Mosquito Traps and Rainfall in YEG

Alexander Ondrus

3/5/2020

Loading & Manipulating Data

The two CSV files that we need for this example are the Rainfall Gauge Results and the Mosquitoes Trap Data. I also load the `tidyverse` library as well.

Note that the mosquito data is *weekly* and the rainfall data is *daily*. That means that I need to aggregate the rainfall data to the same frequency as the mosquito data and then join the two data frames together. Neither date column is immediately recognized as a date, so I use the `lubridate` package to set the format.

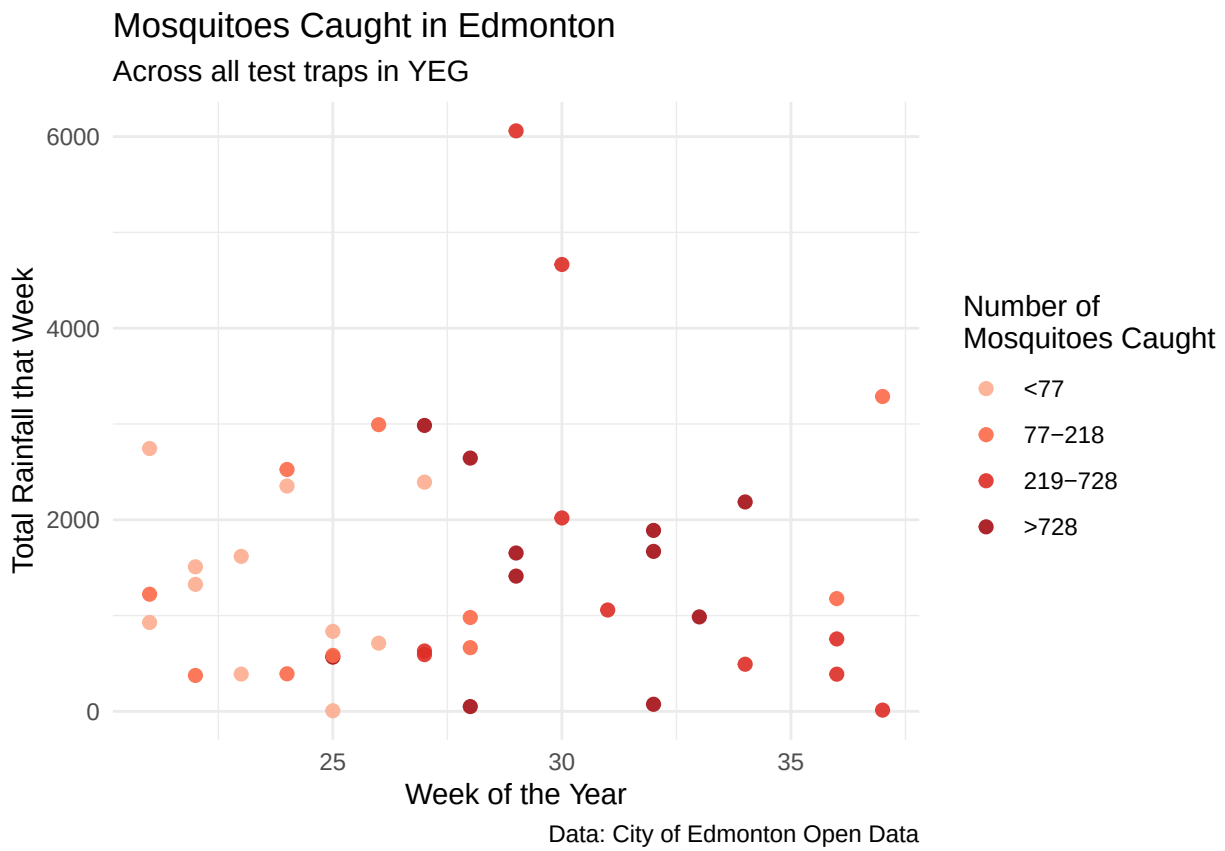
The `lubridate` package also has a `week()` function that returns the week of the year for a given date. I will add these columns to both the mosquito and rain data. I also rename the `YEAR` column in the rain data to `Year` to match up with the mosquito data later on.

To simplify things greatly, I am just going to look at the total rainfall for each week and the total number of mosquitos caught each week. This means that I only need the year, week, and total columns for each data set. Note that the data is spread over multiple rows for each set, so I will need to group and summarise the data as well. To make these commands concise, I will use the pipe operator `%>%`

Now the two data sets are on the same frequency, I can merge them using an *inner join*. This means that I only include rows that have matching values in both data frames for the specified columns.

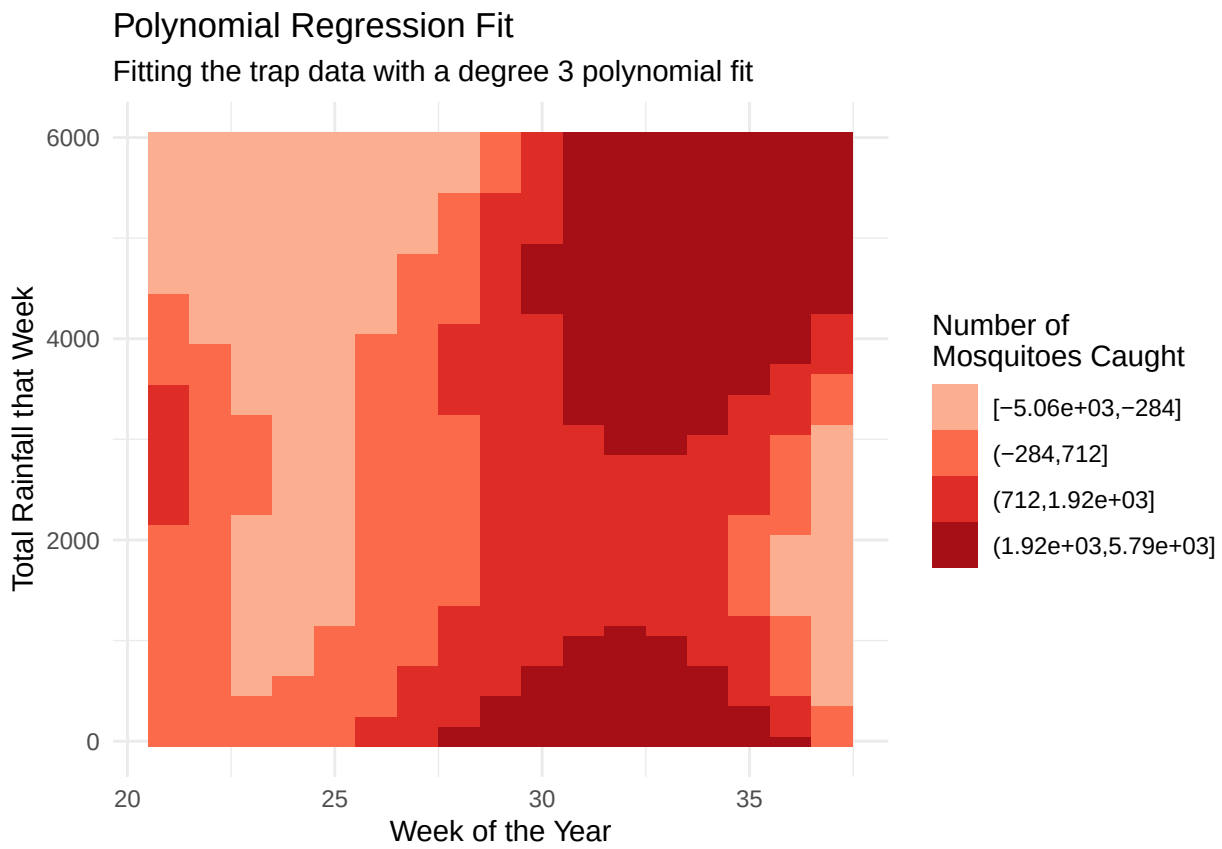
Predictive Modelling

I want to predict the number of mosquitoes caught in the trap in a given week based on the total rainfall that week and the week of the year. R has many, many ways of doing this, but I will show just two. Before I do, let's take a look at the data:



Polynomial Regression

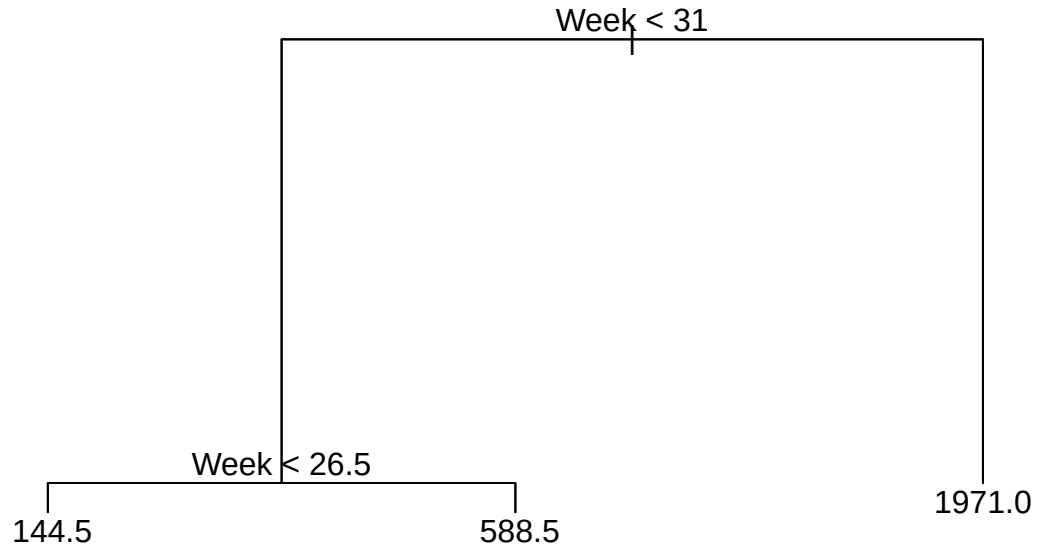
It is difficult to see any linear relationships in the data above, so instead I will try a polynomial regression. I do this using the `polym()` command to tell R that I want to use not only the original two variables, but their powers and intersections up to a specified degree.



Taking a quick look at the legend values, it is easy to see that the polynomial regression fit does not do well at predicting the data at all.

Regression Tree

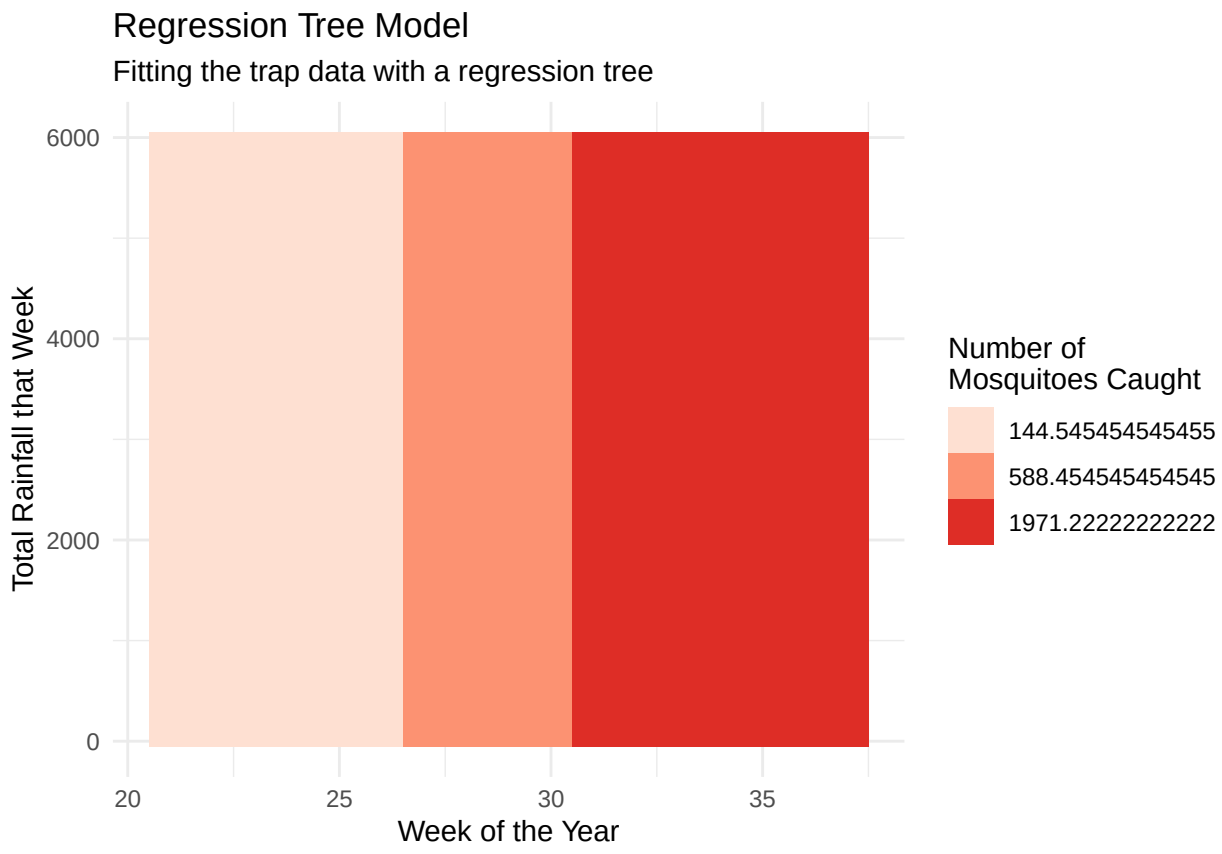
Regression trees segment our predictor space in a manner that minimizes the variation of the output variable and uses the mean of given values to predict the output value in each region. See here for more details. The R package for building regression (or classification) trees is `tree`.



This means that the MSE (or *M*ean *S*quared *E*rror), defined by:

$$\text{MSE} = \frac{1}{n}\sum_{i}\left(y_{\text{predicted}} - y_{\text{actual}}\right)^2$$

is equal to 785 146 for our model. Taking the square root of this we can estimate that the model will be off by 886 mosquitoes. Let's visualize the results in a way similar to linear regression.


Regression Tree Model
Fitting the trap data with a regression tree
Total Rainfall that Week
6000
4000
2000
0
20
25
30
35
Week of the Year
Number of
Mosquitoes Caught
144.545454545455
588.454545454545
1971.22222222222