Project A1 : Fungal invasion of the apple fruit

Auriane Cozic, Thibault de La Taille, Eloi Littner, Audrey Ménaësse

Introduction

Apples are among the best-preserved fruits. However, after a while fungi strains are inevitably starting to grow. In their 1928's study, Gregory et Horne[1], the conditions of infection, the progression of the invasion and the final stage of invasion were studied for two varieties of apple : the Cox's Orange Pippin from Burwell, Cambridgeshire and Bramley's seeding from 6 different localities. The aim of our study is to compare the infecting power depending on the conditions of infection and the resistance to invasion of different varieties of apples and their characteristics.

Exploratory data analysis

Data description

The dataset under study describes the fungal invasion of 35 apples. It contains 7 different variables, namely the characteristics of the apple (variety, weight, radius), the strain of the fungi separated in 7 types (A, B11, B111, C1, C21, C3 and D) and the measurements of the infection (days after infection, fungal radial advance, rate of fungal advance).

The variable *variety* contains several metadata information. Therefore, in order to have explicit variables in our model, we decided to split it into three new variables : the variety of the apple, the year when the experiment was conducted and the storage temperature.

The study involves 2 different varieties of apples and 7 fungi strains. The data results from 5 different experiments started either in 1924 or 1925. 4 of them were performed at the same temperature (12 °C), and the last one at 3 °C. The duration between the infection and the measurements is different for each experiment.

[1]Gregory, F. G., and Horne, A. S. "A Quantitative Study of the Course of Fungal Invasion of the Apple Fruit and Its Bearing on the Nature of Disease Resistance.–Part I. A Statistical Method of Studying Fungal Invasion." Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character, vol. 102, no. 719, 1928, pp. 427-443

Outliers

While exploring the data, we detected three outliers. The Cox apples *18* and *21* have a fungal radial advance greater than the apple radius. As it does not have any physical sense or meaningful explanation, they were removed. The infection did not develop in the apple *17*, as the fungal radial advance is very lower to any other apple. Being the only such case, we considered it as an experimental error and also removed it from the dataset.

Correlations and redundancy

Then, we tried to identify correlations between variables prior to trying to model the infection power, as the presence of highly correlated explanatory variables would complexify the model without bringing any valuable information. Wee noticed that the apples' "weight" and their "radius" were highly correlated (based on Pearson's correlation coefficients), as shown below. This makes sense based on the physical relationship between these two, and we decided hereafter to ignore the "radius". Furthermore, we found a good correlation between the "fungal radial advance" and the "rate of advance", which also makes sense given how the latter was calculated. Since we are more interested in an infection power, independently of the time elapsed, we chose the "rate of advance" as our variable of interest and ignored the "fungal radial advance".

Table 1: Pearson's correlation coefficients

	weight	radius	fungal radial advance	rate of advance	days
weight	1.00	0.97	0.17	0.24	-0.38
radius	0.97	1.00	0.18	0.22	-0.28
fungal radial advance	0.17	0.18	1.00	0.81	-0.29
rate of advance	0.24	0.22	0.81	1.00	-0.71
days	-0.38	-0.28	-0.29	-0.71	1.00

Then, the factor "days" always explained unambiguously the "year", "temperature" and "apple variety" variables, which is coherent with the way the different batches of apples were monitored. Hence, we removed the "days" variable.

Infection modeling

Full data

The exploratory data analysis enabled us to disentangle the non-redundant, independent parameters for modeling the progression of fungal infection in the apple. More precisely, we identified the infection rate of advance as the most reliable reporter of the infection strength

and the variety, *Fusarium* strain, apple weight, storage temperature and year of experiment as potential explanatory variables.

To select the most meaningful variables, we first fitted a full linear model, using all the independent parameters cited above. Then, we used the stepwise AIC (Akaike information criterion) method for model selection. This method serially checks if deleting or adding a variable to the model improves its AIC score until a minimum is reached : the lower the AIC score, the better the model.

This method, when applied on the full data (including the outliers) yielded the following model with AIC score -317.68 :

$$RateOfAdvance = \beta_0 + \beta_1.weight + \beta_2.\mathbb{1}_{variety=Cox} + \beta_3.\mathbb{1}_{temperature=12} + \sum_{i=2}^{7} \beta_{4,i}.\mathbb{1}_{strain=i} \quad (*)$$

Only "year" was removed from the explanatory variables, showing that no seasonality effect can be inferred from the data. Looking at the Cook's distance diagnostic plot confirmed the fact that samples 18 and 21, previously identified as non-sense data, have a strong influence on the model. Hence, we decided to do a further stepwise model selection using filtered data as described in the explanatory data analysis.

Data without outliers

This time, the final model also included the variable "year", but with a tremendously high p-value ($Pr(>|t|) = 0.21363$). To remain consistent with the model obtained when the whole dataset was used, we chose to run a final stepwise model selection without the "year" variable. Removing "year" did not change fundamentally the obtained AIC (-303.1 instead of -303.52 when "year" was included) and even slightly reduced the model's global p-value (0.002044 instead of 0.002403). Furthermore, an ANOVA of the two models yielded no significancy for the addition of year. Our final model is then also described by $(*)$, with parameter values as stated in the table below. To gauge the consistence of our model and the influence of outliers, we included in that table the parameter estimate when fitted on the full data (last column, "FullData"). No important change could be observed.

Table 2: Coefficients of the model obtained by the step AIC method

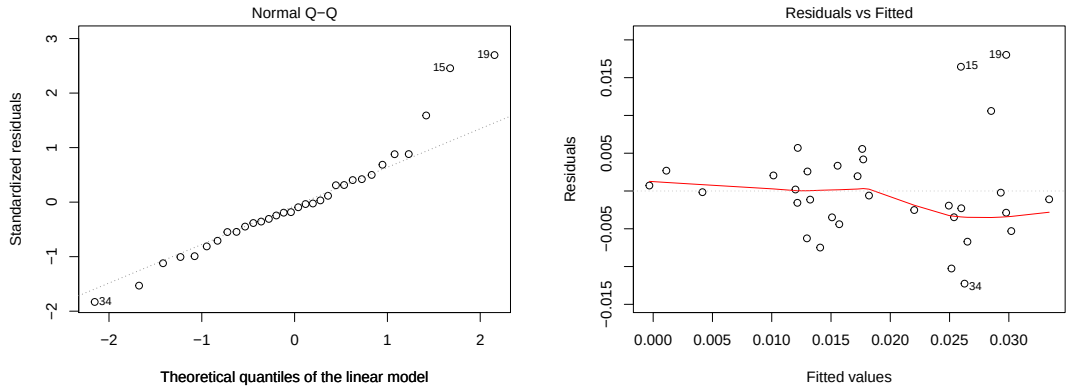
	Estimate	Std. Error	t value	Pr(>|t|)	FullData
(Intercept)	-0.0177	0.0113	-1.5698	0.1307	-0.0197
varietyCox	0.0155	0.0057	2.7347	0.0121	0.0174
strain2	-0.0106	0.0049	-2.1680	0.0412	-0.0106
strain3	-0.0133	0.0052	-2.5443	0.0185	-0.0147
strain4	-0.0001	0.0052	-0.0271	0.9786	0.0047
strain5	0.0031	0.0049	0.6267	0.5373	0.0031
strain6	0.0022	0.0049	0.4514	0.6561	0.0021
strain7	0.0002	0.0052	0.0295	0.9767	0.0053

	Estimate	Std. Error	t value	Pr(>|t|)	FullData
weight	0.0002	0.0001	2.7618	0.0114	0.0002
temperature12	0.0143	0.0038	3.7655	0.0011	0.0172

Evaluation of the model

Residuals

In order to evaluate the model found previously, we first checked the QQ-plot of the fitted data.



(a) QQ-plot (b) residuals

Figure 1: Diagnostic plots

The tail of residuals is composed of a very few data points which are significantly far from the model. These two outliers (15 and 19) were already noticed during the exploratory data analysis, as having particularly advanced infection stage.

The study of the residuals themselves is also important to validate the model. Indeed, if the residuals do not carry any information relevant for the model, their mean should be zero and no pattern should appear, which is the case in the following plot.

Based on these plots, we could confirm that the dataset was then well represented by the linear model described aboc.

Homoscedasticity

To check that the dataset is not biased, we verified that the variance was not dependent on the conditions of the experiment. Dependencies of the residuals on days were plotted, as the *days* variable well differentiates the apples based on the experiment they were part of.

No pattern could be observed on the graph, which confirmed the homoscedasticity of the dataset.

Results and Discussion

We identified 4 characteristics influencing the course of infection.

Apple variety. Infection progressed more slowly in Cox apples than in Bramley apples. This seems to show that they have different resistance to the fungal invasion and that one should prefer Cox's Orange Pippin than Bramley's seeding to avoid infection.

Fungi strain The fungi strain influenced the rate of fungal advance. More precisely, the two B strains (B11 and B111) are less invasive than all others. Different fungi strain can thus have different infecting power.

Temperature Some environmental conditions, as the storage temperature, play a role in the infection. Very cold temperature °C seems to delay the course of infection, when compared to higher temperatures (12 °C). However, only one experiment was conducted at 3 °C with as few as 7 apples. This represents very few amount of data, which explains the high variance associated. It is thus difficult to infer any general conclusion about the ideal temperature, but we can assess that this variable has a significative impact.

Apple weight The apple weight, and by extension its size, has some impact on the speed of the infection. The heavier the apple, the more rapidly the infection progresses. However, this influence is moderate. We suggest it is due to the infection being less limited by the boundary conditions in bigger apples, and that this variable is not related to the resistance of the apple itself.

The variable year is not relevant to explain the infection power. This observation enlights the reproducibility of the experiments, and support the affirmation that the dataset is unbiased.

Conclusion

We were able, in the given dataset, to identify which variables may be meaningful to explain the differences in the invasive power of different fungi strains in two apple varieties under various conditions. By fitting a linear model, we could confidently explain the rate of invasion by a small number of factors: one apple variety appeared to be less susceptible than the other, cold seemed (but this result is still debatable due to the experiment design) to slow the invasion. Fungi strains were separated into two groups, one being significantly less invasive than the other (strains B11 and B111), and the heavier an apple weighs, the more susceptible to infection it is. We were able to assess the independence of the modeled variables, as well as the normality of residuals and the homoscedasticity, therefore validating our approach.

Questions to the teacher

- Our method uses mainly the AIC score for model selection. Do we have to use the ANOVA score for this assignement (as in lab3), or is it acceptable ?
- Do you validate the part when we remove year, even if it was included in the stepAIC return (when run over the filtered dataset) ?