Data analysis 4: The effect of grazing sheep and the age of the stump on the productivity of oak coppice

B030831

June 28, 2021

1 Introduction

In this analysis, we are interested in a study carried out in Greece which looked at the effect on the productivity of oak coppice of allowing sheep to graze around the stumps. It also looked at the effect of the initial age of the stumps. The study consisted of 150 stumps in total, where each stump had a certain number of stems (ranging from 1 stem to a maximum of 22 stems). The height and the diameter of the stems was recorded for each stump, both in 1996 and 2001, in order to compare the evolution under the conditions of interest (sheep grazing and initial age of the stump in 1996). Note that it is possible that the number of stems for a given stump in 1996 differs in 2001. Out of the 150 stumps, there were 50 stumps of age 1 year old, 50 of age 4 and also 50 of age 7, at the start of the study (1996). For each age, the stumps were placed in the same location. Stumps of age 1 in location 1, of age 4 in location 2 and age 7 in location 3. Within each location, 25 stumps were grown in an environment where sheep were allowed to graze around them. Similarly, the 25 other stumps were in an ungrazed environment. Then, within each environment (grazed/ungrazed), there were 5 plots with 5 stumps on each plots. There were thus 30 plots in total. In oak coppicing, one is interested in the total volume of wood. We thus calculate the volume of wood per stem (we take a stem to be a cylinder) in order to calculate the total volume of wood per stump and compare the measures in 1996 to 2001. Firstly, we will explain what are the variables and which model we consider. Then the we will analyze the model, decide whether it is the best fit, and examine what are the effects of stump age and grazing conditions on the oak productivity. Lastly, we will consider a *log* transformation on the response (total volume per stump in 2001) since we will see that the homogeneous variance assumption is violated.

2 Statistical analysis

2.1 Model variables

The interest of the study is to examine whether the initial age of the stump and the grazing condition affect productivity of oak coppice. In other words, we want to compare the measures from 1996 with those in 2001, for each combination of age and grazing conditions. Age and grazing conditions are naturally included as factors. One cannot measure the location effect since we only have one particular age per location. We need to make sure there was no significant difference in the initial stump volume for each age. We cannot take the initial volume as a covariate since we would discard the effect that the initial age had on the initial volume. We thus need to take into account for the age factor in our covariate. Therefore, for each stump volume in location 1, we retrieve the mean stump volume of its corresponding location. Similarly for location 2 and 3. That way, the age factor is accounted for and the remaining difference in initial stump volume (if any) is a random one which can now be taken as a covariate in our model. Lastly, plot number is also treated as a factor variable. We have initial age and grazing conditions as fixed effect factors and the plot effect as a random effect factor. We thus have one continuous variable (the covariate treated as a fixed effect) and 3 factors, which could lead to a very high number of interactions. In our starting model, we only consider the interaction that interests us the most, which is the one between age and grazing conditions. So, our starting model is

$$y_j = \eta x + \alpha_0 + \alpha_1 I_1 + \alpha_2 I_2 + \beta_1 I_3 + \beta_2 I_4 + \beta_3 I_5 + e_j, \tag{1}$$

where x is the covariate. We have that $I_1 = 1$ for age 4 and 0 otherwise, $I_2 = 1$ for age 7 and 0 otherwise. Representing the grazing conditions, we have that $I_3 = 1$ for ungrazed and 0 otherwise. Representing the interaction between age and grazing conditions, we have that $I_4 = 1$ when age is 4 under ungrazed condition and 0 otherwise. Similarly, $I_5 = 1$ for age 7 under ungrazed condition and 0 otherwise. The last term e_j stands for randomness that comes from each plot, $j = 1, ..., 30$.

2.2 Analysis on the raw data

Before starting to analyze to date and fit a model, let us look at two plots of the data.

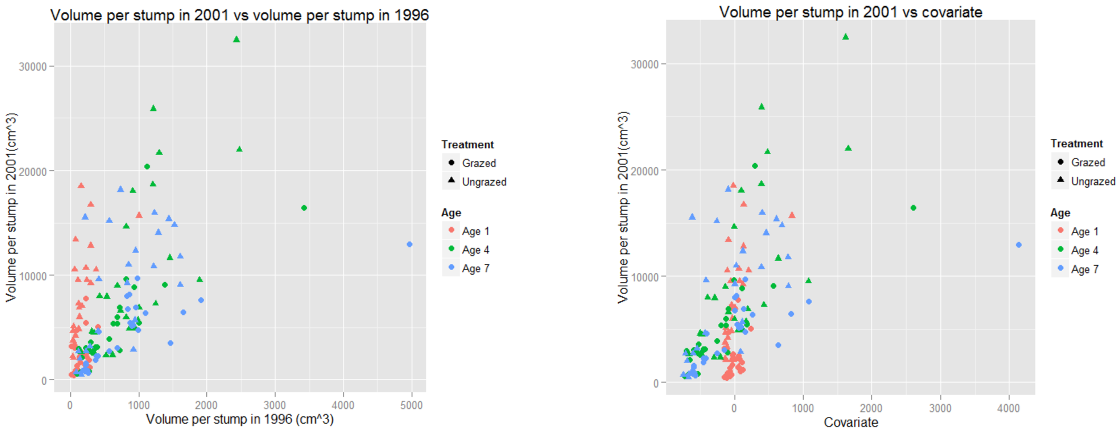

Figure 1: Plot of the volume per stump in 2001 vs the volume per stump in 1996 (right-hand plot) and the covariate (left-hand plot).

Looking at the plot on the left we see that the volume per stump in 1996 is positively skewed while the right-hand plot with the covariate looks more normally distributed. It is more of interest to look at the right-hand side since it is the one with the covariate. We can see that the ungrazed stumps seem to have a higher oak coppice productivity compared to the grazed ones. on the other hand, it is very hard to see if there is an age effect. Let us proceed to fit the model in (1). For each model, we will first look at the fixed effects and then at the random effect once we have decided which fixed effects to keep in our model. The *lmer* R function is used. We get an F value of 0.155 for the interaction between age and grazing conditions. This is very small and we could safely eliminate interaction from the model without further investigation. Nonetheless, let us use a likelihood test ratio to compare the fits of the two models. We get a p-value of 0.8206 and so there is no evidence that the model with interaction offers a better fit than the model without interaction. Our new model is then

$$y_j = \eta x + \alpha_0 + \alpha_1 I_1 + \alpha_2 I_2 + \beta_1 I_3 + e_j. \tag{2}$$

Similarly, we fit this model in R and get an F value of 1.7383 for the age factor. This is not an extreme value and it is thus hard to draw a conclusion by using the F statistics. We will then proceed to a likelihood test ratio in order to compare the fits of the model in (2) and the one without the age factor. We get a p-value of 0.155, there is thus little evidence that age factor is significant. This is a very important result since one of the goal of the study was to decide whether age had a significant effect on the productivity of oak coppice. Therefore the model that offers the best fit to our data is

$$y_j = \eta x + \gamma_1 + \gamma_2 X_2 + e_j \tag{3}$$

where $X_2 = 1$ for ungrazed and 0 otherwise. For model (3), we get an F-value of 64.0 for the effect of the covariate. This is a very large value and thus the effect of the covariate is extremely significant. In other words, there was a large difference due to randomness, in the initial volumes for each stump between the 3 locations. We can say that it is due to randomness since we took into account the effect of the age in our covariate, by retrieving the mean for the relevant age. Similarly, we get an F-value of 13.5 for the effect of the grazing conditions. This is also a large value and we can conclude that it has a significant impact on the oak coppice productivity. We can now look at whether the random effect from the plots is significant. The residual variance for model (3) is 9037460 and the variance for the plots is 10050302. Since the variance for the plots is substantially bigger than the residual variance of the model, we conclude that the plots have indeed some effect. Therefore, we keep it in our model as a random effect. Let us now look at the residuals plot and the normality plot.

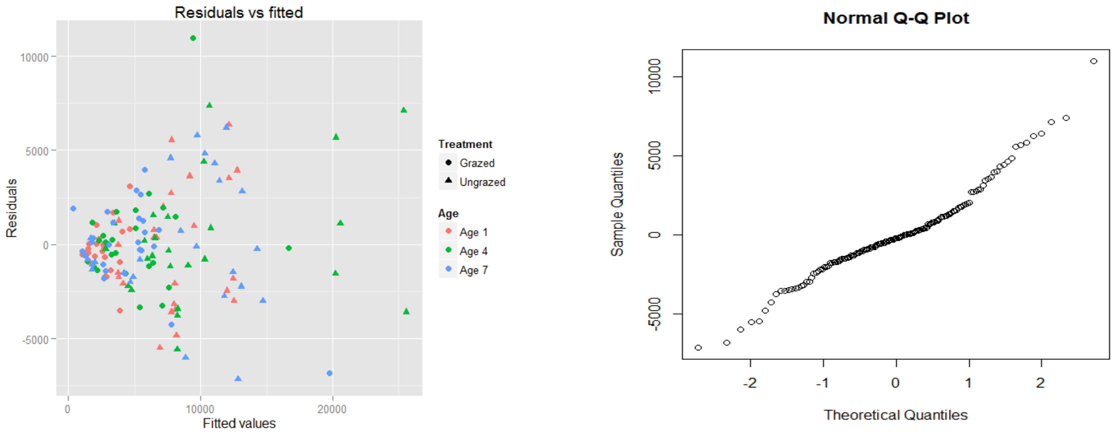

Figure 2: Residuals and Normality plots, respectively.

We see that there seem to be a pattern in the variance, that is the variance increases as the fitted values increase. Look at the right-hand side plot, we see that the that the residuals are not normally distributed as they seem to form a cubic. Let us take a *log* transformation on the response in order to improve the homogeneity of the variance and the fit of the model.

2.3 Analysis after the log transformation

After taking a *log* transformation on the response and having changed the covariate, we fit the same model as in (1). The new covariate is the same as before except that we took the *log* of the total volume in 1996. Using a likelihood ratio test to compare the models with and without interaction, we get a p-value of 0.0358 and there is thus evidence that the interaction effect is significant. We also get an F-value of 89 for the covariate effect, which is very high. All the fixed effects (covariate, age factor, grazing factor and interaction term) are thus kept in our model. For the random effect of the plots, we obtain that the residual variance of the model is 0.2692 and the variance of the plot factor is 0.1742. The variance for the plot is smaller than the residual variance of the model, but it is non negligible, nonetheless. We thus keep the random effect of the plot factor and our model is given by

$$log(y_j) = \eta x_1 + \alpha_0 + \alpha_1 I_1 + \alpha_2 I_2 + \beta_1 I_3 + \beta_2 I_4 + \beta_3 I_5 + e_j, \qquad (4)$$

where the I_i for $i = 1, ..., 5$ have the same meaning as in (1). Note that the same letters are used to denote the coefficients but their estimates are different between (1) and (4).

Let us now look at the residual and normality plots.

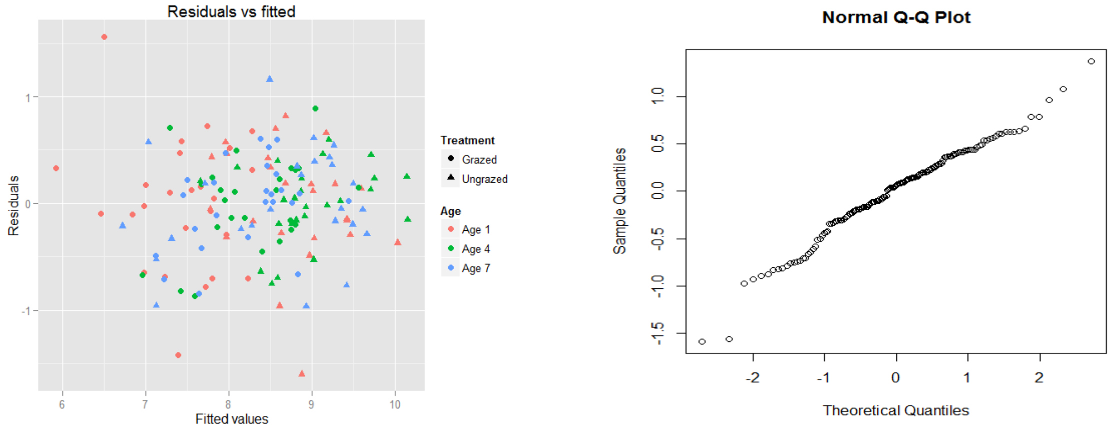

Figure 3: Residuals and Normality plots, respectively.

We see in Figure 3 that after applying the *log* transformation on the response and changing the covariate, the variance is now homogeneously distributed as there is no apparent pattern. The normality plot is also better after the *log* transformation. We can conclude that the assumptions of approximately normally distributed residuals and homogeneous variance are respected and that model (4) is valid. Since we have established that this is the best model for our data, we can now look at the estimated parameters, summarized in the table below.

Parameters	Estimates
η	0.620
γ_0	7.44
γ_1	0.969
γ_2	0.663
β_1	1.31
β_2	0.908
β_3	-0.719

Table 1: Estimate of parameters.

And so, we have that the estimated responses are given by

$$log(\widehat{y_j}) = 0.620x_1 + 7.44 + 0.969I_1 + 0.663I_2 + 1.31I_3 - 0.908I_4 - 0.719I_5 + e_j \qquad (5)$$

We see from equation (5), that the covariate has a positive effect on the oak coppice productivity as expected. Finally, we see that the grazing conditions do have an impact on the productivity. Looking at equation (5), we see that the log of the expected productivity under ungrazed condition is larger by 1.31. This makes intuitive sense since having sheep grazing around the stumps could be a negative influence on the stems's growth. When the stumps have an initial age of 4 year old, we see that the oak coppice productivity is increased by $0.969 - 0.908 = 0.061$. However when the stumps have an age of 7, we see that it has a negative effect on the oak coppice productivity since the interaction term is larger than the contribution of the age factor: $0.663 - 0.719 = -0.056$. Before concluding, let us look at the plot of the data with the transformed response in order to recognize the features of equation (5) in our data.

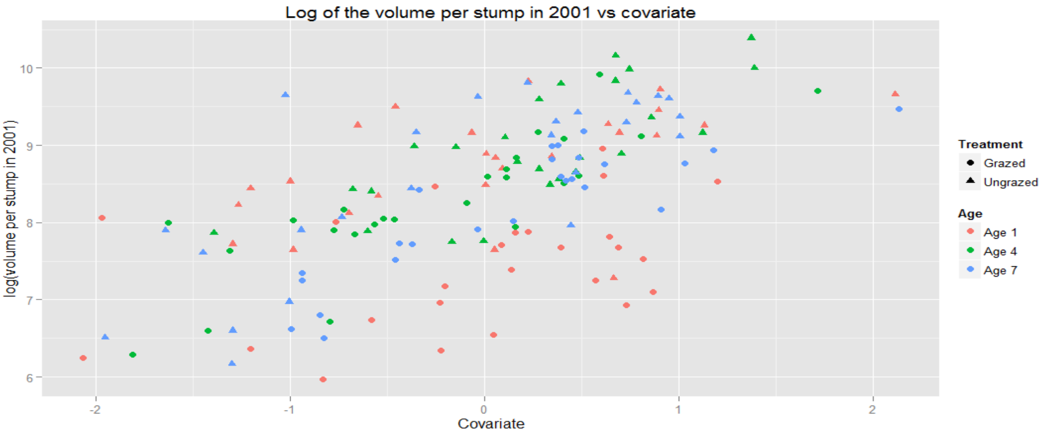

Figure 4: Log of the volume per stump in 2001 vs the covariate for age levels and grazing conditions.

We see from Figure 4 that indeed the *log* of the volume in 2001 is higher under ungrazed conditions. One can also see that the *log* of the volume seems to be higher for age 4 than age 1 and 7, for an equal covariate.

3 Conclusion

We firstly analyzed the data without transforming the response. The independent variables were: covariate, age factor, grazing conditions, plot factor and an interaction term between age and grazing conditions. The plot factor was the only random effect, all other variables were fixed effects. After analyzing this relationship, we found that the age factor and the interaction term were not significant and thus removed from the model. However, after choosing the model that best fits the data, we saw that it did not comply with the homogeneity of the variance and did not have approximately normally distributed residuals, as shown in Figure 2. These assumptions being violated, model (3) was not deemed valid for our analysis. We then proceeded to take the *log* of the response and the *log* of the volume in 1996, in the hope that it would show an improvement. Before looking at the residuals, we needed to choose which model fitted the data best. We found this time that the best fit was by keeping the interaction term. All the fixed effects were significant and the random effect of the plots was also kept in our model. We found that sheep grazing around the stump had a negative effect on the oak coppice productivity. Furthermore, the stumps showed the best productivity when they were 4 year old at the beginning of the experiment in 1996. In conclusion, the best oak coppice productivity is achieved in an ungrazed environment on stumps that are 9 year old (since they were 4 year old in 1996 and the last measurements were made in 2001).