Rick Boelen

Hwang Min Yu

Wen Zeng

MATH 189

05/24/2020

**Homework 4 Data Analysis Report**

# 

# **Introduction**

The Sierra Nevada is a major source of freshwater for most of Northern California. In order to monitor water supply, the Forest Service of the United States Department of Agriculture (USDA) uses a gamma transmission snow gauge in Central Sierra Nevada. This gauge is used to determine the death profile of snow density. However, this gauge may not be entirely accurate as time goes on, due to instrument wear and radioactive decay. Therefore, the goal of our analysis is to develop a procedure to calibrate the snow gauge.

The snow gauge uses radioactive data to make density readings. A radioactive source emits gamma rays in all directions. An energy detector about 70cm away from this source uses a scintillation crystal in order to catch and count the number of photons emitted from the source. The photon pulses are then processed and converted into a measure called “gain,” which should be directly proportional to the emission rate of the source. These are the two variables we will be using in our regression analysis.

In order to calibrate the instrument, polyethylene blocks of 9 different densities are placed in front of the sensor in order to simulate snowpack. The gauge’s readings are then recorded, and this is the data that we are working with.

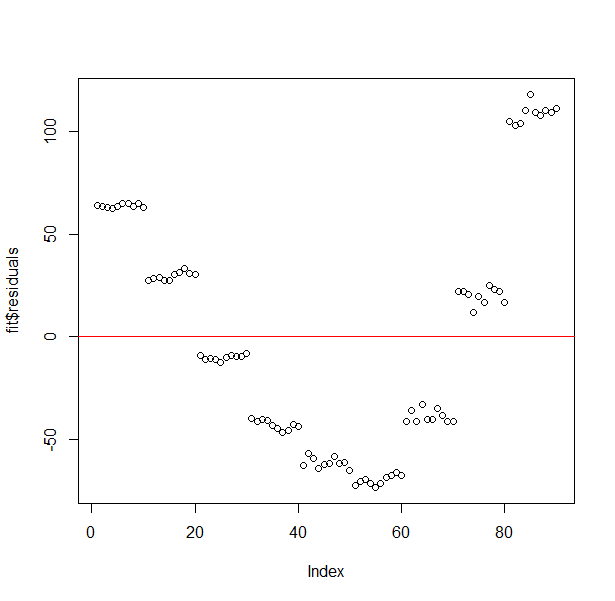
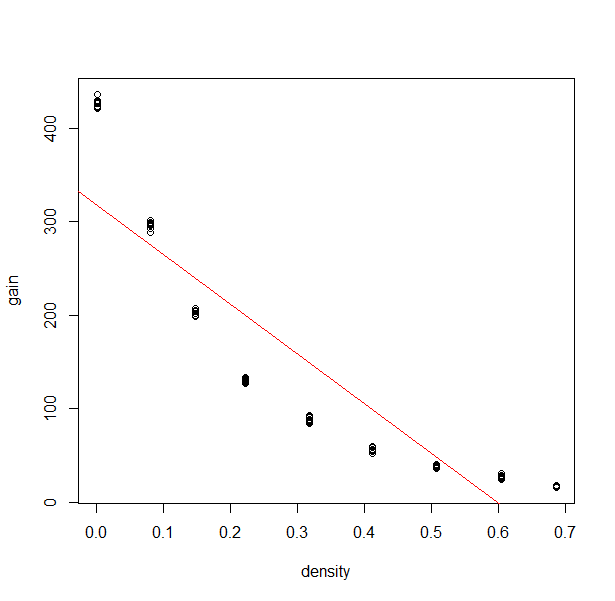
One goal of this analysis is to find a good model to calibrate the gauge to. Our process will be looking at the data using regression. By calibrating the gauge to this model, we should be able to predict the density of the snow-pack given a gain value reading.

Data framing issues with the dataset is the fact that there’s possible wear and tear with the instrument usage and this is hard to take into account within our regression. All we can assume is that there will be possible bias from this challenge. This crosses boundaries between geolocations due to the different advancements in equipment for different locations, we have to keep this in mind when developing our conclusion of generalizing a calibration. Moreover, the timing of measurement changes the values due to conversion changes. One other aspect is the fact that only the 10 middle values of each 30 observations sampled for each block was used in the dataset. Firstly, this leads to possible biases in variance when utilized and also we wouldn’t know how extreme the actual range of these values are.

# **Methods/Analysis**

*Fitting:*

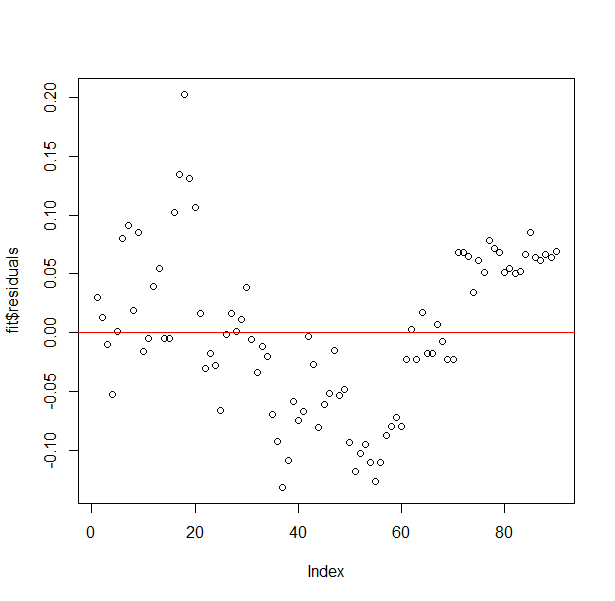
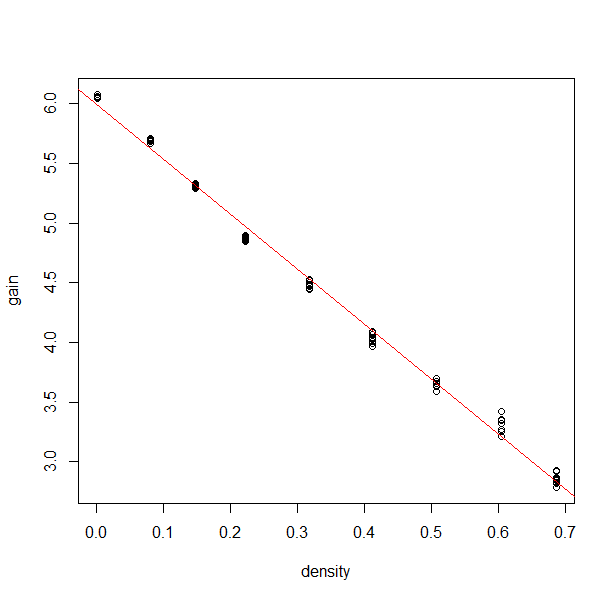
*Method:* We plotted a linear regression of snow gain vs. snowpack density. Then we got a residual plot to see if the data was a good fit or not. We further analyzed



*Analysis:* Our linear regression shows a negative relation with a slope value of -403.3. However, our linear regression demonstrated that we were violating a condition with forming a linear regression model on our two variables. There was not in fact a linear relationship between gain and density. This can be demonstrated through our residual plot giving a quadratic-like shape and not showing signs of being randomly scattered. The quadratic form of the residuals shows that through our observation there was some form of exponential movement in the data. Therefore, we continued looking into creating a transformation on gain that would suffice a linear relation.

*Transformation:*

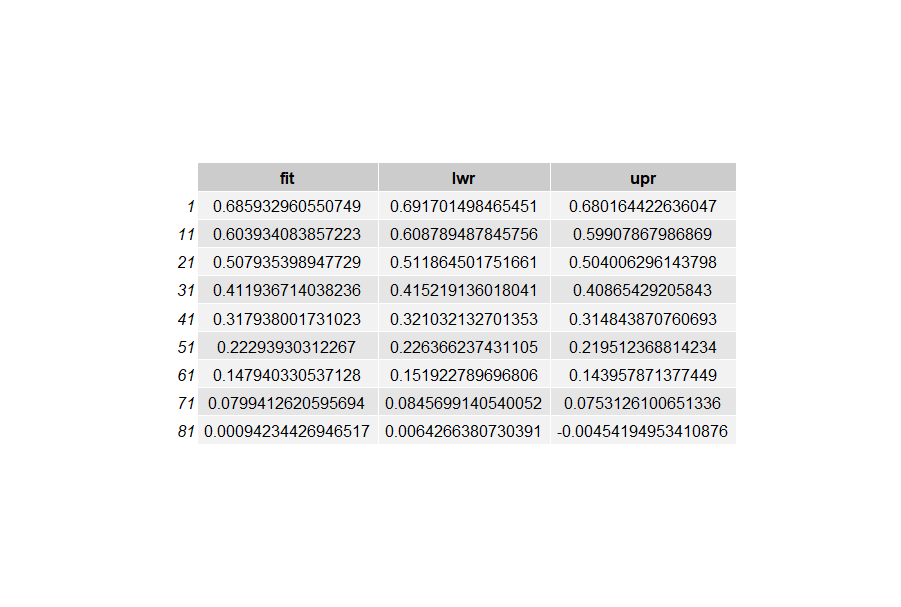
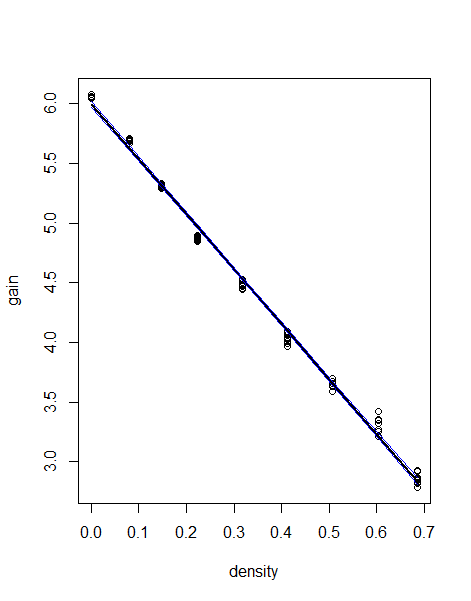
*Method*: There’s a given relation that showed that there was a relation between the number of molecules in line with the detector and the density given a formula relating to the probability of it not getting detected. We utilized this formula and given relations between gain and density to come to the conclusion that the logarithmic function of gain is proportional to density. This can be further clarified in our appendix (1). From there we transformed our linear regression to fit this model and again took a linear regression and a residual plot.



*Analysis*: We found that there was a strong negative linear relation with this transformation and could conduct linear regression. Our slope value showed a value of -4.606. The residuals show closer similarities to a normal distribution than before. (2.2) However, one can see that the densities within the middle have negative residuals, while the outer values have positive residuals. One reason for this could be the fact that there was a mismeasurement of the density values within that group and therefore lead to bias within the middle groups. One aspect that needs to be taken into consideration is whether the measurements were all taken at the same time, as that could lead into different conversions as stated in the introduction.

*Prediction:*

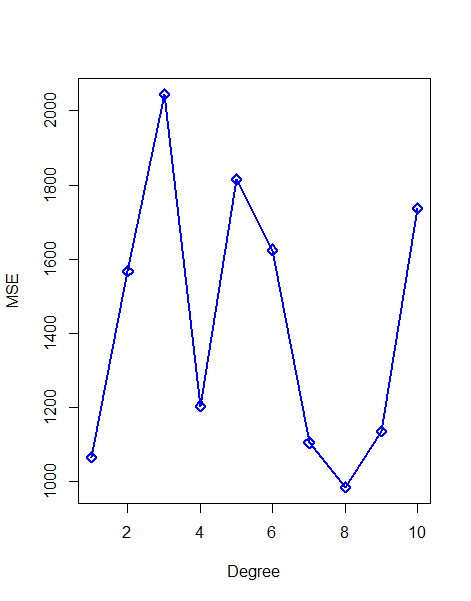
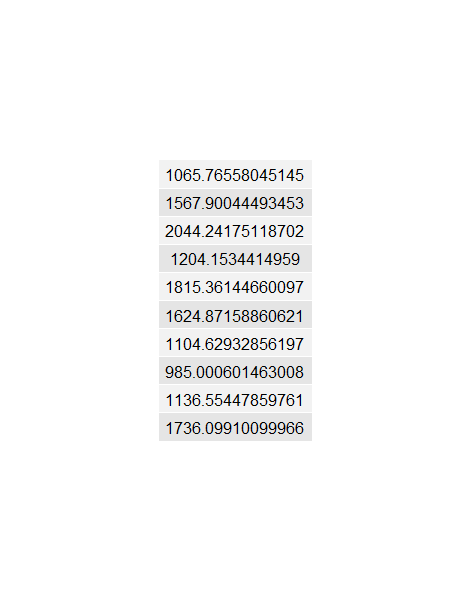
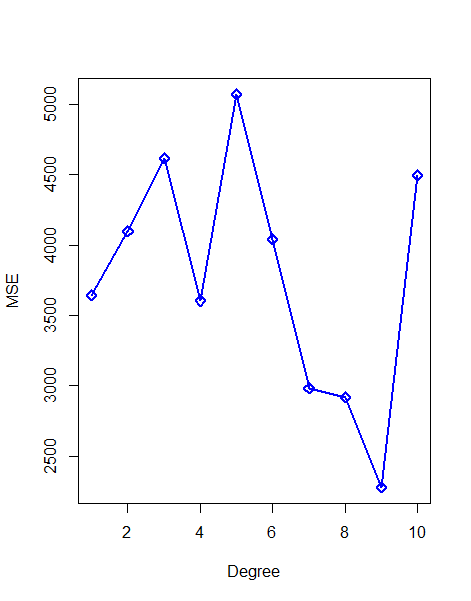
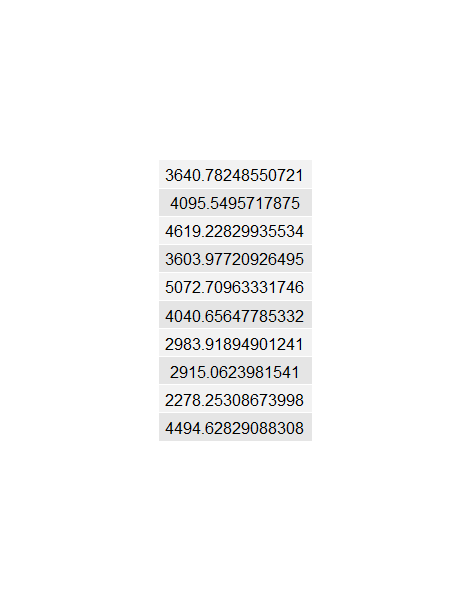
*Method*: From there we developed a procedure to add confidence intervals in order to gain insight into the values that would be produced from our linear regression model. We decided to implement confidence intervals of 95% confidence rather than prediction intervals due to the fact that prediction intervals are used to take into account the sampling of individual observations. We used confidence intervals of gain and produced intervals for density based on inversely reworking the linear regression equation. However, the experiment was conducted to measure these samples multiple times, 30, and this takes that into account and therefore thought it would be suitable to use confidence intervals to gauge a more accurate value.



*Analysis*: This was conducted through our linear regression fit that used the logarithmic function of gain over the density. The confidence intervals are shown in our graph in line with our fitted model; however, these intervals are too small to see visually and this mostly comes down to the fact that there were replicated measurements of the same observation of the density of a block. Our confidence intervals gave a percentage error of 0.84% approximately. Looking at specific values, such as a gain value of 38.6 and 426.7, both would be within the confidence intervals.

*Additional Analysis:*

*Method:* We wanted to see how well our procedures of confidence interval works, so we decided to do a cross validation to check. We used the k-fold cross validation with k = 10. We wanted to compare the mean squared error value of when we omit rows that have density values of either 0.508 or 0.001. We created two training sets where one has omitted the rows of block of density 0.508 and the other has omitted the rows of block of density 0.001. The first graph below is the result for the mean squared error of samples for the training set of omitted rows for block of density 0.508 within the 10 folds in cross validation. The second graph below is the result for the mean squared error of samples for the training set of omitted rows for block of density 0.001 within the 10 folds in cross validation. We then compared the average mean squared error values of each sample set from the 10 folds to see which has the lower value to get a better model.

**

*Analysis:* Our sampling has found that for the training set of omitted rows for block of density 0.508 resulted into the first table above. Our sampling has found that for the training set of omitted rows for block of density 0.001 resulted into the second table above. The average of the mean squared error for the table of omitted rows for block of density 0.508 resulted in 3774.477. The average of the mean squared error for the table of omitted rows for block of density 0.001 resulted in 1428.458. Since the average mean squared error of 1428.458 is lower than 3774.477, the sample set of omitted rows for block of density 0.001 is the better model.

# **Conclusion**

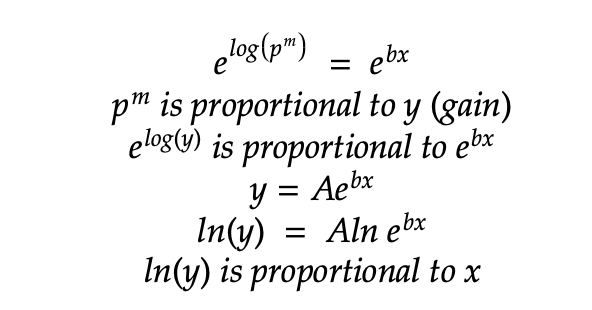
Our objective for this analysis was to find a measurement conversion from a measured gain to the snow-pack density. Firstly, we decided to look at linearly regressing the two variables together, however, our results show a violation of the relation being non-linear. Therefore, we transformed the plot based on proportional relations between the two variables and from there conducted confidence intervals reading our results inversely to find intervals for density based on given gain. Our results showed that given a certain gain we can make an estimate with an uncertainty level of 0.84%. For further research with this calibration method, constant monitoring should be held on their own dataset to check whether the linear regression model still holds a two-fold condition of constant variability and linear relation. If not, this calibration would not apply to their own dataset.

However, other observations showed that the residuals were not completely scattered based on the x axis and this could be due to the measurements being taken at different times leading to different conversions. Moreover, it is important to look at the technology used for this experiment as if measurements were taken again with other instruments, it could lead to a result bias. Weather impacts, such as temperature and pollution, could have a big factor when measuring density of snow.

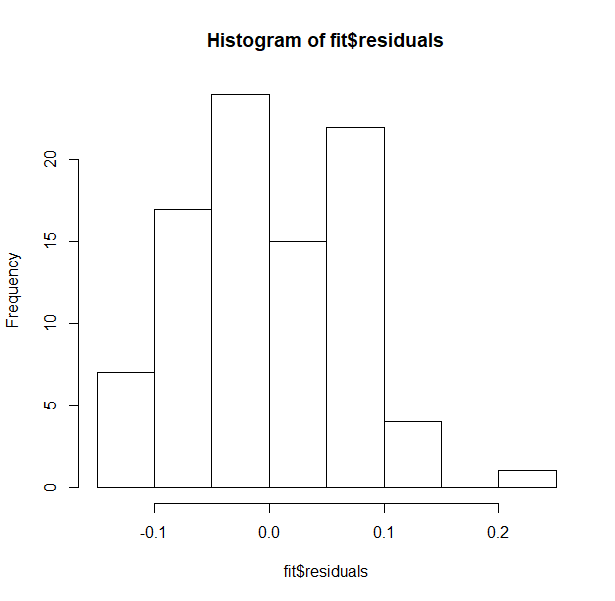
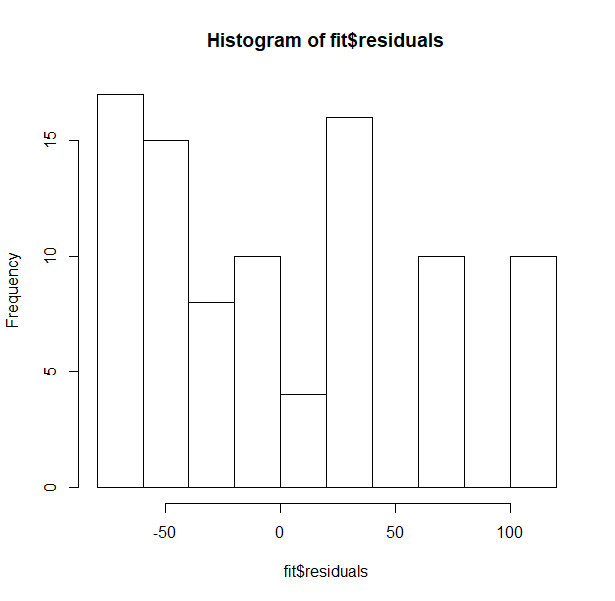
For future analysis, it would be insightful to use instrumental variables in our analysis for variables such as age of equipment or atmospheric temperature. This could be really beneficial in analyzing a more causal relation between the two variables itself and to take into account any possible biases.

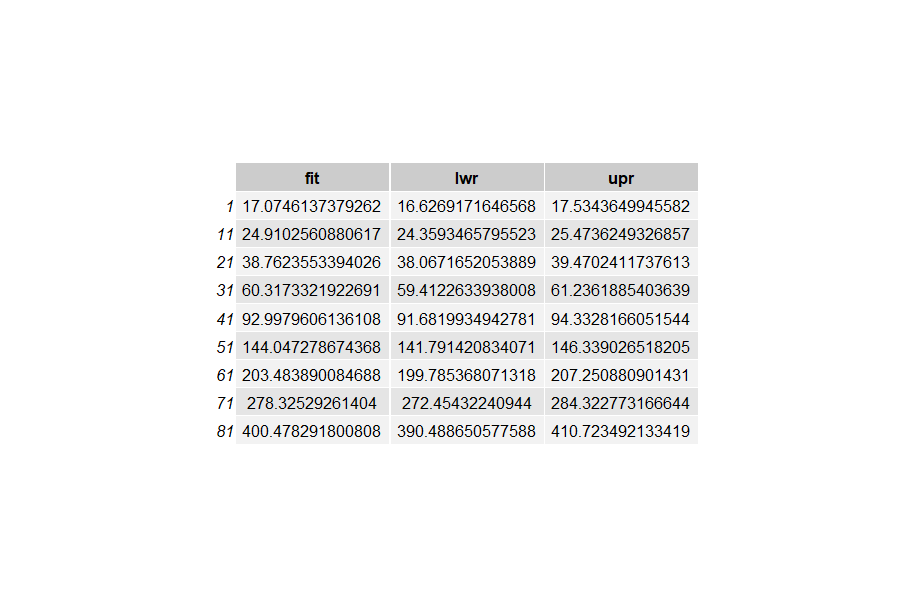
# **Appendix**

1 Relation between gain and density



Histogram of residuals from linear regression of 2.1 x and y 2.2 x and ln y

3 Confidence intervals for gain given a measured density.



**Contribution Statement**

Introduction: Wen

Analysis Writeup: Rick/Hwang

Analysis: Everyone

Conclusion: Rick

Code: Hwang

# 