Adrien Leon

Professor Nichols

MATH 439

18 December 2022

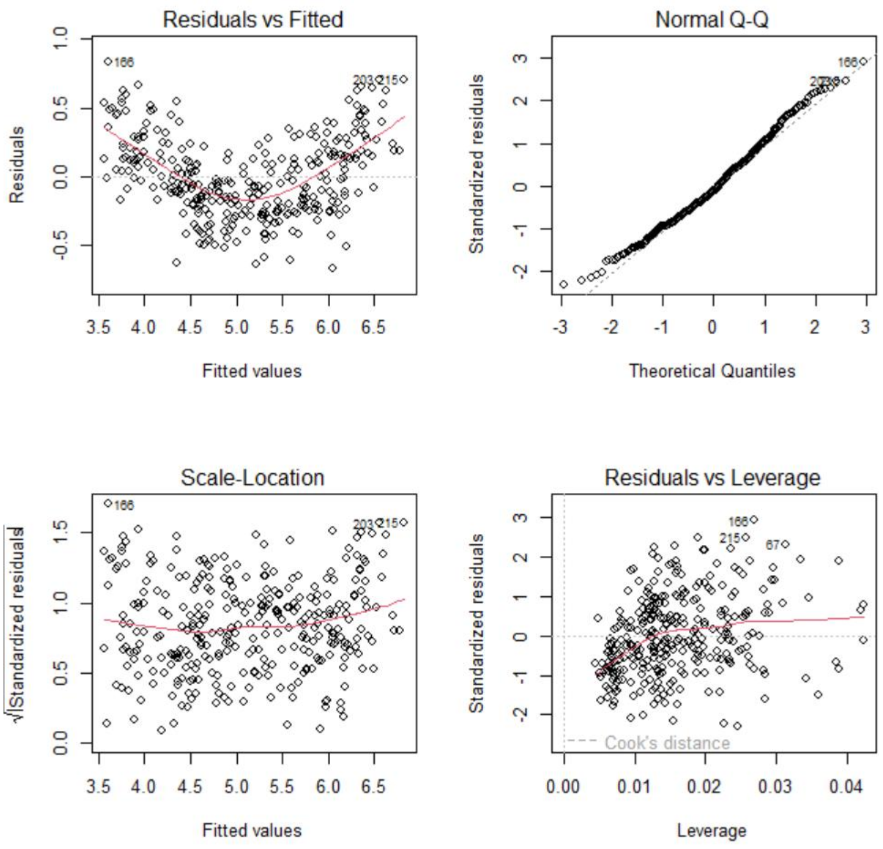
Finals Problem 2

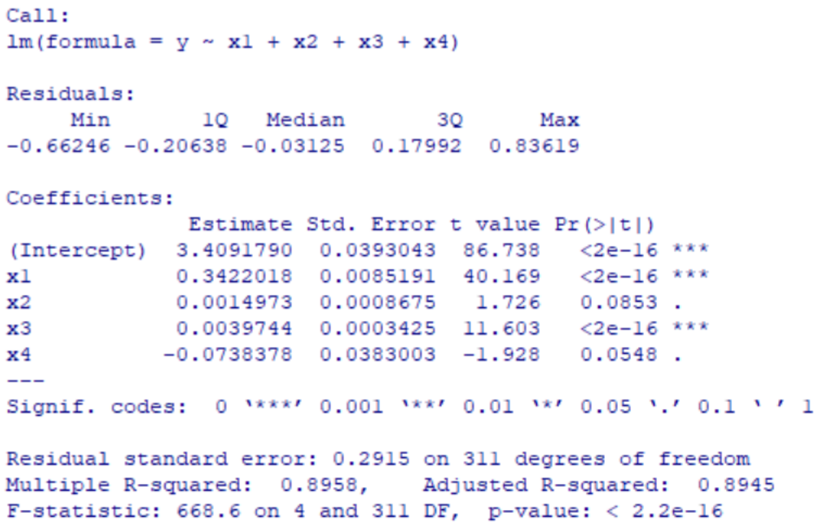
Introduction:

Our data set consists of a single numerical response (magnitude), three numerical predictors (fault depth, fault angle, and fault length), and one categorical predictor on whether a volcano was close in proximity to when an earthquake occurred. We'll determine the best fitting model for predicting earthquake magnitude by utilizing a linear model, random forest, and gbm and compare them all to see which one is the best fitting model to predict future earthquakes. In the end, we'll gain a better understanding of which model is the best fitting for our prediction by showcasing a scenario to demonstrate what a house's predicted magnitude would be if an earthquake occurred near them.

Body:

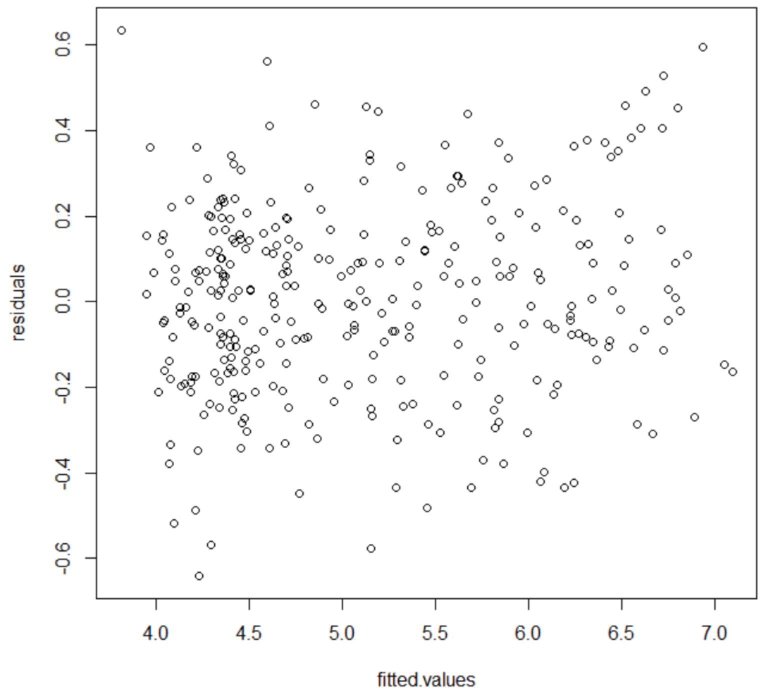
First, I utilized a linear model and checked its residual diagnostics to get an idea for what we're dealing with. This will be our baseline model as we go over other models to compare for which one is the best fit for predicting earthquakes.

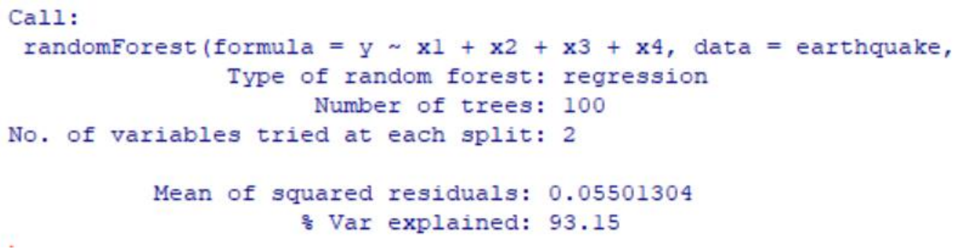




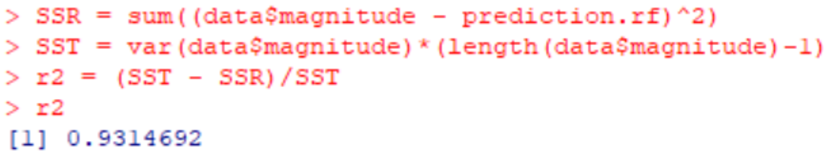
```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.66246 -0.20638 -0.03125  0.17992  0.83619 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.4091790  0.0393043  86.738   <2e-16 ***
x1           0.3422018  0.0085191  40.169   <2e-16 ***
x2           0.0014973  0.0008675   1.726   0.0853 .  
x3           0.0039744  0.0003425  11.603   <2e-16 ***
x4          -0.0738378  0.0383003  -1.928   0.0548 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2915 on 311 degrees of freedom
Multiple R-squared:  0.8958,	Adjusted R-squared:  0.8945 
F-statistic: 668.6 on 4 and 311 DF,  p-value: < 2.2e-16
```

As we can see, from our residual diagnostics and R^2 = 0.8945 for our linear model, so we can make this our baseline to compare to other models that we’ll go over. Next, we’ll consider Random Forest to see whether this is a better predicting model than a linear one for determining future earthquakes for our provided data set.



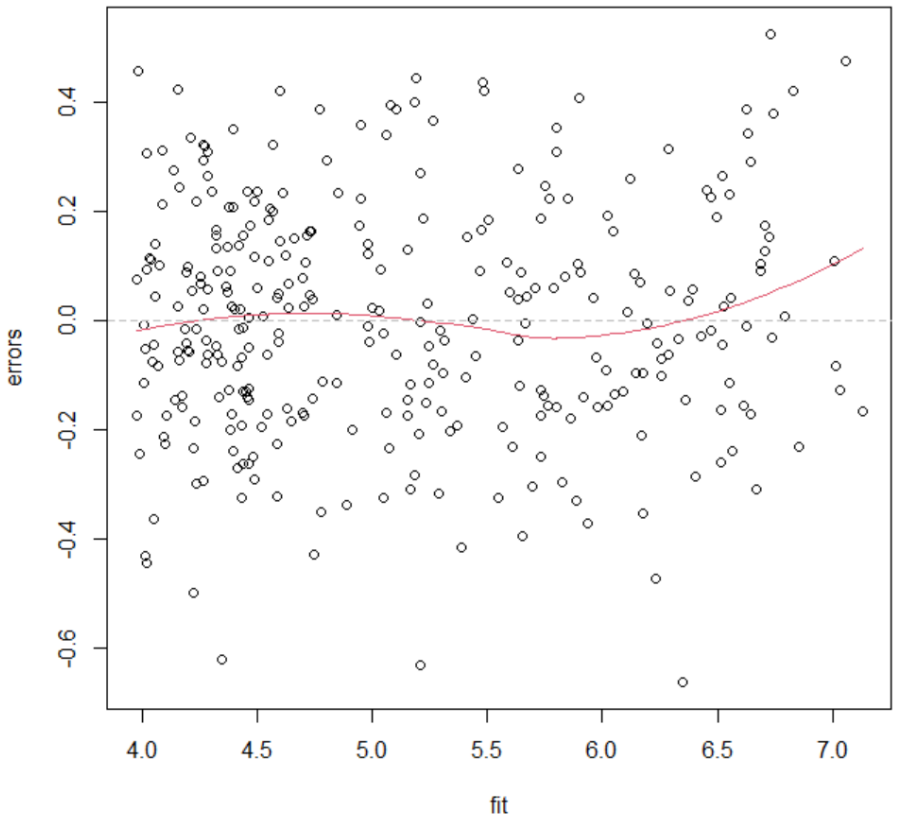


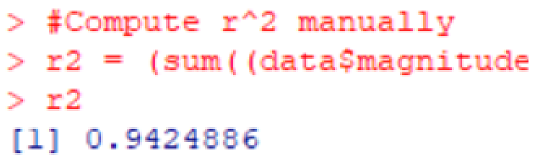
```
Call:
 randomForest(formula = y ~ x1 + x2 + x3 + x4, data = earthquake,
               Type of random forest: regression
                     Number of trees: 100
No. of variables tried at each split: 2

          Mean of squared residuals: 0.05501304
                    % Var explained: 93.15
```



```
> SSR = sum((data$magnitude - prediction.rf)^2)
> SST = var(data$magnitude)*(length(data$magnitude)-1)
> r2 = (SST - SSR)/SST
> r2
[1] 0.9314692
```

From the data above, we can see our R^2 value for a random forest is around 0.931 making it a better fitting model than what our linear model had. This tells us that 93.15% of the variance of the dependent variable being studied is explained by the variance of the independent variable. Meaning that the strength of the relationship between this model and the dependent variables are better than what we calculated for our linear model. Hence, a random forest is our best predicting model so far. Now, we'll try gradient boosting and determine if that model is a better predicting for measuring earthquake magnitude.



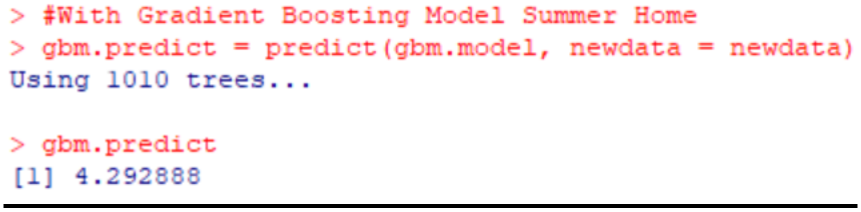


```
> #Compute r^2 manually
> r2 = (sum((data$magnitude
> r2
[1] 0.9424886
```

The residual plot demonstrates the difference between the observed response and the fitted response we used for our model. From the above data and residual plot, we utilized gradient boosting trees to make future predictions of earthquake magnitude with our data and the predictors provided. We can calculate the overall R^2 from this data and see that it’s around 0.9424, representing the proportion of the variance from the dependent variable (magnitude) is supported by the independent variables (fault depth, fault angle, fault length, volcano nearby) in a regression model. Thus, making the gradient boosting model the best predictor that we’ve used so far.

Analysis (Summer Home):

Now, we’ll use our “best model” to predict a scenario where a Summer home is in close proximity to a volcano and near a fault that is 1.6 miles deep and 67 miles long. The angle between the fault and the house is 13 degrees and from our gbm model, we can predict that if an earthquake were to occur near this house, the expected magnitude of this earthquake would be around 4.2929. Utilizing the trees and gradient boosting model, this is the best estimate we came up with for our best fit model for this data set.



```
> #With Gradient Boosting Model Summer Home
> gbm.predict = predict(gbm.model, newdata = newdata)
Using 1010 trees...

> gbm.predict
[1] 4.292888
```

Conclusion:

Overall, we notice that the best model out of the three that we’ve shown so far would be the gradient boosting one where its R^2 was 0.9424886, so it would be considered the “best model” to make our predictions from. This would confidently give us a better understanding of future risks of earthquakes than the other two models that we described between having a linear model and random forest to predict our data. From this, we can see that our final model would be gradient boosting.