Project Applied Biostatistics

ANOVA problem data:A1

Santiago Anton Moreno Ejub Talovic

04/05/2020

Introduction

Our goal is to find a model that describe the rate of fungal invasion of 5 varieties of apples for 7 fusarium strains. Our dataset has 34 observations with 7 variables/columns. The different strains and varieties will be referenced with numbers.

List of variables:

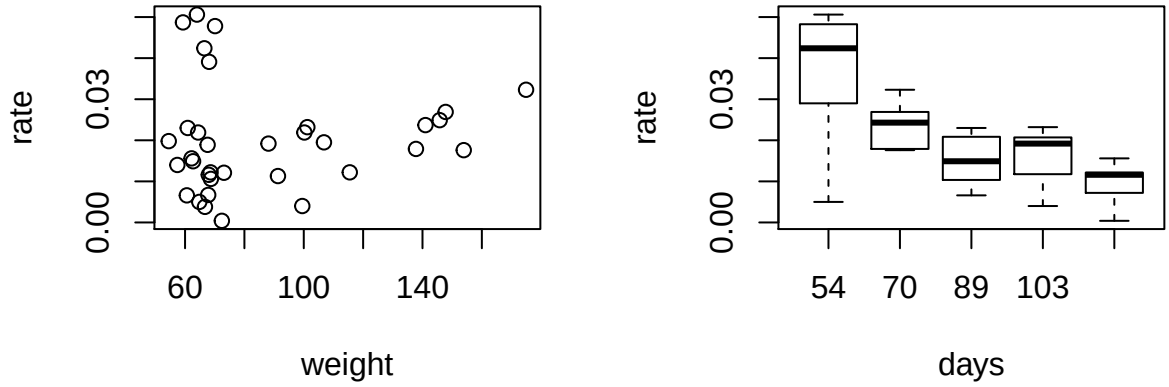
1. Variety
2. Fusarium Strain
3. Days
4. Apple Weight (grams)
5. Radius (cm)
6. Fungal Radial Advance (cm)
7. Rate of advance (cm/day)

List of varieties of the apples:
1=Bramley's Seedling, 1925-26 @12C. 70 Days
2=Bramley's Seedling, 1924-25 @12C. 103 Days
3=Cox's Orange Pippin, 1924-25 @12C. 54 Days
4=Cox's Orange Pippin, 1924-25 @3C. 138 Days
5=Cox's Orange Pippin, 1925-26 @12C. 89 Days

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 for our model. We decide to use the rate, we will justify it later in the notebook.

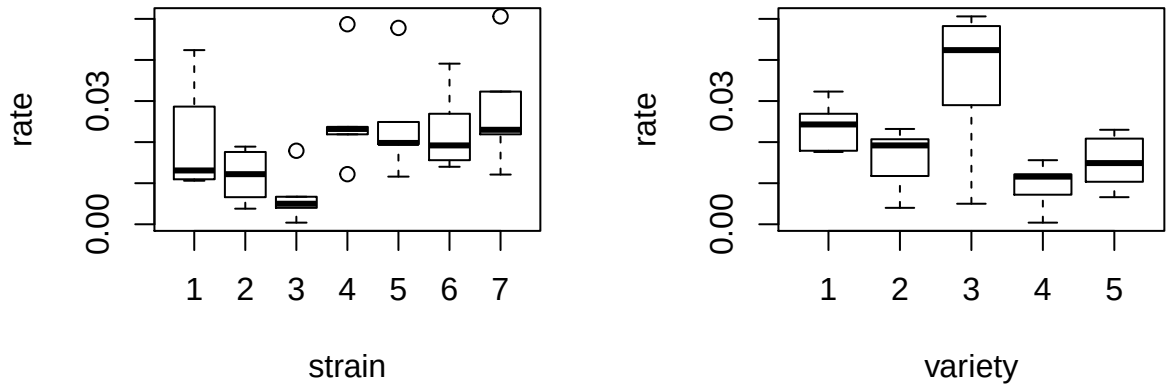
Now we can look at some plots to check if there are some trends we should pay attention to.

Exploratory data analysis



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 reached the same conclusion with the radius, which is to be expected as the weight and radius in our dataset have a correlation of 0.97.

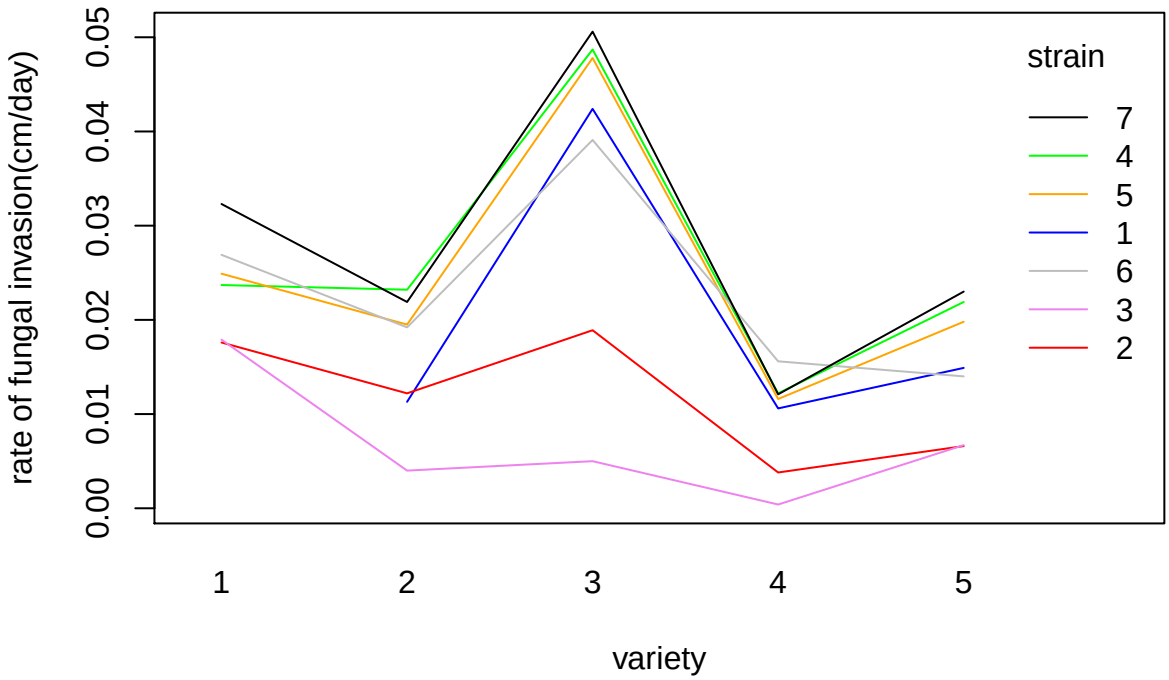
It seems that higher number of days induce lower rates of fungal invasion. However we must keep in mind that apples with the same varieties have the same number of days in our dataset, so we must be conservative with our conclusions.



As we can see in the plots above, the homoskedacity property is violated thus we should be careful when doing analysis. Normality also seems violated as we see that the boxplot for some strains or variety are clearly non symetric which suggest non-normality of the rate.

By taking a deeper look at the description of the data in the introduction, we noticed that variety 3 and variety 5 are the same kind but variety 5 having higher number of days. However not only variety 5 has rates much lower than variety 3, but it also has a lower fungal radial distance. It does not makes sense unless we assume that the conditions of the experience were not exactly the same. For example,

the maturity of the apples could be different at the start of the experience for those two varieties of apples. It 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 have different rates of fungal invasion. We see that all the lines have more or less the same shape but with different scalings except maybe for strain 3.

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

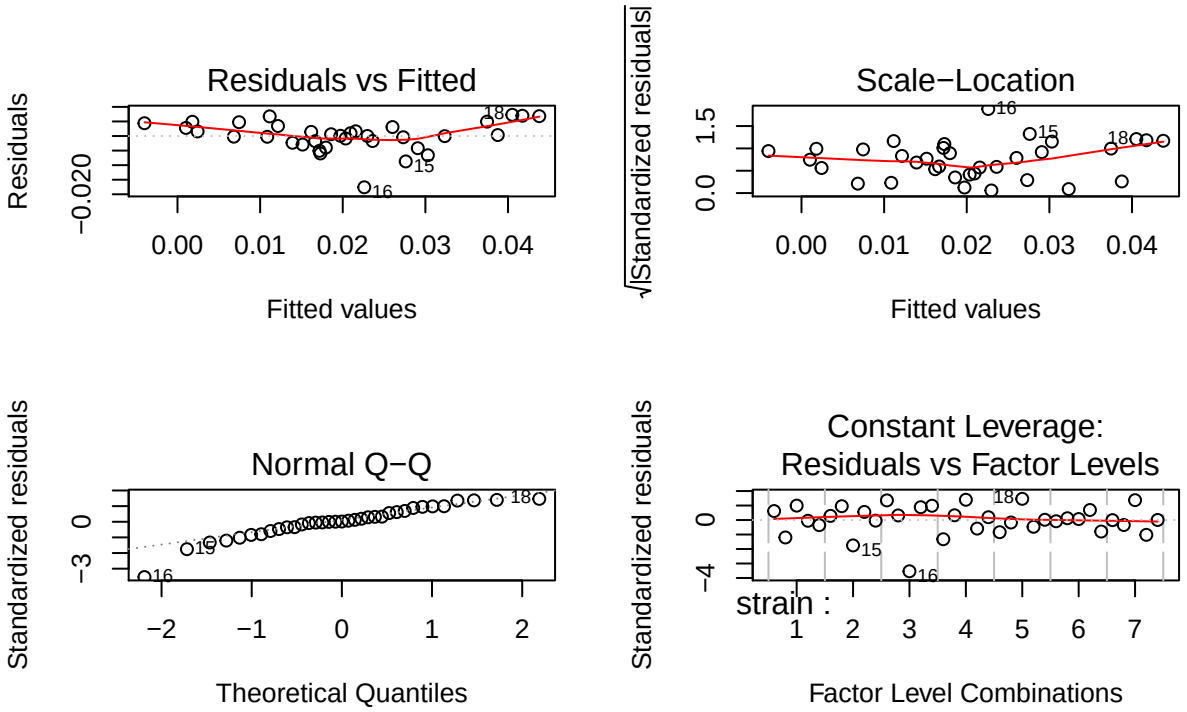
Model fitting

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

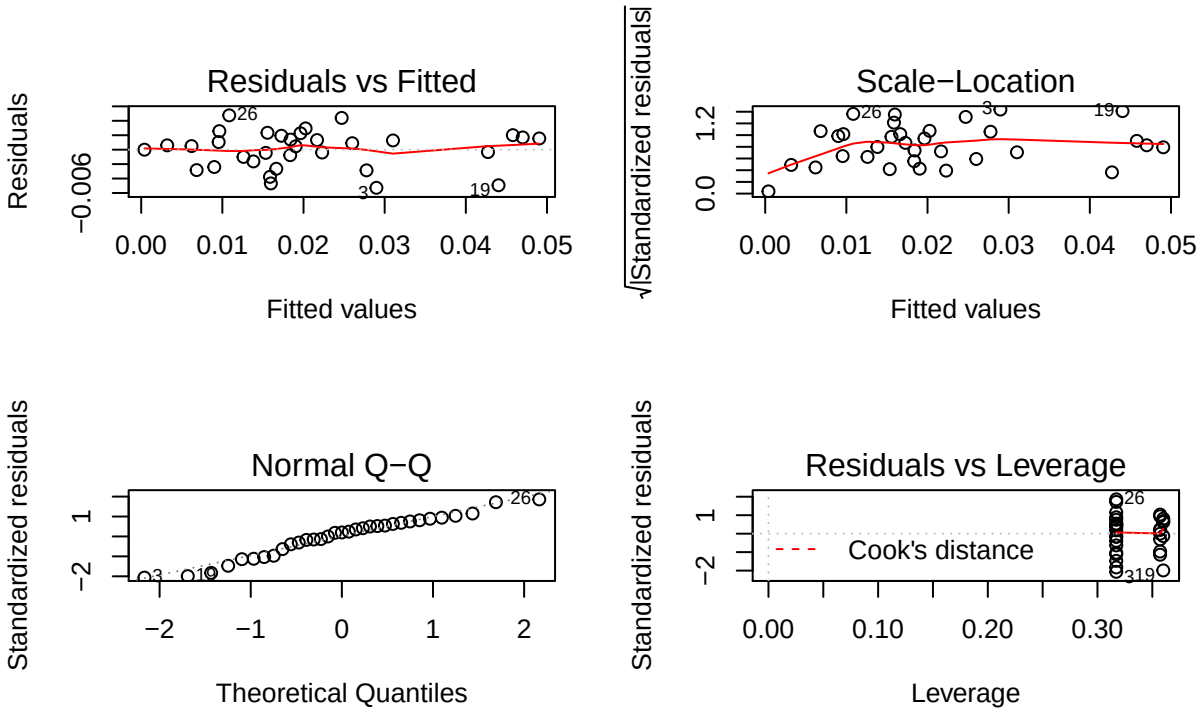
```
##             Df    Sum Sq   Mean Sq F value   Pr(>F)
## strain       6 0.0018670 0.0003112   8.618 4.70e-05 ***
## variety      4 0.0030056 0.0007514  20.810 1.63e-07 ***
## Residuals   24 0.0008666 0.0000361
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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 have too much influence in our variable estimates.



Now the different diagnostic plots suggest that the model fit the data much better now, so our final model formula is: rate ~ strain + variety with the following coefficients:

```
##  (Intercept)      strain2      strain3      strain4      strain5      strain6
##  0.024703714 -0.006343000 -0.009143000  0.004280000  0.003060000  0.001300000
##      strain7     variety2     variety3     variety4     variety5
##  0.006320000 -0.008728571  0.018024286 -0.015157143 -0.009357143
```

We also checked the p-values for the different parameters estimate and it is not unreasonable to set them to 0, but it would not change much in regard to the interpretation of the model or fitting to the data. It would only mean that some strains have the same effect on the rate of fungal invasion than others.

Final model and conclusion

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 tries to predict the rate of fungal expension for a given variety without taking care of the temperature or maturity of the apple. For exemple, we do not know if the variety 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.

In conclusion, we need better quality and higher amount of data to be able to find a good and informative model that can predict the rate of fungal invasion for other apples or at least give usefull information about the resistance of a given variety against a given strain.