# Calibration of Snow Gauge for Snow Density Measurement

Ayaan Saifi and Antony Sikorski

## 0. Contribution Statement

Both Ayaan Saifi and Antony Sikorski contributed evenly to the project. The R code was evenly split, and each author contributed equally to the write-up of the report. The advanced analysis was additionally evenly split between the two. That said, specific primary contributions are as follows:
Ayaan Saifi was the primary contributor for Questions 2, 3, 6
Antony Sikorski was the primary contributor for Questions 1, 4, 5

## 1. Introduction

A snow gauge emits gamma rays at snow in order to calculate its approximate density. In order to correctly calibrate a snow gauge, one must first create an accurate model that maps the gamma ray intensity as a function of density. After a model is found, the data can be reversed, and the gamma ray intensity is mapped to correctly estimate the density of the blocks.
In order to test our snow gauge, we calibrate it using polyethylene blocks of various densities that function extremely similar to snow. Gamma ray measurements from the snow density can be expressed with the function $g = Ae^{\beta d}$, with “g” representing gamma ray gain, “d” representing density, and “$A, \beta$” being constants that we wish to approximate.
We first attempted to fit a linear regression to the raw data, but as the theoretical relationship between gain and density is exponential, we found it much more accurate to create a model using a logarithmic transformation of the data. We first tested the model using formal statistical testing and visual analysis in order to gain a general understanding of the accuracy. We then followed by forward and reverse predictions of gain and density intervals and point estimates to more specifically test the accuracy of our model, and to see if there are any patterns regarding estimating the densities of the blocks with different gains. Since our final model involved using a linear regression on our logarithmically transformed data, advanced analysis was performed in order to see if we could better estimate the transformed data using a cubic prediction to account for a slight, observed curvature. The linear model was overall found to be the most useful for further extrapolation and calibration of snow gauges, demonstrating extremely high accuracy and practicality.

# 2. Basic Analysis

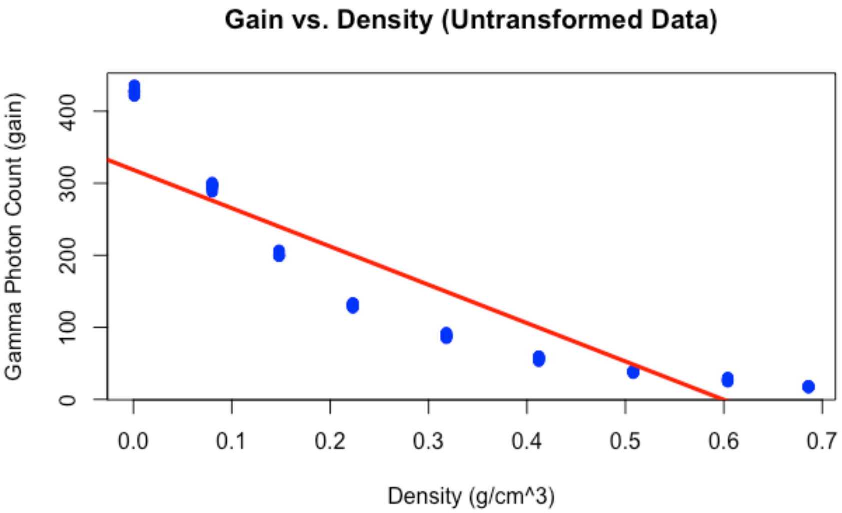
## 2.1. Linear Regression of Raw Data

**Methods:**

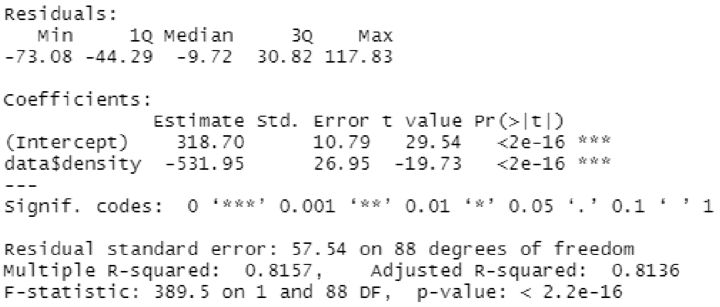
To begin our initial analysis, we first attempt to fit the data with a linear regression to see whether or not this can be a suitable model for calibrating the snow gauge. Using gain Y as our dependent variable, and density d as our independent variable, we create our linear regression of the form $Y = \beta_0 + \beta_1 d + error$. We plot the linear regression line on the plot of the data in order to visually examine the fit. Afterwards, we plot the residuals to attempt to identify any evidence of a pattern, and check the normality of the residuals using both visual analysis and formal statistical testing.

**Analysis:**

We perform our linear regression on the data, and plot it on top of the actual scatter plot to visually examine the fit.

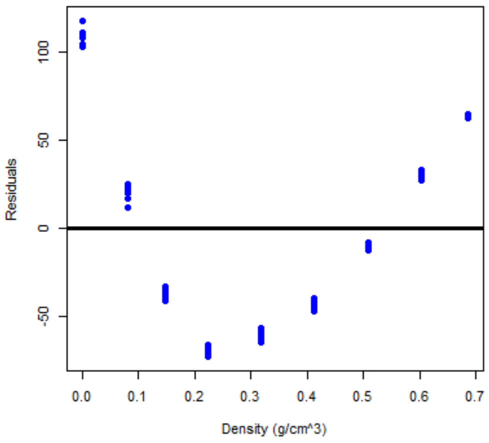


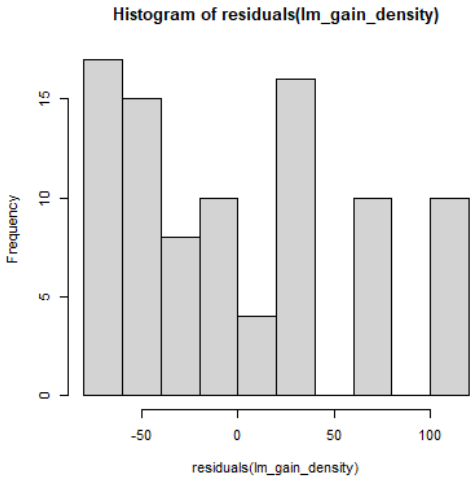
A summary of the fit of the linear regression model is shown below.

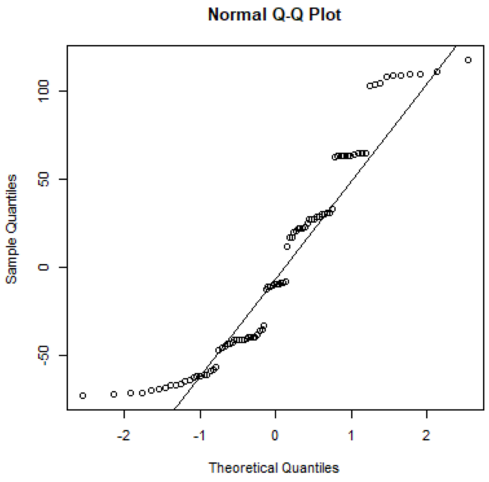


```
Residuals:
   Min     1Q Median     3Q    Max 
-73.08 -44.29  -9.72  30.82 117.83 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)     318.70      10.79   29.54   <2e-16 ***
data$density  -531.95      26.95  -19.73   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 57.54 on 88 degrees of freedom
Multiple R-squared:  0.8157,	Adjusted R-squared:  0.8136 
F-statistic: 389.5 on 1 and 88 DF,  p-value: < 2.2e-16
```

Next, we examine the normality of the residuals with a scatter plot, histogram, and a Q-Q normality plot.







After visual analysis, we perform a one sample Kolmogorov-Smirnov test to see the difference between our residuals and a normal distribution, which results in a low p-value of 0.01693.

**Conclusion:**

We find our linear regression model to be $Y = -531.951d + 318.701$, but it does not appear to be a good fit due to multiple factors. At first glance, although the $R^2 = 0.8157$ value appears high, when the trend line is plotted against the data, it appears to be inappropriate. Our analysis of the residuals shows an evident pattern, and does not appear to be anywhere near normally distributed after a visual inspection. Following this with a Kolmogorov-Smirnov test of the residuals against the normal distribution, and a Q-Q plot against the normal, the visual evidence and extremely low p-value suggest that the residuals are not normally distributed. It appears that a transformation will be needed for a more accurate model.

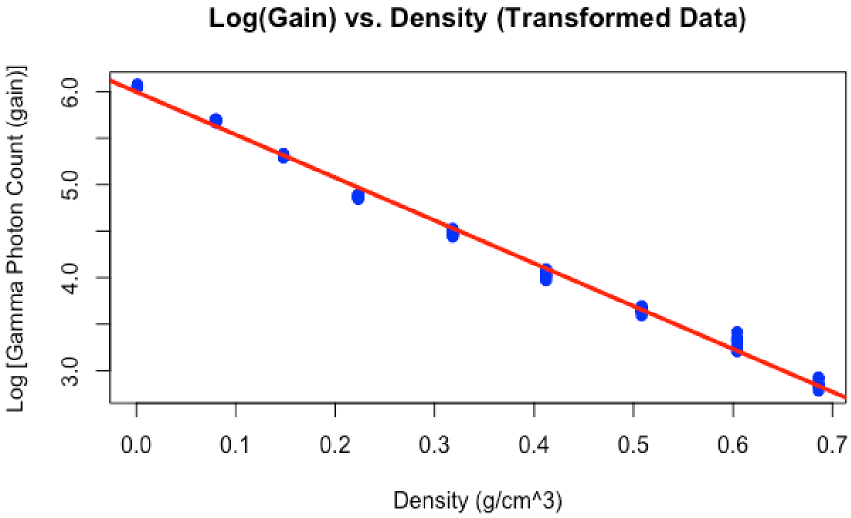
## 2.2. Data Transformation and Appropriate Model Determination

**Methods:**

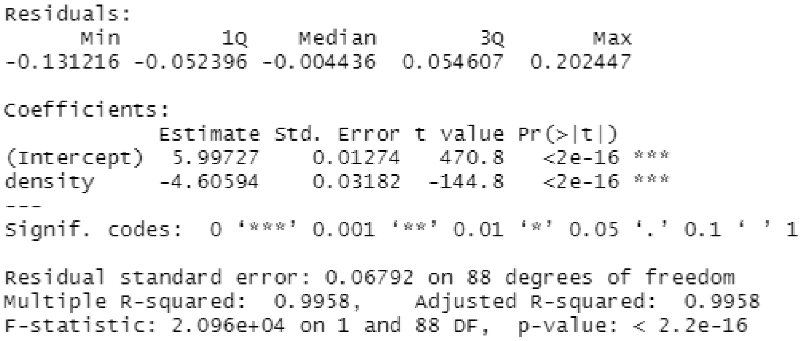
We repeat the procedure that we performed in part one, but this time we transform our data using a logarithmic transformation in order to attempt to create a better model. This is because hypothesized relationship between the gain and the density is of the form $g = Ae^{\beta d}$. Performing a log transform changes the equation to $log(g) = log(A) + \beta d$. Now, if set $Y = log(g)$, we can perform the same procedure to get a linear regression of the form $Y = log(A) + \beta d + error$, in order to more accurately fit the model. In addition to visual assessment and regression testing, we will again perform tests and visual analysis to check the normality of the residuals.

**Analysis:**

We take the logarithmic transformation of the data, and plot the new data along with a new linear regression for visual analysis.

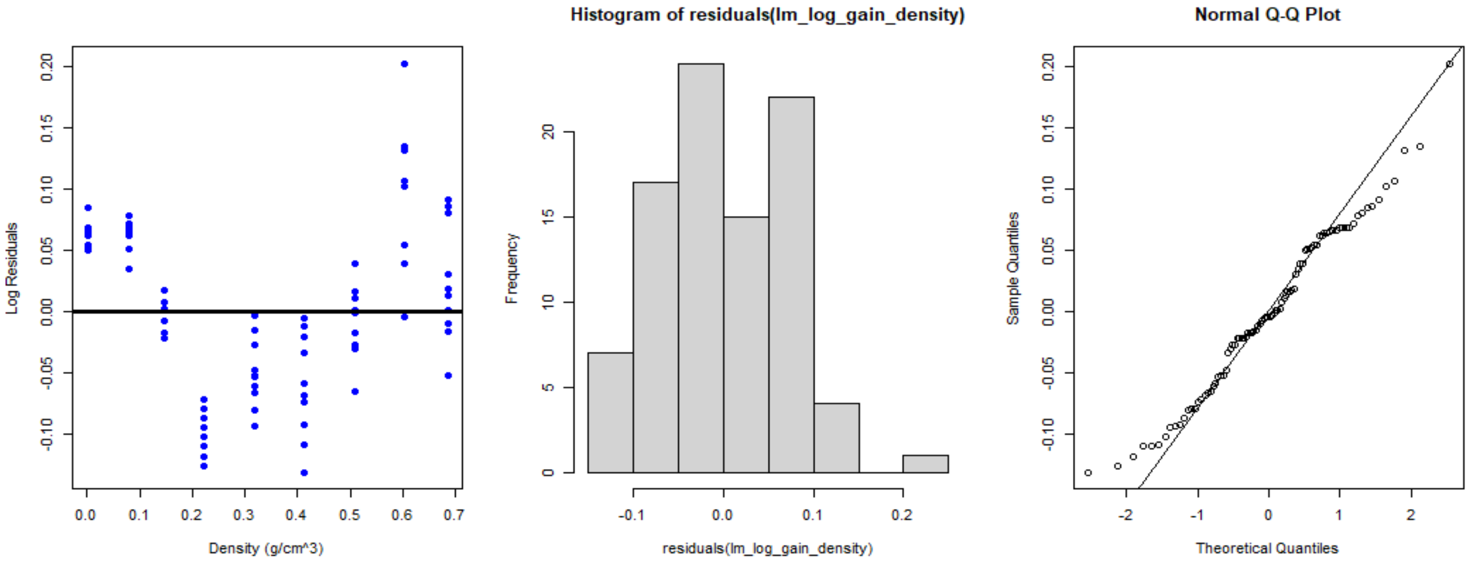


A summary of the regression on the log transformation data is provided below.



```
Residuals:
      Min        1Q    Median        3Q       Max 
-0.131216 -0.052396 -0.004436  0.054607  0.202447 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.99727    0.01274   470.8   <2e-16 ***
density     -4.60594    0.03182  -144.8   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06792 on 88 degrees of freedom
Multiple R-squared:  0.9958,	Adjusted R-squared:  0.9958 
F-statistic: 2.096e+04 on 1 and 88 DF,  p-value: < 2.2e-16
```

We then plot the residual scatter plot, histogram, and Q-Q plot against the normal distribution.



Finally, we perform a one sample Kolmogorov-Smirnov test to see the difference between our residuals and a normal distribution, which results in a high p-value of 0.5892.

**Conclusion:**

It appears that this model is much more accurate than the previous one without the log transformation. The equation for our linear regression of the logarithmically transformed model is $Y = -4.61d + 5.997$. Visual analysis and a high $R^2 = 0.9958$ value show that the model is extremely similar. The residuals appear to have no evident pattern, and the histogram shows that their distribution is approximately normal. This is further confirmed by visual analysis of the Q-Q plot against the normal, which shows a significant improvement over the last model. Finally, the Kolmogorov Smirnov test returns a large p-value of 0.5892, meaning that the distribution of residuals is quite similar to that of a normal distribution. To conclude, the logarithmic transform allows us to create a much more accurate model, which we can later invert to properly model the relationship between the gain and the density.

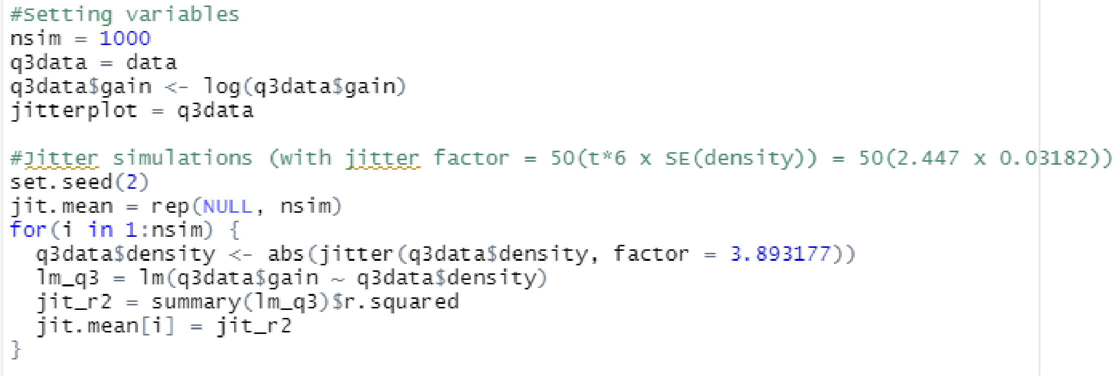
## 2.3. Robustness in Reporting of Density

**Methods:**

We experiment with the idea of the densities of the polyethylene blocks not being reported exactly to see how it will impact the fit of the model. In this case, we simulate some error in the reporting of the blocks by implementing the jitter function in our density data. The parameter of the jitter function is the 95% error margin for the density data from the original linear model, so that the deviations in the data are within a reasonable range of what could be incorrectly reported. We simulate 1000 jittered data sets and find the mean $R^2$ so that we can see the change against our original $R^2$value, which is an excellent indicator of fit. In addition to this, we plot the scatter plot and fit of our model in order to perform visual analysis on how much the incorrect reporting changes the data.

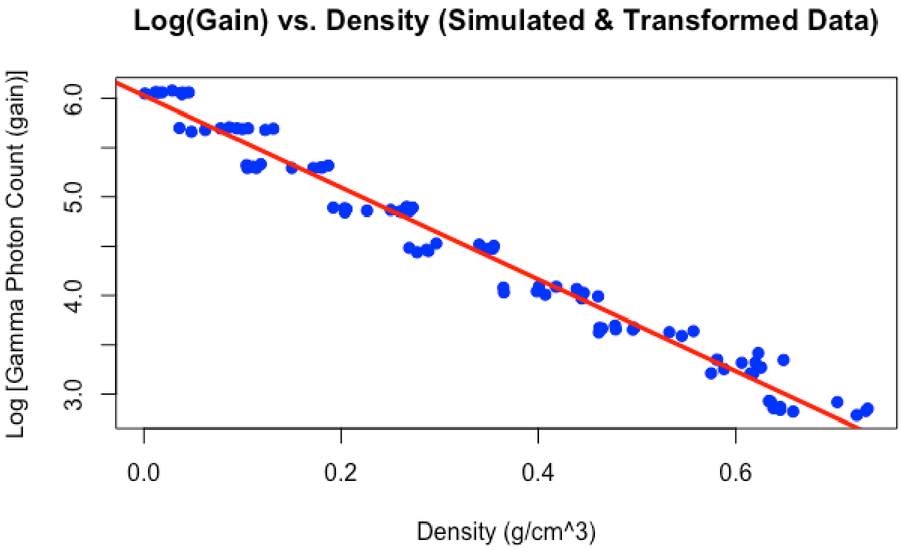
**Analysis:**

This loop creates 1000 simulations and calculates their $R^2$ value. After this we calculate the mean value for all of the simulations.



```
#Setting variables
nsim = 1000
q3data = data
q3data$gain <- log(q3data$gain)
jitterplot = q3data

#Jitter simulations (with jitter factor = 50(t*6 x SE(density)) = 50(2.447 x 0.03182))
set.seed(2)
jit.mean = rep(NULL, nsim)
for(i in 1:nsim) {
  q3data$density <- abs(jitter(q3data$density, factor = 3.893177))
  lm_q3 = lm(q3data$gain ~ q3data$density)
  jit_r2 = summary(lm_q3)$r.squared
  jit.mean[i] = jit_r2
}
```

We find the mean $R^2$ value to be roughly 0.977, and plot the scatter plot of the jittered data against our model to visually understand the fit.



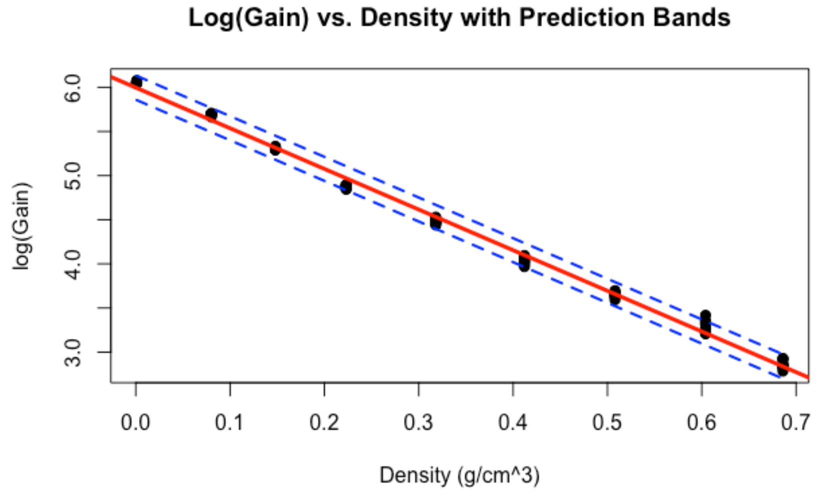
**Conclusion:**

Since there is a high possibility that the density was incorrectly reported, it is important to investigate the possibility. Simulating 1000 jittered datasets and calculating their mean $R^2$ value produces a value of 0.9765475, which is indeed lower than the original estimates where the data was supposedly correctly reported. Despite this, the fit is still quite good, and visual analysis reaffirms the fact that our original model is still a good fit for jittered data, so long as it is incorrectly reported within a reasonable error margin. This further strengthens our confidence in the log transformed linear regression fit model.

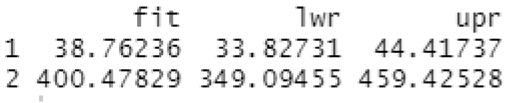
## 2.4. Forward Prediction of Photon Count Gain Values

**Methods:**
In order to further test the accuracy of our model, we construct prediction intervals in order to encapsulate the true prediction values with a probability of 95%. We use prediction bands rather than confidence intervals due to the fact that confidence intervals are only valid for simple random samples, which is not the case with the data that we are analyzing. We further create point estimates and 95% prediction intervals for the values of densities of 0.508 and 0.001 as tests, and analyze the accuracy of such measurements against the range of gain values from the original data we were given for those densities. We hope to observe a pattern in which we see which gain values can be more accurately predicted so we can have a more comprehensive understanding of our model.

**Analysis:**
To begin, we plot our linear regression line, along with the 95% prediction lines, which provide a 95% chance of capturing a new value if it were to be recorded at that density.



After this, we compute the range of the point estimate, along with the 95% prediction range of the gain for the density values of 0.508 and 0.001.



```
        fit       lwr       upr
1  38.76236  33.82731  44.41737
2 400.47829 349.09455 459.42528
```

The following output claims that for a density of 0.508, the approximate expected gain is 38.76, with a 95% interval spanning over (33.83, 44.42). For the density of 0.001, the expected gain is about 400.49, with an interval of (349.09, 459.46).

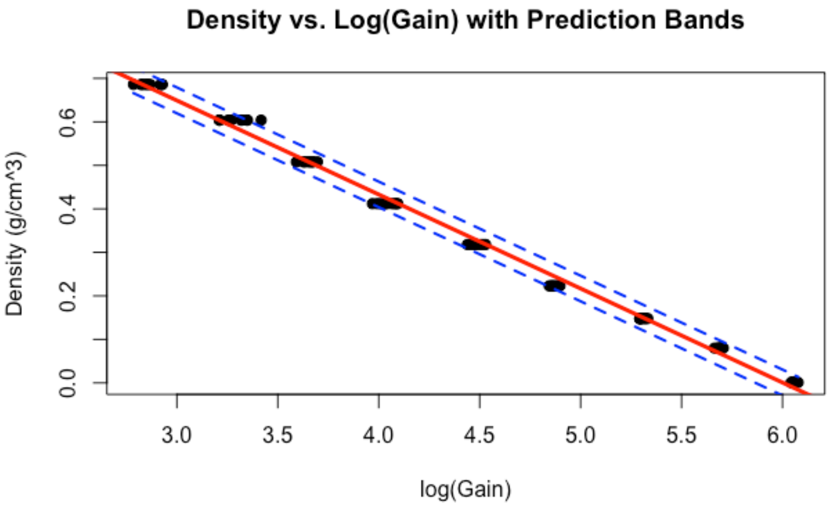
Comparatively, the range of our actual gain values from the data set for a density of 0.508 is (36.3, 40.3), and the range for a density of 0.001 is (421, 436).

**Conclusion:**
Visual analysis strengthens our confidence in our model to the narrow prediction intervals, and the majority amount of the data appears to fall within the bands. Observing the actual range of gain values for the two densities, and comparing them to the prediction bands, it appears that the prediction bands completely encapsulate the actual data values. It appears that it is easier to predict the gain for blocks with higher density compared to those with a lower density. The point estimate for the 0.508 density block lies within the actual gain data range, while the point estimate for the 0.001 density block is off by 20 photons. To further prove this claim, the actual prediction range for the 0.001 density block has a width of about 110, and needs to capture an interval of only 15. Regarding the 0.508 density block, the prediction interval has a width of 11, and captures an actual interval of width 4. This evidence leads us to believe that it is easier to predict the gain for blocks of higher density.

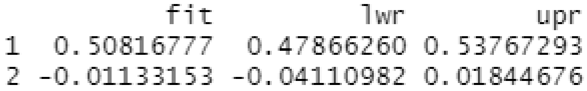
## 2.5. Reverse Prediction of Density Values

**Methods:**
We perform an extremely similar procedure to the one in the previous problem, except to begin this one we flip the axis, and attempt to predict density from gain. We plot the prediction bands on the linear regression with reversed variables to begin with visual analysis. We then find point estimates and 95% prediction intervals for the density values that would come from gains of 38.6 and 426.7. We compare these to the actual densities from data in order to see if there is a pattern in which densities are easier to predict according to our model.

**Analysis:**
First, we plot the regression line, along with the 95% prediction lines .



We then calculate the predicted 95% density range for gains of 38.6 and 426.7.



```
          fit         lwr        upr
1  0.50816777  0.47866260 0.53767293
2 -0.01133153 -0.04110982 0.01844676
```

For a gain of 38.6, the model predicts a density of roughly 0.508, with a 95% prediction range of (0.477, 0.538). For the gain of 426.7, the model predicts a density of -0.011 and a range of (-0.041, 0.018).

**Conclusion:**
Yet again, visual analysis yields a conclusion of an accurate model due to narrow prediction bands and a majority of the data being captured. In addition to this, the reverse predictions show that higher densities can more accurately be predicted from their gain. The point estimate for the density from a gain of 38.6 is nearly exactly the same as the original density of 0.508, while the point estimate for a gain of 426.7 reports a negative value of -0.011, which is close but unrealistic for an actual value of 0.001. Both intervals have a small width and accurately capture the actual densities, but the point estimates make it quite clear that it is easier to predict higher densities from a given gain. Regardless, the model remains reasonably close.

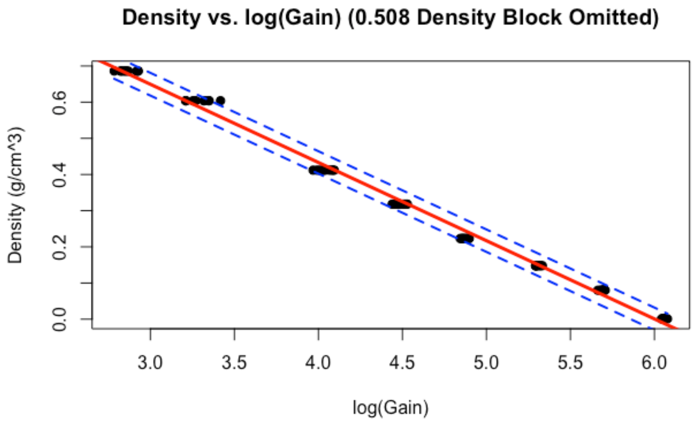
## 2.6. Cross-Validation of Reverse Predictions

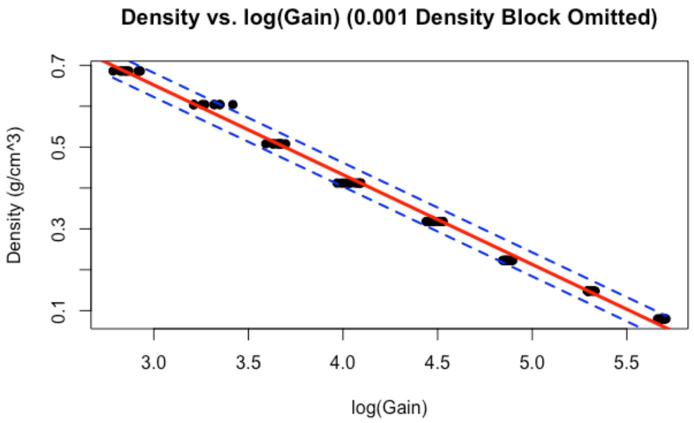
**Methods:**
In order to better understand how accurate our reverse prediction is, we must run a cross-validation of two modified subsets of our transformed data. We first removed all measurements of 0.508 g/cm$^3$ from our transformed data set, plotted the data points on a scatter plot, and constructed a linear regression model to fit the data. We then constructed prediction bands around our linear regression model and used the model to predict a point and interval

estimate of the density of polyethylene blocks with a gain of 38.6 (the average measured gain of blocks with a density of 0.508 g/cm$^3$). We then used the same method in the omission of density measurements of 0.001 g/cm$^3$, eventually using the corresponding linear regression model to predict a point and interval estimate of blocks with a gain of 426.7 (the average gain of blocks with a density of 0.001 g/cm$^3$).

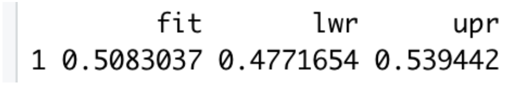
**Analysis:**

After creating two subsets of our transformed data that omit density measurements of 0.508g/cm$^3$ and 0.001 g/cm$^3$, we plotted them with prediction bands as shown below:

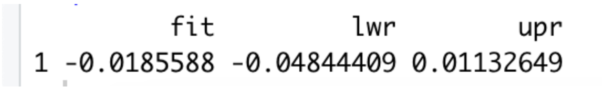




For our 0.508 g/cm$^3$ omission regression, we found that $\beta_1$ = -0.216278, $\beta_0$ = 1.298422, and $R^2$ = 0.9956. Subsequently, for our 0.001 g/cm$^3$ omission regression, we found that $\beta_1$ = -0.219395, $\beta_0$ = 1.310114, and R$^2$ = 0.995. We then used these models to produce the corresponding gain measurements for densities of 0.508g/cm$^3$ and 0.001 g/cm$^3$ respectively:



```
        fit       lwr      upr
1 0.5083037 0.4771654 0.539442
```



```
         fit         lwr        upr
1 -0.0185588 -0.04844409 0.01132649
```

**Conclusion:**

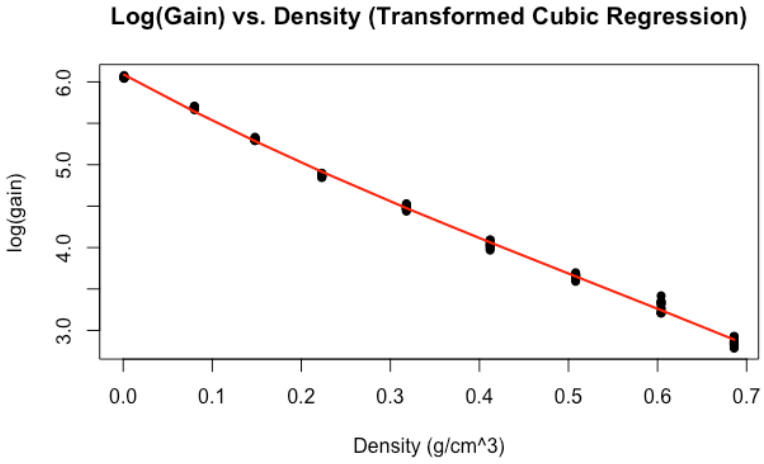
From the first point and interval estimate of our cross-validation, we can see that the density corresponding to an average gain of 38.6 (0.508 g/cm$^3$) lands directly on our point estimate, thus being in the middle of our interval estimate. On the other hand, the density corresponding to an average gain of 426.7 (0.001 g/cm$^3$) does not exactly land on our point estimate, landing within the upper half of our interval estimate. These findings are quite consistent with our point and interval estimates from our original reverse prediction, thus showing that the original reverse prediction was not greatly influenced by the presence of the 0.508g/cm$^3$ and 0.001 g/cm$^3$ density measurements, which were utilized throughout our forward and reverse prediction processes.

# 3. Advanced Analysis

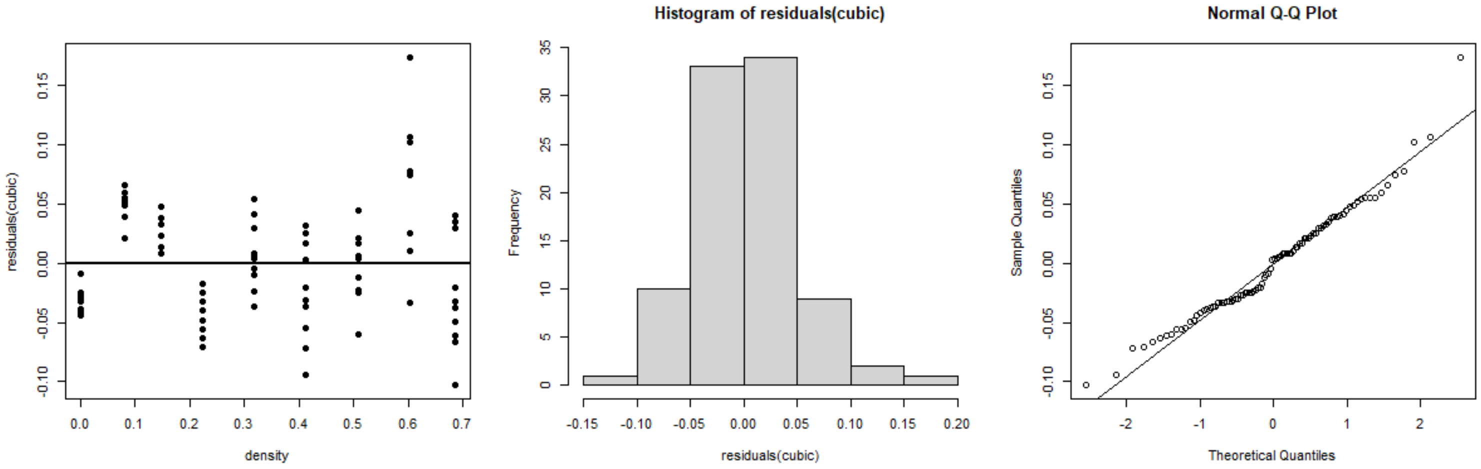
## 3.1 Fitting a Cubic Model to Logarithmically Transformed Data

**Methods:**
Since we have already logarithmically transformed the data and found an accurate linear model for calibration, we wish to address a few minor issues with a more advanced model. For lower densities (higher gain measurements), the intervals and point estimates were more difficult to predict than those for higher densities (lower gain measurements). Due to this fact, we believe the linear model may not be sufficient, and we fit a cubic model to our data to account for the curvature that we suspect to be causing the inconsistencies in predictions of higher and lower densities. We both visually analyze the model, as well as find the fit ($R^2$) value. Finally, we examine the residuals for normality to ensure the viability of the model.

**Analysis:**
First, we plot the model for visual analysis, and calculate an $R^2$ value of 0.998, which demonstrates extremely good fit.



We then follow by testing the normality of the residuals using scatter plots, histograms, and Q-Q plots against the normal.



We then perform a one sample Kolmogorov-Smirnov test to see the difference between our residuals and a normal distribution, which results in a high p-value of 0.3222.

**Conclusion:**
The cubic model appears to be an excellent fit for our logarithmically transformed data, providing an extremely high $R^2$ value of 0.998, with residuals that appear to be roughly normal. It satisfies all of the necessary conditions as a model for this range of gain and density values, and appears to have a slightly higher $R^2$ value than that of the linear model (+0.003). Visual analysis yields that the model accounts for the curvature better than a linear model possibly could, yet further prediction interval testing would yield more accurate results.

## 4. Discussion and Conclusion

The objective of this study was to find the most accurate and applicable regression model for the calibration of a snow gauge that measures the density of the snow by counting the gain, which is the gamma photon count. Our data provided us with gain and density measurements for a snow gauge that measured the densities of polyethylene blocks, which are quite similar to snow in nature. We attempted linear regression on the raw data, and realized that the data must be logarithmically transformed in order for us to more accurately use linear regression. The model proved to be extremely accurate, yet we wished to improve upon the model due to an apparent presence of minor curvature, which was making it more difficult to predict the density of a block from a higher gain. To improve upon the model, we logarithmically transformed the data and fitted a cubic curve to it, which indeed provided a slightly more accurate estimate, but not notably so. Although the cubic model provided slightly higher accuracy and a better fit, we chose a final linear model of logarithmically transformed data to be the most accurate method of calibration for the data. Our model predicts density from gain with an $R^2$ value of 0.9958 by the following equation: $d = -0.216203 log(g) + 1.298013$, where "$d$" represents density, and "$g$" represents gain.

We choose the logarithmic model over the cubic model primarily due to extrapolative purposes. If working with densities that are outside of the range of the current data, it is likely that the cubic model will deviate too far from the data to be applicable. The linear model of the logarithmically transformed data will remain consistent when expanded, and offers much higher practicality considering the fact that our data set is not a good representation of all snow densities in the natural world. The fact that the cubic model has a better fit and is slightly more accurate is negligible due to the miniscule difference between it and it’s linear rival.

This model could be further applied to similar studies, such as the one done by James L. Smith (1) and his team at the Pacific Southwest research station, where they estimate snow density using the albedo method. Rather than using a snow gauge, the measurements are taken from an aerial standpoint, and our model could be much more useful in a setting like this rather than by using a snow gauge that is limited to its ground location. In addition to this, further research and testing could be done on the cubic model of the logarithmic data regression on a larger range of densities in order to see if it is potentially as expandable as the logarithmic one. For now, our chosen linear model of the transformed data provides extreme accuracy for future snow gauge calibrations, with only a slight bit of difficulty predicting lower densities.

# 5. Work Cited

1. Smith, James L.; Halverson, Howard G. 1979. Estimating snowpack density from Albedo measurement. Res. Pap. PSW-RP-136. Berkeley, CA: U.S. Department of Agriculture, Forest Service, Pacific Southwest Forest and Range Experiment Station. 13 p