Case Study 4: Calibrating a Snow Gauge

Dukki Hong (A98058412)

Karl Rummler (A12405136)

Nick Roberts (A11705541)

John Diez (A14751991)

Yiwen Li (A13959913)

Introduction

In this study, we investigate the snow gauge, for which the main source of water for Northern California comes from the Sierra Nevada mountains. To help monitor this 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, California. The gauge is used to determine a depth 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 these replicate measurements on the same volume of snow, researchers can study snowpack 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 the snow can absorb. Analysis of 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. Due to instrument wear and radioactive source decay, there may be changes over the seasons in the function used to convert 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. In this lab, we will develop a procedure to calibrate the snow gauge

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 measurements recorded by the gauge are an amplified version of the gamma photon count made by the detector. We call the gauge measurements the “gain.”

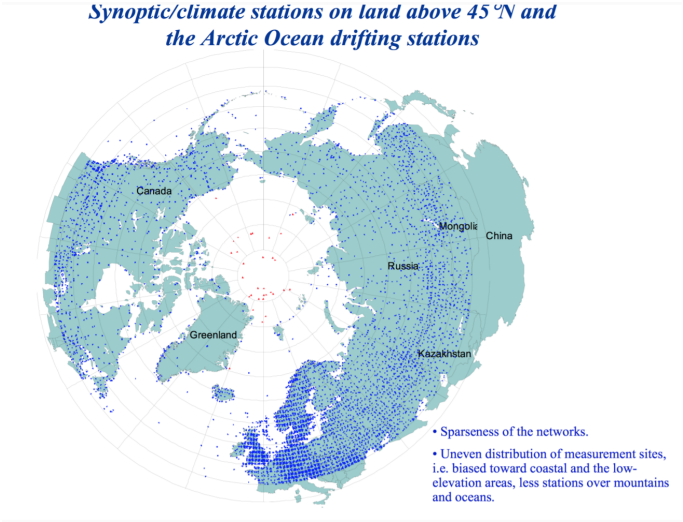
The data available for investigation consist of 10 measurements for each of 9 densities in grams per cubic centimeter (g/cm3) of polyethylene. The measurement reported are amplified version of the gamma photon count made by the detector. We call the gauge measurement the ”gain”. The data available here consists of 10 measurements for each of 9 densities in grams per cubic centimeter of polyethylene.

Density	Gain									
0.686	17.6	17.3	16.9	16.2	17.1	18.5	18.7	17.4	18.6	16.8
0.604	24.8	25.9	26.3	24.8	24.8	27.6	28.5	30.5	28.4	27.7
0.508	39.4	37.6	38.1	37.7	36.3	38.7	39.4	38.8	39.2	40.3
0.412	60.0	58.3	59.6	59.1	56.3	55.0	52.9	54.1	56.9	56.0
0.318	87.0	92.7	90.5	85.8	87.5	88.3	91.6	88.2	88.6	84.7
0.223	128	130	131	129	127	129	132	133	134	133
0.148	199	204	199	207	200	200	205	202	199	199
0.080	298	298	297	288	296	293	301	299	298	293
0.001	423	421	422	428	436	427	426	428	427	429

Table from Literature Stat Labs.

Challenges with the data

Although our data is generally a single calibration of a single snow gauge, a few challenges and issues exist. Operational networks, which is also known as our knowledge base, is not universally spread across the world. For example, the decline of the networks in the northern regions, including Siberia, Alaska and North Canada. Moreover, there are very few stations in the mountain regions. From the figure below, looking into the spread of blue dots on the globe, we can see the problematic distribution. In which case, the uneven distribution of measurement sites may lead to biased data. Also, considering different regions and geological conditions, how to sustain and improve the operational networks would also be a not solvable issue in the short run. Lastly, the standardized methods for calibration and maintenance is different across different regions. This affects our certainty that the snow gauges are efficiently being maintained or corrected throughout the years.



Furthermore, data quality and compatibility across national boundaries may also be an issue. Take consider to: large biases in gauge measurements of solid precipitation, incompatibility of precipitation data due to difference in instruments and methods of data processing and difficulties to determine precipitation changes in the arctic regions. Validation of precipitation data, including satellite and reanalysis products and fused products at high latitudes is questioned somehow. However, with the help of satellite precipitation data and reanalysis on fused products at high latitudes, the challenges are manageable.

Background

Snow Gauges are able to measure solid precipitation throughout multiple areas to keep track of changes across time. With snow gauges located in different temperate fields, their accuracy in snow gains and retention are influenced by environmental factors, and the structure in how the gauge is made. Colli and Lanza were able to linearly model the effects of wind speeds on both unsheltered and sheltered snow gages. They found that a snow gauge's exposure to wind is responsible for significant reduction of their collection performance. To accommodate, they put up two solutions of both a new way to protect the gauges to allow for greater collection, and a reanalyzer or predictive model of gains based on known wind speeds.

Snow Gauges aren't only located within land either, a lot snows occur in the middle of the sea. Due to that, a lot of Buoy type snow gages are sent out all around the world for the seas. Similar to Snow Gauges around the world, these Snow Gauge Buoys have different parts and types, meaning that they might each be affected by

different environmental feedback. In a study by Goodison, they tested how different types of snow gauges are affected by certain environmental factors and how it affects their predicted yields. What they found was that different types (protected vs unprotected) snow gauges had different predicted yields, and factors such as wind affected this. With this in mind, we would expect the same type of variance, possibly even more, for snow Buoy Gages as they are even greatly affected by wind, and the movement of the waves.

Investigations

Scenario 1:

It is important to note that the dataset had it's own problems. The densities of the polyethylene blocks were not reported properly. This implies that when graphed, the data might not fit to the regression line. Furthermore, this would also mean inaccurate measurements of densities. Moreover, since the data was collected from selected regions with no randomness, this condition might negatively impact the data fitting to the least squares regression line. The interest of this investigation is the snow on the Central Sierra Nevada mountain. Hence if data were collected from a single location on the Central Sierra mountain, it would be obvious that there may be biases in the data since the data was only taken from a single location instead of a random location. It is important to note the importance of randomness of the data so the data would not be biased. Lastly, it is important that we satisfy all three conditions for the least squares regression line: Linearity of the explanatory and response variables, Normality of the residuals and Constant variability.

Figure 1 shows the scatter plot of the raw data. Nine polyethylene blocks with their respective densities are plotted in orange. The trend of the data is similar to that of an exponential trend contrary to a linear trend. We will take the log of the gain data to change the data into a linear trend. Figure 2 shows a scatter plot of the density and log of the gain variables. It is clear that the data is more fitted on the least squares regression line. To create an environment where there exists a one-to-one correspondence between the explanatory and the response variable, we will take the average of the replicated measurements to come up with single data for each density. Figure 3 shows a scatter plot of the average of the replicated measurements. We see that the data plotted on Figure 3 is approximately close to the least squares regression line. To be more precise, the correlation coefficient was approximately -0.9985 and r^2 was approximately 0.9969 which means that there is a 99.69% variability in the response variable. Now we have met the first condition for least squares regression line: Linearity of the explanatory and response variables.

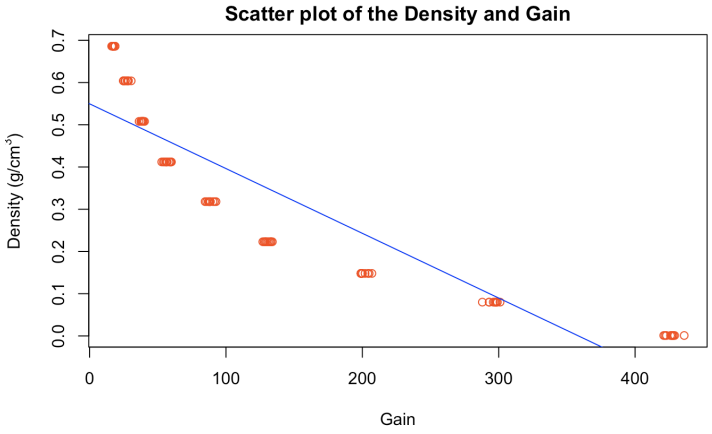


Figure 1: Scatter plot of the raw data

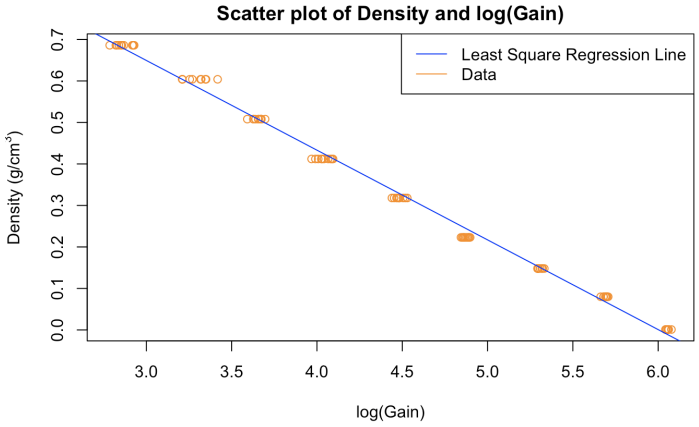


Figure 2: Scatter plot of the log(gain) data

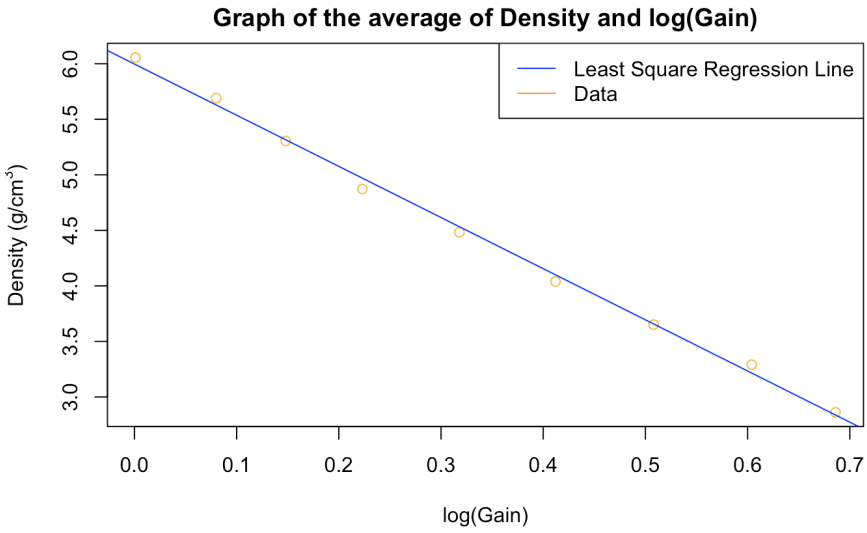


Figure 3: Graph of the average of density and log(gain)

Figure 5 and 6 shows a graph that includes three lines: The Least Squares Regression line (green), Least Absolute Deviation Regression line (blue) and the Median Regression Line (red). The Three different linear lines seem very close to each other and in fact, they seem to overlap each other. Figure 6 shows a zoomed in version of Figure 5. It is apparent in the graph that the three different Regression lines do not overlap each other, rather they are very close to each other. To be more precise, the following is a summary of the slope and intercepts of the three Regression lines:

Least Squares Regression Lines: $\widehat{Density}$	1.2991 - 0.2164*log(gain)
Least Absolute Regression Line: $\widehat{Density}$	1.3029 - 0.2177*log(gain)
Median Regression Line: $\widehat{Density}$	1.3029 - 0.2177*log(gain)

The results above show that all three regression lines are appropriate and good estimators.

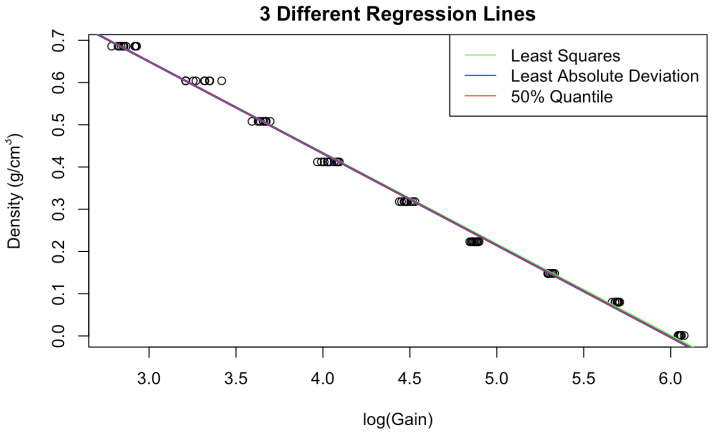


Figure 4: Graph including Least Squares, LAD and Median Regressions

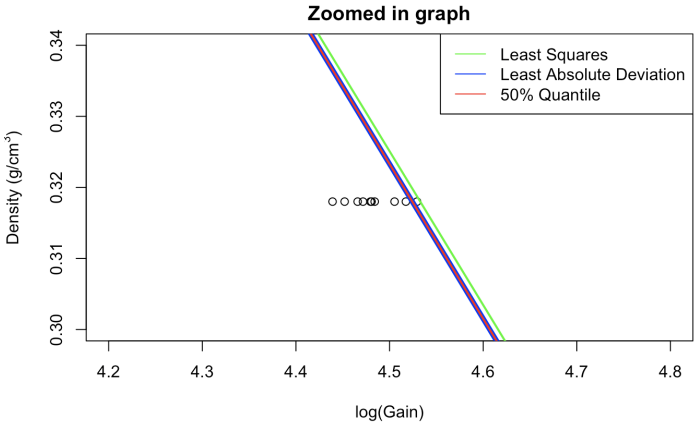


Figure 5: Zoomed in graph of Figure 4

It would be more precise to further check the linearity of the explanatory variables and the response variables. We will proceed by using the three regression lines above. Looking at Figures 6-8, we see that the residual data is scattered around the red line. The data seems to increase from 0 index and then decrease again from around index 60. Additionally, the data points are sparsely scattered around the graphs so it would not be significant of an issue in this case.

Figure 6: Residual plot of Least Squares Regression

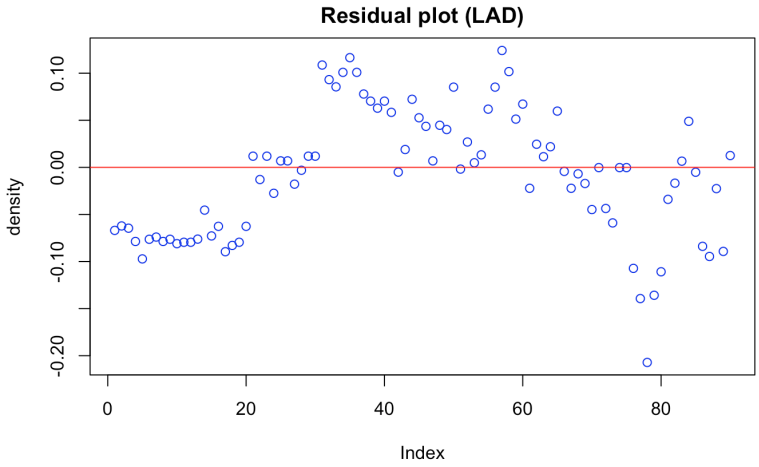


Figure 7: Residual plot of Least Absolute Regression

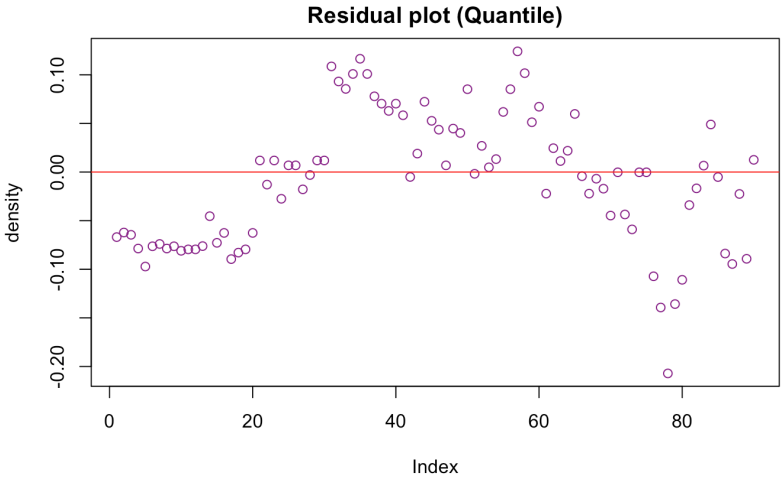


Figure 8: Residual plot of Median Regression

We are left to check the normality and constant variability. We will proceed and check normality first. Figure 9 and 10 shows histogram of the Least Squares Regression Line and a Q-Q plot of the Least Squares Regression Line. It can be seen in Figure 9 that the trend of a histogram resembles that of a normal distribution, a bell-shaped curve. Less data is scattered around the outliers around the value of 0 and more data is concentrated near 0. The data points around the red linear line in Figure 10 are very closely scattered. This shows that the points are approximately normal. Similar conclusions can be drawn for Least Absolute Regression Line and Median Regression Line as both the histogram and the Q-Q plot looks and behaves similar to that of Figure 9 and Figure 10 (Figures 11-14).

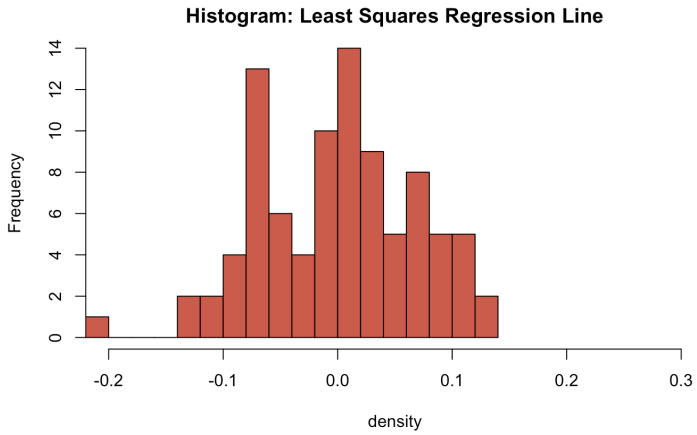


Figure 9: Histogram of Least Squares Regression Line

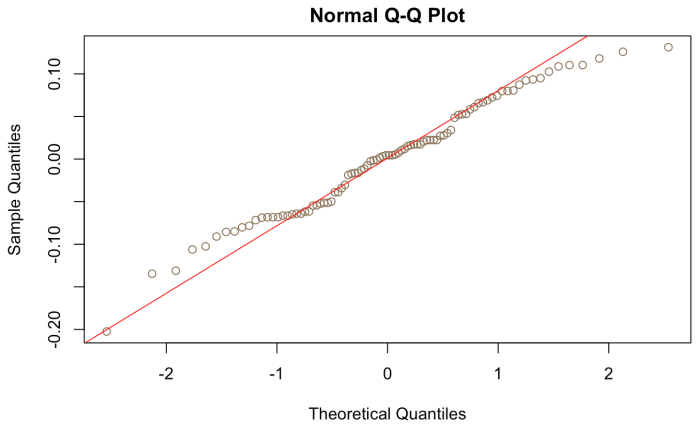


Figure 10: Q-Q plot of Least Squares Regression Line

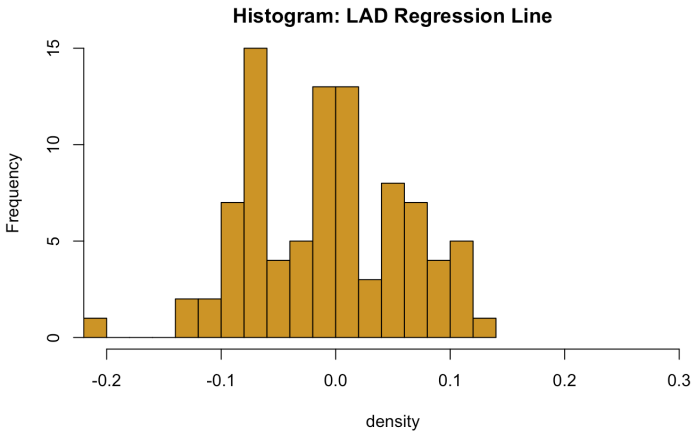


Figure 11: Histogram of Least Absolute Regression Line

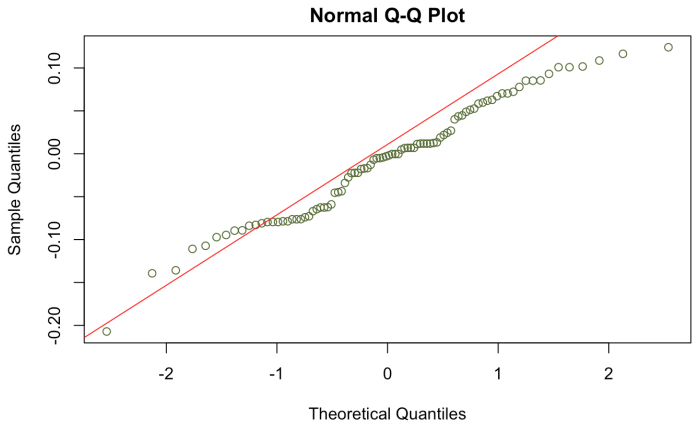


Figure 12: Q-Q plot of Least Squares Regression Line

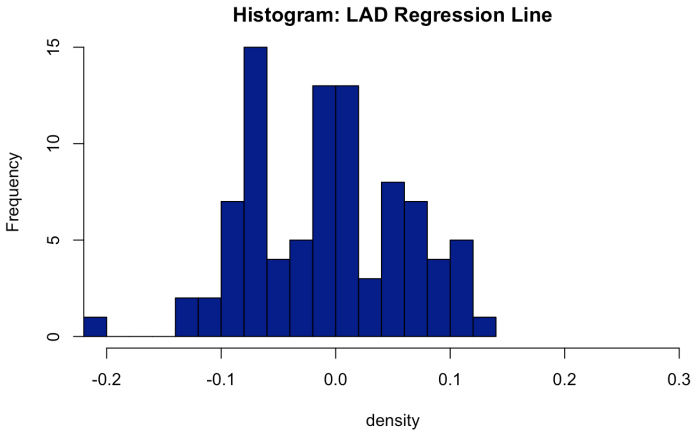


Figure 13: Histogram of Median Regression Line

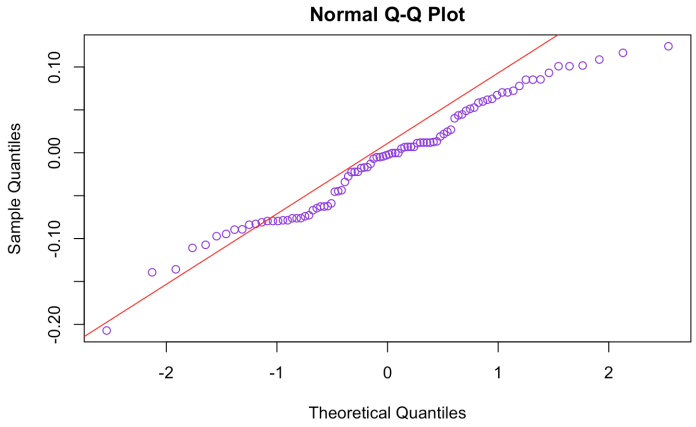


Figure 14: Q-Q plot of Median Regression Line

We are only left to check for constant variability. Figure 15 is an adjusted residual plot of the three regressions: Least Squares, Least Absolute and Median. The data points around the red line are roughly constant and it is similar for all the graphs. This implies that the variability of residuals around the 0 line should be roughly constant as well.

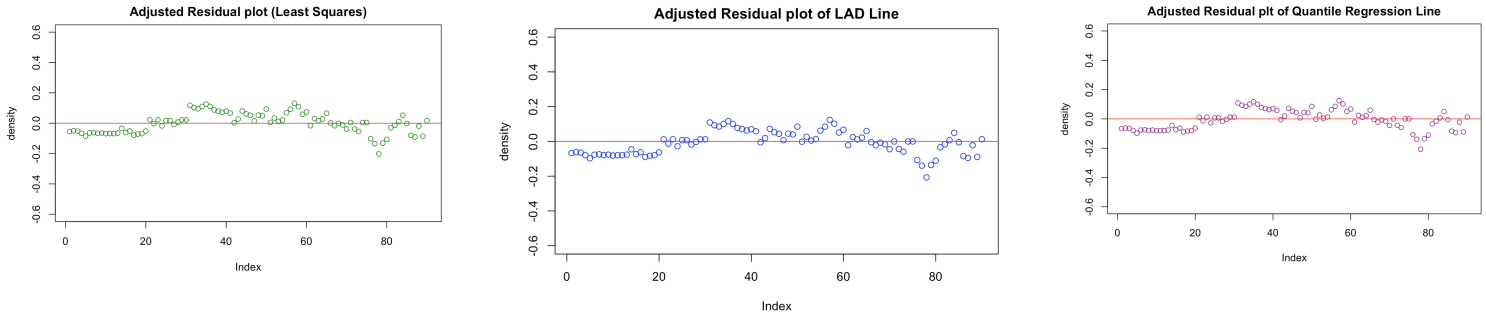


Figure 15: Adjusted Residual plot of the three Regression Lines

Because we have seen that the three conditions for Least Squares Linear Regression, Linearity, Near Normality of Residuals and Constant variability, have been met, we will use such models throughout our analysis.

Scenario 2: Predicting

Our ultimate interest is to answer what the density of the snow-pack is given a gain reading of 38.6 and 425.7. Knowing that the two specific numerical values, 38.6 and 426.7 were chosen because they are the average gains for the 0.508 and 0.001 densities, respectively.

The *Graph 1.* below shows the least square regression of snow density based on previous calculations in basic analysis. This graph gives us a basic idea that the least squares of snow densities are linearly distributed. The least squares regression line is the line that makes the vertical distance from the data points to the regression line as small as possible. It's called a "least squares" because the best line of fit is one that minimizes the variance, which is the sum of squares of the errors. The main point is to find the equation that fits the points as closely as possible.

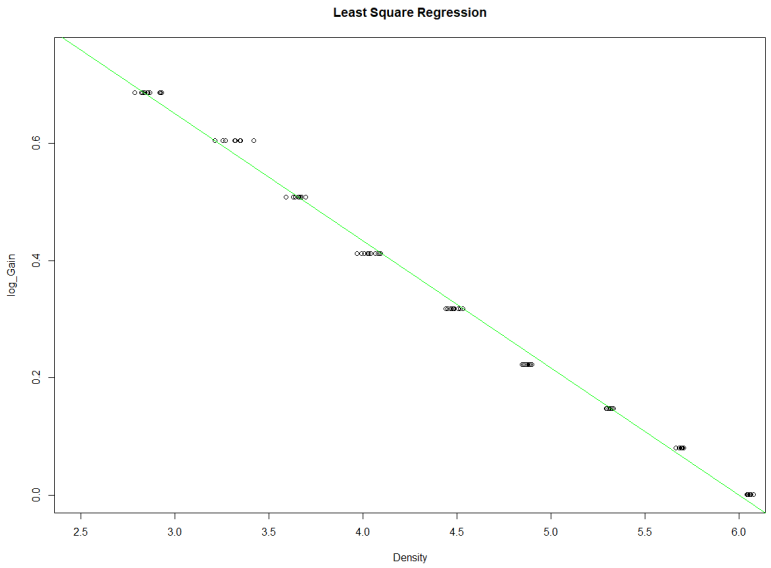


Figure 16: Least Square Distribution

Based on model fitting, we would like to know how good the model predicts the density of the snow-pack based on a given gain reading parameter. With the set up least squares regression model, we make point and interval estimates for the snow-pack density from gain measurements. The *Graph 2.* below follow a procedure for adding bands around least squares line. The *Graph 2.* shows both the bands of confidence intervals prediction intervals. Reading from the graph, both the confidence and prediction interval slightly curved concavely. This meets our expectation because it follows the pattern that the further away an interval is from the center, the wider the interval would be. It changes around the center of the data points.

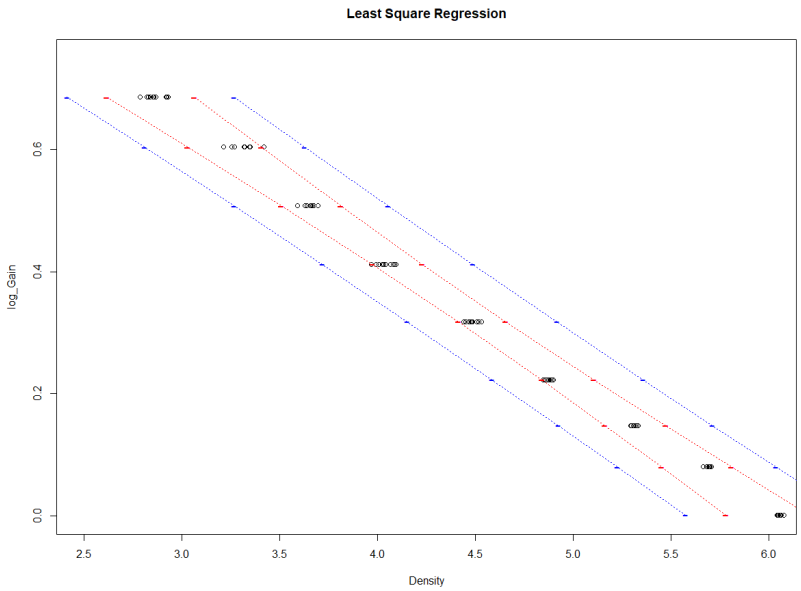


Figure 17: 95% Confidence Interval Bands and 95% Prediction Interval Bands

We consider the log_polynomial model: *log(gain) = a + b * density + c * density * density* and how well log model fits the data. For the point estimate and interval estimate of the gain reading of 38.6. The point estimates, including least squares regression estimate, least absolute deviation regression estimate and quantile estimates provides us expected results, which is around one decimal deviation error with the actual density of the snow-pack, 0.508. The 95% confidence interval estimate of 38.6 does not capture the actual density of the snow-pack while the interval estimate given log gain capture the actual density. Notice that the interval given log gain is wider than interval given gain. On the other hand, the point estimate for gain reading of 426.7 has larger deviation from actual density of the snow-pack while the confidence interval successfully captures the actual density. Notice that there are some negative values in the point estimates of gain 426.7, which indicates small biases and errors exists in the tail of our regression because the snow density cannot be negative.

	Gain Reading of 38.6	Gain Reading of 426.7
Actual density of the snow-pack	0.508	0.001
Least squares regression estimate	0.5089	-0.01277

Least absolute deviations regression estimate	0.5076	-0.01547
25% quantile regression estimate	0.5063	-0.03927
50% quantile regression estimate	0.5076	-0.01547
75% quantile regression estimate	0.5163	0.0009904
95% confidence interval estimate for density given gain	(0.5012, 0.5169)	(-0.02429, -0.001769)
95% confidence interval estimate for density given log gain	(0.4890, 0.5291)	(-0.03465, 0.008619)

Table 1: Estimate Table. in g/cm^3

<u>Scenario 3: Cross Validation</u>

Prediction of Density with a 38.6 gain (Density .508):

Estimating density with a gain reading of 38.6. For a gain of 38.6 the actual density is reported as 0.508, where our least squares estimate gave a value of ~.509 and 95% confidence interval from (.51 +/- .01) and prediction interval of (.515 +/- .015).

Looking at the confidence interval, we notice that its a little larger than when working solely on the training data, this makes sense since its working on data its never seen before. This confidence interval is slightly larger than we've seen but this makes sense as we're predicting an new value. Additionally, when comparing to the confidence interval of a datapoint that's towards the outskirts of the data, we'll see a large increase in the CI, due to increased variance.

Prediction of Density .001, using gain 426 ~= log(gain) ~ 2.6:

Estimating density of .001 is guessing with a gain of 426. However, running least squares on 426 gives a predicted value of -.022. Obviously this point is wrong, but that makes some sense because the actual value is very close to 0. Looking at the confidence intervals for this, we see that the confidence interval is (-.04, -.01) and the prediction interval is from (-.05, .003). These confidence intervals are noticeably wider than the previous confidence intervals calculated with just training and the previous prediction. This makes sense as the prediction is closer to the edge of the data and so as we get farther from the center the intervals will experience higher variance.

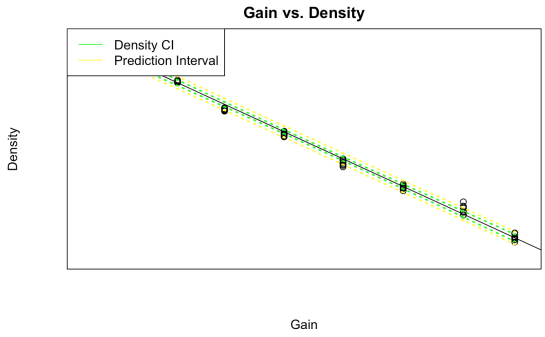


Figure 18: Confidence and Prediction intervals with Density .508 as holdout set

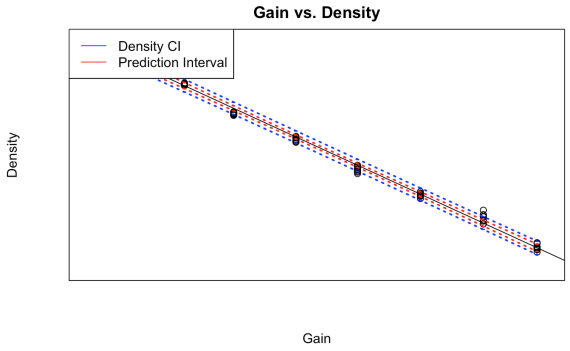


Figure 19: Confidence and Prediction intervals with Density .001 as holdout set

Looking at the figures above, we see that the intervals on the left are clearly tighter than those on the right. This gives a sense of the confidence interval for our predictions. Additionally, these graphs supports our findings in the previous too predictions.

Further analyzing these intervals, we see that the predicted intervals are still capturing the true density which gives a good support to our models. So we see that the model seems to be a decent predictor for the unseen data.

Additional Questions

Questions:

Which Snow Gauge Buoy provides the least error?

Do Snow Gauge Buoys differ in the variability of their measurements of Atmospheric Temperature and Surface Temperature?

With Snow Gauges being measured from soon-to-be snow parts of the ocean all over the world, one would wonder which snow gauge would yield the lowest amount of variability in their measurements of temperatures. There are 9 types of Buoys available and we picked on a subset of four types due to limitations on time. These four are AXIB, ICEB, SVP, and Snow Buoy. In Table 2, we took 2 randomly picked data sets (that had values for our Dependent measure) for each Buoy type in our analysis. Our dependent measure was Surface Temperature and Atmospheric temperature which are typically measured by the Buoys. We created a full graph set that shows a function of how their measurements of the two dependent measures change or is different across time. What we're interested in is the density or length of variability is locating in each specific time frame. If there is a large cluster, then it would indicate higher variability. In the Data, we removed any row that had any missing values for the two dependent measures.

There are many limitations to the analyses we make, as we have not controlled for Longitude and Latitude for each Buoy. As stated before, environmental factors greatly influence the yield and measurements of snow gauges and buoys, but we hope that randomly choosing examples would yield a less biased sample. Also we weren't able to conduct a full analysis and our conclusion will be less comprehensive, but the visualization portion is the strongest.

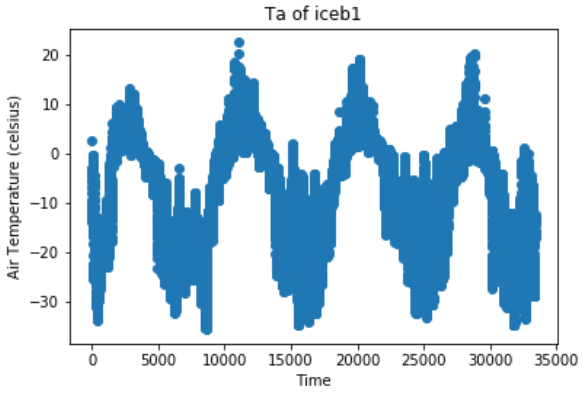
Table 2: Numeric Values for Buoys. TS represents Surface Temperature, TA represents Atmospheric Temperature

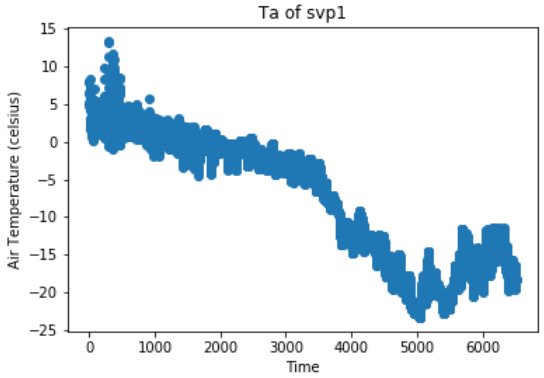
Buoy Name	Buoy Type	TS Mean	TS SD	TS Var	TA Mean	TA SD	TA Var
AXIB1	AXIB	-5.05	7.3	53.26	2.51	25.50	650.55
AXIB2	AXIB	-4.051	3.22	10.38	-8.59	17.26	298.00
ICEB1	ICEB	-6.449	8.903	79.264	-8.357	10.878	118.332
ICEB2	ICEB	-7.171	9.355	87.527	-47.277	17.688	312.889

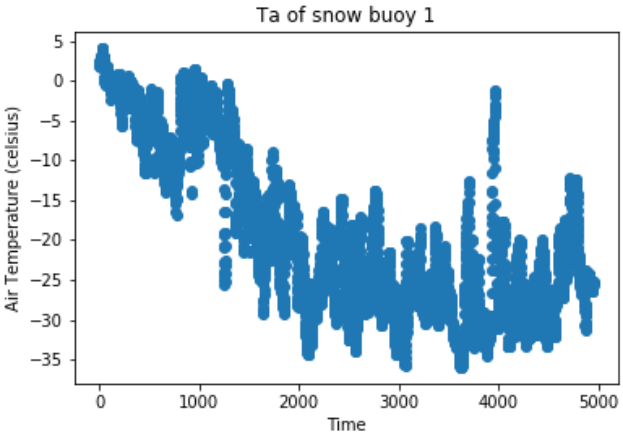
SVP1	SVP	-4.0473	5.813	33.801	-7.308	8.405	70.657
SVP2	SVP	-6.304	6.069	36.837	-11.452	10.368	107.501
SnowBuoy1	Snow Buoy	-5.297	5.0438	25.440	-19.072	10.633	113.066
SnowBuoy2	Snow Buoy	-4.326	3.533	12.489	-20.405	10.195	103.944

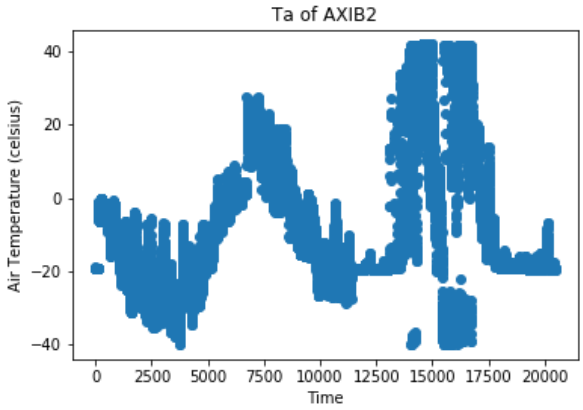
Graph Set:

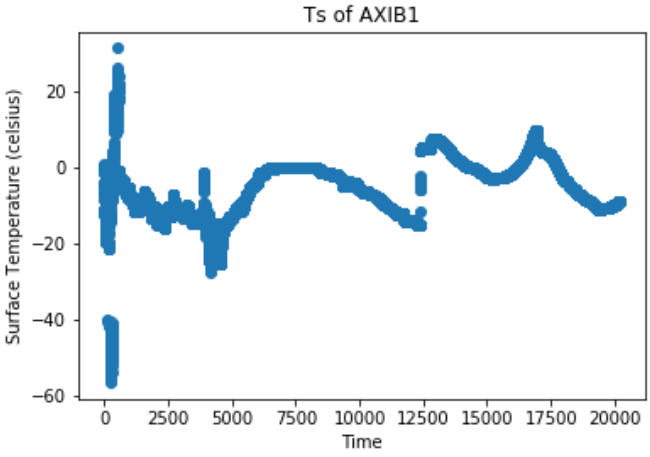
Legend: X axis represent time spanning across multiple days. Y Axis represents either Atmospheric Temperature (TA) or Surface Temperature (TS). The bulkiness typically represents

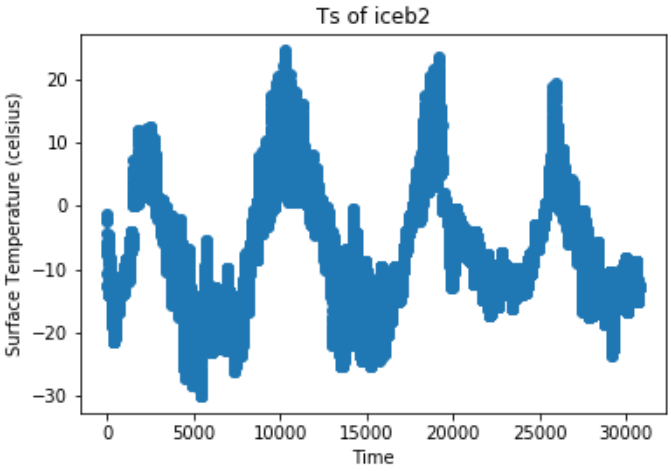


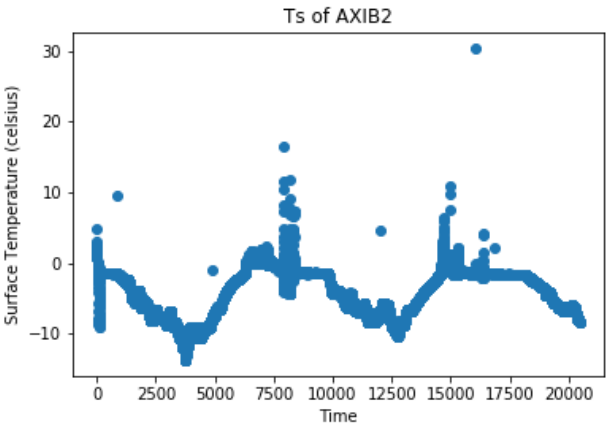



Ta of snow buoy 1
Air Temperature (celsius)
Time


Ta of AXIB2
Air Temperature (celsius)
Time


Ts of AXIB1
Surface Temperature (celsius)
Time


Ts of iceb2
Surface Temperature (celsius)
Time


Ts of AXIB2
Surface Temperature (celsius)
Time

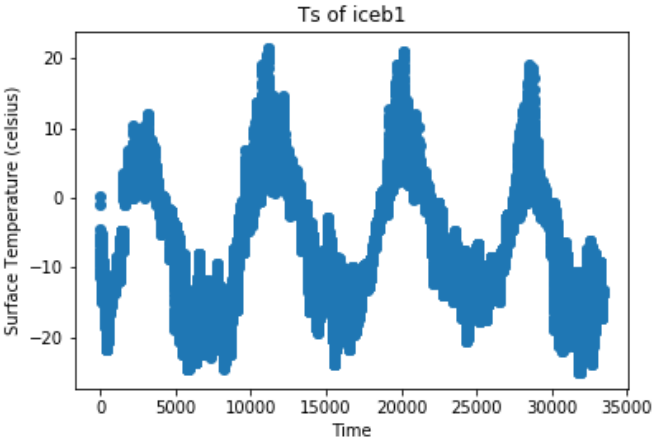

Ts of iceb1
Surface Temperature (celsius)
Time

Figure 20: X axis represent time spanning across multiple days. Y Axis represents either Atmospheric Temperature (TA) or Surface Temperature (TS). The bulkiness typically represents

F/Chi Squared Test on the Variances

To answer the question of which Buoy type yielded the least amount of variability for their measures of TA and TS, we conducted an analysis of the Variances of the four different types of Buoys. To do so, we compared a single Buoy Type's variance and compared it to the overall variance of the three groups combined together. We found that the ICEB had the least amount of variability (most consistent) with a p value of almost 0 for TS, and a non-significantly different variability for its measure for TA. Snow Buoy had the most different variability compared to everyone for both TA and TS. SVP Was significantly different for its TS value, but not the TA values. AXIB was significantly different for TS but not for TA. We would recommend that SB Buoy is the most different one, with the most variability, while ICEB seems to have very consistent variability. Limitations to these results are discussed before and after this analysis.

Discussion/Conclusion

As we said in the intro, the overarching goal with our analysis is to first understand how accurate Snow Gauges measurements on gains are. Due to the scope of the study, we only accessed one specific data set from a single Snow Gauge located in the Sierra Nevada. We received and analyzed a repeated measure dataset. We wanted to calibrate this particular snow gauge through the use of a linear model to be able to better predict the density of snow from a single gain reading. If we succeeded, we would have the capability to predict how much snow there should be based on how much snow gain there was.

From this goal, we created a couple of linear models: LSQ or Least squares regression, Least Absolute Deviations, and 50% median regressions models. After modelling the percent gains, we found that each of these models were good models in predicting density gain for the gauge. The linear model

fits the best with our transformed data (log). Each three of the models produced very similar predictions, showing that our procedure was good and robust. In Scenario two, we created confidence and interval bands for the estimates for density in each of our models and found a good predictor. In scenario three we tested the accuracy of the linear models and the procedure (model fitting) in which we actually calculated.

If we assume that the snow gauge can be best explained by the LAD or linear model, then we safely assume that the linear model is appropriate from the data and consistently produces accurate predictions. However, there is still some worry for the data as the Mean Squared Error, or the average variance for the model is a little high, and can produce some of the main differences or issues in the data. Since the data model provided good enough evidence

For our additional Analysis, we wanted to find out which Buoy would provide the least amount of measurement error. We found and conclude that the SB or Snow Buoy had the most variability in its measurements while ICEB had the least amount of variability. We would recommend future Buoys to be SB, but we had a lot of confounding variables and limitations.

For our additional analysis.

Methods

Linear Regression and simple linear model

Linear regression is one of the most common statistical modeling techniques. In statistics, it is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables. The case of one explanatory variable is called simple linear regression. In this study, we mostly used simple linear regression. The relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the “lack of fit” in some other norm, or by minimizing a penalized version of the least squares cost

function as ridge regression and lasso. Conversely, the least squares approach can be used to fit models that are not linear models.

The simple linear model is E[Y|x] of a random response Y at a known design point x satisfies the relation E[Y|x] = a + b * x. The gaussian measurement model supposes that measurement errors E have mean = 0 and constant variance and are uncorrelated. If we observe pairs (x1, y1),... (xn, yn), then the method of least squares can be used to estimate a and b.

The least squares estimates of a and b are $\hat{a} = \frac{(\sum x_i^2)(\sum y_i) - (\sum x_i)(\sum x_i y_i)}{n\sum x_i^2 - (\sum x_i)^2}$, $\hat{b} = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2}$.

Note that a_estimate and b_estimate are linear functions of the responses yi, so they are linear functions of the errors, even though we don't get to see them. In previous analysis, we showed that a_estimate and b_estimate are unbiased and to find their variances and covariance under the Gauss measurement model. The residuals are also unbiased, and we can think of the residuals as estimates of the errors. And the residual sum of squares has expectation $\mathbb{E}(\sum[y_i - (\hat{a} + \hat{b}x_i)]^2) = (n-2)\sigma^2$.

The residual sum of squares can thus provide an estimate of the variance in the Gauss measurement model. (Stat Lab, D. Nolan, T. Speed, p167)

Properties of Least Squares/Fitting "a line"

The method of least squares is a standard approach in regression analysis to approximate the solution of overdetermined systems. The most important application is in data fitting. The best fit in the least-squares sense minimizes the sum of squared residuals. The sum of squares to be minimized is

$$S = \sum_{i=1}^{n}(y_i - kF_i)^2.$$

The regression line has properties:

- The line minimizes the sum of squared differences between observed values and the predicted values
- The regression line passes through the mean of the X values and through the mean of the Y values

- The regression line passes through the mean of the X values and through the mean of the Y values
- The regression constant is equal to the y intercept of the regression line
- The regression coefficient is the average change in the dependent variable for a 1-unit change in the independent variable. It is the slope of the regression line.

The least squares regression line is the only straight line that has all of these properties

Multiple observation and replicate measurements

Suppose there are 9 distinct values of explanatory variables. For each X, we have 10 replicate measurements. Then if x1,...,xm are the m distinct values, the responses can be denoted by Yij , i = 1,...,m and j = 1,...,k, where Yij is the j th measurement taken at xi. That is, for our simple linear model, $Y_{ij} = a + bx_i + E_{ij}.$. Note that the errors Eij are assumed to be uncorrelated for all i = 1,...,m and all j = 1,...,k. We use replicate measurement to estimate standard deviation and variance that does not depend a lot on the model. If the residual does not fill correctly, we consider the residuals include measurement error as well as model misfit.

Model Misfit

An alternative to the simple linear model is a polynomial model. For example, a quadratic model for the expectation is E(Y |x) = c + d * x + e * x * x. Regardless of the model for the expected value of the response, we can fit a line to the data. For example, say that E(Y|x) is the true model for our data. If we observe pairs (x1, y1),..., (xn, yn), we can fit a line by the method of least squares to these observations: minimize, with respect to a and b, the sum of squares $\sum_{i=1}^{n}[y_i - (a + bx_i)]^2.$

The solutions for a_estimate and b_estimate remain as $\hat{a} = \frac{(\sum x_i^2)(\sum y_i) - (\sum x_i)(\sum x_i y_i)}{n\sum x_i^2 - (\sum x_i)^2},$ $\hat{b} = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2}.$

However, the model has been misfitted. These sample coefficients a_estimate and b_estimate may be biased under the Gauss measurement model, $Y_i = c + dx_i + eu_i + E_i,$ where the Ei are

independent with mean 0 and variance. The expectation of b_estimate need not be d. It can be shown that $\mathbb{E}(\hat{b}) = d + e\left(n\sum x_i u_i - \sum x_i \sum u_i\right) / \left[n\sum x_i^2 - (\sum x_i)^2\right]$.

The residuals then include both measurement error from Ei and model misfit error. If the root mean square is not small in comparison to variance, the residual sum of squares does not provide a good estimate of variance. Residual plots of (xi, ri) help to indicate whether the model is misfitted. Note that even when the model is misfitted, the average of the residuals r_estimate is 0. (Stat Lab, D. Nolan, T. Speed, p167-168)

Maximum Likelihood

In calibration problems, we have known design points, x1, ... xm, and it seems sensible to invert the bands about the regression line to provide an interval estimate of x0. Here we show that under the Gauss measurement model and the additional assumption of normal errors, x0 is the maximum likelihood estimate of x0, and the interval (x_lower, x_upper) is a confidence interval for x0; that is,

consider, for R=1, $\begin{aligned} Y_i &= a + bx_i + E_i \qquad i = 1, \ldots, m \\ Y_0 &= a + bx_0 + E_0, \end{aligned}$ where we now suppose that E0,E1, ...Em are independent normal errors with mean 0 and variance σ2. In our case, a, b, σ2, and x0 are unknown parameters.

An approximate 95% confidence interval for x0 is then the set of all x values that satisfy the inequalities $-z.975 \leq W(x) \leq z.975$. From these inequalities, we see that (x_lower, x_upper) can be interpreted as a 95% confidence interval for x0.

Theory

Regression	Regression Analysis is a statistical process for estimating the relationships among different variables. Regression models involve the following: 1. Independent Variable X 2. Dependent Variable Y 3. Unknown parameters such as β (scalar or a vector) Regression model relates Y to a function of X and β i.e.

$Y \approx f(X, \beta)$. Also note that $f(X, \beta) = E(Y|X)$.
$e_i = y_i - \hat{y}_i$ is the residual, which is the difference between the value of the dependent variable that is predicted by the model and the true value of the dependent variable.

Coefficient of Determination

R^2 is the coefficient of determination. This value can be used to determine the strength of the fit of a linear model. This is calculated by squaring the correlation coefficient. R^2 tells us what proportion of variability in the response variable is explained by the model. The rest of the variability is explained by variables not part of the model or randomness that is part of the data.

Residuals

Error are usually denoted with ε_i and the estimated residual is denoted with $\hat{\varepsilon}_i$. Residuals are estimates of model error, so we use $\hat{\varepsilon}_i$ check whether $\hat{\varepsilon}_i \sim N(0, \sigma^2)$. It is important to note that the variance of $\hat{\varepsilon}_i$ does not equal σ^2. On the other hand, the variance of the standardized residual is σ^2 which implies that we will use standardized residuals to check models.
$\hat{\sigma^2} = \frac{1}{n-\gamma} \sum_{i=1}^{n} \hat{\varepsilon}_i^2$, γ = number of parameters. If γ equals to n then $\hat{\sigma^2}$ will converge to infinity. This indicates that $\hat{\sigma^2}$ is a poor prediction and would mean 'overfitting'.

Least Squares Regression

Least Squares Regression (LSR) is a predictive model of a dependent variable Y given an independent variable X. In the bivariate case, X and Y are both sets of scalar values, however, in the multivariate case, X can be a set of vectors. In any case, the model attempts to fit a linear function, called the regression line, to the data X and Y by minimizing the mean of the squared residuals (i.e., the distance between the regression line and the true data). In squaring the residuals, large residuals are assigned a higher weight in the associated optimization problem. In this sense, Least Squares Regression is an estimator for the conditional mean of the response variable Y. Least Squares Regression guarantees uniqueness of the solution, as the optimization problem is strictly convex.

Least Absolute Deviations Regression

Least Absolute Deviations Regression is a predictive model similar to Least Squares Regression where the absolute value

	function is used instead of squaring the residuals. In this sense, Least Absolute Deviations Regression is an estimator for the conditional median of the response variable Y. The solution is not guaranteed to be unique, as the optimization problem, while convex, is not strictly convex.
Quantile Regression	Quantile regression models the relation between a set of predictor variables and specific percentiles (or quantiles) of the response variable. It specifies changes in the quantiles of the response.
Confidence Interval	Confidence intervals give us a range of plausible values for some unknown value based on results from a sample. This topic covers confidence intervals for means and proportions.
Prediction Interval	A prediction interval is a type of confidence interval (CI) used with predictions in regression analysis; it is a range of values that predicts the value of a new observation, based on your existing model.
Conditions for Linear Model	Linearity: The relationship between the explanatory and the response variables should be linear. Checking using a scatterplot or residual plot is preferred. Normality: 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. Constant Variability: The variability of points around the least squares line should be nearly constant. This means that the variability of residuals around the 0 line should be roughly constant too.
Hypothesis Testing (Linear Regression)	The hypothesis test for linear regression tests the parameters of the model. The null hypothesis states that the parameters should be 0 while the alternatives state that the parameters should not be equal to 0. So small p-values on parameters can tell us how important or how needed these are for prediction.
Cross-Validation	Cross Validation aims to provide stronger guarantees of generality for a given statistical or machine learning model. In practice, this is often done by partitioning data into training,

validation, and testing sets. The training data is used to fit the model, while the validation set is used to tune additional parameters of the model called hyperparameters. The test set is used to evaluate the final model, and no decisions regarding the parameters of the model should be made on the basis of performance on the test set. Cross validation refers specifically to the training and validation sets, however. In cases in which the data are more scarce, k-fold cross validation may be used, in which the data is split into k partitions, and k different models are fit to the data, holding out a different element of the k partitions as the validation set. The performance of the k models evaluated on their respective validation sets are averaged to produce a final validation set performance result.

References

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

Colli, M., Rasmussen, R., Thériault, J. M., Lanza, L. G., Baker, C. B., & Kochendorfer, J. (2015). An Improved Trajectory Model to Evaluate the Collection Performance of Snow Gauges. *Journal of Applied Meteorology and Climatology,54*(8), 1826-1836. doi:10.1175/jamc-d-15-0035.1

Colli, M., Lanza, L. G., Rasmussen, R., & Thériault, J. M. (2016). The Collection Efficiency of Shielded and Unshielded Precipitation Gauges. Part I: CFD Airflow Modeling. *Journal of Hydrometeorology,17*(1), 231-243. doi:10.1175/jhm-d-15-0010.1

Goodison, B. E. (1978). Accuracy of Canadian Snow Gage Measurements. *Journal of Applied Meteorology,17*(10), 1542-1548. doi:10.1175/1520-0450(1978)0172.0.co;2