# MATH 189 Case Study 4

## Snow Gauge

Yuxin Zou, A13996888
Yifei Li, A92082060
Bo Hu, A13805750
Bo Zhang, A13616488
Xuanyu Wu, A13569778

### Literature Review

In all ages, natural disasters happen across the world, threatening millions of individuals' lives and properties. Even worse, human development has driven the amount of and severity of natural disasters up to a new level, thus raising our concern to mother nature, and increasing our awareness of when and in what condition would those disasters arrive. In this case, scientists developed multiple methods to predict different disasters. For example, In Chile, a country where the largest earthquake recorded in the 20th century has taken place, scientists trained neural networks to predict the probability of a certain earthquake(beyond a certain threshold) is going to happen(Reyes). Furthermore, in the study of tornadoes, a combination of universities, research institutions, and private meteorological companies had worked together over the last 20 years to predict and detect the formation of tornadoes. Most of their researches were focused on the physical sciences and technology portion of the warning process, as well as the improvement of warning communication and coordination process that can possibly lead to the further reduction of tornado related injuries and fatalities(Wei). To predict and detect tsunami, a group of scientists designed a sensor network architecture that use the directed diffusion routing protocol as a baseline network and develop several communication mechanisms to improve the accuracy(casey). In this paper, we will see a operation ran by Forest Service of the United States Department of Agriculture (USDA) that measures the snow pack which is used to help monitor the water supply and predict the risk of flooding.

### Introduction and Data Description

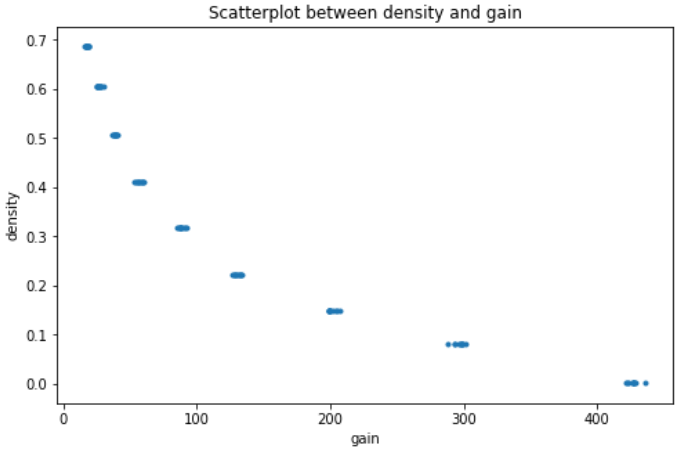
To measure the amount of water supply, a gamma transmission snow gauge was set in the Central Sierra Nevada near Soda Springs, CA. It measures the amount of rain that is absorbed by the snow(the more it absorbs, the more its density is, and if the snow absorb the water up to a threshold, flood takes place). Note that the snow gauge has no effect on the snow during the measuring process, so the data gives an authentic, continuous measurement over a consecutive period of time. The snow gauge receives the gamma ray emissions and converts it to the snow density. And certain adjustments are made in the conversion functions which adjust to the

change of radioactive source decay to ensure accuracy. To collect data, scientists place polyethylene blocks of known densities, which are used to simulated snow, between the two poles of the snow gauge and take readings on the blocks. The middle 10 of 30 measurements taken are recorded in the data, and there are 10 measurements for each of 9 polyethylene blocks with different densities. The first column is the density of the block, and the second column is the gain(the gauge measurement).

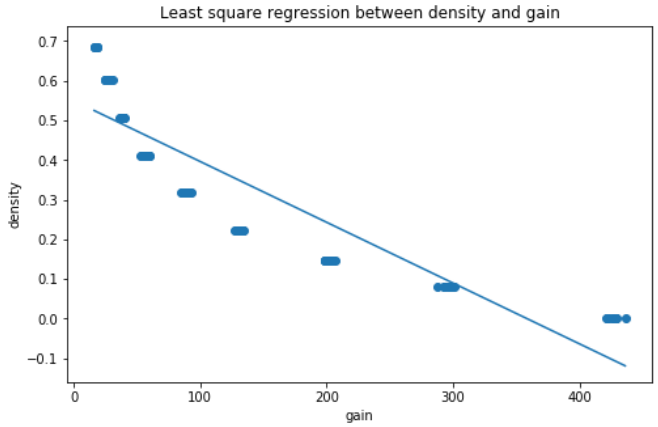
## Investigation

### Fitting

The following scatter plot is the original data we have for density and gain. Knowing that these gains come from measurements from the same 9 polyethylene blocks and each density of block produces multiple gain outcomes. So, graphically, there are 9 groups in the data set grouped by their density.

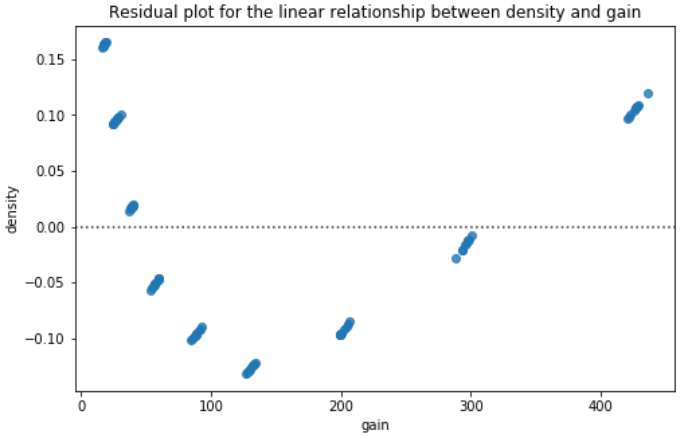



Judging from the scatter plot, there is a nonlinear relationship between gain and density. We can try to fit a linear regression model between the raw data and see how the model performs.



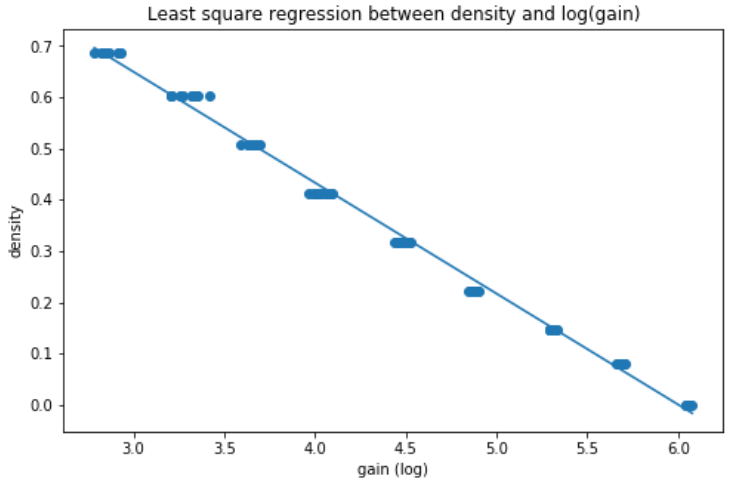

```
{'slope': -0.0015334078316468012,
 'intercept': 0.5497239543095568,
 'rvalue': -0.9031596703485595,
 'pvalue': 4.518580918276382e-34,
 'stderr': 7.76993788865363e-05}
```

**density = -0.001533 * gain + 0.5497**



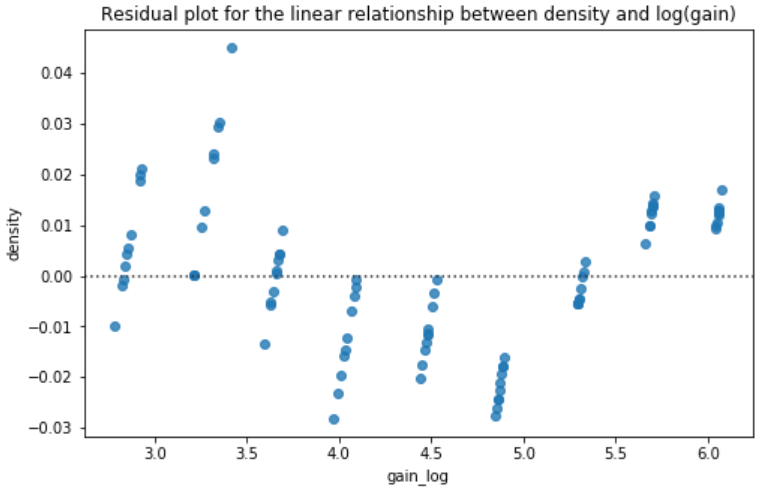

The linear regression line tells us the -0.903 correlation between gain and density is a strong negative linear relationship. And on average, one unit increase of gain will lead to a -0.001533 unit of decrease for density, with a 0.5497 offset. **However, from the residual plot, we can clearly see a nonrandom pattern, which follows a quadratic function. It suggests there could be a model that can fit the data better.**

Next, we try to fit a linear regression between the transformed data. We apply a log function to the gain variable because of its large range and the overall positive concavity.




```
{'slope': -0.2162032010958118,
 'intercept': 1.2980126052584202,
 'rvalue': -0.9979069608362691,
 'pvalue': 1.8572471194586542e-106,
 'stderr': 0.0014935062316818886}
```

**density = -0.2162 * log(gain) + 1.298**




The linear regression line indicates a correlation of -0.9979, which is a very strong negative linear relationship. And on average, one unit of log(gain) increase will lead to a -0.2162 unit of decrease for density, with a 1.298 offset. From the residual plot, the residuals still indicate some patterns as the log of gains vary. However, with a correlation very close to -1, our guess for the unusual pattern is that because these polyethylene blocks are scattered throughout the region, there exists some other extraneous variables that affect the output gain. Since there is no way to use the same block at different places, we cannot control this unknown extraneous variable.

As for the report accuracy of polyethylene blocks densities, an incorrect report of the density will affect the fit of the regression model in a negative way. Since the densities of the 9 unique polyethylene blocks are reported only once, if there were an inaccuracy for one of the

densities, the following measurements of gains will be tied to the wrong density. Assuming the density from one of the blocks is not reported exactly, one ninth of the data are reported inaccurately. As a result, will negatively impact the goodness of fit for our model. In the context of this case, the calibration we intend to get from the model will be falsey and it will worsen the accuracy of future predictions.

Now we consider the scenario where the blocks of polyethylene were not measured in random order. If we do not collect the data randomly, the data collected will be affected by some potential extraneous that affects the outcome of the response variable. In this case, since the locations where the data collected are from the same region, this regression model is therefore representative of this region, so we can only infer the density from this region using the model. Because there may be other extraneous variables in other regions, the model will be biased and the calibration will be inaccurate. So in general, we want to make sure the data collection process is as random as possible.

## Predicting

Ultimately we are interested in answering questions such as: Given a gain reading of 38.6, what is the density of the snow-pack ? or Given a gain reading of 426.7, what is the density of snow-pack? These two numeric values, 38.6 and 426.7, were chosen because they are the average gains for the 0.508 and 0.001 densities, respectively.

Point estimate:

Our transformed linear model is density = -0.2162 * log(gain) + 1.298. Given two gain readings of 38.6 and 426.7, to get point estimates based on our model, we first calculate their log values and then do an addition of 1.298. And since 38.6 and 426.7 are the average gains for the 0.508 and 0.001 densities, our predictions should be around these numbers within some intervals.

```
Gain: 38.6, Predicted density: 0.508168
Gain: 426.7, Predicted density: -0.011332
```

For a gain of 38.6 we get a snowpack density of 0.5082 and for gain of 426.7, we get a density of -0.0113. For the 38.6 gain, our prediction is very close to the actual density. And for the gain of 426.7, the density we get is -0.0113, which is not logical because the density should not be less than 0. We will investigate if the true label 0.001 is within the confidence interval of this point estimate.

Confidence interval bands:

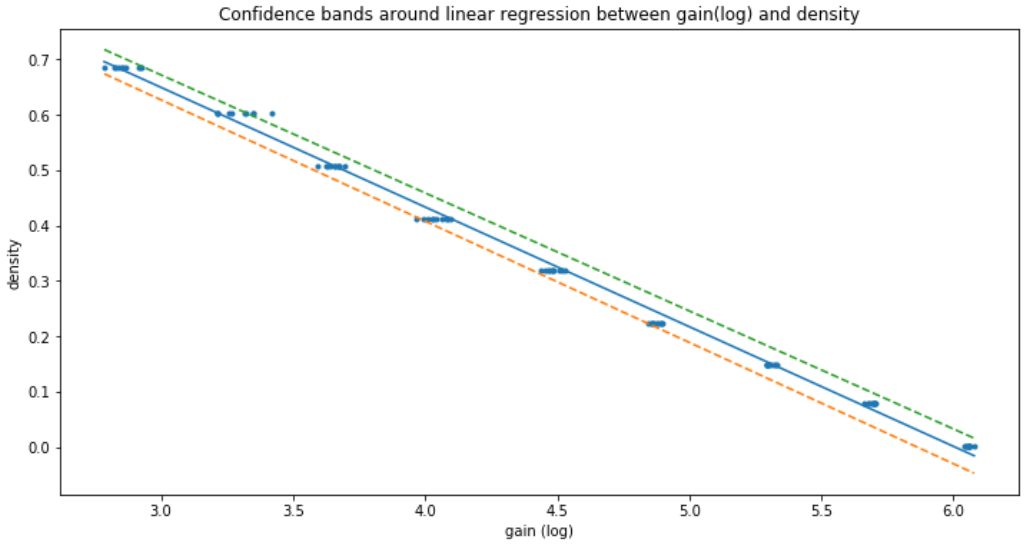
Because there is randomness in our data, we do interval estimates using standard error of the slope and intercept from the transformed linear regression.

```
Call:
lm(formula = density ~ log(gain), data = gauge)

Residuals:
      Min        1Q    Median        3Q       Max
-0.028031 -0.011079 -0.000018  0.011595  0.044911

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.298013   0.006857   189.3   <2e-16 ***
log(gain)   -0.216203   0.001494  -144.8   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01471 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
```

From here we can see the standard error for the intercept is 0.006857, and the standard error for log(gain) is 0.001494. We get the point estimates and add or subtract the standard errors, which are timed by their respective t-values to get the confidence intervals for the whole dataset, that is, $(\widehat{slope} \pm t_{n-2,1-\alpha/2} * SE(\widehat{slope})$ *or* $(\widehat{intercept} \pm t_{n-2,1-\alpha/2} * SE(\widehat{intercept})$. Here the degree of freedom is n-2 because we estimated slope and intercept. We get:

| | **Confidence Interval** |
|---|---|
| **95% ci for slope** | (-0.21917201105784828, -0.21323398894215173) |
| **95% ci for intercept** | (1.2843861533978143, 1.311639846602186) |

Plotting bands along the fitted line with these confidence intervals, we use the first pair (slope, intercept) of values for each interval to get a lower boundary and use the second pair of values to get an upper boundary.



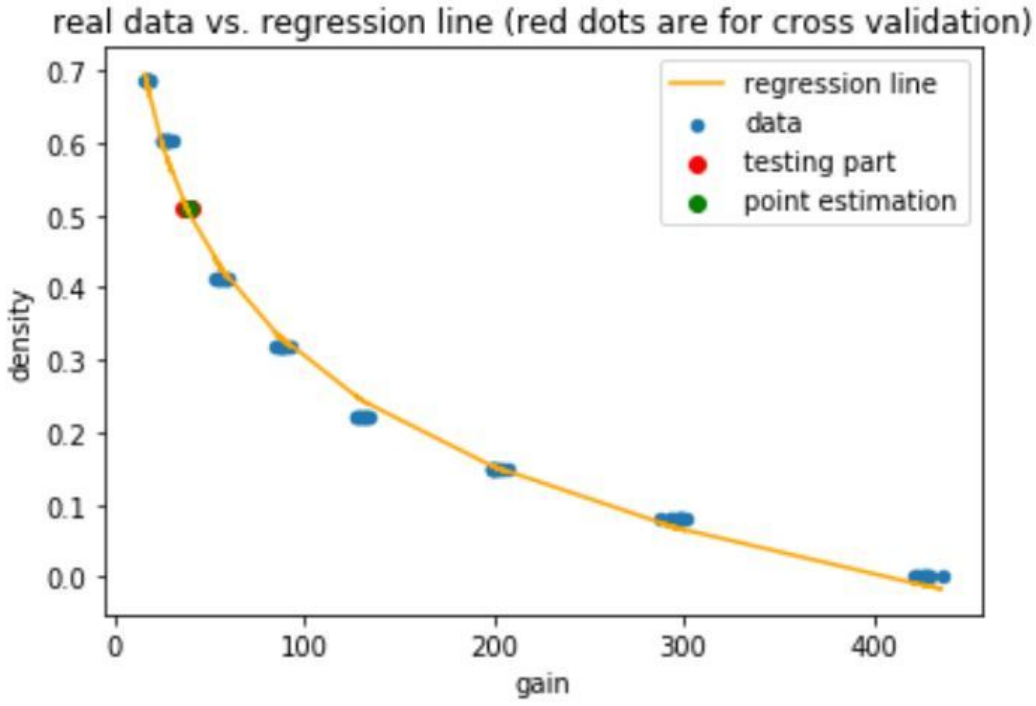

From the graph, we see that only a few of the original data points exceeds the interval, scattering across gain of small values. This is partially because the log function reduce the range

of the same intervals more in higher values, making the higher values of gain to be more squeezed.

### Cross-Validation

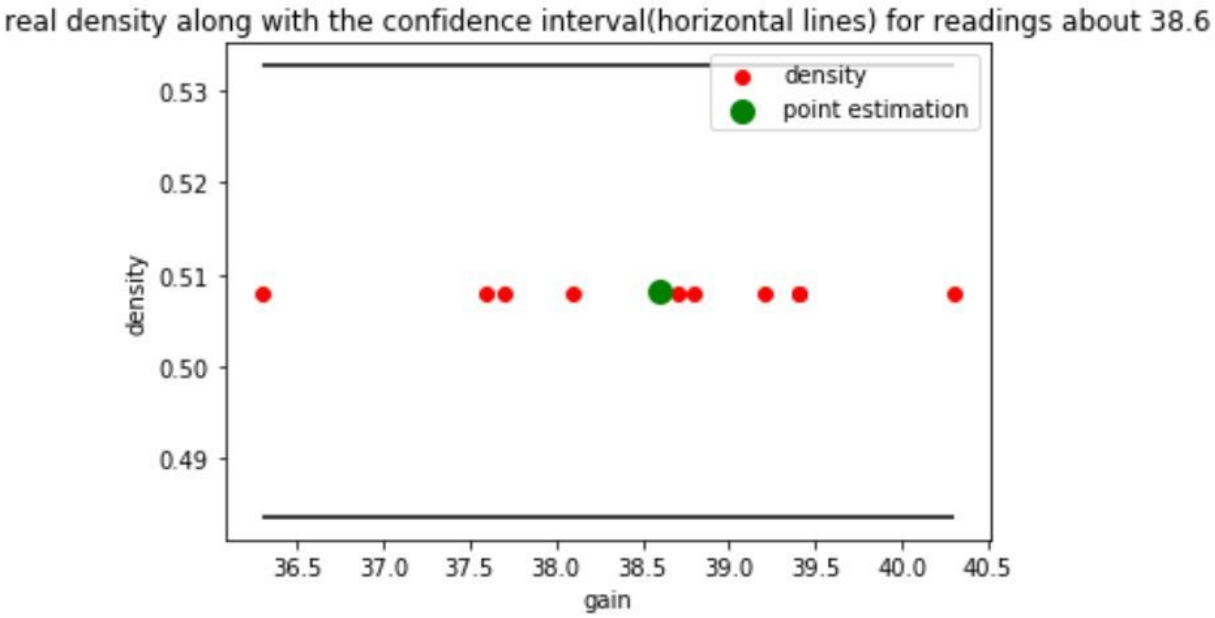
For cross validation test, we first remove the subset of the data where the density is 0.508. Then, we use our x logged linear regression to fit that part of the data, indicated by Figure 3.1.

*Figure 3.1: real data vs. regression line (red dots for cross validation)*




We do a point estimate for those readings that on average are 38.6 by plugging into our linear regression equation. It gives us the prediction 0.5082, which is very close to the true density from the data. Then we calculate the 95% prediction interval, indicated by Figure 3.2.

*Figure 3.2: real density along with confidence interval*



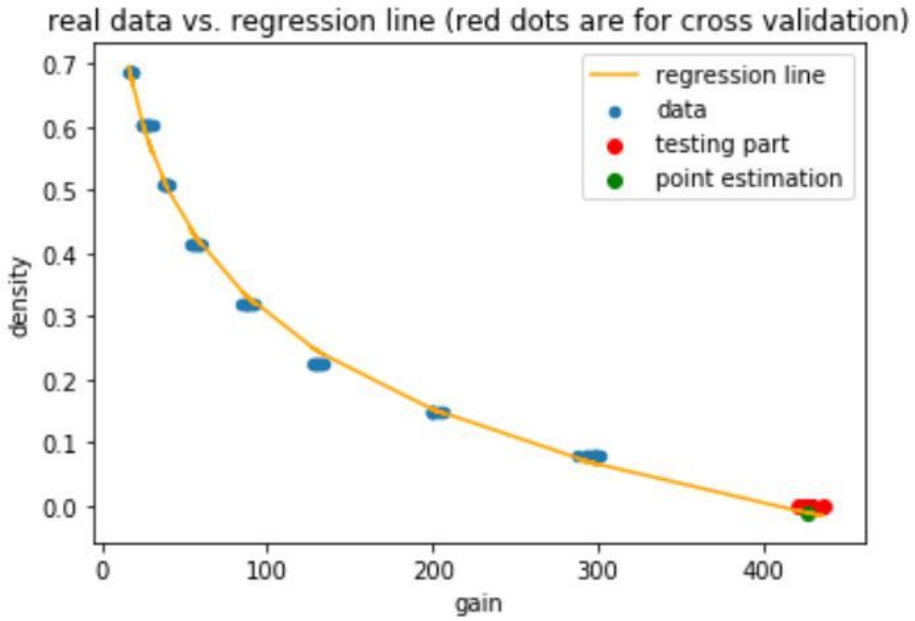

The formula for calculating the prediction interval consists of two parts: slope's confidence interval and intercept's confidence interval. Both come from the equation: $(\widehat{slope} \pm t_{n-2,1-\alpha/2} * SE(\widehat{slope})$ or $(\widehat{intercept} \pm t_{n-2,1-\alpha/2} * SE(\widehat{intercept})$. In this case, since we are using a linear regression to calculate the 95% confidence interval, the t statistic with data size 80 (omitting the testing part) is about 1.991 with degree of freedom 78.

*Table 3.1: Confidence interval for slope and intercept*

| | value |
|---|---|
| **point_estimation** | 0.508169 |
| **95% ci for slope** | (-0.21917732552042124, -0.21322867447957877) |
| **95% ci for intercept** | (1.2843617616509182, 1.311664238349082) |

Degree of freedom 78 comes from the fact that we lose one degree of freedom for each estimator that we used. In linear regression, we have two estimators (slope and intercept), so we deducted 2 from 80. Figure 3.2 shows that the model fits very well for those readings around 38.6. Every data point with density 0.508 is in the confidence interval, and each centers in the middle of the interval. This means all those data are close to our predicted density, so the model works very well in this part. Further on, we do cross validation on another test set whose data has density 0.001 (By Figure 3.3).

*Figure 3.3: real data vs. regression line (red dots for cross validation)*



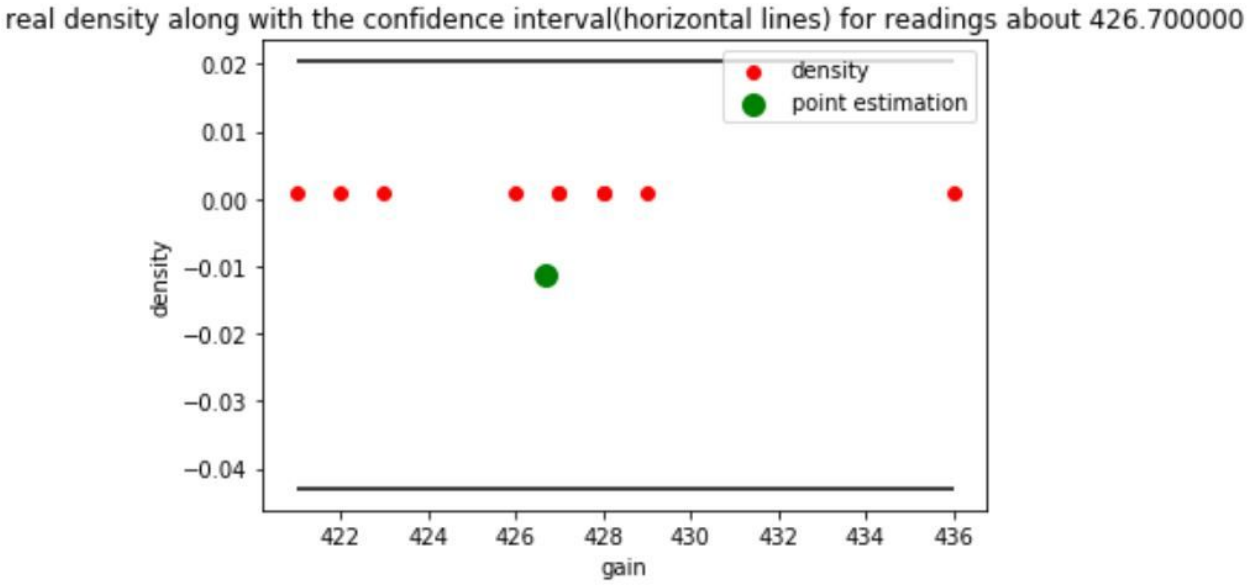

By similar calculation, we could see that the model works not as good as for the 0.508 density's test set.

*Table 3.2: Confidence interval for slope and intercept*

.

| | value |
|---|---|
| **point_estimation** | -0.0113299 |
| **95% ci for slope** | (-0.21917732552042124, -0.21322867447957877) |
| **95% ci for intercept** | (1.2843617616509182, 1.311664238349082) |

*Figure 3.4: real density along with the confidence interval*



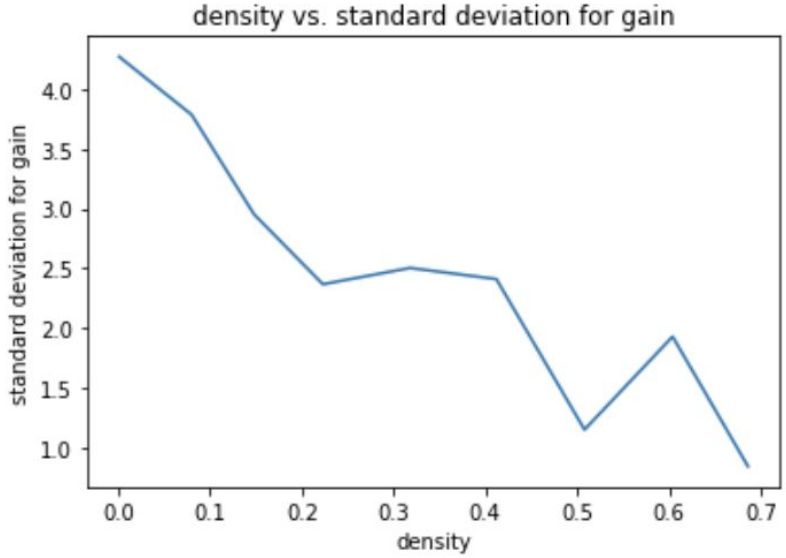

The predicted value underestimated the density (shown in Figure 3.4). From the graph we can see that all those data points are above the point estimation. One explanation might be those data points are at the end of the regression line. Extrapolation might occur. The linear model does not work well here might due to lots of confounding factors. We could have another research on, for example, the location of those readings when the density are at the tails. In that way, we could find out external causes for the increased error in the model. It might be due to the temperature that those machines' working environment or something else. Overall, we should consider this logged gain linear regression model as a good one, because most of the data could be explained by this model well.

## Advanced Analysis

To further investigate the relationship between gain and density, we are interested in the precision, or closeness, of measurements across different densities. Therefore, our question is: how does the precision of gains change along changes of snow densities?

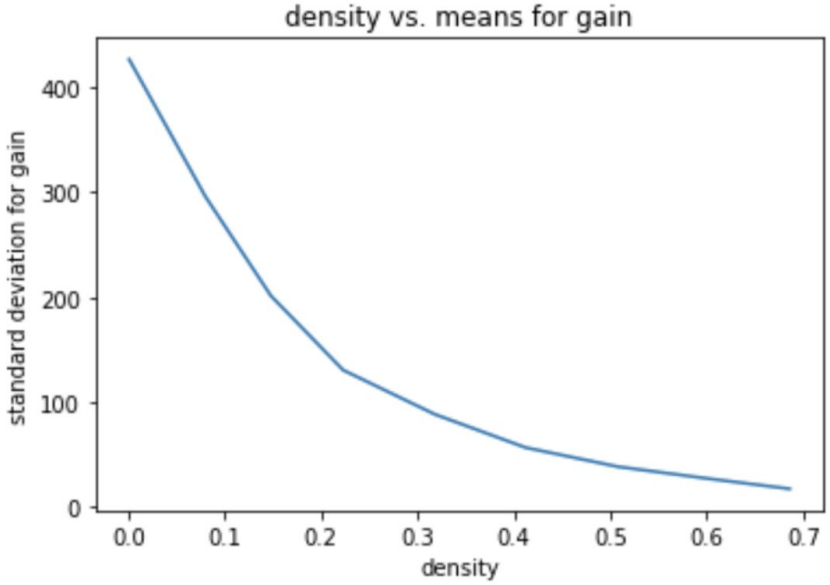
To measure the precision of measurements, we first investigate the standard deviations of gains corresponding to each density. From Figure 4.1, we can see that standard deviation tends to decrease as density increases. This means readings are “closer” to each other when gauges are measuring snow of higher density.

*Figure 4.1: Density and standard deviations of gains*




However, the standard deviations may not tell us the whole story. It is possible that standard deviations are smaller because the values of measurements are smaller! Therefore, it is necessary that we take a look at the actual values of gains across different densities. As shown in Figure 4.2, we can see that gauge readings indeed decrease as density increases. Under such circumstances, measurements of absolute deviations like standard deviations might not be a good tool for finding the true relationship.

*Figure 4.2: Densities and mean values of gains*




Therefore, we decide to analyze the precision using relative deviations - percent deviations. For each value of gain, the deviation is measured by calculating what percent of the value deviates from the mean. Using this strategy, we can do a better job in getting rid of the

effect of the magnitude of measurements. Calculating the mean percent deviation of gains for each density and fit the regression line using least squares, we have Figure 4.3 and 4.4.

*Figure 4.3: Least squares regression between density and mean percent deviation*

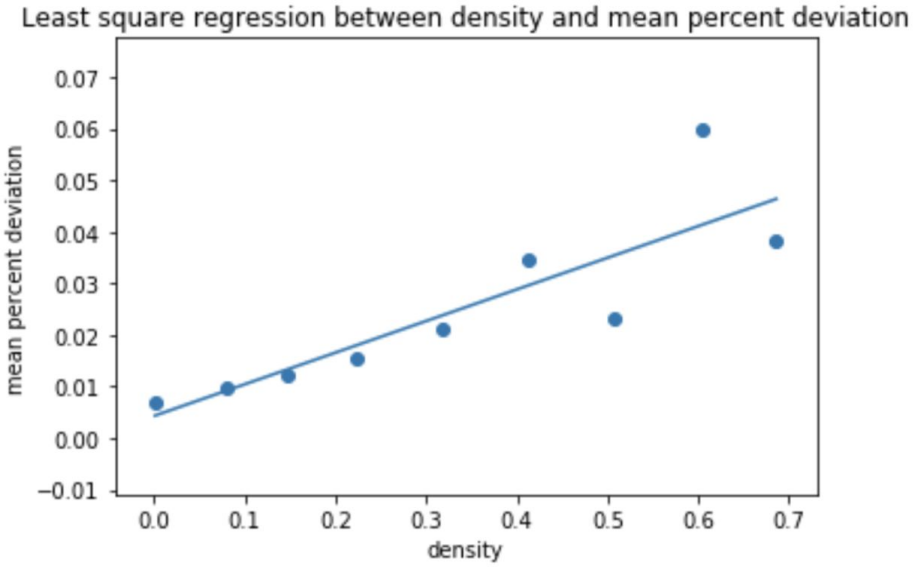
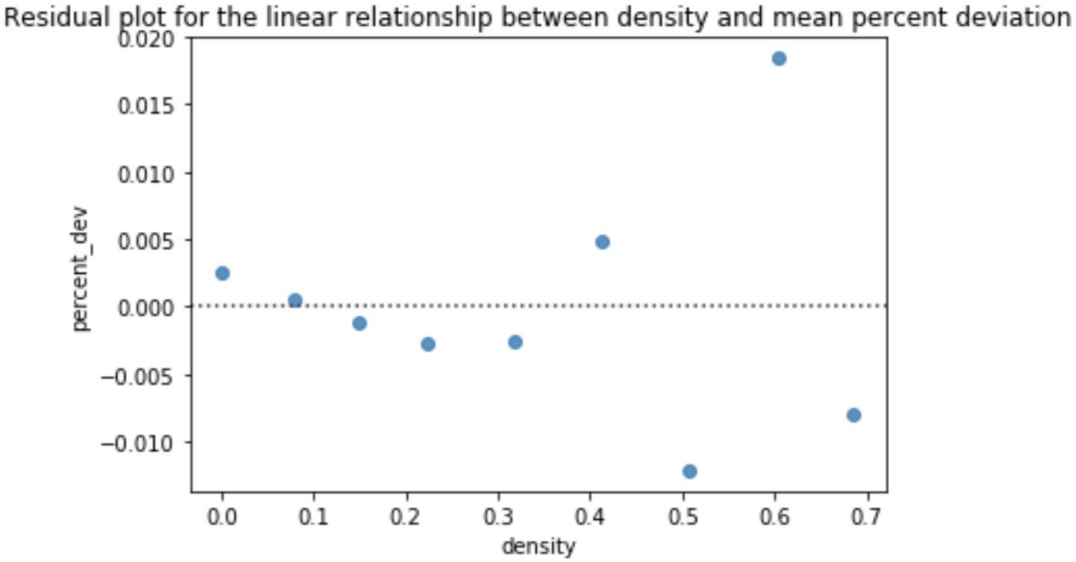



*Figure 4.4: Residual plot for the linear relationship between density and gain*




We can see that in the residual plot, there is no obvious pattern so the linear regression might be a good model. For the fitted line, slope is 0.061 and intercept is 0.004. This means for every 0.1 increase in density, we would expect the mean percent deviation to increase 0.06%. Also, R-squared is 0.742, which means 74.2% of the variability of mean percent deviation can be explained by density.

Therefore, there is still a positive relationship between density and precision of measurements. The reason might be that gamma rays that are sent to the detector may be

scattered or absorbed by the polyethylene molecules between the source and the detector. Because higher reading of gain means higher gamma photon counts, if there are more photons transmitted from the source to the detector, with the number of polyethylene molecules in the air equal, the number of photons that get absorbed or bounced will have a higher variability, which leads to more volatile readings.

## Further Discussion and Conclusion

Through this case, we could see the broad applications for linear regression. Using data given by the machines' readings, we could find a linear relationship with the actual density of the snow gauge. In addition, we could even predict a rough range of density if given another new set of readings for gain. Although there would be some limitations or errors through linear regression, by looking at the cross-validations that we did, we could see that linear regression is still stable to give us a value that is close to the truth. Linear regression is powerful, but we have to keep in mind avoiding extrapolations and be mindful about other confounding variables that could affect the afterall result.

## Method

*Linear Regression:*

Linear Regression is a statistical tool to model the relationship between a scalar response and one or more explanatory variables. If there is one explanatory variable, the fitted line can be represented by $y = \beta_1 x + \beta_0$. The best regression line is found using least squares, that is, minimizing $\sum(\beta_1 x_i + \beta_0 - y_i)^2$, where $x_i$ and $y_i$ are what we observed in our data.

## Theory

*R-squared:*

R-squared is a statistical measure of how close the data are to the fitted line. The definition of R-squared is the percentage of the response variable variation that is explained by the linear model. That is, R-squared = Explained Variation / Total Variation. Thus R-squared is a value between 0% and 100% where 0% indicates that the model explains none of the variability of the response data around its mean and 100% indicates that the model explains all of the variability of the response data around its mean. In general, the higher the R-squared, the better the model fits the data.

*Residual Plot:*

A residual plot is a graph that displays the residuals on the vertical axis and the independent variable on the horizontal axis. From the residual plot, we can determine whether a linear regression model is appropriate for the data. If points in the residual plot are randomly dispersed around the horizontal axis, then the linear regression model is appropriate for the data; otherwise, a non-linear model might be a better choice.

## Work cited:


Reyes, J, et al. "Neural networks to predict earthquakes in Chile." Applied Soft Computing, February 2013, https://www.sciencedirect.com/science/article/pii/S1568494612004656

Wei Yong, et al. "Inverse Algorithm for Tsunami Forecasts." Journal of Waterway, March 2003, https://ascelibrary.org/doi/abs/10.1061/(ASCE)1527-6988(2000)1:2(107)?casa_token=kB Au2Jtv4oYAAAAA%3A8ATWd1f4Lx8e74hoXuVxk6AoeOG3Nys0VIXaSqHAT_-bs9t ZtuRW3gDR53wB_IcdSsR18QAYPg&

Casey, Kenan, et al. "A Sensor Network Architecture for Tsunami Detection and Response." International Journal of Distributed Sensor Networks, January 2008, https://journals.sagepub.com/doi/abs/10.1080/15501320701774675