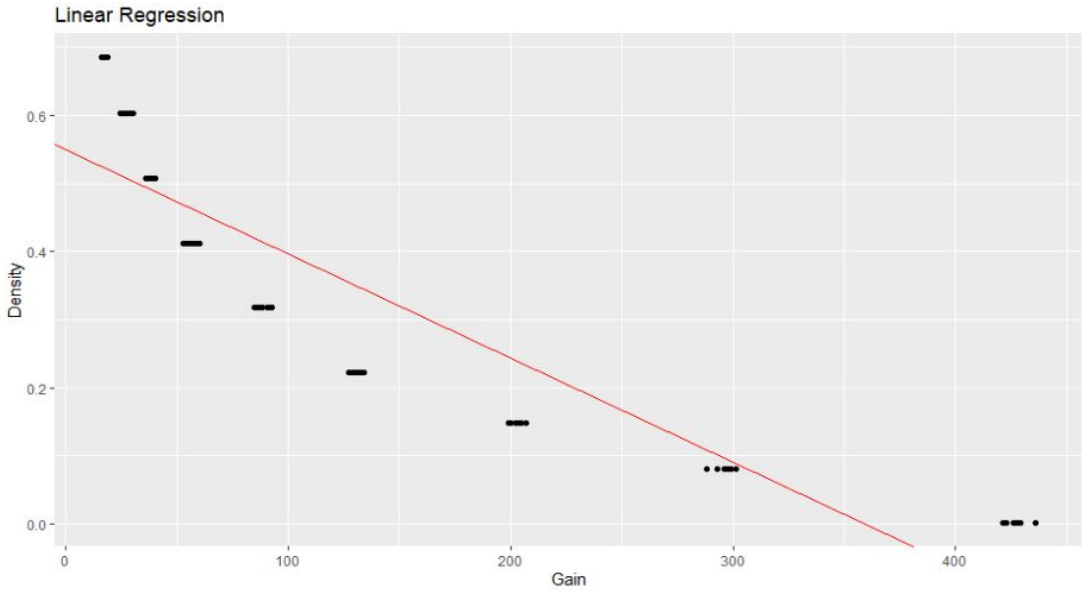
The aim of our investigation is to provide a simple procedure for converting gain into density when the gauge is in operation by varying density and measuring the response in gain. It should be noted that when the gauge is ultimately in use, the snow-pack density is to be estimated from the measured gain. We will use the data to fit the gain to the density, evaluating its fit by the residuals to the least squares line. We then discuss how the inaccurate reporting of polyethylene block densities may affect the fit of the model and its predictions. We then calculate the predicted density of snow packs with gain readings 38.6 and 426.7, the average gains for the 0.508 and 0.001 densities, respectively. This will allow us to develop a procedure for adding confidence intervals around our least squares line that can be used to make interval estimates for snowpack density from gain measurements.

Analysis

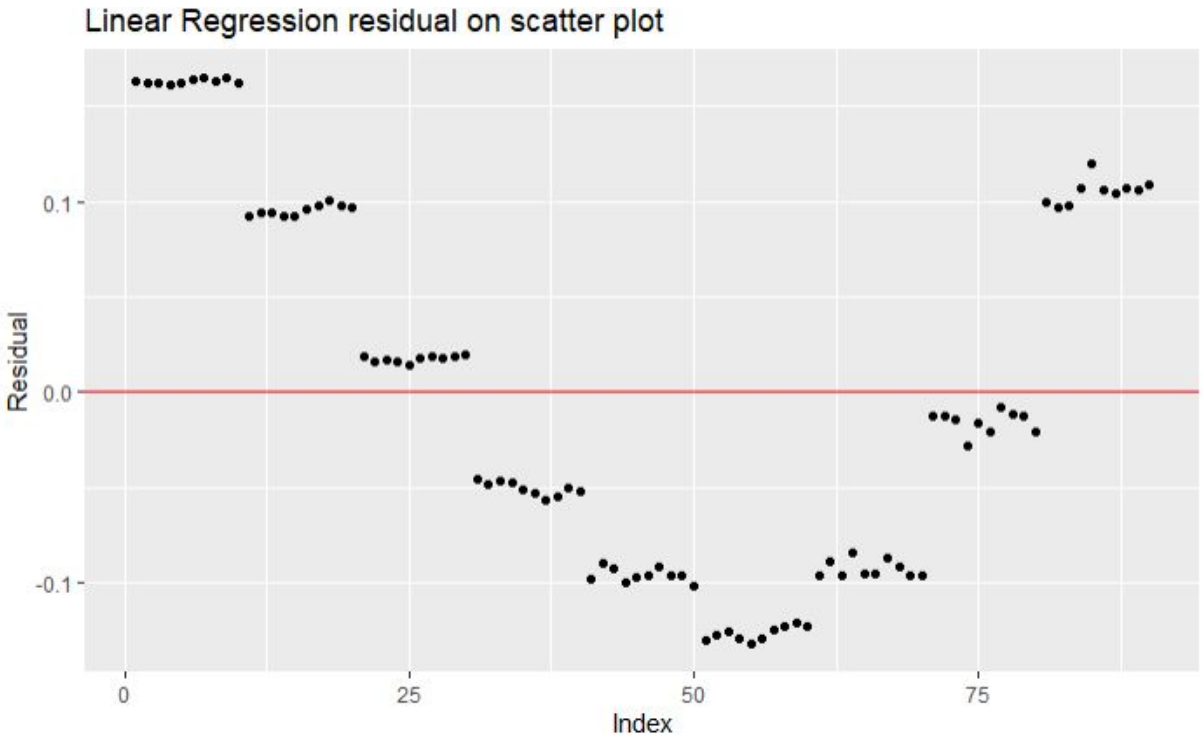
Fitting

* Before transformation

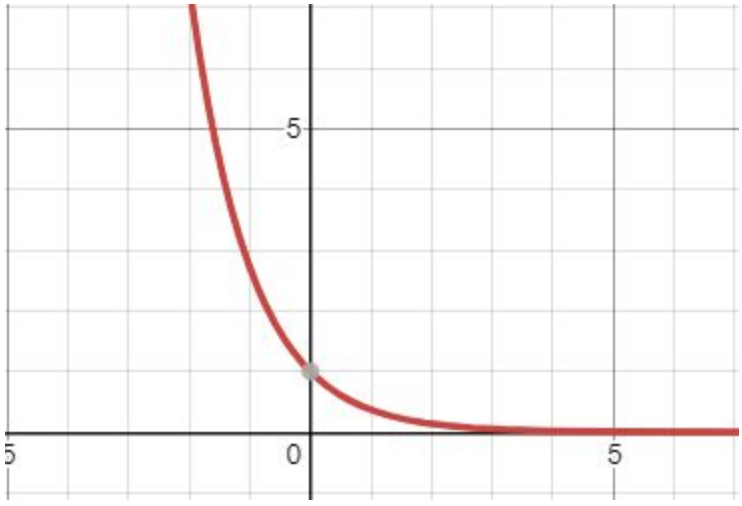
We first take the Linear Regression between gain and density without any transformation on the data as the initial visualization below (Linear Regression).



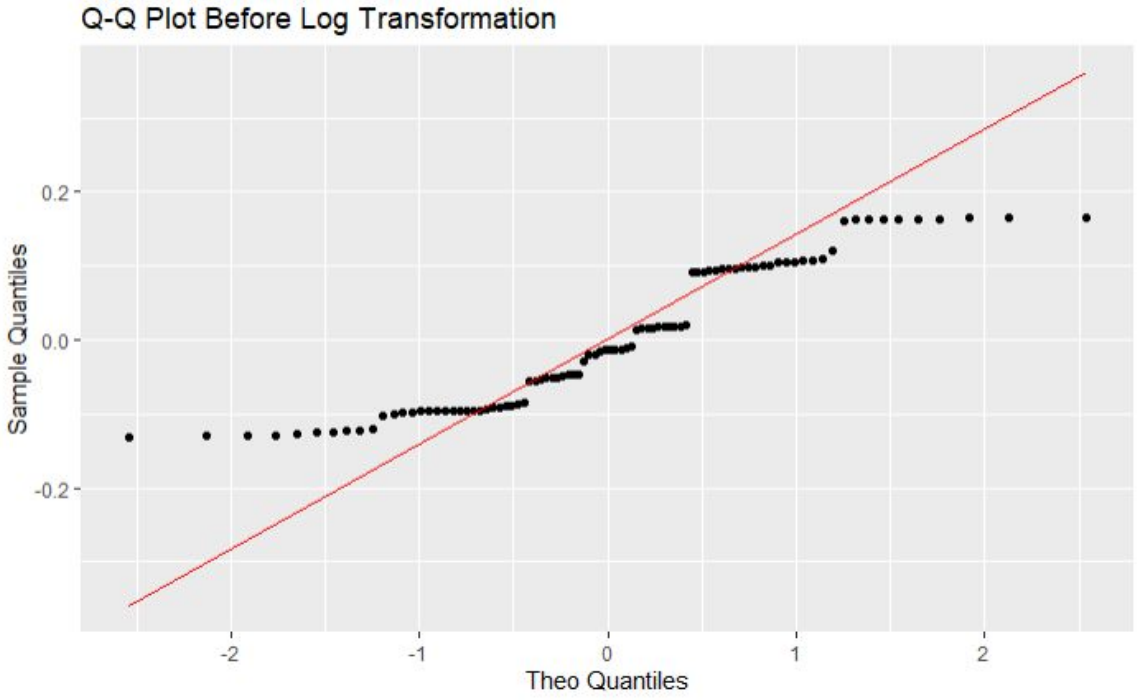
From the residual plot, we can see that the residuals do not follow a typical spread, but rather a noticeable pattern.



The residual follows a pattern and does not seem to fit well to the linear regression line before the transformation. We observed that the data points are non-linear, as the pattern seems to follow an e^{-x} regression as shown below. Instead, we decided to use log transformation to linearize the data.



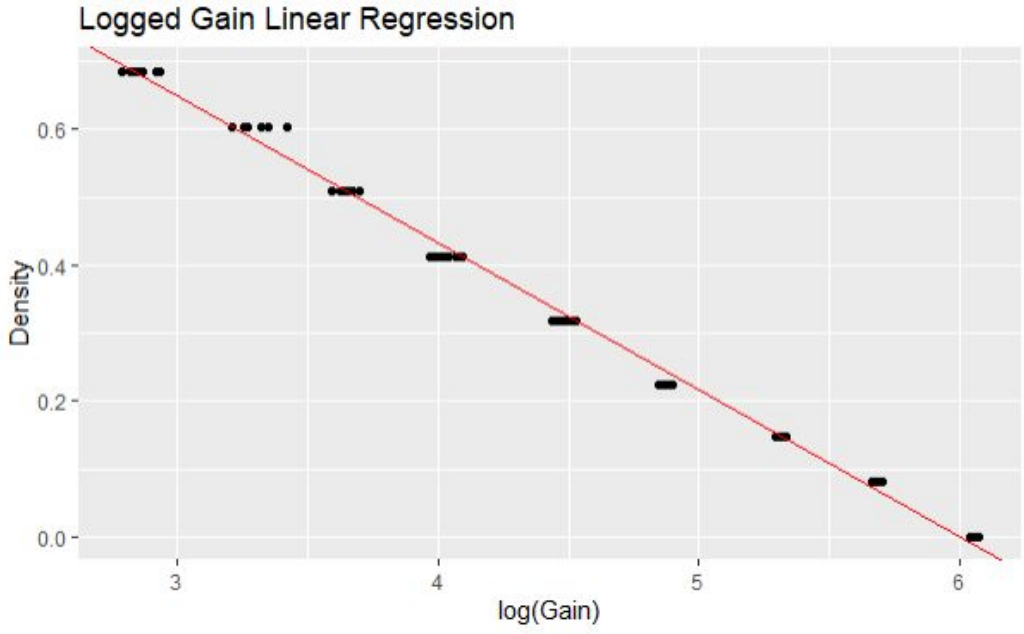
The QQ plot below for the dataset before transformation shows that linear regression is not a good fit. The shape of the data points are significantly flatter than the regression line.



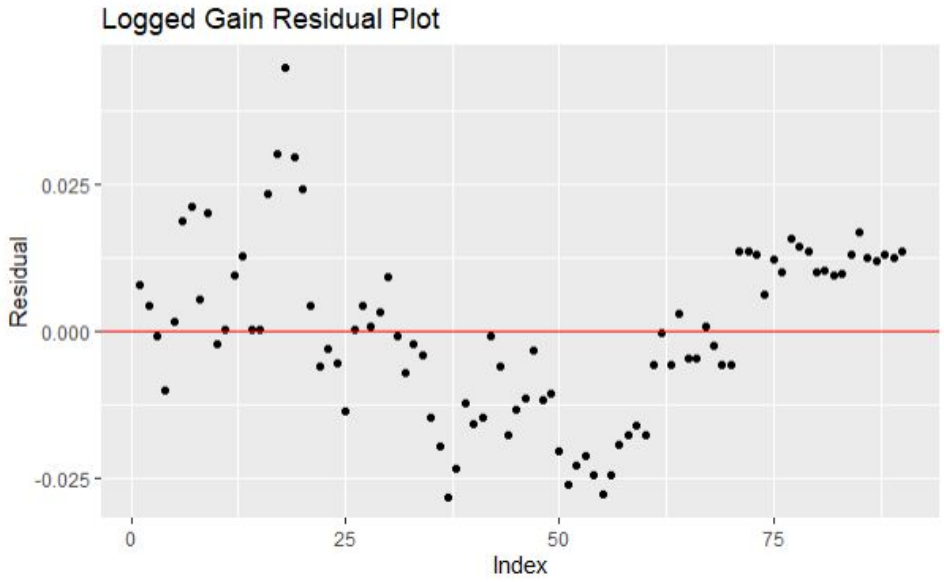
The densities reported have differences below 0.0001. The linear regression line would not be affected much by these minor errors before the transformation.

*After transformation

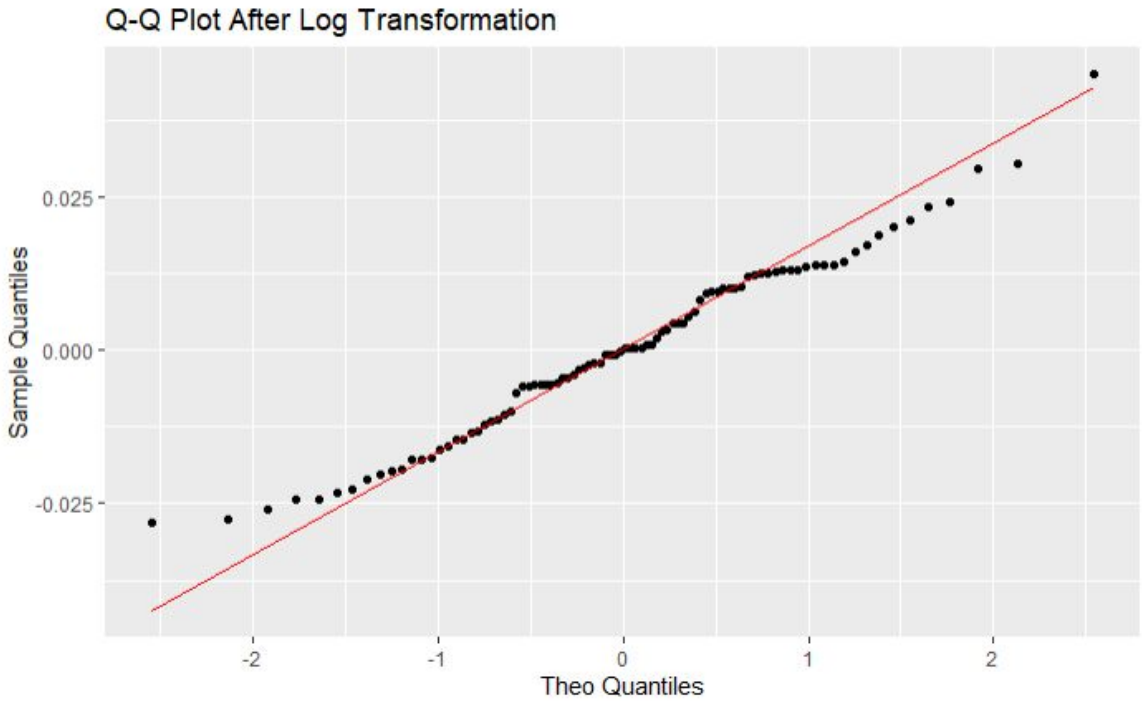
After the log transformation on the data below, the data points adhere more closely to the linear regression line.



The residual distribution appearing below is much more balanced on either side of the line after the log transformation.



The QQ plot below for the dataset after transformation shows that linear regression is a better fit for the data than before the log transformation. The shape of the data points are more representative of the linear regression line in place.

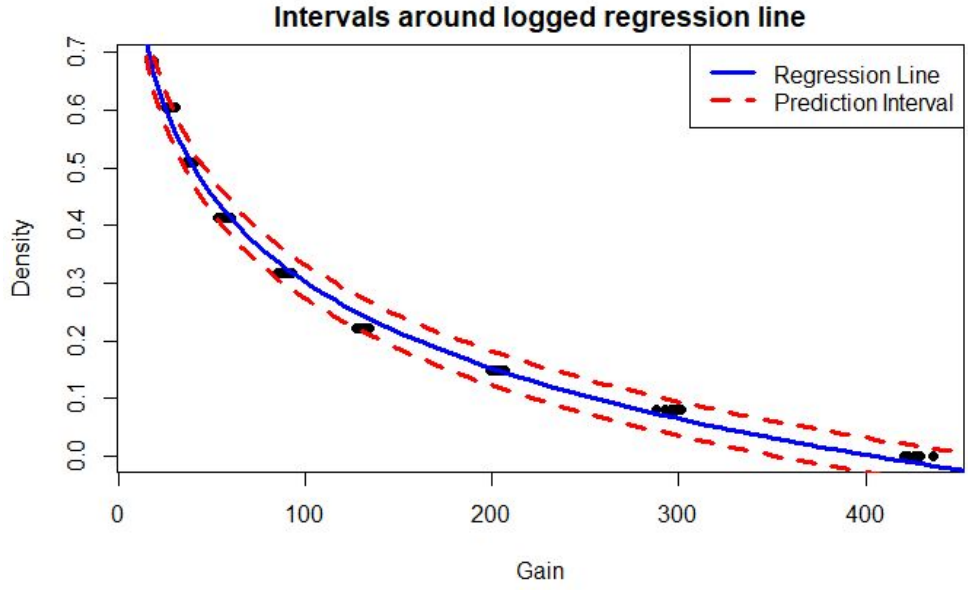


The densities reported have differences that are under 0.0001. The linear regression line would not be affected much by these minor errors after the transformation.

Prediction

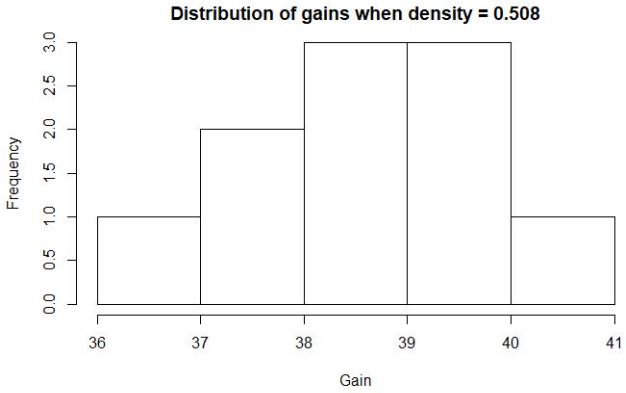
After running analysis on various fitting procedures we end up choosing the log scaled linear regression model as it better fits the data. The intercept of our least squares line ends up being 1.298 while our slope is -0.2162. This means our line of best fit is density = 1.298 - 0.2162 * log(gain). To carry out our prediction analysis we use the gain values of 38.6 and 426.7. These values are chosen because we already know that they are the average gains for the densities 0.508 and 0.001. So when plugging these gain values into our model, we should get densities somewhat close to 0.508 and 0.001.

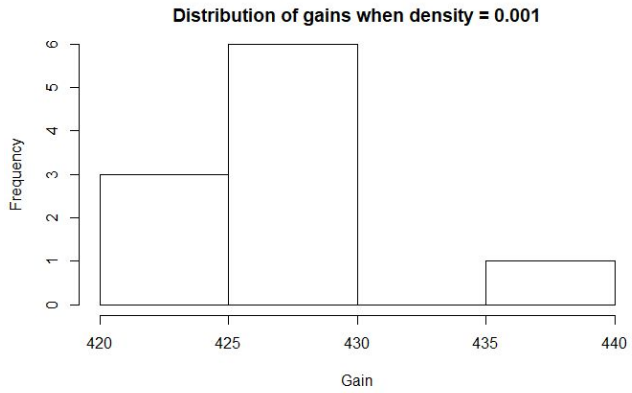
Upon plugging the gain values in, we get a density value of 0.508167 and -0.01133153. We notice that in our predictions, the predicted value for a gain of 38.6 is extremely similar to the actual density we obtain. However, in the case of a gain value of 426.7, we get a negative value. At first glance, this may seem like a big issue. This is because our linear regression model doesn't have a lower bound. It is just a straight linear line that takes whatever value is inputted and finds the input's location on the line. This can be rectified by keeping a lower bound of 0, thus keeping our prediction extremely similar to the actual value. In order to gain further confidence we plot the 95% prediction intervals of our line



The prediction intervals allow us to get confidence with respect to our estimates. We thus learn that in 95% of our estimates, the true density value is between [0.48, 0.54] when we use a gain value of 38.6. For a gain value of 426.7, we get that the true density value is between [-0.04, 0.18]. However, since we can't get negative densities, we lower bound it at 0 thus giving us the prediction interval [0, 0.18]. Our estimates lie within this interval, thus giving us more confidence in our predictions.

Now, we need to acknowledge an important fact. The average gain for densities 0.508 and 0.001 is 38.6 and 426.7. The average implies we have several values for a given density. How do we know how reliable these estimates are? To get an understanding we first plot the distribution of gain values at the densities .508 and .0001





As seen from the 2 distributions they aren't exactly normal. While the left graph has some form of normal shape, it is left skewed as indicated by the higher frequencies at the right. For the right graph there is no semblance of normality. Moreover, we have < 30 gains for a given density value which may make our data biased. Hence, to gain confidence in the average gain value of 38.6 and 426.7, we look to use the bootstrap and BCA method of calculating confidence intervals. This adjusts for bias and can handle non normal data quite well.

With a bootstrap sample of size 1000, we get the confidence interval of [424.8, 429.8] for density = 0.001 and [37.76, 39.14] for density = 0.508. Since our average estimates of 426.7 and 38.6 lie in these intervals, we can be confident that 38.6 and 426.7 are good estimates, and thus good values to use to test our linear regression model for specific densities.

Conclusion

Through our exploration of the effects of transforming the gain data in order to fit it to a least squares line that can accurately predict snow density, we found that the log scaled linear regression model yields the best fit to the data and the most accurate predictions. We found through this model that the line of best fit is *density = 1.298 - 0.2162 * log(gain)*. Using this equation, we found the estimates for the densities corresponding to the average gains. We then validated the fitted model's predictions by using bootstrap and BCA to calculate confidence intervals of density per gain measurement. After using this procedure to adjust for any non-normality of the data, our predicted densities were still in the bootstrap sample's generated confidence intervals. For this reason, we believe our estimates to be valid and our least squares regression line *density = 1.298 - 0.2162 * log(gain)* to be a valid procedure for making interval estimates for snowpack density from gain measurements.

Appendix

BCA Intervals

BCA intervals is a method that is used often in calculating confidence intervals when we have skewed and normal data. It's main advantage over confidence intervals is that it does not need nor rely on the normality assumption that regular confidence intervals use. How it works is that it first computes a regular confidence interval under normality assumptions. After that it transforms the intervals with the help of a bias correction coefficient and acceleration constant which it estimates by comparing the bootstrap samples with a test statistic (we use the mean of samples as our statistic). If the majority of samples are less than the statistic then the CI calculated under normality assumption shifts left. If they are greater than the statistic, the CI shifts right.