Math 439 Final

Andres Pedro

12/11/2021

Executive Summary For this report we were given two datasets to tackle two problems. Problem 1 dealt with gravity data where we aimed to find the best model possible and we verified that the equation hypothesized by the scientist was justifiable. As for problem 2, using the best possible model with an R^2 value of 95%, we wished to predict a specific case scenario given a set of values. We found that the expected magnitude of an earthquake based on the given parameters would be 4.594. It is also important to note that for each problem, we did not remove any data points due to insufficient evidence of a useless high leverage outlier.

Problem 1

Background

For problem one we are given the gravity data set which contains 200 observations of 2 variables, t and P.Here t represents time and P represents position. In this case, we are dealing with very simple data and we wish to construct the best possible model with the idea of comparing our model to the equation given by the scientist to verify if their assumptions were justifiable.

Model

Our first step into finding our best model is to get a feel for what our data looks like. We will begin by visualizing our data to see its shape using a scatter plot.

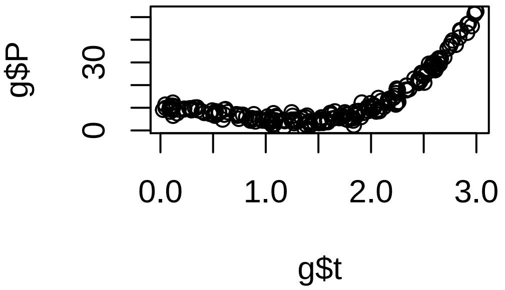

Figure 1: Scatter Plot of t vs P

If we refer to Figure 1, we can see that if we were to follow a linear regression model, we would not get a proper fit. Based on this initial plot, we would guess that our model is either quadratic or cubic especially when considering the formula given by the theoretical physicists. To see this more clearly, we will look at the residual diagnostics if we were to follow a simple linear regression model..

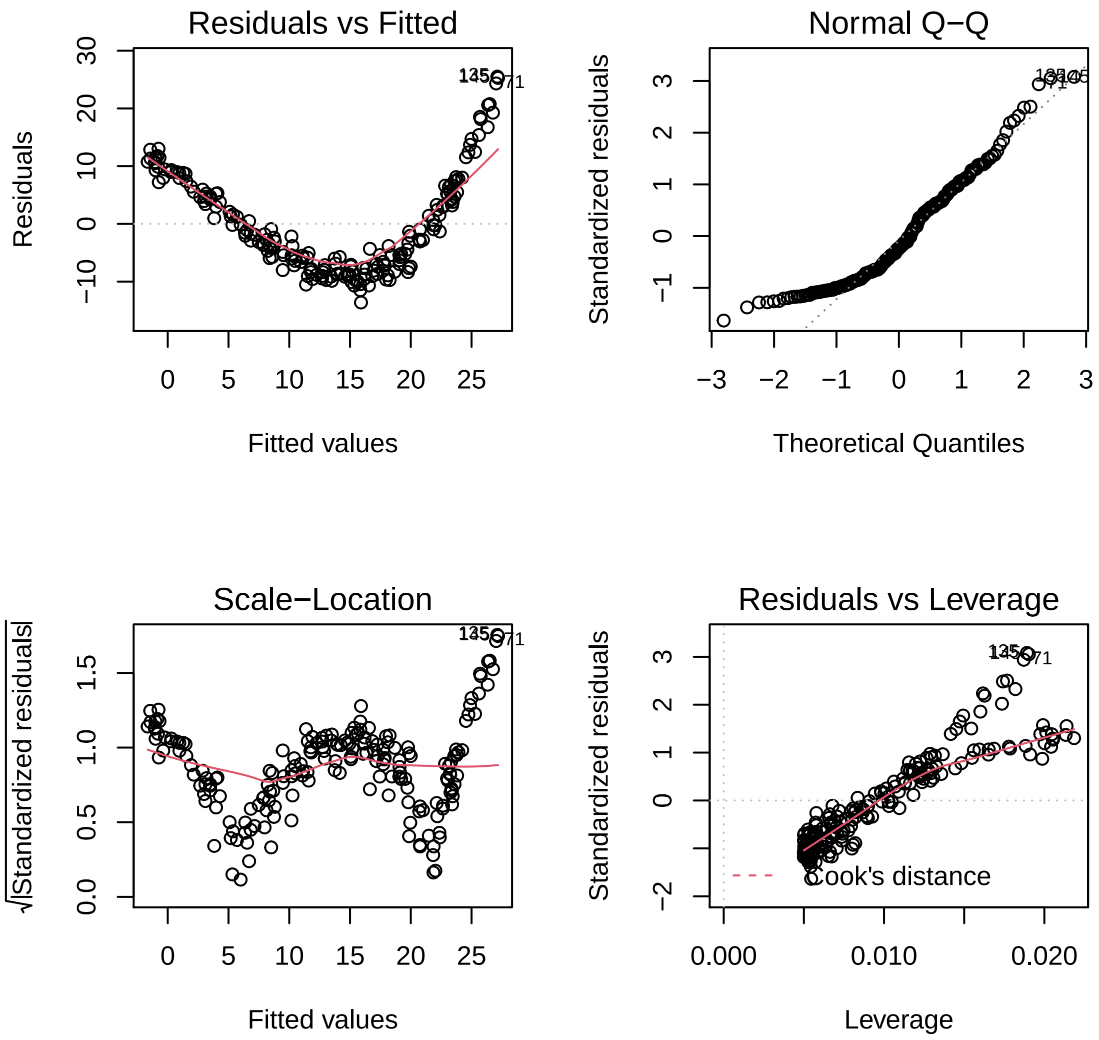

Figure 2: Linear Regression Residual Diagnostics

Taking a look at Figure 2, our graph on the top right clearly shows us that our residuals of our fitted values are not centered around zero. Clearly, our linear model is not the right model for our data. The rest of our diagnostic plots show that based on our model, we do not have constant variance due to our model being highly unaccurate. There are a few ways to approach finding the best model. We can take a look at transformations on x following the equation that is given by theoretical physicists, where we can introduce t^2 and t^3 values. to our model. Additionally, we can use kernel regression. If we go with a kernel regression model, we can visualize how accurate our model fits on our data; however, we will not have an objective measure of fit. On the other hand, if we go with our additional predictors model, we will obtain an interpretable model and an objective measure of fit.

Figure 3 shows us our residual diagnostics of our model $P = \beta_0 + \beta_1 t^2 + \beta_2 t^3$. According to our top left

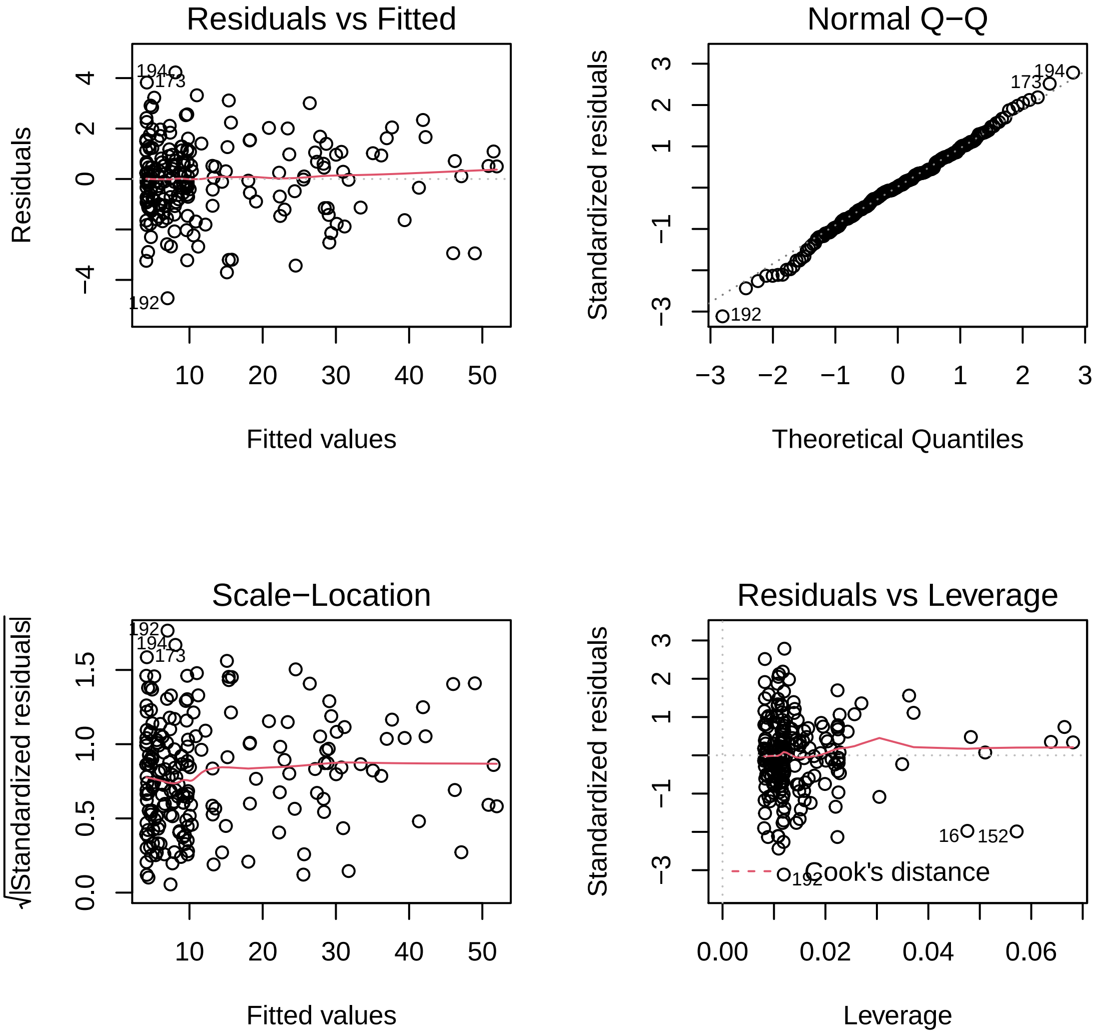

Figure 3: Residual Diagnostics

plot, we can see that the residuals of our fitted values are generally centered around 0. Additionally, if we compare Figure 2 and Figure 3, we can see that our new model is properly centered compared to the residual diagnostics of a normal linear regression model. Most importantly,we can see that the assumptions of the scientists were justifiable. We can see from our QQ plot on the top right that our errors are normally distributed. Our bottom left plot shows that we have constant variance. To verify our model even further, we perform cross validation using an 80/20 split of randomly selected data. After creating our model using our training, we were able to obtain our predicted response based on our testing data. With these predictions, we were able to calculate our objective measure of fit R^2 which worked it to be 98.4%. With all the proof that we have chosen the best model, we can finally interpret our model $P = 9.77 - 9.51t^2 + 4.73t^3$. As part of our interpretation, we took the liberty of creating a 95% confidence interval for our coefficients with the exception of leaving our constant number 9.77. Therefore, we are 95% confident that on average, a one unit increase in t^2 will result in a -8.98 to -10.05 unit decrease in P. Similarly, a one unit increase in t^3 on average will result in a 4.55 to 4.92 unit increase in P. With all this in mind, when comparing our model to the equation given, we can see that the scientists assumptions are justifiable. The coefficients of the scientists' model fall within our calculated 95% confidence interval with the only difference being the constant β_0.

Problem 2

Background

For this problem we are given the earthquake data which contains 316 observations of 5 variables. We have, depth, angle and length and magnitude as numerical variables. Then we have a categorical variable volcano where 1 indicates that the earthquake is in close to proximity to a volcano whereas a 0 indicates that the earthquake was not in close proximity to a volcano. We are looking for the best fitting model to predict the magnitude of an earthquake based on the information given about a neighboring fault in order to predict the magnitude of an earthquake in the future.

Model

For this dataset we are dealing with multiple predictors, so we will need to be more mindful of what model we choose. We know a simple linear regression will not be possible. To get a better understanding of our data, we want to plot our response variable magnitude against some of our predictors to gain some perspective using a scatter plot in hopes of finding some relation whether linear, polynomial, logarithmic,etc.

Figure 4 displays two predictors that stood out since our predictors against our response seemed to take shape. Although it is still not clear, it does seem to be somewhat linear. Predictors fault_length and volcano were more complex in there shapes since fault_length looked more scattered and volcano is a categorical variable. With that being said, we will attempt a linear regression model. If we go based on a simple linear model, we obtain some results with an adjusted R^2 value of about 89%. That sounds great enough as it is; however, it is important to take a look at the residual diagnostics plot to see what is going on.

If we take a look at Figure 5, we can see that although our model is accurate, it is still not the best possible model. Looking at the top left plot, we can see that our residuals are not centered around 0. Our QQ plot on the top right is doing well and our bottom left plot indicates that we do not have constant variance. In order to compensate for this, we transformed some of our predictors and found that the best possible model contained a transformation of fault_depth by turning it into a quadratic. Thus, we obtain a model of the form $magnitude = \beta_0 + \beta_1 fault_depth^2 + \beta_2 fault_length + \beta_3 fault_angle + \beta_4(volcano = 1)$.

Figure 6 displays the residual diagnostics of our potentially best model. Our Residuals vs. Fitted plot shows a huge improvement indicating that our we have chosen a great model with the residuals being centered very close to 0. Our plot on the bottom left seems to follow our red line which indicates constant variance; furthermore, we have now achieved an adjusted R^2 value of about 95% which is a great improvement from our original model. With that being said, we can finally take a look at potentially making predictions with

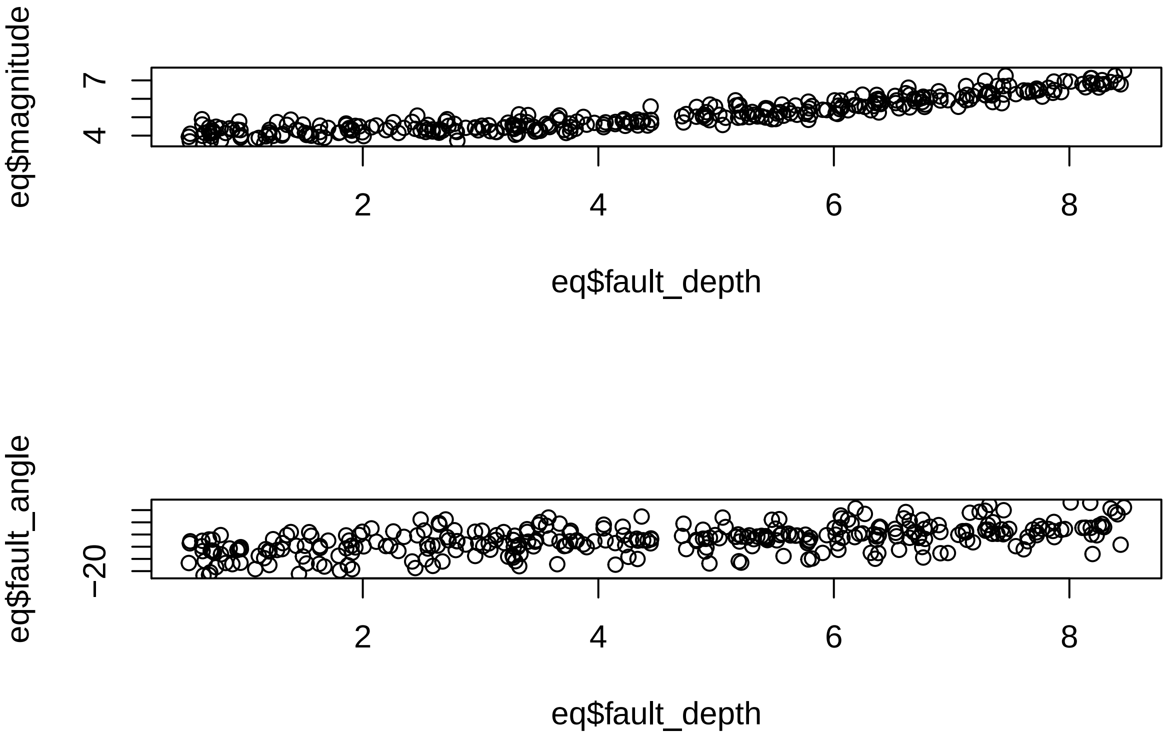

Figure 4: Scatter Plot of Magnitude vs Predictors

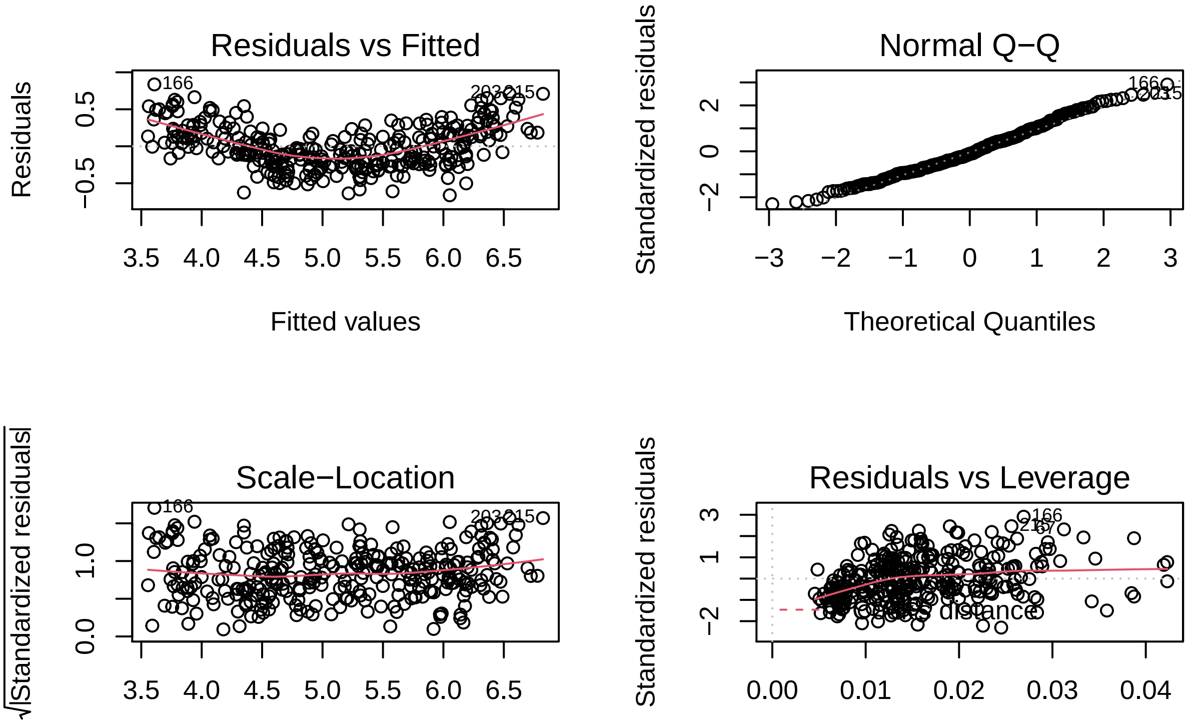

Figure 5: Residual Diagnostics of Multiple Linear Regression

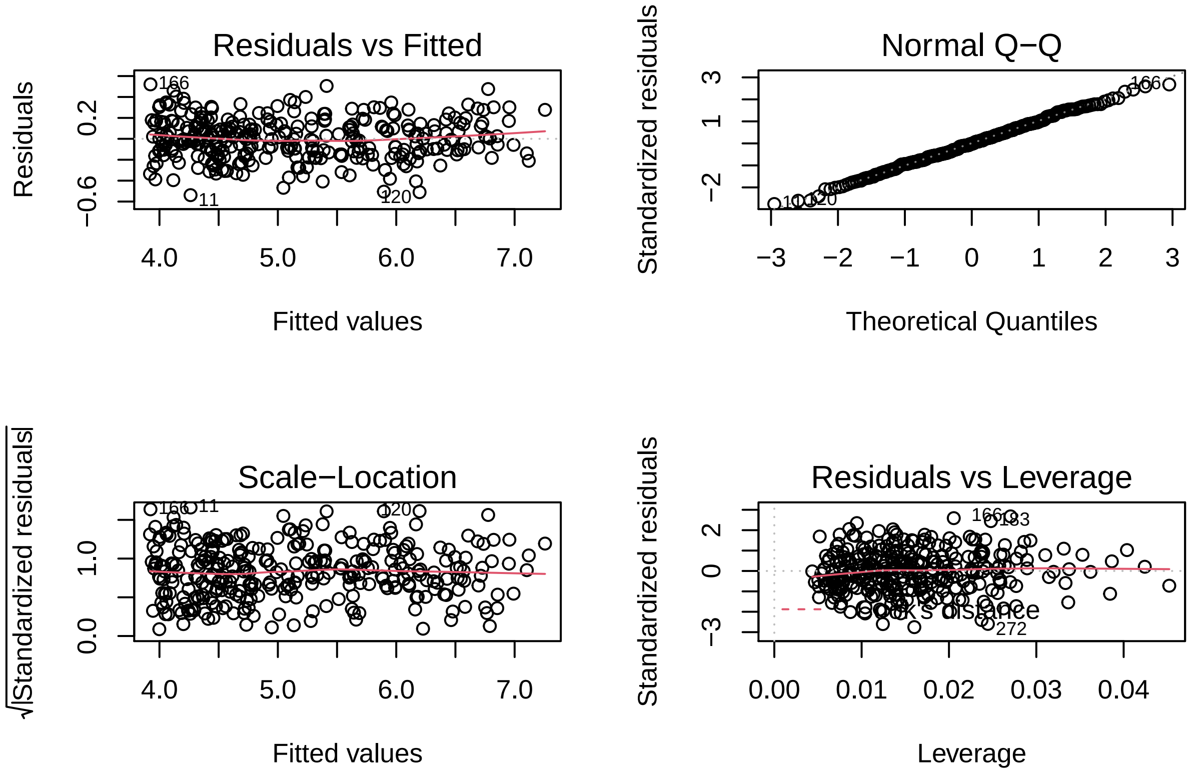

Figure 6: Residual Diagnostics of Best Model

our model $magnitude=3.94+0.039fault_depthˆ2+0.004fault_length+0.001fault_angle-0.037(volcano=1). Thus, we opted for regression trees to properly categorize our model in a clean way.

Prediction

Now that we have found our best model, we can tackle our challenge of making a prediction based on this model; however, to get a better understanding of what is going on visually, we opted for regression trees. There are many tuning parameters to take into consideration, so after messing with different parameters, we decided to go with a cp value of 0.011 after looking at the associated errors and R^2 values of different cp values.Finally, we sought to optimize our tree further to prevent overfitting from occurring with the result being Figure 7.

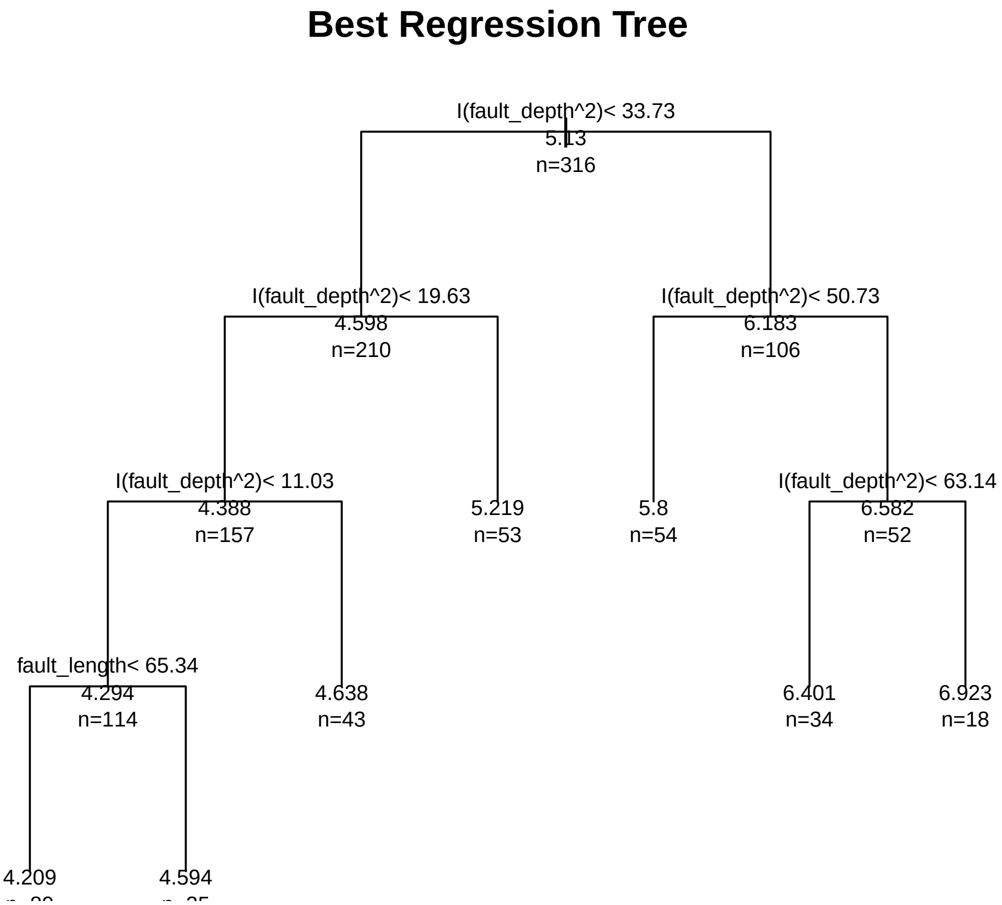

Figure 7: Best Regression Tree

Figure 7 clearly shows how a change in our predictor $faultdepth^2$ based on a range of values will affect our magnitude. If we were to look at the specific case scenario of someone buying a summer home that is located near a fault that is 1.6 miles deep and 67 miles long with an angle of 13 degrees in close proximity to a volcano, we are very confident we can predict the magnitude of the earthquake. We first note that we will not be basing this off of averages, for we are looking at this specific case. Our model predicted that based on these parameters, the tenant can expect to experience an earthquake of magnitude 4.594. If we follow the regression tree in Figure 7, we can see that the square of our fault_depth will be less than 33.73, so we can traverse the left nodes until we reach the bottom left nodes where fault_depth<11.03; however, we know that our fault length will be 67 miles. Following our bottom branch, we know fault_length=67> 65.34, so according to our regression tree, we are confident that our magnitude will be anywhere from 4.294 and 4.594. Our predicted value falls within this range despite it being at the very upper limit.

Conclusion

Overall, for this report we tackled 2 problems where we had similar goals to achieve the best model; however, what we did with our models differed greatly. In Problem 1, we used our best model which was a linear regression model with the transformation of our predictor for the purpose of inference. We found that the assumptions of the scientists was reasonable using our residual diagnostics. For problem 2, we used our best model with the goal of predicting the magnitude of an earthquake near a house with given parameters. In this case, we found our best model with an adjusted R^2 value of about 95% and visualized our model using a regression tree. In that case we found that we could expect the magnitude to be around 4.594.