Case Study 4: Calibrating Snow Gauge

Question

How accurate are gauge measurements, or "gain", to the density of various blocks of snow? Given actual data on the gauge measurements, could we predict the corresponding density?

Hypothesis

We hypothesize that there is a positive correlation between the gauge measurements, or "gain", and the snow density. Specifically, as the gauge measurement value increases, the snow density increases too. We hypothesize that the association will be strong but not exact due to machine or human error.

Literature Review

Precipitation, in general, is one of the most important atmospheric variables for ecosystem research. Because of the extreme importance that precipitation has on climate monitoring, measurement errors for solid precipitation need to be addressed to get more accurate results. According to the literature, measurement errors range from 20% to 50% due to undercatch in windy conditions (Rasmussen). These errors are partly due to the environment, which has a greater impact on measurement accuracies of snow than on rainfall. In particular, the accuracy of a snow measurement is more impacted by local wind than that of a rainfall measurement. In addition, the environment provides redistribution of snow and metamorphosis, both of which result in major temporal and spatial variability. Another issue with snow measurement is the particular design of the measurement mechanisms, whether the design is a spring mechanism and chart recorder, load cells, strain sensors, or some other design. According to the literature, the standard shielded gauge for measurement recorded 150%-200% more precipitation than an unshielded tipping-bucket gauge because of different reasons, one of which is that the snow was removed from the gauge orifice by the wind before it is melted (Rasmussen). Because of this issue, the most significant challenge for the measurement of solid precipitation, regardless of the gauge type or measurement mechanism, is the windy environment where the measurement occurs. Specifically, the different wind metrics, such as speed and temperature, combined with precipitation characteristics and gauge type all influence the measurement of solid precipitation.

Introduction

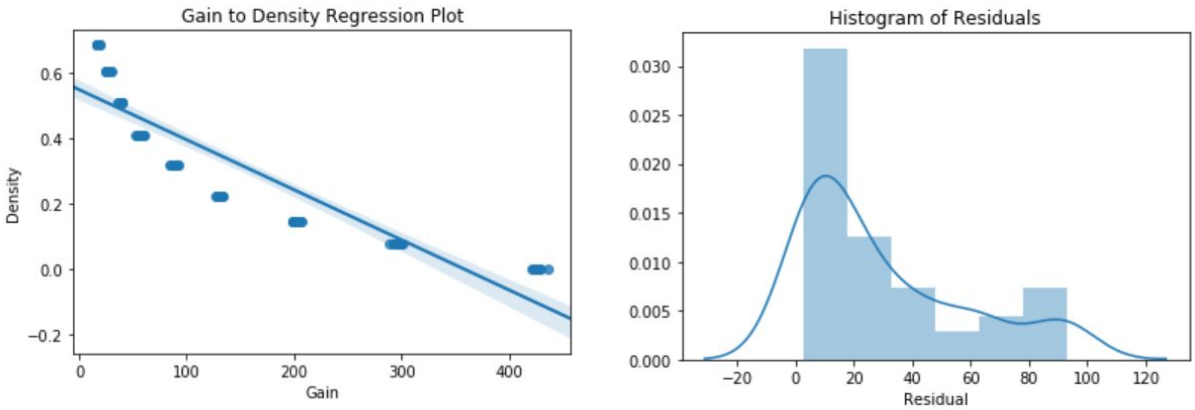
The given dataset (gauge.txt) contains data from a calibration run of the USDA Forest Service's snow gauge located in the Central Sierra Nevada mountain range near Soda Springs, CA. To actually collect data from the snow gauge, polyethylene blocks, which are used to simulate snow, of known densities are placed between the two poles of the snow gauge and take readings on the blocks. For each set of polyethylene blocks, people take 30 measurements, of which only the middle 10 are reported in the dataset. The gauge measurements, or "gain", reported are amplified

versions of the gamma photon count made by the detector. We call the gauge measurement the ”gain”. The given dataset (gauge.txt) consists of 10 measurements for each of 9 densities in grams per cubic centimeter of polyethylene.
In terms of data limitations, there is a decline of the operational networks in the northern regions of the world, such as Alaska and northern Canada. In addition, there are few stations in the mountain regions, and there is some concern on how to sustain and improve the operational networks. In addition, there are data limitations in terms of working with this dataset internationally. Specifically, the difference in instruments and methods of data processing in different nations leads to the incompatibility of precipitation data. In addition, different geographic regions, such as the arctic regions, provide difficulties to determine precipitation changes. Furthermore, there are large biases in the gauge measurements of solid precipitation. From these challenges, we need some form of validation of the precipitation data, such as from satellite data.

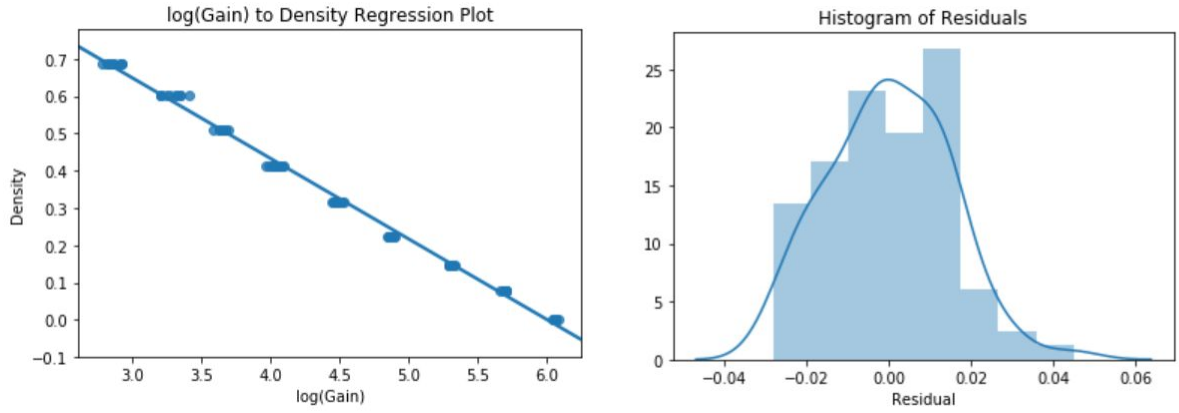
Background
Different geographic regions have various sources of water. The main source of water for Northern California, for example, is the Sierra Nevada mountains. To help monitor the water supply, the Forest Service of the United States Department of Agriculture (USDA) operates a gamma transmission snow gauge in the Central Sierra Nevada near Soda Springs, CA. The gauge is used to determine a death profile of snow density. The snow gauge keeps the snow intact in the measurement process, which allows to perform repeat measurements of the snow-pack. By performing routine measurements of the snow-pack, researchers can study the settlement III of the snow-pack throughout the winter season and the influence of rain on snow. One dynamic of rain on snow is that when rain falls on snow, the snow absorbs the water up to a certain threshold, after which flooding occurs. The denser the snow-pack, the less water the snow-pack can absorb. To determine the amount of influence that the water has on the snow-pack, some data analysis on the snow-pack helps monitor the water supply and flood management.

Basic Analysis

Fitting the Data
We began by reading the data into a pandas DataFrame with columns for gain reading and measured density. A scatterplot of this data with gain on the x-axis and density on the y-axis is as follows:

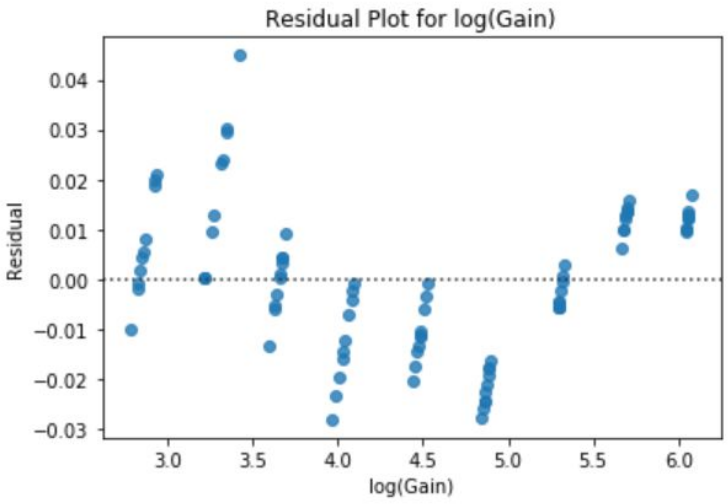


The data does not look linear, so a linear regression model would not be appropriate. From here, we considered a variety of transformations of the data such that the linear regression model would be a more appropriate fit. We eventually determined a logarithmic transformation of gain to be the approach that yielded the most linear scatterplot as well as small and reasonably normal residuals. The regression plot and residual plot appear as follows:



The R^2 of log(gain) to density is 0.9958, indicating a very good fit.

By transforming the data such that the predictor variable is the natural log of the gain and the response is the density, we can see that the result has a linear relationship as well as small and nearly normal residuals. The final condition to run a linear regression is constant variability around the regression line, or homoscedasticity. The homoscedasticity of the log-transformed is not perfect as demonstrated by the residual plot:

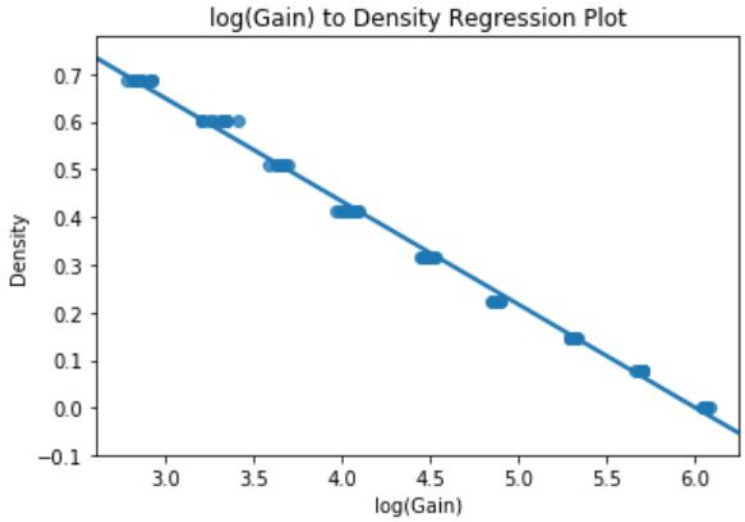


Due to this condition, we should proceed with care when using a linear model to predict density from log(gain). Also, part of the reason for this failure is explained by the context of the problem. Since the data was gathered by taking multiple measurements of gain from snow blocks, groups of data points with the same density yield gain readings that are precise to one another but not necessarily accurate to the linear model. This exacerbates the magnitude of the variability around the regression line. Nonetheless, we should assess the results of our predictions in context of limited homoscedasticity.

Further, it is important to consider other questions about how the data was gathered and reported before we proceed. If the densities of the snow blocks were not reported exactly, this would drastically impact the regression plot and potentially raise questions about the appropriateness of the linear model. Because each measured density corresponds to multiple gain readings, even slight errors in density reporting will lead to significant differences in computed statistics. We also have little information about how and when the data was gathered. If the gain readings were not measured in a random order, possible confounders could include but are not limited to: time of day of reading, time of year of reading, and natural wear of snow gauge.

Predicting Density

Once we have chosen to proceed with a linear regression model for our transformed data, we can use the regression line to generate predictions of density from log(gain). The plot with an overlaid regression line is as follows:

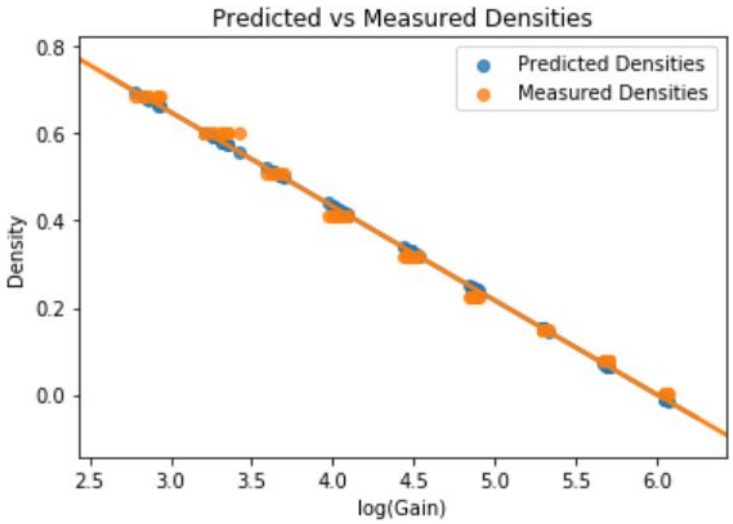



```
Slope: -0.216203201096
Intercept: 1.29801260526
R-Squared: 0.995818302485
```

Based on the linear model *predicted density = m*log(gain) + b*, where *m* is the slope and *b* is the intercept, we have a formula for predicting density from gain. Predictions based on these statistics for the first ten observations are as follows:

	0	1	2	3	4	5	6	7	8	9
gain	17.600000	17.300000	16.900000	16.200000	17.100000	18.500000	18.700000	17.400000	18.600000	16.800000
log(gain)	2.867899	2.850707	2.827314	2.785011	2.839078	2.917771	2.928524	2.856470	2.923162	2.821379
density	0.686000	0.686000	0.686000	0.686000	0.686000	0.686000	0.686000	0.686000	0.686000	0.686000
predicted density	0.677964	0.681681	0.686738	0.695884	0.684195	0.667181	0.664856	0.680435	0.666016	0.688021

The predicted densities appear quite close to the observed densities. This is good, but makes sense as the predictions were based on the observed data. Extrapolations of data outside the range of observed gains will likely be less accurate. The same scatterplot as above with predicted densities overlaid illustrates this relationship:



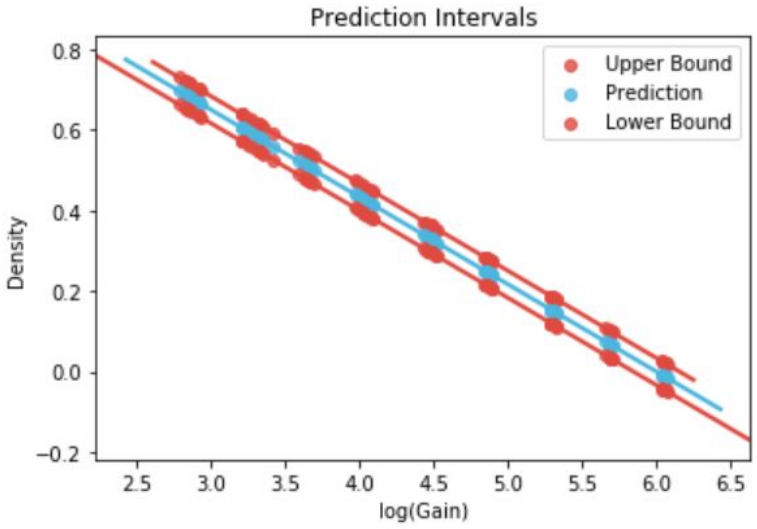
Naturally, these are predictions are therefore come with a designated degree of uncertainty. Evidently, the predictions even within the observed range of gain values are not exact. This variability can be defined by a prediction interval, which is a type of confidence interval generated to assess predictive uncertainty. The formula for this prediction interval is as follows[1]:

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

We can use this formula to generate upper and lower bounds for a prediction interval with 95% confidence with a t-score of around 2.28 and n of 90, with x_i serving as the log-transformed gains and $\hat{y}$ as the predicted densities. The first ten observations are shown below with predictions for density along with the upper and lower bounds of the prediction interval (PI) for their predicted densities:

	0	1	2	3	4	5	6	7	8	9
gain	17.600000	17.300000	16.900000	16.200000	17.100000	18.500000	18.700000	17.400000	18.600000	16.800000
log(gain)	2.867899	2.850707	2.827314	2.785011	2.839078	2.917771	2.928524	2.856470	2.923162	2.821379
density	0.686000	0.686000	0.686000	0.686000	0.686000	0.686000	0.686000	0.686000	0.686000	0.686000
predicted density	0.677964	0.681681	0.686738	0.695884	0.684195	0.667181	0.664856	0.680435	0.666016	0.688021
PI lower bound	0.644170	0.647878	0.652923	0.662045	0.650386	0.633414	0.631095	0.646635	0.632251	0.654203
PI upper bound	0.711757	0.715484	0.720554	0.729724	0.718004	0.700948	0.698618	0.714234	0.699780	0.721840

These predictive bounds can then be displayed over the prediction plot:



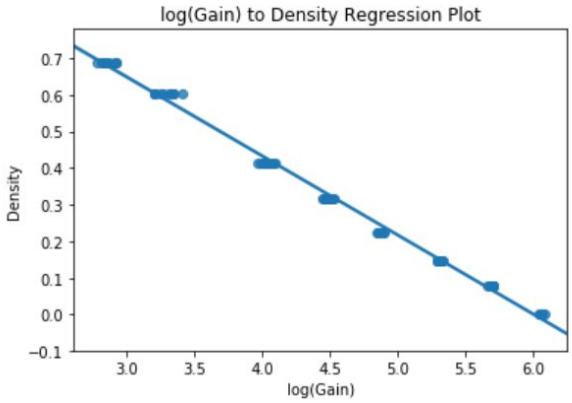
Statistics for the widths of these intervals (upper bound - lower bound) are displayed below:

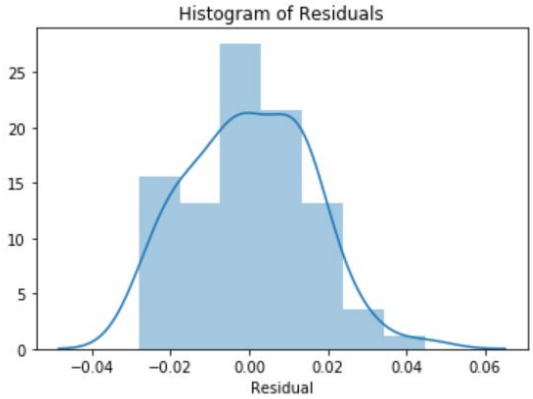
mean	std	min	25%	50%	75%	max
0.0670829	0.000315542	0.066718	0.0667779	0.0669596	0.067251	0.0676786

For example, for a gain reading of 38.6, the predicted density is 0.508, with prediction interval bounds defined by (0.478, 0.542). For a gain reading of 426.7, the predicted density is -0.011, with prediction interval bounds defined by (-0.045, 0.022). These gain readings are the average gains for densities of 0.508 and 0.001, respectively, which are within the prediction bounds.

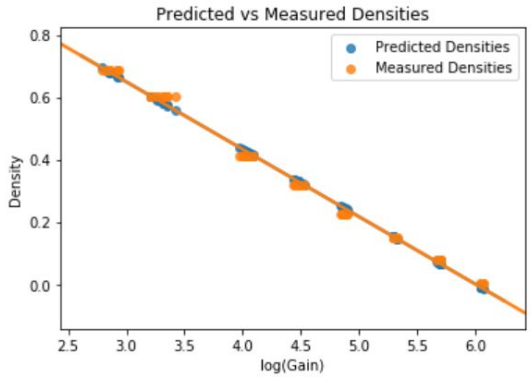
Cross-Validating the Predictions

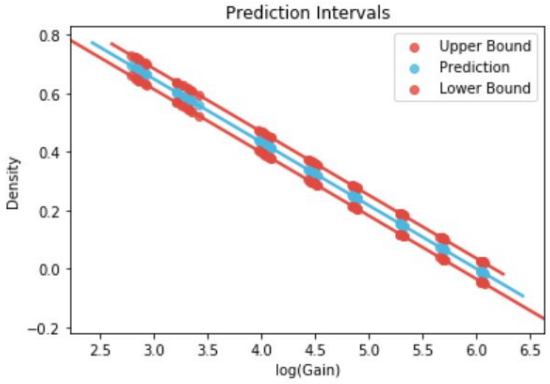
To validate the accuracy of these predictions, we can omit certain blocks of data and see how well our model works to predict the omitted data. If the prediction looks close to what we had, we can say our model is reasonably accurate in that range of gain values. To do this, we first omitted observations with densities of 0.508. Then, repeating the same procedure on this limited dataset yielded the following results:






```
Slope: -0.216278086793
Intercept: 1.29842212289
R-Squared: 0.995559499306
```



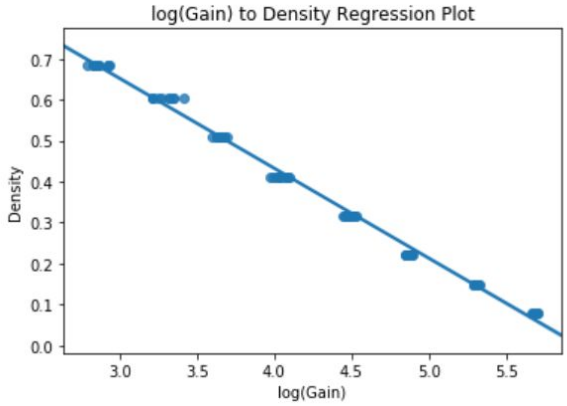


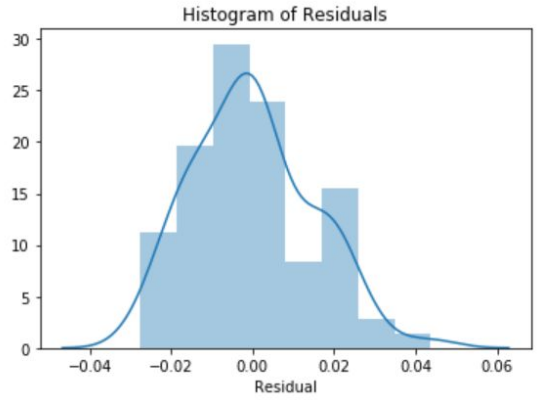
The resulting graphs and statistics are very similar to those generated from the full dataset. The predictions are as follows:

```
Predicted Density for Gain = 38.6: 0.508303709957
Prediction Interval for Gain = 38.6: [0.47309140667979332, 0.5435160132336907]
```

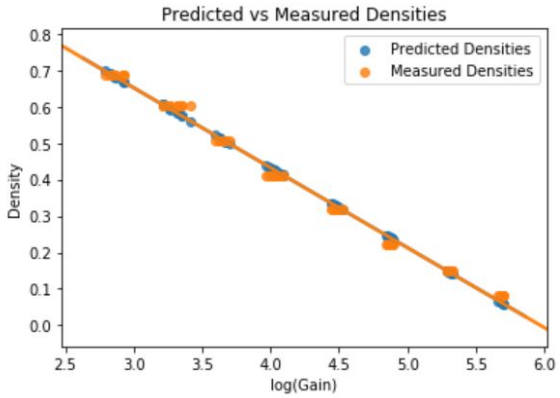
These predictions are quite accurate and thus suggest that our model is a good fit for gain readings around this range.

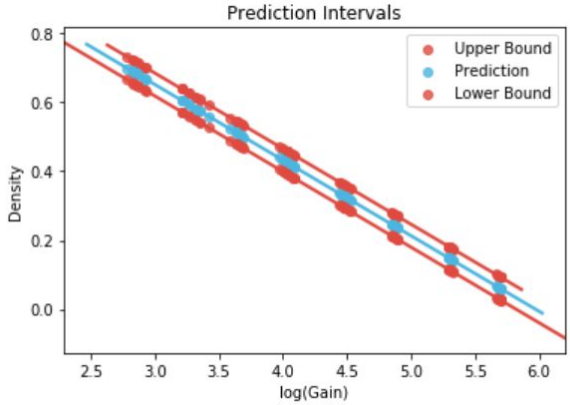
To further validate our model, we can repeat the same steps, this time omitting gain readings with corresponding densities of 0.001. This process yields the following results:






```
Slope: -0.219394805788
Intercept: 1.31011395717
R-Squared: 0.99501508433
```





Again, the results are similar to those from the full dataset. The point estimate is less accurate than that for 38.6 as it is a further extrapolation from the training data. The predictions for a gain reading of 426.7, the average of the 0.001 density blocks, is as follows:

```
Predicted Density for Gain = 426.7: -0.0185587993895
Prediction Interval for Gain = 426.7: [-0.052354129267692624, 0.015236530488656648]
```

[Additional Hypothesis]

For the additional hypothesis, our group decided to take a look at the impact that climate change has had over the last thirty years on snowpack density. Furthermore, from reading a study written by Bohumil M. Svoma, we would like to take a look at the impact that the following features have had on the changing weather conditions over the past thirty years in the western United States: mean air temperature for days without snowfall, the fraction of precipitation falling as snow, the total precipitation, and the mean snowfall density.

We will now give a brief background of the study. An understanding of spring snowpack density is of considerable importance for wet avalanche forecasting, climate modeling, and snow cover modeling. Wet avalanches occur due to the warming and melting of snowpacks with high water content (and therefore density) leading to less cohesion. For many modeling and snow cover reconstruction studies, snow density is essential for determining snow depth, snow water equivalent, and surface albedo (the proportion of the incident light or radiation that is reflected by a surface, typically that of a planet or moon). The purpose of this study was to establish the dominant densification mechanisms in the western United States through the analysis of data sets that largely contain more than 30 years of records

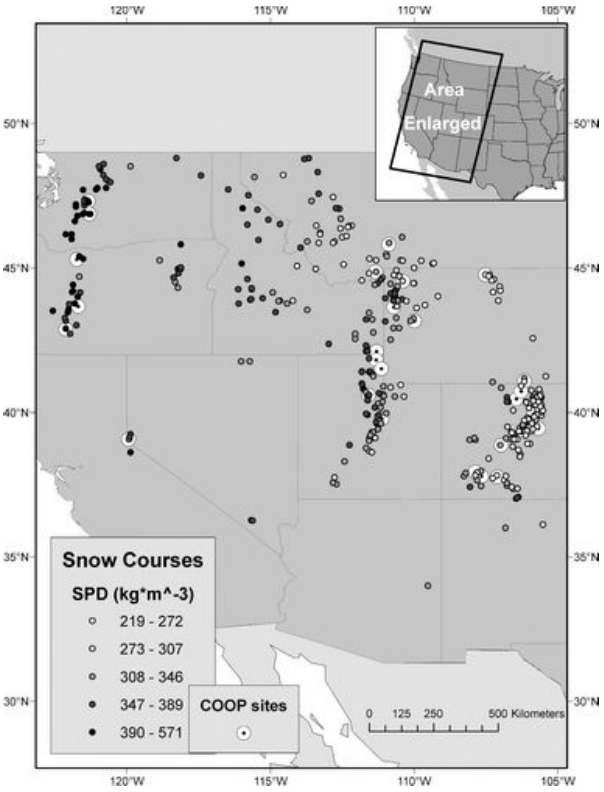


Figure 1: Snow courses (shaded dots indicating average April 1 SPD in kg m−3) and COOP sites (white circle with center dot) with 15 or more years of quality-controlled data. All COOP sites shown here are within 50 m of elevation and 0.5° latitude and longitude of a snow course. All sites depicted were used in the composite analysis described in the methods section. [4]

In the study it is shown that areal composite regression analysis for the western United States indicates a highly significant ($p = 0.005$) positive relationship between winter precipitation total and spring snowpack density. Svoma states, "this relationship weakens in lower elevation regions and coastal regions where warmer winter temperatures are conducive to more frequent rain events and melt events which affect snowpack density and ablate snow cover" (Svoma[4]). The study concludes that the significant positive relationship between precipitation and density is likely due to increased densification rates through gravitational compaction from the presence of greater snow water equivalent resulting from more snowfall.

For each snow course, snowpack density (*SPD*) was calculated for each year. *SPD* is inversely related to the ratio of snow depth to snow water equivalent and is defined by the following equation:

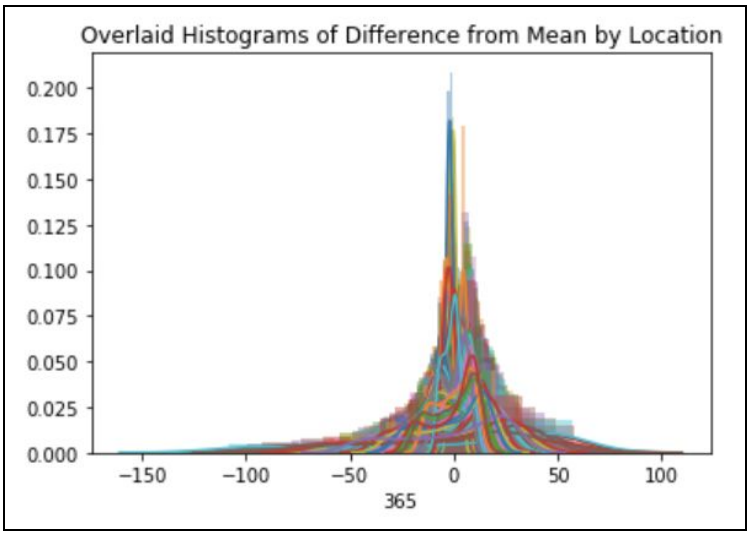
$$\rho_s = \rho_w \frac{SWE}{SD}$$

Here, ρ_s is the density of snow in kg m^{-3}, ρ_w is the density of liquid water. *SWE* and *SD* are in the same units of depth.

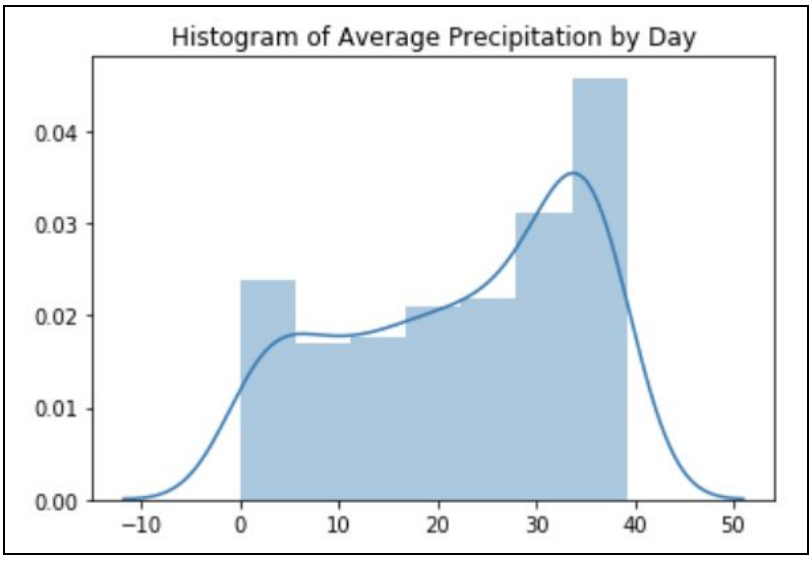
Below is the head of the cumulative rainfall for each unique COOP site over the course of a year:

	name	state	snotel_id	acton_id	shef_id	element	flag	1001	1002	1003	...	0922	0923	0924	0925	0926	0927	0928	0929	0930
0	Adin Mtn	CA	301	20H13S	ADMC1	PREC	U	0.0	0.0	0.0	...	29.1	29.2	29.2	29.2	29.3	29.3	29.3	29.4	29.4
1	Agua Canyon	UT	907	12M26S	AGUU1	PREC	C	0.0	0.0	0.2	...	22.9	23.0	23.1	23.2	23.2	23.3	23.3	23.4	23.4
2	Albro Lake	MT	916	11D28S	ABRM8	PREC	C	0.0	0.1	0.1	...	38.0	38.1	38.2	38.2	38.2	38.3	38.3	38.3	38.5
3	Alpine Meadows	WA	908	21B48S	APSW1	PREC	C	0.0	0.3	0.9	...	155.7	155.7	155.8	155.9	156.2	156.5	156.6	156.8	157.3
4	Anchor River Divide	AK	1062	51K05S	ARDA2	PREC	C	0.0	0.1	0.2	...	29.2	29.4	29.6	29.7	29.8	29.9	30.2	30.2	30.4

We can then overlay the differences from the mean rainfall by location in order to get a sense of the variance in precipitation.



Lastly, we can visualize the average rainfall by in order to see that the distribution is bimodal with a large amount of rainfall coming in the latter portions of each month.



The positive relationship between the *SPD* in the spring and winter snowfall totals suggest that warm spring seasons following wet winters are prime for wet avalanches of high *Snowpack Density*. Svoma concludes the study by stating, “increasing spring snowpack density with winter precipitation is likely due to higher densification rates throughout winter due to increased gravitational compaction from more snow water equivalent.

With more rigorous datasets, we would be able to utilize the latitude and longitude of the COOP sites in order to run a more thorough model on how *SPD and SWE* vary based on location. With a more comprehensive list of variables we could test to see if there is a multiple linear model to predict *SPD* and better estimate the severity of wet avalanches in the coming spring seasons.

[Theory]

Regression

In statistical modeling, regression analysis is the set of processes for estimating the relationships among variables. Among a variety of techniques we can use to analyze a series of variables, our main focus is on the relationship between the dependent variable and one or more independent variables.

Regression models involve the following parameters and variables:

1. The *unknown parameters*, denoted as β (may be represented as a vector or scalar).
2. The *independent* variable(s), X.
3. The *dependent* variable(s), Y.

A regression model relates Y to a function of X and β such that $Y \approx f(X, \beta)$. This function helps us understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed. The approximation is formalized as $E(Y|X) = f(X, \beta)$. The form of this function is sometimes 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, y_i , is the linear combination of the parameters. In simple linear regression for modeling n data points there is one independent variable (x_i) and two parameters β_0 and β_1:

$$y_i = \beta_0 + x_i \cdot \beta_1 + \varepsilon_i$$

Here, ε_i is an error term (residual) that is found by finding the difference between the value of the dependent variable predicted by the model ($\widehat{y_i}$), and the true value of the dependent variable (y_i). Given a random sample from the population, we estimate the population parameters and obtain the sample linear regression model:

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} \cdot x_i$$

There are conditions that must be met in order to construct a linear model.

1. Nearly Normal Residuals: the distribution of the residuals should be nearly normal.
2. Linearity: the relationship between the explanatory and response variables should be linear.
3. Constant Variability: the errors must have constant variance, with the residuals scattered randomly around zero -- in order to check these assumptions, we can use a residuals versus fitted values plot.

Cross-Validation

We use cross-validation in order to assess how well the results from our predictive model will generalize to an independent data set. In this case study our goal is prediction, so we want to estimate how accurately our predictive model will perform in practice. We do so by dividing our original sample into a *training set* and a *testing set*. The known data in our training set is what our model is built on. Then, a dataset of unknown data (our *testing set*) is tested against the model. The goal of cross-validation is to test the model's ability to predict new data that was not used in estimating it. This helps us identify issues with our model like overfitting and selection bias all while giving us insight on how the model will generalize to an independent dataset.

In our linear regression model, we would first leave out a set of data points to use for validation, then perform our model on the remaining data to provide an estimation interval for the data points we just left out. Then, we check where the true data point falls in the aforementioned interval to assess the accuracy of our model. We have a set of real responses of size n ($y_1, y_2, y_3, \ldots, y_n$) that correspond to a set of predictor variables ($x_1, x_2, x_3, \ldots, x_n$). Therefore, if we have some function $y_{predicted} = \alpha + \beta \cdot x$ used to fit our data, we can assess its accuracy by finding the mean squared error:

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - y_{predicted})^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - (\alpha + \beta \cdot x))^2$$

The process of fitting optimizes the model parameters in order to make the model fit the training data as well as possible. If we then take an independent sample of validation data from the same population as the training data, it will generally turn out that the model does not fit the validation data as well as it fits the training data. This difference is likely to be large when the

training data set is small, or when the number of parameters in the model is large - we use cross-validation is a way to estimate the size of this effect.

Using Hypothesis Test for Linear Regression

We can use t-tests to conduct hypothesis tests on the regression coefficients that we obtain from our simple linear regression. We use a statistic based on the t distribution in order to test the two-sided hypothesis that the true slope, β_1, equals some constant value that we assign to $\beta_{1,0}$. We can express our null and alternative hypotheses as:

$$H_0 : \beta_1 = \beta_{1,0}$$
$$H_A : \beta_1 \neq \beta_{1,0}$$

The test statistic that we used for this test is shown below. Here, we consider $\widehat{\beta}_1$ to be the least squares estimate of β_1 and $SE(\widehat{\beta}_1)$ its standard error.

$$T = \frac{\widehat{\beta}_1 - \beta_{1,0}}{SE(\widehat{\beta}_1)}$$

We generate the t value with *n - 2* degrees of freedom (*df = n-2*). The null hypothesis, H_0, is accepted if the calculated value of the test statistic is such that

$$-t_{\frac{\alpha}{2}, n-2} < T < t_{\frac{\alpha}{2}, n-2}$$

where $-t_{\frac{\alpha}{2}, n-2}$ and $t_{\frac{\alpha}{2}, n-2}$ are the critical values of the two-sided hypothesis, $t_{\frac{\alpha}{2}, n-2}$ is the percentile of the t distribution, and α is the significance level.

Coefficient of Determination

We can use the coefficient of determination to evaluate how strong of a fit our linear model is. The coefficient of determination is often denoted as R^2 or the square of the correlation coefficient. This coefficient tells us the proportion of the variance in the dependent variable that is predictable from the independent variable(s) (or from the model). The variability that is left unaccounted for can be explained by either variables that were excluded from the model or by randomness (or *noise*) in the set of data.

In all cases where R^2 is used, the predictors are calculated by least-squares regression. This is done by minimizing $\sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2$ (otherwise known as the explained sum of squares). In this case R^2 increases as we increase the number of variables in the model. This is considered to

be one of the drawbacks of using R^2 where we could keep adding variables in order to increase the value of R^2 .

There is also the case where we will want to utilize $R^2_{Adjusted}$. $R^2_{Adjusted}$ compares the explanatory power of regression models that contain **different numbers of predictors.** The $R^2_{Adjusted}$ is a modified version of R^2 that has been adjusted for the number of predictors in the model. The $R^2_{Adjusted}$ increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance. $R^2_{Adjusted}$ can be negative and it is always lower than the R^2 . We can find $R^2_{Adjusted}$ using:

$$R^2_{Adjusted} = 1 - (\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1})$$

where *n* is the number of cases and *p* is the number of predictors (explanatory variables) in the model. $R^2_{Adjusted}$ applies a penalty for the number of predictors included in the model, so we choose models with higher $R^2_{Adjusted}$ over others.

Errors & Residuals

We denote errors as ε_i (for each point in the data set), and the estimated residual ($\widehat{\varepsilon_i}$) from the least squares line is equivalent to $y_i - (\widehat{\alpha} + \widehat{\beta} \cdot x_i)$. Also, this estimated residual is unbiased such that $E[\widehat{\varepsilon_i}] = \varepsilon_i$. We know that residuals are used as estimates of model error, so we can use $\widehat{\varepsilon_i}$ to check whether or not the residuals are normally distributed: $\widehat{\varepsilon_i} \sim N(0, \sigma^2)$. However, this gives us pause because the variance of the standardized residual is equal to σ^2 . This means that we should look to the standardized residual to do *model checking*. We can write the response as:

$$y_i = \widehat{y_i} + \varepsilon_i \text{ where } \varepsilon_i = y_i - \widehat{y_i} \text{ and } \widehat{y_i} = \beta_0 + \beta_1 \cdot x_i$$

The variances of the respective responses and fitted values are then:

$$Var(y_i) = \sigma^2$$
$$Var(\widehat{y_i}) = h_{ii} \cdot \sigma^2$$

where h_{ii} is going to be the leverage of the i^{th} observation and it can be proved that:

$$Var(\varepsilon_i) = (1 - h_{ii}) \bullet \sigma^2$$

So, the higher the leverage, the lower the residual variance is going to be.

Multiple Observations

Multiple linear regression attempts to model the relationship between two or more explanatory variables and the response variable by fitting a linear equation to observed data. Every value of the independent variable x_i is associated with a value of the dependent variable y_i. The population regression line for p explanatory variables $x_1, x_2, \ldots, x_p$ is defined to be:

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

This line depicts how the mean response (μ_y) changes with the explanatory variables ($x_1, x_2, \ldots, x_p$). The fitted values $b_0, b_1, b_3, \ldots, b_p$ estimate the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$. The line shown above describes how the mean response μ_y changes with the given explanatory variables. These observed values of y are assumed to have a standard deviation of σ . The model we construct will have some deviation so, give n observations, we can formally define the model for multiple linear regression as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_p x_{ip} + \varepsilon_i \text{ for } i = 1, 2, 3, \ldots, n$$

The values that are fit by the equation $b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_p x_{ip}$, as we saw before, are denoted as $\widehat{y_i}$. The variance, σ^2 , can be estimated using $s^2 = \frac{\sum_{i=1}^{n} \varepsilon_i^2}{n-p-1}$. The estimate of the standard error s is the square root of the mean squared error.

Confidence Intervals for Simple Linear Regression

In the case of simple linear regression, we are trying to construct a confidence interval for *E[Y|X]*. We can build a confidence interval for *E[Y|X]* (the *expected value* for *Y* given *X*) using:

$$\widehat{Y} \pm t_{n-2} \cdot s_y \cdot \sqrt{\frac{1}{n} + \frac{(X - \overline{X})^2}{(n-1) \cdot s_x^2}}$$

where s_y is the standard deviation of the residuals. If we want to build a prediction interval, we would have to change our formula above in such a way so as to increase the variability. A prediction is used to give an estimation for a random variable (in this case giving a range for y itself). We guess the expected value, *E[Y|X]*, more precisely than we estimate y because the

estimation of *y* requires the inclusion of the variance that comes from $Y = \alpha + \beta \cdot x + \varepsilon$. The prediction interval for *y* given *x* can be found using:

$$\widehat{Y} \pm t_{n-2} \cdot s_y \cdot \sqrt{1 + \frac{1}{n} + \frac{(X - \overline{X})^2}{(n-1) \cdot s_x^2}}$$

A prediction interval is similar to a confidence interval, except that the prediction interval is designed to cover the random future value of *y*, while the confidence interval is designed to cover the average value of *y*, *E[Y]*, for a given *X*. Although both are centered at $\widehat{y}$, the prediction interval is wider than the confidence interval, for a given *X* and confidence level. This makes sense, since the prediction interval must take account of the tendency of *y* to fluctuate from its mean value, while the confidence interval needs to account for the uncertainty in estimating the mean value.

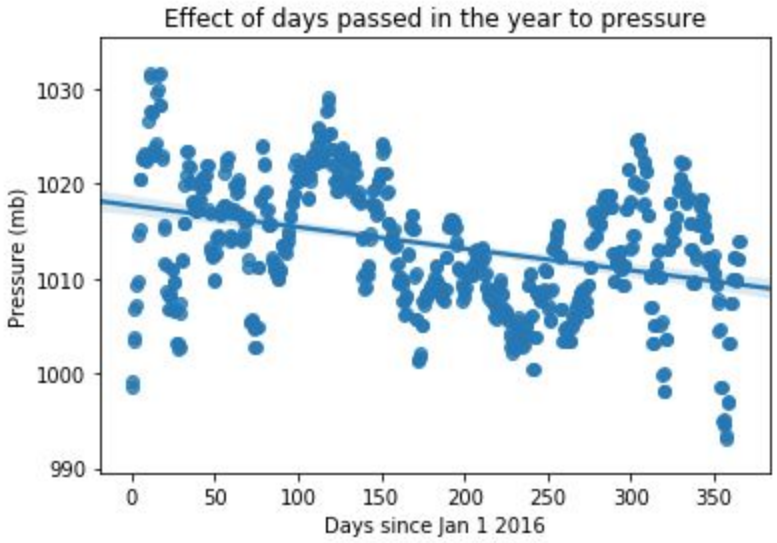
The error in estimating *E[Y]* and $\widehat{y}$ grows as *X* moves away from $\overline{X}$. Therefore, the further *X* is from $\overline{X}$, the wider the confidence and prediction intervals will be. This explains why confidence intervals for the predicted values in a linear regression model tend to be narrow in the middle and wider at the extreme values.

Advanced Analysis

For our advanced analysis, we used data from the National Snow and Ice Data Center. We used the 2016 data for the pressure and temperature fields as described in the buoy reports. On viewing the data, we can see that the data for temperature does not seem accurate. Every record in the table has a temperature of -999.9°C square. It is not possible for us to have a number squared be negative and hence this data is incorrect. Since the data for temperature is not consistent, we will drop the columns related to temperature.
From the 'Predicting snow density using meteorological data' study (rmets), we could see that there was a negative relationship between snow density and the sum of atmospheric pressure with an R-square value of 0.53. We will further investigate how the date and the geographical location affect the atmospheric pressure. We will model our predictions using a Regression model.

We will start by setting a Null Hypothesis stating that no correlation that exists between the atmospheric pressure and the time of year. In turn we can set our alternative hypothesis to be that there is a correlation between the variables. We will accept our alternative hypothesis if our p score is less than 0.05.

For simplicity of calculation we will convert our date time object to the number of days passed in the year since Jan 1 2016, since that is the granularity of our data is every 12 hours, this makes our conversion valid and our data easy to understand.
After plotting the regression line, we got the graph below:



OLS Regression Results

Dep. Variable:	y	R-squared:	0.747
Model:	OLS	Adj. R-squared:	0.747
Method:	Least Squares	F-statistic:	2163.
Date:	Sun, 10 Mar 2019	Prob (F-statistic):	1.34e-220
Time:	00:19:09	Log-Likelihood:	-5601.4
No. Observations:	732	AIC:	1.120e+04
Df Residuals:	731	BIC:	1.121e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	4.1510	0.089	46.506	0.000	3.976	4.326

Omnibus:	533.547	Durbin-Watson:	0.000
Prob(Omnibus):	0.000	Jarque-Bera (JB):	44.400
Skew:	0.003	Prob(JB):	2.28e-10
Kurtosis:	1.793	Cond. No.	1.00

Our predictor variable is the days in the year passed since Jan 1, while our response variable is the atmospheric pressure (mb). From the descriptive statistics, we can see that we have a p value

of 0, hence we will be rejecting the null hypothesis and we will now believe the alternate hypothesis. Now that we know there is a linear relationship between the variables, we can model an equation for the regression line, where the number of Days passed since Jan 1 2016 is X (the explanatory variable) and Pressure is Y (the response variable):

Pressure = 1017.719 - 0.023 * Days passed since Jan 1 2016
Our gradient of 0.023 indicates that for each day after new year our pressure decreases by 0.023mb.

Next, we will conduct a similar hypothesis to check if the longitude and latitude have a linear relationship with the atmospheric pressure. Again we can see that the p value is 0, leading us to believe that there is a linear relationship.
If we were to draw an equation for the linear model, it would be:
Pressure = 1018.811 + 0.006 * Longitude - 0.079 * Latitude

OLS Regression Results

Dep. Variable:	Pressure(mb)	R-squared:	0.995
Model:	OLS	Adj. R-squared:	0.995
Method:	Least Squares	F-statistic:	2.441e+07
Date:	Sun, 10 Mar 2019	Prob (F-statistic):	0.00
Time:	00:54:02	Log-Likelihood:	-1.5137e+06
No. Observations:	264252	AIC:	3.027e+06
Df Residuals:	264250	BIC:	3.027e+06
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Lon	0.0973	0.001	70.607	0.000	0.095	0.100
Lat	12.5428	0.004	3550.003	0.000	12.536	12.550

Omnibus:	130038.198	Durbin-Watson:	0.033
Prob(Omnibus):	0.000	Jarque-Bera (JB):	13964.535
Skew:	-0.012	Prob(JB):	0.00
Kurtosis:	1.874	Cond. No.	5.54

It is very interesting to see that the atmospheric pressure has a positive correlation with the longitude, but a negative correlation with the latitude. The negative correlation with the latitude is expected, since a place away from the equator receives less sunlight (lower temperature) and we also know that lower temperatures lead to lower atmospheric pressure. Our data is however slightly skewed, since the range of latitudes in our dataset is just 20°, while the range of our longitudes is just 90°.

Having an R-squared value of 0.995 suggests that a lot of variability in our data is explained by this regression model.

On combining our two models, We will expect to see a positive correlation between the pressure and the longitude, but a negative correlation between the latitude and the days passed since the first day of the year. In the end the equation for our linear regression model would be:

Pressure = 1022.984 + 0.006 *Longitude - 0.079 * Latitude - 0.023 * Days passed since Jan 1 2016
For every degree our longitude increases, the pressure increases by 0.006mb. Similarly for every increase in one degree on the latitude our pressure decreases by 0.079mb. From above we saw that for every day after Jan 1st, our pressure decreases by 0.023mb.

OLS Regression Results

Dep. Variable:	Pressure(mb)	R-squared:	0.995
Model:	OLS	Adj. R-squared:	0.995
Method:	Least Squares	F-statistic:	1.640e+07
Date:	Sun, 10 Mar 2019	Prob (F-statistic):	0.00
Time:	00:22:49	Log-Likelihood:	-1.5126e+06
No. Observations:	264252	AIC:	3.025e+06
Df Residuals:	264249	BIC:	3.025e+06
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Lon	0.0962	0.001	70.133	0.000	0.094	0.099
Lat	12.4002	0.005	2657.585	0.000	12.391	12.409
hours	0.0630	0.001	46.552	0.000	0.060	0.066

Omnibus:	104666.595	Durbin-Watson:	0.032
Prob(Omnibus):	0.000	Jarque-Bera (JB):	13175.466
Skew:	-0.019	Prob(JB):	0.00
Kurtosis:	1.907	Cond. No.	9.50

We were able to achieve a similar R-squared in the last model. But solely due to the simplicity of the model, we will prefer to use that model.
The accuracy of our model would potentially increase if we had accurate information about the temperatures and the amount of sunlight absorbed by the snow, as that affects the atmospheric pressure.

Data Limitations

In terms of data limitations, there is a decline of the operational networks in the northern regions of the world, such as Alaska and northern Canada. In addition, there are few stations in the mountain regions, and there is some concern on how to sustain and improve the operational networks. In addition, there are data limitations in terms of working with this dataset internationally. Specifically, the difference in instruments and methods of data processing in different nations leads to the incompatibility of precipitation data. In addition, different geographic regions, such as the arctic regions, provide difficulties to determine precipitation changes. Furthermore, there are large biases in the gauge measurements of solid precipitation. From these challenges, we need some form of validation of the precipitation data, such as from satellite data.

Conclusion

From Scenario 1: Fitting the Data, we see that the original data given in the dataset does not follow using a linear regression model since the original data is not linear. To make the data linear, we performed a logarithmic transformation of gain to normalize the data and see relative changes in gain, as opposed to absolute changes in gain given in the original data. Because of our logarithmic transformation of the gain data, we see a linear relationship between log(gain) and density, as well as nearly normal residuals. In addition, we see that the logarithmic transformation does not result in homoscedasticity, or constant variability around the regression line. This means that using a linear model to predict density from log(gain) is most likely not the best option.

From Scenario 2: Predicting Density, we use the linear model of log(gain) to density to predict density. As we see from the first ten observations from the datasets, we predicted densities really close to the observed densities. Since predictions come with a degree of uncertainty, we use the equation for the prediction interval to calculate a prediction interval with 95% confidence, using the logarithmic transformed gain data and the predicted densities.

From Scenario 3: Cross-Validating the Predictions, we omit certain blocks, or parts, of data to validate the accuracy of our predictions. Specifically, we omitted observed data with density values of 0.508, which we predicted in the previous scenario for a gain reading of 38.6. We see

that our predictions are really accurate, which suggests that our model is a good fit for gain readings around this range. Afterwards, we found similar results when omitting data with density values of 0.001.

From our advanced analysis, we performed data analysis to determine if there was a correlation between the atmospheric pressure and the time of year. From our initial analysis, we found that there is a linear relationship between these variables since we computed a p-value of 0, thereby rejecting our null hypothesis that there is no correlation between these variables. Similarly, we find that there is a linear relationship between longitude and latitude since we also computed a p-value of 0. Interestingly, we find that there is a negative correlation between atmospheric pressure and latitude but a positive correlation between atmospheric pressure and longitude.

From all of our analysis, we recognize that the original data in the given data sets do not follow a linear model, but by performing a logarithmic transformation on the gain data, we find a reasonable linear relationship between log(gain) and density. Additionally, we find that other variables, such as atmospheric pressure, time of year, latitude, and longitude, should also be considered when collecting this type of data.

Works Cited

[1]https://newonlinecourses.science.psu.edu/stat501/node/274/
[2]https://journals.ametsoc.org/doi/full/10.1175/BAMS-D-11-00052.1
[3]https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/met.40
[4]https://www.tandfonline.com/doi/full/10.1657/1938-4246-43.1.118