Calibrating a Snow Gauge
Yong Liu, Yunlin Tang, Jian Jiao
MATH 189, Spring 2020, HW4

I. Introduction

The main source of water for Northern California comes from the Sierra Nevada mountains. In order to monitor this water supply, the Forest Service operates a gamma transmission snow gauge to determine a depth profile of snow density. However, the gauge does not directly provide the measurement of snow density; instead, the density reading is converted from a measurement of gamma-ray emissions. To adjust the conversion method for updating the density readings, scientists will conduct a calibration run each year. The data used in this lab are from a calibration run of the USDA Forest Service's snow gauge. In the run, several polyethylene blocks with known densities were placed between the radioactive source and the detector. For each block, the middle 10 measurements of the gamma photon count were included in our data, which is the variable "gain"; in other words, there are 10 measurements for each of 9 densities in g/cm^3 of polyethylene. In order to develop a procedure to calibrate the snow gauge, we will first perform the data analysis by fitting a regression line on the data with different fitting models; then we will employ the regression model to predict data.

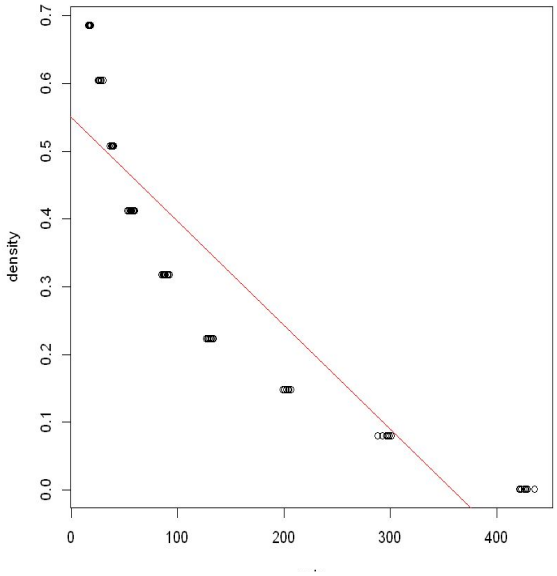
II. Analysis

A. Fitting

In this step, we first try to fit a least-squares regression line directly on the data. Then we determine a transformation on the explanatory variable "gain" and again fit the model to the transformed data. The general goal in this step is to explore the performance of different linear regression models on the data and discuss their issues by providing the residual plots respectively. In addition, we also inspect the outliers in the data in order to check how they affect the fits if some values of the response variable "density" are not reported exactly.

The Simple Linear Regression Model

First, fit the simple linear regression line which is optimized by the least-squares Zon the original data without any transformation. This simple linear regression model has an intercept of 0.549 and the slope of -0.0015. These values give information as such: the gain with the value of 0 is expected to have 0.549 density and a 1% increase in gain is associated with a 0.0015% decrease in % of density. Then by plotting the fit on the data and the residual plots of this fit, it is clear that this regression model roughly follows the trend of data; however, it is robust and underfitting our data.



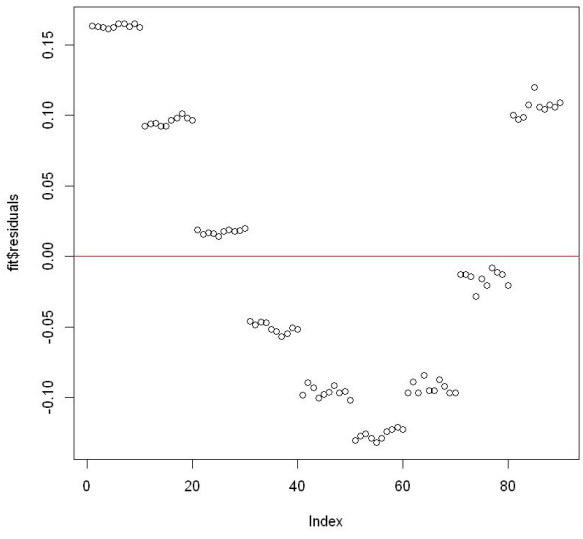
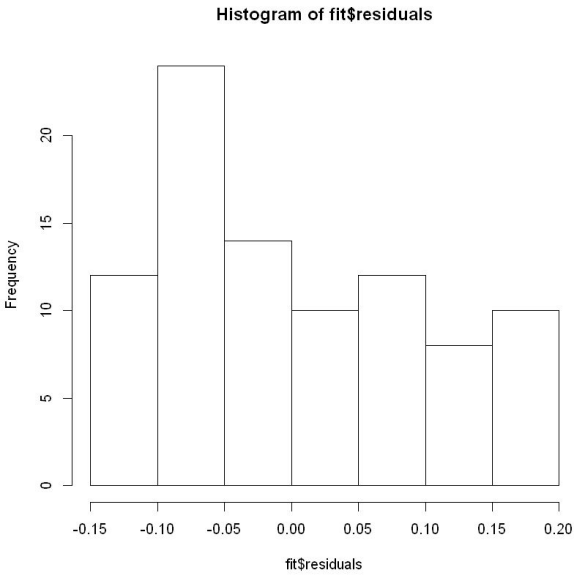


Figure 1.1 (left): Scatter Plot between Gain and Density, fitted by the Simple Linear Regression
Figure 1.2 (right): Residual Plot of the Simple Linear Model



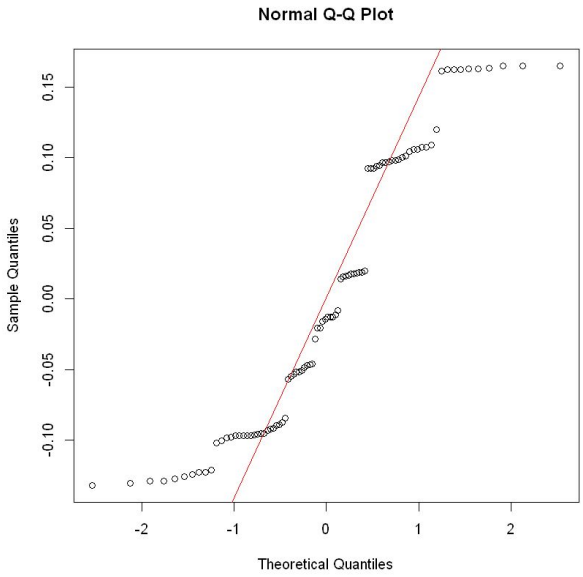
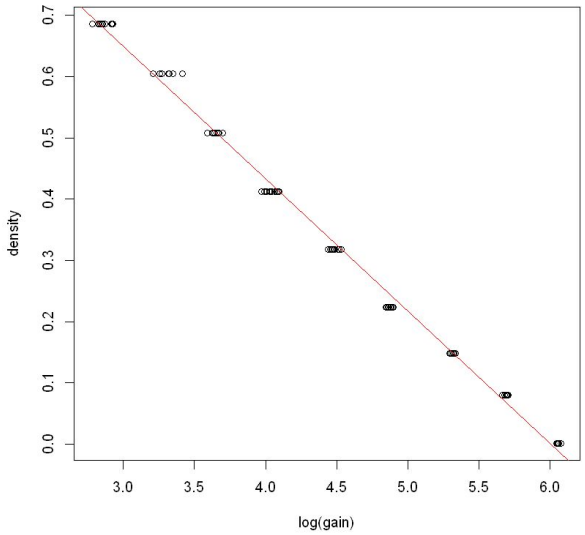


Figure 1.3 (left): Histogram of the Residuals of the Simple Linear Model
Figure 1.4 (right): Q-Q Plot between the Residuals and the Normal Quantiles

The scatter plot shows how the model fits the data. Then use the residuals of the fit to examine the conditions for the least-squares line: linearity, nearly normal residuals, and constant variability. For inspecting the linearity and constant variability, the residual plot (figure 1.2) reveals that there are a nonlinear trend and a non-constant variability of the residuals which do not satisfy the conditions. In addition, by plotting the histogram and normal Q-Q plot of the residuals (figure 1.3, 1.4), we see that the residuals are not nearly normal distributed. Thus we determine to transform the explanatory variable in order to produce a better linear regression fit.

The Linear Regression Model After Log Transformation

The physical model for gain and density $e^{m \log p} = e^{bx}$ also confirms that the simple linear model is inappropriate since it suggests the nonlinear relationship as $E(g) = ce^{bx}$ where g represents the gain, c is a constant, and x represents the density. Then take the natural logarithm of both sides in this equation, obtain $log(E(g)) = log(c) + bx$ which reveals the linear relationship between the gain g and the density x. Therefore, we determine to apply a log transformation on the explanatory variable gain then again fit the least-squares regression line on the data. The fit model obtained has an intercept of 1.29 and a slope of -0.21, which indicates a negative linear relationship between the explanatory variable and the response variable.



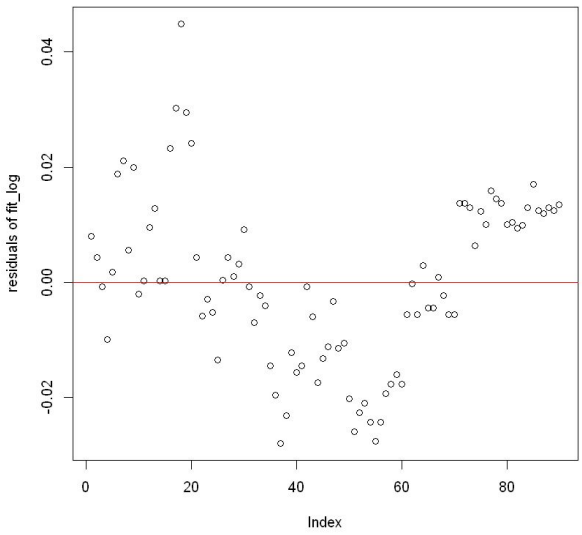
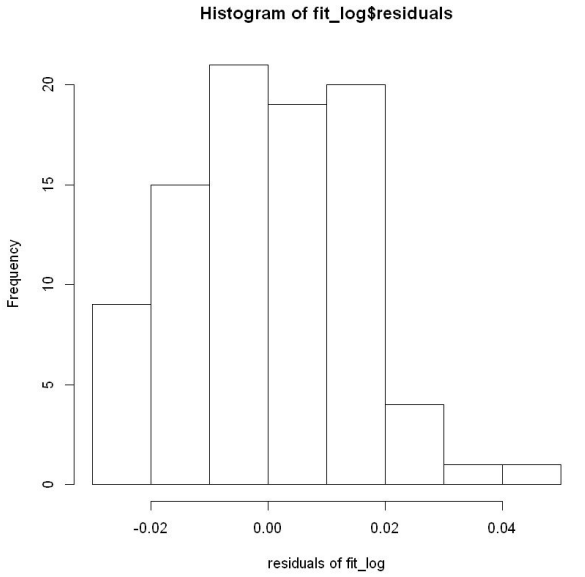


Figure 2.1 (left): Scatter Plot fitted by the Simple Linear Regression after Log Transformation
Figure 2.2 (right): Residual Plot of the Simple Linear Model after Log Transformation



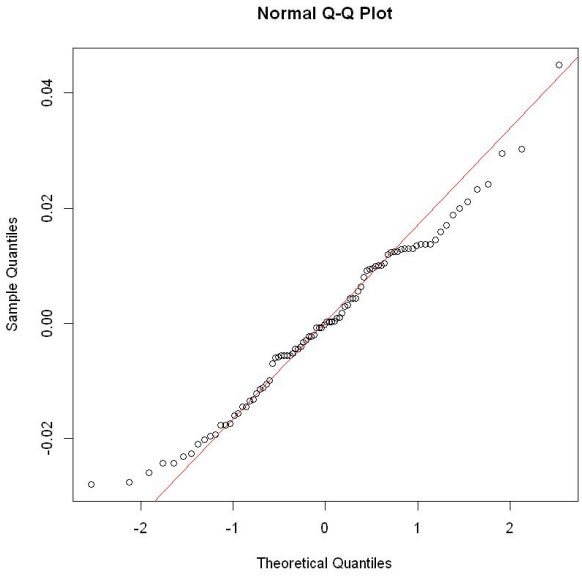


Figure 2.3 (left): Histogram of the Residuals of the Simple Linear Model after Log Transformation
Figure 2.4 (right): Q-Q Plot between the Residuals and the Normal Quantiles after Log Transformation

The above graphs show the performance of the new linear regression model which transforms the explanatory variable gain by taking the logarithm. The scatter plot between log gain and density indicates the new model is fitted better with the trend of data. The normal Q-Q plot (figure2.3) and histogram of residuals (figure 2.4) also indicate this new model meets the requirement of nearly normal residuals. However, by inspecting the residual plot (figure 2.2), two of the conditions of least squares line: linearity and constant variability are still not satisfied, where the plot reveals the non-linear trend and non-constant variability of the residuals.

The Polynomial Regression Model

In comparison, we choose to build a complex model with polynomial regression which is also optimized by the least-squares. To visually process the model selection of polynomial regression,

we draw out the scatter plot between gain and density which is overlayed by actual fits of the polynomials ranging from degree 3 to 7.

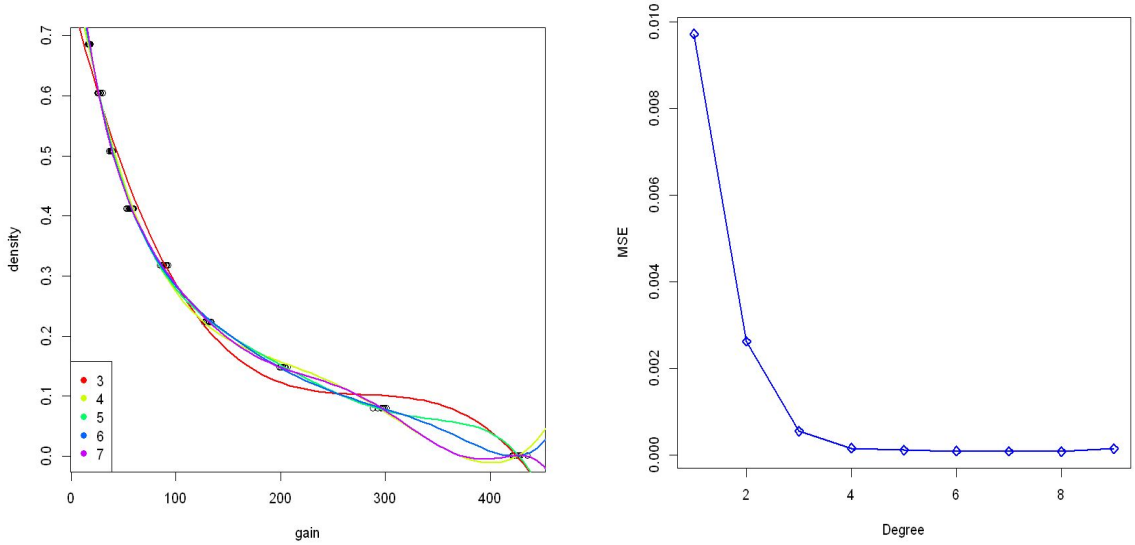


Figure 3.1 (left): Scatter Plot Overlaid by Different Choices of Degrees on Polynomial
Figure 3.2 (right): Scatter Plot of Polynomial Regression with Degrees from 1-10 and their MSEs

After comparing the fitted line of different degrees in polynomial regression, the degree of 4 will be one of the most appropriate candidates among these choices since it fits the trend of data well and has a relatively lower complexity among other degrees. Furthermore, by using cross-validation on the polynomial regression model, each run calculates the mean squared error between prediction and actual value with the corresponding degree. As shown in figure 3.2, the cross-validation MSE achieves a minimum at degree 4 as the optimal degree. Therefore by choosing a degree of four, create a new linear polynomial regression under the optimization of least-squares. This new model transforms the explanatory variable gain into four values: x, x^2, x^3, and x^4 with a total of five coefficients (the weight terms). Then by graphing the residual plot and normal Q-Q plot we can inspect whether this new model satisfies the conditions for the least-squares line.

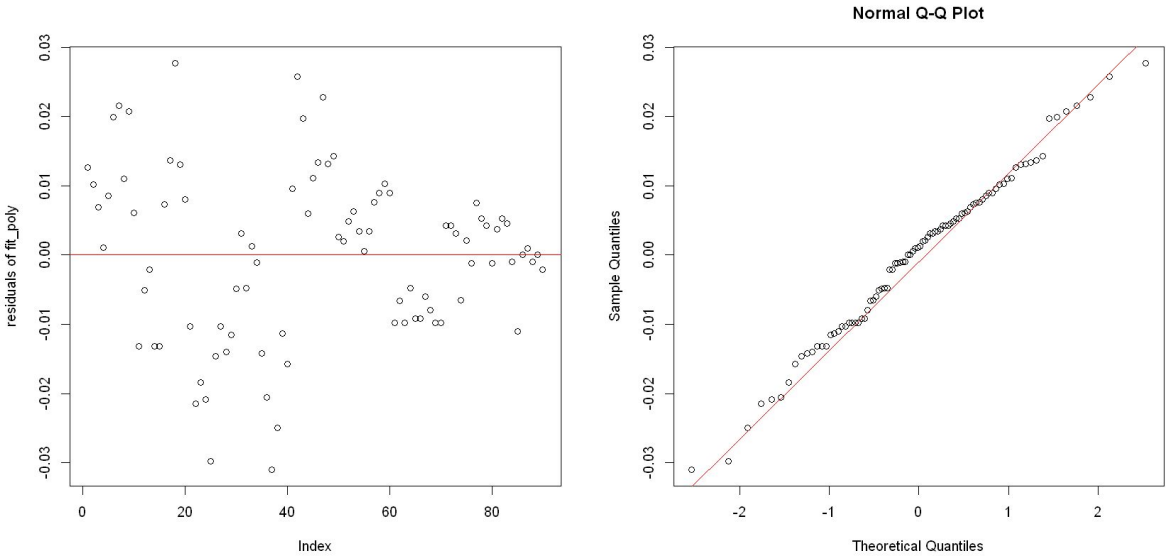


Figure 3.3 (left): Residual Plot of the Polynomial Model
Figure 3.4 (right): Q-Q Plot between the Normal Quantiles and Residuals of Polynomial Model

Both of the two graphs shown above indicate the polynomial model with the degree of four meets the requirements of the conditions: linearity, nearly normal residuals, and constant variability. However, there are still some concerns and issues while employing this model for further predicting the data. For example, the high complexity of this new model would potentially lead to the problem of overfitting. The R-squared value of this model is 0.9971, which indicates that 99.71% of the variability in the gain among density is explained by the model. Nevertheless, high R-squared value also indicates that this new model has the potentiality to describe the noises rather than the genuine relationships in the real population. Thus, in comparing to the other two models mentioned above (simple linear regression and the log-transformed), this model is less robust and more dependent on the training data given. It will be less likely for this model to generalize while some unknown data are added and needed to be predicted.

In addition, to consider how the fit could be affected if there were errors in the density measurements, we choose to inspect the outliers in the given data. From the normal Q-Q plots provided above (figure 1.4, 2.4, and 3.4), it is appropriate to suspect that there are some outliers that potentially affect the fits since they are deviating from the theoretical Q-Q line. After removing these outliers, the first model which is the simple linear regression has changed the intercept from 0.549 to 0.537 and slope from -0.0015 to -0.0014; the log-transformed model changed its intercept from 1.298 to 1.290 and slope from -0.216 to -0.214. To further investigate, we decide to manually change some values of densities in order to make them as "errors". In the original density, we changed the third entry of density from 0.686 to 0.8, and the 63rd entry from 0.148 to 0.4 in order to see how the fits will be affected if there is a relatively large error in the response variable.

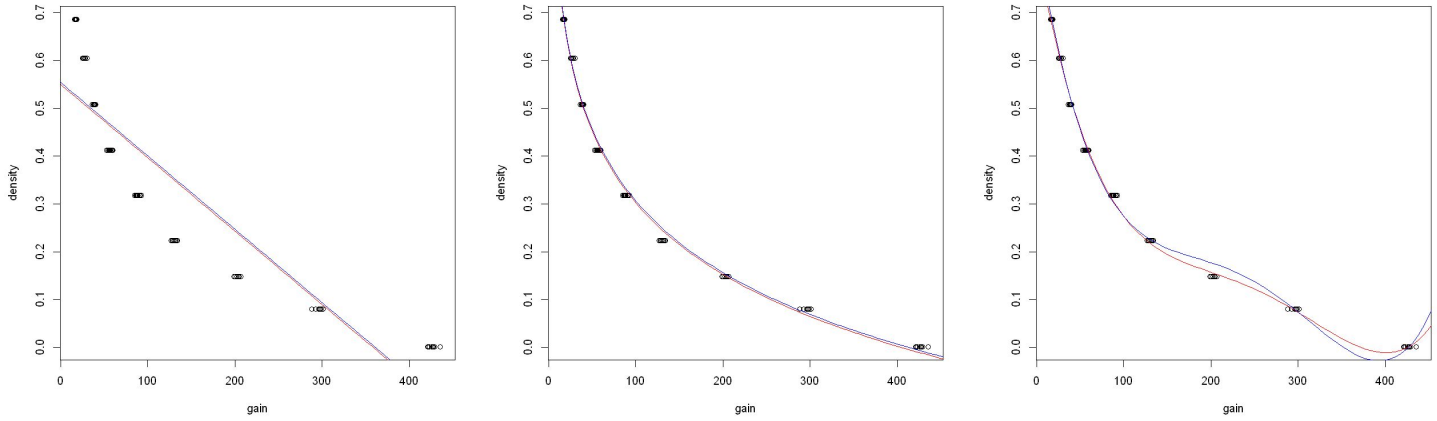


Figure 4: Comparisons between Original Fit and Error Fit for Three Models (Linear, Log, Polynomial)

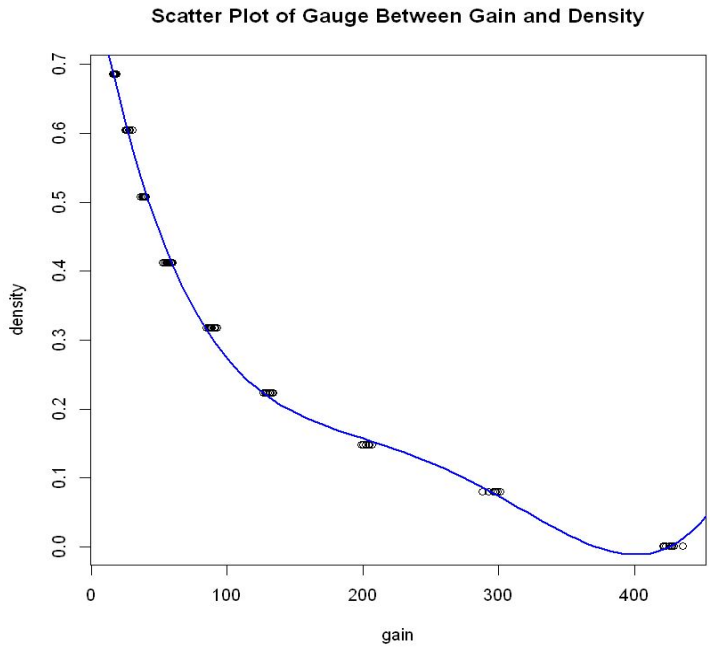
For each model, derive a new model by fitting on the data with some error in the densities, then use both the original and the new model to predict the original data. Figure 4 shows the results after plotting both models on the original data, where the red line is the old fit and the blue line is the new fit on errors. It is clear to conclude that both simple linear regression and regression with log transformation are relatively robust and fitting well on the trend of data; however, the polynomial regression model is slightly deviating from the trend since it also considers the significance of errors (random noises).

In conclusion, by inspecting both the satisfaction of conditions in least-squares and the fit on the trend of data, we determine to use the polynomial regression model with a degree of four as our final model. This model meets all the conditions of the least-squares thus we expect it will

produce the best performance in predicting data among these three models mentioned above. In addition, although it has the potentiality of overfitting, its complexity makes it flexible enough to refit the data.

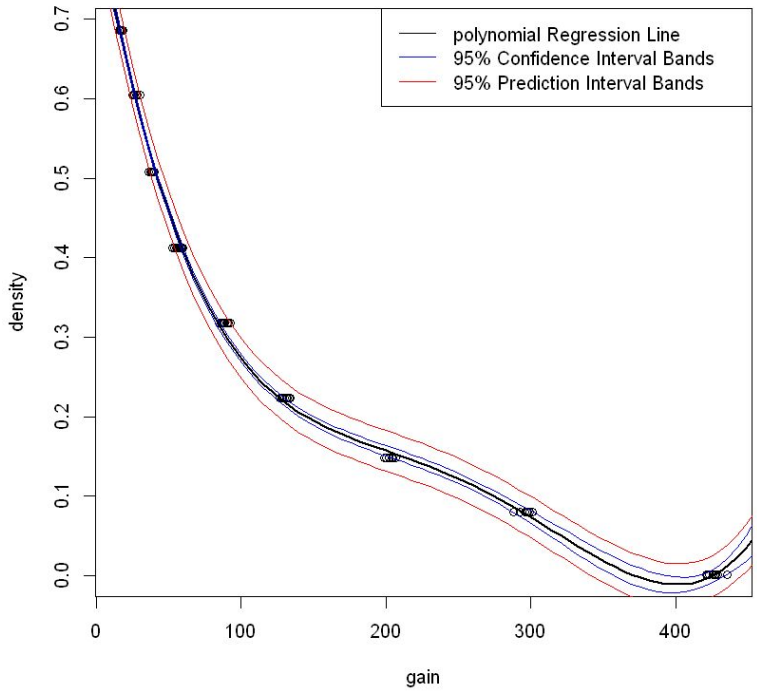
B. Predicting

Based on the all the linear regression model we fit in the last part, we decided to use a polynomial regression model to make predictions.



The figure above is the polynomial regression line and the scatter plot of . This line will give us the predicted density with giver gain. For example, the predicted density of gain=38.6 is: 0.523250800770545, the predicted density of gain=426.7 is: 0.000665083971871816

In order to have a interval estimate of density by given gain , we did two interval estimate here: confidence interval and prediction interval and we think prediction interval is a better : As you can see from the graph below :

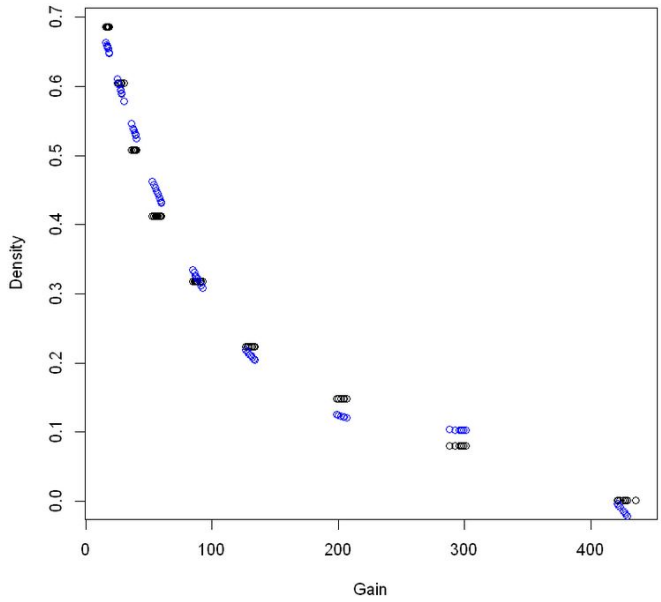


From this figure, we can notice that the prediction interval is much wider than the confidence interval. This is because the confidence interval did not include the variability of the density, instead it tells us the estimated mean of density by given gain. It should be noticed that the prediction and confidence intervals are similar in that they are both predicting a response, however, they differ in what is being represented and interpreted. Confidence interval is good for predicting a random variable. The prediction interval is a good estimator for an observation.

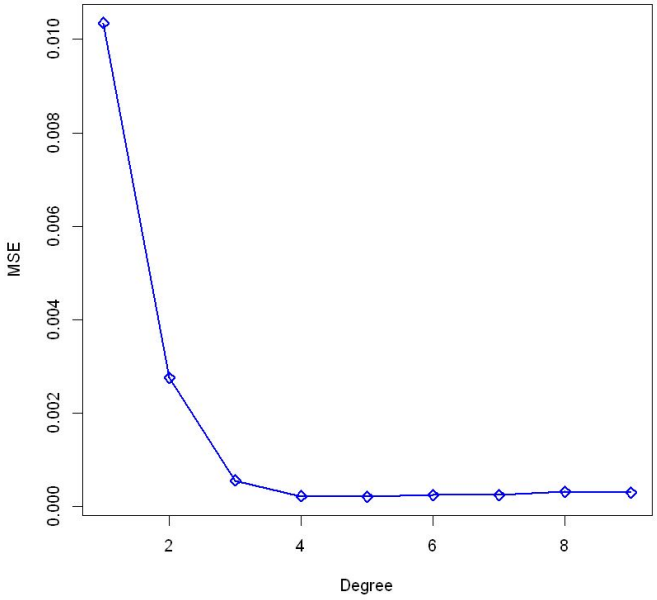
In conclusion, since the predictor is the gain which is a set of observed values with known density, we have to take into account the variability of the predictor which is gain.

C. Advanced Analysis

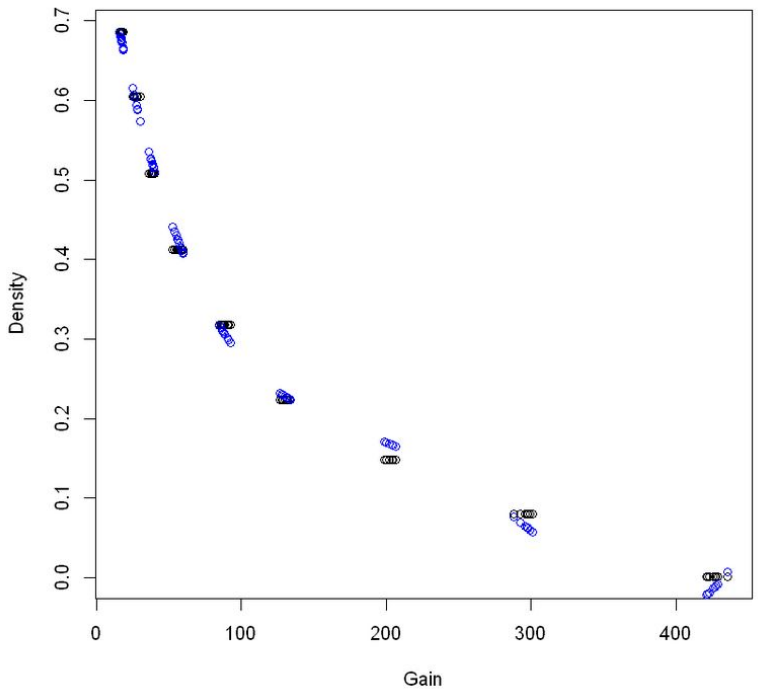
For this part, we planned to use a new model for the data and find the best parameter for the model to fit data. The first fit, as shown below, uses a support vector machine with polynomial kernel and a degree of 3.



It fits the data well, presenting a roughly the same shape as the data, but we want to improve this model. Therefore, we create a function which calculates the mean squared validation error to evaluate the model. We use k-fold technique to split the data into training sets and validation sets and calculate MSE for each fold.



From the graph, we can see that the MSE comes to its local minima when degree = 4. After 4, the validation error starts increasing. Using degree = 4, we can see the model fits data more accurately.



III. Importance and Conclusion

Our group successfully constructed a procedure for converting gain into density by creating linear models and predicting.

For the fitting part, we tried a simple linear regression model, a linear regression model with log transformation and a polynomial regression model. We found the polynomial regression model to be the best model by evaluating the fit error for all three models.

For the predicting part, we made some predictions and think the interval estimate is also needed here. Finally , we chose the prediction interval estimate since we need to take the variability of gain into consideration.

We also discovered a new model, support vector machine, in the advanced analysis part. We found that svm with degree = 4 has the lowest validation error, also this error is less than the polynomial regression model we created in the fitting part.

The limitation of our project is that it could not be generalized to data other than the present dataset since our research focuses on the given dataset with only 90 observations.

IV. Appendix: see the code file

V. Contribution Statement

Yunlin worked on the introduction and fitting.
Jian worked on advanced analysis, k-fold validation tests and conclusions.
Yong worked on the prediction part.