Building of a predictive model of the exploitable volume of timber with the regression method

1 Abstract

Using the trees' dataset of R in the SMIR package we will construct a model that will allow us to predict the exploitable volume of timber of new black cherry trees when their height and diameter are given.

2 Introduction

Below we present the data that will use to build the model. We will study three different models, and also show how automatic model selection can be done and finally, we will explain how to select a specific model using the cross-validation method that compares the prediction errors.

3 Data presentation

The trees file contains 31 cases of cherry trees in which case the following variables are recorded: the exploitable tree of volume (in cubic feet), the height (in feet), and the diameter of trees(in inches). To process the data of this file we convert the volume to lt, the height and diameter in cm.

4 Statistical data analysis

4a Model 1

Table 4.1 presents the correlations of observations of interest to us, using the Pearson correlation coefficient and the cor()function.

Table 4.1: Correlations of observations with Pearson coefficient metric.

	V	H	D
V	1.0000000	0.5982497	0.9671194
H	0.5982497	1.0000000	0.5192801
D	0.9671194	0.5192801	1.0000000

From Table 4.1 we can see that the highest correlation of cherry timber volume is the one with diameter D. Figure 4.1 illustrates graphically the relationship between them.

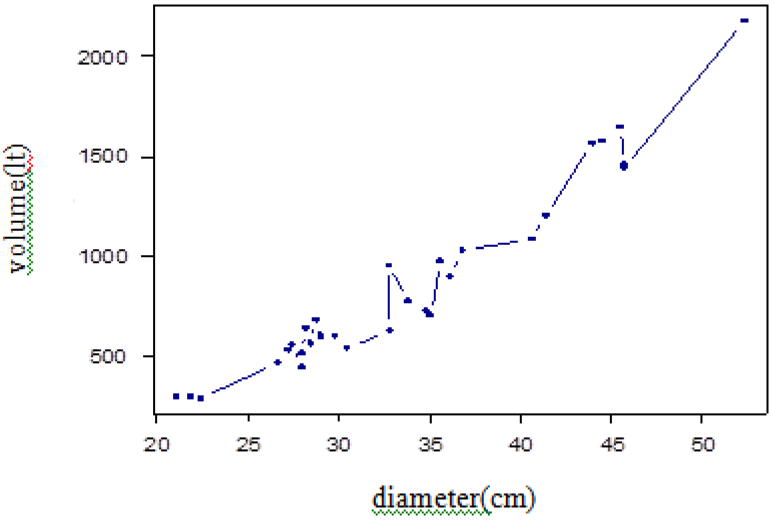

Figure 4.1: Tree volume in relation to the diameter.

We will assume that the relationship between them is linear and we will start the analysis with model 1: V= b_0 + b_1 D. The regression of the volume V on the diameter D, gives the following results: b_0 =-1046.122 and b_1 =50.476. The p-value=7.62*10^{-12} ≅ 0 rejects the hypothesis H_0 : b_0 =0, while the p-value=2*10^{-16} ≅ 0 rejects the hypothesis H_0 : b_1 =0, hence the diameter D adequately interpret the volume V. This is confirmed by the implementation of confint() function, where we can see that zero is not included in the corresponding confidence intervals. The R^2 = 0.9353 is satisfactory as is supported by the F- test (p-value= 2.2*10^{-16} ≅ 0) and the fact that the tilt of the straight line is significant. The leverage ratio is 0.1935484, so we assume that we do not have outliers.

4b. Model 2

We will try the automatic model selection to see if the model 2: V= b_0 + b_1 D + b_2 H, is better than the last. We will use the backward elimination method. The automatic selection consider that the model 2 is better than the model 1, thus we can see that we have improved value for the coefficient of determination R^2 = 0.948. The p-value = 0.0145 rejects the H0 hypothesis: b_2 =0, so height also contributes to the volume, having a coefficient b_2 =0.3152. The values of coefficients b_0= -1642.0276 and

b_1=52.4883 change slightly compare to model 1. The Harvey-Collier test gives p-value = 0.004542, which means we have a deviation from linearity, that is, our model is not exactly linear. The Breusch-Pagan test gives p-value = 0.2911, so our model does not have homoscedasticity. Durbin-Watson and Breusch-Godfrey tests with p-value = 0.00917 and p-value = 0.5517, respectively, show that there is no autocorrelation between the coefficients. Finally, from the stats.d() function for the residuals, we observe that model 2 is a normal distribution as it can be seen in Figure 4.2.

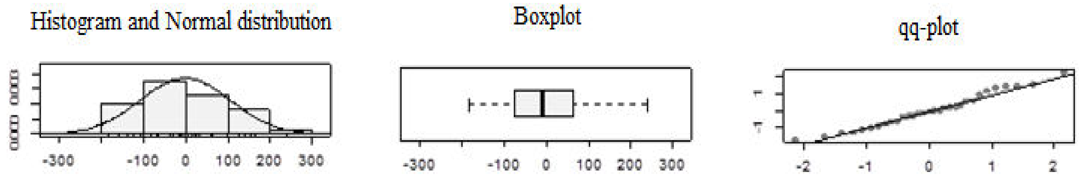

Figure 4.2: Graphs of Model 2.

4c. Model 3

We will try and another model. The volume of a tree can be simulated either by the volume of a cylinder: V= $\pi \left(\frac{D}{2}\right)^2$ H, either by the volume of a cone: V= $\frac{\pi}{12}$ H D^2 . In both cases, if we logarithm, a relationship of the form will emerge: logV= b_0 + b_1 logD + b_2 logH. The regression of the logarithm of volume logV on logH +logD gives us the following results: b_0 =-8.95360, b_1 =1.98265 and b_2=1.11712. The p-value=1,3*10^{-6} ≅ 0 rejects the hypothesis H_0 : b_0 =0, the p-value=2*10^{-16} ≅ 0 rejects the hypothesis H_0 : b_1 =0 and the p-value=7.81*10^{-6} ≅ 0 rejects the hypothesis H_0 : b_2 =0. This is confirmed by the implementation of confint() function, where we can see that zero is not included in the corresponding confidence intervals. We ascertain that the coefficients of logD and logV are very close to those given by the theory of geometry for the volume of a tree. We also can see that R^2= 0.9777 is larger than the previous models, although the values of the determination coefficients are not directly comparable. Finally, because F = 613.2 >> F_0.05 = 13.3615 the empirical predictability criterion is met. Harvey-Collier test gives p-value = 0.2875, which means we have linearity. Breusch-Pagan test gives p-value = 0.2557, so our model does not have homoscedasticity. The Durbin-Watson and Breusch-Godfrey tests with p-value = 0.4863 and p-value = 0.7039, respectively, show that there is no autocorrelation between the coefficients. Finally, from the stats.d()function for the residuals, we observe that model 3 is a normal distribution as it can be seen in Figure 4.3.

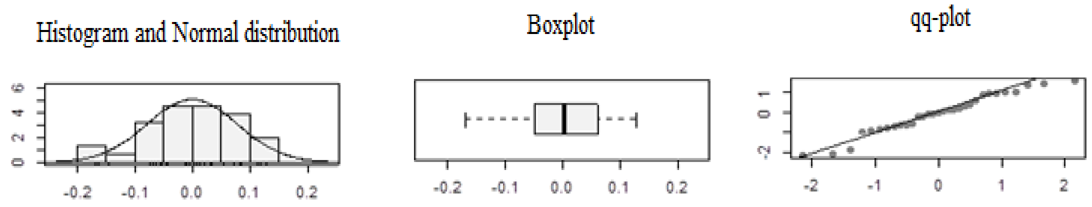

Figure 4.3: Graphs of Model 3.

4d. Model selection

We choose the model that makes the best predictions according to some measures. We decide to check the corrected R^2 values, the normality of the residues and the mean square prediction error of the models 2 and 3. We exclude model 1 because automatic model selection has already selected the model 2 against model 1. The above measures are presented in Table 4.2.

Table 4.2: Comparison of models.

	R^2	Shapiro-Wilk p-value	MSE
Model 2	0.94795	0.6439824	0.1279619
Model 3	0.9776784	0.2782375	0.006320406

We observe that both models have a normal residuals distribution, but Model 3 has smaller errors after the transformation is implemented. Therefore, Model 3 is more suitable for predictions.

5 Conclusions

By applying the regression method on tree dataset we constructed a linear, non-homoskedastic model of logarithms of all variables (model 3), which enables us to predict the exploitable volume of cherry tree timber when we are given their diameter and height. This model is consistent with the theory of geometry.