儿 Case Study 4: Calibrating Snow Gauge

Yu-Chun Chen, A13356506
Yanyu Tao, A13961185
Bolin Yang, A92111272
Shuibenyang Yuan, A14031016
Haoming Zhang, A14012520
Mingxuan Zhang, A13796895

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. The gauge is used 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 this 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, the snow absorbs the water up to a certain point, after which flooding occurs. The denser the snow pack, the less water it can absorb. Analysis the snow pack profile may help with monitoring the water supply and flood management. The gauge does not directly measure snow density. The density reading is converted from a measurement of gamma ray emissions.

In this case study, we are going to search for calibrate the snow gauge in order to measure the accurate snow density readings by using the gamma ray emissions. Because instrument wear and radioactive source decay, there may be changes over the seasons in the functions used to cover the measured values into density readings. To adjust the conversion method, a calibration run is made each year at the beginning of the winter season.

Before we go into our own investigation, some literatures are reviewed. Kenneth J. Condreva states that in order to forecast more accurately snow runoff and allocate resources appropriately, the new invention using the secondary cosmic gamma ray is generated, the fundamental technology in snow gauge. Because some areas are difficult to access snow, the new invention is designed to operate remotely to determine the water equivalent of snow by telemetry system to accumulate date as a function of time and transmit (Condreva). In “A Cosmic-Ray Snow Gauge,” the author indicates the examination of the cosmic-ray snow gauge to show the effectiveness of it. One pair of type-A detectors has been placed during the snow season. One is inside the building and another one is on the outside ground. The result is that the water equivalent values (W) of two detectors are same amount, which means that W values are primarily influenced by the cosmic radiation (M. Kodama, S. Kawasaki, and M. Wada). Davide Bavera and Carlo De Michele investigate the spatial distribution of snow water equivalent (SWE) over a mountain basin at the end of the snow accumulation season using a minimal statistical model (SWE-SEM). Firstly, they calculate the local SWE estimates at snow gauges, then the spatial distribution of SWE using an interpolation method. And then using the linear

regression of the first two order moments of SWE with altitude. This method is applied to the Mallero basin, which can also be used in our lab.

Data:

The data 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 of known densities between the two poles of the snow gauge and taking readings on the blocks. The polyethylene blocks are used to simulated snow. For each polyethylene blocks, 30 measurements are taken. Only the middle 10 are reported here. The measurement reported are amplified version of the gamma photon count made by the detector. We call the gauge measurement the "gain". The data available here consists of 10 measurements for each of 9 densities in grams per cubic centimeter of polyethylene.

density	gain
0.6860	17.60
0.6860	17.30
0.6860	16.90
0.6860	16.20
0.6860	17.10
0.6860	18.50
0.6860	18.70
0.6860	17.40
0.6860	18.60
0.6860	16.80
0.6040	24.80
0.6040	25.90
0.6040	26.30
0.6040	24.80
0.6040	24.80
0.6040	27.60
0.6040	28.50
0.6040	30.50
0.6040	28.40
0.6040	27.70

Figure 1: Data Set

Our data also exists challenges with procuring a large quantity and diversity of snow gauge data beyond the scope. First challenge is operational networks, which is also called knowledge base. The networks have declined in the northern regions, including Siberia, Alaska and N. Canada, and the mountain regions. The figure below clearly shows the sparseness of the networks. And then the question is to sustain and improve the operational networks.

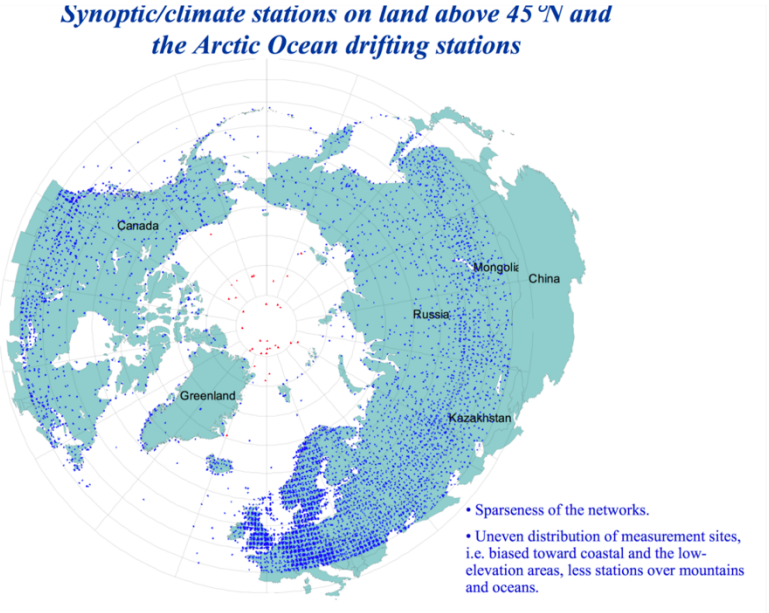

Figure 2. Sparseness of the Networks

The second challenge is the data quality and compatibility across national boundaries. There are large biases in gauge measurements of solid precipitation. And the precipitation data is incompatible due to difference in instruments and methods of data processing, which is also difficult to determine precipitation changes in the arctic regions.

However, there is a way to solve the challenge, which is validation of precipitation data, including satellite and reanalysis products and fused products at high latitudes.

Background:

The snow gauge is a complex and expensive instrument as the research tool instead of monitoring the water supply. The snow gauge has the broad functions in snow-pack settling, snow-melt runoff, avalanches and rain-on-snow dynamics. Gauges are used in many locations, like Idaho, Colorado, Alaska, Russia, Mongolia, China, Japan etc. The California gauge, approximately diameter 62 meters and elevation 2099 meters, is in the center of a forest opening. The snow pack reaches an average depth of 4 meters each winter (Bradic).

The snow gauge is formed by a cesium-137 radioactive source and an anergy detector mounted on separate vertical poles approximately 70cm apart. The radioactive source emits gamma photons at 662 kilo-electron-volts (keV) in all directions. And a scintillation crystal in detector counts those photons from the source to the detector crystal. Photons can generate pulses transmitted by a cable to amplified and transmitted via a buried coaxial cable to the lab. In the lab, the stable signal has the function of correcting for temperature drift and converting to a

measurement called “gain”. And The snow pack density typically ranges between 0.1 and 0.6 g/cm^3(Bradic).

The gamma rays sent in the direction of the detector will be scattered or absorbed by the polyethylene molecules between the source and the detector. The denser polyethylene, the fewer gamma rays reaching the detector. The physical model of the relationship between the polyethylene density and the detector readings is complex. The simplified one is that a gamma ray on route passes a number of polyethylene molecules, depending on the density of the polyethylene. A molecule can either absorb it or allow it to pass. If each molecule acts independently, the probability of gamma ray reaching the detector is p^m, where p is the chance that the molecule will not absorb or bounce the gamma ray, and m is the number of molecules in a straight-line path from the source to the detector. So the probability function is that $e^{m \cdot \log(p)} = e^{b \cdot x}$, where x the density (Bradic).

Investigations:

Scnario 1: Fitting

Before we make scatter plot or do regression to visiliaze the dataset, we want to discuss the nature and potential problems of our dataset. Since our data all come from measurement, it definitely contains some errors in measuring. There could be two major problems.

First, the densities of the polyethylene blocks might not be reported exactly. This could due to man-made recording error or the wear of our equipment. No matter what causes it, it will have a great effect on the regression line that we generate to predict rain fall. Specifically, it would make our predictions for density from the gain inaccurate and resulting in a departed regression line from the true regression line, which could generate a far cry value from the true values that associated with the specific density. And due to our small sample size (only ten replicate records for each unique polyethylene block), the effect of incorrect densities will be huge, as it would not be treated as outliers or be observed during the regression time.

Second, the blocks of polyethylene might not be measured in random order. According to the Sparseness of the networks graph above, there is a geological relationship between each site of polyethylene. As a result of that, there must be a relationship between the recordings of the gains and the locations of the blocks of polyethylene. Therefore, the regression line that we generate might not able to generalize to the area that we want it to prediect. Since our data are from a calibration run of the USDA Forest Service’s snow garuge located in the Central Sierra Nevada mountain range near Soda Springs, if the data only comes from a small portion of the area without randomly selected, the regression model that we generate would not be able to correctly predict the densities from the gain of the polyethylene blocks. In conclusion, if the data or the blocks of polyethylene are not chosen in random order, the model would be biased.

Thus, it is important to check if two potentinal problems could occur before we make any analysis or do regression on it. For the rest of the scenario 1 or the following scenarios, we assume that our dataset does not have such problems, and ready to precede any analysis.

Our task is to fit the gain of the nine polyethylene blocks to the measured densities in our dataset. Before we make any regression plot, we first make a scatter plot on the 10 data for each different density of polyethylene. The following figure 3, scatter plot, has "gain" as the explanatory variable, and "density" as the response variables. From the overall trend, the plot shows that the data does not have a linear relationships, thus it is not appropriate to use the linear regression model to fit our data.

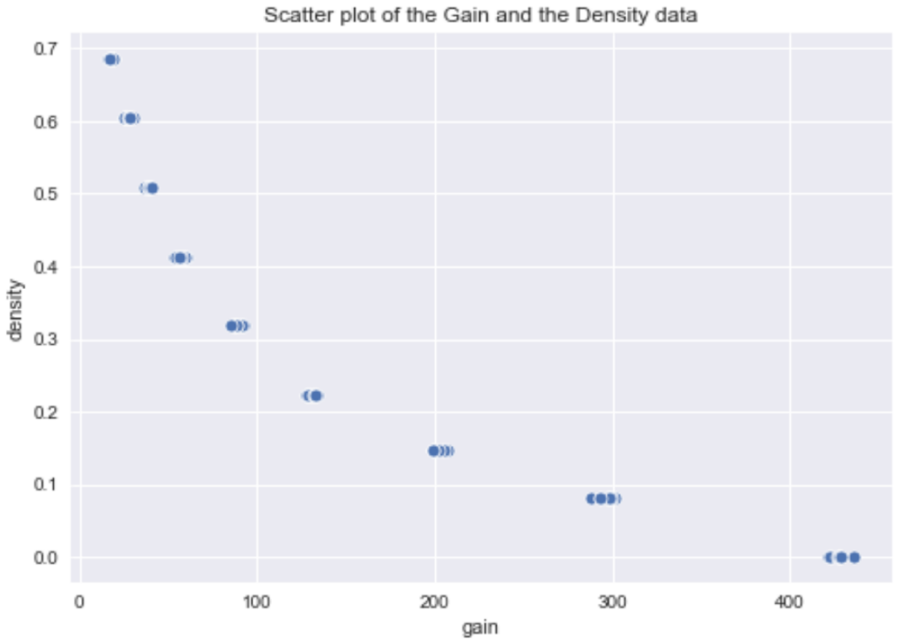

Figure 3: Scatter plot of the Gain and the Density data

Even the plot shows that it is not linearly correlated, we observe that there is a exponential relationship between the variables. Thus, we make a logarithmic transformation of the "gain" variable to further explore the relationship between the variables. The figure 4 is the scatter plot of the Density and the log-transformed Gain data. We observe that there is a linear trend in the plot, indicating that the there might be a linear relationship between the density and the log transformed gain. Thus, we decide to use linear regression model to fit the data.

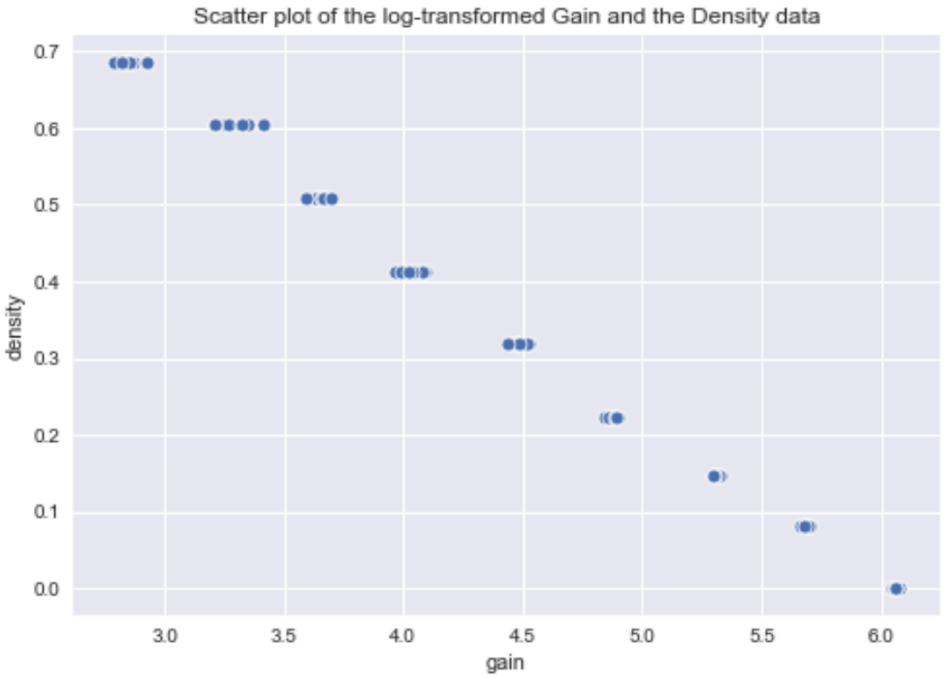

Figure 4: Scatter plot of the log-transformed Gain and the Density data

Since in the dataset, we have multiple values for gains for the same density. However, linear regression only allow one to one relationship, thus we compute the mean value for each density, and again we plot a new graph for the density with mean log gain in figure 5, just for visualization. (Noted, we still use the whole dataset for the latter regression part.)

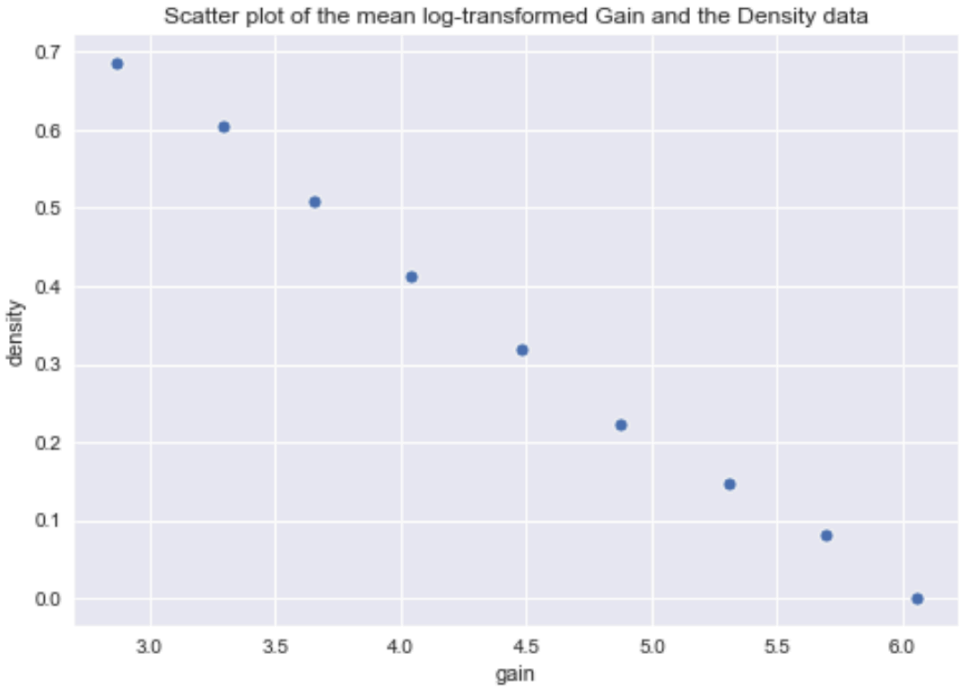

Figure 5: Scatter plot of the Density and the mean log-transformed Gain data

Now we are ready to fit our dataset with linear regression model to predict density from logarithmaically transformed gain. In order to make sure that our model is correct, we use two different models on the dataset, one is least square regression line, and least absolute deviations regression line. From figure 6, we can see that the transformed data successfully fit into the linear model for both linear models. The intercept and the slope are similar.

Least Absolute Deviation Regression Line: density = -0.2155 * log(gain) + 1.2955
Least Square Regression Line: density = -0.2162 * log(gain) + 1.2980

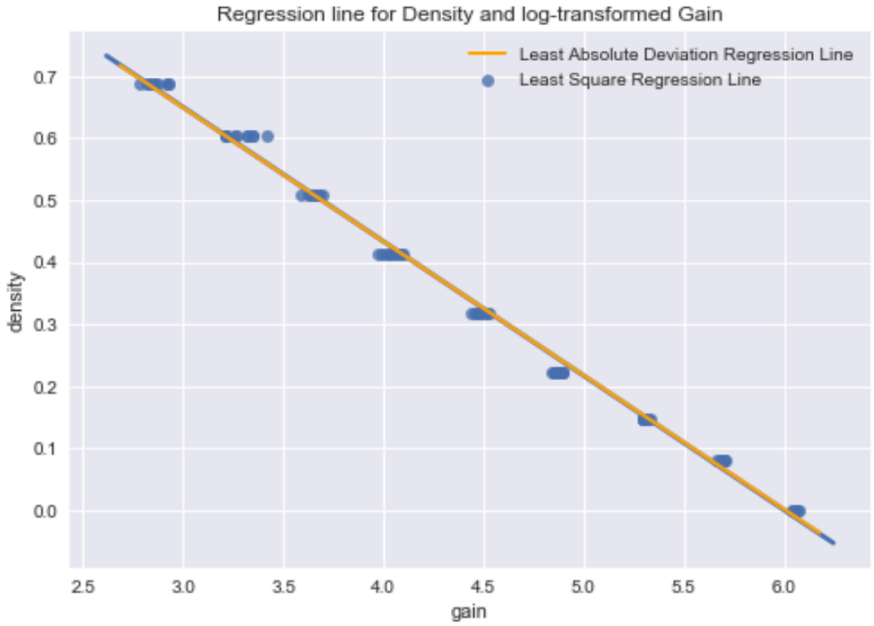

Figure 6: Regresson plot of the Density and the log-transformed Gain data

Next, we want to check conditions to make sure that the linear model is appropriate. We are going to check the linearily, nomality and the homoscedasticity of our model.

We want to make sure that the data fit the linear model well, the linearily of the graph and homoscedasticity of the residual are needed to be checked. To check linearily of our model, we plot residual plots for both linear regression models. In figure 7 and figure 8, we obseve that the residuals stay close to the middle, and there is no significant outliers, indicating that the two linear model indeed fit our data well. Even though there is small trend in the residual plot, it is not strong, thus we conclude that variability of residuals around the 0 line appears roughly constant as well. Therefore, our linear model follows homoscedasticity.

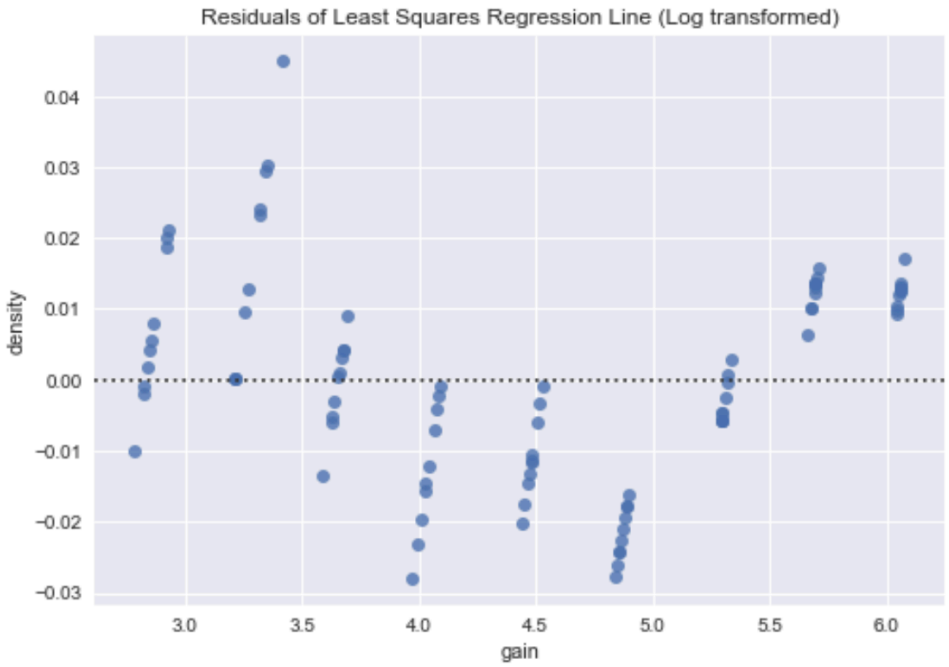

Figure 7: Residuals of Least Square Regression Line (Log transformed)

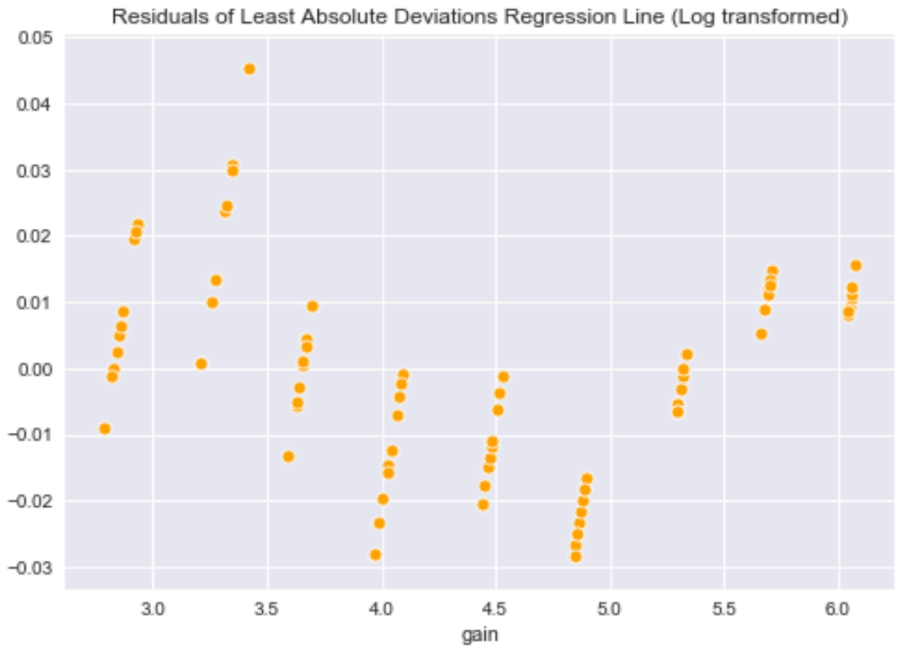

Figure 8: Residuals of Least Absolute Deviations Regression Line (Log transformed)

The final thing we want to check is the normality of the residuals, the figure 9 and 10 are the Histogram and QQ plots of residuals for both Least Square Regression and Least Absolute Deviation Regression. From figure 9, the distribution of the residuals looks normal, which is

supported by the qq plots in the figure 10. It shows that both qq plots line stick to the normal line quite well, without any shapes. Thus, we conclude that the linear model is indeed the correct model to use fit log gain to density in our dataset.

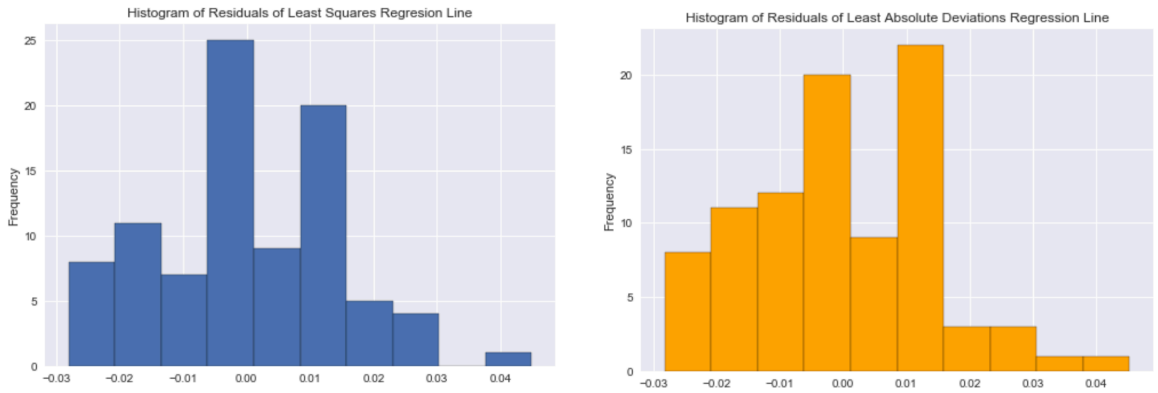

Figure 9: Histogram of residuals for both Least Square Regression and Least Absolute Deviation Regression

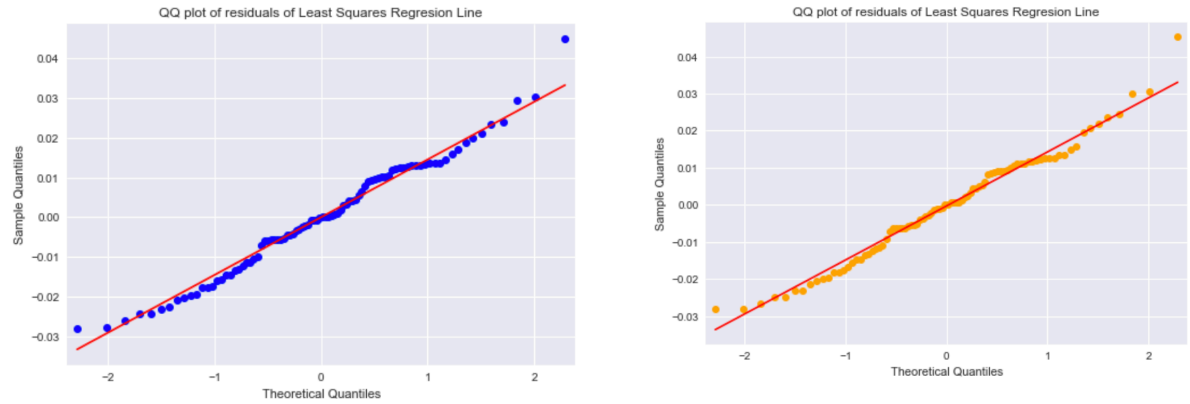

Figure 10: QQ plots of residuals for both Least Square Regression and Least Absolute Deviation Regression

Scenario 2: Prediction

Ultimately, we are interested in predicting the snow density given a gain reading. Having our linear regression models, we are now able to predict the density with a given gain. In this scenario, we will make predictions of snow density given gain readings of 38.6 and 426.7. These two numeric values are chosen because they are the average gains for the 0.508 and 0.001 g/cm^3 densities respectively. Thus, we expect our density predictions for the two gains to be close to 0.508 and 0.001 g/cm^3.

Below is the fit plot for snow density and logrithmic gain with our least squares regression line and 95% confidene intervals and prediction intervals. As 95% confidence interval

gives a prediction range of the true population mean density given logarithmic gain, the 95% prediction interval give a prediction range of future density given logarithmic gain. Since the prediction for future gain has to take the uncertainty in the population mean and the randomness in data scattering process into account, we expect the prediction intervals to be wider than the confidence intervals. Thus, in the plot, we can see that the prediction intervals are a lot wider than the confidence intervals as what we expected.

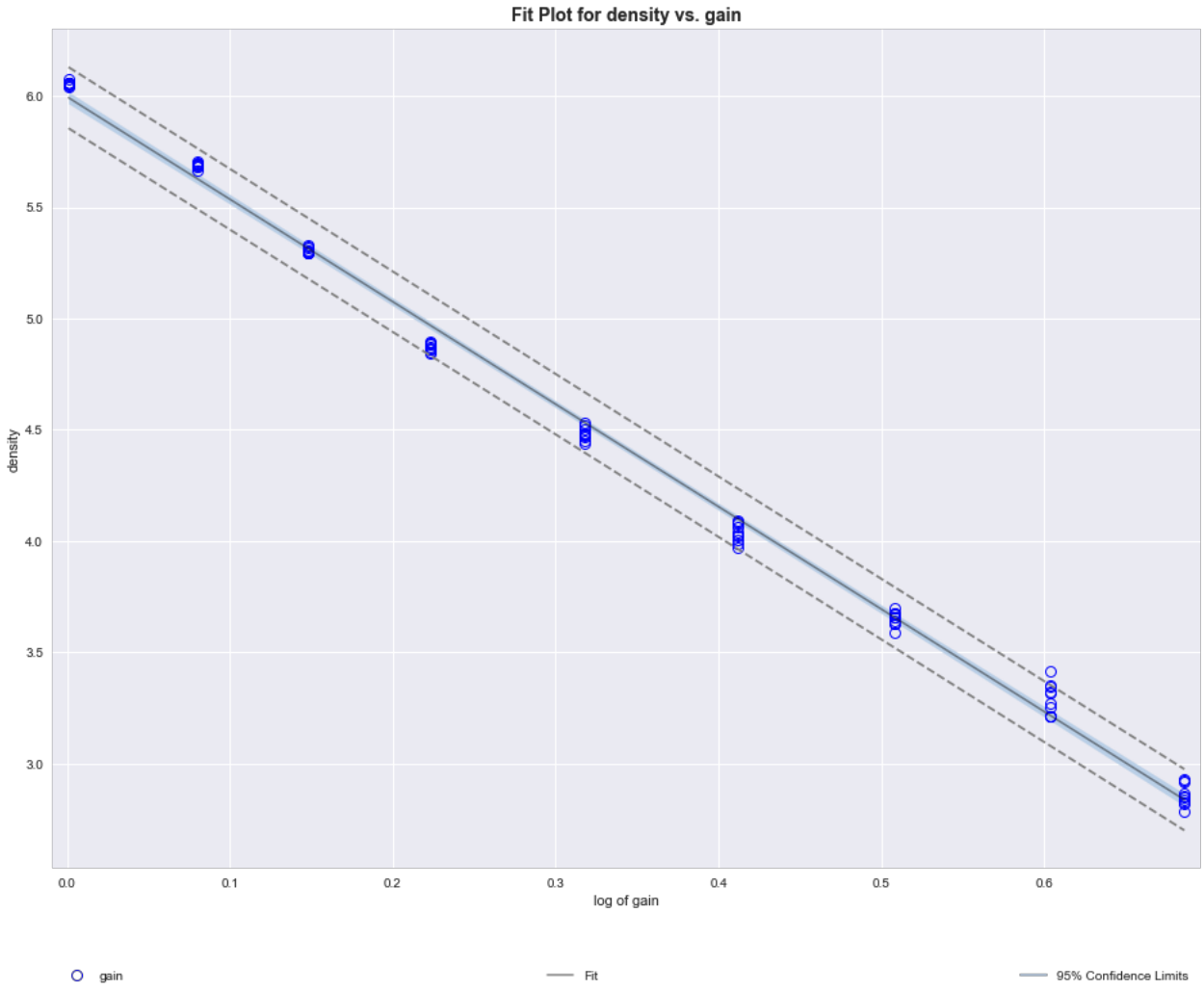

Figure 11: Fit Plot for Density and Log of Gain

We now calculate point and interval predictions of density given gain based upon the above intervals and functions defined earlier in scenario 1.

Gain reading of 38.6 (0.508 g/cm^3 as actural density in dataset):

By using the function of the least square regression line, we get the point estimate in g/cm^3 as 0.50817. By using the function of the least absolute deviations regression line, we get the point estimate in g/cm^3 as 0.50795. Also, we can calculate that the 95% confidence interval in g/cm^3 is (0.50777, 0.50857), and the true mean density falls within the interval. The 95% prediction interval in g/cm^3 is (0.50517, 0.51116), and the density of next data point falls in the interval.

For the predictions given gain reading of 38.6, both least square and least absolute deviations produce point estimates that are similar to the acutral density, and both confidence interval and prediction interval contains the true mean density.

Gain reading of 426.7 (0.001 g/cm^3 as actural density in dataset):

By using the function of the least square regression line, we get the point estimate in g/cm^3 as -0.011332. By using the function of the least absolute deviations regression line, we get the point estimate in g/cm^3 as -0.010028. Also, we can calculate that the 95% confidence interval in g/cm^3 is (-0.0083092, -0.011902). The 95% prediction interval in g/cm^3 is (-0.0083092, -0.014354).

For the predictions given gain reading of 426.7, both least square and least absolute deviations produce point estimates that are close to the acutral density. However, note that the actual density is too close to 0 that our models produce a negative values and negative intervals. These negative values are falsly predicted because of some small errors in our regression models.

Nevertheless, using the same procedure above, we can make further predictions of density given other gain readings.

Scenario 3: Cross-Validation

From the investigation above, we see that there is a linear relationship between density and the log transformation of measured gains. However, we have no way to test the precision of the regression model we used above except using the predicion and confidence interval. Thus, we will divide our data into train set and test set, in which the train set is used to fit the regression model while the test set is used to serve as a check to validate the regression model we get.

First, we will define our test set as the gains for which the polyethylene blocks have a density of 0.508 g/cm^3. The regression model we compute for the train set would be: $\widehat{density} = -0.216 \times \log(gain) + 1.298$. A plot for our regression model, along with confidence interval, is shown below.

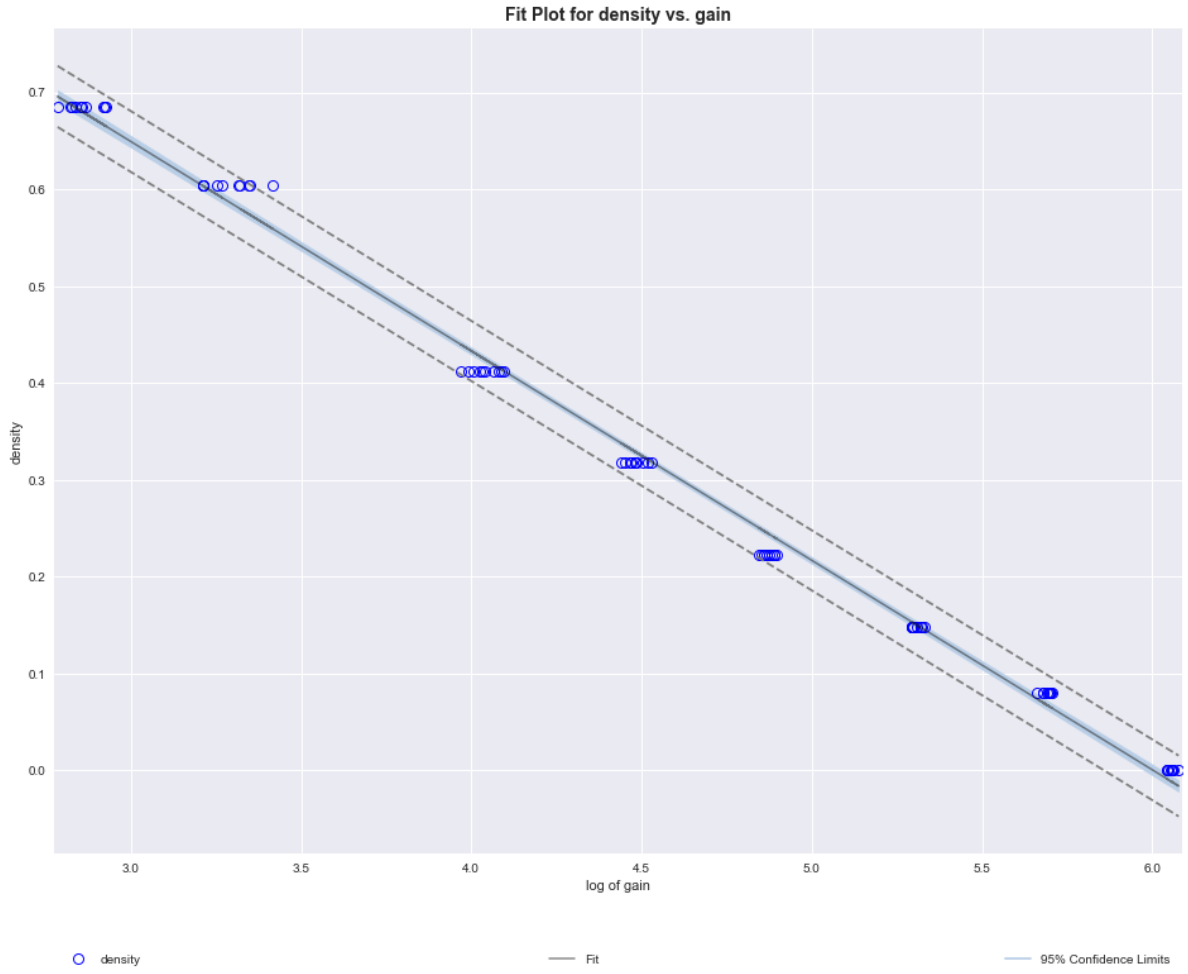

Figure 12. Fit Plot for Density and Log of Gain for Train Set

Now we will use the gain reading of 38.6, the average gain reading for polyethylene blocks with density 0.508, to test this new regression model we get. For the gain reading of 38.6, our regression model would predict the density to be 0.508304, which is close to the actual density of the polyethylene blocks. The confidence interval is (0.507822, 0.508786), while the prediction interval is (0.505018, 0.511589). We can see that the actual density of 0.508 is in the prediction interval. Thus, we can say that the regression model we obtain from the train set of data is a good predictor for density of test set of data.

We will repeat the same process, except this time we will define the test set as gains for which the polyethylene blocks have a density of 0.001 g/cm^3. The regression model we compute for the train set then would be: $\widehat{density} = -0.219 \times \log(gain) + 1.310$. A plot for our regression model, along with confidence interval, is shown below.

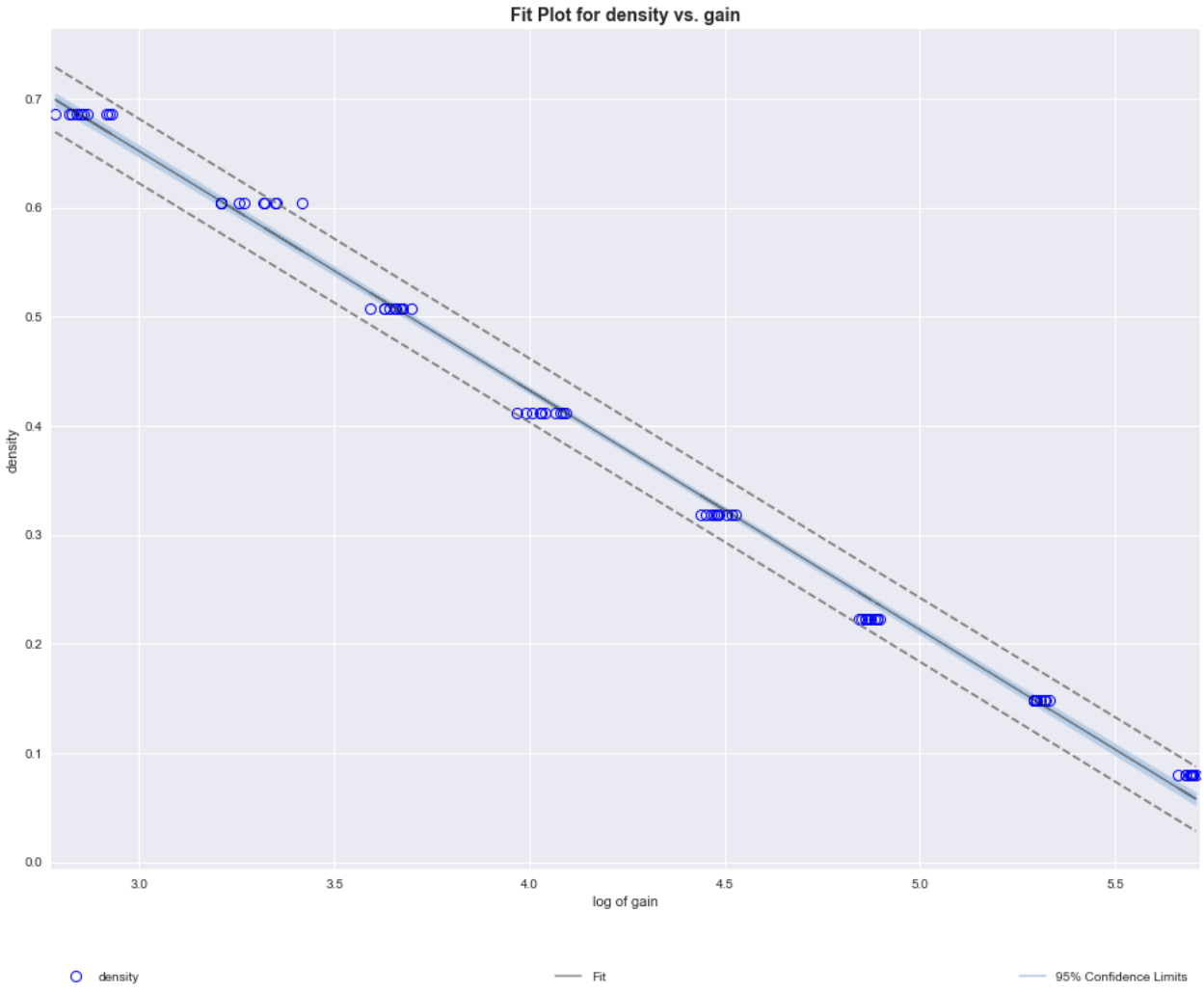

Figure 13 . Fit Plot for Density and Log of Gain for Train Set

This time, we will use the gain reading of 426.7, the average gain reading for polyethylene blocks with density 0.001, to test this new regression model we get. For the gain reading of 426.7, our regression model would predict the density to be -0.0186, while the actual density is 0.001. The confidence interval is (-0.019406, -0.017711), while the prediction interval is (-0.0221604, -0.014957). We can see that the acutal density, which is 0.001, is not in the prediction interval. Thus, we believe that the regression model we obtain from the train set of data is not a good prediction for test set of data. We see that this time the model as a predictor for density is less accurate, comparing to the regression model we get from the first train set. This might be resulted from the fact that 0.001 is the smallest density in our data, and eliminating it along with gain readings obtain from it will affect our model more than eliminating gain readings for polyethylene blocks with density 0.508, which is not and edge value for densities in our data.

From these two instances of cross-validation, we see that although using training set of data which defines gain readings from polyethylene blocks of density 0.508 as test set is able to produce an accurate linear regression model for the testing data, the training set of data which defines gain readings from polyethylene blocks of density 0.001 as test set does not produce the same accurate model.

Additional Hypothesis:

After our analysis above for density and gain, along with literature review, we are also interested in the relationship between latitude and temperature, and thus will make use of data '64503600.csv' from the Full Resolution Data folder to investigate on the possible linear relationship between latitude and temperature. We would like to see how the change of latitude will affect temperature.

As we first compute the correlation coefficient of latitude and temperature, we got a value of -0.832312, showing possible negative linear relationship between latitude and temperature. We then divide our data into train set and test set, with test set size 0.2 of the complete data, and perform linear regression. Our resulting model is $\widehat{Temperature} = -0.832 \times Latitude + 59.39$. The p-value is extremely close to 0, indicating a strong linear relationship between temperature and latitude. With such a small p-value, we reject our null hypothesis that latitude and temperature has the linear relationship by chance. Moreover, the R-squared value of 0.7038 suggests that the model can explain approximately 70.38% of the variability in resulting Temperatures. The regression model, along with its 95% prediction interval, is shown below in *Figure 14*.

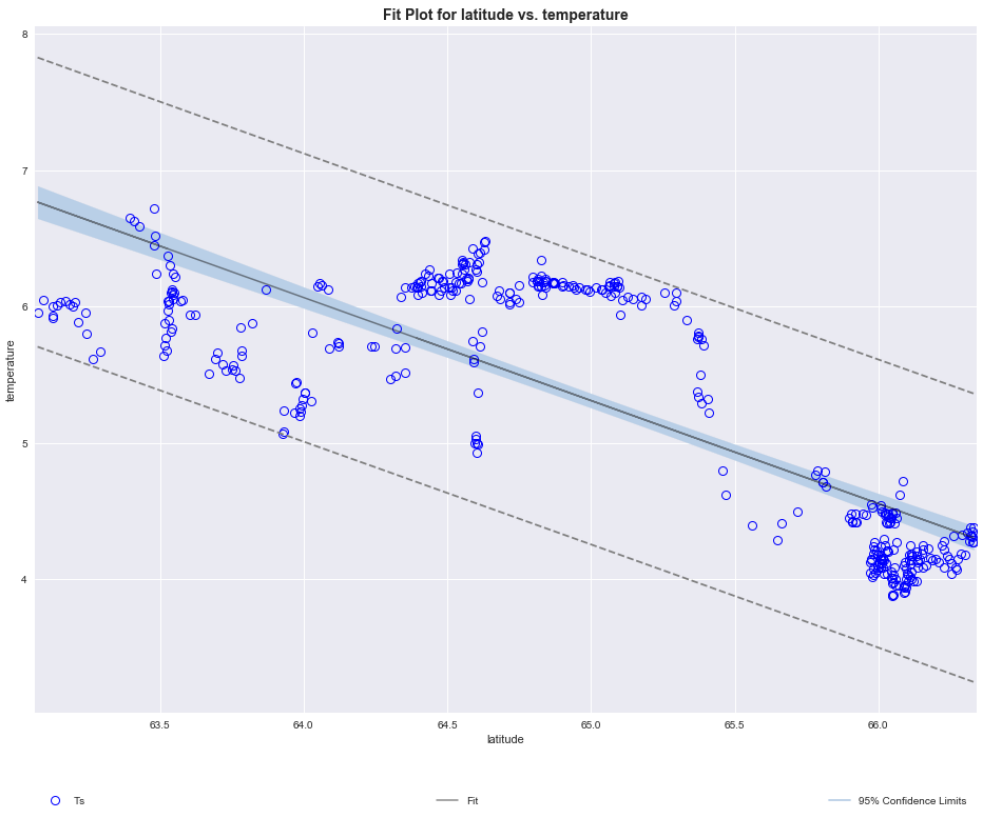

Figure 14. Regression Model for Latitude v.s. Temperature

The residual plot and QQ plot for the residuals of predicted temperatures by test set latitude and actual test set temperatures are shown below. From the residual plot, we can see that the

residuals are stochastic. On the other hand, we can see that the residuals are approximately normal, as the QQ plot of residuals approximately follows the strait line that indicates stardard normal distribution. As a result, we conclude that our regression model for latitude and temperature satisfies constant variability and normality.

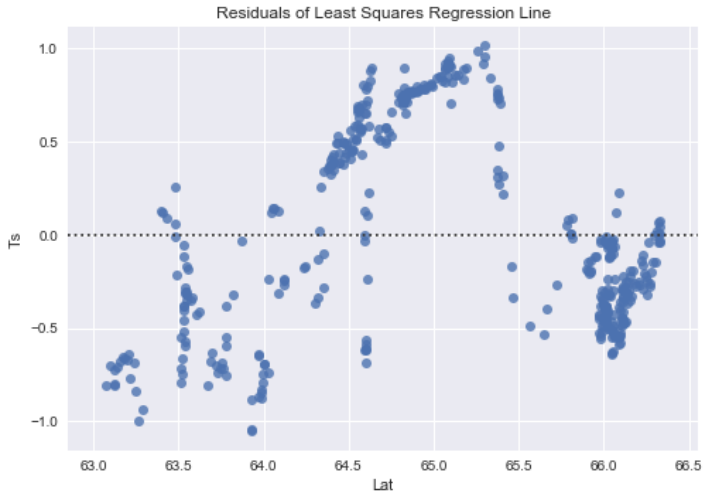

Figure 15. Residual Plot of Predicted Temperatures by Test Set Latitude and Actual Test Set Temperatures

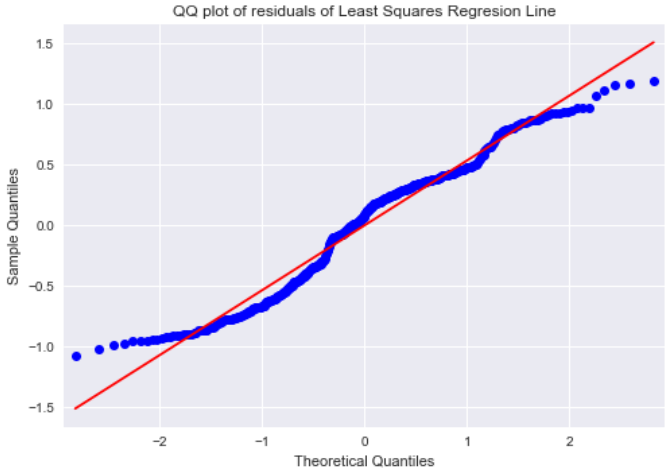

Figure 16. QQ plot of Residuals of Predicted Temperatures by Test Set Latitude and Actual Test Set Temperatures

Discussion & Analysis:

The goal of this project is to calibrate the the data we got from the calibration run of the USDA Forest Service's snow gauge located in the Central Sierra Nevada mountain range near

Soda Springs in order to predict the density of the snow thus having an idea of the snow-pack settlement over the course of the winter season and the dynamics of rain on snow.

Before all the scenerios, we have to talk about the limitation of our data. The dataset we got is quite small, compared to all the snow guage in the area, thus having the risk of being biased or inaccurate in the regression model and the parameters we choose. In the scenario 1, we not only talk about how inaccurate report of the densities of the polyethylene blocks might influence our model and also talk about what will happen if the blocks of polyethylene are not measured in random order. Those are also big potential problems in the dataset. Since we cannot allocate more data to our dataset, we assume all those problems do not occur, and ready to precede our analysis. If we have more resources and more data, the model to predict rainfall will be more precise.

In scenario 1 fitting, we want to find the model that best fit the gain in our data to the density. We first plot a scatter plot of the data, we found that instead of a linear relationship, we observe a exponential relationship, thus we log transformed the gained and make a second scatter plot. In the second plot, we see that there exits a linear relationship between the mean log transformed gain and the density. Thus, we choose two different linear model to fit the data. And both data fit well and robust. We have checkeed the linearity of the graph, the homoscedasticity of the residual and the normality of the residuals. They all supports that the linear model is a good fit for the data and the problem. Thus, we use linear model to precede.

In scenario 2 prediction, since we have the linear model on hand, we can try to predict the snow density with the logarithmic gain for bothe least square regression and least absolute deviation regression models. We plot the 95% confidence interval and 95% prediction interval. . Since the prediction for future gain has to take the uncertainty in the population mean and the randomness in data scattering process into account, we expect the prediction intervals to be wider than the confidence intervals. We observe that for two different gain readings, the point estimate and the true mean density falls within the intervals, thus we believe that the randomness could be predicted by our models.

In scenario 3 cross-validation, we want to use cross validation to furthur test the precision of the regression moedel we used. We divide the data into train set and test set, in which the train set is used to fit the regression model while the test set is used to serve as a check to validate the regression model we get. Fo the gain reading of 38.6, we observe that the actual density of 0.508 is in the prediction interval, thusm we can say that the linear regression is a good fit. For that gain reading of 426.7, our regression model would predict the density to be -0.0186, while the actual density is 0.001, which is not in the prediction interval. We believe this is due to the fact that the density is too small, the smallest in the data we got, thus could have a not so strong position in the loss of regression models we got. But in general, our model would perform well.

In additional hypothesis, we construct a linear model predicting temperature from latitude. Since in the previous analysis and in the literature review, we believe that there is a correlation between the temperature and the density of the snow, and temperature might have a relationship with the latitiude of the gauage site. Thus, we want to figure out the relationship and might use it to fine tune our model in the future to predict snow density. We use the additional data from the Full Resolution Data folder to investigate on the possible linear relationship between latitude and temperature. With p-value 2.37e-95, we indeed find a strong negative relationship between the temperature and the latitude of the site. The residual and qq plots also support that. Thus, we can

conclude that the with higher latitude, the temperature will be lower, thus resulting in a higher density of the snow.

In conclusion, we found that linear model is a good fit for the problem and the dataset. Even with testing and validation, the model produce a consistant and accurate prediction. Thus, by assuming that there is no potential problems in the data, we conclude that our calibration are able to predict the density from the gain correctly, and thus will be able to predict the water supply in northern California.

Theory

- **Correlation:**
  - Describes the strength of the linear association between two variables.
  - Takes value between -1 (perfect negative) and +1 (perfect positive).
  - A value of 0 indicates no linear association
- **Least Square Regression:**
  - The method of least squares is a standard approach in regression analysis to approximate the solution of overdetermined systems.
  - $y_i = ax_i + b$ where a is slope and b is interception
- **Features of Least Square Regression:**
  - Residual is the difference between the observed $y_i$and predicted $\widehat{y_i}$.
    - $e_i = y_i - \widehat{y_i}$
  - Two ways to find the small residuals:
    - 1 Norm: least absolute deviations
      - Minimize the sum of magnitudes (absolute values) of residuals:
        - $\sum_{i=1}^{n} abs(e_i)$
    - 2 Norm: least square regression
      - Minimize the sum of squared residuals – least squares:
        - $\sum_{i=1}^{n} {e_i}^2$
- **Least Squares Regression:**
  - Least Squares Regression is a method for approximating the relationship between the explanatory variables X and response variable Y by minimizing the sum of squared residuals, which is the difference between an observed value, and the fitted value provided by a model. The Least Squares Regressions gives a fitted straight line, called the least squares line. The least squares lines are given in forms of $\hat{y} = \beta_0 + \beta_1 X$, where β_0 represents the intercept, β_1 is the slope, $\hat{y}$ is the predicted value of the dependent variable y and X is the value of independent variable.
  - The least square lines hold the following properties:
    - The line minimizes the sum of squared difference between the observed value (y) and fitted value ($\hat{y}$).
    - The intercept β_0 is where the regression line intersects the y-axis.
    - The slope β_1 represent the increase of predicted value $\hat{y}$ for every unit increase in independent variable X.

        - The line pass through the point $(\bar{X}, \bar{y})$.
        - For each dataset (X, y), the least squares line is unique.
- **Prediction and Extrapolation:**
    - Prediction is plugging the value of X in the linear model equation to predict the value of the response variable for a given value of the explanatory variable X.
    - Extrapolation is applying a model estimate to values outside of the realm of the original data
- **Conditions for The Least Square Line:**
    - Linearity: The relationship between the explanatory and the response variable should be linear. This can be checked by plotting the scatterplot or the residual plot.
    - Nearly normal residuals: The residuals should be nearly normal. This condition may not be satisfied when there are unusual observations that don't follow the trend of the rest of the data (outliers). This property can be checked by plotting the histogram or normal probability plot of residuals.
    - Constant variability: The variability of points around the least squares line should be roughly constant. This implies that the variability of residuals around the 0 line should be roughly constant as well. This property can be checked by plotting the histogram or normal probability plot of residuals.
- **R^2:**
    - The strength of the fit of a linear model is most commonly evaluated using R^2.
    - R^2 is acalculated as the square of the correlation coefficient.
    - It tells us what percent of variability in the response variable is explained by the model.
    - The remainder of the variability is explained by variables not included in the model or by inherent randomness in the data.
- **Types of outliers in linear regression:**
    - Outliers are points that lie away from the cloud of points.
    - Outliers that lie horizontally away from the center of the cloud are called high leverage points.
    - High leverage points that actually influence the slope of the regression line are called influential points.
    - In order to determine if a point is influential visualize the regression line with and without the point. Does the slope of the line change considerably? If so, then the point is in fluential. If not, the it is not an influential point.
- **Inference for linear regression:**
    - HT for the slope:
        - We always use a t-test in inference for regression.
        - Test statistic, T = (point estimate – null value) / SE
        - Point estimate = b_1 is the observed slope
        - SE_{b1} is the standard error associated with the slope.
        - Degrees of freedom associated with the slope is df = n-2, where n is the sample size.
            - We lose 1 degree of freedom for each parameter we estimate, and in simple linear regression we estimate 2 parameters, $\beta_0 \; and \; \beta_1$.

- CI for the slope:
  - Inference for the slope for a single-predictor linear regression model:
    - Hypothesis test:

$$T = \frac{b_1 - null\ value}{SE_{b_1}} \quad \text{df} = \text{n} - 2$$

    - Confidence interval:

$$b_1 \pm t^*_{df=n-2} SE_{b1}$$

    - The null value is often 0 since we are usually checking for any relationship between the explanatory and the response variable.
    - The regression output gives b_1, SE_{b1}, and two-tailed p-value t-test for the slope where the null value is 0.
    - We rarely do inference on the intercept, so we'll be focusing on the estimates and inference for the slope.
- (1-α)100% confidence interval for mean response $\beta_0 + \beta_1 x$:
  - $[\hat{\beta}_0 + \hat{\beta}_1 x - t_{\frac{\alpha}{2},n-2} \cdot s\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}},\ \hat{\beta}_0 + \hat{\beta}_1 x + t_{\frac{\alpha}{2},n-2} \cdot s\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}]$
- (1-α)100% prediction interval for new response $Y = \beta_0 + \beta_1 x + \varepsilon$:
  - $[\hat{\beta}_0 + \hat{\beta}_1 x - t_{\frac{\alpha}{2},n-2} \cdot s\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}},\ \hat{\beta}_0 + \hat{\beta}_1 x + t_{\frac{\alpha}{2},n-2} \cdot s\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}]$
- Caution:
  - Always be aware of the type of data you're working with: random sample, non-random sample, or population.
  - Satistical inference, and the resulting p-values, are meaningless when you already have population data.
  - If you have a sample that is non-random (biased), inference on the results will be unreliable,
  - The ultimate goal is to have independent observations.

- **Multiple linear regression:**
  - is used to explain the relationship between one continuous dependent variable and two or more independent variables.
    - Multiple variables: y and $x_i\ for\ i = 1,2,3$ ...
  - $y_i = \sum_{i=1}^{p} a_i x_i + b$ where
    $a_i is\ the\ slope\ coefficient\ for\ each\ the\ independent\ variable, b\ is\ the\ interept\ term$

Works Cited

Bardic, Jelena. "Chapter 5: Calibrating a Snow Gauge." MATH 189 Lecture, UC San Diego. Lecture.

Wada, M., Kodama, M. and Kawasaki S (1975). "A Cosmic-Ray Snow Gauge." International Journal of Applied Radiation and Isotopes, 1975, Vol.26, pp. 774-775.

Pergamon Press.

Condreva, Kenneth J (1995). “Method For Detecting Water Equivalent of Snow Using Secondary Cosmic Gamma Radiation.” *United States Patent.*

Bavera, D., & De Michele, C. (2009). “Snow Water Equivalent Estimation in the Mallero Basin Using Snow Gauge Data and MODIS Images and FieldWork Validation.” *Hydrological Processes, 23,* 1961–1972. https://doi.org/10.1002/hyp.7328