California Polytechnic State University San Luis Obispo

Wildfire Size Regression Model

Michael Earl, Olivia Wallin, Claire Freeman, Nathan Stanley

Stat 324

Executive Summary

After careful analysis of these data, we determined the most useful regression equation to be Log(area) = 0.9630 - 0.01763 ISI based on the F-test of model utility being F =3.30 with a p-value of 0.07. We also settled on this model as it had a F-statistic of 1.22 for the lack-of-fit test so we were able to conclude that linear regression was appropriate in this case. We transformed the response variable, area, with a log base 10 according to the Box Cox ladder of powers to correct for the violation of the normal errors assumption in the original data seen in figure 1 in the appendix. We also excluded wildfires with an area of zero, reasoning that these values skewed our dataset and further violated the normal errors assumption as 247 out of 517 of the observed fires had an area of 0 hectares. Our predictor, ISI stands for initial spread index and is a unitless term and therefore it does not have an intuitive interpretation like the other variables would. This was part of the reason we initially resisted using this variable. It is "a numerical rating of fire spread immediately after ignition"[1]. When we back transform our regression equation it took the form Area =$10^{0.9630}10^{-0.01763}$. Since $b_1 < 0$ we can subtract the 10 to b_1 from one and multiply it by 100% to get 3.978% in order to interpret the area in its original units. The interpretation is as follows: Each additional unit increase in initial spread index is associated with a 3.978% decrease in the predicted value of wildfire area in hectares. This interpretation does not make sense as one would expect that as ISI increases, indicating a faster rate of spread, so would the predicted area for a wildfire. While this model is statistically significant, it leaves much to be desired. From what we know about fires, one would expect many of the variables included in this dataset, i.e. wind and rain, to be significant predictors of wildfire area. Yet, the analysis tells us there is not significant relationship between these variables and wildfire area.

[1] https://fire.ak.blm.gov/predsvcs/fuelfire/fwidefined.php

Our best explanation for this lack of a relationship is there is an unaccounted-for variable. We theorize that location of the fire could be messing up our data. If the fire were to break out in an easily accessible part of the forest, it would be easy for firefighters to extinguish the fire and the variables we analyzed would be rendered useless.

Methods and Findings

In this dataset, our response variable is area, measured in hectares. Our explanatory variables are month, day, FFMC, DMC, DC, ISI, temperature, relative humidity, wind and rain. Month is a qualitative predictor that ranges from January to December. Day is also a qualitative predictor that ranges from Monday to Sunday. Fine fuel moisture code (FFMC), duff moisture code (DMC), DC (Drought Code) and initial spread index (ISI) are quantitative predictors that are calculated based on the temperature, relative humidity, wind and rain. Calculations for each of these predictors can be found in the appendix. Temperature is measured in degrees Celsius, relative humidity is measured in percentage points, wind is measured in kilometers per hour and rain is measured in millimeters. Temperature, relative humidity, wind and rain are all recorded at noon on any given day.

In order to get a feel for the data, our first step was rather naive and consisted of us trying to fit the model with the response variable area and every explanatory variable. We then looked at the F-test of model utility which had a small value of 0.94 and a corresponding p-value of 0.498 which meant that we could not conclude that all the predictors were collectively significantly useful for predicting area. Furthermore, our $R_{adj}^2 = 0.00\%$ which means we were explaining no variation in the data with our model. One interesting thing that we did observe while fitting this model was that even though FFMC, DMC, DC and ISI were calculated using combinations of temperature, relative humidity, wind and rain, there was no evidence of multicollinearity in our initial model as evidenced by the largest VIF (for DC) being relatively small at only 7.91. From here we decided to utilize the best subsets procedure instead of randomly choosing variables to save time. We also made the decision to not use FFMC, DMC, DC and ISI as predictors in an attempt to simplify our models. We did not include the qualitative predictors in the best subsets procedure as that created too many issues instead opting to manually fit each qualitative predictor to the identified candidate models.

Through the best subsets procedure, we identified two candidate models. The first candidate model used only temperature as a predictor. We selected this model because it had the largest R_{adj}^2 which was a pitiful 0.77%. We also selected it because it had a Mallows Cp of $C_2 < 2$. This model also had the smallest residual standard error at 63.412. The second model we chose used temperature and wind as a predictor. We selected this model because it had the second largest R_{adj}^2 which 0.70% , a Mallows Cp of $C_3 < 3$. This model also had the second smallest residual standard deviation at 63.433. We should take the time to note that neither R_{adj}^2 or the residual standard error exhibited much difference between the models selected in the best subsets procedure. R_{adj}^2 ranged from a low of 0.40% to a high of 0.77% and the residual standard error ranged from a low of 63.412 to a high of 63.530. Also, all C_p's were less than the respective p's.

Despite the small R_{adj}^2, took our first candidate model that used only temperature and fit it along with day. We found that the F-test of model utility for this model was slightly larger than the model in which we used every possible predictor. The F-statistic was 1.36 and had a corresponding p-value of 0.220. We were unable to conclude that day and temperature are collectively significantly useful for predicting wildfire areas. This model had a R_{adj}^2 equal to 0.00%. We then tried to fit the model using temperature and month as predictors and found that this model had an F-test of model utility that was even smaller than the model we fit that used every possible predictor with F = 0.64 with a p-value of 0.805. Once again, we found that we could not conclude that month and day are collectively significantly useful for predicting wildfire areas. This model had an R^2_{adj} equal to 0.49%. There was evidence of multicollinearity in this model with the months August and September having a large VIF of 17.57 and 16.25 respectively.

At this point we decided to try our luck with our second candidate model and fit a model that used temperature and wind and day as predictors. We found that the F-test of model utility for this model was F = 1.32 with a corresponding p-value of 0.230 which told us that we could not conclude that temperature, wind and day are collectively significantly useful for predicting wildfire areas. The R^2_{adj} for this model was equal to 0.50%. We then fit the model using temperature, wind and month as predictors. We found that the F-test of model utility was an abysmal F = 0.66 with a p-value of 0.805. This meant that we could not conclude that temperature, wind and month are collectively significantly useful for predicting wildfire areas. There was evidence of multicollinearity in this model with the months August and September having a large VIF of 18.18 and 17.15 respectively.

We wanted to see whether we should drop the wind and the month predictors from our model so we performed the partial F-test. The calculation was as follows:

$$\frac{2070848 - 2055945}{14 - 2} \% \frac{2055945}{517 - 2} = 0.311$$

The F-statistic had a p-value from an F distribution with df_{num}=12 and df_{denom} = 515 that equaled 1-0.0125 = 0.9875. This means that we cannot conclude, at the 0.05 level, that month and wind significantly contribute to the prediction of wildfire area when temperature is already in the model. We concluded that dropping month and wind from the model is most appropriate. We redid the test but this time with day instead of month. The calculation was as follows:

$$\frac{2070848 - 2048235}{9-2} \% \frac{2048235}{517-2} = 0.812$$

The F-statistic had a p-value from an F distribution with df_{num}=12 and df_{denom} = 515 that equaled 1 - 0.3618 = 0.6382. This means that we cannot conclude, at the 0.05 level, that day and wind significantly contribute to the prediction of wildfire area when temperature is already in the model and should drop day and wind from the model.

Knowing that we should drop both qualitative predictors and wind from our model we decided to fit the model using only temperature as a predictor for wildfire area. Temperature was a significant predictor in this model with a t-test of 2.23 and a p-value of 0.026. R^2 in this model was 0.96%, the highest we have seen out of any model up to this point. The lack of fit test was large with an F-statistic of 5.31 and a small p-value of <0.001 so we concluded that linear regression was not appropriate in this case. We proceeded to see if adding a quadratic term would improve our model. It did not help our model, and instead increased the value of the F-statistic for the lack of fit test to F= 5.33. We went back to our model using only temperature as a predictor and examined the 4-in-1 plot (figure 1 in index) and concluded that normality was violated since the residuals did not fall on the straight line of the normal probability plot. Because of this, we decided that it would be appropriate to perform a transformation on area. We decided to perform a log transformation based on the box-cox transformations. There were many wildfire areas recorded as zero hectares in the dataset, presumably small fires that rounded down to 0 hectares. To prevent undefined values, we added 0.01 to every area measurement before transforming. We then redid the best subsets procedure and identified 2 candidate models. The first candidate model used temperature and wind as a predictor. We selected this model because it had the largest R_{adj}^2 equal to 0.70%. We also selected it because it had a Mallows Cp of $C_3 < 3$.

This model also had the smallest residual standard error at 1.4752. The second model we chose used temperature wind and rain as predictors. We selected this model because it had the second largest R_{adj}^2 which was 0.54% , a Mallows Cp of $C_4 < 4$. This model also had the second smallest residual standard deviation at 1.4765.

We fit the first candidate model using temperature and wind along with the day. We found the F-test of module utility was 0.86 with a p-value of 0.008 so we did not further investigate this model. We found the F-test of model utility was 2.22 with a p-value of 0.008. This small p-value meant that we were able to conclude that temperature, wind and month were collectively significantly useful for predicting wildfire areas. We were happy to have a model that finally worked that feeling was short lived when we examined the 4 in 1 plot (figure 2 in index) and found that there was still a violation with normality since the residuals did not fall on the straight line of the normal probability plot. There also appeared to be a violation of constant error variance and linearity due to the weird line underneath the data points in the residual vs fits. At this point, we decided to remove the fires that had zero area as we believed these observations skewed our data. While it is not the best statistical process, this is an exploratory study and we reasoned that it would be better to manipulate our data a bit and get a model where the LINE conditions are met so we can make predictions and create confidence intervals.

With the fires with zero area removed, we performed a log transformation on area (not adding 0.01) and did the best subsets procedure. This yielded some interesting results with the R-squared adjusted equaling 0 for every candidate model. Things were going so poorly we decided to reintroduce FFMC, DMC, DC and ISI in a last-ditch effort to get a usable model. This seemed to work and we identified the candidate model that used DMC ISI and wind. We fit this model and found that the F-test of model utility was 2.07 with a p-value of 0.105. This large p-value meant that we were not able to conclude that DMC, ISI and wind were collectively significantly useful for predicting the log of wildfire area. ISI had a partial t-test of -2.24 with a p-value of 0.0261 so we performed the partial F-test to see if we could drop wind and DMC from the model. The calculations were as follows:

$$\frac{116.822 - 115.565}{4 - 2} \% \frac{115.565}{270 - 2} = 1.458$$

The F-statistic had a p-value from an F distribution with $\mathrm{df}_{\mathrm{num}}$=2 and $\mathrm{df}_{\mathrm{denom}}$ = 268 1 - 0.7655 = 0.2345. This means that we cannot conclude, at the 0.05 level, that DMC and wind significantly contribute to the prediction of wildfire area when ISI is already in the model and we should drop DMC and wind from the model.

Since we concluded that DMC and wind should be dropped from the model we proceeded to fit the model with only ISI as a predictor for log of area. We found that the F-test of model utility was 3.30 with a p-value of 0.07 so we concluded that, at the 0.05 level, ISI is a significant predictor of log wildfire area. We then looked at the lack of fit test and it had a value of 1.22 with a p-value of 0.125 so we concluded that linear regression was acceptable in this case. Finally we examined the 4-in-1 plot (figure 3 in appendix). We concluded that the errors were normally distributed since the residuals fell along the line of the normal probability plot. We concluded that linearity wasn't violated because the residuals were randomly scattered around the zero line. We also concluded that constant error variance wasn't violated because the spread of residuals stayed the same as fits increased. However there is now an issue with autocorrelation as seen in the time series plot. We believe that the first order positive autocorrelation is due to the data on fires being collected in the same forest and that the fires are spatially close to one another. The Durbin-Watson statistic was 0.9297. The Durbin-Watson test bounds for p-1 = 1 and n =100, actual n was 270 but table only went to 100, are d_l = 1.65 and d_u

= 1.69. Since D = 0.9297 < d_l = 1.65 we can conclude that first order positive autocorrelation is present in the errors. We will proceed with this model but acknowledge that any intervals may be inaccurate due to the violation of the independent errors assumption. We found that there were a considerable amount of outlying observations. There were a total 20 unusual observations with 15 outliers with respect to log area, 6 outliers with respect to ISI. Only one fire was an outlier with respect to both log area and ISI. This was fire 122 with a log area of -0.76955 (untransformed area was 0.17 hectares) and an ISI of 22.7. Finally we looked for influential observations by examining cooks distance. We got the value F(1 - 0.5 ; 2 ; 270-2) = 0.3929 and F(1 - 0.8 ; 2 ; 270-2) = 0.1811. We then created an index plot of Cook's distances (figure 4 in appendix) in appendix. None of Cook's distances were above either of these values so none of the observations had major, or moderate, influence on the entire fitted regression function.

Now that we had our model, the last thing we did was attempt to predict the size of a wildfire. We found that the ISI ranged from a low of 0.800 to a maximum of 22.700. The first quartile was 6.8, the median was 8.4 and the third quartile was 11.4. We used a random number generator to pick a value from the interquartile range and got 9.3 as our ISI. We plugged this into our model and got a 95% prediction interval of (-0.5033, 2.1013). This is the confidence interval for log base 10 of area so we raised the lower and upper bound to the power of 10. We got the prediction interval of (0.3128, 126.2699). We interpreted this as we are 95% confident that a fire that has an initial spread index of 9.3 will have an area between 0.3138 and 126.2699 hectares. This would be for a fire in a forest somewhere in Portugal.

Appendix

Initial Spread Index Calculation

Fine Fuel Moisture Code Calculation

Drought Code Calculation

Duff Moisture Code Calculation

Figure 1

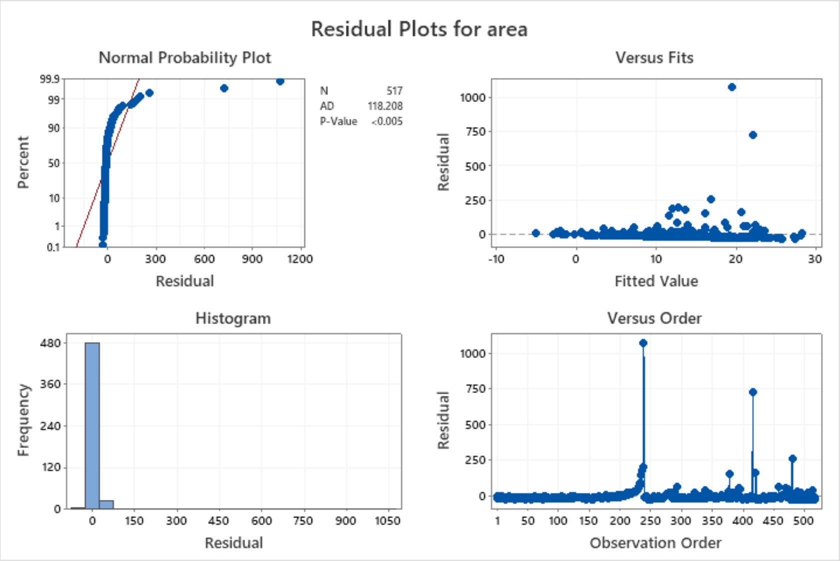


Figure 2

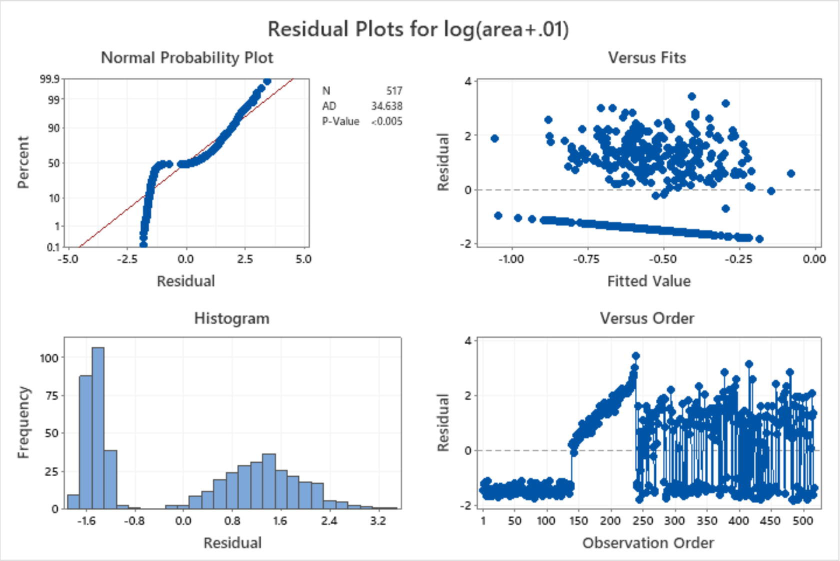


Figure 3

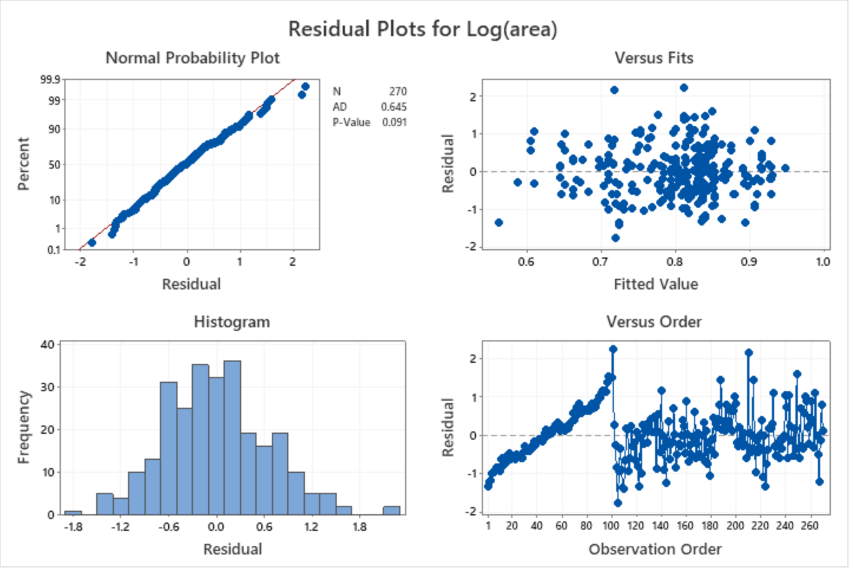
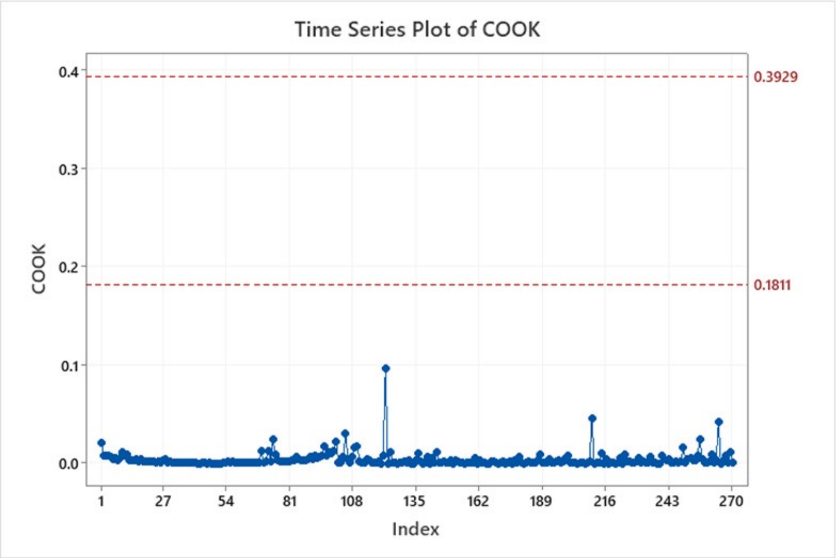


Figure 4


Time Series Plot of COOK
COOK
Index