# Predicting the Potential Spread of Forest Fires Using Climate Data

## 1. Abstract

Over the past few decades, forest fires have become increasingly common and more intense on a global scale.[1] Since wildfires are destructive in both economic and social terms, it is of critical importance to determine which environmental components are responsible for their occurrence. To get a better grasp on the causes of forest fires, we investigated the relationship between various explanatory variables on the Initial Spread Index (ISI) (outcome variable) for the Portuguese forest fires dataset.[2] We performed weighted least squares linear regression of ISI on temperature, wind, and season. After dropping a handful of outliers from the training data, the model's adjusted R-squared was approximately 0.61, which indicates that approximately 61% of the variability in ISI can be explained by these predictor variables. At the 5% significance level, temperature, wind, and season were found to be statistically significant. However, the results of these t-tests for regression coefficients may not be valid because there does not appear to be much of a relationship between ISI and these variables as evidenced by the scatter plots. In addition, the constant variance assumption for the residuals is violated, which means that our outlier analysis may not be valid. During validation, our model performs reasonably well (after removing a single outlier from the validation dataset).

## 2. Introduction

Forest fires are natural disasters occurring all across the planet that can cause immense damage to both the natural environment and humans. Since the occurrence of forest fires is hard to predict and (generally) cannot be controlled by humans, it is important to use the data we currently have to best predict the possibility of future forest fires happening. Global climate change is expected to accelerate in the decades ahead. One consequence of climate change is an expected increase in both the frequency and intensity of wildfires. In order to better prevent destruction and loss of life and property from forest fires, it is essential to understand which factors are primarily responsible for their occurrence and intensity. To study this phenomenon, we examined the effect of several predictor variables on the Initial Spread Index (ISI) (response variable) for the Portuguese forest fires dataset, whose observations were recorded at Montesinho Natural Park in Portugal.[2] ISI is a unitless quantity that measures the potential spread of a wildfire, which incorporates "fuel moisture for fine dead fuels and surface wind speed."[3, 4] The overarching goal of this work is to identify the most important factors that affect ISI. Although most of these factors cannot be directly intervened upon (i.e. season, temperature, wind, etc.), actions can still be taken to prepare for and mitigate the impact of fire-conducive weather conditions.

## 3. Background

The primary objective of this research is to construct a model that accurately predicts the potential for a forest fire to spread using data from a nature park in Portugal. The entire dataset was originally collected from Montesinho Natural Park, from the Trá-os-Montes northeast region of Portugal, over the time period from January 2000 to December 2003.[2] There are two databases: the first database was collected by the inspector that was responsible for the Montesinho fire occurrences and the second database was collected by the Bragança Polytechnic Institute, with several weather observations that were recorded with a 30 minute period by a meteorological station located in the center of the Montesinho Park.[2] In the previous lab report, we added a new variable to improve the accuracy of our model, as we considered that the previous model was not the best for predicting ISI; the R-squared value was 0.33, which is not very close to 1. We used the variable 'month' to create a new variable 'season.' Since 'season' is a categorical variable with 4 levels, it was much easier to consider the relationships with other variables using boxplots. Also, we decided to remove outliers to improve the accuracy of the model.

## 4. Modeling and Analysis

First, we looked at the correlation matrix and scatter plots (Figure 1) to determine the relationships between ISI and the other variables: temp (temperature, measured in ºC), wind (measured in km/hr), RH (relative humidity, measured in %), day, rain (measured in mm/m$^2$), and area (burned area, measured in hectares) from the training dataset. All data points were colored by season (pink for Winter, green for Spring, blue for Summer, and purple for Fall). Initial Spread Index (ISI) is not significantly correlated at all with day (correlation = 0.011), rain (correlation = 0.049), and area (correlation = 0.036). In addition, since there are no apparent patterns in these scatter plots, we decided to drop these variables from the model. Observe that ISI has a moderately strong positive relationship with temperature, a weak positive relationship with wind, and a weak negative relationship with relative humidity, which are supported by the correlation coefficients and the scatter plots. Note that some sub-correlation entries are marked as NA, which is due to one variable having a standard deviation equal to zero for that particular season (i.e. all winter observations had no rain, so the standard deviation of the rain variable for winter is zero, and thus the correlation between ISI and rain for winter is undefined).

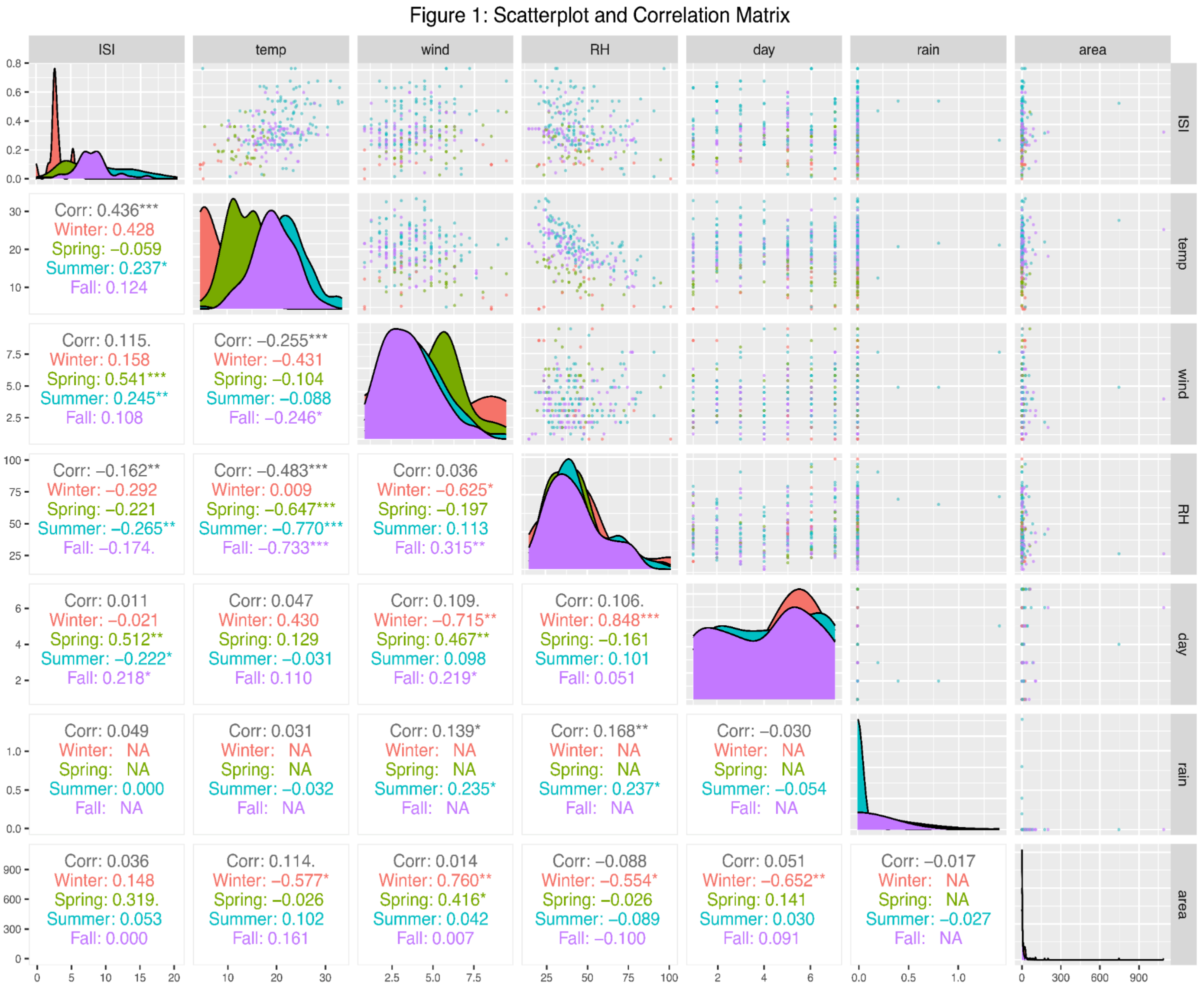

**Figure 1: Scatter plot and correlation matrix of select variables, points colored by season.**

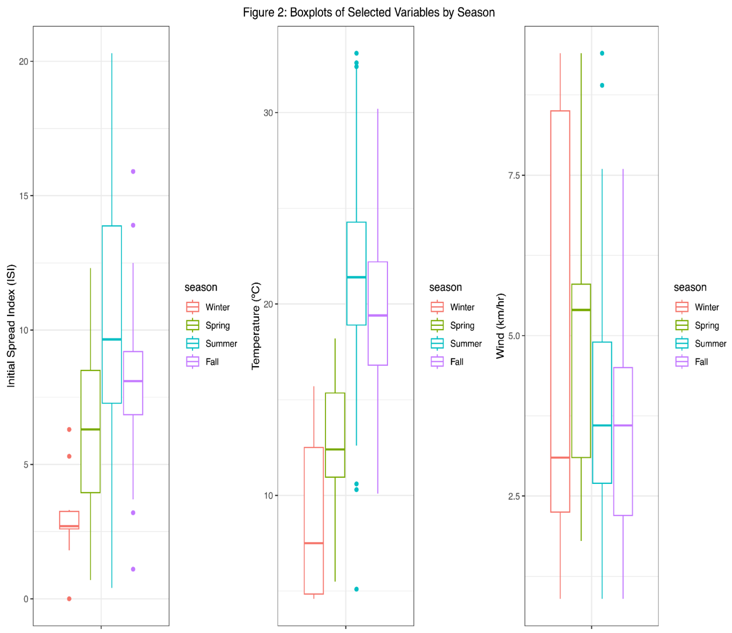

**Figure 2: Boxplots of initial spread index (ISI) values (Left), temperature (Middle), and wind (Right) by season in the training dataset.**

We examined the relationship between the categorical variable season and other continuous variables (ISI, temperature, and wind) by generating boxplots. In the left boxplot, we observe that ISI (outcome variable in our model) differs considerably by season. Therefore, the season variable was included as a predictor in our model. Notice that some ISI values were outliers (both low and high) in the Winter and Fall seasons. In the middle panel, the boxplot of temperature by season shows that temperature varies considerably by season. The boxplots of winter and spring appear to be much lower than Summer and Fall. This suggests that the temperature is lower during Winter and Spring than in Summer and Fall, which is apparently true at Montesinho Natural Park. The boxplot furthest to the right shows the relationship between wind and season. The boxplot of Winter is comparatively taller than for the other seasons, which suggests that wind during Winter tends to be highly variable. The relationship between wind and season is weaker compared with the relationships between season and ISI and season and temperature.

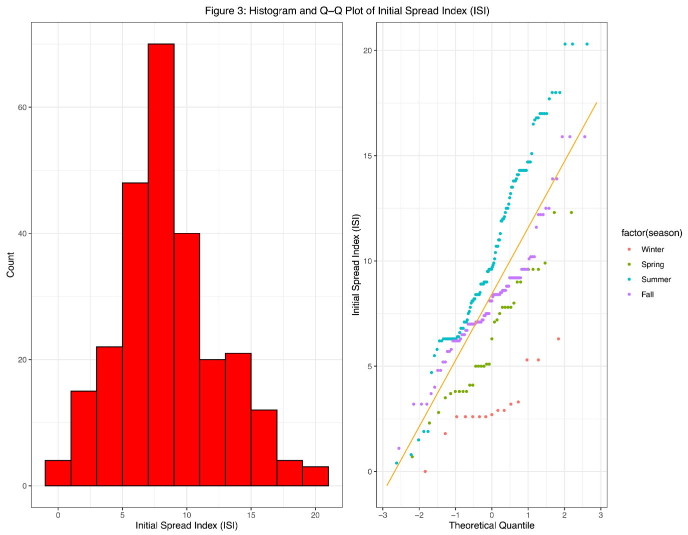

**Figure 3: Histogram of ISI values in the training dataset (Left). Q-Q plot of ISI values in the training dataset, colored by season (Right).**

We examined the distribution of ISI in the training dataset by generating a histogram and Q-Q plot (Figure 3). ISI appears normally distributed but skewed slightly to the right. In the Q-Q plot, as the quantiles increase, several points deviate from the Q-Q line, indicating once again that the distribution of ISI is skewed right. By season, it appears that Summer observations tend to fall above the line, while Spring and Winter observations fall below the line. At lower (mostly negative) quantiles, the Fall observations seem to fall above the Q-Q line, while at higher quantiles, the Fall observations fall below the line. There are also three apparent outliers at the upper extreme with ISI values > 20.

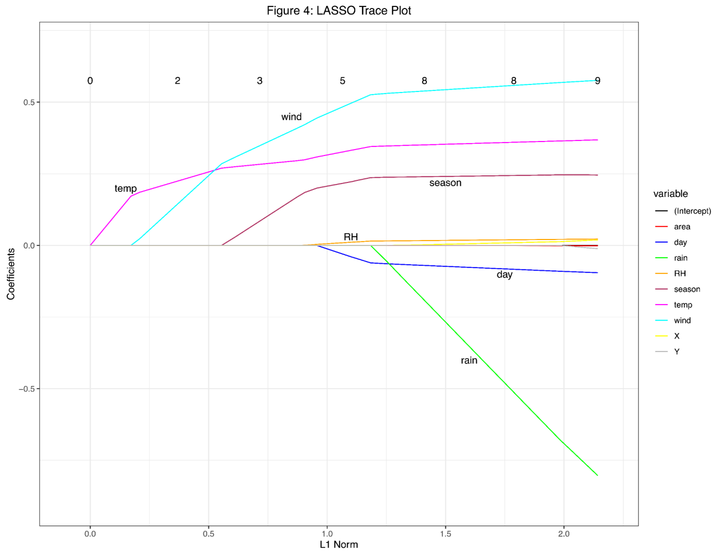

**Figure 4: LASSO trace plot used in model selection to identify the most important variables in the training data.**

We constructed a LASSO trace plot (Figure 4) to assist with the model selection process. Since temperature and wind diverge from the x-axis at small L1 norm values, these two variables are considered the most important, and we included them in our model. Note that categorical variables (such as season) may not be interpreted correctly by LASSO. Therefore, we didn't make any decisions about categorical variables from this plot. In addition to all the variables already dropped from the model during the scatter plot and correlation matrix analysis phase, relative humidity was also dropped from the model based on the LASSO trace plot.

**Table 1: Ordinary Least Squares Regression Summary**

| Variable | | Estimate | Standard Error | t-Statistic | P-Value |
|---|---|---|---|---|---|
| (Intercept) | | -0.81674 | 1.13269 | -0.721 | 0.47154 |
| Temperature | | 0.18017 | 0.04805 | 3.749 | 0.00022 |
| Wind | | 0.49891 | 0.11667 | 4.276 | 2.69E-05 |
| Season | Winter | Reference | | | |
| | Spring | 2.43919 | 1.03958 | 2.346 | 0.019732 |
| | Summer | 5.52504 | 1.09541 | 5.044 | 8.71E-07 |
| | Fall | 3.71688 | 1.05245 | 3.532 | 0.000491 |

Multiple R-squared: 0.3332, Adjusted R-squared: 0.32

**Table 1: Summary of ordinary least squares regression from the training dataset.**

After viewing the LASSO trace plot, an ordinary least squares regression analysis, shown in Table 1, was performed as an initial model to determine whether temperature, wind, and season are significantly predictive of ISI. Note that we decided to train and validate the model, so the dataset (n = 517) was split in half to conduct internal validation. The purpose of splitting our dataset in half is to use one dataset to conduct variable selection and the other to evaluate the final model. For clarity, it should be mentioned that the 'season' variable is categorical with 'Winter' as the reference group for making comparisons in the regression analysis. Since a few different models are explored in this paper, the hypothesis tests explained below will be utilized several times.

Looking at the model as a whole, the global null hypothesis ($H_0$) is the F-test, with all beta coefficients equal to zero. The alternative hypothesis ($H_A$) in this case would be that at least one beta coefficient is not equal to zero. With a sample size of 259 (half of the observations in the original dataset), the overall regression has an R-squared value of 0.3332 and F-statistic (df: 5, 253) of 25.28 (p < 2.2e-16), which is greater than the critical F-value at the 0.05 level and the global null hypothesis is rejected. The variation in ISI can be explained in some way by at least one variable in the training model. The parameter estimate of the intercept is -0.82 (t = -0.72, p = 0.47, stderr = 1.13) and the adjusted R-squared value is 0.32, meaning that 32% of the variation of ISI can be explained by the covariates in the training dataset. Since the R-squared and adjusted R-squared values are not very close to 1, the overall training model may not be the best at predicting ISI, even though the model is statistically significant at the 0.05 level (for the F-test).

Now we must look at the t-tests for the parameters in the training model. For each covariate, the null hypothesis ($H_0$) is that there is not a significant linear relationship between the independent variable and ISI (adjusting for other predictors). The alternative hypothesis ($H_A$) is that there is a significant linear relationship between the independent variable and ISI (adjusting for other predictors). In the training model, the null hypothesis can be rejected for all of the covariates in the model. Since the p-values for all covariate t-statistics are less than 0.05, we can conclude that at the 5% significance level, there is a significant linear relationship between each of the three covariates [individually] and ISI, adjusting for other predictors, for the training dataset.

The parameter estimate for temperature is 0.18 (t = 3.75, stderr = 0.05, p = 0.0002, df = 253), meaning that a one degree Celsius increase in temperature is accompanied by a 0.18-unit increase in ISI, holding wind speed and season constant. The parameter estimate for wind is 0.50 (t = 4.28, stderr = 0.12, p = 2.69e-05, df = 253), meaning that a one km/h increase in wind is accompanied by a 0.50-unit increase in ISI, holding temperature and season constant. The parameter estimate for Spring, compared to Winter, is 2.44 (t = 2.35, stderr = 1.04, p = 0.02, df = 253), meaning that the mean difference in ISI between Spring and Winter is approximately 2.44 units, holding temperature and wind speed constant. The parameter estimate for Summer, compared to Winter, is 5.53 (t = 5.04, stderr = 1.10, p = 8.71e-07, df = 253), meaning that the mean difference in ISI between Summer and Winter is approximately 5.53 units, holding temperature and wind speed constant. The parameter estimate for Fall, compared to Winter, is 3.72 (t = 3.53, stderr = 1.05, p = 0.0005, df = 253), meaning that the mean difference in ISI between Fall and Winter is approximately 3.72 units, holding temperature and wind speed constant. *Note these interpretations of the t-statistics in the training model are not necessarily valid, though, since there does not appear to be much of a relationship between ISI and the variables of interest in the scatterplot matrix.

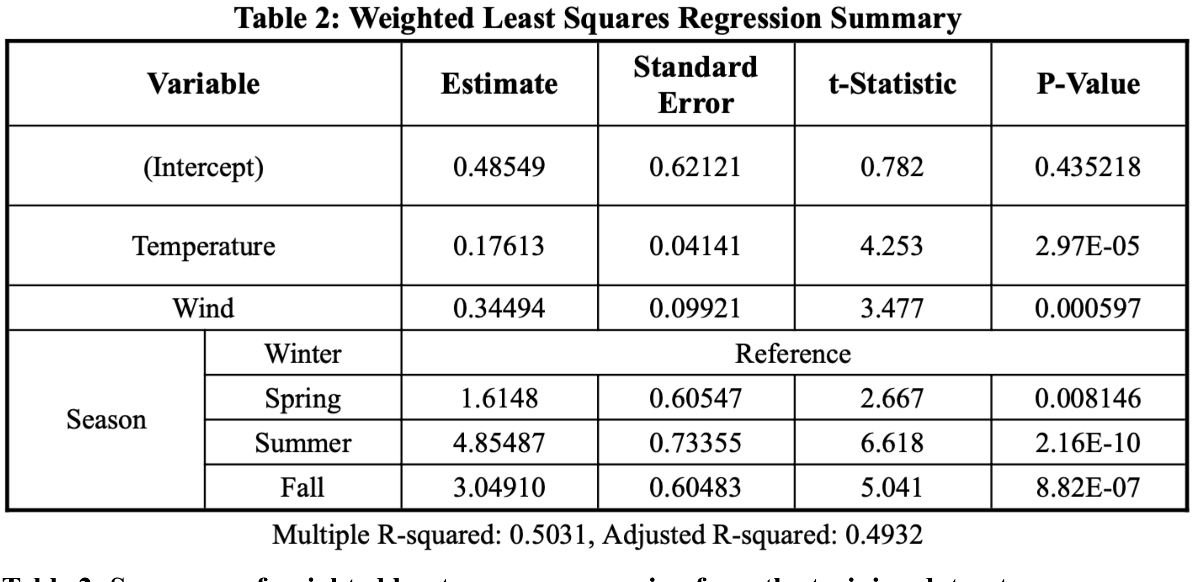

**Table 2: Weighted Least Squares Regression Summary**

| Variable | | Estimate | Standard Error | t-Statistic | P-Value |
|---|---|---|---|---|---|
| (Intercept) | | 0.48549 | 0.62121 | 0.782 | 0.435218 |
| Temperature | | 0.17613 | 0.04141 | 4.253 | 2.97E-05 |
| Wind | | 0.34494 | 0.09921 | 3.477 | 0.000597 |
| Season | Winter | Reference | | | |
| | Spring | 1.6148 | 0.60547 | 2.667 | 0.008146 |
| | Summer | 4.85487 | 0.73355 | 6.618 | 2.16E-10 |
| | Fall | 3.04910 | 0.60483 | 5.041 | 8.82E-07 |

Multiple R-squared: 0.5031, Adjusted R-squared: 0.4932

**Table 2: Summary of weighted least squares regression from the training dataset.**

Previous diagnostic procedures performed on the initial ordinary least squares regression model showed non-constant error variance in the standardized residuals. Using a weighted least squares multiple regression model (Table 2) instead may help limit biased standard errors and provide more accurate estimates with less noise. To perform this analysis using the training dataset, temperature, wind, and season were again included in the regression model as potential predictors of ISI. With a sample size of 259, the overall weighted least squares regression has an R-squared value of 0.5031 and F-statistic (df: 5, 253) of 51.22 (p < 2.2e-16), which is greater than the critical F-value at the 0.05 level and the global null hypothesis is rejected. The variation in ISI can be explained in some way by at least one variable in the weights least squares model. The adjusted R-squared value is 0.49, meaning that 49% of the variation of ISI can be explained by the covariates in the model. Since the R-squared and adjusted R-squared values are closer to 1 and the global F-statistic is more significant in this model than in the ordinary least squares model, the weighted least squares model is considered better at predicting ISI.

Looking at the t-tests for the parameters in the model, the null hypothesis can once again be rejected for all of the covariates in the model. Since the p-values for all covariate t-statistics are less than 0.05, we can conclude that at the 5% significance level, there is a significant relationship between each of the three covariates [individually] and ISI, adjusting for other predictors. A one degree Celsius increase in temperature is accompanied by a 0.18-unit increase in ISI, holding wind speed and season constant. A one km/h increase in wind is accompanied by a 0.34-unit increase in ISI, holding temperature and season constant. The mean difference in ISI between Spring and Winter is approximately 1.61 units, holding temperature and wind speed constant. The mean difference in ISI between Summer and Winter is approximately 4.85 units, holding temperature and wind speed constant. The mean difference in ISI between Fall and Winter is approximately 3.05 units, holding temperature and wind speed constant. *Note that the effect estimates are less extreme (closer to 0) in the weighted least squares model compared to the ordinary least squares model.

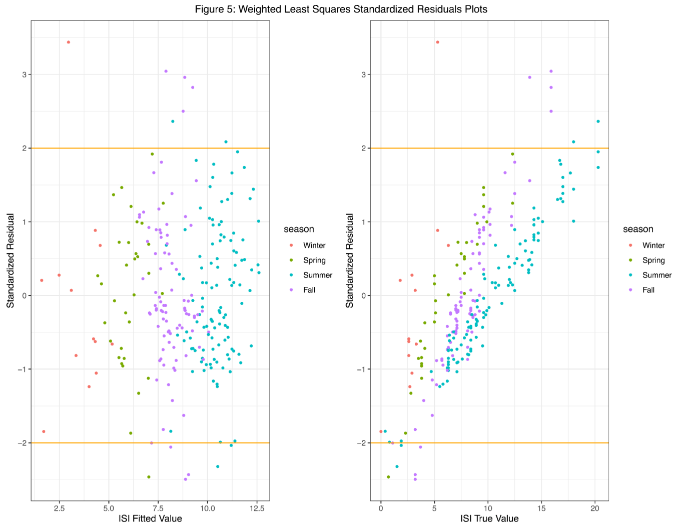

**Figure 5: Plots of standardized residuals vs. ISI fitted values in the training data (Left). Plot of standardized residuals vs. ISI true values in the training data (Right).**

The standardized residuals of the ISI fitted values from the weighted least squares regression, shown on the left in Figure 5, seem to follow some patterns. In particular, there is relatively greater variability in the standardized residuals at higher ISI fitted values, indicating that the constant variance assumption does not hold. Additionally, there are at least one or two relatively extreme [potential] outliers near the top of the plot with standardized residuals greater than or equal to 3. These observations stray from the several other data points outside of the upper cutoff line (standardized residuals greater than 2). There are also several standardized residuals below the -2 cutoff, but none deviate as extreme in respect to one another. Although most of the data points are scattered between the horizontal lines of standardized residuals less than ±2, the data still follow a pattern, as mentioned before, and the constant variance assumption is violated.

The standardized residuals of the ISI true values, shown on the right in Figure 5, also seem to follow a pattern. There appears to be a positive trend in the data, with lower standardized residuals at smaller ISI true values and higher standardized residuals at greater ISI true values. This pattern differs from that of the ISI fitted values, indicating that the trend of the model is not correctly specified. It should be noted that the potential outliers identified in the interpretation of the ISI fitted value residual plot are also present in the true value plot with the same residual values, but located at different relative points on the x-axis.

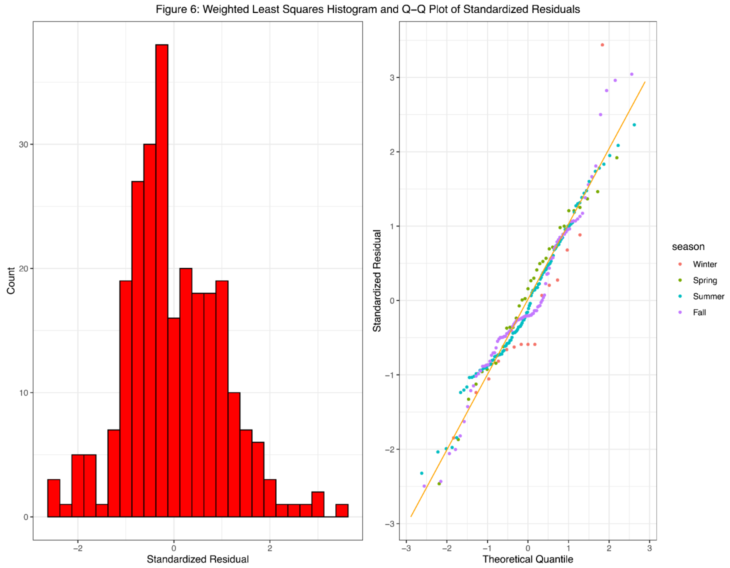

**Figure 6: Histogram of standardized residuals in the training dataset (Left). Q-Q plot of standardized residuals in the training dataset, colored by season (Right).**

In Figure 6, we examined the distribution of the standardized residuals from our weighted least squares training model. The histogram shows that the standardized residuals generally follow a normal distribution with a skew to the right. There seems to be some problematic points in the left tail of the distribution as well. In the Q-Q plot, as the quantiles increase in magnitude, several points deviate from the Q-Q line (primarily at the right tail). There are also three potential outliers at the upper extreme with standardized residuals greater than or equal to 3. *Note that these observations were also identified in Figure 3 (Histogram and Q-Q plot of ISI).

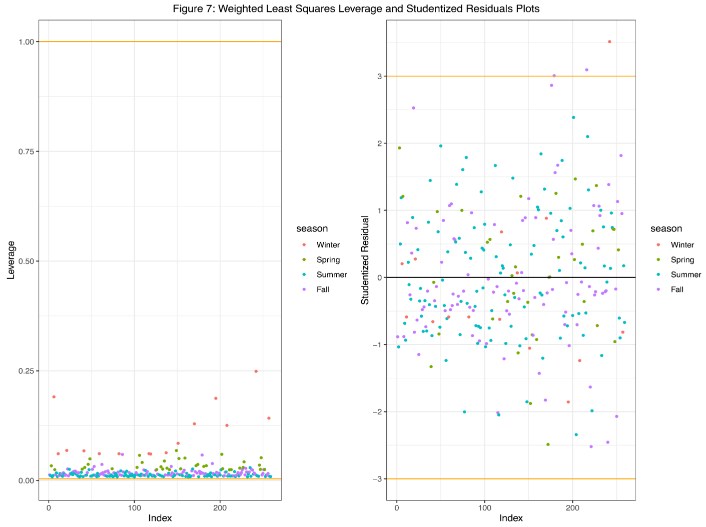

**Figure 7: Plot of leverages to identify outliers in the x-space in the training dataset (Left). Plot of studentized residuals to identify outliers in the y-space in the training dataset (Right).**

In Figure 7, we searched for potential outliers in the training dataset. The plot of the leverages is shown on the left to visualize outliers in the x-space, and there is a point with a leverage greater than or equal to 0.20 that may be considered an outlier. It should be noted that by season, Winter observations tend to have higher leverages than those of the other three seasons. The studentized residual plot, shown on the right, can be used to identify residuals with a magnitude greater than 3 as outliers in the y-space. Here, there are three points with studentized residuals greater than three that must be flagged as potential outliers. However, considering that the constant variance of the errors assumption is violated (Figure 5), this outlier analysis might not be completely valid.

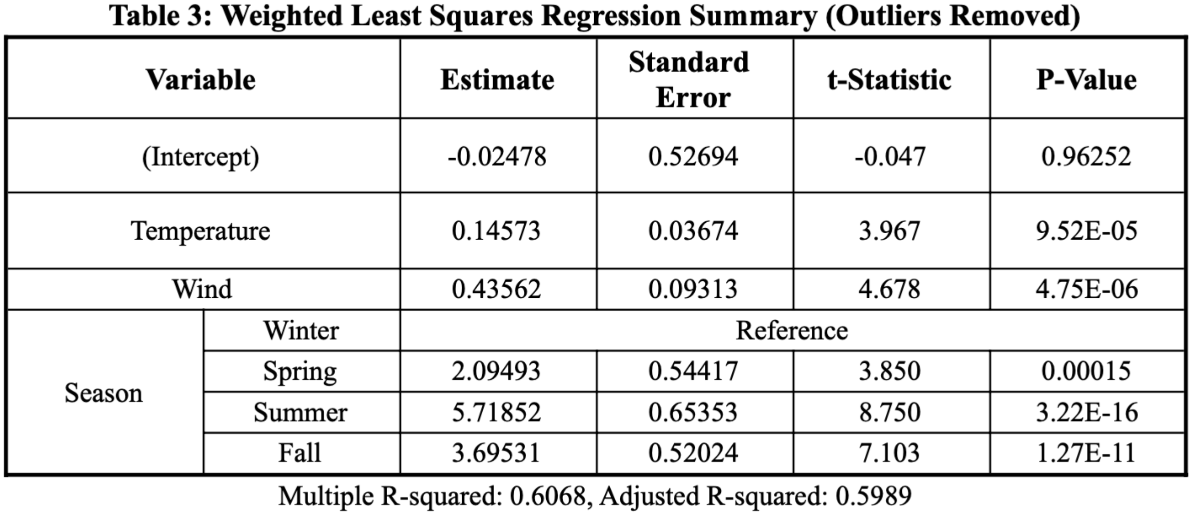

Table 3: Weighted Least Squares Regression Summary (Outliers Removed)

| Variable | | Estimate | Standard Error | t-Statistic | P-Value |
|---|---|---|---|---|---|
| (Intercept) | | -0.02478 | 0.52694 | -0.047 | 0.96252 |
| Temperature | | 0.14573 | 0.03674 | 3.967 | 9.52E-05 |
| Wind | | 0.43562 | 0.09313 | 4.678 | 4.75E-06 |
| Season | Winter | Reference | | | |
| | Spring | 2.09493 | 0.54417 | 3.850 | 0.00015 |
| | Summer | 5.71852 | 0.65353 | 8.750 | 3.22E-16 |
| | Fall | 3.69531 | 0.52024 | 7.103 | 1.27E-11 |

Multiple R-squared: 0.6068, Adjusted R-squared: 0.5989

**Table 3: Summary of weighted least squares regression after removing three outliers from the training dataset.**

As mentioned above, the leverage and studentized residuals plots showed that our dataset had some possible outliers. We defined an outlier as having a leverage greater than 0.20 or having a studentized residual greater than 3 in absolute value. We ended up identifying 3 outliers in our training data. Consequently, we decided to remove these three outliers and refit our model with the remaining 256 data points (Table 3). With a sample size of 256, the overall weighted least squares regression with outliers removed has an R-squared value of 0.6068 and F-statistic (df: 5, 250) of 77.15 ($p < 2.2e-16$), which is greater than the critical F-value at the 0.05 level and the global null hypothesis is rejected. The variation in ISI can be explained in some way by at least one variable in this new weights least squares model. The adjusted R-squared value is 0.599, meaning that 59.9% of the variation of ISI can be explained by the covariates in the model. Since the R-squared and adjusted R-squared values are closer to 1 and the global F-statistic is more significant in this model than in either the ordinary least squares model or the previous weighted least squares model, this final model is the best at predicting ISI.

Looking at the t-statistics for the parameters in the final model, all the p-values are less than 0.05, so the null hypothesis is rejected. Once again, we conclude that at the 5% significance level, there is a significant relationship between each of the three covariates [individually] and ISI, adjusting for other predictors. A one degree Celsius increase in temperature is accompanied by a 0.15-unit increase in ISI, holding wind speed and season constant. A one km/h increase in wind is accompanied by a 0.44-unit increase in ISI, holding temperature and season constant. The mean difference in ISI between Spring and Winter is approximately 2.09 units, holding temperature and wind speed constant. The mean difference in ISI between Summer and Winter is approximately 5.72 units, holding temperature and wind speed constant. The mean difference in ISI between Fall and Winter is approximately 3.70 units, holding temperature and wind speed constant. *These are the final effect estimates that our conclusions will be based upon.

## 5. Prediction

To see how well the refitted model predicts ISI, we examined the MSE and relative MSE in the training and validation datasets. We initially split our dataset in half (n = 259 for training dataset, n = 258 for validation dataset) to see our model's validity. Both the MSE and relative MSE explain how close the predicted values of ISI are to the true ISI values; therefore, the lower the MSE and relative MSE are, the better the model is predicting the data.

Overall, our model can predict ISI relatively well in both the training and validation datasets. The MSE (10.37) and relative MSE (0.1144) in the training dataset are comparatively lower than those of the validation dataset (MSE = 20.33, relative MSE = 0.1801) (Table 4). The higher values of MSE and relative MSE in the validation dataset suggest that there are larger residuals (in magnitude) in the prediction of ISI in the validation dataset than in the training dataset. However, we noticed that the validation dataset had one major outlier (ISI = 55). As a result, we removed it after some consideration of how well the model can predict ISI without it. Our model predicts ISI relatively better once the outlier is removed as seen in the significant decreases in MSE and relative MSE for the validation dataset (12.40, 0.1226, respectively). Observe that the MSE and relative MSE of the validation data after removing the outlier are still higher than the MSE and relative MSE in the training dataset, which indicates that the model does not perform as well in the validation data compared with the training data.

**Table 4: Weighted Least Squares Validation Summary**

| Measure | Training Data | Validation Data | Validation Data (Without Outlier) |
|---|---|---|---|
| Mean Square Error (MSE) | 10.37 | 20.33 | 12.40 |
| Relative Mean Square Error (Relative MSE) | 0.1144 | 0.1801 | 0.1226 |

**Table 4: Summary of validation in the training dataset (after removing three outliers), original validation dataset, and modified validation dataset (after removing one outlier).**

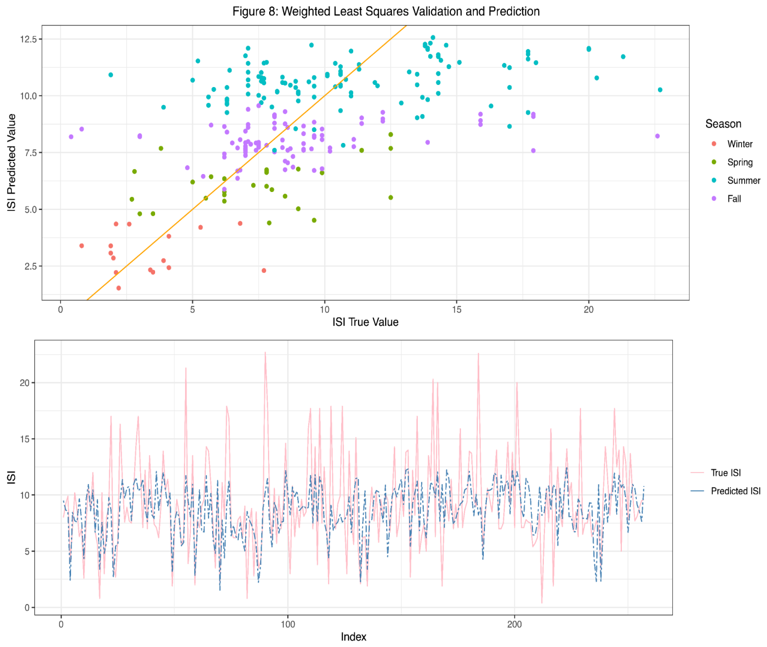

**Figure 8: Plot of ISI predicted values vs. ISI true values in the validation dataset, colored by season, and overlaid with the 45° line (Top). Comparison of ISI true values (pink) and predicted values (blue) in the validation dataset (Bottom).**

To further examine how well our model does in predicting ISI, we compared the predicted ISI values from the validation dataset to the true ISI values. In the top plot of Figure 8, we plotted the predicted ISI values against the true ISI values, overlaid with the 45 degree line and colored by season. For lower ISI true values, we saw that there was a linear trend and noticed that most of the points stay close to the line. The linear trend in the beginning suggests that our model predicts lower ISI values extremely well. However, our model does not predict higher ISI values well. As the ISI values increase, we noticed that the points start to deviate from the 45 degree line and spread out. In particular, our model consistently underpredicts these highest ISI values.

The bottom plot of Figure 8 compares the true ISI values (pink) with the model's predicted ISI values (blue) for the validation data. The plot also shows that our model does well in predicting the lower ISI values since the predicted ISI values and the true ISI are extremely close. Once again, we noticed that the model underpredicts the higher ISI values, confirming what we saw in the scatter plots. Hence, we concluded that our model does not predict ISI quite as well as we thought when we initially looked at the model's adjusted $R^2$.

**6. Discussion**

Initially, we had three main goals when we started this project. We wanted to look at the most important predictors to examine the outcome of interest (ISI), search for any possible confounders, and examine any potential effect measure modification. With variable selection and various analyses, we found that the most important covariates that predict ISI are: wind, temperature, and season. However, we did not cover the topics of confounding and effect measure modification (interaction between two variables) in this class. If we were given the opportunity, we would conduct further analyses to see if there was any interaction between two variables and whether there was any confounding in the results.

Recall that our final model includes wind, temperature, and season to predict the initial spread index (ISI). The reason we chose these three variables is that the response variable, ISI, does not have significant linear relationships with any of the other variables shown in the scatter plot and correlation matrix. Among all the variables included in the dataset, temperature has the highest correlation coefficient with ISI, which is 0.436. The adjusted $R^2$ of our final model (using weighted least squares) is around 60%; this demonstrates one strength of the model. Even though the constant variance assumption of our linear model does not hold since there is a pattern in the standardized residuals vs. ISI fitted values plot, the normality assumption seems to hold since the Q-Q plot basically follows a straight line with some minor deviations. As we can see in the histogram in Figure 6, the distribution of standardized residuals seems to be bimodal (i.e. two peaks in the histogram). This could indicate that there should actually be two separate models. We may need to divide the training data to look further into this problem.

It is important to address the limitations of our work. Since we only examined fire conditions at a single natural park in Portugal, these results may not be generalizable to other countries, continents, or climates in general. Furthermore, since this dataset was collected during the early 2000s, the results may not even be generalizable to Montesinho Natural Park today due to the changing climate.

In order to improve our model from previous lab reports, we first removed several outliers from the training dataset before moving into the validation dataset to assess our model. We also decided to use LASSO for variable selection and weighted least squares (the weights are calculated using the absolute values of the residuals from the ordinary least squares model) to fit the model. Overall, the final model turns out to predict the occurrence of forest fires (using ISI as a proxy) reasonably well.

## Works Cited:

[1] MacCarthy, J., Tyukavina, S., Weisse, M., & Harris, N. (2022, August 17). *New Data Confirms: Forest Fires Are Getting Worse*. World Resources Institute. Retrieved November 12, 2022, from https://www.wri.org/insights/global-trends-forest-fires

[2] P. Cortez and A. Morais. *A Data Mining Approach to Predict Forest Fires using Meteorological Data*. In J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimaraes, Portugal, pp. 512-523, 2007.

[3] *Fire Weather Index (FWI) System | NWCG*. www.nwcg.gov/publications/pms437/cffdrs/fire-weather-index-system. Accessed 22 Oct. 2022.

[4] Van Wagner, C. E., and T. L. Pickett. 1985. Equations and FORTRAN program for the Canadian forest fire weather index system. Forestry Technical Report No. 33. Ottawa, Environment Canada, Canadian Forestry Service, Petawawa National Forestry Institute.

## 7. Appendix

R Code Relevant Snippets

```
# Scatterplot and Correlation Matrix
data <- data.frame(ISI, temp, wind, RH, day, rain, area)
ggpairs(data, upper = list(continuous = wrap("points", alpha = 0.5, size = 0.3)),
        mapping = ggplot2::aes(color = factor(season, labels = c("Winter", "Spring", "Summer", "Fall"))),
        lower = list(continuous = wrap('cor', size = 4))) +
        theme(axis.text = element_text(size = 7)) +
        labs(title = "Figure 1: Scatterplot and Correlation Matrix")

# Box Plots
legend_title <- "season"
data1 <- data.frame(ISI, season)
data1_melt <- melt(data1, id = "season")
p1 <- ggplot(data1_melt, aes(x = variable, y = value, color = factor(season))) + geom_boxplot() +
theme_bw() +
      theme(axis.text.x=element_blank()) + labs(x = "", y = "Initial Spread Index (ISI)") +
      scale_color_discrete(legend_title, labels = c("Winter", "Spring", "Summer", "Fall"))

data2 <- data.frame(temp, season)
data2_melt <- melt(data2, id = "season")
p2 <- ggplot(data2_melt, aes(x = variable, y = value, color = factor(season))) + geom_boxplot() +
theme_bw() +
      theme(axis.text.x=element_blank()) + labs(x = "", y = "Temperature (ºC)") +
      scale_color_discrete(legend_title, labels = c("Winter", "Spring", "Summer", "Fall"))

data3 <- data.frame(wind, season)
data3_melt <- melt(data3, id = "season")
p3 <- ggplot(data3_melt, aes(x = variable, y = value, color = factor(season))) + geom_boxplot() +
theme_bw() +
      theme(axis.text.x=element_blank()) + labs(x = "", y = "Wind (km/hr)") +
      scale_color_discrete(legend_title, labels = c("Winter", "Spring", "Summer", "Fall"))

grid.arrange(p1, p2, p3, nrow = 1, top = "Figure 2: Boxplots of Selected Variables by Season")

# Histogram of Response Variable (ISI)
ggplot(ISI_data, aes(x = ISI, color = season)) +
  geom_histogram(binwidth = 2, color = "black", fill = "red") +
  labs(x = "Initial Spread Index (ISI)", y = "Count") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))

# Q-Q Plot for Response Variable (ISI)
ggplot(ISI_data, aes(sample = ISI, color = factor(season))) +
  stat_qq(size = 1) +
  geom_qq_line(color = "orange") +
  labs(x = "Theoretical Quantile", y = "Initial Spread Index (ISI)") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_color_discrete(labels = c("Winter", "Spring", "Summer", "Fall"))

# LASSO
x <- model.matrix(ISI~X+Y+season+day+temp+RH+wind+rain+area, firedata_train)
y <- ISI
fit <- glmnet::glmnet(x, y, alpha = 1)

# Plot coefficients vs. lasso penalty
pallete <- c('black', 'red', 'blue', 'green', 'orange','maroon','magenta','cyan', 'yellow', 'gray')
Lasso <- autoplot(fit, cex = 0.5, xlim = c(-0.1,2.2), ylim = c(-0.9,0.7)) +
        scale_colour_manual(values = pallete) +
        labs(title = "Figure 4: LASSO Trace Plot") +
        theme_bw() +
        theme(plot.title = element_text(hjust = 0.5)) +
        annotate("text", x = 0.0, y = 0.2, label = "temp") +
        annotate("text", x = 0.75, y = 0.45, label = "wind") +
        annotate("text", x = 1.5, y = 0.2, label = "season") +
        annotate("text", x = 1.75, y = -0.13, label = "day") +
        annotate("text", x = 1.5, y = -0.4, label = "rain") +
        annotate("text", x = 1, y = 0.05, label = "RH")
Lasso

# Ordinary Least Squares Model (Using Training Data)
# ISI vs. Temperature, Wind, Season (categorical)
model_train <- lm(ISI~temp + wind + factor(season))
summary(model_train)

# Weighted Least Squares (Using Training Data)
# Calculate fitted values from a regression of absolute residuals vs predictors
wts <- 1 / fitted(lm(abs(residuals(model_train))~temp + wind + factor(season)))^2

# Fit a WLS model using weights = 1 / (fitted values)^2
wls_train <- lm(ISI~temp + wind + factor(season), weights = wts)
```

```
summary(wls_train)

# Standardized Residuals vs. ISI Fitted Values Plot
ggplot() + geom_point(data = data_diagnostics_train, aes(x = fitted_train, y = stand_resid_train, col =
season), size = 1) +
  geom_hline(yintercept = 2, color = "orange") + geom_hline(yintercept = -2, color = "orange") +
  labs(x = "ISI Fitted Value", y = "Standardized Residual") +
  scale_y_continuous(breaks = seq(-3, 3, 1)) +
  theme_bw() +
  scale_color_discrete(labels = c("Winter", "Spring", "Summer", "Fall")) +
  theme(plot.title = element_text(hjust = 0.5))

# Standardized Residuals vs. ISI True Values Plot
ggplot() + geom_point(data = data_diagnostics_train, aes(x = ISI, y = stand_resid_train, col = season),
size = 1) +
  geom_hline(yintercept = 2, color = "orange") + geom_hline(yintercept = -2, color = "orange") +
  labs(x = "ISI True Value", y = "Standardized Residual") +
  scale_y_continuous(breaks = seq(-3, 3, 1)) +
  theme_bw() +
  scale_color_discrete(labels = c("Winter", "Spring", "Summer", "Fall")) +
  theme(plot.title = element_text(hjust = 0.5))

# Histogram of Standardized Residuals
ggplot(data_diagnostics_train, aes(x = stand_resid_train)) +
  geom_histogram(binwidth = 0.25, color = "black", fill = "red") +
  labs(x = "Standardized Residual", y = "Count") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))

# Q-Q Plot
ggplot(data_diagnostics_train, aes(sample = stand_resid_train, color = season)) +
  stat_qq(size = 1) +
  geom_qq_line(color = "orange") +
  labs(x = "Theoretical Quantile", y = "Standardized Residual") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_color_discrete(labels = c("Winter", "Spring", "Summer", "Fall"))

# Plot of Leverages (to detect outliers in the x-space)
ggplot() + geom_point(data = data_diagnostics_train, aes(x = index_train, y = leverages_train, color =
season), size = 1) +
  geom_hline(yintercept = 1 / length(index_train), color = "orange") + geom_hline(yintercept = 1, color =
"orange") +
  labs(x = "Index", y = "Leverage") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_color_discrete(labels = c("Winter", "Spring", "Summer", "Fall"))

# Plot of Studentized Residuals (to detect outliers in the y-space)
ggplot() + geom_point(data =  data_diagnostics_train, aes(x = index_train, y = student_resid_train, color
= season), size = 1) +
  geom_hline(yintercept = -3, color = "orange") + geom_hline(yintercept = 3, color = "orange") +
  geom_hline(yintercept = 0, color = "black") +
  scale_y_continuous(breaks = seq(-3, 3, 1)) +
  labs(x = "Index", y = "Studentized Residual") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_color_discrete(labels = c("Winter", "Spring", "Summer", "Fall"))

# Remove Outliers from Training Data
ISI_rev <- ISI[-c(which(student_resid_train > 3 | student_resid_train < -3 | leverages_train > 0.20))]
temp_rev <- temp[-c(which(student_resid_train > 3 | student_resid_train < -3 | leverages_train > 0.20))]
season_rev <- season[-c(which(student_resid_train > 3 | student_resid_train < -3 | leverages_train >
0.20))]
wind_rev <- wind[-c(which(student_resid_train > 3 | student_resid_train < -3 | leverages_train > 0.20))]

# Weighted Least Squares (After Removing Outliers)
# Calculate fitted values from a regression of absolute residuals vs predictors
wts_rev <- 1 / fitted(lm(abs(residuals(model_train_rev))~temp_rev + wind_rev + factor(season_rev)))^2

# Fit a WLS model using weights = 1 / (fitted values)^2
wls_train_rev <- lm(ISI_rev~temp_rev + wind_rev + factor(season_rev), weights = wts_rev)
summary(wls_train_rev)

# Remove largest residual in Validation Data (nearly 3x the second largest residual in magnitude)
predict <- predict_valid$fit
predict_valid_rev <- predict[-c(which(resid_valid > 15))]
resid_valid_rev <- resid_valid[-c(which(resid_valid > 15))]
ISI_valid_rev <- firedata_valid$ISI[-c(which(resid_valid > 15))]
season_valid_rev <- firedata_valid$season[-c(which(resid_valid > 15))]

# Plot Initial Spread Index vs. Prediction for Validation Data
ggplot(data = test, aes(x = ISI, y = Prediction, color = Season)) + geom_point() +
  geom_abline(intercept = 0, slope = 1, color = "orange") +
  labs(x = "ISI True Value", y = "ISI Predicted Value") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_color_discrete(labels = c("Winter", "Spring", "Summer", "Fall"))

# Further Comparisons of Predicted Values vs. True Values for Validation Data
ggplot(data = test, aes(x = Index)) +
  geom_line(aes(y = ISI, color = "ISI")) +
  geom_line(aes(y = Prediction, color = "Prediction"), linetype = "twodash") +
  scale_color_manual(name = element_blank(), labels = c("True ISI","Predicted ISI"),
             values = c("pink", "steelblue")) + labs(y = "") +
  labs(x = "Index", y = "ISI") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```