Xueru Xie, Yichi Zhang, Jinsong Yang

Professor Schwartzman

MATH 189 Report 4

24 May 2020

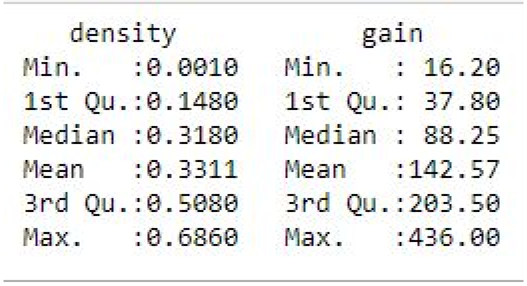
Prediction Model of Calibrating Snow Gauges

In northern California, the water supply comes from the Sierra Nevada mountains. The USDA has snow gauges that use gamma transmission to measure the density of snow that contain water. These snow gauges do not interfere with the snow because meteorologists stick them onto the ground. Snow absorbs water from the rain, and less water will be absorbed when the snow is dense. So the snow gauges can measure the density of snow packs to manage water supply and management. In this investigation, we will read data from a calibration run of the USDA Forest Service's snow gauge. The data consists of 90 measurements of density and gain values. Using a linear regression model, we are going to fit the gain data and predict the density given a gain reading from the snow gauge.

Dataset

Our data gauge.txt contains two columns: density and gain. Since the snow gauge is a gamma transmission, it contains radioactive elements. The gain column is values that are converted from those transmissions. The density column is the density of the snowpack. The density and detector readings relationship comes from a complex physical model, but we are going to use linear regression to fit gain values. Before actually applying the linear regression model, we first take a look at the data. The data are all numbers, and there are no missing values.

The data cleaning is complete. Next, we want to see the summary of the data, and as well as the unique values of density, just to see what they look like.

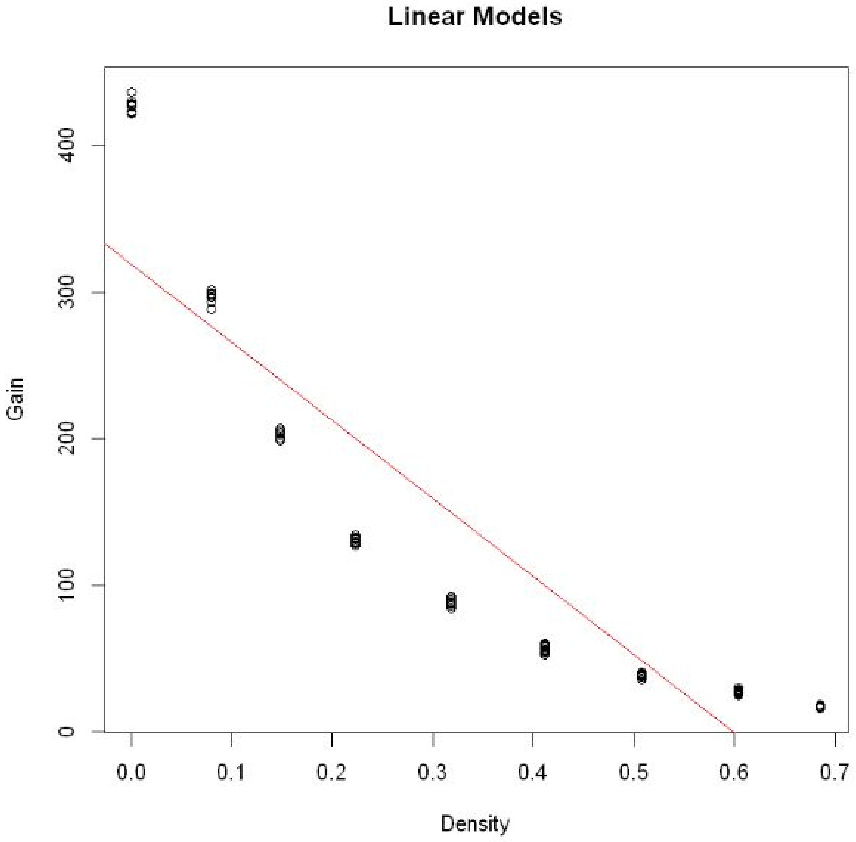


```
    density            gain       
 Min.   :0.0010   Min.   : 16.20  
 1st Qu.:0.1480   1st Qu.: 37.80  
 Median :0.3180   Median : 88.25  
 Mean   :0.3311   Mean   :142.57  
 3rd Qu.:0.5080   3rd Qu.:203.50  
 Max.   :0.6860   Max.   :436.00  
```

From these results, we can see that the gain value can range from a small number like 16.2 to a big number like 436. We also see that the density values contain only a few numbers, but they all appear a lot of times.

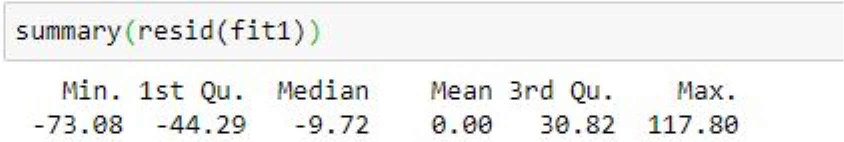
Fitting

After skimming through the data, we start to regress the line to the data and plot the fit. Doing this will allow us to find a general y=mx+b formula of the dataset. As a result of fitting the linear modeling, the y-intercept appears to be 318.7, and the slope is negative -532.0. So the formula is gain = 318.7 + (-532.0) * density.



From the linear model above, we can predict the density values for any given new gains. It could be given the gain of 100, the density would be about 0.4. But before doing that, we can see that although the predicted line captures most of the data points, the points are very far away from the line. This means that the model may not be very accurate. So, we want to check the residuals to detect any problems with the fitting line because the plots may contain unwanted numbers that would influence the model negatively.

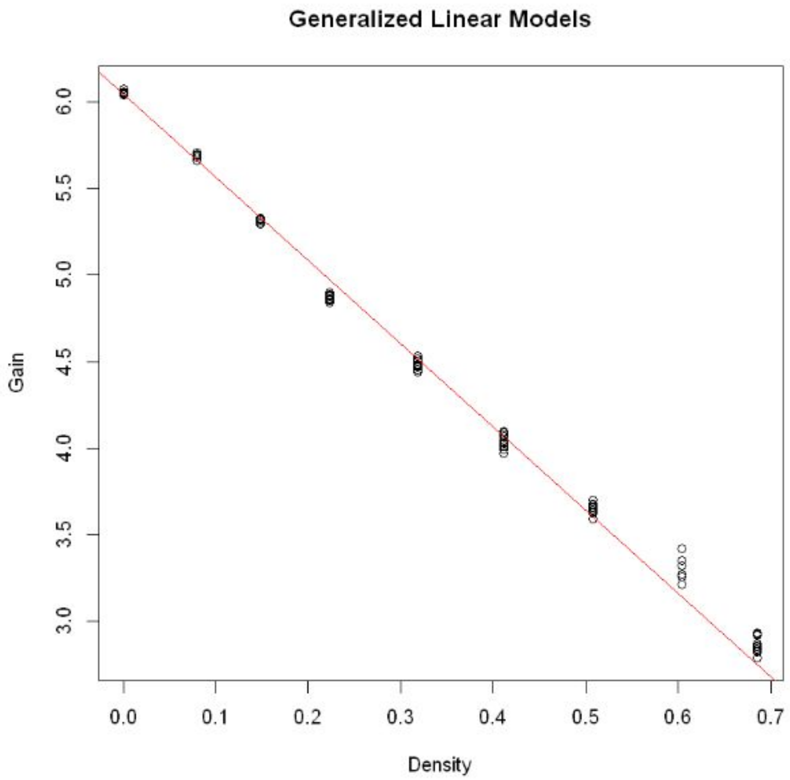
Here we summarize the residuals and find that such large residuals are not good enough.

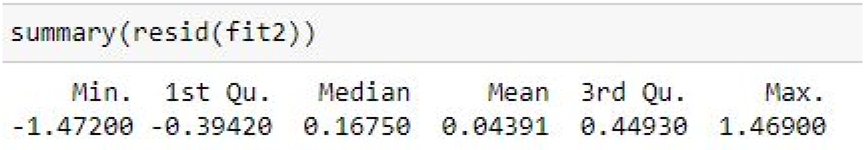


```
summary(resid(fit1))

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 -73.08  -44.29   -9.72    0.00   30.82  117.80 
```

So, if the densities of the polyethylene blocks (the radioactive elements) are not reported exactly, the fit will not be accurate predicting the new gain values. In this case, we decide to pick a more complex model by transforming the gains. Instead of fitting the gain values, we will fit the transformation of gains. From the graph of the linear model above we can see that the points are distributed hyperbolically, with points approaching each axis. Therefore we decide it is reasonable to use log transformation to the gain values. After applying the new logarithmic model with glm function in R, we get the new intercept 6.045 and the slope -4.810. Looking back at the lecture, we agree to this new model because the detector uses gamma rays transmission. The probability of a gamma ray arriving at the detector is e^(m log p) = e^(bx), where p is the probability, x is the density, b is the y-intercept, and m is the slope. This formula is a log equation, which is what we used for the modeling. And such a formula indicates that for every 0.1 unit (use 0.1 instead of 1 since density<1 always) increase in the density would result in e^(0.1*-4.81)= 61.9% decrease in the gain.



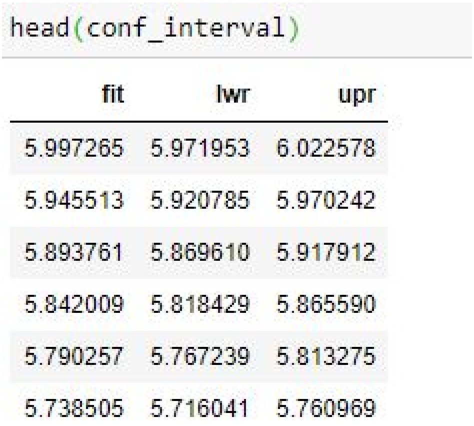


```
summary(resid(fit2))

     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
-1.47200 -0.39420  0.16750  0.04391  0.44930  1.46900 
```

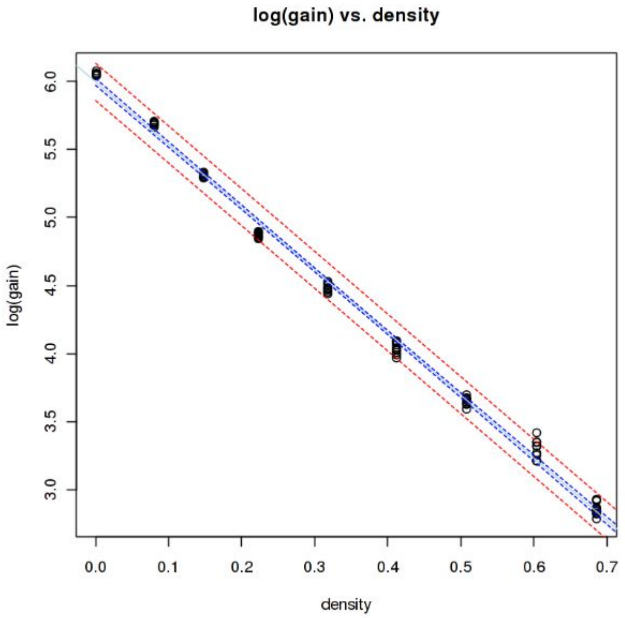
Predicting

Besides developing a prediction line to the dataset, we also want to add the confidence intervals around the least squares lines. Predicting the estimated density is great, but the value isn’t always correct. Instead, we want to use a range of values that could capture the density as close as possible. In our case, we use the CIs to make interval estimates for the snow-pack density from gain measurements. We use 90% of CI, and here are the results of the first five confidence intervals of the predicted gains.



```
head(conf_interval)
```

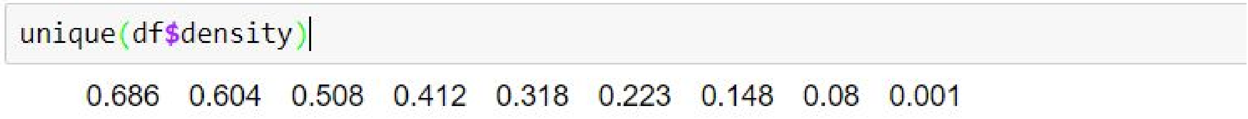
fit	lwr	upr
5.997265	5.971953	6.022578
5.945513	5.920785	5.970242
5.893761	5.869610	5.917912
5.842009	5.818429	5.865590
5.790257	5.767239	5.813275
5.738505	5.716041	5.760969



Data Limitations

Our limitation for this dataset would be not enough of data values. Also, the prediction model might not exactly predict the values, since there might be extreme density values from outliers, or when the device was faulty and could not record the gain measurements correctly.

Below we check the density column and notice that among 90 observed values, there are only 9 unique values of density, while values in the gain column vary.



```
unique(df$density)
```

```
0.686   0.604   0.508   0.412   0.318   0.223   0.148   0.08   0.001
```

For our linear/generalized linear model, a single value of density always points to a unique value of gain or transformation of gain. In other words, our prediction function f(x) is from R to R while our real data are otherwise. This makes sense because the experiment is conducted in a way that several measurements of gain are taken for the same density. Such a characteristic means that our predicting model would always end up having marginal error, since the model can only produce one predicting value with one density value, and therefore, each measurement of gain only increases the absolute margin of error in total. This model can be improved if the density values can be more varied, and each time for each density value, we take the average number of measured gains instead of all measurements.

Conclusion

In summary, we developed a model that could predict the sno-pack density given gain measurements. At first we used the linear regression model, but the great residuals caused us to shift to another method. Eventually, we applied the logarithmic model to our data, and the residuals decreased.

Appendix

In this report, student Xueru Xie wrote the introduction and fitting part of the analysis report. Student Yichi Zhang had the responsibility to write all the R codes. Student Jinsong Yang completed the predicting and conclusion part of the report. During our procedure, we encountered a problem when the linear model had too much residuals, but we were able to reduce that by transforming our data.