Project Applied Biostatistics

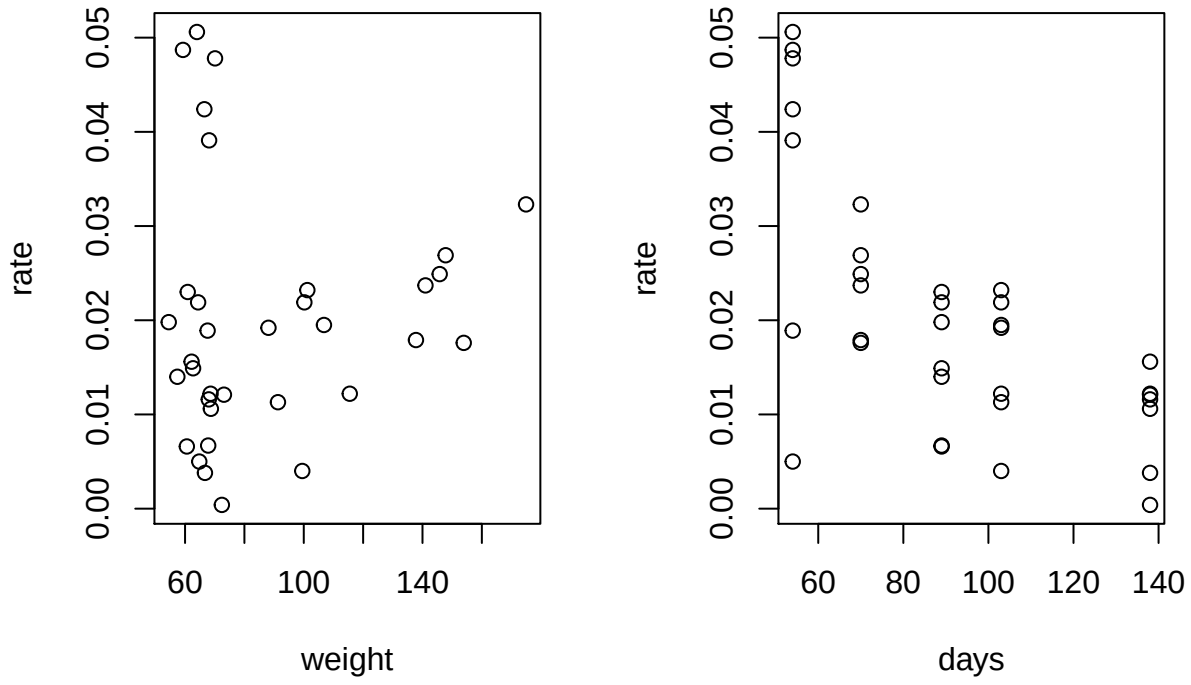
Santiago Anton Moreno

16/03/2020

Our goal is to find a model that describe the rate of fungal invasion of 5 varieties of apples for 7 fusarium strains.

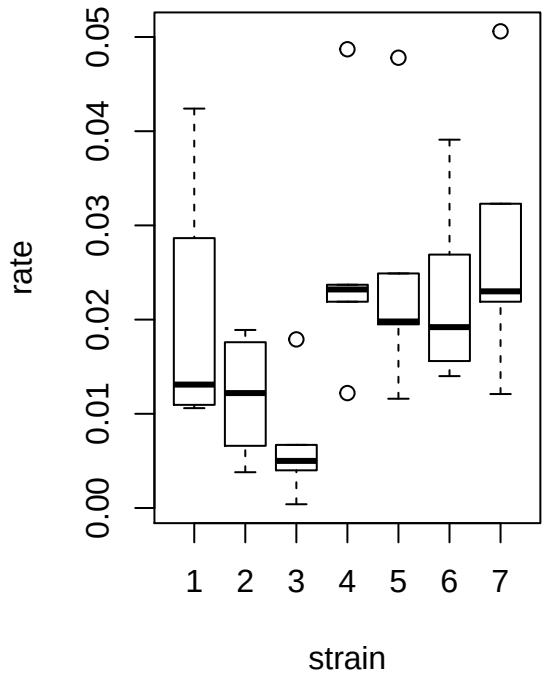
As the rate of fungal expension and the fungal radial advance are redundant information, we have to choose which one will we discard and which one will we use as the response. We decided to use the rate, we will justify it later in the notebook.

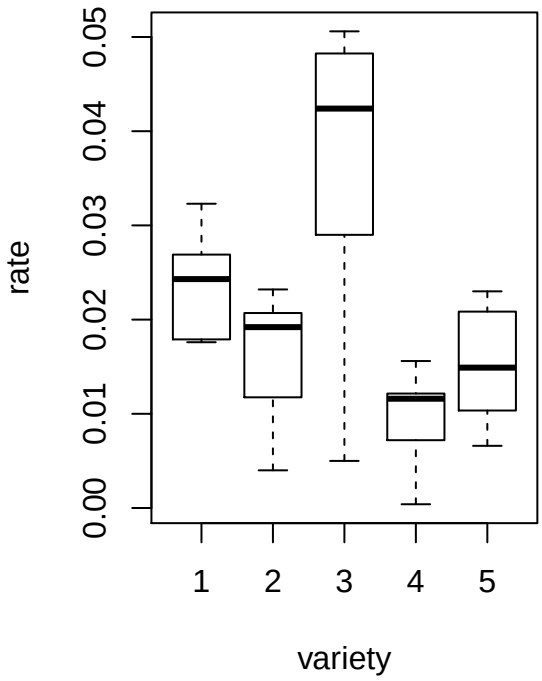
Now we can look at some plots to check if there are some trends we should pay attention to.



It is hard to tell exactly but it appears there is no correlation between the rate and the weight. Also there is higher variation when the weight is low. We also did this plot for the number of days and reached the same conclusion, which is to be expected as the weight and radius in our dataset have a correlation of 0.97.

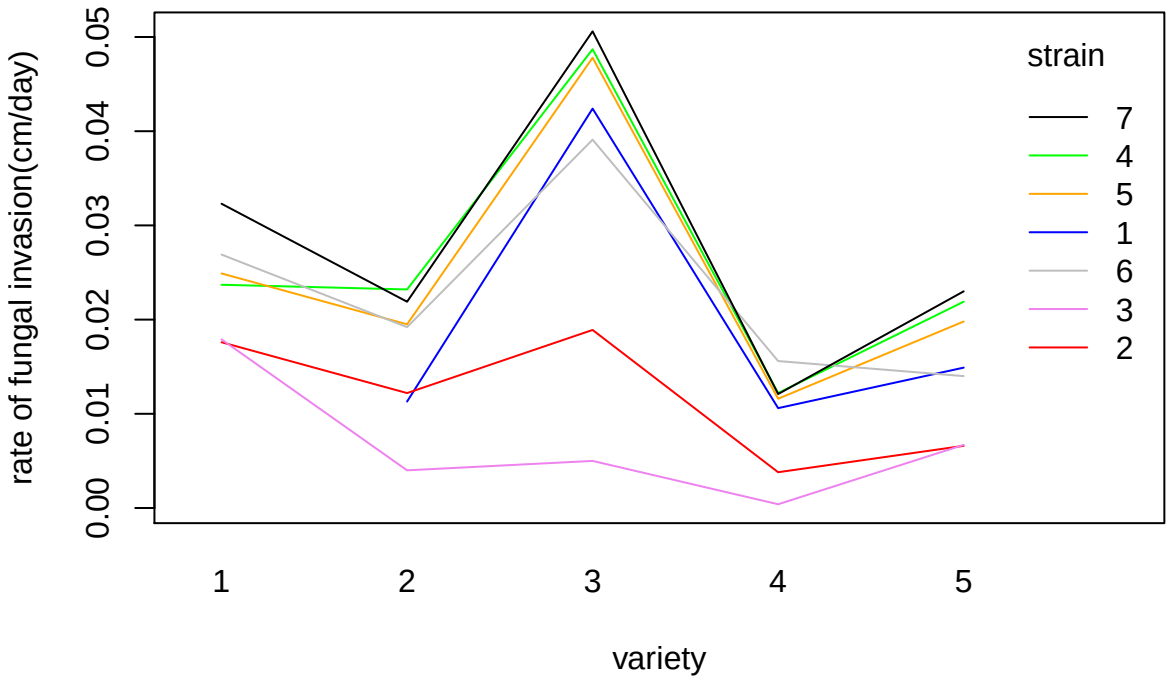
It seems that higher number of days induce lower rates of fungal invasion. However we must keep in mind that apples with the same varieties have the same number of days in our dataset, so we must be careful with our conclusions.





As we can see in those two plots the homoskedacity property is clearly violated thus we should be careful when doing analysis. The strain factor seems to be more informative than the apple variety. Normality is also violated as we see that the boxplot for some strains or variety are clearly non symetric which suggest non-normality of the rate. Indepence should be preserved since the study was done cleanly.

By taking a deeper look at the description, we noticed that variety 3 and variety 5 are apples with the same race and variety 5 has higher number of days. However not only variety has rates much lower than variety 3, but variety 5 also has a lower fungal radial distance. It does not makes sense unless we assume that the conditions of the experience were not the same. For exemple, the maturity of the apples could be different at the start of the experience for those two varieties of apples. This justifies working with “varieties” of apples instead of race of apples and number of days.



There are several things we may notice from this plot. The general rate of fungal invasion varies a lot depending on the variety of the apple. Also different strains induce have different rates of fungal invasion. We see that all the lines have more or less the same shape but with different scalings, which suggest that a model that include only strain and variety should be decent.

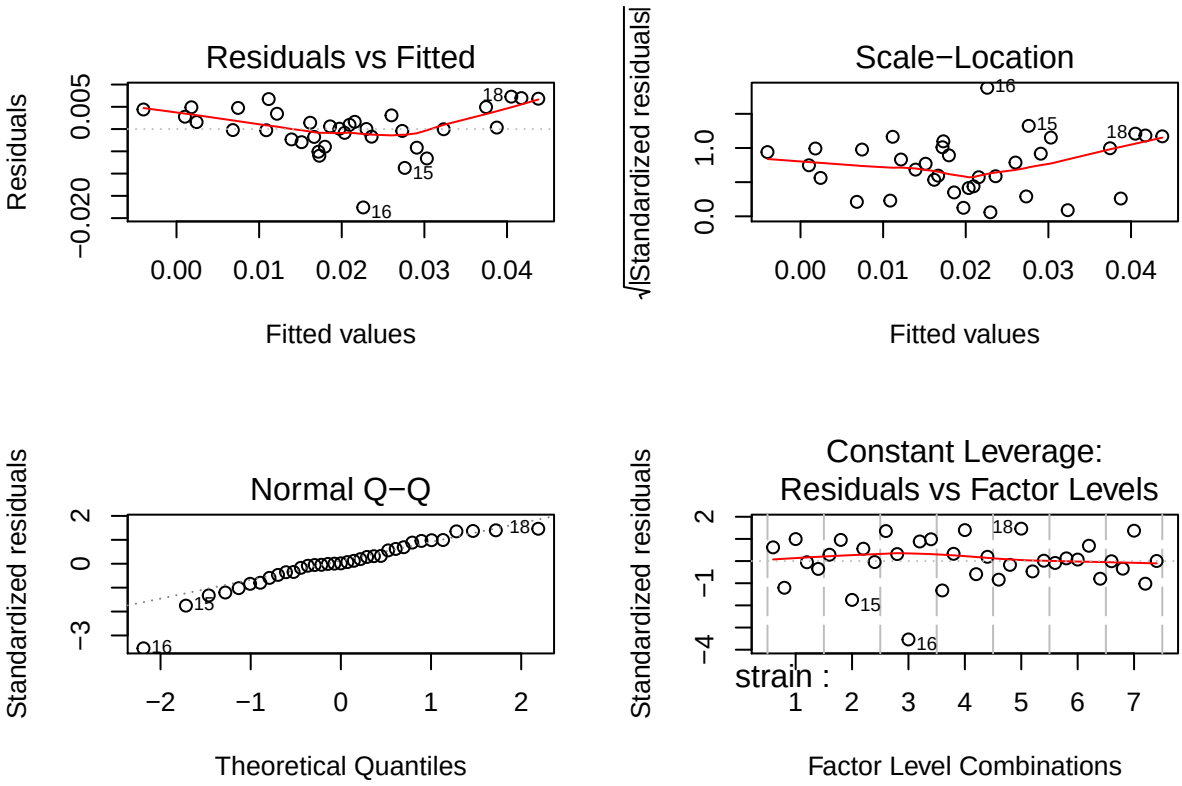
We can see on this plot that there is no data point for variety 1 and strain 1. This was suspicious, so we checked with the dataset online and the first row of the dataset is indeed missing. So we manually reinsert the missing value.

Now we can try to fit our model now. We start by doing an analysis of variance.

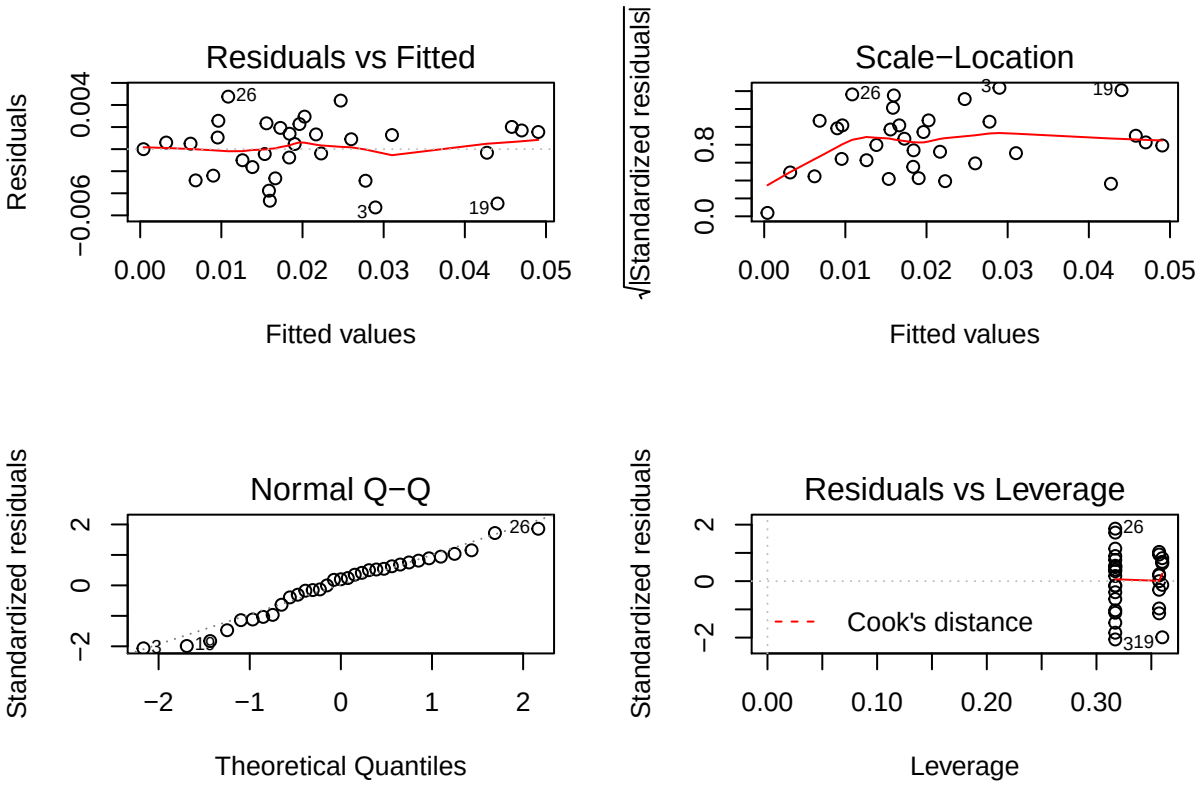
```
##             Df    Sum Sq   Mean Sq F value   Pr(>F)
## strain       6 0.0018670 0.0003112   8.618 4.70e-05 ***
## variety      4 0.0030056 0.0007514  20.810 1.63e-07 ***
## Residuals   24 0.0008666 0.0000361
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By looking at the p-values, we can conclude that strain and variety variables must be included in our model. The backward elimination suggest that weight,radius and number of days are not significant.

The fact that the number of days is insignificant also suggest than the rate is constant in time, which justify using the rate of fungal invasion as our response instead of the fungal radial advance.



The row 16 is clearly an outlier. We decided to remove it along with the row 15 (also an outlier) because they also have too much influence in our variable estimates. #dire qu'ils sont trop influents ca justifie de les enlever un peut, jsp pas pouquoi mais le cook distance plot est remplacé par un truc chelou donc pas sure que se soit vrai, mais les valeur en strain 3 and strain 2 et variety 3 change pas mal en enlevant ces points



Now the different diagnostic plots suggest that the model fit the data much better now, so we decide to use it as our final model.

There are numerous shortcomings to our model. The first one is the clear lack of data, having only one data point per variety and strain is cleary too low, which means any outlier greatly influence our model.

Another problem is that our model does not include any notion of "special interaction" between some stains and variety. For example, it is possible some varieties of apple have more immunity against some kind of fusarium strains than others. This might explain why we have outliers with our model.

Even though we concluded that the rate is constant in our dataset, it is still possible that we were wrong considering the low amount of data we have and that apples with the same varieties have the same number of days. In this case, using the rate as our response is not really relevant.

The final problem is that our model does not give us much insight on the fungal expension. Our model try to predict the rate of fungal expension for a given race without taking care of the temperature or maturity of the apple. For exemple, we do not know if the race of the apple matters a lot or if any variation we might see between those races is caused by some other factor like the maturity of the apples at the start of the experience.