Christine Hong, A12763457, Math 189
Wesley Partridge, A12791650, Math 189
Matthias Smyrl, A13014876, Math 189
Tong Yoo, A12611985, Math 189
Cristian De Leon, A10803160, Math 189

Case Study 4
https://github.com/msmyrl/MATH189_CS4

Introduction

The main source of water for Northern California comes from the Sierra Nevada mountains where it runs into the Central Valley watersheds. In order to help monitor the water supply, the Forest Service of the United States Department of Agriculture (USDA) utilizes a gamma transmission snow gauge in the Central Sierra Nevada near Soda Springs, California. The gamma transmission snow gauge is used to determine a depth profile of snow density. Analyzing the depth profile of snow density is crucial as excess water that rain falls onto snow and melts goes into the watershed reservoirs to be used.

The technology used does not disturb the snow in the measurement process; therefore, the same snowpack can be measured repeatedly. Researchers are able to study the same snowpack settlement over and over again throughout the winter and the dynamic of rain on the snow with these replicate measurements. When analyzing the dynamic of rain on snow, the rain falls on the snow, and the snow absorbs as much of the water as possible before flooding occurs. Hence, the denser the snowpack the less water is being absorbed. Analysis of the depth profile of the snowpack can help with monitoring the water supply and flood management into the watersheds.

The gamma transmission snow gauge does not directly measure the snow density; instead, gamma rays are emitted from the technology where the measurement is converted into density readings. Due to instrument wear and radioactive decay, there may be changes over the seasons in the functions used to cover the measured values into density readings. In order to counteract this problem and adjust the conversion method, a calibration run is made each year at the beginning of the winter season to ensure that measurements are as accurate as possible. Developing a procedure to calibrate the snow gauge will help track the water supply and flood management.

Data

The data for the study are from a calibration run of the USDA Forest Service's snow gauge located in the Central Sierra Nevada mountain range near Soda Springs, California. The calibration run consists of placing polyethylene blocks, simulating snow, of known densities between the two poles of the snow gauge and taking measurement readings on the blocks. For each block, 30 measurements are taken, which are discrete numbers. However, only the middle

10 measurements are reported for the case study. The data available consists of 10 measurements for each of the 9 densities in grams per cubic centimeter of polyethylene. The measurement reported are amplified version of the gamma photon count made by the detector in which will be called the "gain".

There are several challenges with the data such as the different data quality and compatibility across national borders. There are large biases in gauge measurements of solid precipitation as well as incompatibility of precipitation data due to difference in instruments and method of data processing. Additionally, there are difficulties in determining precipitation changes in the arctic regions. Hence, the data from this study may not be accurate or applied in other parts of the world.

Background

The gamma transmission snow gauge is a complex and expensive instrument; therefore, it is not practical to set up a broad network of snow gauges in the area in order to monitor the water supply. Alternatively, the gauges are used as a research tool to help study snowpack settling, snowmelt runoff, avalanches, and rain-on-snow dynamics. These snow gauges exist in Idaho, Oregon, Colorado, Alaska, and the Sierra Nevada, California. The data in this case study is from the Central Sierra snow gauge. This gauge is located in the center of a forest opening that is about 62 meters in diameter. The laboratory site is at 2099 meters elevation and is subject to all major high altitude storms which regularly deposit 5-20 centimeters of wet snow, and the snowpack reaches a depth of 4 centimeter each winter, on average.

The snow gauge mechanism consists of a a cesium-137 radioactive source and an energy detector mounted on separate vertical poles about 70 centimeters apart. The radioactive source emits gamma photons, gamma rays, at 662 kilo-electron-volts (keV) in all directions. The detector contains a scintillation crystal which counts the photons going through the 70-centimeter gap from the source to the detector crystal. The photons that reach the detector crystal generate a pulse which are transmitted by a cable to the recordings. The signal is then stabilized, adjusted for accuracy like temperature drifts, and converted into a measurement called "gain." The "gain" should be directly proportional to the emission rate. The density usually ranges between 0.1 and 0.6 gram per centimeter squared.

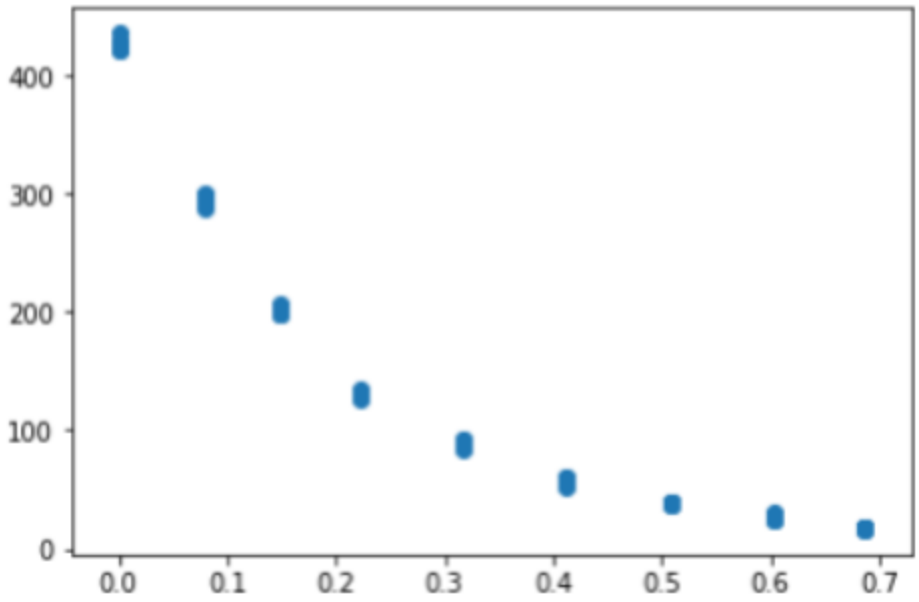
The gamma rays that are emitted from the radioactive source are sent out in all directions. The rays sent in the direction of the detector may be scattered or absorbed by the polyethylene molecules between the source and the detector. Fewer gamma rays will reach the detector with denser polyethylene.

There are complex physical models for the relationship between the polyethylene density and the detector readings; however a simplified version of the model that may be workable for the calibration problem is described in the following. A gamma ray on the way to the detector passes a number of polyethylene molecules, and the number of molecules depends on the density of the polyethylene. A molecule may either absorb the gamma photon, bounce it out of the path

to the detector, or allow it to pass. If each molecule acts independently, then the chance that a gamma ray successfully arrives at the detector is p^m, where p is the chance a single molecule will neither absorb nor bounce the gamma ray, and m is the number of molecules in a straight-line path from the source to the detector. This probability can be expressed as $e^{mlogp} = e^{bx}$, where x, the density, is proportional to m, the number of molecules.

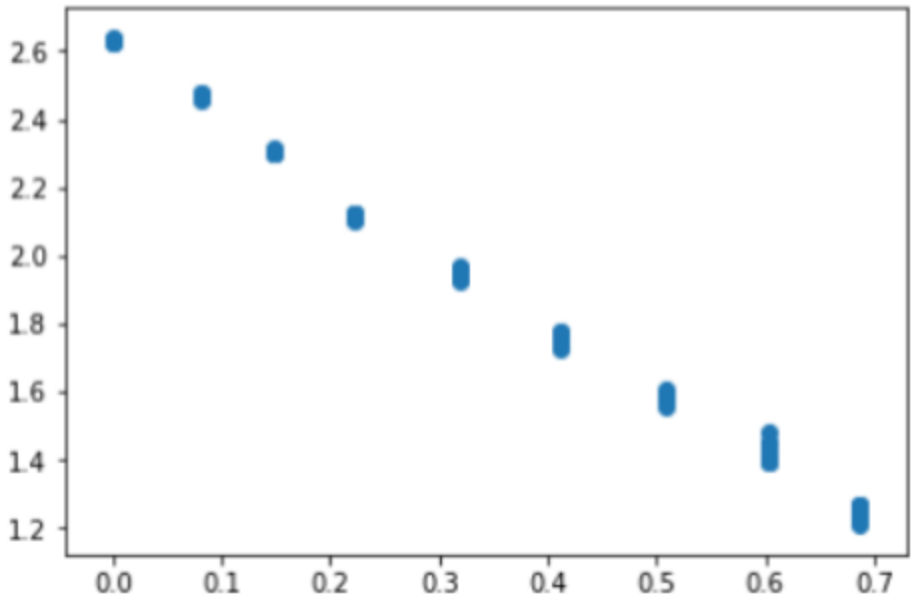
Investigation

Figure 1: Scatter Plot of Gauge Versus Density



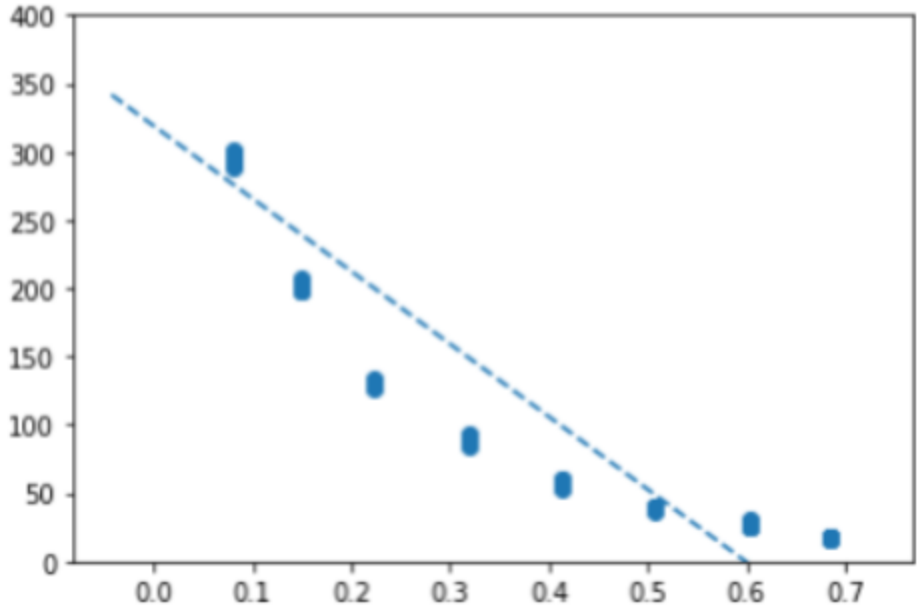
In Figure 1, it is observed that there is a clear relationship between the gauge and density of snow, although the relationship is of an exponential and not linear nature. To see a linear relationship, rather than considering the snow gauge, one may consider the log of the snow gauge measurements to see a linear relationship.

Figure 2: Log(Gauge) Versus Snow Density



In order to visualize a linear rather than exponential relationship between the gauge and snow density as in figure 1, the log of the gauge is taken and plotted against the snow density. Doing so, a linear relationship is observed between the log of gauge measurements and the snow density.

Figure 3: Least Squares Fit of Gauge Versus Snow Density



A least squares line is fitted to the measurements of the gauge versus the snow density; or in other words, the gauge measurements is linearly regressed on the snow density.

Figure 4: Table of OLS Regression of Gauge(Gain) on Density

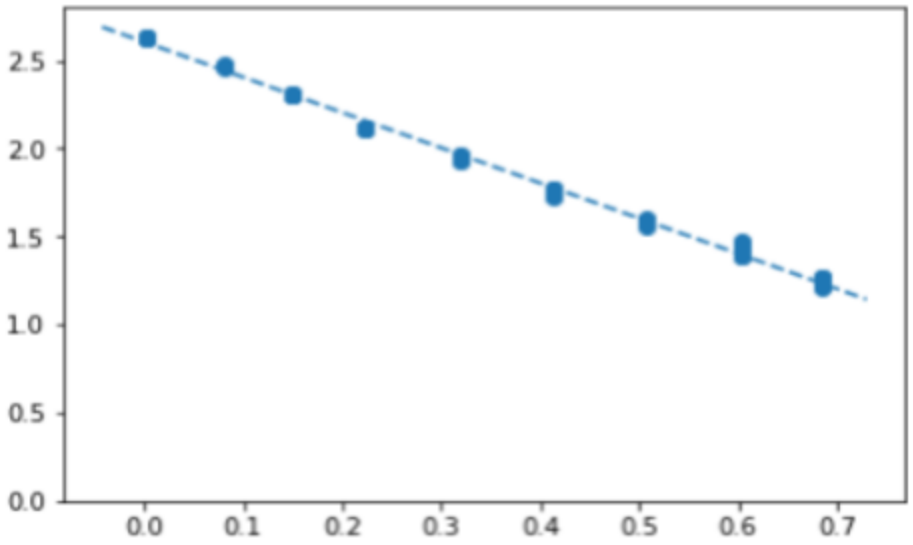
Source	SS	df	MS	Number of obs	=	90
				F(1, 88)	=	389.48
Model	1289407.73	1	1289407.73	Prob > F	=	0.0000
Residual	291334.995	88	3310.62494	R-squared	=	0.8157
				Adj R-squared	=	0.8136
Total	1580742.72	89	17761.1542	Root MSE	=	57.538

gain	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
density	-531.9507	26.9545	-19.74	0.000	-585.5171	-478.3843
_cons	318.7015	10.7907	29.53	0.000	297.2572	340.1457

Figure 4 shows the results of regressing the gauge (gain) on snow density, and since there is an exponential relationship between gain and density, the coefficients are expectedly high. The t value of the coefficient of density is statistically significant, as it is greater than 1.96, allowing

one to reject the null hypothesis that the density coefficient is statistically insignificant in explaining changes in the gain of the gauge. The p-value of 0 also reinforces this notion that the density coefficient is statistically significant as a causality of changes in gain. The R-squared value of 0.8157 supports the claim that the OLS line accurately fits the data.
To answer questions of prediction, the density of the snowpack may be estimated using the linear regression equation in Figure 4. If the gauge reads 426.7, by solving 426.7= 318.7015-531.9507x for x, x yields .49053466 as the snow density given a gain of 426.7. Given a reading of 38.6, the snow density may be acquired by solving for x in a similar equation 38.6= 318.7015-531.9507x to find that the corresponding snow density equals -.10457788.

Figure 5: Least Squares Fit of Log(Gauge) Versus Density



A least squares line is also fitted to the plot of log(gauge) versus density reading to linearize the exponential relationship between gain and snowpack density.

Figure 6: Table of OLS Regression of Log(Gain) on Density

Source	SS	df	MS	Number of obs	=	90
				F(1, 88)	=	20956.09
Model	18.2327453	1	18.2327453	Prob > F	=	0.0000
Residual	.076563989	88	.000870045	R-squared	=	0.9958
				Adj R-squared	=	0.9958
Total	18.3093093	89	.205722576	Root MSE	=	.0295

loggain	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
density	-2.000333	.0138181	-144.76	0.000	-2.027794	-1.972873
_cons	2.604579	.0055318	470.84	0.000	2.593586	2.615573

Figure 6 shows the numerical results of regressing log(gain) against the snowpack density, and since the exponential nature of the relationship has been neutralized by taking log(gain), as a result the R-squared value increases significantly to a 0.9958. Since the log(gain) is taken rather than the gain itself, the magnitude of the density coefficient falls significantly, though the variable itself remains statistically significant. To forecast with the log(gain) regression, one would need to take the inverse of the log value in order to get the true value.

Figure 7: Confidence Bands Around Least Squares Regression

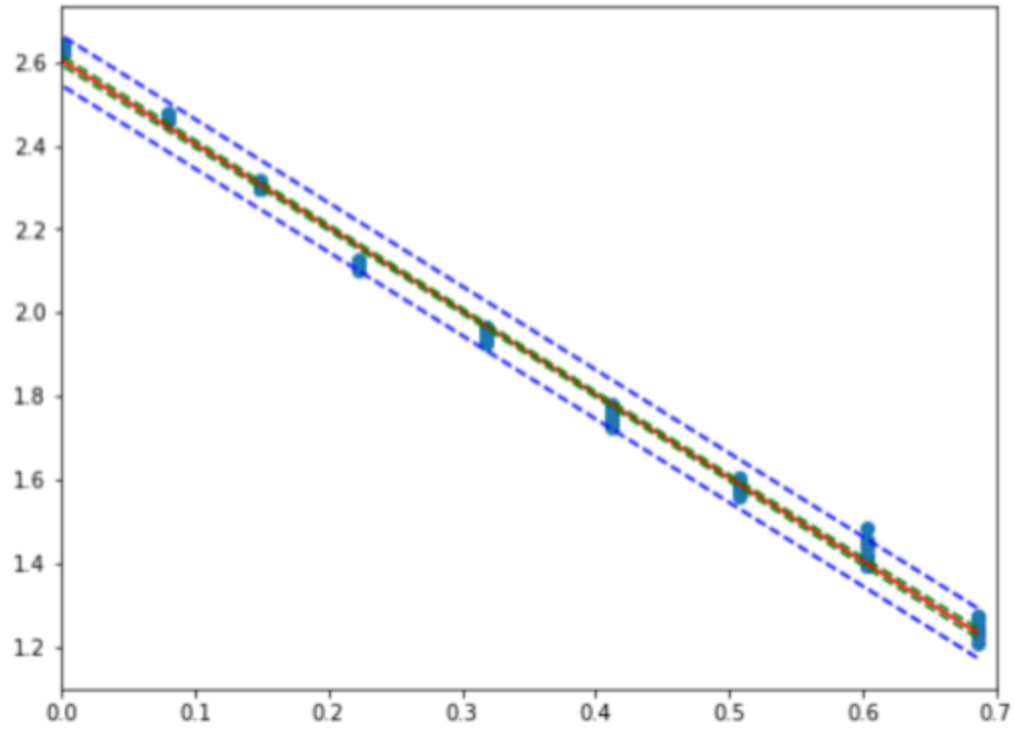


Figure 8: Table of Least Squares Regression with Density .508 Omitted

Source	SS	df	MS	Number of obs	=	80
				F(1, 78)	=	17487.59
Model	16.8038544	1	16.8038544	Prob > F	=	0.0000
Residual	.074950342	78	.000960902	R-squared	=	0.9956
				Adj R-squared	=	0.9955
Total	16.8788048	79	.213655757	Root MSE	=	.031

loggain	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
density	-1.999121	.0151173	-132.24	0.000	-2.029217	-1.969025
_cons	2.604525	.0058165	447.78	0.000	2.592945	2.616105

In Figure 7, confidence bands are constructed around the least squares regression line, for the purposes of building a confidence interval around every data point. The y axis represents the

log(gain) and the x axis represents the snowpack density. To cross validate this model all data points with a density of .508 were omitted. Using the average gain of 38.6 a confidence interval for the density given by the reading of 38.6 gain was constructed. First a point estimate of density pertaining to the gain of 38.6 was observed, and from there the regression equation given by the table in Figure 8 was used to yield the confidence interval [0.471 and 0.507].

Theory

Linear regression is a statistical method that involves modeling a linear relationship between a response variable, and one or more explanatory variables. This tool helps identify a relationship between variables. This relationship is quantified in the form of a correlation, where correlation is a value between -1 and 1. Linear regression is a powerful tool therefore in order to use it properly one should first ensure that there is good reason to believe a relationship between two variables exists. It is possible to check for a possible relationship by plotting a scatterplot. The scatterplot should be a clear indication of whether there is a dependency to be measured. Once a dependence among the variables is established one can begin to use the science of linear regression to calculate a linear model that quantifies the dependence between a response variable and one or more explanatory variables.

Residuals tell us the distance between a fitted line and any individual observation used to estimate the fitted line. Say, for example, that a set of observations y_i is estimated by the linear equation $y = \alpha + \beta x_i$. The residual equals $| y_i - y_{i\,hat}|$. Where $y_{i\,hat}$ is the estimated value by the fitted line and y_i is the actual observed value. Residuals are a good way to measure if a line is a good fit for the data. If the residuals are large the model likely isn't a good fit to the data. Therefore we will learn techniques used to figure out how to minimize these residuals. Typically in a linear model residuals are represented by the greek letter epsilon and are often referred to as the error term. Thus, we can say, $\varepsilon_{i} = | y_i - y_{i\,hat}|$. A good model will look to minimize the epsilon$_i$'s. Essentially, there are two choices one can either minimize the sum of magnitudes of the ε_i's (i.e. $|\varepsilon_1| + |\varepsilon_2| + \ldots + |\varepsilon_k|$), or one can minimize the sum of squared residuals (i.e. least squares $\varepsilon_1^2 + \varepsilon_2^2 + \ldots \varepsilon_k^2$). Least squares is the most commonly used method because it is easier to compute by hand and by software.

The model for the line estimated by least squares is as follows:

$$y\ hat = \beta_0 + \beta_1 x_i hat$$

Where, yhat is the predicted value (response variable), β_0 is the intercept, β_1 is the slope, and X is the explanatory variable. The exact values of β_0 and β_1 are difficult to estimate therefore we use the point estimation technique to estimate b_0 and b_1. The slope can be calculated as standard deviation of the response variable over the standard deviation of the explanatory variable, then multiply this ratio by the correlation. This is shown below.

$$b_1 = sy/sx * R$$

Where sy is standard deviation of the response variable, sx is the standard deviation of the explanatory variable and R is the correlation between the two variables. The formula to estimate the intercept is

$$b_0 = b_1 X \text{ hat} - y \text{ hat}.$$

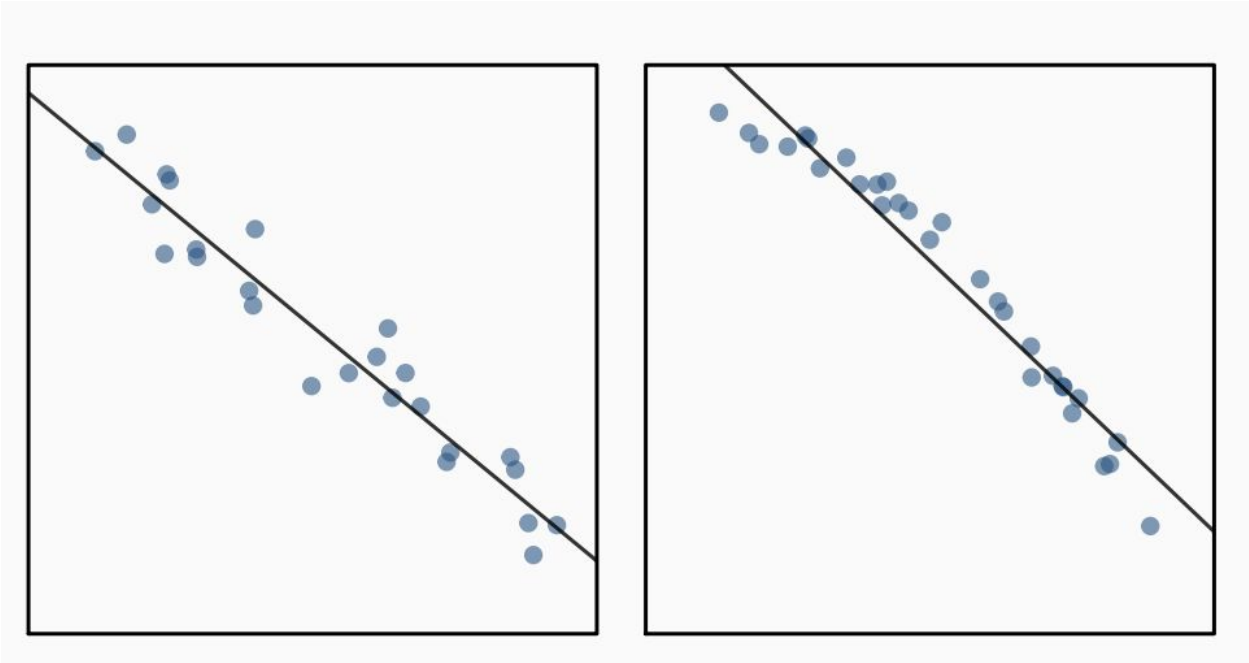
Where, b_1, is the value calculated above. X hat is the calculated mean of the explanatory variable and y hat is the calculated mean of the response variable. The intercept is where our fitted line intercepts the y-axis, at times this is meaningful and other times it's simply an insignificant part of our model. The intercept should be interpreted as the expected value of the response variable when the explanatory variable is 0. The meaningfulness of the intercept is entirely dependent on what is being modeled for example if we're trying to explain the percent of people that live in poverty via high school graduation rates then our intercept is not so meaningful because there does not exist a state that failed to graduate at least one student. Using the intercept in this case would be an extrapolation because it is not realistic to expect that it would be useful. However, if one were to regress the binary variable sex on wages, the intercept would be very meaningful. When the explanatory variable (i.e. sex = 0) the response variable would be the average wage of men or women (whichever gender we assigned the value of 0). Similarly, one is able to get a predicted value by entering different probable values into the explanatory variable. Thus given the model from above with our point estimates instead of true parameters:

$$y \text{ hat} = b_0 + b_1 \, x_i \text{hat}$$

We are able to predict values of yhat by entering in different and probable values for X. When a model is used to predict values beyond what is reasonably predictable based on the observed data there is a risk of making an extrapolation. This leads to faulty assumptions and should be avoided. This least squares line is very useful, however to use this line three conditions must be satisfied.

In order to use simple linear regression using least squares one must ensure (1) linearity, (2) residuals should follow a normal (or nearly normal) distribution, and (3) the observations should have a constant variance. These conditions are important because they qualify any claims or conclusions that are made from a set of data. Thus if a set of data is used to make certain claims or conclusions, but the criteria above are not met then the claims are worse than useless they are dangerous.

The linearity condition states that the relationship between the response variable and the explanatory variable is linear. One way to check this is to graph a scatterplot of the data and check whether the observations seem to have a linear shape. The following are examples of observations that take a linear shape, and observations that do not.

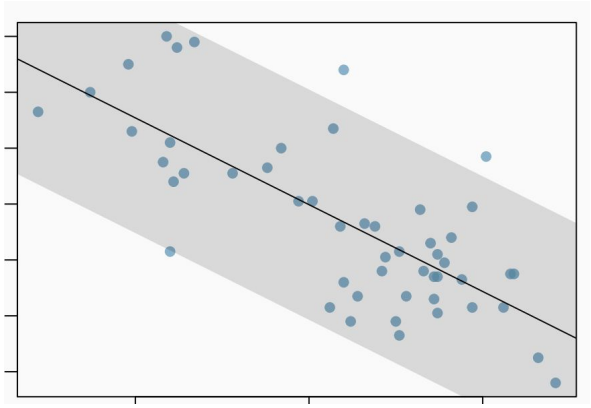


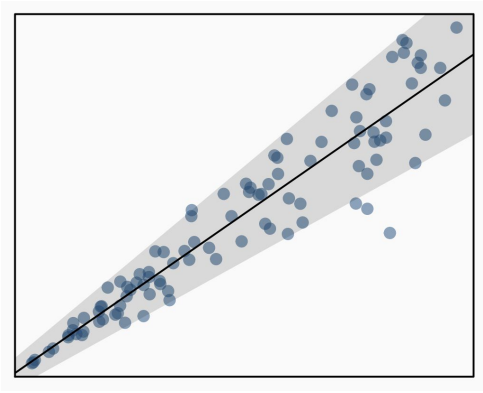
Another method for checking the linear relationship is a residual plot. This plot can be checked to see if the observations are linear by checking if the observations are randomly scattered with the same magnitude above and below 0. The linearity condition is important because linear regression creates a linear model that fits the data. If there is no linear relationship between the two variables the model developed by linear regression will lead to faulty conclusions about the data. The Linear regression technique should be avoided until one has observed a linear relationship between the variables to be used in the regression.

Residuals should follow a nearly normal distribution. This means that the e_i's mentioned above will follow a normal distribution (close enough). Normally distributed residuals are important in regression analysis because they imply that our model is consistently accurate for estimating the response variable. In other words if residuals are normally distributed then the mean of all the e_i's is zero, and hence the $E(e_i)=0$, for any i (I.e. |yi - yi hat| is 0 on average). Clearly one can see why normally distributed residuals would be ideal, however, in practice, this isn't a realistic expectation. In practice it's not realistic to expect normal residuals because that would imply that the model is a perfect fit and thereby this condition might be overly strict. Instead, the residuals should follow a normal distribution well enough. This condition is violated most commonly when there are unusual observations in the data. One can check whether this condition is satisfied by checking normality with a Q-Q plot of the residuals, a residuals plot, or a histogram of the residuals. While normality isn't necessary straying too far from normal can be dangerous and lead to faulty conclusions and a faulty model.

The variability of the points around the estimated least squares line is constant. Constant variability is important for the least squares model because it implies that the error tolerance is

similar on all portions of the data. This implication allows us to validate the predictive power of the data. If variance were not constant then the standard errors would be questionable and the predictive power of our model would be limited, if still possible. Constant variance can be checked by plotting a scatterplot of the data along with the estimated least squares line. If the points seem to be scattered about the least squares line within a constant magnitude then the data has constant variance. Included below are examples of constant and non constant variability in dataset.





Checking the variability of a dataset is important for linear regression because it validates our use of standard errors, and other conditions that make it possible to use least squares to make useful predictive model.

R^2 is a measure of fit. It quantifies how good of a fit a particular model is at predicting the given observed data. R^2 is the correlation coefficient R squared. R-squared will always take on values from zero to one. Low R-squared values imply a model is a bad fit, whereas larger R-squared values imply a better fit. This value tells what percentage of the variability in the response variable is coming from the explanatory variable. In essence this value tells what percentage of the response variable is explained by the model. Therefore 1-R-squared accounts for the percentage of the response variable that is explained by other variables not included in the model. R-squared would clearly be a useless value if three conditions above are not met.

In order to have a predictive model it is necessary to first construct confidence intervals. Confidence intervals are bounds that restrict all the data within an upper and lower bound. These confidence intervals are usually placed on point estimates. These confidence intervals are constructed using a specific t-value that corresponds to an arbitrary alpha value and the degrees of freedom in calculating the point estimate. The alpha value corresponds to the significance level one wants to estimate. For example a 95% confidence interval would correspond to an alpha value of 0.05 and when degrees of freedom approaches infinity the corresponding t-stat is 1.96. Thus, to construct a confidence interval the following values are needed, t* which denotes the t-statistic, b_1 which denotes the point estimate and SE which denotes the standard error. Given these values a confidence interval is then given by:

$$(b_1 - t^* \times SE, b_1 + t^* \times SE).$$

A confidence interval can then be used to determine the interval an estimated value should lie within given a particular confidence level. Thus, higher confidence levels will result in confidence intervals with less magnitude and larger confidence levels will result in confidence intervals with larger magnitude.

Conclusion

In order to create a procedure to help calibrate these snow gauges, we must use regression analysis, since the data collected from these machines will not be fully consistent, as inaccuracies grow with wear, age, and usage, as well as vary between different machines. Regression helps to estimate and group the readings using maximum likelihood estimators and least-squares, and to use these results to give accurate analysis of the calibration of the snow gauge.