Measuring Snow-Pack Density From Gain

Contribution Statement: Jerry and Nathan worked on the Fitting problem and the Introduction & Data, while Charlie worked on the Predicting problem. All 3 members contributed to writing the Conclusion.

Introduction

To monitor our source of water in relation to snow pack density, countries all over the world operate snow gauges to monitor snow density and flooding. By detecting the number of gamma rays passed through polyethylene molecules, the snow gauge converts the measurement of gamma ray photon counts into the gauge measurement known as “gain”. The operation uses 9 polyethylene blocks of varying densities with 10 middle measurements of their densities. Together, these tools can help us detect the snow density and therefore analyze the current water supply.

However, the snow gauge seasonally requires calibration due to instrument wear and radioactive source decay. In this lab, we aim to use linear regression to convert gain into density while the gauge is in calibration. We ask, how do we create the best regression line to the gain data and develop a procedure for predicting gain with confidence intervals? In our analysis, we fit a regression line as well as higher degree polynomial models and discuss fit accuracy in relation to errors in density measurements. We analyze the effects of overfitting the model and conclude that a model degree of 2 is the best fit. Then, we fit an appropriate confidence interval around our least squares line that can be used to make interval estimates for the snow-pack density from gain measurements. In the end, we found that if our data is scaled logarithmically we can fit the transformed data with a linear regressor and then build 95% confidence intervals around the linear regression line that will predict where new data points will occur.

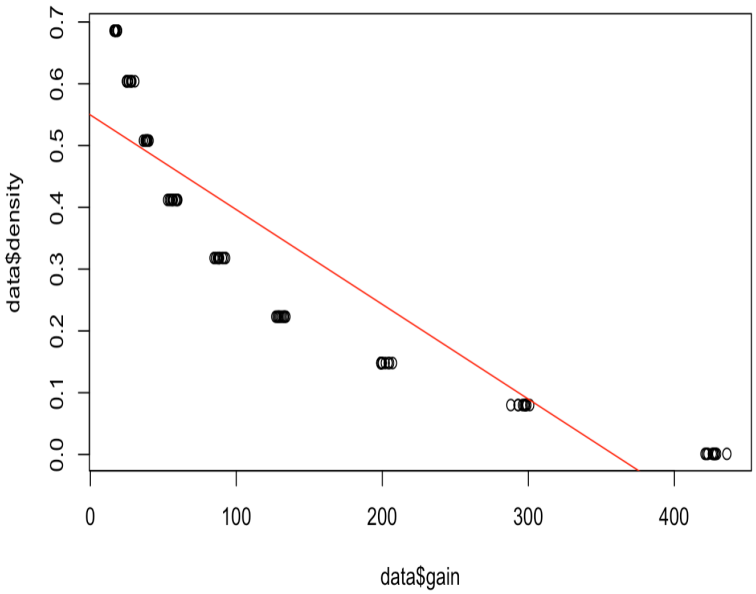
Data

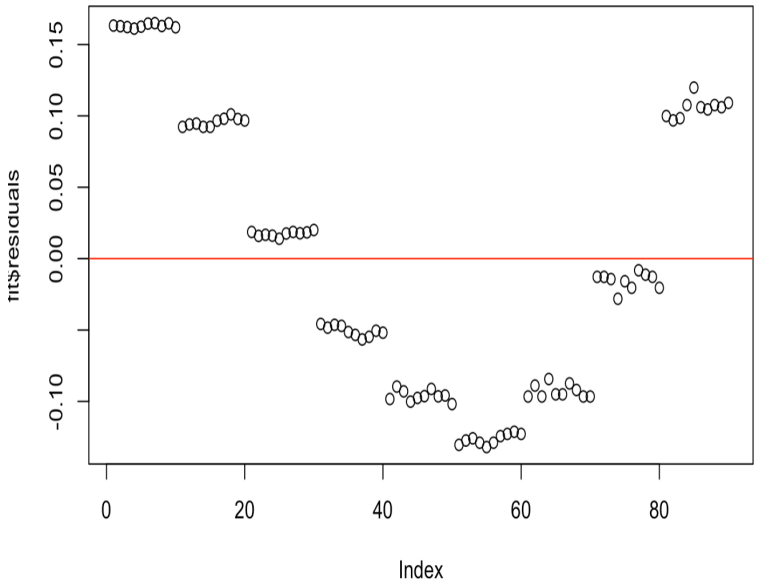
We are given data from a calibration run of the USDA Forest Service's snow gauge located in the Sierra Nevadas. The run used polyethylene blocks of known densities and the researchers took readings on the blocks and took 30 measurements each. In our data, only the middle 10 measurements from each of the 9 blocks of polyethylene are included. The gain column is another name for the gauge measurement of the gamma photon count. The density

column are known densities of the polyethylene blocks in grams per cubic centimeter of polyethylene, or g/cm^3. For 9 blocks of polyethylene with 10 measurements each, there are 90 rows of data available to us.

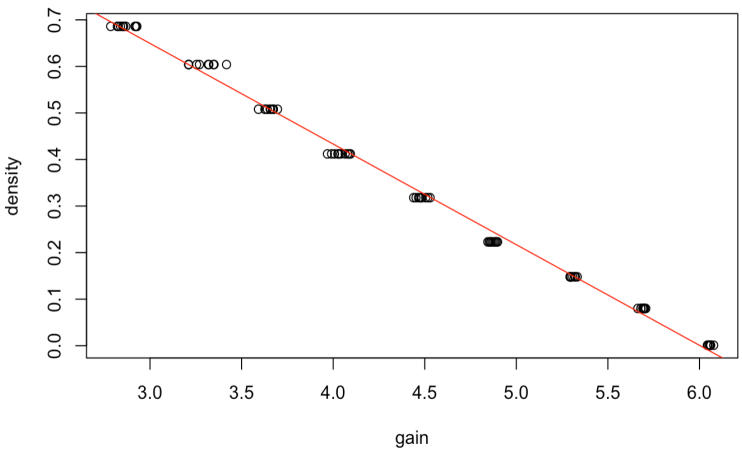
Analysis

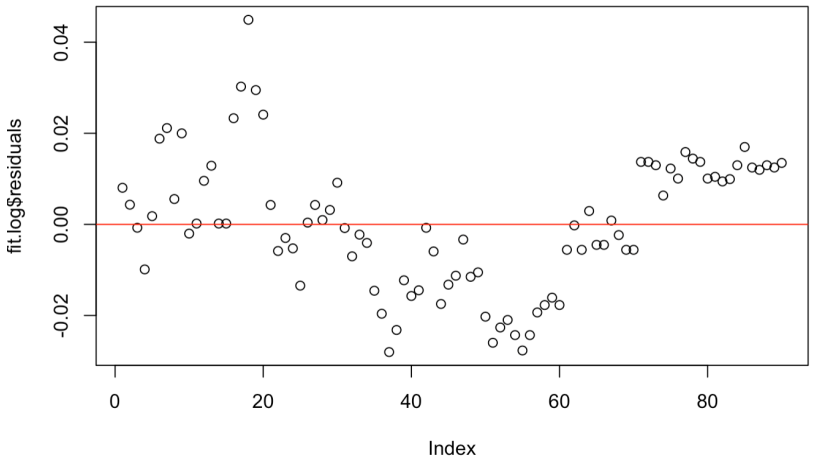
Fitting



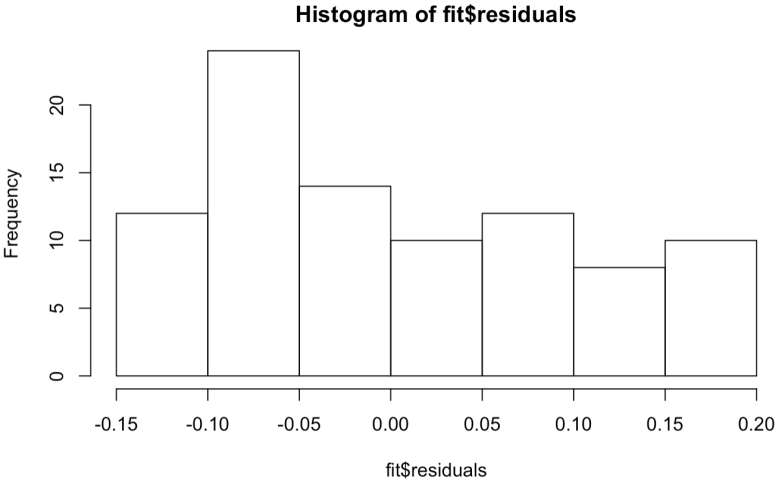


Looking at the plot of gain and density fitted with a regression line, we can see that there is not a linear fit between gain and density. The residual plots also show a larger variance between predictions and observed points, with residuals varying around 0.15 and -0.15.



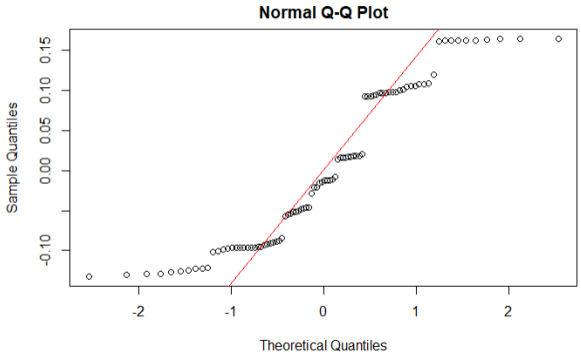


After transforming the gain data to a logarithmic scale, there is now a clear linear fit between the log transformed gain and density. The residuals also show a smaller variance between observed points and predictions between 0.04 and -0.04 now. The R^2 value also increased from 0.816 on the original data to 0.996 on the log transformed data.




```r
# linear relationship on original data
summary(fit)$r.squared
# linear relationship on log transformed data
summary(fit.log)$r.squared
```

```
[1] 0.8156974
[1] 0.9958183
```



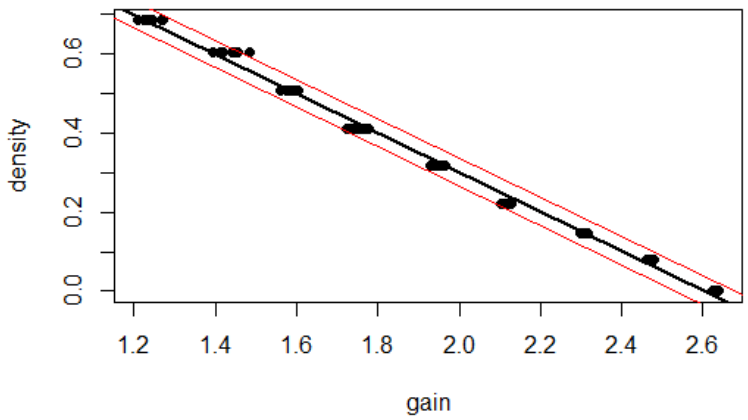
Looking at the residual plot of our log transformed fit, we can see all the points are more or less symmetrically distributed, and they all have a tendency to cluster towards the middle of the plot. All the points clustered around lower values on the residual axis, around 0.04 and -0.04. Additionally, there aren't any clear patterns in the distribution of residual points, meaning the residuals display constant variability. Lastly, we analyze the histogram of residuals and see that there is an approximate normal distribution.

If the densities of the polyethylene blocks were not reported exactly, this would lead to lower accuracy of the linear fit. Lower or higher values than the observed values would lead to higher residuals. It would also distort the distribution of residuals into a more spread shape with more variance.

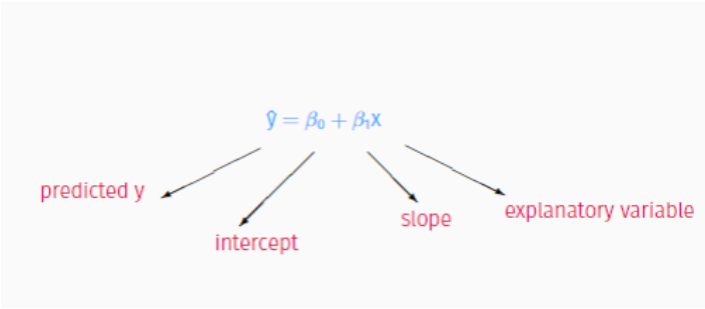
Predicting

We can now use the information gathered from the previous question to implement our transformed linear model to make predictions. Ultimately we are interested in answering questions such as: Given a gain reading of 38.6, what is the density of the snow-pack? or Given a gain reading of 426.7, what is the density of snow-pack? We develop a procedure to add confidence intervals around our least squares line that can be used to make interval estimates

for the snow-pack density from gain measurements. Since our goal is to give the range in which future observations can be thought most likely to occur, we create prediction intervals in R rather than confidence intervals which tells us where the mean of future observations are most likely to reside.



The red lines in the graph above are our prediction intervals and are centered around our regression line. Our prediction interval tells us with a confidence level of 95% where a new data point will lie. Our best guess where a data point will lie is the regression line, which explains why our interval is centered around the regression line. However, there is going to be variance with where our data point will lie, and so our prediction interval helps us capture where it will be with a degree of confidence. It is important to remember that our linear model is transformed with a logarithmic function on the x-axis, and to interpret our model we have to un-transform it. The model we are trying to fit is of the form below, and from our least-squares line we come up with an intercept of 1.2980 and a slope of -0.4978.




```
Call:
lm(formula = density ~ gain)

Coefficients:
(Intercept)         gain
     1.2980      -0.4978
```

Again, to untransform our data, we have to take the log of our X variable before multiplying by our slope, and adding our intercept. For example, to use our model to predict the density of a snowpack given a gain reading of 38.6, we do: 1.2980 +(-.4978 * log(38.6)) = .508 kg/m^3

We know that several measurements of gain were taken from polyethylene blocks of known density so we create 90 values of gain we wish to predict the snowpack density with a confidence interval. This simulates how the data was collected where only the middle 10

measurements from each of the 9 blocks of polyethylene are included so we have a total of 90 measurements.

1.000000	7.730337	14.460674	21.191011	27.921348
34.651685	41.382022	48.112360	54.842697	61.573034
68.303371	75.033708	81.764045	88.494382	95.224719
101.955056	108.685393	115.415730	122.146067	128.876404
135.606742	142.337079	149.067416	155.797753	162.528090
169.258427	175.988764	182.719101	189.449438	196.179775
202.910112	209.640449	216.370787	223.101124	229.831461
236.561798	243.292135	250.022472	256.752809	263.483146
270.213483	276.943820	283.674157	290.404494	297.134831
303.865169	310.595506	317.325843	324.056180	330.786517
337.516854	344.247191	350.977528	357.707865	364.438202
371.168539	377.898876	384.629213	391.359551	398.089888
404.820225	411.550562	418.280899	425.011236	431.741573
438.471910	445.202247	451.932584	458.662921	465.393258
472.123596	478.853933	485.584270	492.314607	499.044944
505.775281	512.505618	519.235955	525.966292	532.696629
539.426966	546.157303	552.887640	559.617978	566.348315
573.078652	579.808989	586.539326	593.269663	600.000000

fit	lwr	upr
1.298013e+00	1.265750631	1.3302745791
8.558441e-01	0.825569726	0.8861184640
7.204403e-01	0.690553612	0.7503269363
6.378195e-01	0.608114666	0.6675242594
5.781875e-01	0.548587634	0.6077873463
5.314974e-01	0.501964099	0.5610306451
4.931213e-01	0.463632398	0.5226101316
4.605411e-01	0.431082542	0.4899995629
4.322336e-01	0.402795956	0.4616712104
4.072070e-01	0.377783558	0.4366304121
3.847791e-01	0.355365017	0.4141932280
3.644607e-01	0.335052265	0.3938691470
3.458888e-01	0.316483243	0.3752943961
3.287868e-01	0.299381934	0.3581917223
3.129390e-01	0.283533092	0.3423449620
2.981740e-01	0.268765611	0.3275823169
2.843531e-01	0.254941223	0.3137649875
2.713629e-01	0.241946600	0.3007792292
2.591092e-01	0.229687692	0.2885306630
2.475129e-01	0.218085588	0.2769401151
2.365070e-01	0.207073430	0.2659405164
2.260343e-01	0.196594081	0.2554745550

The figure to the right tells us what our regression line predicts for our corresponding gain value in the figure to the left with a prediction interval with a lower and upper bound. For example, with a gain reading of 7.73, we are 95% confident that our interval of (.8255, .88611) captures the snowpack density. We have now developed a procedure to create confidence intervals around our predictions to make interval estimates of snowpack density based off gain readings.

Gain	Density(estimate)	Prediction Interval
38.6	0.508	(0.479,0.538)
426.7	-0.01	(-0.041, 0.018)

Prediction Accuracy and Error Rates (Advanced Analysis):

A simple correlation between the actuals and predicted values can be used as a form of accuracy measure. A higher correlation accuracy implies that the actuals and predicted values have similar directional movement, i.e. when the actuals values increase the predicted values also increase and vice-versa. First we create a side by side table to compare values:

	actuals <dbl>	predicteds <dbl>
1	0.686	0.677963682
2	0.686	0.681680734
3	0.686	0.686738350
4	0.686	0.695884260
5	0.686	0.684194753
6	0.686	0.667181233
7	0.686	0.664856445
8	0.686	0.680434603
9	0.686	0.666015714
10	0.686	0.688021459

1-10 of 90 rows Previous

	actuals	predicteds
actuals	1.000000	0.997907
predicteds	0.997907	1.000000

From the code input in R, we can see that we have a 99.79% prediction accuracy.

We calculate other error rates such as Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, and Mean Absolute Percentage Error. The table shows low error rates across the board.

mae	mse	rmse	mape
0.0117129444	0.0002117179	0.0145505292	1.4128703617

Conclusion:

From our analysis of the fitting of the data, we can see that the original data did not have a linear relationship between the predictor and response variables. However, after scaling the data logarithmically, the data can be fit with a linear regressor. Using this information, we were able to make confidence intervals around a regression line that would estimate where new data points would occur with 95% confidence and found that with 95% confidence gains of 38.6 and 426.7 would result in between (0.479,0.538) and (-0.041, 0.018) densities respectively. Something to think about considering our data is that each density has only a few data points. Consequently, if there is an outlier within the data for one density that outlier could heavily skew the data.