Shiyin Liang
Darren Chang
Anthony Martinez
HW#4

INTRODUCTION

For their water, millions of people in Northern California depend on the snowfall in Sierra Nevada mountains. Thus, in order to monitor water supply and anticipate potential floods, the USDA operates a snow gauge out of Central Sierra Nevada to measure snow density. The snow gauge does not measure snow density directly but is rather calculated from the snow gauge's emission of gamma rays. However, as the snow gauge experiences wear and tear, the calculation method needs to be recalibrated every passing year. The objective of this project is to develop a procedure for calibration or converting the snow gauge readings into snow density.

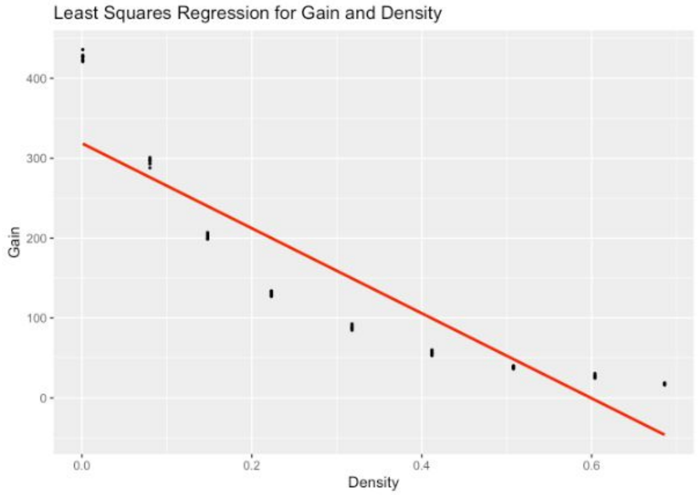
This project uses data collected from one such USDA calibration run of the snow gauge in Central Sierra Nevada. During this calibration run, the snow gauge took measurements from nine polyethylene blocks, each with different densities to simulate snow. The data contains ninety observations: ten measurements called "gain" gathered from each density.

ANALYSIS

1. **FITTING**

   Since there is only one independent variable (density) and one dependent variable (gain), we will start our search for a calibration formula in simple linear regression or, more specifically, of the form: $\widehat{y} = \beta_0 + \beta_1 x$ where β_1 is the slope, β_0 is the intercept, x is density value and $\widehat{y}$ is predicted gain. In Figure 1, we used the built-in lm() function to find the least squares regression line of $\widehat{y} = 318.7 - 531.95x$ for our data. The RMSE value is 0.0966 and the R^2 value of 0.8157 which tells us that about 81% of the variability in gain is explained by the model. While these values don't seem terrible, you can see that in Figure 1, none of our observations lie on the prediction line. In fact, the relationship between density and gain does not seem to be linear at all as the convexity in the data points suggests an exponential decay relationship.

   Figure 1:



   To definitively check if least squares regression is appropriate for this dataset, we need to check for linearity between density and gain, normality of the residuals, and homoscedasticity. For linearity, we found the residuals by subtracting the predicted gain from the observed gain for each datapoint and plotted Figure 2 and 3. Figure 2 shows that density and gain is not linear as their residuals are clearly not random and seem to follow a hyperbolic pattern. The variability of the points around the y = 0 line is also not very constant which means the dataset also fails the condition of homoscedasticity. Figure 3 shows that this dataset fails the normality condition as the data points do not follow the QQ plot and kurtosis value was 1.715 which is quite far from 3, the value that denotes normality. The residuals of this model shows that the dataset at its current form is not suitable for least squares regression.

Figure 2:

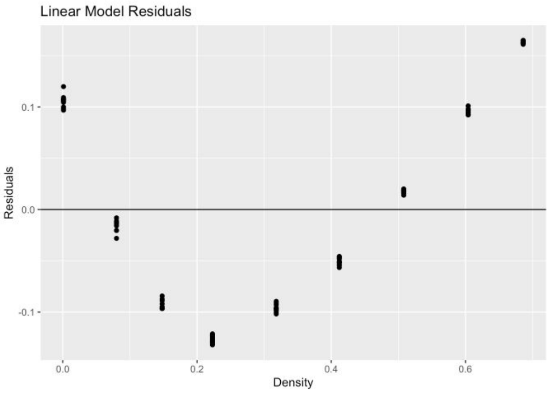
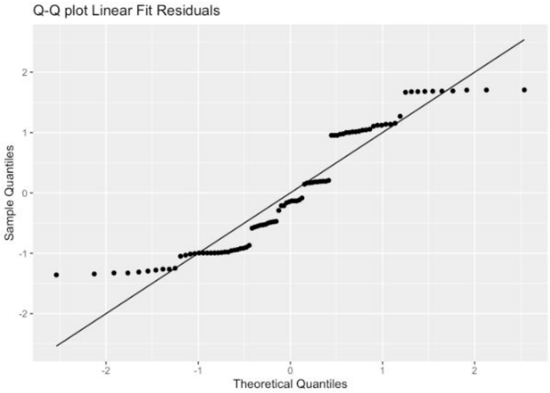


Figure 3:



Since the relationship between density and gain is clearly not linear and seems to be more exponential decay, we log transformed the gain variable to force the relationship between the two variables to be more linear and thus, more suitable for least squares Regression. As you can see, Figure 4 shows that least squares regression fits better for log transformed gain as mosts points lie on the prediction line or lies close to the Prediction line. The least squares prediction line is $\hat{y} = 1.298 - 0.216x$. This line suggests that for every unit of increase in density, the log of gain decreases by about 0.21 units and when density is 0--or there is no snow--the log of gain is about 1.3. The R^2 value for this line is 0.9958 which means this model explains 18% more of the variability than our first linear fit. The RMSE value of 0.01455053 tells us that the average error between predicted values and observed values is about 0.015 units, which is much better than the average error in our original linear fit. In Figure 5, we can see that there is more randomness in the distribution of the residuals for this model than the previous one. Figure 6 is a QQ plot and it seems to suggest that the data follows a somewhat normal distribution. We verify this with a kurtosis value of 2.88 which is fairly close to the ideal normal value of 3.

Figure 4:

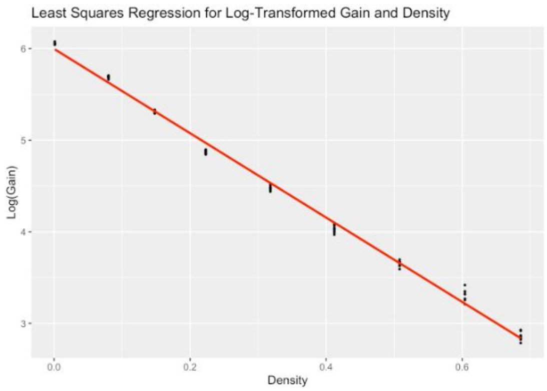


Figure 5:

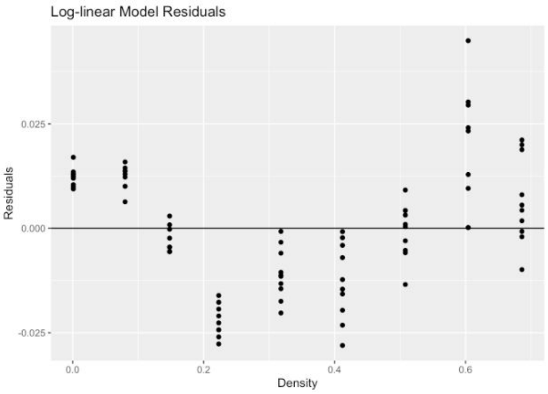
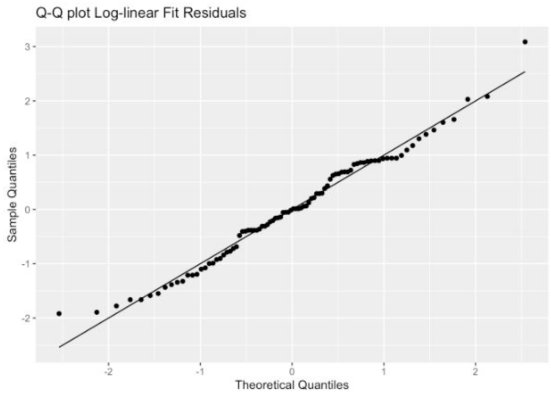


Figure 6:



CONCLUSION

In order to use least squares regression, all three conditions of linearity, constant variability and residual normality must be met. Fitting a least squares regression line onto non-transformed data was clearly not suitable as all three conditions failed. Comparing the residuals plots for each fit, log-transforming the gain variable meets the three conditions better than not log-transforming. Judging by R^2 and RMSE, log-transformed least squares regression also performs better.

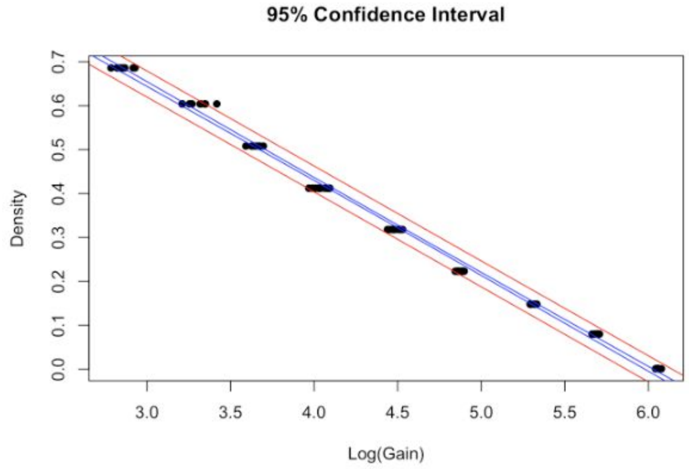
If the densities of the polyethylene blocks were not reported exactly, it may be harder to find a least squares line that can fit through our observations. Since the densities are fairly close to each other in value, there may be many outliers and deviations from the general trend created at each density and the correlation between density and gain may be weakened. Additionally, not reporting the densities exactly will affect the regression slope as measurement error in the independent variable causes the regression slope to be biased towards zero.

2. PREDICTING

According to the log transformed averages simple regression line--$\widehat{y} = 1.299 - 0.217x$ -we found that if the snow gauge measures a gain of 38.6, the density of the snow should be about 0.508. Similarly, if the snow gauge measures a gain of 426.7, the model predicts the density of the snow to be -0.01. Since it is not possible to have a negative density, we assume negative values returned by the model means snow density is 0 or there were no snow. Since we know that 38.6 is the average gain for 0.508 density and 426.7 is the average gain for 0.001 density, we can see that our model is good at predicting density. However, gain values from 0.001 density may predict 0 based on the prior assumption, which is not desirable as the scenarios of no snow and little snow are very different.

In Figure 7, we included a 95% confidence interval--the blue lines--to show the range of where the "true" density would most likely lie at each log-transformed gain and a prediction interval--the red lines--to show the range of where predictions of density would most likely lie. For this model, the 95% confidence interval says that there is 95% chance that the true density for each log-transformed gain is within ± 0.059 of the prediction line.

Figure 7:



In Figure 8, we included the range of where the true gains should lie within for each of the 9 known densities used in the calibration run to generate this dataset. The gain predictions for each density is fairly similar to the gains observed for each density in our dataset. It does however, start performing poorly at the density values of 0.0800 and 0.0010. This makes sense because we did see that our model predicted negative density for large gain values like 426.7 above. In other words, our model does not predict for densities close to 0 so well.

Figure 8:

Density	Range of Possible of Log-Transformed Gain Values	Range of Possible Gain Values
0.6860	2.811023 - 2.864163	16.62691886 - 17.53437079
0.6040	3.192916 - 3.237644	24.35935586 - 25.47363511
0.5080	3.639352 - 3.675547	38.06716121 - 39.47024118
0.4120	4.084501 - 4.114738	59.41228362 - 61.23616828
0.3180	4.518326 - 4.546829	91.68199379 - 94.33280422

0.2230	4.954357 - 4.985926	141.7914052 - 146.3390222
0.1480	5.297244 - 5.33393	199.7854419 - 207.2508717
0.0800	5.607471 - 5.65011	272.454329 - 284.3227396
0.0010	5.967399 - 6.01792	390.4886877 - 410.723402

CONCLUSION

Based on our log-transformed linear model, there is a 95% chance that actual density values lies with 0.059 units of the prediction density value. Since snowpack density typically ranges between 0.1 and 0.6g/cm^3, the difference of 0.059 g/cm^3 can actually result in a large difference in the amount of water available for the year. We have also included a table with the ranges of gain values and the density associated with them to be used as calibration for the snow gauge.

3. **Advanced Analysis: Support Vector Regression (SVR)**

After fitting a least squares regression line to our data, we were able to see that it fits our data fairly decently but improvements could be made with other regression methods such as SVR. Since SVR fits a curved line to data, the predictions made by SVR may be more accurate for our dataset.

We begin by fitting a SVR model on our sample data as shown in Figure 9. The black dots and the red dots represent actual values and predicted values respectively. We can see here that the values of our predictions are much closer to our actual data than when compared to the least squares regression line.

When creating our SVR model, we use a Radial Basis Function (RBF) kernel. The kernel function transforms our data from non-linear space to linear space. The kernel allows the SVR to find a fit and then data is mapped to the original space. Also, for this model we have 5 support vectors. The goal for the support vectors is to help find the best fit for our model.

However, in order to get the best fitting SVR model for our data, we tune the model by varying maximum allowable error and cost parameters. When tuning, we evaluate the performance of 1100 models i.e. for every combination of maximum allowable error and cost parameter. We found out that our best model has values epsilon - 0 and cost - 6. At these values, the Mean Squared Error (MSE) is the lowest. Our best model also has a Root Mean Squared Error (RMSE) of 7.611187. This is an unusual value since the RMSE is high, but the SVR model fits our data very well.

Lastly, in Figure 10 we plot our initial SVR model (in blue) and our tuned SVR model (in red). We can see that they are somewhat similar. They both fit our data very well. However, the tuned model will still give better predictions than the initial model.

Figure 9:

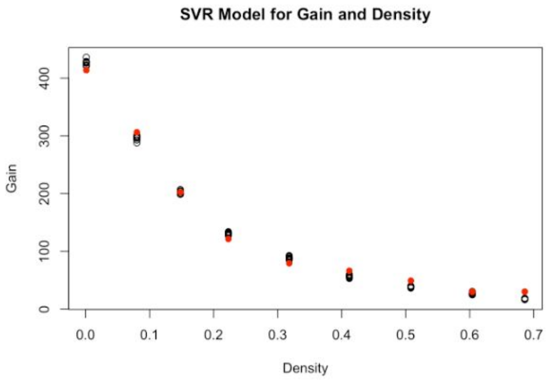
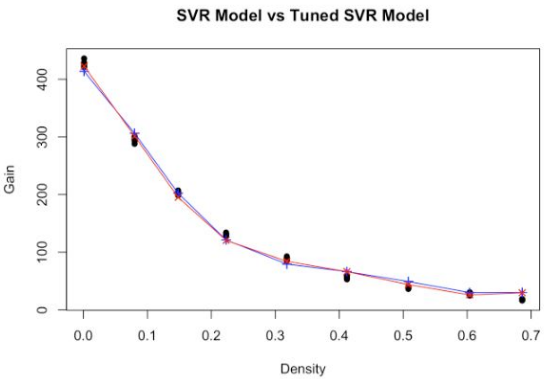


Figure 10:



CONCLUSION

After creating two SVR models, we can see that both models will give better predictions than a least squares regression line. In other words, the new models more closely resemble our given data. This leads to more reliable predictions.

CONCLUSION

Our objective for this report was to provide a procedure for calibrating snow gauge readings into snow density readings. We first used a least squares regression line to describe how our response variable (gain/snowgauge readings) changes as our explanatory variable (density) changes. However, the conditions for least squares were not met for our data so, we had to log-transform the gain variable to build a stronger linear regression model. With our log-transformed least squares regression we found a fairly good prediction model of $\widehat{y} = 1.298 - 0.216x$ where x is the density value and y is the predicted log-transformed gain value. We also found out it is likely the actual density values lie with 0.059 units of the prediction density value when the above model is inverted for $\widehat{x}$. Figure 8 contains our procedure for calibration or converting the snow gauge readings into snow density. This should be fairly accurate for the snow gauge in Central Sierra Nevada at the very least.

However, since 0.059g/ cm^3 can amount to a large difference in the annual water supply, we feel that it is best if we could try methods other than linear regression to build stronger prediction models. Judging by RMSE, we found out that a SVR model fits the data much better than the least squares regression line. In the future, if we were to improve on this project, we could try fitting a polynomial regression or expand on our SVR model instead of linear regression since the data seems to follow a hyperbolic curve. Polynomial regression may give us a stronger fit and fix the problem of predicting a negative gain from very low densities such as 0.001 that we have with linear regression. Another area for improvement is that we could repeat the calibration run but get more than ten measurements per density so that we can lower the impact noise has on the models we build and include snow gauges elsewhere in the United States so we can generalize our calibrations to more than the specific snow gauge in Central Sierra Nevada.. We can also introduce more polyethylene blocks with a larger range of densities and record gain measurements for no snow density so that we can have more density baselines to build our prediction model off of and to check the accuracy of the prediction model. Another idea is to gather data for snow gauges at different ages. Since the wear and tear is the primary reason why snow gauges need to be recalibrated every year, we should try to see if there is a relationship between measurement of gain and the age of the snow gauge. Adding another independent variable and using multiple regression may also help us build a stronger prediction model. We also noticed in the initial distribution of our data the gains in the 420 - 430 range were influential points that were causing the linear regression to be steeper. If those points were excluded we would observe a linear regression model that is more accurate and flatter. If we were to improve this analysis it may be a good idea to remove that cluster of points in order to have a far more accurate linear regression model.

APPENDIX

1. gauge.txt
2. https://www.kdnuggets.com/2017/03/building-regression-models-support-vector-regression.html
3. https://www.rdocumentation.org/packages/POT/versions/1.1-7/topics/qq
4. https://ggplot2.tidyverse.org/reference/

CONTRIBUTIONS

Darren: code for part 1 and part 2
Shiyin: introduction, part 1 analysis, part 2 analysis and conclusion
Anthony: coded Advanced Analysis, wrote Advanced Analysis