Calibrating A Snow Gauge

Joshua Castro, Kevin Elkin, Sneheil Saxena, Yuka Sukamaki, Tony Wu

INTRODUCTION

Northern California's main source of freshwater is located within the Sierra Nevada mountains. In an effort to regulate the supply of water, the Forest Service of the United States Department of Agriculture operates a gamma transmission snow gauge in the Central Sierra Nevada near Soda Springs, CA. This gauge determines a depth profile of snow density.

The snow gauge does not disrupt the natural process of the examined area of snow in the measurement process, meaning the gauge is able to measure the same snowpack a number of times. With replicate measurements compiled from the same volume of snow, researchers are able to study the nature of snow-pack-settlement throughout the winter season while examining the relationship of rain on snow. It is known that as rain falls on snow, the snow absorbs the water until a certain capacity, causing flooding if more rain were to fall on it. As the snow pack increases with density, we find that the capability of the snowpack to absorb water decreases. This pushes us to further analyze the interactions of snowpack with rainfall, in order to progress tactics in monitoring flood management and water supply.

The snow gauge indirectly measure snow density as follows: Gamma ray emissions are used as a medium to measure density, from which the readings are converted into usable numbers. We find that there may be changes in functions used to convert the measured values into density readings throughout the seasons, due to factors such as radioactive source decay and instrument usage wear. A calibration run is conducted at the beginning of each winter season to adjust the conversion method. In this lab, we will develop a procedure to calibrate the snow gauge.

DATA

The data we have is collected from a calibration run of the USDA Forest Service's snow gauge located in the Central Sierra Nevada mountain range near Soda Springs. Polyethylene blocks of known densities, which are used to simulate snow, are placed between the two poles of the snow gauge and are used to take readings in an effort to examine inconsistencies. Thirty measurements are taken on each block. The reported measurements are amplified versions of the gamma photon count made by the detector, which we call the "gain". Our data consists of 10 measurements taken for 9 densities in grams per cubic centimeter of polyethylene.

With this in mind, we find challenges with our data. Firstly, there is a decline in networks located within the northern regions, including locations such as Siberia, Alaska, and Northern Canada. We already have a few stations in the mountain regions, but the reason why we conduct this study is to find ways to sustain and improve the limited operational networks. We also find differences in the quality and compatibility of data across national boundaries. There are evident biases in gauge measurements of solid precipitation. This incompatibility of precipitation data is due to the difference in instruments and data processing methods. We also find it especially difficult to determine precipitation changes in the arctic regions.

BACKGROUND

The snow gauge is a complex and expensive instrument. It is not feasible to establish a broad network of gauges in the watershed area in order to monitor the water supply. Rather, the gauge is primarily used as a research tool.

The snow gauge has utilized in study regarding various scenarios such as snow-pak settling, snow-melt runoff, avalanches, and the interactions of rain-on-snow. Gauges are placed in Idaho, Colorado, Alaska, Russia, Mongolia, China, Japan, and more. We find that the gauge in California is located in the center of a forest opening that is roughly 62 meters in diameter. The laboratory site is at 2099 meters elevations and is subject to all major high altitude storms, which regularly deposit 5-20 centimeters of wet snow. The snowpack reaches an average depth of 4m each winter. The snow gauge consists of a cesium-37 radioactive source and an energy detector mounted on separate vertical poles approximately 70 cm apart.

The lift mechanism at the otp of the poles raises and lowers the source and detector together. The radioactive source emits gamma photons, also known as gamma rays, at 662 kilo-electron-volts in all directions. The detector contains a scintillation crystal which counts those photons eating through the 70-cm gap from the sources to the detector crystal. The pulses generated by the photons that reach the deter crystal are transmitted by a cable to a preamplifier, and then further amplified and transmitted via a buried coaxial cable to the lab. There, the signal is stabilized, corrected for temperature drift, and converted to a measurement that we have termed as the "gain." This should be directly proportional to the emission rate. We find that the snowpack density typically ranges between 0.1 and 0.6 grams per cubic centimeter.

In terms of the physical model, we find that gamma rays emitted from the radioactive source are sent out omnidirectionally. Those sent in the direction of the detector may be scattered or absorbed by the polyethylene molecules located between the source and the detector. With denser polyethylene, we find that fewer gamma rays reach the projector.

For a simpler version of the model, we find that a gamma ray on route to the detector passes a number of polyethylene molecules. This number of molecules is dependent on the density of the polyethylene, so a molecule can either absorb the gamma photon and bounce it out of the path to the detector or allow it to pass. If each molecule were to act independently, then we find that the chance a gamma ray arrives at the detector is p^m ,where p denotes the probability that a single molecule will neither absorb nor bounce the gamma ray and m is the number of molecules moving in a straight line path from the source to the detector. We express this probability as $e^{m*log(p)} = e^{bx}$ where x, the density, is proportional to the number of molecules, m.

Additionally, we would like to investigate the current temperature and pressure trends of the Arctic ocean. Correlation between time, temperature, and pressure in the Arctic troposphere has been distinguished in a previous study. [Highwood] The study has shown patterns in the temperature and pressure trends of the Arctic troposphere, denoting what months of the year typically correlate with temperature maxima, minima, and the general cycle. It has been found that we generally find a

temperature minima in the Arctic troposphere within the months of December, January, and February, in which the study declares a -.58 correlation coefficient with 98% significance. From this, we want to examine temperature and pressure trends on the surface of the earth, as follows in our additional investigations.

HYPOTHESIS

1. We want to investigate whether or not linear models are appropriate figuring out whether or not the current calibrations of the snow gauge are appropriate in examining the water levels of Northern California's supply in Soda Springs, CA.
    a. We hypothesize our investigations will show linear models will be appropriate for the scenarios to be discussed.

2. Additional: Our hypothesis is that we will be able to predict the pressure through the recorded temperature readings of the Arctic region.

SCENARIO 1: FITTING

Before beginning our analysis, we must note potential drawbacks of the data set: (i) if the densities of the polyethylene blocks are not reported exactly and (ii) if the blocks of polyethylene were not measured in random order.

Addressing (i), this will negatively affect our fit of the regression line to predict the density from the gain. An inaccurate reading of the density would cause a shift in our regression line thus leading to inaccurate prediction. Since our data set only has 10 readings per polyethylene block, a misreading of one density will affect our gain readings for that block by a factor of ten.

Addressing (ii), we see that we gather our data from Central Sierra Nevada mountain range, thus this represents our population and we are making predictions of the snow densities here. Thus, we must ensure that we measure blocks across various, random orders to be representative of the population. If we were to not measure in random order, our calibration of the snow gauge would be bias, thus not representative of our region of interest.

We want to fit the gain of the polyethylene blocks to the densities in our data set. We see in Fig. 1 that there are nine different densities with 10 readings each, resulting in the data points being grouped so closely together. We want to predict the snow density based on the gain readings. Thus, we will use linear regression with density as the explanatory variable to predict gain as the response variable.

To perform linear regression, we must satisfy three conditions: (1) the data must be linear, (2) nearly normal residuals, (3) constant variability. So, we must check if the relationship between the explanatory variable and the response variable should be linear. Also, we must show that the residuals are nearly normal. This condition may not be satisfied when there are unusual observations that don't follow the trend of the rest of the data (Bradic). Finally, we must show the variability of points around the least squares line are roughly constant, which implies that the variability of residuals around the zero line should be roughly constant as well.

Looking at the raw data plotted, we can clearly see that there is no linear association; instead, we see a logarithmic curve. Thus, we will transform the gain of the data to get a linear model so that we can perform linear regression. We do this by taking the logarithm of the response variable, depicted in Fig. 2.

Fig. 1 Scatterplot of the data set

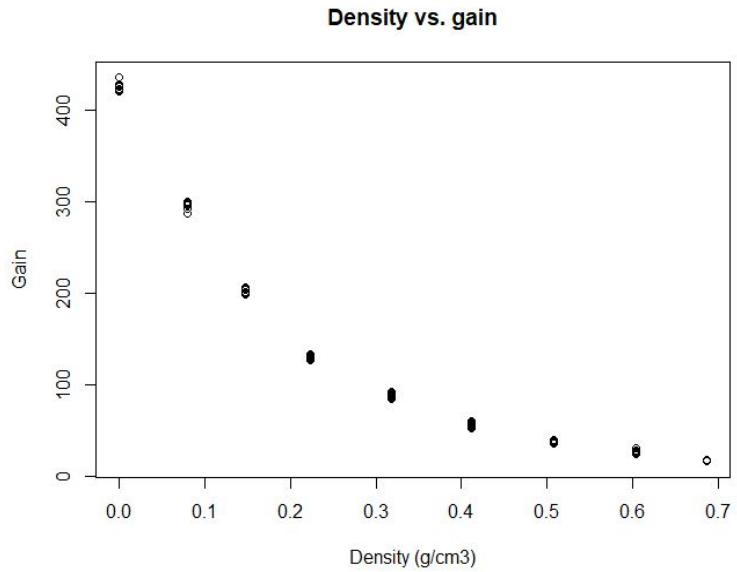
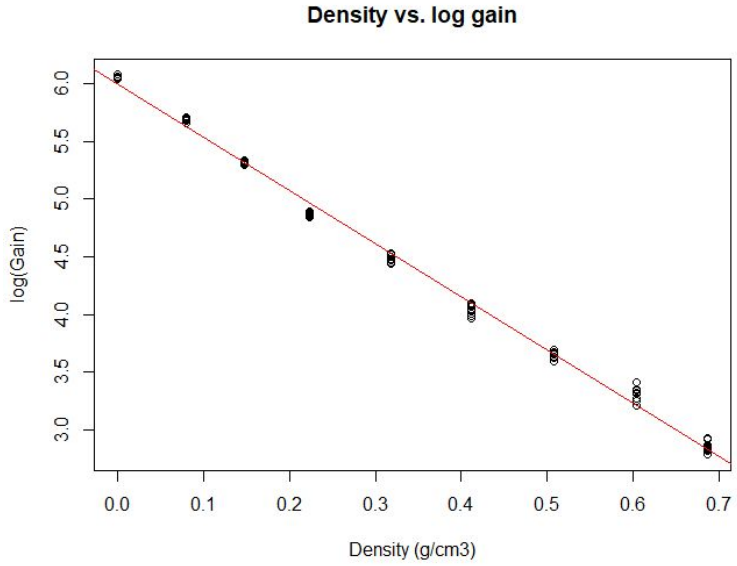
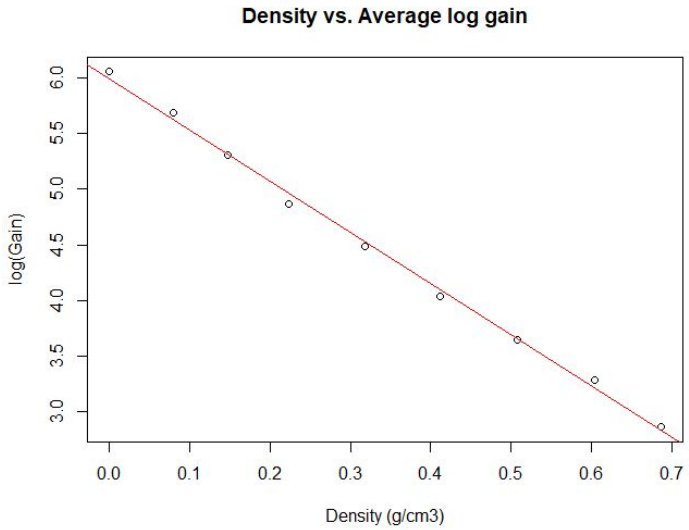


Fig. 2 Scatterplot of the data set where we take the logarithm of the gain with the best fit least square line



We clearly see a strong, negative linear association of the transformed data. Taking the average of the log gain for each polyethylene block (Fig.3), we still see a linearly association of the data. Calculating the correlation coefficient r of the log gain data, we get -0.9979, thus the square of the correlation coefficient (r^2) is 0.9958. So, about 99.58% of the variability in the gain is explained by a linear model. Thus, we can conclude that we satisfy the condition of linearity to best fit a line. So, using linear regression, we calculate that the best fit line is represented by $\widehat{log(gain)} = 5.997 - 4.606 * density$. Interpreting this value, we see the estimated average log gain of the gamma photon count made by the detector is 5.997 and the estimated average log gain of the gamma photon count made by the detector is 4.606 lower per density in grams per cubic centimeter.

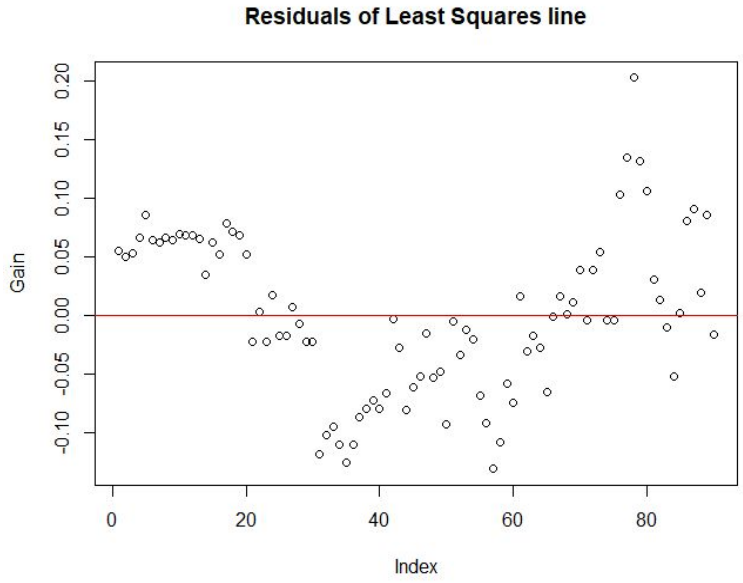
Fig. 3 Scatterplot of the data set where we take the average of the logarithm of the gain for each polyethylene block.



We must also be able to verify nearly normal residuals to perform linear regression. Upon looking at the scatterplot of the residuals (Fig. 4), we see that there may be a slightly parabolic pattern in the data. However, the pattern does not appear very strong, thus we will continue our analysis of linear regression, with caution.

To analyze the normality of the residuals, we will look at the histogram of the residuals (Fig. 5) as well as the QQ plot (Fig. 6). In the histogram, we see that the shape is somewhat normal but may be somewhat skewed with an outlier. Thus, we calculate that the skewness and kurtosis of the residuals are 0.1598 and 2.7085, respectively. So, we see that the skewness is close to zero and the kurtosis is fairly close to three, thus giving us strong evidence that our residuals are normal. Looking at the QQ plot, we see that the sample quantiles do not have any large deviations from the theoretical quantiles line. We see the tail ends have a slight deviation, but does not provide strong evidence of nonnormality. Thus, we conclude that our residuals are nearly normal and we can further continue our analysis of linear regression.

Fig 4. Scatterplot of residuals of the least squares line



Lastly, we must check for constant variability of points around the least squares line. So, looking at the residuals of the least squares line (Fig. 4), we concluded that there is no obvious pattern of the residuals. So, we conclude that the variability of residuals around the 0 line is roughly constant, thus we can conclude we have constant variability of the data points around the least squares line. Therefore, we meet all three conditions of a linear model and will use this throughout our analysis.

Fig 5. Histogram of the residuals of the least squares line

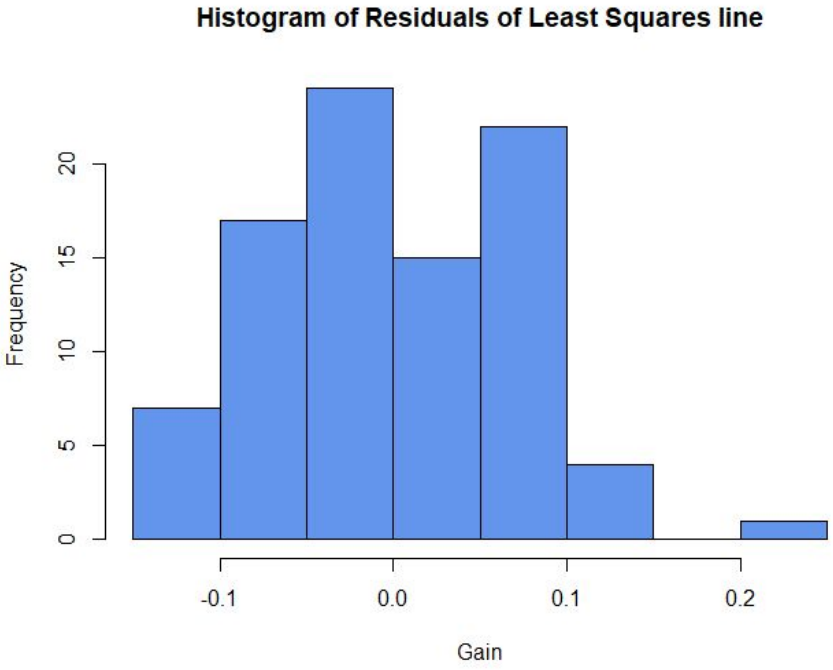
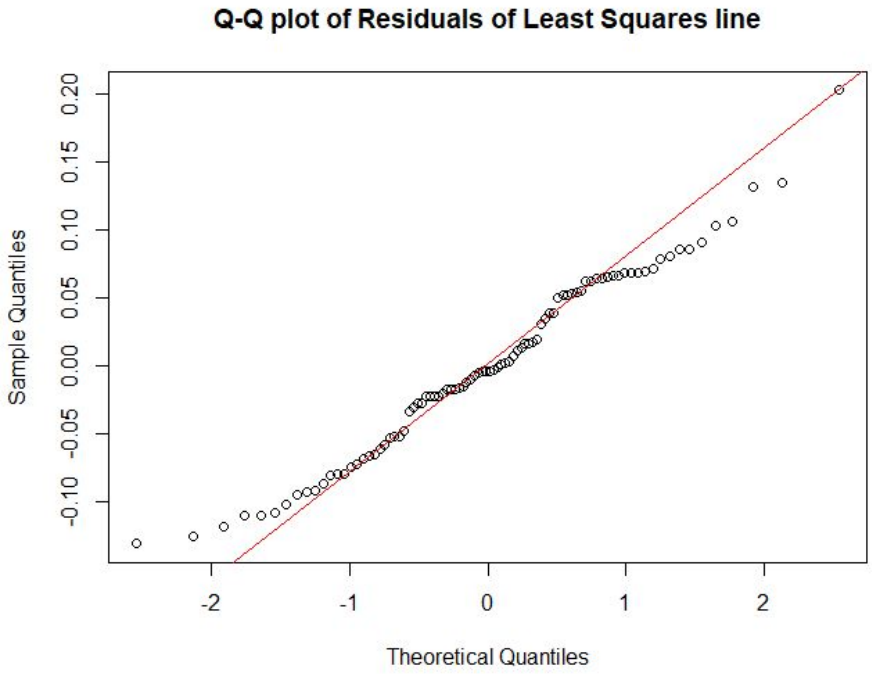


Fig 6. Q-Q plot of the residuals of the least squares line



SCENARIO 2: PREDICTING

After finding our linear model(s) we will predict the density given a specific gain reading. Note that this is the inverse of our linear model Scenario 1. We are asked to predict the density for the following two gain readings: 38.6 and 426.7. Based of these gain readings, we hope to be able to find the snow's density with a high degree of accuracy. Additionally, it is important to mention that these gain readings were chosen due to the fact that they represent the average gains for 0.508 and 0.001 density polyethylene blocks respectively. Our procedure and methods/functions used to make our later calculations and predictions are described in detail in Table I. below:

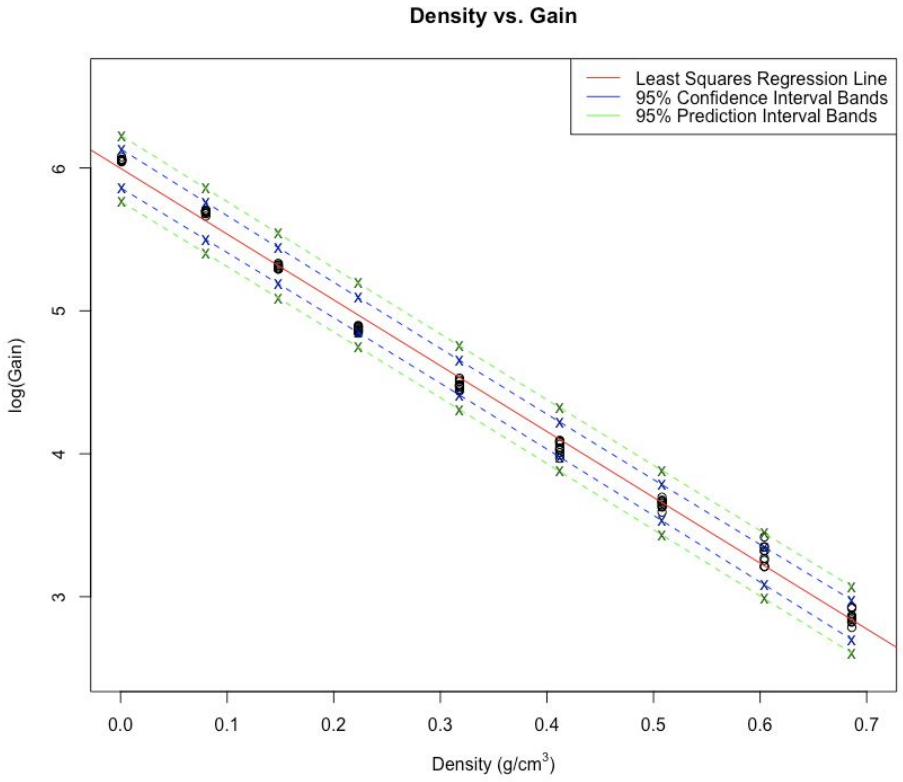
Table I: Functions used to analyze and predict the snow gauge

Function Name	Description
LogGainPredict()	Predict the log(gain) density using least squares regression model
LeastSquareDensityPredict()	Inversely predicts density given the gain using least squares regression
ConLogGainUp()	Creates the upper bound for a 95% confidence interval for log(gain) when given the density
ConLogGainLow()	Creates the lower bound for a 95% confidence interval for log(gain) when given the density

PriLogGainUp()	Creates the upper bound for a 95% prediction interval for log(gain) when given the density
PriLogGainLow()	Creates the lower bound for a 95% prediction interval for log(gain) when given the density
ConDensity()	Creates the 95% confidence interval and outputs a vector of the confidence interval in the form of (lower, upper)
PriDensity()	Creates the 95% prediction interval and outputs a vector of the prediction interval in the form of (lower, upper)

After running our code we get the following graph constructed in Fig.7 with the least squares regression line shown in red, the 95% confidence interval shown in blue, and the 95% prediction interval shown in green. Our estimates in Table II and Table III will use many of the functions contained in Table I.

Fig. 7: A figure representing the density vs. gain data from the dataset which contains a least square regression line to fit the data as well as 95% confidence and prediction intervals



We know that the 95% confidence interval gives us an interval for the true population mean log(gain) given the density; the 95% prediction interval on the other hand will provide us an interval for future predicted log(gain) when given the density. Upon examination of the plot we notice that the 95% prediction intervals are wider than the 95% confidence intervals. This makes sense because the prediction intervals must account for uncertainty in the population in addition to randomly scattered data.

Given this information, we can now use our other functions in our code to calculate the point and interval predictions of the density for the gain readings of 38.6 and 426.7

Table II: A table containing point estimates and the 95% confidence and prediction intervals for a gain reading of 38.6

Gain Reading of 38.6	
Actual Density in dataset	0.508 g/cm^3
Least Squares Regression line point estimate	0.509 g/cm^3
95% Confidence Interval	(0.494, 0.525) g/cm^3
95% Prediction Interval	(0.466 0.553) g/cm^3

Table III: A table containing point estimates and the 95% confidence and prediction intervals for a gain reading of 426.7

Gain Reading of 426.7	
Actual Density in dataset	0.001 g/cm^3
Least Squares Regression line point estimate	-0.013 g/cm^3
95% Confidence Interval	(-0.032, 0.006) g/cm^3
95% Prediction Interval	(-0.058 0.032) g/cm^3

Upon examining our findings in Table II. and Table III. we notice that our least square regression line point estimate produces an estimate that is very close to the actual density contained in the dataset for the gain readings of 38.6 and 426.7. It is important to note that the estimates may contain negative values (specifically for a gain reading of 426.7). We know that these readings are false because density can't be negative; this is likely due to us using a linear regression line to estimate the density.

SCENARIO 3: CROSS-VALIDATION

In Scenario 1 and Scenario 2 we concluded that a linear model was best suited to represent the data we were given. Additionally, we predicted both confidence and prediction intervals for the gain readings we were given. The findings we found in Scenario 2 are useful, however, we do not know how reliable and accurate our predictions for these gain readings are. In order to validate the accuracy of our findings we must separate the original dataset into both training data and testing data to perform cross-validation so we can test the accuracy of our linear model.

Let us omit the set of data points corresponding to the polyethylene block with the density 0.508 g/cm^3. We will use the dataset with this omitted as our training data to produce a linear least squares regression model and 95% confidence and prediction intervals in Fig. 8. We will use the point(s) omitted in this dataset and evaluate it on the test dataset (the part omitted).

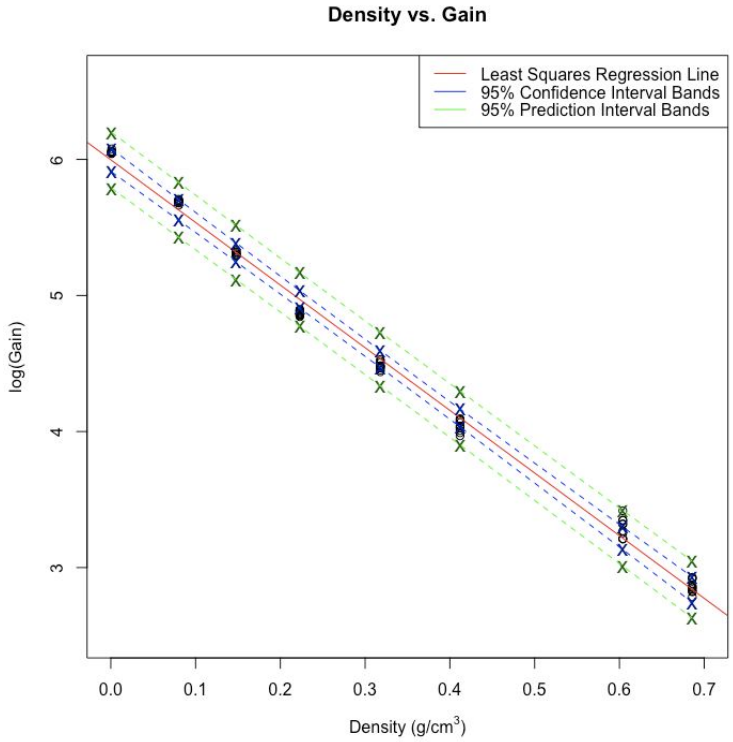
After examining the new linear model created it is difficult to notice any difference between Fig. 7. We will calculate the point and interval estimates for the density of a block with an average gain reading of 38.6 (the average gain reading of the polyethylene block we omitted).

We notice that in Table III. that the 95% confidence and prediction intervals are larger than the ones shown in Fig. 7 which makes sense due to there being less data available in the dataset. The actual density of the block we omitted is 0.508 g/cm^3 and the linear model that used the training data found an estimate of 0.509 g/cm^3. In addition to this we found that the prediction was well represented based off of our prediction interval of (0.462 0.556) g/cm^3. Thus, we can say that the linear model produced from the training data is a very good predictor for the testing data.

Table IV: A table containing point estimates and the 95% confidence and prediction intervals for a gain reading of 38.6 when the set of points containing the polyethylene block with the density 0.508 g/cm^3 is omitted.

Gain Reading of 38.6	
Actual Density in dataset	0.508 g/cm^3
Least Squares Regression line point estimate	0.509 g/cm^3
95% Confidence Interval	(0.492, 0.526) g/cm^3
95% Prediction Interval	(0.462 0.556) g/cm^3

Fig. 8: A figure representing the density vs. gain data from the dataset with the set of data points corresponding to the polyethylene block with the density 0.508 g/cm^3 omitted. The graph contains a least square regression line to fit the data as well as 95% confidence and prediction intervals



Let us omit the set of data points corresponding to the polyethylene block with the density 0.001 g/cm^3. We will use the dataset with this omitted as our training data to produce a linear least squares regression model and 95% confidence and prediction intervals in Fig. 9. We will use the point(s) omitted in this dataset and evaluate it on the test dataset (the part omitted).

After examining the new linear model created it is difficult to notice any difference between Fig. 7. We will calculate the point and interval estimates for the density of a block with an average gain reading of 38.6 (the average gain reading of the polyethylene block we omitted).

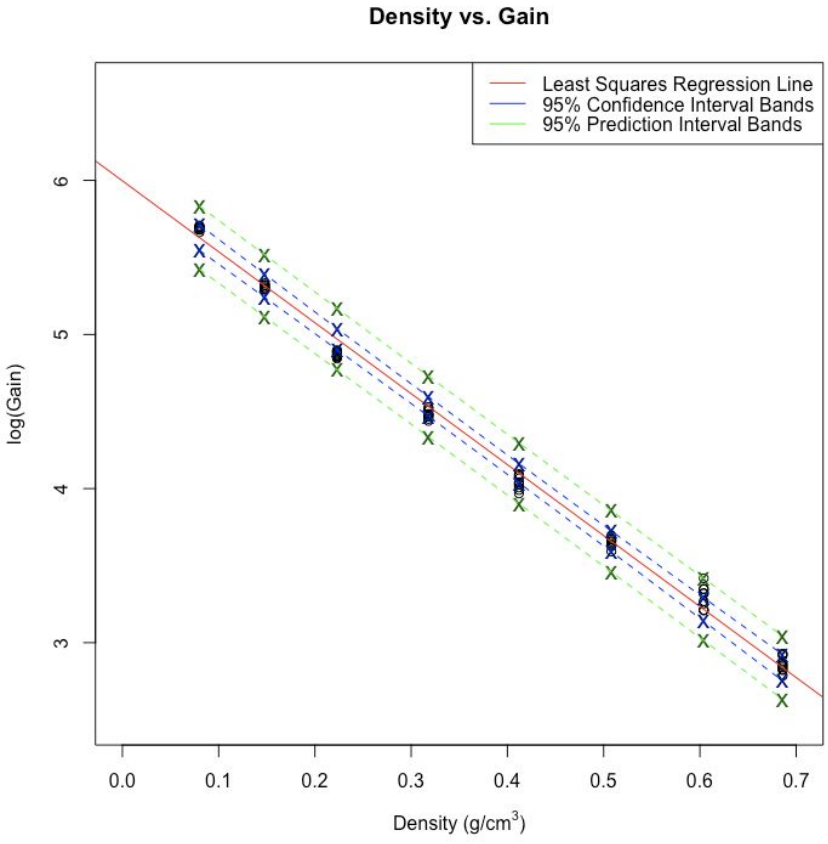
We notice that in Table VI. that the 95% confidence and prediction intervals are larger than the ones shown in Fig. 7 which makes sense due to there being less data available in the dataset. The actual density of the block we omitted is 426.7 g/cm^3 and the linear model that used the training data found an estimate of 426.7 g/cm^3. In addition to this we found that the prediction was well represented based off of our prediction interval of (-0.062 0.035) g/cm^3. Thus, we can say that the linear model produced from the training data is a very good predictor for the testing data.

Based on our two iterations of cross-validation we can conclude that our linear model is very strong and will produce predictions and estimations reliably and accurately.

Table V: A table containing point estimates and the 95% confidence and prediction intervals for a gain reading of 426.7 when the set of points containing the polyethylene block with the density 0.001 g/cm^3 is omitted.

Gain Reading of 426.7	
Actual Density in dataset	0.001 g/cm^3
Least Squares Regression line point estimate	-0.013 g/cm^3
95% Confidence Interval	(-0.034, 0.007) g/cm^3
95% Prediction Interval	(-0.062 0.035) g/cm^3

Fig. 9: A figure representing the density vs. gain data from the dataset with the set of data points corresponding to the polyethylene block with the density 0.001 g/cm^3 omitted. The graph contains a least square regression line to fit the data as well as 95% confidence and prediction intervals



ADDITIONAL INVESTIGATION

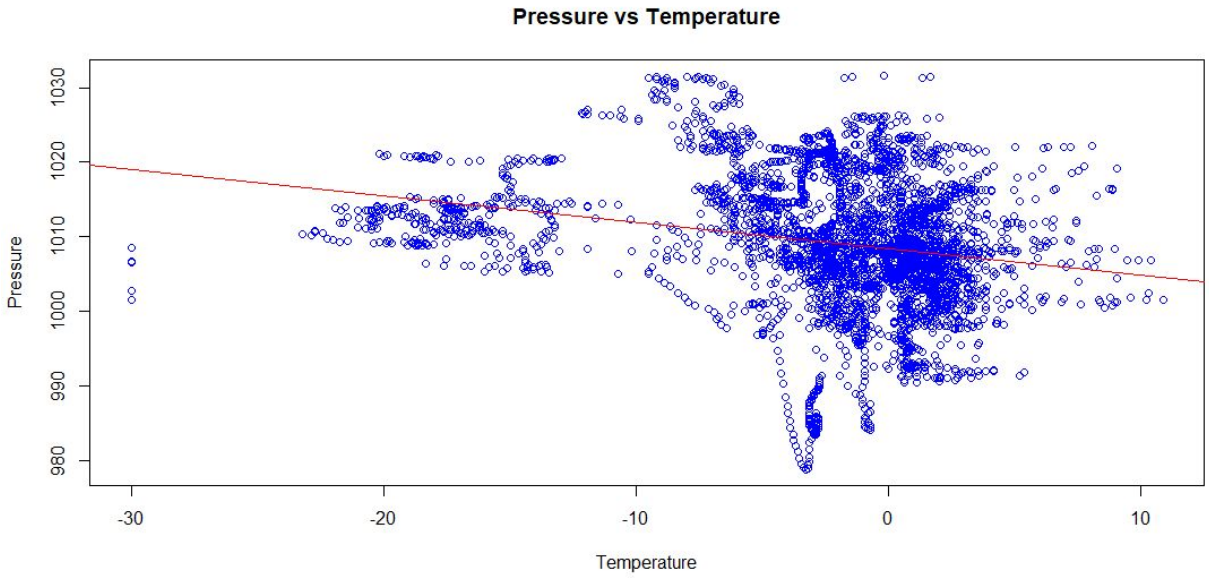
As described above, we are trying to see whether we can predict the pressure given the temperature. We're using data about what the pressure and temperature values were in the Arctic as observed using a buoy in the year 2018.

We can check what the correlation between the pressure and the temperature is. The correlation coefficient, for two random variables X and Y can be defined as corr(X, Y) = $\frac{Cov(X, Y)}{\sigma_X \sigma_Y}$, where Cov(X, Y) is the covariance of X and Y, and $\boldsymbol{\sigma}_X$ and $\boldsymbol{\sigma}_Y$ are the standard deviations of X and Y respectively. The value of corr(X, Y) lies between -1 and 1, where the closer it is to either end, the stronger the correlation and potential dependence between X and Y. A value of 0 indicates that there is no correlation between the two variables.

The correlation coefficient is -0.2293887, which doesn't indicate a strong correlation. The square of the correlation coefficient is 0.05261917568, which means that only about 5% of the variability in the pressure can be explained by the temperature.

We then plot a graph of the temperature and pressure while trying to fit a line:

Fig. 10 Scatterplot of data comparing the temperature to pressure readings.



In order to predict the pressure, given the temperature, we must be able to fit our data. Plotting our data (Fig. 10), it is clear that the data does not have a clear linear pattern. Thus, it is not reasonable to use linear regression to predict the pressure based on temperature. Although not linear, we calculate a least square line that predicts $\widehat{pressure} = 1008.3673 - 0.3552 * temp$. We see that there is not a big change in pressure based on temperature. To further investigate, we will look at the residuals of this least square line and its normality.

The data set is ordered based on time readings, so we see in Fig. 11 that the residuals do in fact follow a pattern in temperature change and pressure. We see that the pressure oscillates based on different temperatures and that there is not constant variability among the residuals. Thus, we do not have a very good read on the estimation in error of this comparison.

Looking at the histogram of the residuals (Fig. 12), we see that it is slightly skewed left and in the QQ plot (Fig. 13), we see there is slight skewness in the left tail end. We calculate the skewness to be -0.3417885 and kurtosis to be 3.665469. Therefore, we can see that the residuals do not follow the normal distribution, so we can conclude that there is very weak linear correlation between the pressure and temperature.

Fig. 11 We use the least squares line best fit for the data to plot the residuals, where each data point is sorted by time.

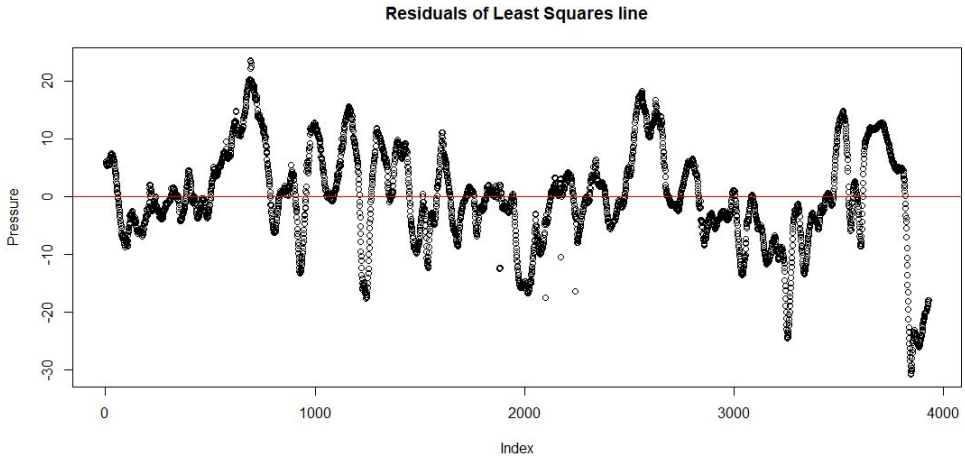


Fig. 12 Histogram of the residuals of the least squares line

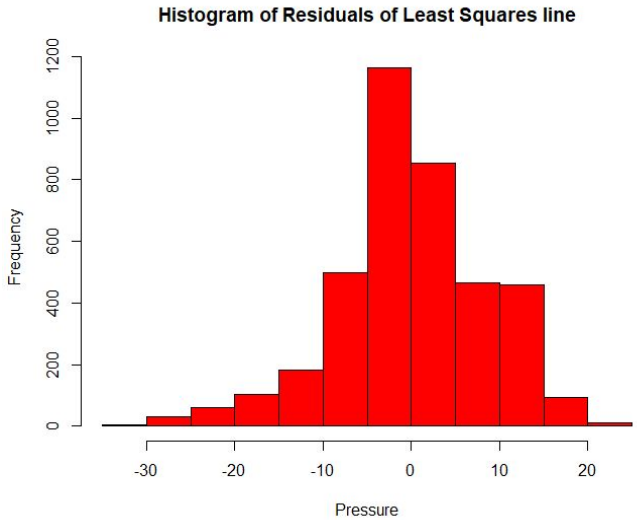
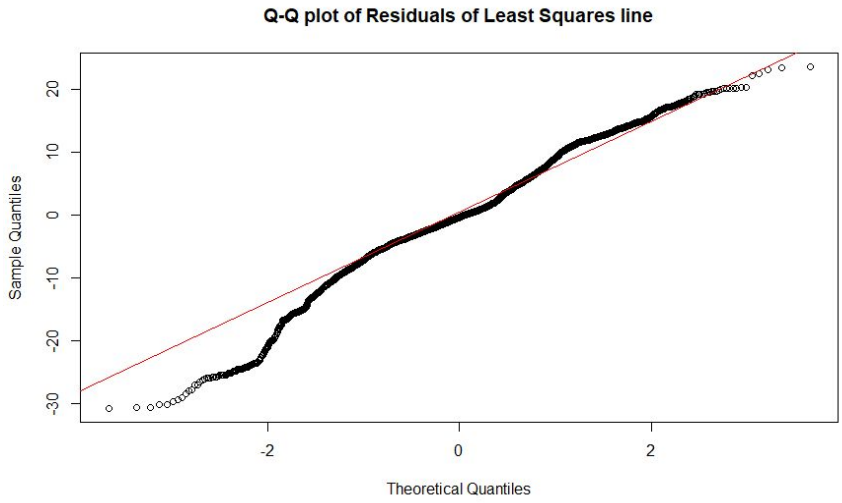


Fig. 13 QQ plot of the residuals of the least squares line



DISCUSSION & CONCLUSION

We go back through each scenario in order to discuss our investigations.

In our investigation of Scenario 1, while we did not immediately see a linear relationship between the density and the gain, we observed a logarithmic curve and thus we were then able to fit the density to the log transformation of the gain, seeing a strong, negative linear association of the transformed data. The square of the correlation coefficient (r^2) was 0.9958 which meant that about 99.58% of the variability in the gain is explained by a linear model. Our calculated best fit line obtained via linear regression is represented by $\widehat{log(gain)} = 5.997 - 4.606 * density$. We find that residuals are nearly normal and there is constant variability within the data, thus we can conclude that the data is, indeed, linear and will be able to use linear regression to analyze and predict data.

As for our investigation of Scenario 2, we now find that we are able to predict the density of the snow with high accuracy, given a specific gain reading. Using the information calculated from our dataset, we find that the predictions made through our computation are in fact accurate as the least squares regression line point estimate closely resembles the findings given in our data for the gain readings of 38.6 and 426.7. Hence, we are able to conclude that use of a linear model is appropriate for our given dataset despite the logarithmic transformation. We continue as follows in Scenario 3.

In Scenario 3, we used our linear models and statistical techniques in order to cross-validate the accuracy of our model with the original dataset. By omitting certain values from our dataset in generating linear models, we wanted to examine whether or not our models will give us the expected omitted values when we input the missing information. In both instances, our models create similar results as described in the omitted data. With this, we validate that our linear models are appropriate in our analysis of the snow gauge.

With these scenarios in mind, we find that we are able to use linear models in order to understand the results of the snow gauge, despite the logarithmic transitions. With this in mind, drawbacks in our data

include the statements made in Scenario 1, as the data may only be representative of the Sierra Nevada region, which we must keep in mind in distributing our results.

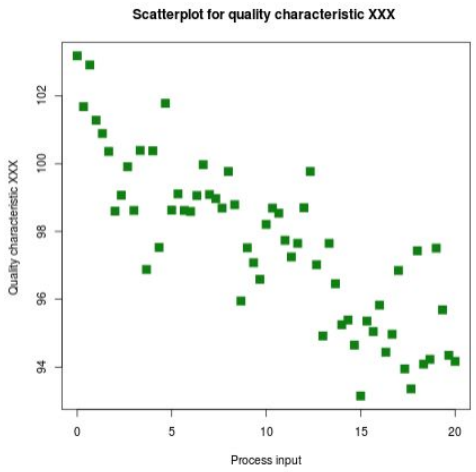
For the additional investigation, we found the correlation coefficient to be -0.2293, indicating that there isn't a strong correlation between the temperature and the pressure. The data did not seem to have a linear relationship and mathematically, it did not meet the conditions to use linear regression either, so we conclude that there is a very weak linear correlation between the temperature and the pressure, thus we cannot reliably predict the pressure based on the temperature.

METHODS AND THEORY

1. Regression Modeling
   a. Line fitting, residuals, correlation

Scatter plot

A scatter plot or scatter plot is a type of mathematical diagram that uses the Cartesian coordinates to show the values of two variables for a data set.



It is used when one or more variables is under the control of the experimenter. If there is a parameter that is increased or decreased systematically by the experimenter, it is called an independent variable or control parameter and is usually represented along the horizontal axis. The measured or dependent variable is usually represented along the vertical axis. If there is no dependent variable, any variable can be represented on each axis and the scatter diagram will show the degree of correlation between the two variables.

A scatter diagram can suggest several types of correlations between the variables with a given confidence interval. The correlation can be positive, negative, or null. You can draw an adjustment line in order to study the correlation between the variables. An equation for the correlation between the variables can be determined by adjustment procedures. For a linear correlation, the adjustment procedure is known as linear regression and guarantees a correct solution in a finite time.

One of the most powerful aspects of a scatter plot, however, is its ability to show the nonlinear relationships between the variables. In addition, if the data is represented by a simple relationship mixing model, these relationships are visually evident as overlapping patterns.

Residual

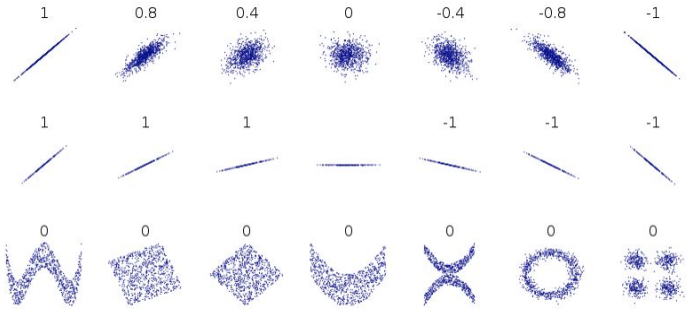
When thinking of statistical models, the words "error" and "residue" come out very similar words. We must distinguish. For example, in equation

(1) $y_i = \mu + 2i$ and

(2) $y_i = \beta 0 + \beta 1 x_i + 2i$,

the true values of the parameters μ and β0 and β1 are unknown. This is not an accurate value. In that sense, “error” refers to the part of the data that can not be explained using the true parameters μ, β0 and β1 etc. Therefore, we can never know the true parameter, so we can never know the value of "error". The part of the data that can not be explained by "estimates of true structure" is called "residue". In equation (2), for example, the ith "error" 2i is 2i = yi-(β0 + β1xi) (yi, xi are obtained from the data, β0, β1 are unknown constants), and i If the second “residual” is ei, then ei = yi − (βc0 + βc1xi) (yi and xi can be obtained from the data, and βc0 and βc1 can be calculated from the data from the estimated value). It is important to keep in mind that “error” is a value that you can never know and “residual” is a value that can be calculated in practice. In other words, “residual” can be thought of as an estimate of “error”.

Correlation



The correlation coefficient is an index that measures the strength of the linear relationship between two random variables. The correlation coefficient is a dimensionless quantity and takes values of real numbers between -1 and 1 inclusive. When the correlation coefficient is positive, the random variable has positive correlation, and when the correlation coefficient is negative, the random variable has negative correlation. Also, when the correlation coefficient is 0, the random variable is said to be uncorrelated.

For example, the unemployment rate and the real economic growth rate of developed countries are strongly negatively correlated, and the correlation coefficient is relatively close to -1.

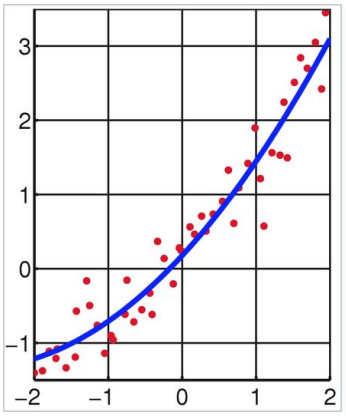
The correlation coefficient takes a value of ± 1 only when and when the two random variables have a linear relationship. If the two random variables are independent of each other, the correlation coefficient is 0, but the reverse is not true.

Usually, simply speaking the correlation coefficient indicates Pearson's product moment correlation coefficient. The test of Pearson product moment correlation coefficient is a (parametric) method that assumes a normal distribution of deviations, but Spearman's rank correlation coefficient, Kendall's rank phase is a nonparametric method that does not put such assumptions in mind Relational numbers are also commonly used.

b. Fitting a line by least squares regression

In the least squares method, when approximating a set of numerical values obtained by measurement using a specific function such as a linear function assumed from an appropriate

model or a logarithmic curve, the assumed function is good for the measured value The approximation is performed by a method of determining a coefficient that minimizes the sum of squares of residuals so as to be an approximation, or by such a method.



In the least squares method, the measured data y is the sum of the model function f (x) and the error ε.

It can be expressed as

$$y = f(x) + \varepsilon .$$

The measured data of physical phenomena include errors, which include systematic errors and accidental errors. Of these, accidental errors are often expected to be normally distributed if they are derived from microscopic phenomena of the signal path in the measurement. There is also a way of thinking that even if it is difficult to identify the reason for the error such as social survey, it is expected that the error will be a normal distribution.

It should be noted that the model function obtained by the least squares method is not plausible if the error does not follow a normal distribution. If the random error is not normally distributed, if the systematic error is not large enough to be included in the model function, the measured data may include outliers that are greatly deviated from the normal distribution.

Including the above, the following assumptions have been made on the theoretical basis of the least squares method.

There is no bias in the error of the measured value. That is, the average value of the errors is zero.
The variance of the measurement error is known. However, the different value may be used for each measurement data.
Each measurement is independent of one another and the covariance of the error is zero.
The errors are normally distributed.
There is known a model function f that includes m parameters, and there are parameters that can reproduce the true value of the measurement without approximation errors.

For the sake of simplicity, let us assume that the measured values are distributed in a two-dimensional plane of x, y, and the assumed distribution (model function) is of the form y = f (x). Assume that the function f to be assumed is represented by a linear combination of known functions g (x). That is,

$$f(x) = \sum_{k=1}^{m} a_k g_k(x)$$

For example, gk (x) = xk-1 is a polynomial approximation, and in particular, when m = 2, $f(x) = a_1 + a_2(x)$ It is an approximation by the straight line a_2 x.

Now, suppose that there is a set of the following set of numerical values obtained by measurement.

$$\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$$

Assuming that the distribution of (x, y) follows the model function y = f (x), theoretical values assumed are (x1, f (x1)), (x2, f (x2)),. .., (xn, f (xn)), and the residual with the actual measurement value becomes | yi − f (xi) | for each i. The magnitude of this residual is also the distance between (xi, yi) and (xi, f (xi)) on the xy-plane.

Here, the estimate of the variance of the error from the theoretical value is the sum of squares of the residuals

$$J = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

Therefore, it is sufficient to define the assumed distribution f so that J is minimized.

In order to do that, since the above equation can be regarded as a function with a_k as a variable, let J be a partial derivative of a_k with zero. Solve the m simultaneous equations obtained in this way, and determine a_k .

In the case of a linear equation
As a simpler example, let the model function be a linear function,

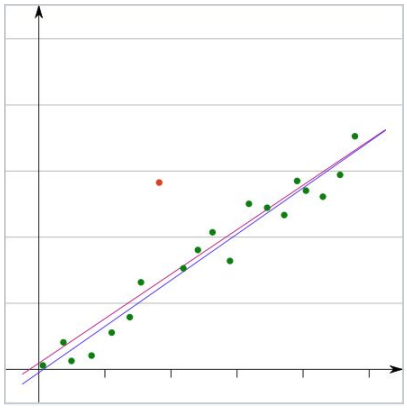
$$f = ax + b$$

Then, a and b can be obtained by the following equations:

$$a = \frac{n\sum_{k=1}^{n} x_k y_k - \sum_{k=1}^{n} x_k \sum_{k=1}^{n} y_k}{n\sum_{k=1}^{n} x_k^2 - \left(\sum_{k=1}^{n} x_k\right)^2}$$

$$b = \frac{\sum_{k=1}^{n} x_k^2 \sum_{k=1}^{n} y_k - \sum_{k=1}^{n} x_k y_k \sum_{k=1}^{n} x_k}{n\sum_{k=1}^{n} x_k^2 - \left(\sum_{k=1}^{n} x_k\right)^2}$$

c. Outliers in linear regression
Outliers are values that deviate significantly from other values in the statistics.
Although it differs conceptually from an outlier due to a measurement error, a recording error, etc., it may not be practically distinguishable.



↑The red point is the outlier.

For a sample that you want to test for outliers, find the test statistic by dividing the deviation by the unbiased standard deviation is greater than the significance point.

$$\tau_1 = \frac{x_1 - \mu}{\sigma}$$

In the simple method, 2 or 3 is considered as a significant point. In other words, it is an outlier if it is outside $\mu \pm 2 - 3\ \sigma$.

There are two ways to test, *Smirnov-Grabbs test* and *Thompson test* .

More precisely, we use the *Smirnov-Grabbs test*, assuming a normal distribution.

Let $\tau = \frac{(n-1)t}{\sqrt{n(n-2)+nt^2}}$ be the significance point, where n is the number of samples, the required significance level is α and $\alpha / n \times 100$ percentile of the t distribution with n-2 degrees of freedom is t. This expression is used recursively. That is, only the sample that deviates the most is tested, and if it is determined to be an outlier, the sample that deviates second is tested using $n - 1$ samples excluding it, and then the outlier is detected Repeat until no longer.

The *Thompson test* uses $t = \frac{\tau\sqrt{n-2}}{\sqrt{n-1-\tau^2}}$.

On the contrary to the Smirnov-Grabbs test, the significance level α_1 is often determined from the test statistic τ_1 of the sample value through t_1 for the convenience of the formula. If n is large enough, the result is the same as the Smirnov Grubbs test.

d. Inference for linear regression

Inference of β_0 and β_1 : Hypothesis test and confidence interval

1. Hypothesis test of the slope $H_0 : \beta_1 = \beta_{10}$ or

the intercept $H_1 : \beta_0 = \beta_{00}$

Theorem: Let $\widehat{\beta_0}, \widehat{\beta_1}$ be the LSE or MLE of β_0, β_1 .

Let $S^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - (\widehat{\beta_0} + \widehat{\beta_1} x_i))^2$

Then (1) $T = \frac{\widehat{\beta_1} - \beta_1}{s / \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \sim t_{n-2}$

(2) $T = \frac{\widehat{\beta_0} - \beta_0}{s \sqrt{\sum_{i=1}^{n} x_i^2} / \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \sim t_{n-2}$

Proof: (1) Sice $\widehat{\beta_1} \sim N(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2})$

$$T = \frac{\widehat{\beta_1}-\beta_1}{\sigma/\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \cdot \frac{\sigma}{S}$$

$$= \frac{\widehat{\beta_1}-\beta_1}{\sigma/\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \div \sqrt{\frac{(n-2)S^2}{\sigma^2}/(n-2)} \sim t_{n-2}$$

Also, $\widehat{\beta_1}$ and S^2 are independent.

(2) can be proved similarly by using $\widehat{\beta_0} \sim N(\beta_0, \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i-\bar{x})^2})$ and $\widehat{\beta_0}$ and S^2 are independent.

Test for the slope β_1 in the regression line.

Step1: State a null hypothesis H_0 and an alternative hypothesis H_1

$H_0 : \beta_1 = \beta_{10}$ versus

$H_1 : \beta_1 \neq \beta_{10} (or\ \beta_1 > \beta_{10},\ \beta_1 < \beta_{10})$

Step2: The test statistic $T = \frac{\widehat{\beta_1}-\beta_{10}}{s/\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \sim t_{n-2}$ if H_0 is true.

Calculate the observed test statistic t.

Step3: Find the critical region.

If $H_1 : \beta_1 > \beta_{10}$, *critical region* $= \{t \geq t_{\alpha,n-2}\}$

If $H_1 : \beta_1 < \beta_{10}$, *critical region* $= \{t \leq -t_{\alpha,n-2}\}$

If $H_1 : \beta_1 = \beta_{10}$, *critical region* $= \{t \geq t_{\alpha/2,n-2}\ or\ t \geq t_{\alpha/2,n-2}\}$

Step4: We can have conclusion.

Confidence interval of β_0 and β_1 when σ^2 is unknown.

The (1-α) 100% confidence interval for β_1 is

$$\left(\widehat{\beta_1} - t_{\alpha/2,n-2} \cdot \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}, \widehat{\beta_1} + t_{\alpha/2,n-2} \cdot \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right)$$

The (1-α) 100% confidence interval for β_0 is

$$\left(\widehat{\beta_1} - t_{\alpha/2,n-2} \cdot \frac{s\sqrt{\sum_{i=1}^{n} x_i^2}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}, \widehat{\beta_1} + t_{\alpha/2,n-2} \cdot \frac{s\sqrt{\sum_{i=1}^{n} x_i^2}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right)$$

2.Multiple Observations

Multiple regression

Multiple regression analysis is one of multivariate analysis. Two or more independent variables in regression analysis. One independent variable is called single regression analysis.

The generally used least squares method and multiple regression of generalized linear models are mathematically a type of linear analysis and are mathematically similar to analysis of variance etc. By selecting a plurality of appropriate variables, it is possible to create a prediction equation which is easy to

calculate and has a small error. The coefficient of each explanatory variable of the multiple regression model is called a partial regression coefficient. The degree of influence on the objective variable does not indicate the partial regression coefficient, but the standardized partial regression coefficient indicates the degree of influence on the objective coefficient. The following relational expressions are known.

$$SPRC = PRC \times \frac{SDEV}{SDRE}$$

SPRC: Standardised Partial Coefficient
PRC: Partial Regression Coefficient
SDEV: Standard Deviation of Explanatory Variable
SDRV: Standard Deviation of Response Variable

Coefficient of Determination

The coefficient of determination is a value representing how much the independent variable can explain the dependent variable. Sometimes called contribution rate. It is used as a measure of the fitness of the regression equation obtained from sample values.

Definition

There is no clearly agreed definition of the coefficient of determination. According to Tarald O. Kvalseth, there are eight types of definitions that require attention. However, it seems that it is generally define the following equation (sample value <measured value, observed value> is y, estimated value by regression equation is f).

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}$$

That is, it is a value obtained by subtracting one obtained by dividing the sum of squares of the residuals by the sum of squares of differences from the average of the sample values, from 1; It should be noted that the least squares method is a parameter selection method that maximizes this definition.

In the case of general linear regression, the following equations are equivalent, and they may be defined as defined equations.

1. The variance of the estimates divided by the variance of the sample values

$$\frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}$$

2. Squared correlation coefficient between the actual value and the estimated value

$$\frac{(\sum_i (f_i - \bar{f_i}) \cdot (y_i - \bar{y_i}))^2}{\sum_i (f_i - \bar{f_i})^2 \cdot \sum_i (y_i - \bar{y_i})^2}$$

In cases other than linear regression, when it is required to pass through the origin, in the case of regression by methods other than the least squares method, these equations may not necessarily be equivalent to the above definition, so care must be taken.

Degree of freedom adjusted decision coefficient

The above definition of the coefficient of determination tends to be better as the number of explanatory variables is increased. Therefore, with the number of explanatory variables as p and the number of

samples as N, the following degrees of freedom may be adjusted, which is called a coefficient of adjustment with adjusted degrees of freedom (adjusted R^2).

$$R'^2 = 1 - \frac{\sum_i (y_i - f_i)^2/(N-p-1)}{\sum_i (y_i - \bar{y})^2/(N-1)}$$

In addition, although it is set as "the number of explanatory variables", when not using linear regression, for example, also using the term of a square or the term of a cube with respect to the same explanatory variable, adjustment for that much is also needed. It may be said that the number of parameters excluding constant terms.

Sum of squares

In statistics, residual sum of squares is the sum of squares (squares) of residuals. It is also called residual sum of squares, sum of squared residuals (SSR) or sum of squared errors of prediction (SSE). Residual sum of squares is a measure to assess the discrepancy between the data and the estimated model. Small RSS values indicate that the model fits the data exactly.

In general, total sum of squares = sum of squares explained + residual sum of squares.

1. For models with a single explanatory variable, RSS is given by
$$RSS = \sum_{i=1}^{n} (y_i - f(x_i))^2$$
2. At this time, yi is the value of the i-th variable, xi is the value of the i-th explanatory variable, and f (xi) is the predicted value of yi. In a standard linear simple regression model, $y = a + bx_i + \varepsilon_i$ (where a and b are coefficients, y and x are dependent and independent variables, respectively, and ε is an error term). The residual sum of squares is a sum of squares of estimators of ε i and is expressed by the following equation.
$$RSS = \sum_{i=1}^{n} (\varepsilon_i)^2 = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$
At this time, α is an estimated value of the constant term a and β is an estimated value of the regression coefficient b.

Collinearity

If there is a very strong correlation between independent variables or there is a linear dependent variable relationship, analysis is not possible or the reliability is low even if the result is obtained. In this case, it is said that there is multicollinearity.

If you do not perform variable selection, it is better not to include those with high correlation among the independent variables.

If they are not independent (for example, if both variables A and B and their sum C = A + B are included) the analysis will fail.

In some cases, the sign of the correlation coefficient between each independent variable and the dependent variable may not coincide with the sign of the partial regression coefficient. This is because the partial regression coefficient is determined from the viewpoint of "predicting", so the sign is determined by the relationship between the variables included in the multiple regression equation. This happens because there is a mixture of highly correlated independent variables (some variables may be

over-predicted with other variables). . However, since such a thing is inconvenient in considering the causal relationship, it would be better to search multiple regressions excluding independent variables whose signs do not match.
Let rii be the element of the inverse of the correlation coefficient matrix between independent variables, then

$$R = \sqrt{1 - 1/r^{ii}}$$

It is a multiple correlation coefficient when predicting the independent variable i with the remaining independent variables. Therefore, a large number indicates that it is not suitable as an independent variable.
The same thing as this, but sometimes 1 / rii is called tolerance, and rii is called variance expansion coefficient. In this case, it is better to exclude independent variables with low tolerance (large variance expansion factor).
Backward-elimination /p − value approach
One of the methods to select a combination of highly explanatory variables in multiple regression analysis. In the variable group prepared in advance, inappropriate variables are sequentially eliminated to improve the accuracy of multiple regression analysis.

On the other hand, the method of adding more and more variables from the place without variables is called the variable increasing method, and the method of increasing and decreasing the variables is called variable increasing and decreasing method (stepwise method). In pay statistical software, etc., the method of increasing and decreasing variables is often used, but in normal Excel, since the function is not equipped as a standard, the variable reduction method is often used. The procedure is as follows.

1. Create a correlation matrix using the "correlation" of the analysis tool. In order to prevent multicollinearity, variables with a correlation coefficient of 0.9 or more are excluded from explanatory variables at this stage (usually leaving the cause system). In addition, multicollinearity is the partial regression coefficient of a variable that results in the calculation result of partial regression coefficient becoming unstable if two or more explanatory variables with strong correlation are included in the explanatory variable, and it should be positive correlation by itself. It means that it becomes difficult to interpret the result, such as the sign inversion becoming negative. As a coping method, remove one of the highly correlated explanatory variables. However, when the purpose is only prediction and the interpretation of the coefficients is not important, multicollinearity does not have to worry much.

2. Perform regression analysis from the analysis tool using all those explanatory variables.

3. Create a regression equation by removing one explanatory variable with the largest P value among the results, and repeat this.

4. Select the model with the largest correction R2. Check if the P value of each explanatory variable of the model is smaller than approximately 5% (or 10%).
Logistic Regression

The general linear model is one of the linear models used in statistics. Among linear models, those whose residuals follow a multivariate normal distribution are general linear models and those whose distribution is arbitrary are generalized linear models. Both can be abbreviated as GLM, but in the R language the generalized linear model is lm () and the generalized linear model is glm ().
It is expressed by the following equation.

$Y = XB + U.$
In this equation, Y is a multivariate data matrix, X is a design matrix, B is a matrix containing predicted parameters, and U is a residual. The residual follows a multivariate normal distribution.

General linear models include several statistical models such as analysis of variance (ANOVA), analysis of covariance (ANCOVA), multivariate analysis of variance (MANOVA), multivariate analysis of covariance (MANCOVA), linear regression, t-test, F-test, etc. It is built in. If the number of rows in Y is 1 (ie, the dependent variable is 1), the general linear model can also be applied to multiple regression analysis.

ACKNOWLEDGEMENTS

Bradic, Jelena. "Chapter 5: Calibrating a Snow Gauge" Math 189, UC San Diego, 13 Mar 2019. Lecture

Bradic, Jelena. "Chapter 6: Introduction to Linear Regression " Math 189, UC San Diego, 11 Mar 2019. Lecture

"Coefficient of determination." *Wikipedia*, Wikimedia Foundation, 27 Feb. 2019, en.wikipedia.org/wiki/Coefficient_of_determination

"Collinearity." *Wikipedia*, Wikimedia Foundation, 27 Jan. 2019, en.wikipedia.org/wiki/Collinearity

Highwood, E. J., Hoskins, B. J., & Berrisford, P. (2000). Properties of the Arctic Troposphere. *Royal Meteorological Society,126*(565), 1515-1532. Retrieved from https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.49712656515.

IABP Drifting Buoy Pressure, Temperature, Position, and Interpolated Ice Velocity, Version 1. (1979, January 19). Retrieved from https://nsidc.org/data/G00791

International Arctic and Buoy Programme. (n.d.). Retrieved from http://iabp.apl.washington.edu/data.html

"Least squares." *Wikipedia*, Wikimedia Foundation, 14 Mar. 2019, en.wikipedia.org/wiki/Least_squares

"Linear regression." *Wikipedia*, Wikimedia Foundation, 2 Mar. 2019, en.wikipedia.org/wiki/Linear_regression

"Logistic regression." *Wikipedia*, Wikimedia Foundation, 16 Mar. 2019, en.wikipedia.org/wiki/Statistical_hypothesis_testing

"Outlier." *Wikipedia*, Wikimedia Foundation, 4 Mar. 2018, en.wikipedia.org/wiki/Outlier

"Residual." *Wikipedia*, Wikimedia Foundation, 17 Mar. 2019, en.wikipedia.org/wiki/Residual

"Stepwise regression." *Wikipedia*, Wikimedia Foundation, 11 Dec. 2018, en.wikipedia.org/wiki/Stepwise_regression

"Sum of squares" *Wikipedia*, Wikimedia Foundation, 5 Jun. 2018, en.wikipedia.org/wiki/Sum_of_squares