CASE STUDY 4: CALIBRATING A SNOW GAUGE

May 18, 2018

Chengyu Chen (A14051607), 2nd Year Applied Mathematics; Data Science, MATH 189
Chenyue Fang (A13686794), 2nd Year Probability and Statistics, MATH 189
Daniel Lee (A13726312), 2nd Year Probability and Statistics; Data Science, MATH 189
Xinran Wang (A13564644), 2nd Year Probability and Statistics, MATH 189
Yuqi Wang (A13532155), 2nd Year Applied Mathematics, MATH 189
Ning Xu (A92061610), 3rd Year Probability and Statistics; Economics, MATH 189

INTRODUCTION

The main source of water in northern California comes from the Sierra Nevada mountains. To help monitor the water supply, the Forest Service of the United States Department of Agriculture (USDA) operates a gamma transmission snow gauge in the Central Sierra Nevada near Soda Springs, CA. This snow gauge is used to determine a depth profile of snow density. Since the snow gauge does not disturb the snow in the measurement process, the snowpack can be measured over and over again. Using 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 the snow.

When rain falls on snow, the snow absorbs the water up to a certain threshold, after which flooding occurs. Denser snow absorbs less water. Therefore, analysis of the snowpack profile may help with monitoring the water supply and thus preventing water shortages and flooding.

The snow gauge does not directly measure snow density, but calculates a density reading based on a measurement of gamma ray emissions. Due to instrument wear and radioactive source decay, the gauge's measurements may stray over time. Thus, a calibration run is made each year at the beginning of the winter season.

In this paper, we will develop a procedure to calibrate the snow gauge in order to ensure accurate measurements and thus accurate snow density readings.

THE DATA

The data used in our analysis comes from a calibration run of the USDA Forest Service's snow gauge located in the Central Sierra Nevada mountain range near Soda Springs. The calibration run consists of placing polyethylene blocks (proxies for snow) of known densities

between the two poles of the snow gauge and taking readings on the blocks. For each polyethylene block, 30 measurements are taken. Only the middle 10 are reported, since beginning and ending measurements tend to produce inaccurate readings. The measurements reported are amplified versions of the gamma photo counts made by the detector. These measurements are referred to as the "gain."

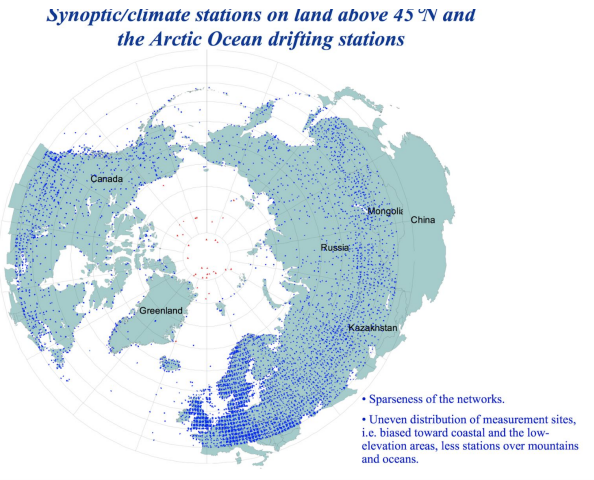
Our data consists of 10 measurements for each of 9 densities in grams per cubic centimeter of polyethylene. Thus, the data consists of 90 rows with columns "density" and "gain."

Excerpt of Dataset

	density	gain
1	0.686	17.6
2	0.686	17.3
3	0.686	16.9
4	0.686	16.2
5	0.686	17.1
6	0.686	18.5
7	0.686	18.7
8	0.686	17.4
9	0.686	18.6
10	0.686	16.8
11	0.604	24.8
12	0.604	25.9
13	0.604	26.3
14	0.604	24.8
15	0.604	24.8
16	0.604	27.6
17	0.604	28.5
18	0.604	30.5
19	0.604	28.4
20	0.604	27.7
21	0.508	39.4
22	0.508	37.6

Our data consists of a single calibration run from a single snow gauge. However, there exists challenges with procuring a large quantity and diversity of snow gauge data beyond the data used in the scope of our analysis.

First, operational networks, colloquially our "knowledge base," are not evenly spread out across the world. There has been a decline of networks in northern and mountainous regions, e.g., Siberia, Alaska, and northern Canada. This sparseness is illustrated in the figure below.



Furthermore, sustaining and improving these operational networks has been a challenge in the past.

Second, data quality and compatibility differs across national boundaries. This includes large biases in gauge measurements of precipitation, incompatibility of precipitation data due to differences in instruments and data processing methods, and difficulties in determining precipitation changes in arctic regions. However, some of these challenges have been redressed via validation of precipitation data using satellite and reanalysis products and fused products at high latitudes.

BACKGROUND

Location

The snow gauge is an instrument used to gather and measure snow precipitation. The snow gauge helps researchers study many aspects of snow-related weathers including rain-on-snow dynamics, snowpack settling and snowmelt runoff. Today, the snow gauge is widely used in many areas, including California, where a gauge is set up at the center of a forest opening with diameter 62 meters and elevation 2099 meters. The snowpack at this site is an of average 4 meters in depth every winter (Bradic).

Physical Model

The snow gauge has a radioactive source of cesium-137 and an energy detector on two poles about 70 centimeters apart. The radioactive source emits gamma photon in all directions. The energy detector has a scintillation crystal which counts photons from the radioactive source. The polyethylene molecules between resource and detector scatter and absorb photons. The

denser polyethylene molecules, the fewer photons reach detector. In our model, assuming each molecule acts independently, the probability that a gamma photon reaches detector is p^m , where p is the probability that one molecule does not absorb or bounce the photon and m is the number of molecules in one straight path between the source and the detector. Readdressing the probability gives us the formula $e^{m \cdot log(p)} = e^{b \cdot x}$ where x is the density of molecules (Bradic).

The pulses which is produced by photons to reach crystal are transmitted to a preamplifier by cable. After the pulses are amplified, they are transmitted to the lab by cable. In the lab, the signal is stabilized, corrected and converted to a measurement called gain. The snowpack density is typically between 0.1 and 0.6 g/cm^3 (Bradic).

Literature Review

According to a research about density of freshly fallen snow in the central rocky mountains by Judson et al., researchers present snow density distribution for six sites in mountains of Colorado and Wyoming in order to predict snow density. The sites in the research are at wind-protected area in forests in order to minimize side-effects from wind, and all data are collected once daily in early morning from all sites. The histogram of frequency distribution shows that densities are ranged from 10 to 257 kg/ m^3 . Steamboat Spring, the site with lowest elevation, has lowest average density of 72 kg/ m^3 . Teton Pass, the most northern site in the research, has snow density of 82 kg/ m^3 . In order to find relationship between snow density and temperature, researchers collect data of snow density and temperature at the same day at Dry Lake for four winters, then they use a scatter plot to predict snow density as a function of temperature. The result shows that there is a negative relationship between air temperature and snow density with correlation coefficient of 0.52, which is moderately strong. Furthermore, in order to test effect of new snow depth on snow density, researchers compare a group of 70 snowfalls of depth above 30 cm with a group of 381 snowfalls of depth below 15 cm The average density for two groups are 75 kg/ m^3 and 74 kg/ m^3 respectively, and no clear relation is found between snow depth and density (Judson et al.).

INVESTIGATIONS

Scenario 1: Fitting

Before we begin our analysis of our dataset, we will address potential issues with the data collection.

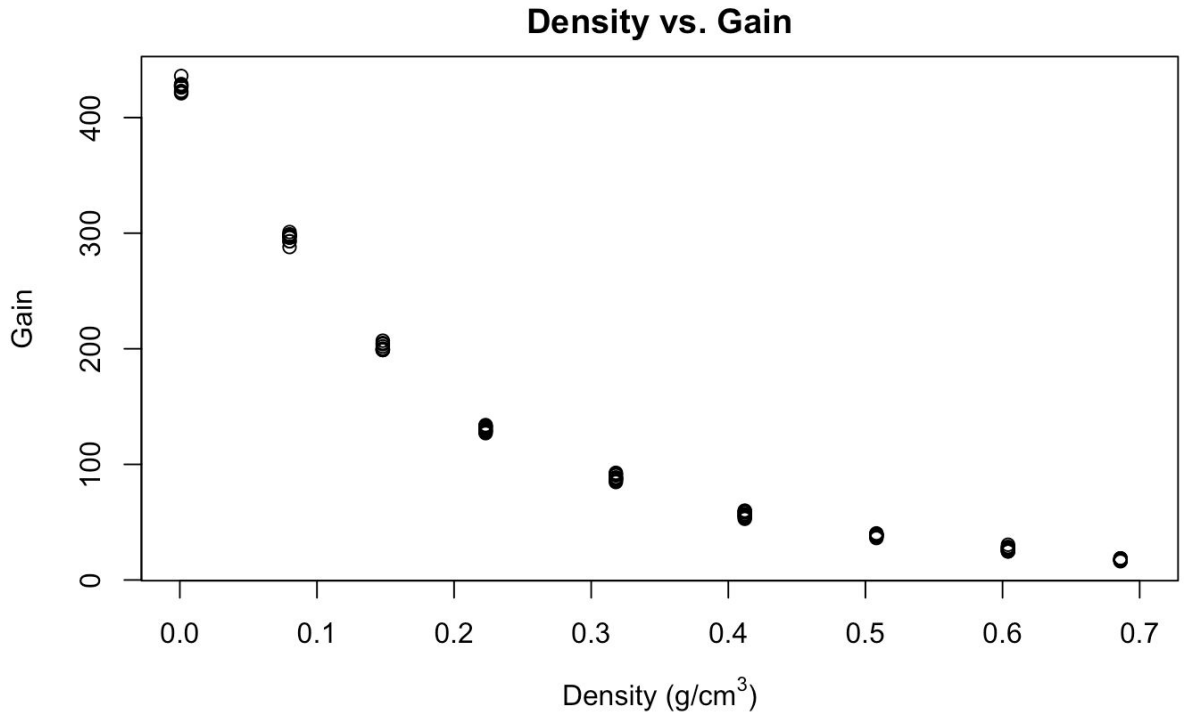
First, if the densities of the polyethylene blocks were not reported exactly, this could negatively impact the fit of our regression line predicting density from gain. A departure in

recorded densities would cause a departure in our regression line, which would result in inaccurate calibration and predictions. Since our dataset consists of measuring 10 gain readings from each unique polyethylene block (replicate measurement), a departure in one density record would amplify the departure of gain readings for that polyethylene block by a factor of 10.

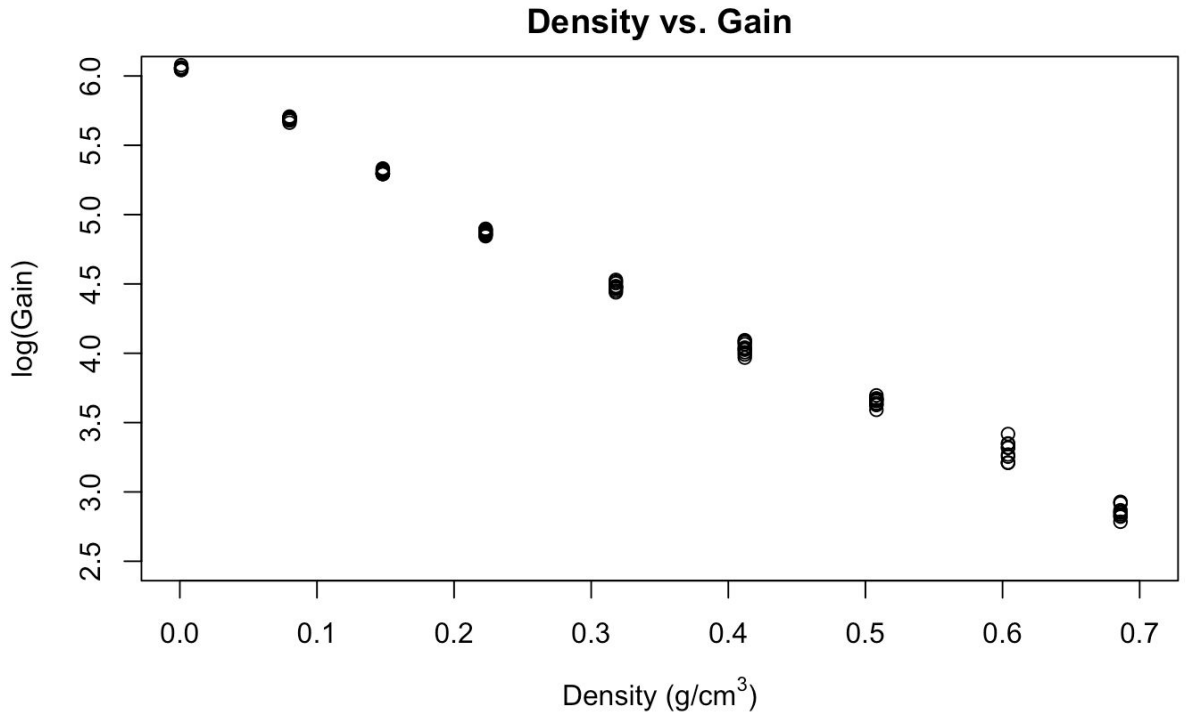
Second, if the gain readings of polyethylene blocks were not measured in random locations, this could also negatively impact the fit of our regression line. Our region of interest is the Central Sierra Nevada mountain range, and this is the region where future predictions of snow densities will be made. Therefore, our dataset should be representative of the entire mountain range as a whole. If the gain readings were measured in only one region of the mountain range, our calibration would be representative of that area only, and could give inaccurate predictions for snow densities in other regions of the mountain range. In other words, the gauge calibration would be biased and not representative of our region of interest. Therefore, it is crucial that gain readings were measured in random locations, e.g., uniformly across the Central Sierra Nevada mountain range.

Our first task is to fit the gain of the nine polyethylene blocks to the measured densities in our dataset.

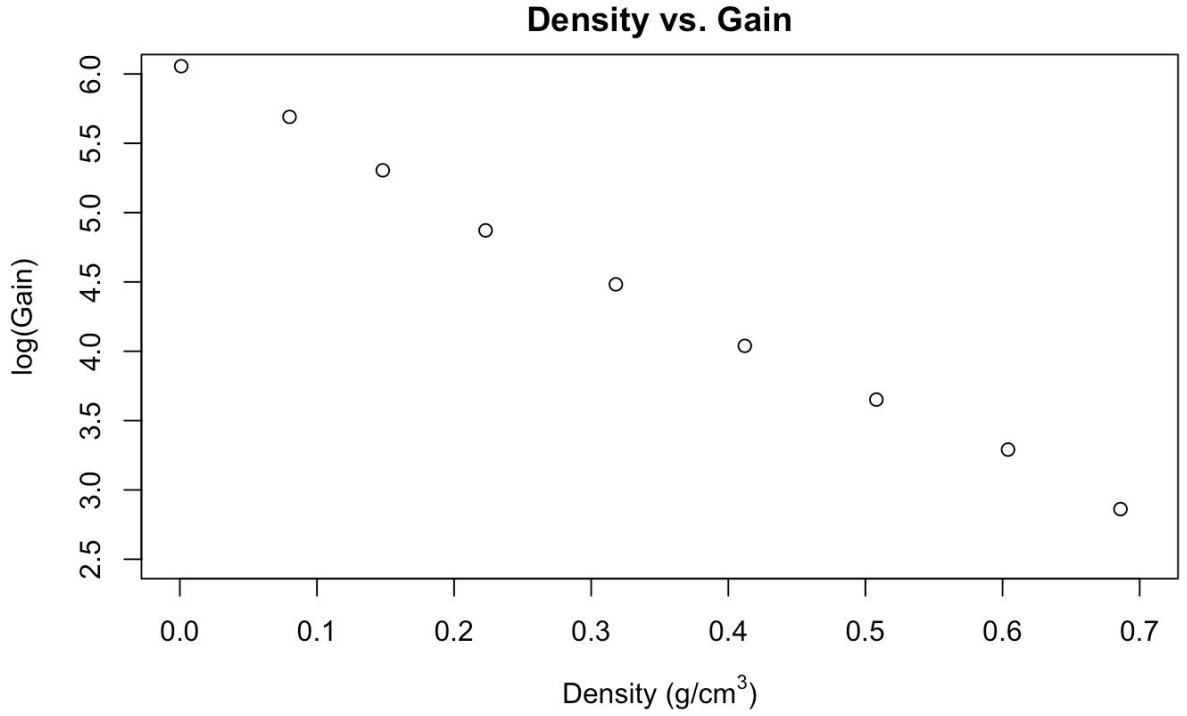
The scatter plot below depicts our raw dataset. There are nine unique densities corresponding to the nine unique polyethylene blocks. Each unique density is mapped to 10 gain readings for each corresponding polyethylene block, resulting in the “grouped” data points seen in the scatter plot. Note that although our ultimate goal is to predict snow densities from given gain readings, it is crucial that we do not invert the explanatory and response variables, as this would violate our assumptions for linear regression with regards to replicate measurements. Therefore, for the reminder of our analysis, we will visualize our data with density as the explanatory variable on the x-axis to predict gain as the response variable on the y-axis.



Upon first glance it is apparent that the data does not follow a linear association, but rather an exponential trend. An assumption of linear regression line is that the data must have a linear association. Thus, before any further analysis, we will apply a logarithmic transformation to our response variable (gain) in order to achieve a linear association, as depicted below.

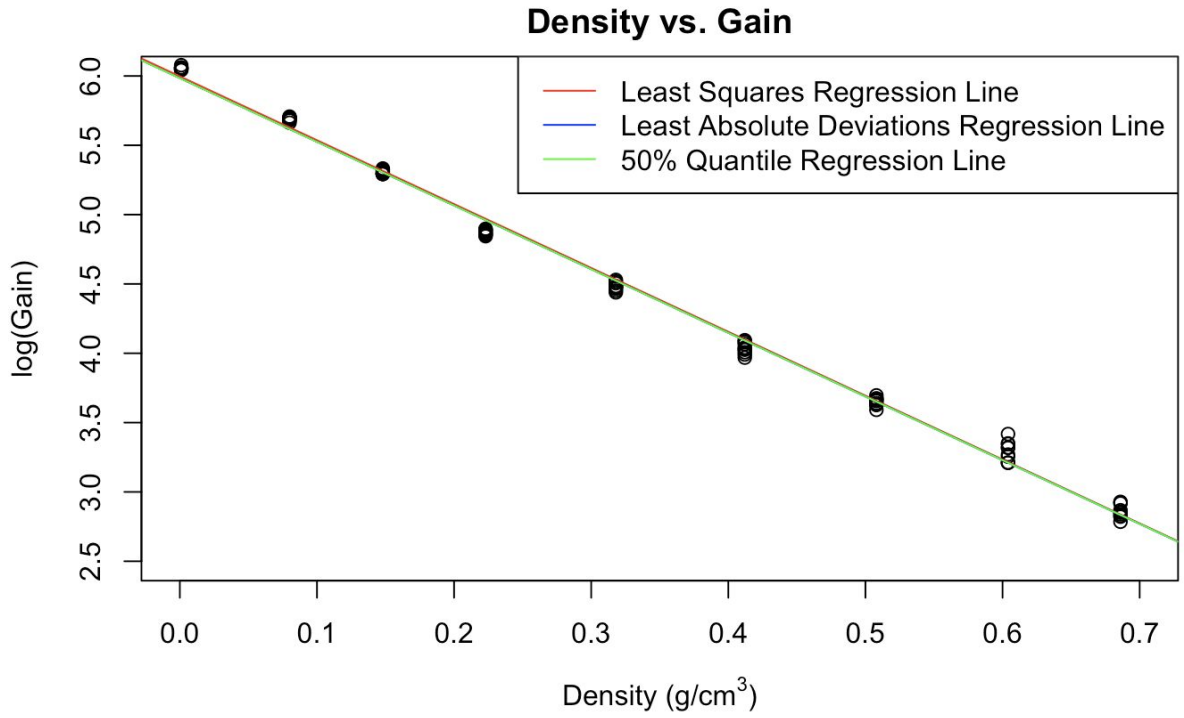


Since our dataset consists of replicate gain measurements for each polyethylene block, we must average the replicate measurements in order to have one aggregated gain reading for each polyethylene block. This is a necessary step since an assumption for linear regression is that there must be a one-to-one correspondence between explanatory and response variables. Note that weighting data points is not a necessary step for our fit, since there are an equal amount of weight, .i.e., number of replicate measurements, for each polyethylene block. Also note that although our linear model is based only on these nine aggregate data points, we will still use the full dataset in further analysis. The aggregated data points are depicted below



Now we are ready to fit our linear regression line to predict gain (logarithmically transformed) from density. Upon first glance, it is apparent that the transformed data follows a strong, positive, linear relationship. This is supported by a very strong correlation coefficient (r) of -0.998, and coefficient of determination (r^2) of 0.997. Thus, 99.7% of the variability in our response variable (gain) is explained by a linear model. Therefore, a linear model appears appropriate.

We will use three variations of linear regression in order to make our analysis more robust and replicative: least squares regression, least absolute deviations regression, and 50% quantile (median) regression. The three regression lines are depicted below.



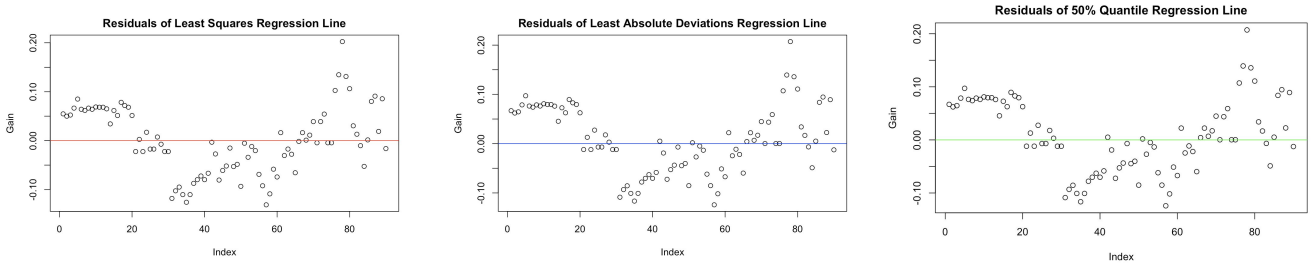
Upon first glance, it is apparent that all three regression lines produce extremely similar predictions. They share extremely similar slopes and intercepts:

- Least Squares Regression Line: $\widehat{log(gain)} = 5.997 - 4.606 \times density$
- Least Absolute Deviations Regression Line: $\widehat{log(gain)} = 5.985 - 4.594 \times density$
- 50% Quantile Regression Line: $\widehat{log(gain)} = 5.985 - 4.593 \times density$

This further supports the fact that a linear model is appropriate, and that all three regression lines appear robust as estimators.

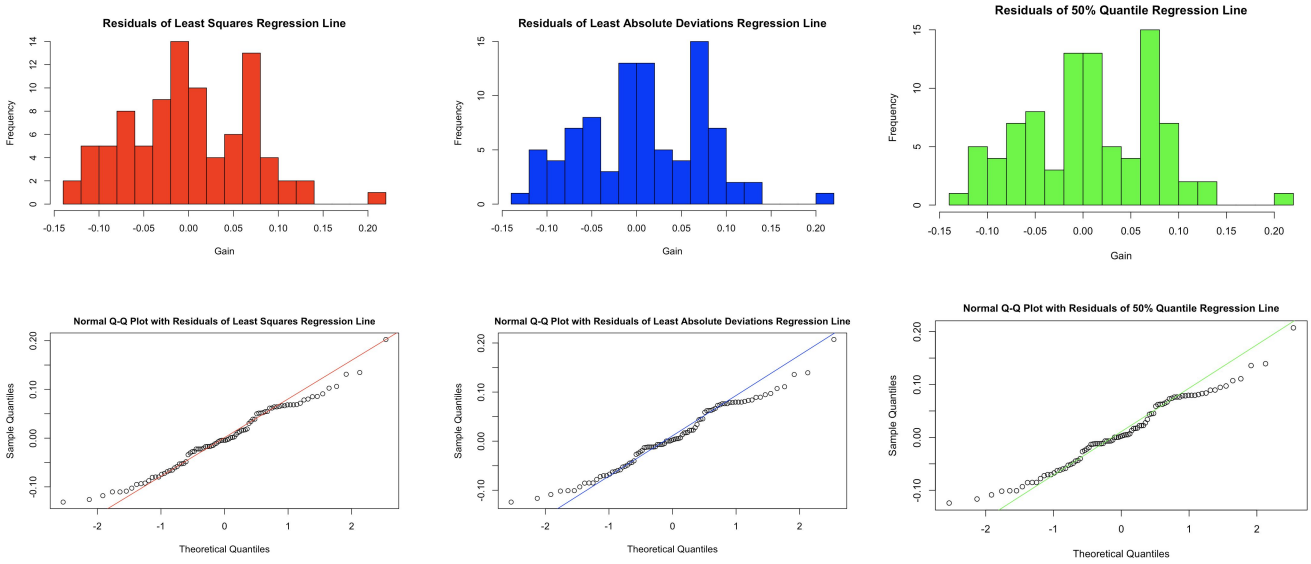
Now, we will check the three conditions for a valid linear model: linearity, normality of residuals, and constant variability.

It has been shown earlier that the relationship between the transformed explanatory variable (density) and response variable (gain) is linear. In particular, the transformed data follows a strong, positive, linear relationship. To further check linearity, the residual plots for all three regression lines are depicted below.



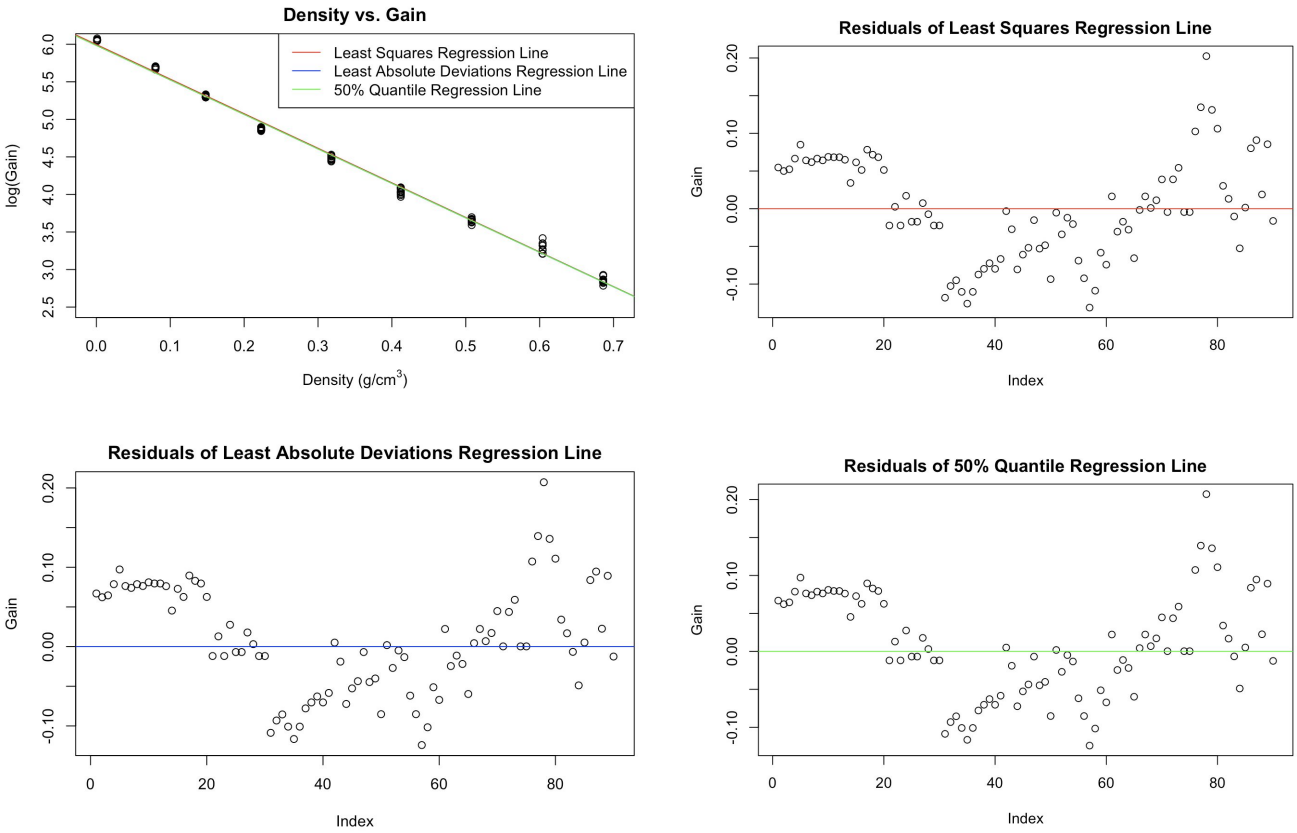
There appears to be a slight parabolic pattern in the residual plots, which may lead to minor problems with our fit. However, since the pattern is not very strong, we will continue with our analysis using a linear model.

To check normality of residuals, histograms and normal Q-Q plots of the residuals for all three regression lines are depicted below.



The histogram of the residuals appears approximately normal. Furthermore, although the tails of the Q-Q plots depart from the normal distribution, the majority of the residuals follow an approximate normal distribution.

To check constant variability (homoscedasticity) of our linear model, we can refer to the previous figures, depicted again below.



The variability of points around the least squares appears roughly constant. This implies the variability of residuals around the 0 line appears roughly constant as well. Therefore, our linear model follows homoscedasticity.

Therefore, because all three regression lines (least squares, least absolute deviations, 50% quantile) meet the linear model conditions of linearity, normality of residuals, and constant variability without overt problems, these linear models are viable and will be used throughout our analysis.

Scenario 2: Predicting

Now that we have our linear models, we will use these models to predict density given a gain reading. Note that this is a prediction inverse to our linear models (which predict gain given density), and we must exercise caution when making these inverse predictions in order to not violate the assumptions for linear regression with regards to replicate measurements.

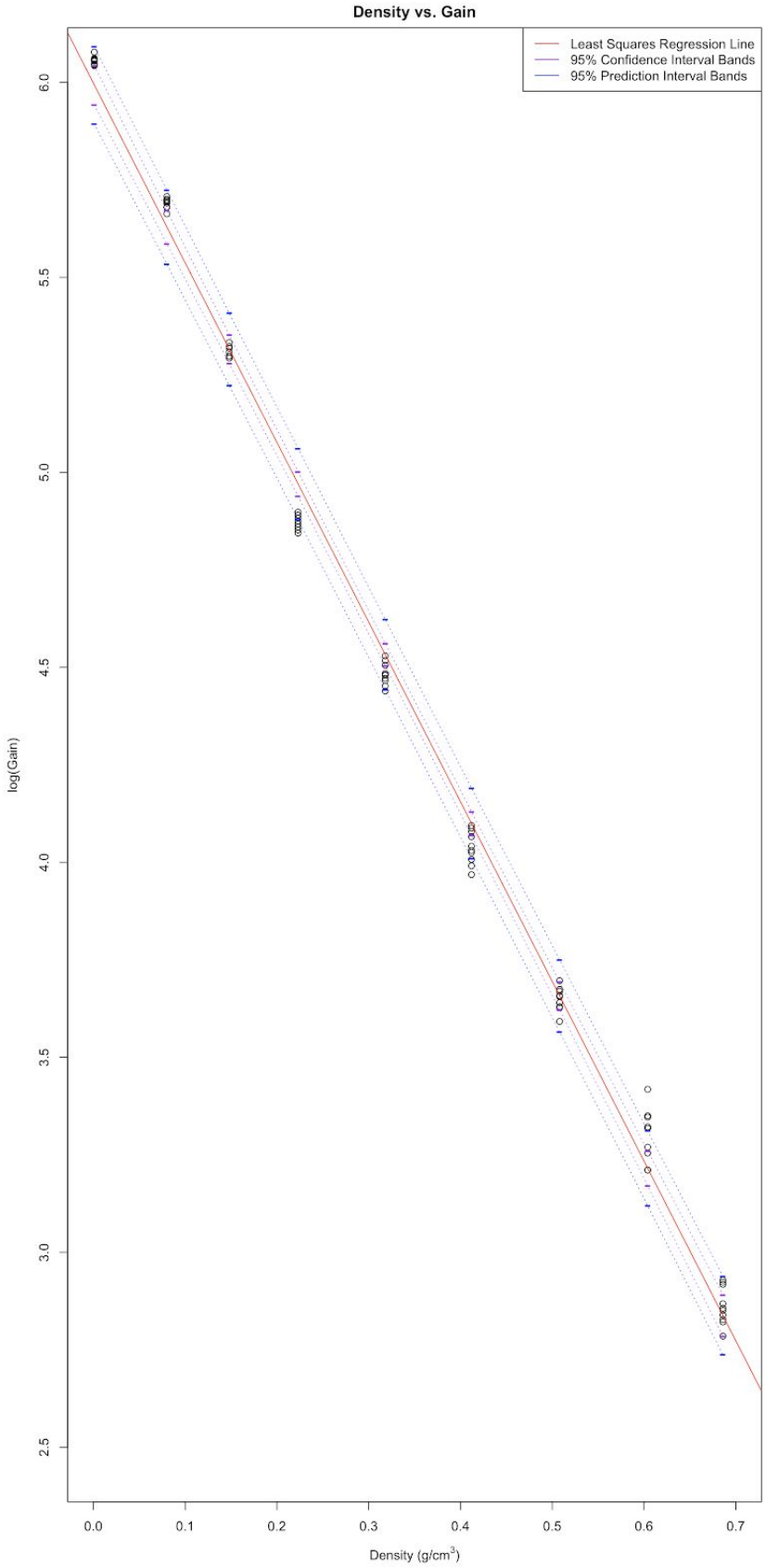
Examples of predictions that will be made using our linear model is: Given a gain reading of 38.6 or 426.7, what is the snow's density? Note that these gain readings were chosen because

they are the average gains for the 0.508 and 0.001 density polyethylene blocks respectively. Therefore, we should expect our predicted density to be in the neighborhood of 0.508 and 0.001 respectively.

To make these calculations, we define a suite of functions to use for our analysis:

Function Name	Description
`PredictLogGain()`	Predicts log(gain) given density, using least squares regression.
`PredictDensityLeastSquares()`	Inversely predicts density given gain, using least squares regression.
`PredictDensityLad()`	Inversely predicts density given gain, using least absolute deviations regression.
`PredictDensityQuant()`	Inversely predicts density given gain, using 50% quantile regression.
`LogGainCiLower()`, `LogGainCiUpper()`	Constructs 95% confidence interval for log(gain) given density, using least squares regression with pooled standard deviation.
`LogGainPiLower()`, `LogGainPiUpper()`	Constructs 95% prediction interval for log(gain) given density, using least squares regression with pooled standard deviation.
`DensityCi()`	Constructs 95% confidence interval for density given gain, using least squares regression with pooled standard deviation.
`DensityPi()`	Constructs 95% prediction interval for density given gain, using least squares regression with pooled standard deviation.

A plot of our original prediction of log(gain) given density is depicted below, with our least squares regression line and corresponding 95% confidence and prediction intervals.


Density vs. Gain
Least Squares Regression Line
95% Confidence Interval Bands
95% Prediction Interval Bands
log(Gain)
Density (g/cm^3)
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

Note that although our ultimate goal is to construct point and interval estimates for density given gain, we will not invert the plot in order to not violate assumptions for linear regression with regards to replicate measurement. Instead, we will calculate these points and interval estimates using our inverse functions on a case-by-case basis.

Looking at the plot, the prediction intervals are wider than the confidence intervals. The 95% confidence intervals provide an interval for the true population mean log(gain) given density, while the 95% prediction intervals provide an interval for future predicted log(gain) given density. Thus, it is expected that the prediction intervals are wider than the confidence intervals, as prediction intervals must account for uncertainty in the population mean as well as random data scatter. Both the confidence and prediction interval bands are curved concavely as expected, since these bands "revolve" around the center (mean) of the data points, meaning the further an interval is from the center, the wider the interval must be.

Given these confidence and prediction intervals, we can now calculate point and interval predictions of density given gain by using the inverse prediction functions defined earlier.

For a gain reading of 38.6:

- Actual corresponding density in dataset: 0.508 g/cm^3.
- Least squares regression line point estimate: 0.509 g/cm^3.
- Least absolute deviations regression line point estimate: 0.508 g/cm^3.
- 50% quantile regression line point estimate: 0.508 g/cm^3.
- 95% confidence interval: (0.501, 0.517) g/cm^3.
    - 95% confidence this interval contains the true mean density for the given gain.
- 95% prediction interval: (0.489, 0.529) g/cm^3.
    - 95% confidence this interval contains the density of the next data point with the given gain.

For a gain reading of 426.7:

- Actual corresponding density in dataset: 0.001 g/cm^3.
- Least squares regression line point estimate: -0.013 g/cm^3.
- Least absolute deviations regression line point estimate: -0.015 g/cm^3.
- 50% quantile regression line point estimate: -0.015 g/cm^3.
- 95% confidence interval: (-0.024, -0.002) g/cm^3.
    - 95% confidence this interval contains the true mean density for the given gain.
- 95% prediction interval: (-0.035, 0.009) g/cm^3.
    - 95% confidence this interval contains the density of the next data point with the given gain.

It is apparent from these predictions that our least squares, least absolute deviations, and 50% quantile regression lines produce extremely similar point estimates, as expected. Note that point and interval estimates very close to the axes, i.e., predictions for gain readings of 426.7, may contain negative values. These predictions are certainly false, as density cannot be negative, and is due to small errors in our regression lines.

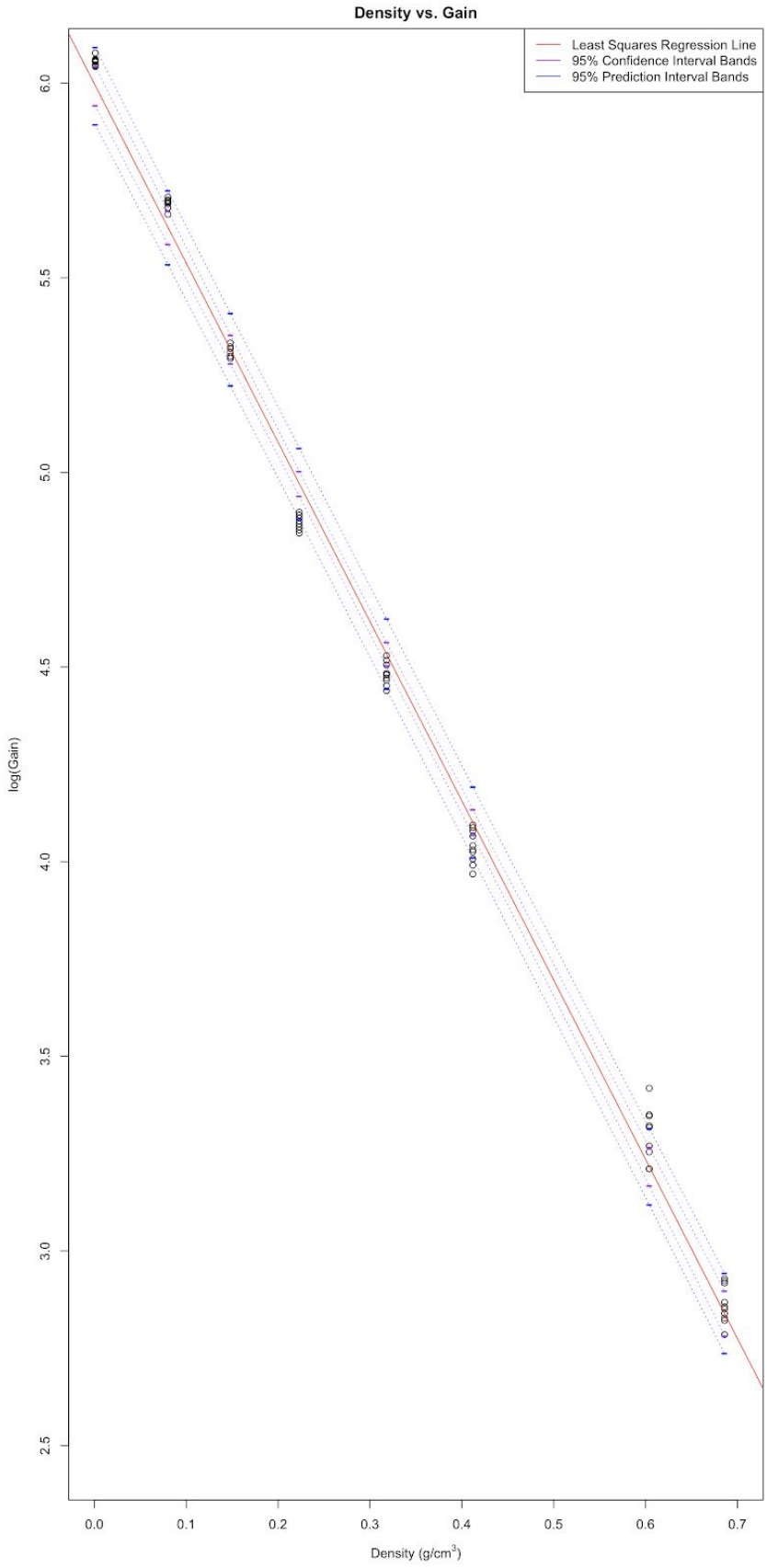
Further point and interval estimates of density given gain can be produced by our regression lines and inverse prediction functions on a case-by-case basis.

Scenario 3: Cross-Validation

Earlier, we found that a linear model fits our data well, including least squares, least absolute deviations, and 50% quantile regression models. With these models, we made both direct and inverse predictions with associated confidence and prediction intervals. However, we did not directly test the accuracy of our predictions and have no way of knowing the quality of our predictions asides from our confidence and prediction intervals.

Therefore, we will now separate our original dataset into training data and testing data to perform cross-validation in order to test the accuracy of our linear model.

We first omit the set of measurements corresponding to the polyethylene block with density 0.508 g/cm^3. Using this modified dataset as our training data, we produce a new linear model (specifically least squares regression) using the same estimation/calibration procedures as for the original dataset. A plot of the linear model from our training data is depicted below, with our least squares regression line and corresponding 95% confidence and prediction intervals.


Density vs. Gain
Least Squares Regression Line
95% Confidence Interval Bands
95% Prediction Interval Bands
log(Gain)
Density (g/cm3)
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

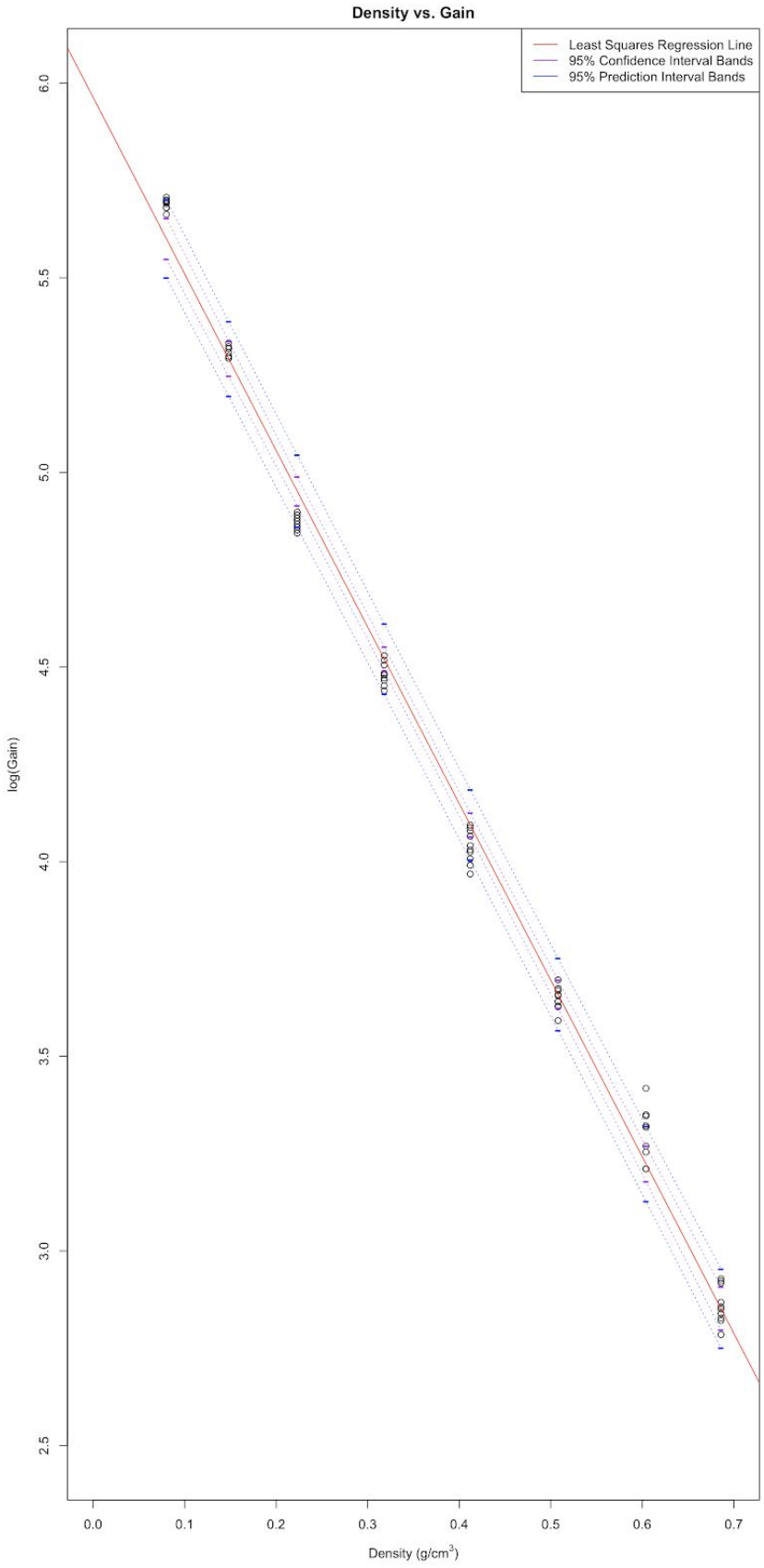
Upon first glance, there is almost no difference between this linear model and our original linear model. We then calculate the point and interval estimates for the density of a block with an average gain reading of 38.6, which is the average gain reading of the polyethylene block we omitted. In this case, the polyethylene block we omitted contains our training data that we will use to cross-validate our predictions.

For a gain reading of 38.6:

- Actual corresponding density in dataset: 0.508 g/cm^3.
- Least squares regression line point estimate: 0.509 g/cm^3.
- 95% confidence interval: (0.501, 0.518) g/cm^3.
    - 95% confidence this interval contains the true mean density for the given gain.
- 95% prediction interval: (0.489, 0.530) g/cm^3.
    - 95% confidence this interval contains the density of the next data point with the given gain.

The width of these confidence and prediction intervals are slightly larger compared to the confidence and prediction intervals of our original data, which is expected due to less data available to make a prediction. The actual density of the polyethylene block we omitted is 0.508 g/cm^3, while the linear model using the training data was able to produce an estimate of 0.509 g/cm^3. Furthermore, the actual density is well captured by our prediction interval of (0.489, 0.530) g/cm^3. Thus, the linear model produced from the training data is a very good predictor for our testing data.

We will now repeat this procedure with the polyethylene block with density 0.001 g/cm^3 containing our testing data. A plot of the linear model based on our training data is depicted below, with our least squares regression line and corresponding 95% confidence and prediction intervals.


Density vs. Gain
Least Squares Regression Line
95% Confidence Interval Bands
95% Prediction Interval Bands
log(Gain)
Density (g/cm^3)
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

Again, upon first glance, there is almost no difference between this linear model and our original linear model. We then calculate the point and interval estimates for the density of a block with an average gain reading of 426.7, our training data.

For a gain reading of 426.7:

- Actual corresponding density in dataset: 0.001 g/cm^3.
- Least squares regression line point estimate: -0.021 g/cm^3.
- 95% confidence interval: (-0.035 -0.007) g/cm^3.
    - 95% confidence this interval contains the true mean density for the given gain.
- 95% prediction interval: (-0.045, 0.003) g/cm^3.
    - 95% confidence this interval contains the density of the next data point with the given gain.

Again, the width of these confidence and prediction intervals are slightly larger compared to the confidence and prediction intervals of our original data, which is expected due to less data available to make a prediction. The actual density of the polyethylene block we omitted is 0.001 g/cm^3, while the linear model using the training data was able to produce an estimate of -0.021 g/cm^3. Compared to the previous cross-validation procedure with the polyethylene block with density 0.508 g/cm^3 being the training data, the predictions of this iteration of cross-validation is less accurate. However, the actual density is still captured by our prediction interval of (-0.045., 0.003) g/cm^3. Thus, the linear model produced from the training data is a good predictor for our testing data.

These two iterations of cross-validation were able to produce a linear model (using the training data) which accurately captured the prediction using the testing data. Therefore, our linear model is robust and has a high degree of accuracy and reliability.

Additional Scenario: Temperature, Day Of Year, and Latitude

For the additional hypothesis, we plan on investigating how temperature correlates with the day of year and latitude. We used one of the dataset from Full Resolution Data “64506420.csv” to make the investigation. Judson’s research shows that there is a negative relationship between air temperature and snow density with correlation coefficient of 0.52. By constructing a regression model, we wanted to see how temperature, an influencing factor to snow density, is changing due to the time of year and the latitude of the location.

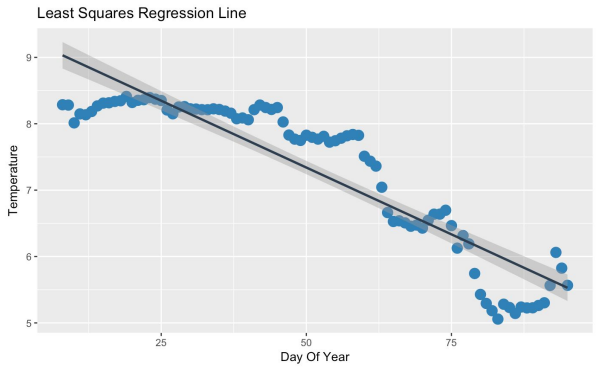
The null hypothesis is that there is no correlation between each pair of the parameters. The alternative is that the correlation exists. In our model, P-value is a measure of the probability of having a non-zero slope by chance. The decision rule is to reject the null hypothesis when

p-value is smaller than 0.05.

Prior to fitting the data, we selected our information of the air temperature and day of year of each data entry and sorted the dataset based on Day of Year. Since the Day of Year has a range of number between 8 to 95, we know that the data was recorded from January to early April. The scatter plot below depicts our raw dataset with a fitted line fitting the air temperature based on the day of year. The summary statistics below shows how good the model fitted our data. The estimated slope coefficient for the fit is -0.04, with a p-value significantly small, suggesting a negatively correlated linear relationship between DOY and temperature. Having an R-squared value of 0.83 also suggests that approximately 83% of variability in the response variable "temperature" is explained by the model.

The regression model is as following:

$$\widehat{Temperature} = 9.346 - 0.04 \times DOY$$

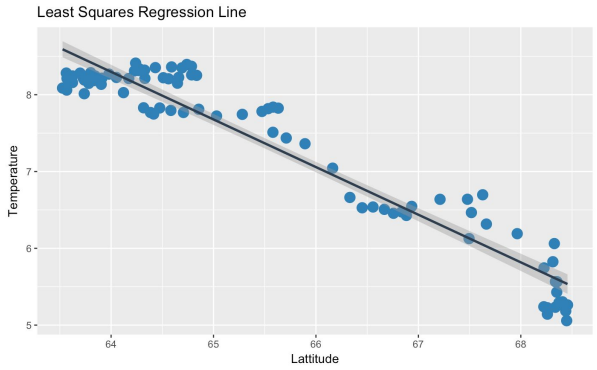



```
Call:
lm(formula = data.Ts ~ DOY, data = data.avg)

Residuals:
     Min       1Q   Median       3Q      Max
-0.95439 -0.34633  0.03139  0.35247  0.84693

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.353218   0.114242   81.87   <2e-16 ***
DOY         -0.040253   0.001989  -20.23   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4741 on 86 degrees of freedom
Multiple R-squared:  0.8264,    Adjusted R-squared:  0.8244
F-statistic: 409.4 on 1 and 86 DF,  p-value: < 2.2e-16
```

The scatter plot below depicts our raw dataset with a fitted line fitting the air temperature based on the latitude of the location, indicating the relationship between temperature and latitude. The estimated slope coefficient for the fit is -0.6, with a p-value significantly small, suggesting a negatively correlated linear relationship between latitude and temperature. Having an R-squared value of 0.92 suggests that approximately 92% of variability in the response variable "temperature" is explained by the model. Having a larger R-squared value suggests that more variability is explained by this model.

The regression model is as following:

$$\widehat{Temperature} = 47.99 - 0.62 \times latitude$$




```
Call:
lm(formula = data.Ts ~ data.Lat, data = data.avg)

Residuals:
     Min       1Q   Median       3Q      Max
-0.51377 -0.26924 -0.04201  0.24403  0.64902

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  47.9901     1.2672   37.87   <2e-16 ***
data.Lat     -0.6202     0.0193  -32.14   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3155 on 86 degrees of freedom
Multiple R-squared:  0.9231,    Adjusted R-squared:  0.9222
F-statistic:  1033 on 1 and 86 DF,  p-value: < 2.2e-16
```

Lastly, we want to see if both parameters were independently influencing the air temperature. We used multiple regressions for this fitting model. The estimated slope coefficient for DOY is -0.0097, and the estimated slope coefficient for latitude is -0.49, both with p-values significantly small, suggesting a negatively correlated relationship among latitude, DOY, and temperature. Since we have more than one explanatory variable, we look at the adjusted R-squared value. Having an adjusted R-squared value of 0.93 suggests that approximately 93% of variability in the response variable "temperature" is explained by the model. Although in this model, we having a slightly larger R-squared value by 0.01, due to the purpose of model's simplicity, the earlier model fitting temperature with latitude is preferred.

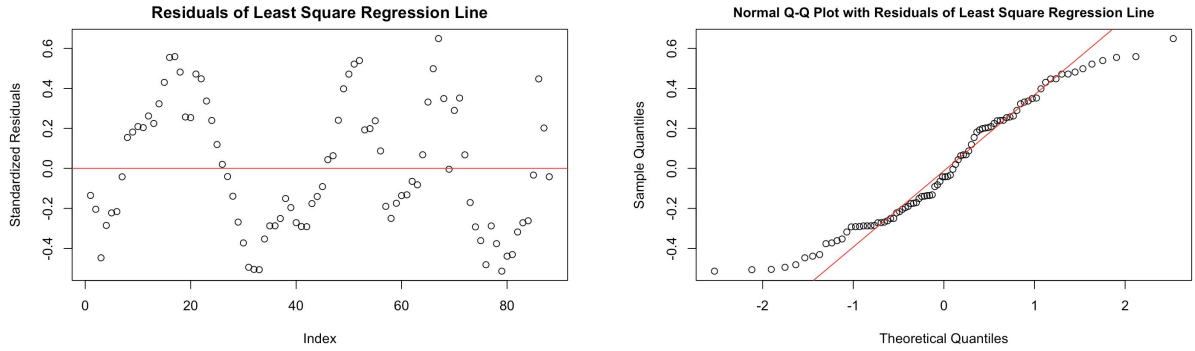
The regression model is as following:

$$\hat{Temperature} = 40.0496 - 0.0097 \times DOY \; - \; 0.4916 \times latitude$$

```
Call:
lm(formula = data.Ts ~ DOY + data.Lat, data = data.avg)

Residuals:
     Min       1Q   Median       3Q      Max
-0.60721 -0.22496 -0.00537  0.25822  0.61356

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 40.049638   2.673979  14.978  < 2e-16 ***
DOY         -0.009756   0.002936  -3.322  0.00132 **
data.Lat    -0.491566   0.042805 -11.484  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2985 on 85 degrees of freedom
Multiple R-squared:  0.932,     Adjusted R-squared:  0.9304
F-statistic: 582.1 on 2 and 85 DF,  p-value: < 2.2e-16
```

Inference:

Overall, we decide to use the second model, a linear model predicting temperature from latitude. Having a significantly low p-values for the slope coefficient indicates that the probability of having such correlation by chance is small, which means the null hypothesis should be rejected. Therefore, although there is correlation between Temperature and Date Of Year, Temperature and Latitude, we find the correlation between latitude and temperature the most appropriate.



The left graph above is the graph of residuals for least square regression line, which reveals the distribution of residuals is stochastic, indicating that the samples are meaningful for investigation. The right graph above is a Normal Quantile-Quantile plot with residuals of least squares regression line, since the residuals forms roughly a straight line which indicates that the residual plot of least squares line is approximately normal. Since the line meets the linear model conditions of linearity, normality of residuals, and constant variability without overt problems, this regression models is appropriate. By having a model and will be used throughout our analysis.

DISCUSSION

Scenario 1, 2, and 3

Our ultimate goal is to provide a simple procedure for converting gain into density when the gauge is in operation. In particular, we aim to calibrate a snow gauge in the Sierra Nevada mountains via a linear model in order to predict snow density from a gain reading. We created three linear models: least squares regression, least absolute deviations regression, and 50% quantile (median) regression models. We were able to show that a linear model is appropriate for our logarithmically transformed data, and all three linear models were shown to produce

extremely similar predictions, indicating robustness and replicability of our procedure. To quantify the uncertainties of our predictions, we constructed confidence and interval bands around our least squares regression line. Using this information, we were able to produce both point and interval estimates for density given any gain reading. Finally, to test the accuracy of our linear models and the procedure used to produce them, we performed two iterations of cross-validation. In both iterations, the linear model from the training data was able to accurately make predictions on the testing data.

After all of this testing and validation, we can not only conclude that our linear model is not only appropriate for the data, but can also conclude that our linear model produces consistently accurate predictions. Therefore, the current calibration is accurate enough to predict snow densities given gain readings from the snow gauge, and hence can be used to predict the water supply in northern California.

Additional Scenario

Based on the previous investigations, we found a significant, negative correlation between temperature and latitude. Therefore, we constructed a linear model predicting temperature from latitude. From our literature review, Judson et al. found that snow density is strongly correlated with temperature. We aim to develop a model relating snow density to latitude. However, since the datasets are different, the regression formula from that scholarly research predicting snow density from temperature cannot be directly compared to our investigation. From our model, we discovered that temperature is negatively correlated to the latitude of locations. Combining with what the Judson found out in their research, our hypothesis is that as latitude increases, the temperature drops, and the snow density becomes higher. In order to test this, we would need a more rigorous dataset containing the relative air humidity, number of times with wind velocity above 4 ms^{-1}, the amount of sunlight absorbed by the snow cover, etc. (Judson). With more rigorous datasets, we would be able to relate these variables with latitude and test if there is a multiple linear model that we can use to predict snow density from latitude of the site, along with more comprehensive variables.

THEORY

Regression modeling

Regression analysis is a set of statistical processes for estimating the relationships among variables. Regression models involve the following parameters and variables:

1) The unknown parameters, denoted as β, which may represent a scalar or a vector.
2) The independent variables, X.

3) The dependent variables, Y.

A regression model relates Y to a function of X and β. $Y \approx f(X, \beta)$. The approximation is formalized as $E(Y|X) = f(X, \beta)$. The form of the function f must be specified. Sometimes the form of this function is based on knowledge about the relationship between Y and X that does not rely on the data.

In linear regression, the model specification is that the dependent variable, yi, is a linear combination of the parameters. For example, in simple linear regression for modeling n data points there is one independent variable: xi, and two parameters $\beta 0, \beta 1$: yi = $\beta 0 + \beta 1$ xi + εi. The residual, ei = yi - $\widehat{yi}$, is the difference between the value of the dependent variable predicted by the model, $\widehat{yi}$, and the true value of the dependent variable, yi.

Coefficient of Determination (R^2)

The strength of the fit of a linear model is most commonly evaluated using R^2, which is calculated as the square of the correlation coefficient. It tells us what percent of variability in the response variable is explained by the model. The remainder of the variability is explained by variables not included in the model or by inherent randomness in the data.

Residuals

Errors are denoted by ε_i, the estimated residual $\widehat{\varepsilon_i}$ from the least squares line is equal to $Y_i - \widehat{a} - \widehat{b}xi$ and is unbiased which means $E[\ \widehat{\varepsilon_i}\ ] = \varepsilon_i$. Residuals are estimates of model error, so we can use $\widehat{\varepsilon_i}$ to check whether $\widehat{\varepsilon_i} \sim N(0, \sigma^2)$. However the variance of $\widehat{\varepsilon_i}$ doesn't equal to σ^2. it depends on i, even if the model is perfect. However the variance of the standardized residual is equal to σ^2, which means we can look at standardized residuals to do model checking.
$\widehat{\sigma^2} = \frac{1}{n-\gamma} \sum_{i=1}^{n} \widehat{\varepsilon_i}^2$ where γ is the number of regression parameters. If the number of regression parameters equal to n then $\widehat{\sigma^2} \rightarrow \infty$ which is going to be a poor prediction and it will also leads to "overfitting".

Least Squares regression

Least squares linear regression is a method for predicting the value of a dependent variable Y, based on the value of an independent variable X. Linear regression finds the straight line, called the least squares regression line, that best represents observations in a bivariate data set. Suppose Y is a dependent variable, and X is an independent variable. The population

regression line is $Y = \beta 0 + \beta 1 X$, where $\beta 0$ is a constant, $\beta 1$ is the regression coefficient, X is the value of the independent variable, Y is the value of the dependent variable.

The regression line has the following properties:

1) The line minimizes the sum of squared difference between observed values (y) and predicted values ($\widehat{y}$).
2) The regression line passes through the mean of the X values and through the mean of the Y values.
3) The regression constant ($\beta 0$) is equal to the y intercept of the regression line.
4) The regression coefficient ($\beta 1$) is the average change in the dependent variable (Y) for a 1-unit change in the independent variable (X). It is the slope if the regression line.

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

Least Absolute Deviations Regression

The least absolute deviation model tries to minimizes the absolute value of the residuals, for example MAE= $\min \sum_{i=1}^{n} \left| y_i - \widehat{y_i} \right|$ this provides a robust solution when outliers are present. However, sometimes there is no unique solution and in fact an infinite number of different regression lines are possible.

Quantile Regression

Quantile regression aims at estimating either the conditional median or other quantiles of the response variable. It is the extension of linear regression and we use it when the conditions of linear regression are not applicable. Any real-valued random variable X can be characterized by its distribution function $F(x)=\mathbb{P}(X \leq x)$. Then for any $\tau \in (0, 1)$, $F^{-1}(\tau)=\inf\{x:F(x) \geqslant \tau\}$ the τ th quantile of X. In another word, $\mathbb{P}(X \leq F^{-1}(\tau))=\tau$, that is, the chance the random variable is less than $F^{-1}(\tau)$ is τ.

We can find all quantile information by minimizing the empirical quantity. If we think about the first order equation for the median : $\mathbb{E}[½ \operatorname{sign}(\lambda - X)]=0$. It means when $X>\lambda$, the contribution of X is $-1/2$, and when $X<\lambda$, the contribution of X is 1/2. Because of this equal contribution, we want to find a λ with X distributed symmetrically around it. In another word, $\frac{1}{2}\mathbb{P}(X \leq \lambda) - \frac{1}{2}\mathbb{P}(X>\lambda)=0$.So if we want to find a general τ-th quantile instead, we basically want to find a λ such that $\mathbb{P}(X \leq \lambda)=\tau$. Then $\mathbb{P}(X>\lambda)=1-\tau$, and $(1-\tau)\mathbb{P}(X \leq \lambda) - \tau\,\mathbb{P}(X>\lambda)=0$. Integrating this equation, we get the minimization problem: $\min_\lambda \mathbb{E}\,\rho_\tau\ (X-\lambda)$ where

$\rho\tau(a)=\{a\tau$ if $a>0$; $a(\tau-1)$ if $a\leq 0$. We call $\rho\tau(a)$ the τ-th quantile loss function. Sometimes we do not only care about the quantile for the whole data set, but also want to know the quantiles at each location. Therefore, we model the τ-th quantiles as: $Q(\tau|Xi)=X\,i^T\ \beta(\tau)$, that is, at different location Xi, we have different quantile for the response Yi. We have the quantile regression:

$\min\beta\ \frac{1}{n}\sum\rho\tau(Yi-X\,i^T\ \beta)$.

Conditions for a Linear Model

1) Linearity: The relationship between the explanatory and the response variable should be linear.
2) Nearly Normal Residuals: The residuals should be nearly normal.
3) Constant Variability: The variability of points around the least squares line should be roughly constant. This implies that the variability of residuals around the 0 line should be roughly constant as well, which is also called homoscedasticity.

Confidence and Prediction Intervals

A confidence interval provides a range of values which is likely to contain the population parameter of interest. It is used to provide estimation to the parameter. In our cases we are trying to construct confidence interval for E[Y|X]. However, a prediction bounds is used to give estimation for random variable, in this case it gives a range for y itself. The difference between the confidence interval and the prediction interval is the standard error. Because we guess the expected value of E[Y|X] more precisely than we estimate y itself. Estimation y requires including the variance that comes from the true error term which is $Y = a+bx+\varepsilon$. The variance is calculated as

$\mathrm{Var}(\widehat{Y}-Y)=$

$$Var(\widehat{a}+\widehat{b}x-Y)=Var(\widehat{a}+\widehat{b}x-a-bx-\varepsilon)=Var((\widehat{a}-a)+(\widehat{b}-b)x-\varepsilon)\ =\sigma^2(1+\frac{1}{m}\frac{(x-\bar{x})^2}{\sum_{i=1}^{m}(xi-\bar{x})^2})$$

Where the prediction interval is calculated by $(\widehat{a}+\widehat{b}x)\pm t_{m-2,1-\alpha/2}\widehat{\sigma}\sqrt{1+\frac{1}{m}+\frac{(x-\bar{x})^2}{\sum_{i=1}^{m}(xi-\bar{x})^2}}$ it works when m is large. Thus the prediction interval will be wider than a confidence interval which is calculated as $(\widehat{a}+\widehat{b}x)\pm t_{m-2,1-\alpha/2}\widehat{\sigma}\sqrt{\frac{1}{m}+\frac{(x-\bar{x})^2}{\sum_{i=1}^{m}(xi-\bar{x})^2}}$. Moreover, confidence intervals for the predicted values in an linear regression tend to be narrow in the middle and fat at the extremes because the effect of uncertainty in slope at some value x is multiplied by how far you are from

the mean (x-$\bar{x}$) . So the interval will be a minimum at $\bar{x}$, when x becomes far apart from $\bar{x}$, the standard error will become larger, making the prediction interval wider.

Outliers in Linear Regression

Outliers are points that lie away from the cloud of points. Outliers that lie horizontally away from the center of the cloud are called high leverage points. High leverage points that actually influence the slope of the regression line are called influential points.

Hypothesis Testing in Linear Regression

We always use a t-test in inference for regression. b1 is the observed slope, and SE_{b1} is the standard error associated with the slope, and degrees of freedom associated with the slope is df= n-2 where n is the sample size because we lose 1 degree of freedom for each parameter we estimate, and in sample linear regression we estimate 2 parameters, b1 and b0.

Hypothesis test can be constructed that H_0 : b1=0, H_a : b1 ≠ 0 The null hypothesis is often 0 since we are checking for any relationship between the explanatory and the response variable. Then, we generate the t value: T= $\frac{b1-null\ value}{SE_{b1}}$ with df=n-2. If t is bigger than the critical value, we can conclude that there is relationship between the explanatory and the response variable. We can then generate the confidence interval for the slope by using the formula: $b_1 \pm t_{df=n-2} SE_{b1}$.

Linearity between Two Variables

$H_o : b = 0$ (*significance test*)
If we fail to reject H_0 then it means X is not linearly associated to Y.
The test statistic is equal to $\frac{\widehat{b}}{SE(\widehat{b})}$ ~student tn-2. SE($\widehat{b}$) - standard error. P-value = 2P(T> $T_{observed}$). Reject Ho if |T|> $T_{observed}$. Reject Ho if p-value < α .

Model-Misfit

To check whether a model is misfit, we first calculate the residual sum of sequences(RSS), RSS= $\sum_{i=1}^{m}\sum_{j=1}^{k}(Y_{ij}-\widehat{Y}i)^2 = \sum_{i=1}^{m}\sum_{j=1}^{k}(Y_{ij}-\overline{Yi})^2 + k\sum_{i=1}^{m}(\widehat{Y}i-\overline{Yi})^2$ where $\widehat{Y}i = a + bxi$ and the first part of the equation is measurement error(MESS), the second part is model misfit sum of sequences (MMSS). MESS has m(k-1) degrees of freedom, MESS~ $X^2 df$ and MMSS~

$X^2 m-2$. MESS won't change with change of model only MMSS indicates the fitness of model. If $\varepsilon_{ij} \sim N(0,\sigma^2)$ means the model fit well. $\frac{KMMSS/(m-2)}{MESS/m(k-1)} \sim F_{m-2,m(k-1)}(Fisher's\ distribution)$. We reject there is a model misfit if F>3.

If a model is misfit an alternative in linear model is to use a polynomial model. For example a quadratic model E[Y|X] = c+dx+ex^2 . The two-variable linear model, say U=X^2 therefore E[Y|X,U] = c+dx+eu, Yi=c+dx_i + eu_i + ε_i . For $\widehat{a},\widehat{b}$ in least square equation where $\widehat{b}$ is the slope of the regression line. If Xi is uncorrelated with Ui then it means that $\widehat{b}$ is unbiased. Otherwise $\widehat{b}$ is biased, meaning LS regression line fit well.

Multiple Observation

Suppose we have m distinct values of explanatory variables X. For each Xi, we have k replicate measurements, then we have the formula Y_{ij} =a+ $bX_i + \varepsilon_{ij}$ where i=1,2,…,m and j=1,2,...,k. Here, X is the explanatory variable, and Yij is the dependent variable which is jth measurement taken at Xi. We will assume that εij is uncorrelated for all i and j. We use replicate measurement to estimate σ^2 that does not rely too much on the model. If the residual is filed incorrectly, the residuals include measurement error from εij as well as model misfit. Suppose Y_{ij} = c+dX_i + $e\mu_i + \varepsilon_{ij}$ but simple linear model is fit by least squares. Then, Y_{11} , Y_{12} ... Y_{1k} are k uncorrelated. E[Y_{11}] =E[c+dX_1 +e $\mu_1 + \varepsilon_{11}$]=c+dX_i +e μ_i and var(Y_{11}) = σ^2 . From $(S_1)^2$ = $\frac{1}{k-1}\sum_{j=1}^{k}(Y_{1j} - \overline{Y})^2$ and $\overline{Y} = \frac{1}{k}\sum_{j=1}^{k} Yij$, we get $S_1{}^2$, $S_2{}^2$ …, $S_m{}^2$ (no regression), so $S_{pooled}{}^2 = \frac{1}{m}\sum_{i=1}^{m} Si^2$.

Cross-Validation

Cross-validation is a technique used to assess how accurate a predictive model will perform in practice by dividing the original sample into training set (known data) on which model is trained and testing set (unknown data) on which model is tested. In cross-validation, we first leave a set of data points out for validation, then perform our model on the remaining data to provide an estimation interval for the data point that we just left out. Finally, we check where the true data point falls in the estimation interval to assess accuracy of our model. In our linear regression model, we have n real response values $y_{1,} ...y_n$ corresponding to x_1 ,... $x_{n.}$ Suppose we have the function y= $\alpha + \beta x$ to fit the pattern. Then, we can assess the model by calculating mean squared error (MSE), where MSE = $\frac{1}{n}\sum_{i=1}^{n}(y_{actual} - y_{predicted})^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$.

WORKS CITED

Bradic, Jelena. “Chapter 5: Calibrating a Snow Gauge.” MATH 189 Lecture, UC San Diego, 10 May 2018. Lecture.

Judson, Arthur, and Nolan Doesken. "Density of Freshly Fallen Snow in the Central Rocky Mountains." *American Meteorological Society*, 1 July 2000, doi:10.1175/1520-0477(2000)081%3C1577:DOFFSI%3E2.3.CO;2.