Math 189 - Case Study #4

Calibrating a Snow Gauge

March 16th, 2019

Name	PID
Erin Werner	A12612584
Emma Choi	A12635909
Talal Alqadi	A13816618
Ella Lucas	A13557332
Samantha De La Torre	A13300273

1.) Introduction

Northern California's main source of water comes from the Sierra Nevada mountains. The Forest Service of the United States Department of Agriculture (USDA) monitors the water supply by operating 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 measurement process does not disturb the snow, so the same snowpack can be measured multiple times. These replicate measurements on the same volume of snow allow researchers to study snowpack settlement over the winter season and the effects 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 snowpack the less water it can absorb, so analyzing the snowpack profile might help with managing the water supply and flood management. Although the gauge doesn't directly measure snow density, it's output is converted from a measurement of gamma ray emissions to a density reading. Due to instrument wear and radioactive source decay, the functions used to cover the measured values into density readings may change over seasons. There is an annual calibration run at the beginning of winter season to adjust the conversion method . The goal of this lab is to provide a simple procedure for converting gain into density when the gauge is in operation.

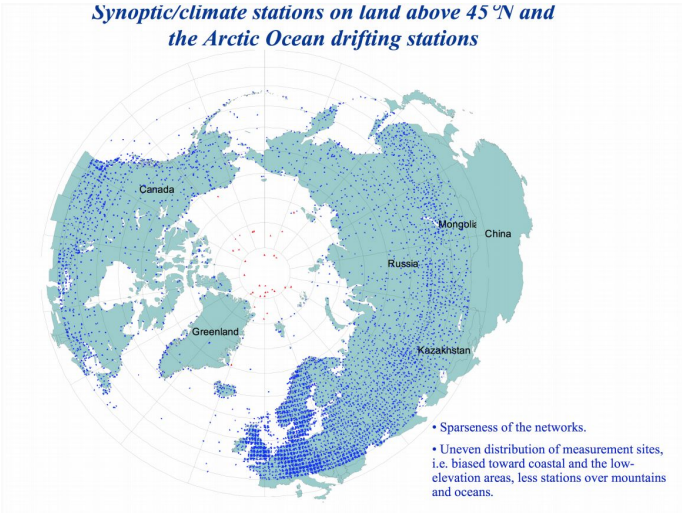
2.) Data

The collected data is from a calibration run of the USDA Forest Service's snow gauge, which is located in the Central Sierra Nevada mountain range near Soda Springs. The run consisted of placing polyethylene blocks of known densities between the two poles of the snow gauge and taking readings on the blocks. The polyethylene blocks were used to simulate snow. For each of the polyethylene blocks, 30 measurements were taken. But, only the middle 10 are reported here. The measurements reported are amplified versions of the gamma photon count made by the detector. So, we will call the gauge measurements the "gain". The available data here consists of 10 measurements for each of the 9 densities in grams per cubic centimeter of polyethylene.

Yet, there are some challenges that come with the data. One challenge is the operational networks, which serve as our knowledge base. There was a decline of the networks in the northern regions, including Siberia, Alaska and Northern Canada. There are also very few stations in the mountain regions. So we need to consider how to sustain and improve the operational networks. Another challenge is the data quality and compatibility across national boundaries. There are large biases in gauge measurements of solid precipitation. This results in incompatibility of precipitation data due to difference in instruments and methods of data processing. There are also difficulties to determine precipitation changes in the arctic regions. A final challenge is then the validation of precipitation data, including satellite and reanalysis

products and fused products at high latitudes.

Figure 2.1. Locations of climate stations that reflect the challenges that come with this data.



3.) Background

Since the gauge is a very complex and expensive instrument, it is not feasible to establish a broad network of gauges in the watershed area to monitor the water supply. It's used primarily as a research tool. It has helped to study snowpack settling, snowmelt runoff, avalanches and rain-on-snow density. The gauge in California and from our data is located in the center of a forest opening that is about 62 meters in diameter. The laboratory site is at 2099 meters elevation and is subject to all major high altitude storms which usually deposit 5-20 cm of wet snow. The average depth of the snow-pack is 4 meters each winter.

The snow gauge consists of a cesium-137 radioactive source that emits gamma photons in all directions, and an energy detector that counts those photons eating through the 70cm gap from the source to the detector crystal. The lift mechanism at the top of the poles raises and lowers the source and detector together. The radioactive source emits gamma photons also called gamma rays at 662 kilo-electron-volts (keV) in all directions. The detector contains a scintillation crystal which counts those photons eating through the 70-cm gap from the source to the detector crystal. The pulses generated by the photons that reach the crystal are transmitted by a cable to a preamplifier and then further amplified and transmitted via a burial coaxial cable to the lab. The signal is stabilized, corrected for temperature drift, and converted to the "gain." It should be directly proportional to the emission rate. The typical snowpack density is between 0.1 and 0.6 g/cm3.

The radioactive source emits gamma rays in all directions. The ones towards the detector may be scattered or absorbed by the polyethylene molecules between the source and the detector. The denser polyethylene with stop more gamma rays from reaching the detector. There are

complex models for the relationship between polyethylene density and the detector readings, but a simplified version may be workable for the calibration problem of interest. A gamma ray onroute to the detector passes a number of polyethylene molecules and that number depends on the density of the polyethylene. A molecule may either absorb the gamma photon, bounce it out of the path to the detector, or allow it to pass. The chance of the gamma ray successfully arriving at the detector is p^m where p is the chance a single molecule will neither absorb nor bounce the gamma ray, and m is the number of molecules in a straight line path from the source to the detector. The density, x, is proportional to m the number of molecules. The probability can be expressed by the formula below:

$$e^{m \log p} = e^{bx}$$

4.) Investigations

The goal of this lab is to provide a simple procedure for converting gain into density when the gauge is in operation. This experiment was conducted by varying density and measuring the response in gain, but when the gauge is ultimately in use, the snow-pack density is to be estimated from the measured gain.

a.) Fitting - Linear Model

i.) Fitting Gain and Density

To begin our investigation, we generated a scatter plot that consists of the 90 sets of gains and densities, along with the fitted linear regression line, shown in Figure 4.1.1. By applying the linear regression model to the data, the corresponding red line represents the minimized squares of residuals as a linear relationship of gains and snow density. In order to check the fit of the linear regression line, Figure 4.1.2, 4.1.3, and Figure 4.1.4 check for the linearity between the response and explanatory variables, for the normality of residuals of the regression, and the variability of points around the least square line.

Figure 4.1.1. Linear regression.

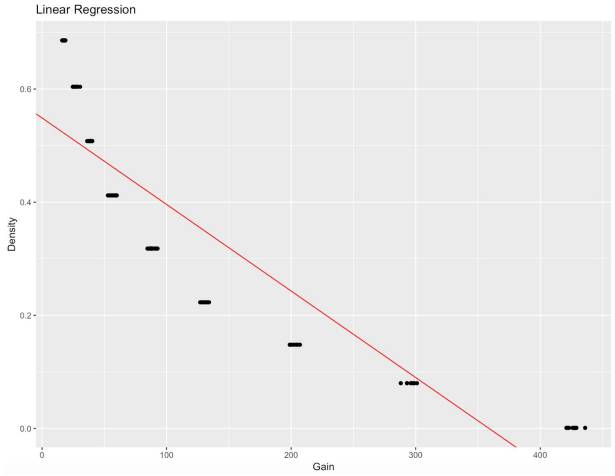


Figure 4.1.2. Residual plot.

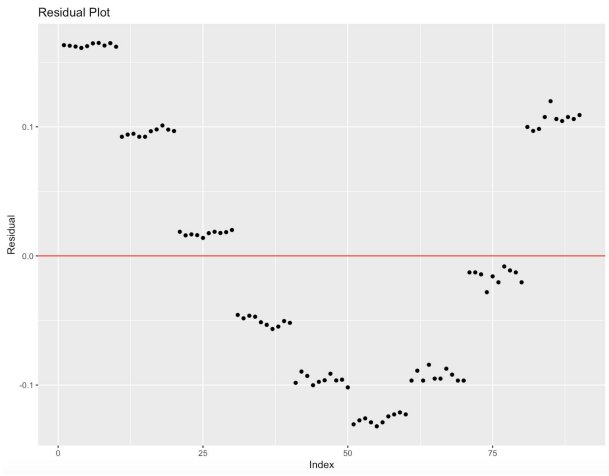


Figure 4.1.3. Histogram of residuals.

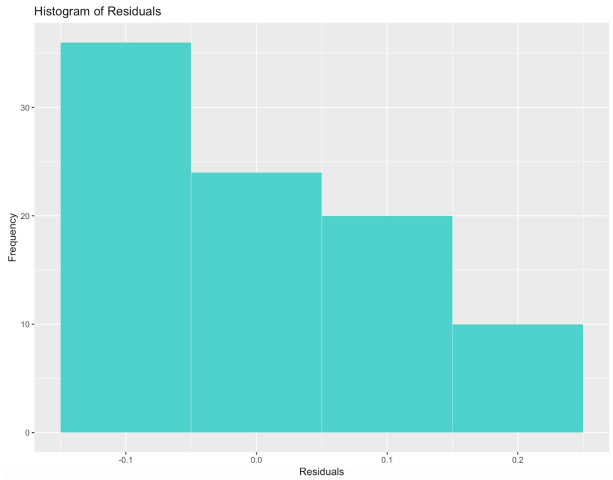
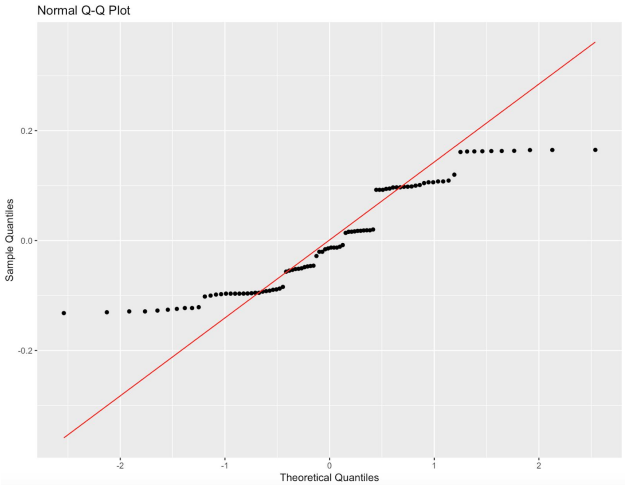


Figure 4.1.4. Q-Q plot of residuals.



First, Figure 4.1.1 breaks the first condition of linearity between gain and density as the data appears to be parabolic, which is significantly nonlinear. Both Figure 4.1.2 and Figure 4.1.4 reveal that the data does not fit the normal line (represented in red) very well. This breaks condition two. Furthermore, Figure 4.1.3 reveals a histogram that is nonsymmetric with a right skew. This reinforces the notion that the data does not follow a normal distribution. As a result, condition three is not met when looking at Figure 4.1.2 because there is a strong nonrandom pattern about the zero slope line, indicating that the data is heteroskedastic.

ii.) Fitting Log(Gain) and Density

In order to try and find a better fit, the densities and the log of gains were fit into a linear regression model. The results are plotted as a scatterplot in Figure 4.1.5. It represents the 90 sets

of log transformed explanatory variable with the density, as well as its resulting fitted least squares line. Once again, Figures 4.1.6, 4.1.7, and 4.1.8 verify the fit of the linear regression by checking for the linearity between the response and explanatory variables, the normality of residuals of the regression, and the variability of points around the least squares line.

Figure 4.1.5. Transformed linear regression.

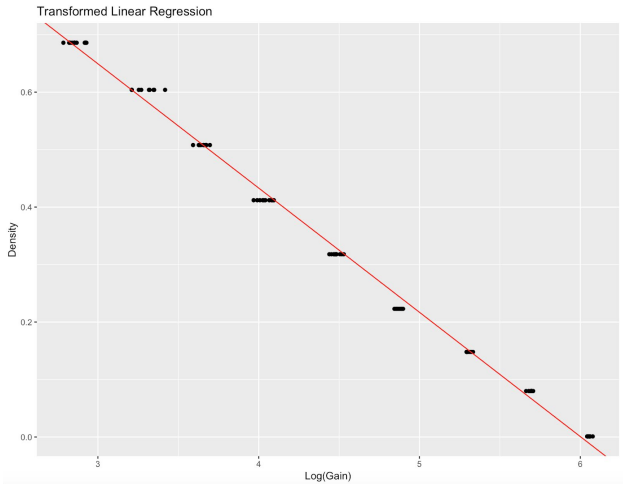


Figure 4.1.6. Transformed residual plot.

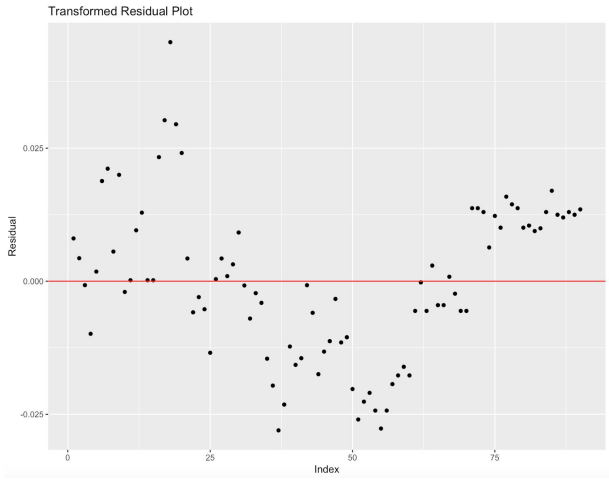


Figure 4.1.7. Transformed histogram of residuals.

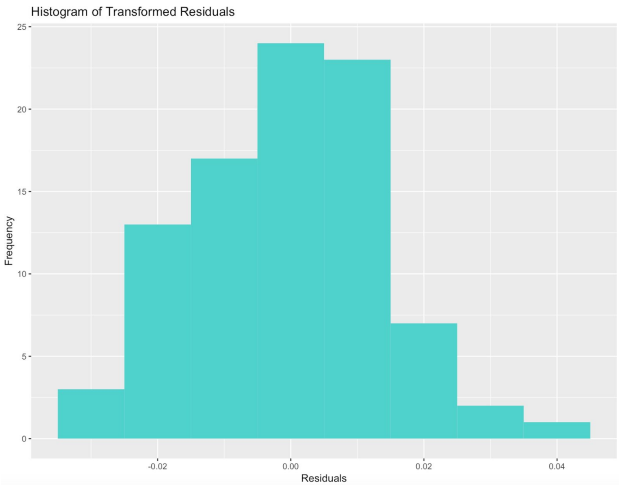
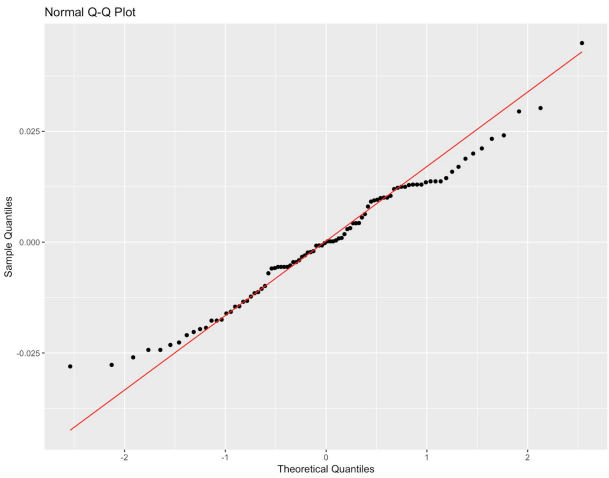


Figure 4.1.8. Transformed Q-Q plot.



First, Figure 4.1.5 shows that the first condition is now satisfied, as the transformed data fits a linear model more closely compared to the raw data. But, Figure 4.1.7 reveals that the residuals still do not follow the normal distribution because the histogram is slightly skewed. Also, Figure 4.1.8 shows that the Q-Q Plot still does not completely follow the normal line. Yet, the differences are small enough to conclude that the second condition is roughly followed. As there is no strong relationship between the explanatory variables on the x-axis and the residuals seen in Figure 4.1.6, condition three, concerning homoscedasticity, is satisfied. Therefore, a

fitted linear regression model can apply to the relationship between log(gains) and the densities of the data.

- *Do the residuals indicate any problems with the fit?*

The fit of gain and densities is insufficient due to heteroskedasticity and the non-normal distribution of the residuals, as demonstrated in the plots above. But, the transformed data of log(gain) and density satisfies these conditions roughly enough to conclude that a fitted linear regression model applies.

- *If the density of the polyethylene blocks are not reported exactly, how might this affect the fit?*

The densities that are reported vary slightly as all differences are under 0.0001. So, the model densities are rounded up to 0.001. As a result, the least squares straight line would not change drastically, especially its slope.

- *What if the blocs of polyethylene were not measured in random order?*

By not measuring the blocks randomly, it would be possible for a confounding factor to disrupt the results of the data. For example, if the data were to be sorted from high to low density, a confounding factor due to environmental changes or the order of the measurements can cause a resulting relationship between the density and gain that may not be true. The log(gain) and density relationship has a negative linear relationship, so it is fair to assume the relationship was due to taking the measurements in a specific order of density, rather than a true negative correlation. Thus, randomness within the experiment is important.

b.) Prediction and Interval Estimates

The fitted linear regression line had an estimated value of -0.216203 for the coefficient on log(gain). This can be interpreted to mean an increase of log(gain) by 1 results in an estimated decrease in the density by 0.216203 g/cm^3. The negative slope for this regression makes sense because an increase in gain means more gamma photons were able to pass through the snow and reach the detector. This indicates that the snow has a lower density of molecules that could block the photons from reaching their destination. The y-intercept of the linear regression model is estimated to be 1.298013. The exact interpretation of the intercept is that at a log(gain) reading of 0, the density of the snow would be 1.298013 g/cm$_3$. This interpretation is not useful though, since snow would never be this dense.

The values for the slope and the intercept help compile the following equation to predict density(y) given gain(x): $y = -0.216203log(x) + 1.298013$. The gain readings of 38.6 and 426.7 can be plugged directly into this equation to predict the density of the snow. The density for a gain reading of 38.6 is 0.5081678 g/cm^3, and the density for a gain reading of 426.7 is

-0.01133153 g/cm^3 . We know density cannot by negative, so the negative value for the gain reading of 426.7 can be interpreted as the density being 0 g/cm^3, meaning there is no snow on the ground.

The log(gain) model and density is approximately linear, so it is easy to fit confidence bands. The goal is to predict the density given a gain reading, and as a result prediction intervals will be used rather than confidence intervals. Prediction intervals are used to determine where the next data point sample should fall, while confidence intervals are used to determine where the population mean likely lies. A 95% prediction interval can be interpreted as the range of values where the next data point should fall 95% of the time. The result of the prediction intervals is a lower and upper bound for each value of gain, giving us Figure 4.2.1.

Figure 4.2.1. Prediction interval of density given log(gain).

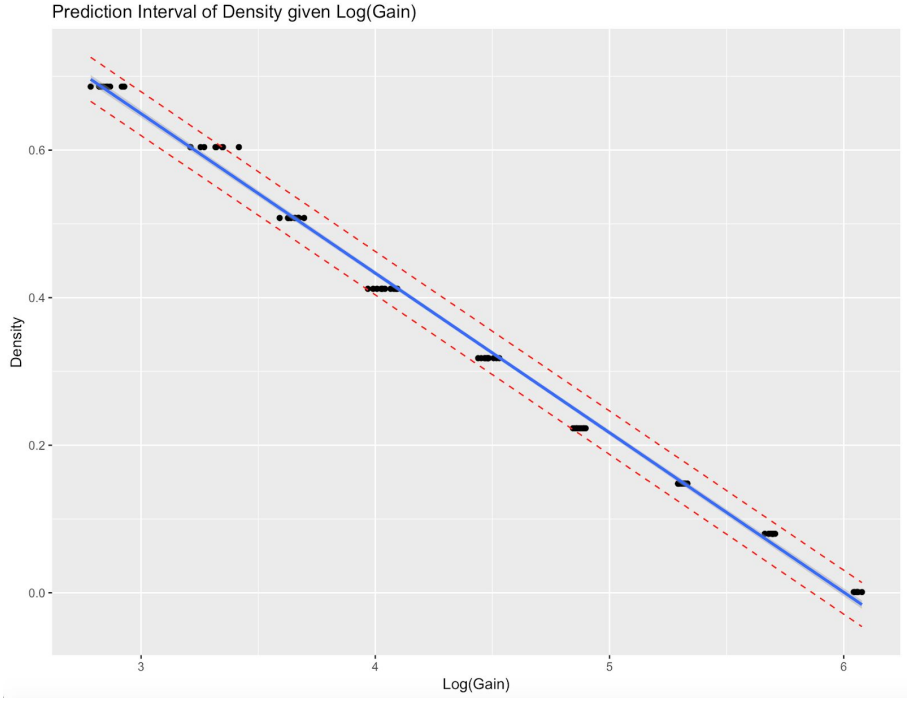


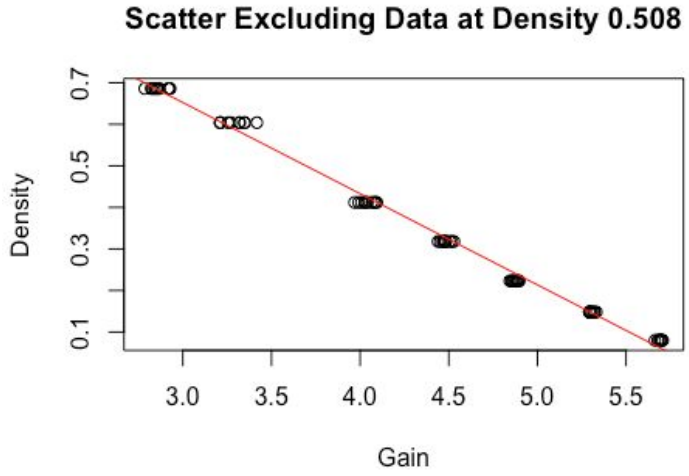
Figure 4.2.1 illustrates the relationship between gain values and the density of the snow. The log transformation is applied to the gain values to create a linear model. The blue line is the fitted regression line, and the two red dashed lines are the lower and upper bounds using the 95% prediction interval.

The gain value of 38.6 in the prediction function gives us an interval of [0.4786626, 0.5376729]. Hence, we are allowed to say if there is a gain reading of 38.6, there is a 95% chance that the density of the snow falls between 0.4786626 and 0.5376729 g/cm^3 . If the gain reading is 426.7, the interval will then be [-0.04110982, 0.01844676]. Since negative values indicate an actual value of 0, this interval can be shortened to [0, 0.01844676].

c.) Cross Validation

The fitted regression line from the previous section produced the following equation for predicting the density (y), given the gain(x): $y = -0.216203 log(x) + 1.298013$. In order to check the accuracy of our prediction method, we omitted the set of measurements at the block of density at 0.508, applied our estimation method, and produced an interval estimate for the density of a block with an average reading of 38.6. *Figure 4.3.1* shows the scatter of data when omitting the data corresponding to the density at 0.508 with a linear regression line, which was added due to the clear linear correlation. The equation of the linear regression line did not change very much when the block of data at density 0.508 was excluded, it becomes $y = -0.219531 log(x) + 1.310849$.

Figure 4.3.1. Scatter plot.



Two conditions were checked to see how well the regression line in *Figure 4.3.1* fit the data, excluding the block corresponding to the density 0.508: (1) the points in the residual plot are dispersed randomly above and below the line x=0. (2) the residuals from the data do align with normal data. The residual plot in *Figure 4.3.2* shows that the residuals are randomly scattered around the horizontal axis and therefore indicates that linear regression is a good fit. The qq plot in *Figure 4.3.3* and the histogram of residuals in *Figure 4.3.4* both serve to compare the residuals to normal distributions because if the residuals are normal then linear regression is a good fit. Based on the distributions present in those figures we concluded that the snow gauge data excluding the block corresponding to density 0.508 follows a linear model.

Figure 4.3.2. Residual plot.

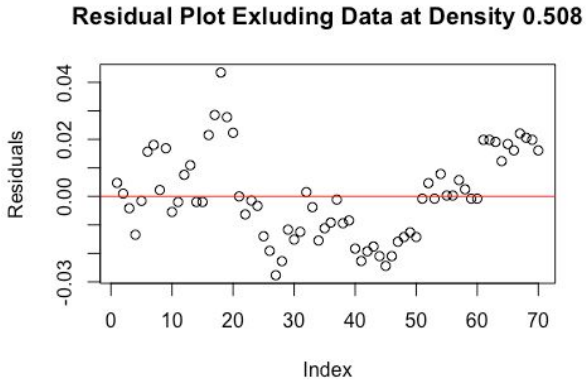


Figure 4.3.3. QQ plot.

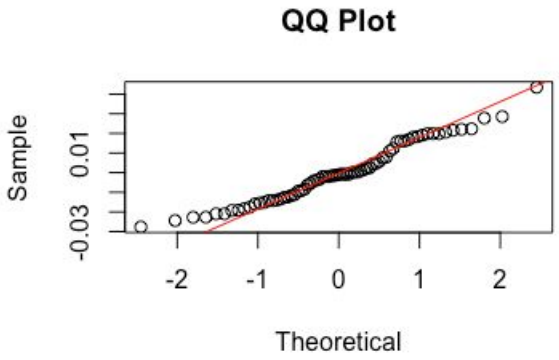
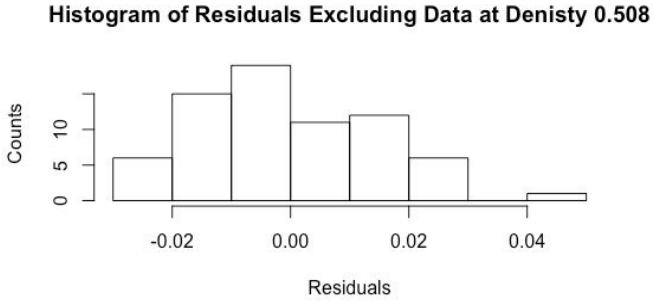


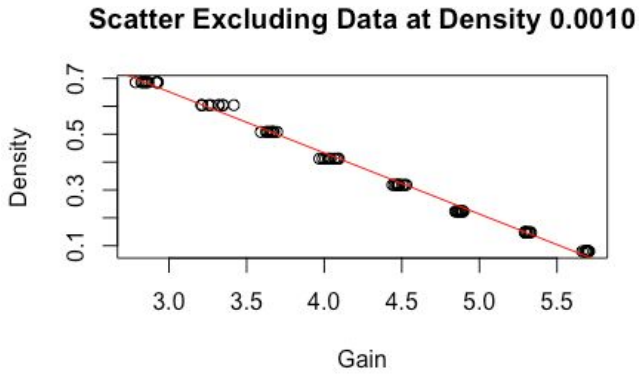
Figure 4.3.4 Histogram of residuals.



After determining that a linear model was, in fact, a good fit for the data excluding the values at density 0.508, we created a 95% prediction interval using 38.6 as the average. This means that the density corresponding to the gain value of 38.6 is 95% likely to fall within the interval [0.4777083 0.5399846]. When 38.6 is plugged into the equation for the regression line in *Figure 4.3.1* the resulting density is 0.5088469, which is very close to 0.508, the value in the original data set corresponding to a gain of 38.6 that we omitted in this section to test our prediction method. Additionally, the value 0.5088469 falls in the middle of the prediction interval. This makes sense because we set the interval estimate using an average of 38.6.

To further test our prediction method, the same tests were done excluding the data corresponding to the density at 0.0010. The regression line used to fit the data in *Figure 4.3.5* has the equation y = -0.219395*log*(x) +1.310114.

Figure 4.3.5 Scatter plot.



The random distribution of residuals in *Figure 4.3.6* and the alignment of the residuals with a normal distribution in *Figures 4.3.7* and *4.3.8* indicate the linear model fits the data, even when excluding the data corresponding to densities at 0.0010.

Figure 4.3.6. Scatter plot.

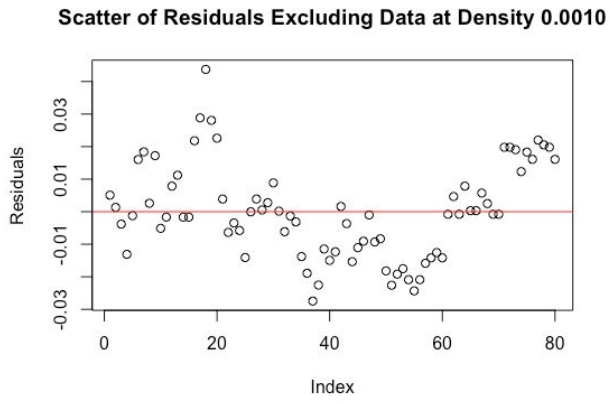


Figure 4.3.7. QQ plot.

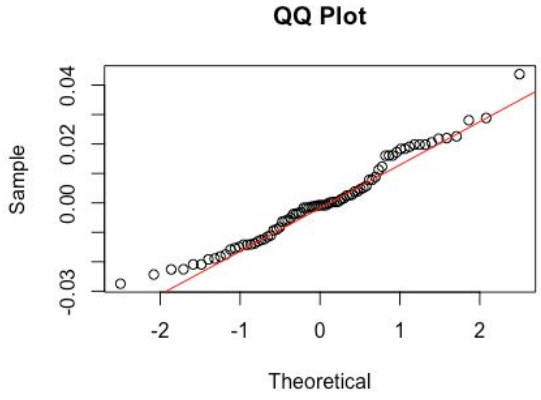
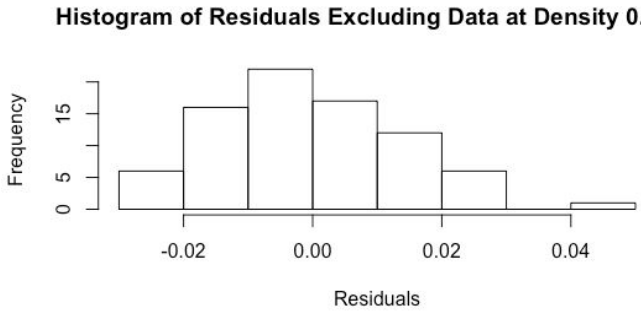


Figure 4.3.8. Histogram of residuals.

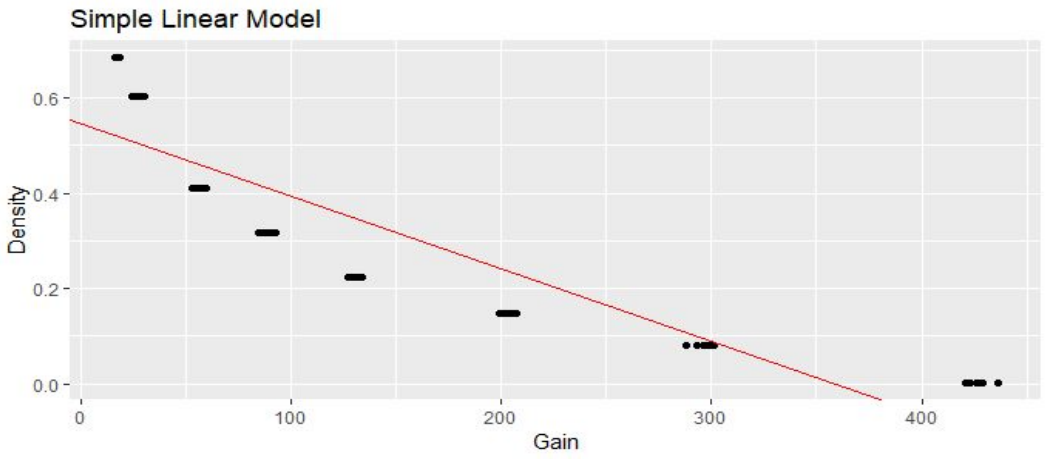


Since it was established that the data still follows a linear model when excluding the block of data corresponding to the densities at 0.0010 a 95% prediction interval can be found using 426.7 as the average for the interval estimate. The prediction interval is [-0.04844409, 0.01132649] meaning that density corresponding to the gain 426.7 is 95% likely to fall within that interval. The predicted density for the gain 426.7 is -0.01855993, which lies in the middle of the prediction interval and is fairly close to 0.0010, the observed density at gain 426.7.

d.) Alternative Models

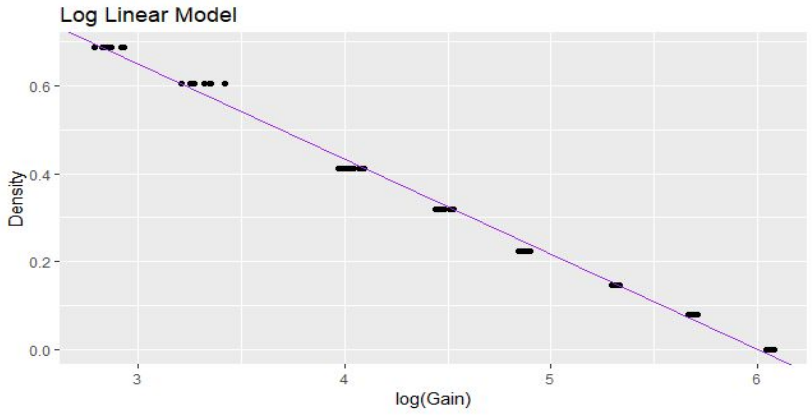
Other alternative regression models can be applied to test the fitness of that predictive model. Similar to the previous section, Cross Validation, density points with 0.508 are excluded out of the training set. Four linear regression models are attempted in this section: Simple Linear Model, Log Linear Model, Polynomial Regression, and Local regression (or LOESS).

Figure 4.4.1. Scatter plot with fitted line in simple linear model.



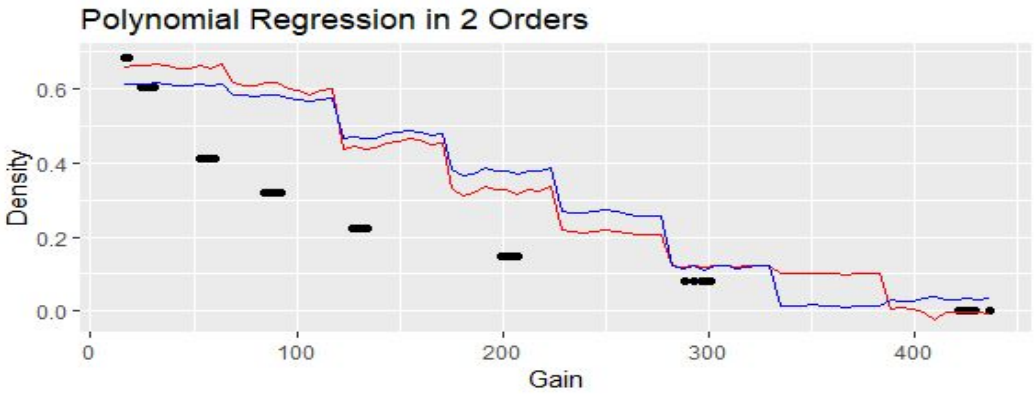
In Figure 4.4.1, the simple linear regression model is applied to the training set. With adjusted R-squared = 0.7986, this predictive model has a 79.86% variation in density when compared to variation in gain. The p value is equated to be extremely small, which indicates that changes in density's value may be related to gain's changes. In addition to that, the mean squared error was equated to 0.01, not a very low value. The summary statistics suggest an average predictive model from the simple linear regression. But, when looking at Figure 4.4.1, it is easy to notice that this model is non-linear, and doesnt fit.

Figure 4.4.2. Scatter plot with fitted line in log linear model.



The second model attempted is the log linear model. A quick look at Figure 4.4.2 shows the upgrade in predictiability when compared to the simple linear model. Summary statistics show that adjusted R-squared is calculated to be 0.9955, meaning that this model can explain 99.55% of the variance in the training set. In addition to that, MSE = 0.000233, indicating that there is a higher accuracy of prediction than the simple linear model. Figure 4.4.2 shows that the scatter plot of the log model shows a linear relationship, and a high chance of predictability.

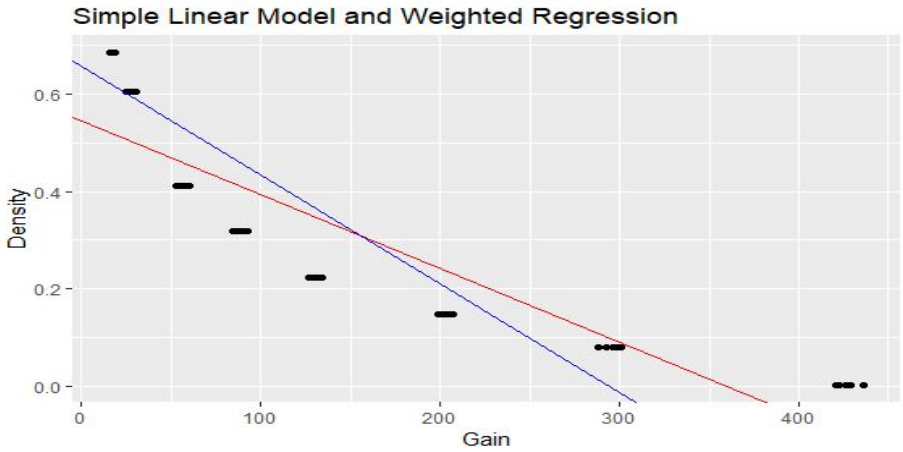
Figure 4.4.3. Scatter plot with 2 fitted lines in Polynomial Regression model, second degree shown in blue, and third degree shown in red



A different type of regression by looking at higher orders can help explain the data when the linear relationship is lacking. Using polynomial regression, we calculate the R-squared in the second order. R-squared was calculated to be 0.9469 with MSE = 0.002718294. The second order of the polynomial model is statistically significant, and the calculated R-squared is a good predictability value. The third order of the polynomial regression leads us to an R-squared = 0.9907, and a MSE = 0.0004. The third order R-squared and MSE is definitely better than the second order, which is also confirmed in Figure 4.4.3. In addition to that, when compared to

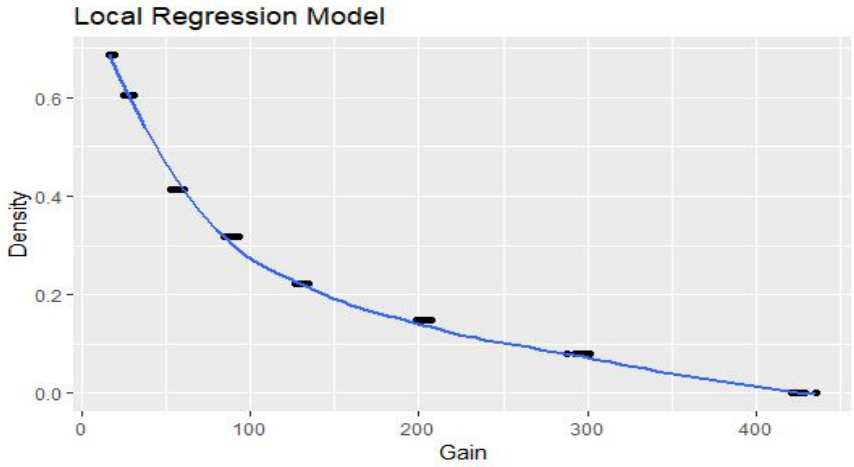
R-squared of log linear model, third order is a contender, with 99.55% vs 99.07 %, respectively. Although both summary statistics show a promising model, Figure 4.4.3 shows that both orders aren't a good fit. This allows us to confirm that the log linear model is the best fit up until now.

Figure 4.4.4. Scatter plot with a fitted line in simple linear model, shown in red, and weighted regression, shown in blue.



To counter model misfitting, a weighted regression model can be used with assigned weights to have an equal effect on the independent variable. In this case, with gain and density have a linear relation, weights can be assigned to be reciprocal of gain. R-squared is assigned to be 0.7742 with a mean squared error of 0.0001. Although the MSE is a good value, R-squared is not. The latter can be seen in Figure 4.4.4, where it is easy to see that the weighted regression is not a very good fit of a line when compared to the simple linear regression model, which itself proved to be a bad fit earlier.

Figure 4.4.5. Scatter plot with fitted line in local regression.



Similar to the methods above, when a linear relationship is not applicable and simple regression models are not fitting accurately, a Local Regression (LOESS) is very useful. In this local regression model, R-squared is calculated to be 0.9973, and the MSE is calculated to be 0.00014. Both, indicating this line is a good fit. Thus, we graph the points with the fitted line in local regression in Figure 4.4.5. When compared to the log linear model, the Local Regression Model has a better R-squared value with 0.9973 vs 0.9955, and a better MSE with 0.0001 vs 0.0002. Both models seem really similar, but with the slightly better MSE and R-squared values, the Local Regression model is accepted as the best fit model.

5.) Theory

a.) Simple Linear Model

The objective of this case study was to identify if the snow gauge data fit a linear model in order to determine a method of predicting the response variable, density, given the explanatory variable, gain. The correlation between two variables can be described by their correlation coefficient, which falls between -1 and 1, where 0 means they have no correlation and 1 and -1 means they are 100% correlated. Three conditions must be met for a prediction method using a linear model to hold (1) there must be a linear relationship between the explanatory and response variables (2) the residuals must align with a normal distribution and (3) the residuals must be variable around the horizontal axis. The linear model is as follows,

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

where Yi = alpha + beta(Xi) is the line in y-intercept form showing the relationship between the response variable (Yi) and the explanatory variable (Xi), and epsilon is the error.

b.) Residual Plots

Residual plots display the differences between the actual data and the modeled data. The real data is the fit with the addition of the residual. The best fit line has the smallest residuals. So, the least-squares line, which is the minimum error sum of squares, is the best way to model that.

Residual:

$$e_i = Y_i - \hat{Y}_i$$

Least-Squares Line:

- Minimize Error Sum of Squares (SSE):

$$\mathrm{SSE}(b_0, b_1) = \sum_{i=1}^{n} \left[Y_i - (b_0 + b_1 X_i)\right]^2$$

- Result: Notice the least-squares regression line passes through the means of X and Y:

$$\hat{\beta}_1 = \frac{SS_{XY}}{SS_X} \qquad \text{and} \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

c.) Prediction Intervals

Given the level of an explanatory variable the prediction interval is used to calculate a range of values that the response variable will fall into. The prediction interval is usually wider than the confidence interval since it accounts for the variability in each variable and the uncertainty of the mean. Therefore encompassing all of the predicted values, rather than just the values that are likely. The prediction interval is calculated using the following equation:

$$\hat{y}_h \pm t_{(\alpha/2, n-2)} \times \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right)}$$

Where MSE is the mean square error, y-hat is found using the fitted regression equation, and the t-value is calculated with n-2 degrees of freedom.

d.) Model Misfit

In the event that the established conditions required to fit a linear regression model are not met, then the model is said to be a misfit for the data. In this section, we will digress into details on the existing conditions, which will aid in recognizing when a misfit occurs.

i.) Polynomial Regression

In statistics, polynomial regression is a form of analysis in which the relationship between the independent variable (x) and the dependent variable (y) is modeled as an nth degree polynomial in x. This fits a nonlinear relationship between x and the conditional mean of y, E(y|x). Since the regression function, E(y|x) is linear in unknown parameters estimated from the data, polynomial regression is considered a special case of multiple linear regression.

$$\mathbb{E}(Y \mid X = x) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p$$

Polynomial regression model in its general form.

Notice how the model has properties by appearance similar to a linear combination of x values, which indicates the ability to transform the function to its matrix form as a system of linear equations.

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^m \\ 1 & x_2 & x_2^2 & \dots & x_2^m \\ 1 & x_3 & x_3^2 & \dots & x_3^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^m \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

Polynomial regression model as a system of linear equations.

Features:

- x: design matrix
- Y: response vector
- B: parameter vector
- E: vector of random errors

$$\hat{\vec{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\,\mathbf{X}^\mathsf{T}\vec{y},$$

The vector of estimated polynomial regression coefficients using OLS.

In linear regression, coefficients are assessed using the least squares method. Due to the aforementioned property of the polynomial regression function being linear in unknown parameter estimations, coefficients are likewise assessed using the least squares method:

$$\begin{aligned} \text{minimize} \quad & \mathrm{SSE}(b_0, b_1, \dots, b_p) = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i - \cdots b_p X_i^p)^2 \\ \text{over} \quad & b_0, b_1, \dots, b_p \in \mathbb{R} \end{aligned}$$

Depiction of the least squares method for estimating polynomial regression coefficients.

Resulting solutions are unique if and only if the number of unique X's is greater than or equal to the number of parameters + 1.

Although polynomial regression is viewed as a special case of multiple linear regression, interpreting the polynomial regression model once it has been fit to the data takes a different perspective. Interpreting the effects of individual coefficients is difficult, since the underlying monomials can be highly correlated, which makes it difficult to isolate a specific effect. It is generally more informative to consider the regression function as a whole, once it's fitted. At that point, confidence bands can be employed to determine uncertainty. To highlight this means to assess the model fit (or misfit), consider the following example:

- Quadratic model, E[Y|X]=a+bX+cX^2
- Multivariate linear representation, set U=X^2, E[Y|X,U]=a+bX+cU

- Estimate b using least squares on linear model, E[Y|X]=d+eX (idea is that b and e should be the same if U and X are not correlated).
- This idea will be proved incorrect, because U=X^2, so the estimate of e is biased.
- E[b^] = b + c (n $\sum_{i=1}^{n}$ XiUi - ($\sum_{i=1}^{n}$ Xi) ($\sum_{i=1}^{n}$ Ui)) / (n $\sum_{i=1}^{n}$ Xi^2 - ($\sum_{i=1}^{n}$ Xi)^2) tends towards 0 if Xi and Ui are not correlated.

ii.) Weighted Regression

Weighted linear regression exits as a general form of ordinary least squares where the errors covariance matrix is allowed to differ from an identity matrix. This idea violates the fundamental assumption that the least squares model is the homoskedasticity of the error term (estimated by the residual of the model). This assumption ordinarily keeps the information in each data point equal. Unsurprisingly, this level of consistency does not always translate to data in practice - including with the data motivating our study. From the analysis contained in Investigations, we can see the relationships between variables as they are fitted to the model. It can be seen that the residuals of the linear model seem to present a pattern, so we conclude that the independent variable (gain) influences the dependent variable (density), so the variance must be unequal. To handle this data, we assign a weight to each data point as is the practice with weighted regression. When the errors are uncorrelated with each other and the variance is equal, b^ is a best linear unbiased estimator (BLUE). The minimization process differs in that we no longer minimize the residual sum of squares as in the single linear model, rather the computation looks like this:

$$\text{minimize} \quad \mathrm{SSE}(b_0, b_1, \ldots, b_p) = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i - \cdots b_p X_i^p)^2$$
$$\text{over} \quad b_0, b_1, \ldots, b_p \in \mathbb{R}$$

Formula for the minimization of the weighted sum of squares.

Using the weighted least squares model with our data improves accuracy when dealing with heteroskedastic data. Assigning larger weight values to unequal data when fitting a linear model ensures that the fitted values depend more on those points, and the idea is that this will correspond more closely to the pattern the data follows. Weights being inversely proportional to the variance at each level yields the most precise parameter estimates possible. Furthermore, ordinary least squares requires that the error terms are iid, yet when faced with heteroskedasticity OLS would not be efficient as it would no longer be the maximum likelihood estimator. To correct this and obtain an efficient MLE, it is necessary that the variance of each epsilon is known and that the weights are set to 1/(sigma)^2.

By design, weighted regression also comes with some noteworthy disadvantages. It is worth mentioning again that the biggest assumption is also the biggest disadvantage, because it is so definitive - the theory behind the method is all based on the assumption that weights are exactly known. In real applications, this is almost never the case, so estimated weights are used instead, which can affect the results of analysis when weights must be estimated using only a few observations.

iii.) Local Regression

LOESS curve fitting (local polynomial regression) is a method for fitting a smooth curve between two variables, or fitting a smooth surface between an outcome and up to four predictor variables. The method is nonparametric as linearity assumptions of conventional regression methods have been relaxed. Instead of estimating specific parameters, the focus is around fitted points and their standard errors estimated with respect to the whole curve. This enables investigators to measure both overall uncertainty as well as how well the estimated curve fits the population curve. The function describes variation of data in different regions. An aspect that makes this method advantageous is that it does not require a global function to fit a model, only to fit localized subsets of the data. The assumptions are as follows:

- The mean of y can be approximated around point x by a small class of parametric functions in polynomial regression.
- Errors in estimating y are iid with a mean of 0.
- Bias and variance are traded off by the settings of span and degree of the polynomial.

The mathematical representation of local regression:

$$\min_{\beta} \sum_{i=1}^{n} w_i(x)(y_i - \vec{x} \cdot \beta(x))^2$$

Optimizing data with local regression.

Another feature of local regression is that the weight function must exist proportionally to some kernel function. While there are variations, the most common choice is tri-cubic kernel:

$$K(x_i, x) = \left(1 - \left(\frac{|x_i - x_0|}{h}\right)^3\right)^3$$

Tri-cubic kernel function to enforce that f is smooth as a function of x0.

While the model is advantageous because it doesn't require the specification of a function to fit a model to all of the data in the sample, making it very flexible, it makes less efficient use of data compared to other least squares methods. LOESS curve fitting requires large, densely sampled data to produce good models. This is due to the fact that the method relies on the local data structure when performing the local fitting. This makes the fitting more costly, and prevents there from being a simple mathematical formula to represent the model.

e.) Transformation

In cases where the two variables are not linear, we can apply a transformation to make them so. Since the linear relationship in the simple linear model showed that they were not a linear fit, we can apply a log transformation to the linear equation, similar to the log linear model. So, with G representing gain and x representing density:

$$F(G) = a * e^{bx}$$

By applying log on both sides:

$$log(F(G)) = log(a * e^{bx}) \rightarrow log(F(G)) = log(a) + bx$$

Thus, we now have a linear relationship between the two variables. To adjust for measurement errors with Gauss measurement error model, we add Us to the previous equation:

$$log(F(G)) = log(a * e^{bx} + U)$$

where U is the error changes with respect to x. However, this does not assist us. We try multiplying the error W instead:

$$log(F(G)) = log(a * e^{bx} W) \rightarrow log(F(G)) = log(a) + bx + \log(W)$$

Now this equation seems most similar to the gauss model.

f.) Confidence and Prediction Bands

In statistical analysis, confidence bands are used to represent the uncertainty in a given estimation of a curve/function based on limited or noisy data. A prediction band represents the uncertainty about the value of a new data point on the curve, also based on limited/noisy data. Confidence and prediction bands are frequently used to visualize the results of regression analysis.

Confidence bands have a close relation to confidence intervals, which are used to represent the uncertainty in the estimate of a single numerical value. Comparatively, confidence intervals are narrower at the single point they refer to than a confidence band, which is intended to represent many points.

The mathematical process of obtaining confidence and prediction bands goes as follows. The response (y) can be predicted at a value (x) using the linear equation, $\hat{y} = \hat{a} + bx$. Based on the variance R=1, the following predictive interval is recorded:

$$\mathrm{Var}(y_0 - \hat{y}_0) = \sigma^2 \left[\frac{1}{r} + \frac{1}{mR} + \frac{(x_0 - \bar{x})^2}{R \sum (x_i - \bar{x})^2} \right]$$

Prediction for the response value.

This computation results from expanding upon the linear equation. 1/r is derived from the variation of y about the line, 1/mR represents the uncertainty value of estimating â and b.

Next, establishing bounds on the function achieves a prediction interval of 99% for the function's ŷ value, based on x. To better understand the meaning of the bound functions below, we will describe the meaning of unseen variables and notable features:

- $Z_{0.995}$ represents the 0.995 quantile of the standard normal distribution. This represents the standard deviation of residuals.

$$(\hat{a} + \hat{b}x) + z_{0.995}\, \hat{\sigma} \left[1 + \frac{1}{m} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]^{\frac{1}{2}}$$

Formulation of right bound.

$$(\hat{a} + \hat{b}x) - z_{0.995}\, \hat{\sigma} \left[1 + \frac{1}{m} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]^{\frac{1}{2}}$$

Formulation of left bound.

To illustrate the difference between the prediction bands and confidence bands, observe the confidence interval for a + bx:

$$(\hat{a} + \hat{b}x) \pm z_{0.995}\,\hat{\sigma}\left[\frac{1}{m} + \frac{(x - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]^{\frac{1}{2}}$$

Confidence interval.

While at face value, the equations appear very similar they differ in several dimensions. The confidence interval here is smaller because by design and definition, it excludes the variability of y about a + bx. Furthermore, they are implemented for two different purposes. The confidence interval represents the accuracy of the estimated mean (y) at point x. The prediction interval represents an observation - namely the mean of a collection of observations held at point x. Bands must be adjusted based on the average number of observations (r). Executing the equations above provides an interval for either a singular future reading, or the average of several future readings.

g.) Maximum Likelihood

With a regression model with known $X_1, \ldots, X_n$ inputs, we can provide an interval under the Gauss additive measurement model, and the maximum likelihood estimate of X_0. In this case, we show that under the Gauss measurement model with normal errors, χ_0 is the maximum likelihood estimate of X_0, and its confidence interval is (X_1, X_u). Thus with 2 outputs:

For i = 1, …. , n :

$$Y_i = a + bx_i + E_i$$
$$Y_0 = a + bx_0 + E_0$$

$E_1, \ldots, E_n$ represent the normal errors with zero mean value and variance σ^2. With 4 unknown parameters, in terms of the normals errors, the log likelihood is:

$$l(a, b, x_0, \sigma) = (m + 1)(log(\sigma)) - \tfrac{1}{2\sigma^2}\,[\Sigma(y_i - a - bx_i)^2 + (y_0 - a - bx_0)^2]$$

So, the maximum likelihood estimate χ_0 satisfies $y_0 - \alpha - \beta\chi_0 = 0$. α *and* β are respectively maximum likelihood estimators of a and b. With these estimators satisfying the equations, this shows that there is an approximate normal distribution. Thus, the confidence interval (X_1, X_u) can be interpreted as a 95% confidence interval ($\alpha = 0.025$).

6.) Conclusion

Out of the models tested and analyzed throughout our study, a conclusion can be drawn on which is the best fit for the USDA Forest Service's snow gauge data. Based on the results from running a linear model, log linear model, polynomial regression, and local regression, local regression proved to be the best fit (*Figure 4.4.2).* Log linear was a close second fit, with close (but worse nonetheless) R-squared and MSE values.