Snow Gauge Calibration

Lu Xu and Zehui(Barry) Zhang

Author Contributions

In this homework, both Lu and Barry are contributed into the process of doing this homework. We complement each other's ideas and works. Specifically Barry contributed to the code aspect more, and Lu did slightly more doc work. But roughly equal weight. After we finished all the work, we went through all the staff homework together and revised some of it.

Index

1.1 Introduction

The United States Agriculture department's forest Service wants to moinot the water supply so they operate a gamma transmission snow gauge. This gamma transmission snow gauge is located in the Sierra Nevada mountains near California, Soda Springs. The snow density's depth profile is determined by gauge. By analyzing them, will given the manager will give the information about how the climate change, flood, and water supply. Gamma ray emissions were used to convert the measurement to density. Each season they may change the function to convert the measured values to density. This is caused by the change in temperature or configuration of the detection. Every year there will be a calibration run to adjust the conversion method in winter. Therefore, we need to calibrate the snow gauge from the data by using the procedure. The calibration process is first known as the function of the density of the polyethylene blocks that are measured by the gamma ray intensity, second there are maps density to gamma ray intensity function, and third we use the function inverse to map gamma ray intensity to density. There is a physical model for it. Radioactive sources emitted gamma rays to the polyethylene block and some of the molecules may be absorbed or scattered. Less of the gamma rate for the denser polyethylene. One way to solve the calibration problem is to count the number of polyethylene molecules. This would determine the density of the polyethylene. P is the probability that rays get through simple molecules. P^m is the probability that rays get through to the detector. D equal to cm is the density and the g=Ap^m is the instrument gain. A is e^beta d is the gamma ray measurement and A is greater than zero and beta is less than zero and both coefficients are unknown. Since the beta is less than zero we know it is a decay exponential. Then the calibration main goal is to know what the coefficients debta and A. Y equal to log g can be the same as X equal to d which gives me that Y equal to beta0 + beta X + error and they both can be inverted to a new function density d to solve for observed gain g.

The main goal of this study is to develop a procedure to calibrate the snow gauge from data. This procedure could convert the gain into density while operating the gauge. The final gauge will be the one used to measure gain. For the study, we will use the regression, prediction, testing and confidence intervals, and the longitudinal data to find the procedure. In the given data set, we have 9 density and 10 measurements.

Data

Data was based on the single snow gauge calibration run. There are emitters, polyethylene blocks ,and detectors. The emitter will emit gamma rays to the polyethylene blocks and some of them will be absorbed or detected. Those gamma photons in the detector are counted and it is called "gain" which is the gauge measurement. Our data was made of 9 densities with 10 measurements in each.

2.0 Basic Analysis

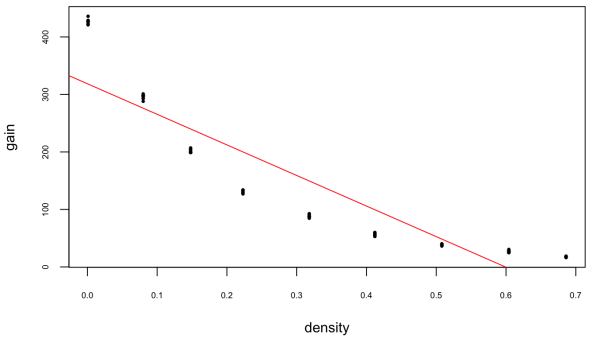
2.1 Raw Data

Method

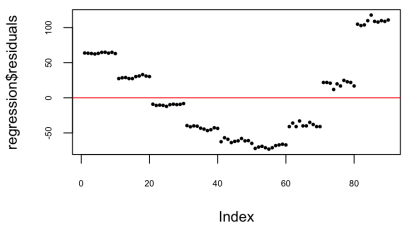
In R, Using the data set that has 9 density and 10 measurements for each density to fit the linear regression model. Then we plot it out and check whether the transformation is needed.

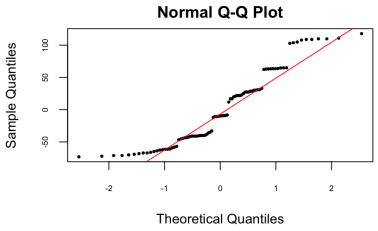
Analysis

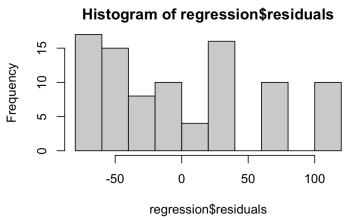
```
lm(formula = dat$gain ~ dat$density, data = dat)

Residuals:
   Min     1Q Median     3Q    Max
-73.08 -44.29  -9.72  30.82 117.83

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   318.70      10.79   29.54   <2e-16 ***
dat$density  -531.95      26.95  -19.73   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 57.54 on 88 degrees of freedom
Multiple R-squared:  0.8157,    Adjusted R-squared:  0.8136
F-statistic: 389.5 on 1 and 88 DF,  p-value: < 2.2e-16
```



The intercept is 318.7 and slope is -531.95 which given the function of Y = -531x+318.7 where x represents the density. We could see the graph, it shows that there is a negative correlation between the gain and density. It makes sense that when we have denser polyethylene there will be less gamma rays to reach the detector. From this regression line we have some problems with it. The R-squared is 0.8136 which is not close to 1.It means roughly 81.36% gains could be explained by this linear model. That is to say the measurement is not really fit to the regression line. Also from the actual values, we could see that it is not close to the predicted line which is shown as the red line in the graph(the best fit linear line). For example, the density 0.07 given us the gain is 281.4635 which is not really close to the actual one.







We can check the linearity, constant variance, and normality. They are helping to avoid the fit bias, variance of the estimated slope and intercept, and evaluate the statistical significance of the slope and intercept. For this residual plot, residuals are not really normally distributed around the red line and the histogram also proves this. For the Normal Q-Q plot, it shows heavy tails on the graph and does not fall in the straight line. This indicates that it is not linear. Sample Quantiles

and theoretical Quantiles are not linear relationships. Also from the histogram of regression residual we could see that the graph is extremely skewed data. Therefore, in order to get rid of outliers we need to use transformation for this data set.

Conclusion

By plotting the fit and fitting a regression line to the data, the graph shows that the distribution is not a normal distribution and there is an extremely skewed and non-normal distribution in the graph. Therefore using the Transformation will help the model to get rid of the outliers.

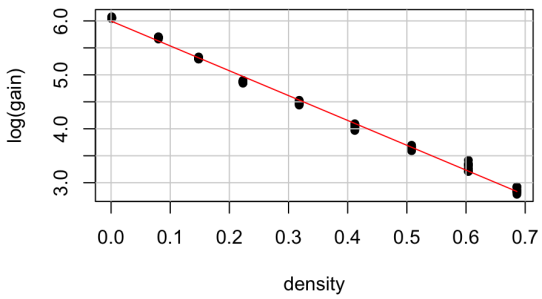
2.2 Transformed Data

Method

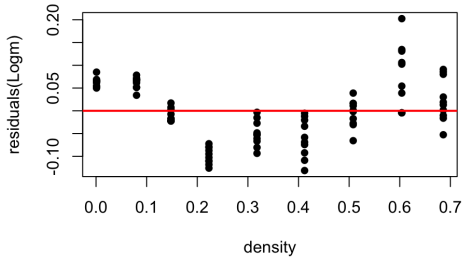
In this part, we are going to use log transformation to transform the data and fit the data to the model. Then plot the new fit, examine the residuals then compare with the residuals to the previous one. And check if log transformation works well.

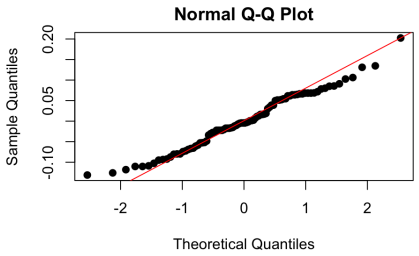
Analysis

```
Call:
lm(formula = log(gain) ~ density, data = dat)

Residuals:
      Min        1Q    Median        3Q       Max
-0.131216 -0.052396 -0.004436  0.054607  0.202447

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.99727    0.01274   470.8   <2e-16 ***
density     -4.60594    0.03182  -144.8   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06792 on 88 degrees of freedom
Multiple R-squared:  0.9958,    Adjusted R-squared:  0.9958
F-statistic: 2.096e+04 on 1 and 88 DF,  p-value: < 2.2e-16
```



After the log transformation, we got the new linear function of 5.997 as their intercept and -4.606 as its slope. The function will be rewritten as Y = -4.606x+5.997. This function is really fit to the line, we still have a negative relationship between the density and gain but it was more fit to the line. The adjusted R-squared is 0.9958 which means that the measure is really close to the fitted regression line. Roughly 99.58% percentage log(gain) can be explained by this model. Below is the residual and normal Q-Q plot.





From the graph we could see that The theoretical quantiles and the sample quarries are roughly straight. Also, it has a constant variability and the residuals are normal with no extreme outliers. All the points fit into the prediction line. This transformation actually helps to avoid the fit bias and provide a much better model to fit these data. Also, it makes sense that gauge measurement will be better for converting in this procedure than the previous one. The physical model of the gamma ray measurement is g = Ap^m then plot the log into the p and then we will get it equal to Ae^betad. The d represents the density which equals the number of molecules times some constant. Lowercase g is equal to Ap^m are the are the instrument gain that combine from the probability of detection to some constant. The A must be greater than zero and beta must be less than zero. The gamma ray did have decays exponentially which showed in the graph. From the data that are really fit to the line which means this line will give a good procedure for converting measurement. Therefore Y= Beta p + Beta X will equal to Y = -4.606x+5.997 is the new function for observed gain by an estimated new density.

Conclusion

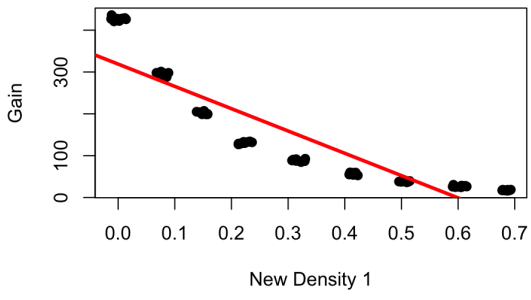
After we made a transformation and exam the residuals are normal distribution without outliers and the data are fit to the new transformed data so Y = -4.606x+5.997 where x presents the density as a new procedure for converting.

2.3 Robustness

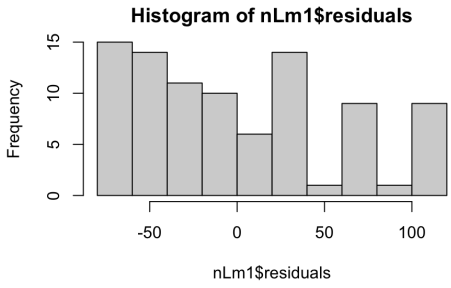
Method

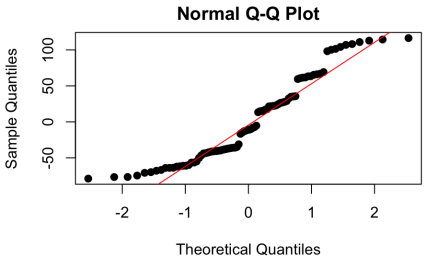
Assume the densities of the polyethylene blocks are not reported on the actual number. We are going to do a simulation test more than once by adding a random error into the data set for both transformed model and the linear model and then see the effect to the fit.We are making data sets and doing a stimulation test on those 2 data sets then doing the transformation.

Analysis

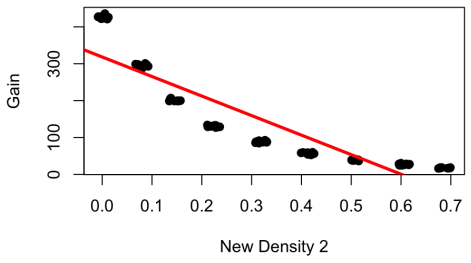



```
Call:
lm(formula = dat$gain ~ ndensity1, data = dat)

Residuals:
   Min     1Q Median     3Q    Max
-78.78 -43.20 -10.96  34.31 116.42

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   318.96      10.69   29.83   <2e-16 ***
ndensity1    -533.10      26.73  -19.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 57.04 on 88 degrees of freedom
Multiple R-squared:  0.8189,    Adjusted R-squared:  0.8168
F-statistic: 397.9 on 1 and 88 DF,  p-value: < 2.2e-16
```

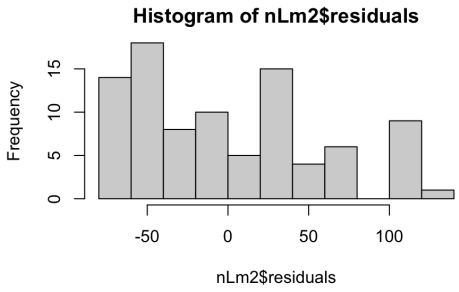


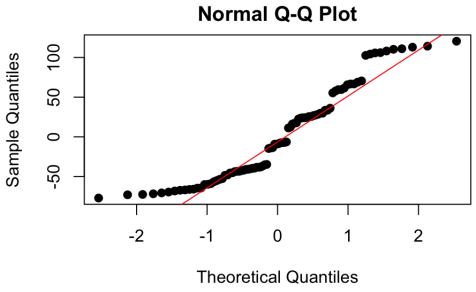


These graphs are all about the first data set created with original data points which is non transformed. From this simulation test, we could see that the overall shape is really similar to the 2.1 raw data.The intercept is 318.96 and the slope is -533.10. The intercept and slope do not change much after the simulation test, compared to the funcion we get in 2.1 raw data. There are only slight differences compared to the original one. There is only a slight change in the slope but overall it still keeps the negative correlation between the density to the gain. Based on the Q-Q plot we also could see that it had some unmatched points, such as the one nearly -3, and it is not normal distribution from the histogram graph. The R squared value is 0.8168 was not close to 1 so it needs to use the log transformation.



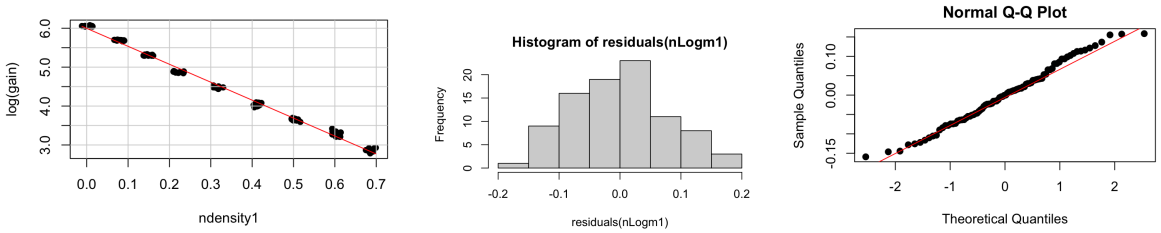

```
Call:
lm(formula = dat$gain ~ ndensity2, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-77.089 -45.448  -8.946  32.587 120.481

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   318.36      10.82   29.42   <2e-16 ***
ndensity2    -529.56      26.95  -19.65   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 57.74 on 88 degrees of freedom
Multiple R-squared:  0.8144,    Adjusted R-squared:  0.8123
F-statistic: 386.1 on 1 and 88 DF,  p-value: < 2.2e-16
```



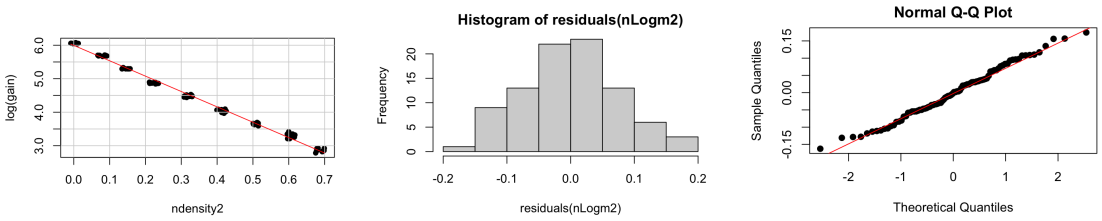


This second data set is created from the original and it is also non-transformed. We are also doing the simulation test on this one. The intercept and slope is 218.36 and -529.56 which is also similar to the original one. The trend is kindly similar to the original one and not a normal distribution. Also both data sets are really similar. It is also similar to the data set 1 that

R-squared is not nearly to 1 so it is not close to the fitted regression line so we need to use the log transformation.



After doing the log transformation for the dataset 1, it shows that the graph is really fit to the line without any extreme outliers. The residual graph shows that the error terms are normally distributed. The histograms are symmetric and normally distributed, too. The adjusted R-squared is 0.994 which is really close to 1 so it is close to the fitted regression line.



From the data set 2 log transformation simulation, it is also a normal distribution data with no extreme outliers, and error terms are normally distributed. The adjusted R-squared is 0.9947 which is close to 1 so this is also really fit to the regression line. Both two graphs show that the log transformations help the data set to get to the normal distribution and most accurate procedure for converting. By comparing both of the log transformation simulations to the original one, there is not much difference and also means there does not have much fluctuation between them.

Conclusion

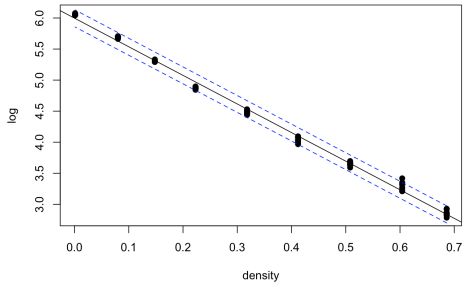
After doing the simulation test to find out if the polyethylene blocks are not reported exactly how does it affect the fit. When we do a simulation data set for it, we find out there is not much difference on the graph, it only has slightly skewed and little bit slope change in the graph. Also after the log transformation, we will always get the normal distribution and do better on predicte the procedure.

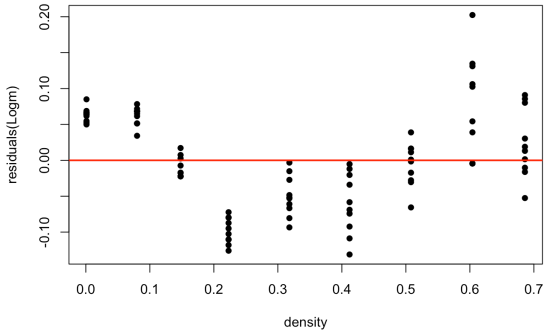
2.4 Forward Prediction

Method

In this part, we will use R to find out, produce, point estimates and uncertainty bands for predicting the gain function of the measured density. By using the model created in the previous part, we could easily find the point estimates and the prediction intervals.

Analysis





Given Density	Fit	lower	upper
0.508	38.76236	33.82731	44.41737
0.001	400.4783	349.0946	459.4523

This chart is about the predicted values for a given density value and the lower and upper bundle. In the second analysis question, we found that the R square value here is 0.9958 which means 99.58% of the variability in the transformed gain is explained by this model. Thus the predicted value by this model is quite reliable. As the graph shown above, the blue dot line represents the prediction intervals of this model. And the black line is the best fit line of the model we made before. Apparently, almost all dots are included between two blue dot lines. This implies that almost all expected values are reliable which correspond to the high R square value. Looking through the graph, the expected balck dot at the place around 0.6 is overpassing the expectation interval a little bit. Thus some gains around here are less accurate than others. Recall the residual plot in the second question. Also, we could find that the region of the residual around 0.6 is larger than the rest of them. Moreover, residuals of predicted gains around 0.20 to 0.25 does not have any overlap of the red line, which means the prediction around this place always has some errors and is not accurate all the time. This condition also occurred around 0 to 0.1. Therefore, in general, around the density around 0 to 0.1, and 0.2 to 0.25, the predicted value always has an error, and predicted values for density around 0.6 have the largest possible overpassing the prediction interval.

Based on the model we set before, the prediction value is the exp(5.99727 - 4.60594* density). For the densities of 0.508, the prediction value is 38.76236, and range is (33.82731, 44.41737). For the densities of 0.001, the prediction value is 400.4783 and range is (349.0946, 459.4523).

The difference between the upper and lower for densities of 0.508 is around 10.59, for 0.001 this number becomes 110.36, which is much larger than the prior one. This situation matches with the conclusion we made before that it is harder to accurately predict the gain when the given density is around 0 to 0.1. Larger range means higher possibility that the prediction is not accurate.

Conclusion

After we forward predict the gains when density is given, we could find that almost all predictions are under the reliable and acceptable region which is the prediction interval.While it is not hard to find that is the given density is 0 to 0.1 and 0.2 to 0.25, the prediction will always have a tiny error and when the density goes to 0.6, there is high possibility that the prediction is not accurate. And the much larger range of measured gains of 0.001 matches the conclusion mentioned above and the larger the range is, the less accurate it is.

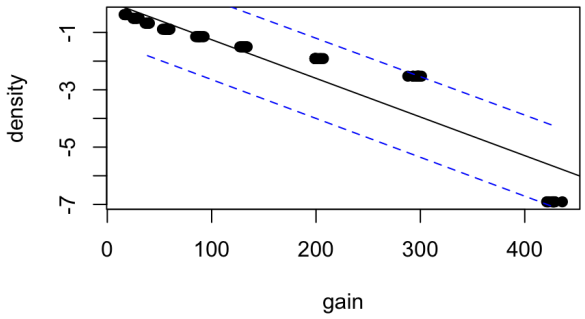
2.5 Reverse Prediction

Method

Based on the prediction we had made in 2.4, we will do an invert prediction to it. Then doing an invert to the uncertainty bands to produce a prediction interval and point estimate for the density. After reverse predictions to get the true density values. Find the densities that are harder to predict.

Analysis

```
Call:
lm(formula = log(density) ~ gain, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3279 -0.2491 -0.2061  0.1932  1.4364

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.09586    0.10706   0.895    0.373
gain        -0.01348    0.00055 -24.511   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6915 on 88 degrees of freedom
Multiple R-squared:  0.8722,    Adjusted R-squared:  0.8708
F-statistic: 600.8 on 1 and 88 DF,  p-value: < 2.2e-16
```



After doing the reverse prediction, we get an intercept of 0.096 and the slope is -0.013. Then Y= 0.096 - 0.013x where x is the gain. We get 0.654 density for the 38.6 gain measurement and the range is 0.163 to 2.617. For the 426.7 we got the density of 0.003 and the variance is 0.001 and 0.014 for the lower bound and upper bond. By comparing the true density values, the measurement gain of 38.6 is 0.146 higher than the true density. The measurement gains 426.7 compared to the true density; it is only 0.002 higher. Therefore, the lower gain measurement will make it harder to prejudice the true density values by inverting the function. I found out that the

larger densities are harder to predict than the lower densities. This is because the larger density has less measurement of gain. From the graph we could see that when gain gets larger the density is a little bit off the fit line. Therefore it makes sense that larger densities are harder to predict. The R-squared is 0.8708 which means that the measurements measure that is really close to the regression line.

Conclusion

After the reverse prediction, we got the new density 0.654 and 0.003 for the gain measurement 38.6 and 426.7. There is more difference in density for the small gain measurements compared to the larger gain measurements. The larger density is harder to predict than the lower density because they have less gain measurement.

2.6 Cross-Validation

Method

To avoid the measurement corresponding to the densities 0.508 and 0.001 which was included in the fitting. We are going to remove those two densities from the original data set and make a new prediction with an average reading of 38.6.

Analysis

```
Call:
lm(formula = log(sub1density) ~ sub1gain, data = sub1)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3056 -0.3024 -0.1853  0.2939  1.4343

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.1591790  0.1241015   1.283    0.203
sub1gain    -0.0136849  0.0006024 -22.717   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7277 on 78 degrees of freedom
Multiple R-squared:  0.8687,    Adjusted R-squared:  0.867
F-statistic: 516.1 on 1 and 78 DF,  p-value: < 2.2e-16

        fit       lwr      upr
1 0.6913853 0.1598598 2.990206
```

```
Call:
lm(formula = log(sub2density) ~ sub2gain, data = sub2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.159267 -0.082428 -0.004974  0.088868  0.148180

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.3841150  0.0161395  -23.80   <2e-16 ***
sub2gain    -0.0075370  0.0001145  -65.82   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0939 on 78 degrees of freedom
Multiple R-squared:  0.9823,    Adjusted R-squared:  0.9821
F-statistic:  4332 on 1 and 78 DF,  p-value: < 2.2e-16

       fit       lwr       upr
1 0.509134 0.4215615 0.6148983
```

After applying the calibration procedure of the data without 0.508 density, we got the intercept, 0.159, and the slope, -0.0137, then the function equal to Y=0.159-0.0137x where the x represents the measurement of the gain and Y is the density. When we have a gain measurement of 38.6, we have the actual density equal to 0.691 and the interval falls between 0.159 to 2.990. The R-squared is 0.867 which is still not high enough, so that it does not really fit the regression line. Then by repeating the same process but omitting 0.001, we have the intercept equal to -0.384 and the slope is -0.008. By converting to the function, it will be Y=-0.384 -0.008x where y

represents the density and x represents the gain. For the average reading block 38.6, we get the actual density of the block is 0.509 and falls in the interval 0.422 to 0.614. The R-squared is 0.9821 which is really fit to the regression line because it is almost one. Which also means that avoiding the measurement corresponding to 0.001 will help to have better prediction.

Conclusion

After omitting the 0.508, we get the actual density equal to 0.691 for the gian measurement of 38.6 and the interval is 0.159 to 2.990. Then we omit 0.001 which gives us the actual density 0.509 and interval 0.422 to 0.614 for the measurement of 38.6. R-squared comparison tells us that omitting 0.001 gives better prediction than omitting 0.508.

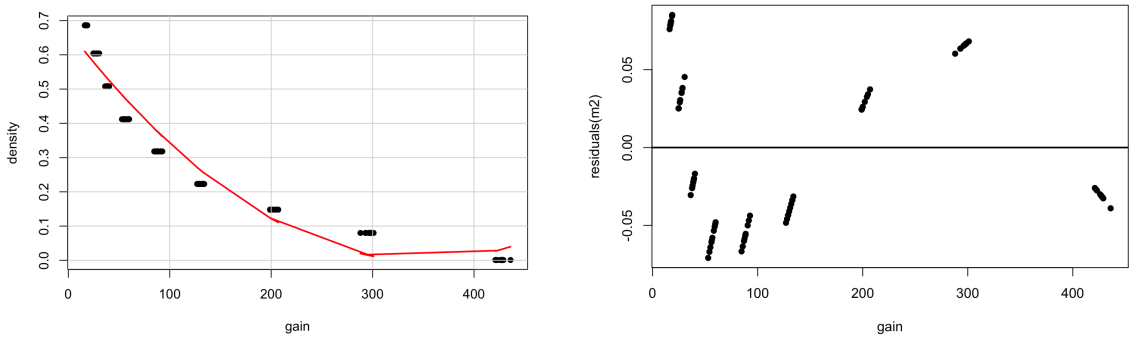
3.0 Advanced Analysis

In the previous analysis, we only used the log transformation, but are there some better models? In this part, we will find if polynomial regression works here. Specifically, what quadratic regression and cubic regression behave compared with log regression

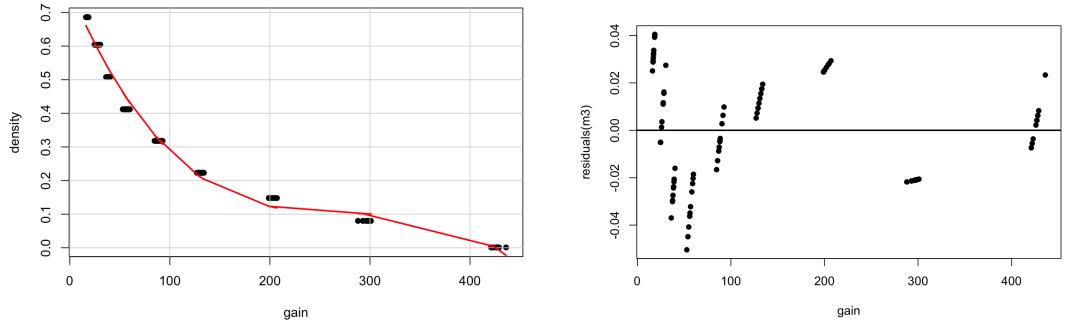
Method

By using R, create two brand new models which are quadratic regression and cubic regression respectively. Then plot their residual plots and scatter plots with a regression line. Finally compare the residual of prediction value for given gain is equate to 38.6 and their R square values.

Analysis



Scatter plot and residual plot for quadratic model



Scatter plot and residual plot for cubic model

	Quadratic model	Cubic model	Log model
R square	0.9508	0.9899	0.9958
Residual	0.02264383	0.02483875	0.1460813

Having a look at scatter plots of quadratic and cubic models, apparently, the cubic model is a little bittle better than the quadratic model. This is because higher degrees may have a chance to overfit the data. But here, there is such a problem due to the sample size and the degrees are still in the low region. Both residual plots seem to provide the indication of homoscedasticity and normality. Residuals are roughly normally distributed around the black line which means 0 residual. Comparing the R square value and difference between real value and the prediction for 38.6 of these three models, it is surprising that the Log model which we created and used in the previous part has the largest R square value. It means the log model can accurately predict the highest percentage of the density. In other words, roughly 99.58% of the density can be accurately predicted by this model if gain is given. This number is95.08% and 98.99% respectively for the other two models. However, even possessing the largest R square value, the Log model has the largest residual, the difference between the real value and prediction. Instead of the residual around 0.02 for the other two, the residual for Log model is around 0.146 which is much larger. Therefore, in general, the cubic model has 0.9899 R number and 0.248 residual. In this comparison, the cubic model is relatively better than the other. But these models are almost the same.

Conclusion

Considering the polynomial regression model, both quadratic model and cubic performance are similar. Quadratic model has a relatively lower R square, but has a slightly smaller residual, the difference between real value and the prediction. Conversely, the cubic model has a higher R square value and the slightly higher residual. When comparing the R square value and the residual with the log model used in the entire analysis, polynomial regression does not perform much better than the log model. Log model has the highest R squared value, but the residual is a little bit larger than both polynomial regression models.

4.0 Conclusion and Discussion

From the Sierra Nevada mountain we need to operate a gamma transmission snow gauge to determine the depth profile of snow density. This will help to monitor the flood management, water supply, and climate change study source. Snow Density is not directly measured by the gauge. It needs to be converted from a measurement of gamma ray emissions. The measurement function needs to be changed over seasons so for this study we develop a procedure to estimate the snow gauge from data. This procedure is used to cover gain into density when it is operated by the gauge. When we use the data and fit a regression line to the data, we find out that it is not normally distributed by seeing the R-squared that is not close to 1.Then we are doing a log transformation to avoid those outliers and get a better prediction function. We get the new function equal to $Y = -4.606x+5.997$ where x represents the density.After the log transformation we see the r-squared is 0.9958 which is close to 1. Then the measure is really fit to the regression line. Then when we are finding the densities of the polyethylene block are not reported exactly, we see it will not have a big effect on the fit after doing the transformation. We will always get rid of the outlier by doing the log transformation so the polyethylene blocks are not reported exactly and will not affect the fit after doing the log transformation. When we do the forward predictions , we find out that almost all predictions are under the reliable and acceptable region which is the prediction interval. When the density goes to 0.6 there is always going to be some error which means that when there is a high possibility ,the prediction will most likely be not accurate. And the much larger range of measured gains will give the less accurate it is. After reverse prediction, we find that there is more difference in density for the small gain measurements compared to the larger gain measurements to the true density values. Also, The larger density is harder to predict than the lower density. To omit the set of measurements correspond to the block of density 0.001 and 0.508 separately, we find that by omitted 0.001 will help to predict a better procedure for converting gain into density because it has R-squared equal to 0.9821 which is really close to 1.

From the advanced analysis, we see that the log models are having bigger R-squared but the residual is a little bit larger than polynomial regression models. Overally, the procedure we will use to convert the gain into density is $Y= 0.096 - 0.013x$ where x is the gain.

Based on this procedure for converting gain into density, I think we also check other polyethylene blocks to see the difference between the other polyethylene blocks and compare them after log transformation.