Snow Gauge Calibration Procedure

Author Contributions:
James Lu: R code, writeups / plot for questions 1 - 4, conclusion for questions 1 - 4, discussion
Matin Ghaffari: R code, Introduction, writeups / plot for questions 5 - 6, advanced analysis, discussion, and conclusion for questions 5 - 6

Introduction

In our analysis we will be examining the data from the calibration of a snow gauge, which provides snowpack profile information. This snowpack profile information is helpful in the area of conservation, particularly for the purposes of monitoring water supply, managing floods, and studying climate change. The snow gauge instrument allows us to indirectly measure snow density through gamma ray emissions that are used to calculate the density through a conversion method. Additionally, this conversion method has conversion factors that may potentially change over time due to various reasons such as instrument wear and physical configurations. Therefore, because of the wide variety of potential factors that may change during any given time, our analysis of this snow gauge data is intended to develop a procedure to calibrate the snow gauge for converting gain into density when the gauge is in operation.

Our Analysis uses a dataset "gauge.txt" that has observations from a single calibration run which uses polyethylene blocks to simulate snow for 9 various densities and with 10 measurements for each block. Then these blocks that are placed between the two poles of the snow gauge have gamma rays transmitted through them in order to measure the intensity of the photons detected after passing the blocks. As a result, the dataset has two columns for the continuous numerical variables of density and gain, where the density corresponds to densities of the polyethylene blocks in cubic centimeters while gain corresponds to the gauge's measurement of amplified versions of gamma photon counts made by the detector. These 90 rows of observations are recorded by the Forest Service of the USDA who collected these values from a single calibration run at the beginning of the winter season from a gamma transmission snow gauge in the Central Sierra Nevada near Soda Springs, California.

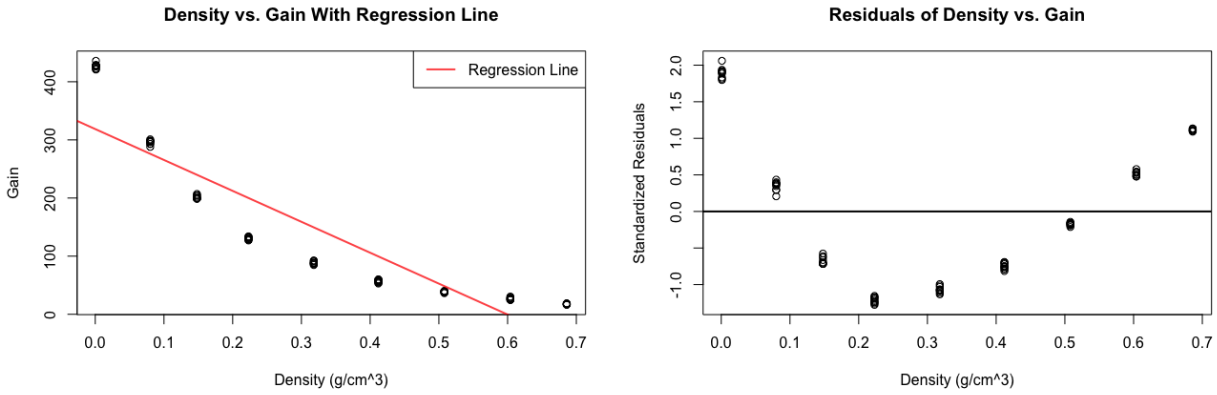
Furthermore, since the gamma rays that are emitted from the radioactive source may be scattered or absorbed by the polyethylene molecules between the source and the detector, we investigate how denser blocks allow less gamma rays to be detected. Subsequently, the gamma ray measurement decays exponentially with density, and we may solve for approximate exponential decay expression coefficients for our calibration. Ultimately, in our analysis we fit a regression line and transform the data to reasonably produce point estimates and confidence bands for predicting the gain as a function of measured density. We then use these for forward predicting with our fitted model in order to predict the gain using the original scale as a function

of the measured density, in addition to a reverse prediction where we map the measured gain to what the density would have been, which we also cross validate. As a result, our analysis helps to provide a procedure for converting gain into density when the gauge is in operation.

Primitive Regression

To begin our analysis we attempt to fit a regression line to a scatterplot of density vs gain in order to examine the fit of our model. To do this we used a least squares regression line created by minimizing the sum of squared residuals.

Figure 1: First plot is a scatter plot of density vs gain with a least squares regression line in red. Second plot is of the standardized residuals of the observed values to our regression line.



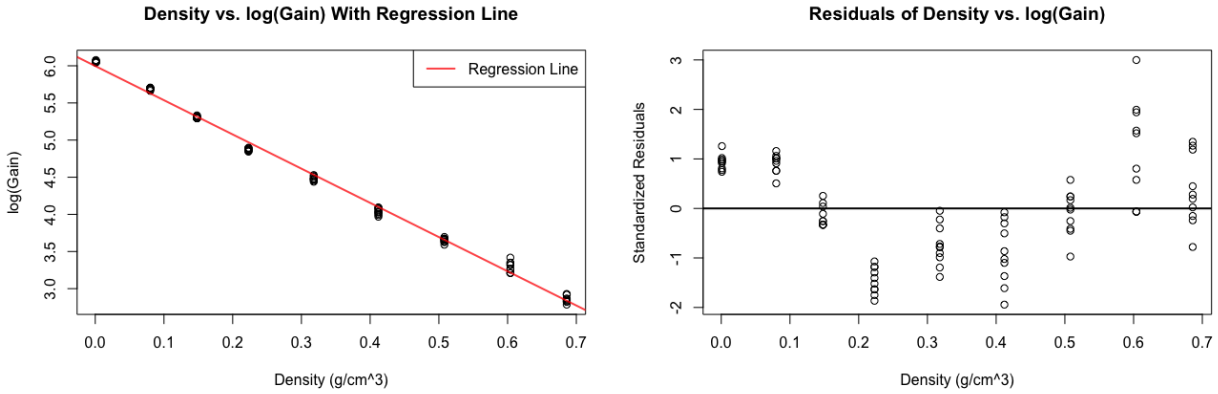
From our generated plots we can see that our line is not a good fit to the data as it violates the first condition of a least squares regression line which is linearity. It is clear in both the density vs gain scatter plot and the standardized residuals plot that the data is non-linear in its current state and a least squares regression line would not be suitable for this application. However, we can perform a transformation on the data to attempt to make it linear and more suitable for a least squares regression line.

Linearizing The Data

To continue with our analysis we try to find a suitable transformation for our given dataset in order to be able to utilize a least squares regression line. We can begin this process by examining the equation for the gamma ray measurement, $g = Ae^{\beta d}$ where A > 0 and $\beta < 0$ are unknown coefficients and d is the density. To linearize this equation with respect to density we can take the log of both sides giving us $log(g) = log(A) + \beta d$ and if we let $Y = log(g)$ and $X = d$ we have $Y = \beta_0 + \beta x + err$. After performing this transformation on our

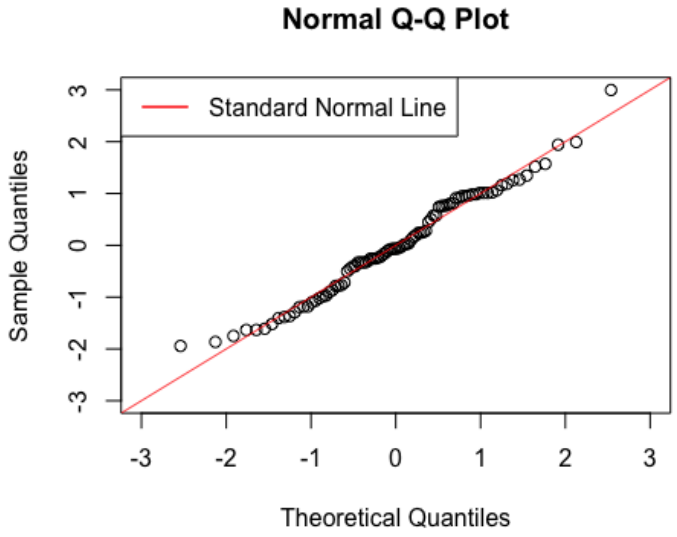
gain, we generated another scatterplot of density vs gain along with it's residuals to examine if least squares is suitable for this dataset.

Figure 2: First plot is a scatter plot of density vs log(gain) with a least squares regression line in red. Second plot is of the standardized residuals of the observed values to our regression line.



After examining the scatter plot in Figure 2 we found that our transformation did indeed make the data set more linear, so to further ensure that our least squares regression line is suitable we need to ensure that the variability in the data is relatively constant. To do this we look at the second plot of the standardized residuals where empirically it appears that variability is not constant, however we can justify our model theoretically. Our justification for not meeting the assumption of constant variability comes from the fact that our gain function is exponential, so theoretically our chosen log transformation is ideal for linearly modeling this data set. Finally we examine the normality of the residuals to see if there are unusual trends that do not follow the rest of the data, to do this we generate a QQ plot against a standard normal.

Figure 3: A QQ plot of the residuals against the standard normal

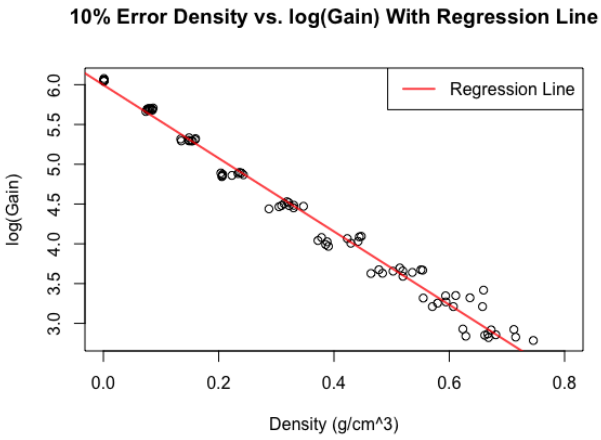


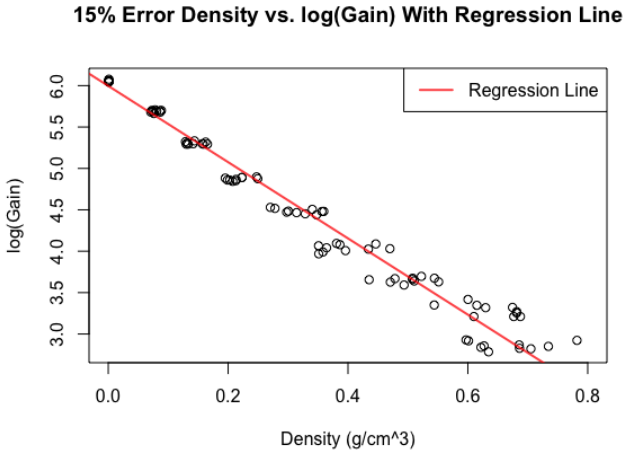
After examining our normal QQ plot we found that the residuals do indeed follow a standard normal and our model is suitable for a least squares regression line after our proposed transformation.

Testing Robustness

To test the robustness of our fit we added varying levels of noise to each recorded density. To calculate our noise we randomly sampled a number from the range [-x, x] where x is k% of the current density. For our tests we used k = 10, 15, and 20 and applied these k's to each density in 3 different simulations. After adding our noise we then compared our line of fit to these new values and compared the R^2.

Figure 4: Three scatter plots displaying our noisy densities against our old regression line. Left: Densities with errors of up to 10%, Right: Densities with errors of up to 15% Bottom: Densities with errors of up to 20%





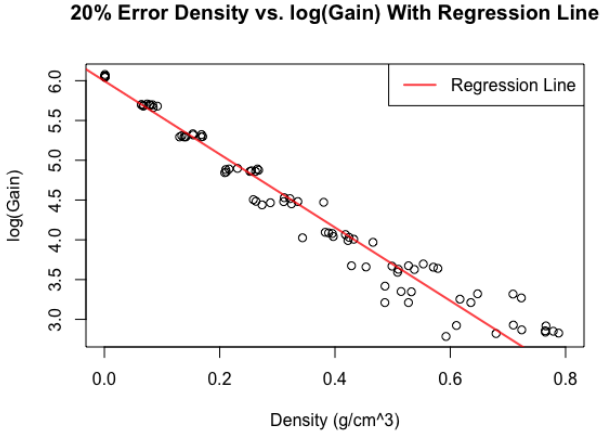


Table 1: Table showing the R^2 values from each simulation

Density Used	R^2
Original	0.996
10% error	0.988
15% error	0.970
20% error	0.958

After examining Figure 4 and Table 1 it is clear that even with varying levels of error, our original regression model is still a good fit for these data points. To confirm our strength of fit we examine the R^2 values of our new data points to the old regression model. The R^2 statistic shows how much of the variability in gain can be explained by our model. We found that even with up to 20% error our R^2 remains at 0.958 which is not a significant change from the original R^2 of 0.996.

Forward Prediction

To continue our investigation we will attempt to back test our model with values that have been observed to see if our model can be reliably used for prediction. To do this we generate a 95% prediction interval that gives us a range of where we can expect the value to fall and we use point based and interval based comparisons to test our model. First we begin with a visualization of the prediction band, and then we will test specific values to empirically compare our model against the data.

Figure 5: Scatter plot of density vs log(gain) with a least squares regression line in red and the 95% prediction interval in blue, and 95% confidence interval in green

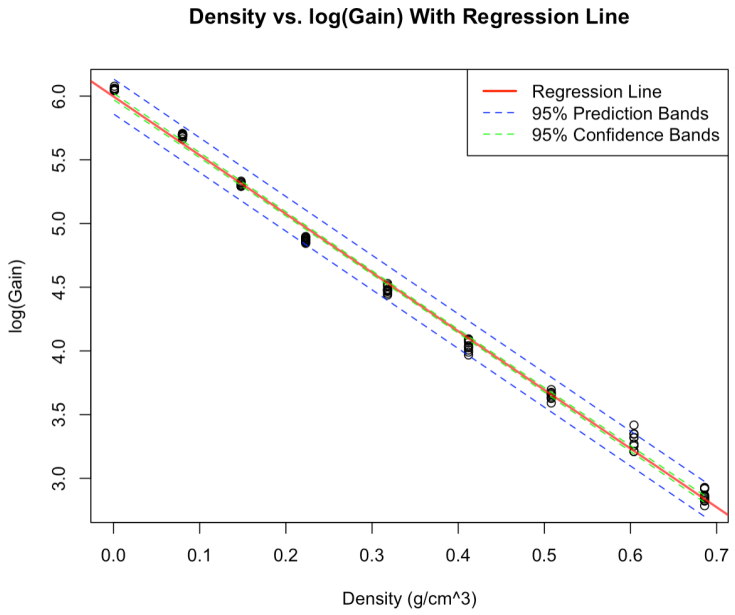


Table 2: Table comparing observed gain measurements and estimated gain measurements using density values of .508 and .001

Density Value (g/cm^3)	**Point Based Gain Estimate**	**Prediction Range**	**Observed Gain**	**Observed Range**
0.508	3.66	(3.52, 3.79)	3.657	(3.59, 3.7)
0.001	5.99	(5.86, 6.13)	6.056	(6.04, 6.08)

After generating our 95% prediction bands in Figure 5, we can see that the bands are close to the regression line and do a good job of capturing the data which implies that our model should generate good predictions. To empirically test this hypothesis we plug in observed density values and compare the predicted gain to the observed gain. In Table 2 we can see that our point based estimates are extremely close to our observed gain, and our prediction ranges match up well to our observed ranges. To test if predicting some gains are more accurate than other gains we examine the mean squared error of each gain at density values .508 and .001.

Table 3: Table comparing the mean squared errors of each gain at density levels .508 and .001

Density Value (g/cm^3)	**Mean Squared Error of Gain**
0.001	0.0041
0.508	0.00084

After measuring the mean squared error at each density value, we found that there are some gains that can be more accurately predicted than others. Our observations are detailed in Table 3, where we can see that the mean squared error for gains predicted at density .508 are much smaller than they are for .001 which suggests that some gains are better predicted than others.

Reverse Prediction

Next in our analysis we will invert the forward prediction and uncertainty bands that we just determined above to produce point estimates and prediction intervals for the density that correspond to the gain measurements 38.6 and 426.7. We do this by re-arranging our prediction line and interval formulas to have them now be X in terms of Y where it was previously Y in terms of X in the forward prediction, so that we can now plug in response values for Y in order to get an X value as a point estimate**.**

Figure 6: Scatter plot of log(gain) vs density with a least squares regression line in red and the 95% prediction interval in blue, and 95% confidence interval in green

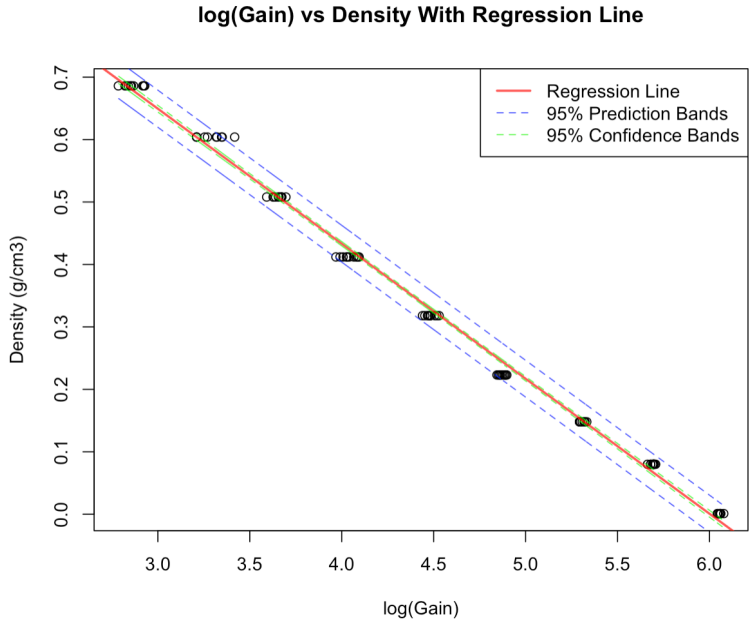


Table 4: Table comparing density measurements from mapping the measured gain to what the density would have been

Gain Value	**Point Based Density Estimate (g/cm3)**	**Actual Density Value in Data (g/cm3)**	**Prediction Range**	**Confidence Interval**
38.6	0.5081678	0.508	(0.4786626, 0.5376729)	(0.5042423, 0.5120933)
426.7	-0.01133153	0.001	(-0.04110982, 0.01844676)	(-0.01695305, -0.005710022)

Table 5: Table comparing the mean squared errors of each density at gain value 38.6 and 426.7

Gain Value	**Mean Squared Error of Density**
38.6	0.0001560127
426.7	3.829551e-05

We can see from the values above that the reverse prediction yields values that are quite close to the expected values for the true densities. Additionally we see that the mean square error of the inverse for density was lower, however we see that 38.6 is harder to predict since it has a wider 95% intervals and a greater mean square. Despite these supportive values it is important to

check the validity of this model in the next step where we cross-validate with the training data that helps us determine if these inputs are highly influential to our prediction.

Cross-Validation

Next in our analysis we will be performing cross validation in order to assess the reliability and effectiveness of our model's predictions using confidence and prediction bands, since it is possible that our reverse prediction may be influenced by the fact that the measurement corresponding to the densities 0.508 and 0.001 were included in the fitting. Therefore in order to mitigate this undesired outcome we may omit the measurements corresponding to these block densities of 0.508 and 0.001 to see how influential these corresponding values may be to the predictions made by our model. Therefore, we run the same calibration method as above but with these omitted values to cross validate and compare with our previous values intervals from when these block densities were included.

Figure 7: Scatter plot of density vs log(gain) with least squares regression line in red and the 95% prediction interval in blue, and 95% confidence interval in green, for when omitting density blocks of 0.508 (right hand side) and 0.001 (left hand side) in order to cross validate.

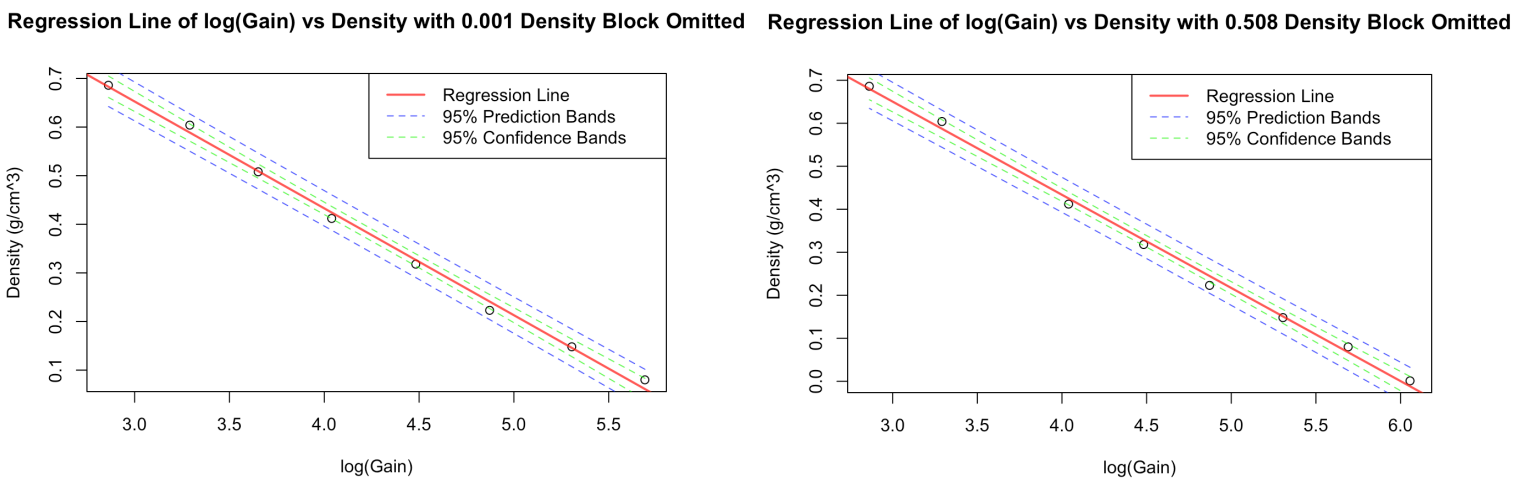


Table 6: Table comparing density measurements from cross-validation when the density block of when the corresponding values for the density block of 0.508 is removed.

Gain Value	**Point Based Density Estimate (g/cm^3)**	**Actual Density Value in Data (g/cm^3)**	**Prediction Range**	**Confidence Interval**
38.6	0.5087579	0.508	(0.4670858, 0.5504299)	(0.4910544, 0.5264613)
426.7	-0.01176274	0.001	(-0.05592688, 0.03240141)	(-0.03472629, 0.01120082)

Table 7: Table comparing density measurements from cross-validation when the corresponding values for the density block of 0.001 is removed.

Gain Value	**Point Based Density Estimate (g/cm^3)**	**Actual Density Value in Data (g/cm^3)**	**Prediction Range**	**Confidence Interval**
38.6	0.5090319	0.508	(0.4719027, 0.5461611)	(0.4944873, 0.5235765)
426.7	-0.01921668	0.001	(-0.06225272, 0.02381937)	(-0.04539046, 0.006957112)

We can see that in both the cases when we remove the values corresponding to the density block of 0.508 as well as when we do the same for the same set with 0.001, that both of the corresponding measures just slightly change due to the fewer data points that consequently make the intervals slightly wider. Furthermore, our point based estimates continue to be close to actual, and the actual value is captured well within the range of our 95% prediction interval. This indicates that our linear model with this new training data cross validates our predictions and is a good predictor of the observed data.

Advanced Analysis

In a further analysis we would like to more accurately model the relationship between the gain and the density by using polynomial regression to better fit the data. We can see from earlier in our analysis with the first 3 figures that represent plots such as the residuals, which depict the violation of linearity and suggest that the data has a nonlinear relationship. Furthermore, we see in figure 4 that when random noise is added to our regressed data that the cone shaped spread of the variance becomes evident, indicating the violation of constant variability. Thus these open pathways for potential violations when meeting the conditions for the least squares line regression since it appears that the data follows a non-linear trend, and as a result we can use the benefit of polynomial regression. In such a model for a predictor, X, is :

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_h X^h + \epsilon,$$

Where h is the degree and where β is the regression coefficients which allows for a nonlinear relationship between X and Y (response variable) (1).

Figure 8: Polynomial Regressions of degree 1, 2, 3, 4, 5 & 6 overlaid to show the differences in their fits with the data, which we see best fit with degree 4 and fits the poorest with degree 1.

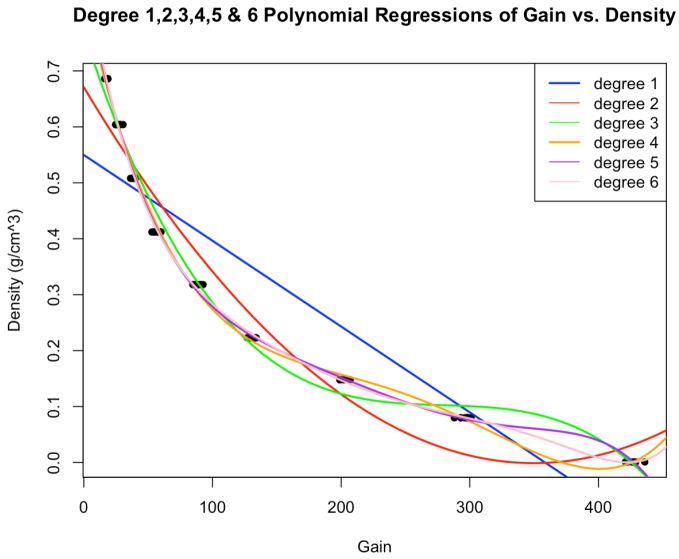
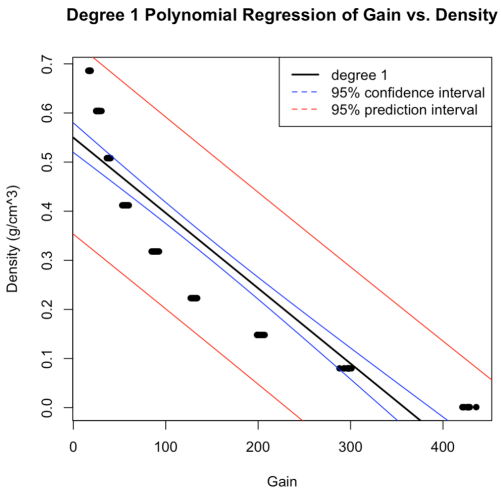
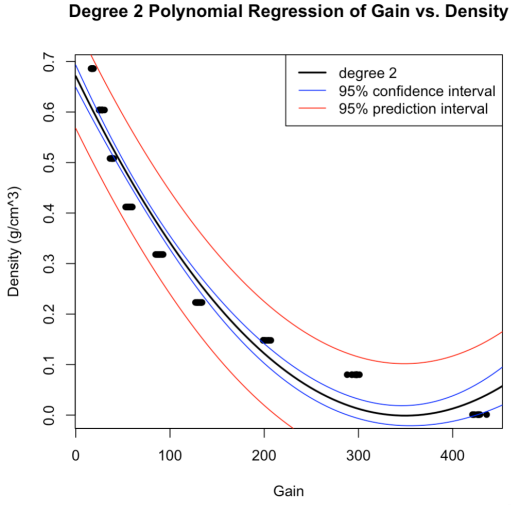
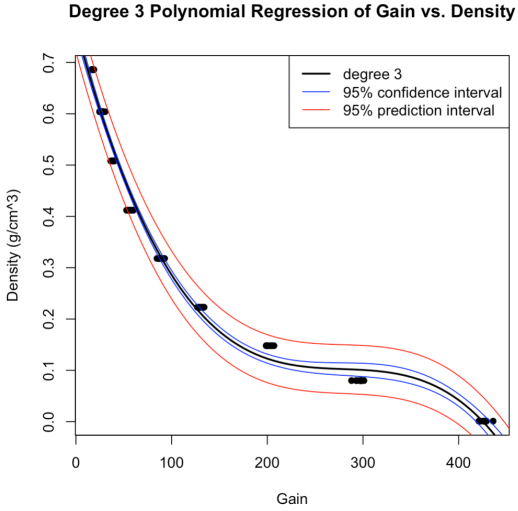
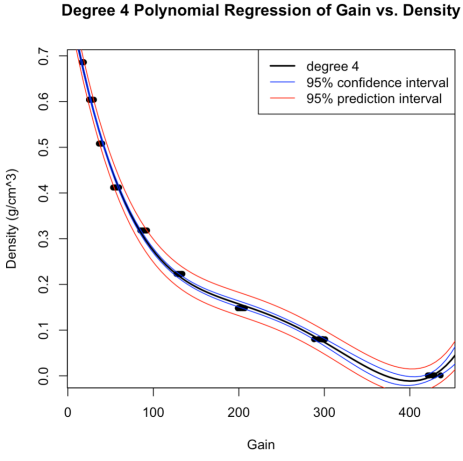


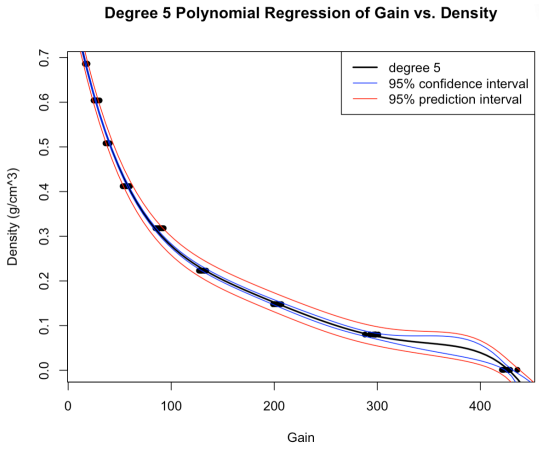
Figure 9: Individual polynomial regression for degree 1, 2, 3, 4, 5 & 6 with 95% confidence intervals in blue and 95% prediction intervals in red, to better show when overfitting occurs when our degree is too large which occurs with arguably degree 5 or 6 since in the last 2 plots we see the confidence and prediction lines begin deviating from the regression line.

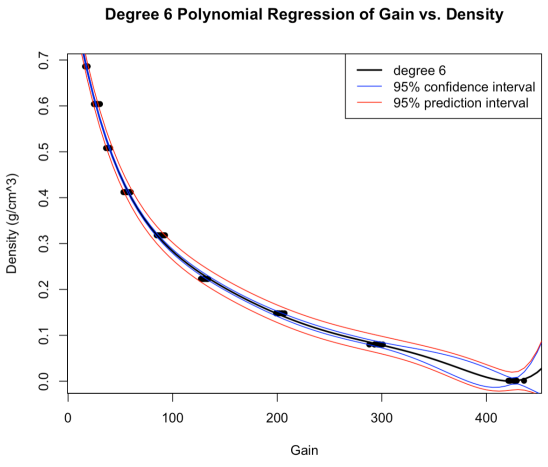












Therefore, we can see from the plots above that generally up until a certain degree that with polynomial regression non-linear relationships are better fitted with the greater value of degree. However that is until we encounter overfitting. In this case we can argue that the best fits appear with degree 4 or 5 but with either 5 or 6 we begin to see that despite the regression fit that the 95% confidence and prediction bands begin to deviate from the regression line indicating overfitting. This occurs when attempting to estimate too many parameters from the sample than the data allows for optimally, since overfitting would instead represent the noise rather than the relationship in the data (2). Therefore it is important to test multiple degrees in order to narrow down what degrees give the best fits before overfitting occurs.

Conclusion and Discussion

For the first section of our analysis we generated a basic least squares regression line for our dataset, however we found that the dataset wasn't suitable to be modelled by least squares as it is. Originally the dataset followed a non-linear trend and if we wanted to model the data using least squares regression we would need to transform it to make it linear.

After determining the need for a transformation, we began to try and find different transformations that could linearize the data. We found that the equation for the gamma ray measurement was an exponential function, and theoretically if we took the log of the function it would give us a more linear dataset. After performing the log transformation on our data we empirically tested it to make sure that it followed the requirements for least squares regression: linearity, constant variance, and normal residuals. We found that it empirically met the linearity and normal residuals requirements, however the variance did not. We believe that the empirical failure of constant variance can be justified by the fact that our model is theoretically ideal for the measurement function used to collect the data.

With our model chosen and tested we wanted to test the robustness of our fit if random variance is introduced into the data. We did this by adding up to 10%, 15%, and 20% of random error to each density in the data and testing how well our model fit to this new data. We found that even after adding this noise, our model still maintained a very high R^2coefficient (.958) and did not have significant change from the original R^2(.996).

After deciding that our model fit our data and was robust enough to handle random variance, we decided to compare point based and interval based estimations to the actual observed values. We used the known density values of .508 and .001 and found that both our point based and interval based estimations were very close to their observed counterparts and seemed to generate accurate predictions. We also found that some gains tend to have better estimations than others by testing the mean squared error of the gains at densities .508 and .001.

Next we invert this forward prediction to produce point estimates and prediction intervals for the density that correspond to the gain measurements 38.6 and 426.7. We then see that the reverse prediction yields values that are quite close to the expected values for the true densities.

Additionally we see that the mean square error of the inverse for density was lower, however we see that 38.6 is harder to predict since it has a wider 95% intervals and a greater mean square.

Lastly, In our analysis we performed cross validation on our model and training data in order to assess the reliability and effectiveness of our model's predictions using confidence and prediction bands. We remove the values corresponding to 0.508 and 0.001 block densities in order to determine how much influence they have on our predictions when observing the reverse predictions. Ultimately this indicated that our linear model with this new training data that omits 0.001 and 0.508 block densities, cross validates our predictions and is a good predictor of the observed data.

In discussion, our dataset allowed us to create a model that predicts snow density from a measured gain, however our model is based upon data that was gathered from a single snow gauge using nine densities of polyethylene blocks. I think that our data is slightly biased as it is limited to just one sample of gauges and blocks whereas the ideal dataset would have values taken from multiple different gauges and various polyethylene blocks. Furthermore, the data is limited in the sense that it is not generalizable to other climates, altitutedes, and environments since our data only is representative of Soda Springs, California. Thus it would be more ideal in future works to have data from various regions to make this calibration procedure more generalizable to other environments rather than just California. Additionally another limitation is that we only have 9 measurements per density block, which is less than 12 which may suggest that these data points are not normal under the central limit theorem balance condition (or less than 30 for highly skewed data). Lastly further research can be done using different transformations on the data for linear regression, different methods of linear regression (ie. LAD regression), or other non-linear regression methods such as polynomial regression in order to more accurately represent the relationships amongst the data.

Works Cited

1. "7.7 - Polynomial Regression." *7.7 - Polynomial Regression | STAT 462*, The Pennsylvania State University, online.stat.psu.edu/stat462/node/158/.
2. Frost, Jim, et al. "Overfitting Regression Models: Problems, Detection, and Avoidance." *Statistics By Jim*, 5 Apr. 2021, statisticsbyjim.com/regression/overfitting-regression-models/.