Probabilities and Statistics Project

Cirstea Ionela-Madalina, Cirstea Natasa-Alexandra, Puiu Ana Maria Alexandra

February 2020

1 Trees data frame description

This data set provides measurements of the diameter, height and volume of timber in 31 felled black cherry trees. Note that the diameter (in inches) is erroneously labelled Girth in the data. It is measured at 4 ft 6 in above the ground.

[1] Girth - numeric Tree diameter (rather than girth, actually) in inches
[2] Height - numeric Height in ft
[3] Volume - numeric Volume of timber in cubic ft

2 Requirements

1. Using the given data frame execute descriptive statistic operations: mean, median, variance, quantiles, box plot, interpretations.

2. Using the given data frame build two regression models (simple and multiple) choosing the predictor and the result variable as you consider. Add to the given data set at least one variable that can be used in more than one regression model. Generate its data using a suitable distribution. Justify your choices and make interpretations. Which model is more suitable? Give at least two arguments.

3. Choose a distribution different from the ones you studied and create two visual representations of the mass function/density (depending on the type of random variable) and the cumulative distribution. What properties can you identify? Which is the practical usage of this distribution? What kind of phenomena does it model?

3 R packages

1. data sets - we'll be using the "trees" data frame, in order to build two linear regression models with one or multiple predictors
2. ggplot2 - we'll use this package to build plots of our models
3. GGally - this package extends the functionality of ggplot2. We'll be using it to create a plot matrix as part of our initial exploratory data visualization
4. scatterplot3d - we'll use this package for visualizing more complex linear regression models with multiple predictors

4 Descriptive statistics

Descriptive statistics is used to describe the characteristics of a data set, which may represent a whole population or a sample of it. Descriptive statistics are divided into measures of central tendency (mean, median) and measures of variability (standard deviation, minimum / maximum, asymmetry).

Descriptive statistics operations

- **Mean** is a central data trend, more precisely a number around which they are spread and which can estimate the value of the whole set. In R, it is calculated by means of the mean () function, which receives as a parameter a data set.
- **Median** is the value that divides a data set into two equal parts, the number of terms on the left being equal to the one on the right when the data is arranged in ascending or descending order. The corresponding function in R is median().
- **Variance** represents the square of the standard deviation (the average deviation of the values in a data set from the average). It can be calculated using the var() function.
- **Quartiles** are values that divide a data set into quarters. They are the averages of the first (first quartile) and last half (third quartile), respectively the average of the whole set (second quartile). They are 25%, 50% and 75%. The function used is the quantities().
- **T**he Boxplot presents the five-digit summary of a dataset: the minimum, the first quartile, the average, the third quartile, and the maximum. Therefore, it is a measure of how well-distributed the data is in a set.

Interpretations of the values obtained

- **Mean and Median**
  - **Girth:** Data are slightly sloping to the right (positive slope distribution), with the mean ($\simeq 13.25$) being higher than the median ($\simeq 12.9$).
  - **Height:** The data are relatively balanced (symmetrical distribution), the mean ($= 76$) being equal to the median ($= 76$).
  - **Volume:** Data are slightly tilted to the right (positive slope distribution), the mean ($\simeq 30.17$) being higher than the median ($\simeq 24.2$).
- **Variance**
  - The small variance suggests that the values are close to the mean and close to each other, in this case being **Girth** ($\simeq 9.85$) and **Height** ($\simeq 40.6$).
  - The large variance indicates the opposite: the values are far from mean and far one from another - **Volume** ($\simeq 270.2$).
- **Quartiles**
  - **Girth:** The values of the quartiles (11.05, 12.9 and 15.25) indicate that the variance is higher between the higher values of the set than between the smaller ones and that the distribution is slightly inclined positively.
  - **Height:** Since the median is the average of the first and last quartiles (72, 76, 80), the variance of values is uniform, a fact also supported by the equality between the mean and the median.
  - **Volume:** The values of the quartiles being 19.4, 24.2 and 37.3, we can deduce there exists a greater variance between the higher values of the set than between the smallest ones and that the distribution is positively inclined.
- **Boxplot**
  - **Girth:** The boxplot indicates that there are no outliers among the values and that the variation is higher between the larger values.
  - **Height:** No outlier is identified in the boxplot, and the variance is relatively symmetrical.
  - **Volume:** The boxplot suggests that there is only one outlier and that the variance is greater between the larger values.

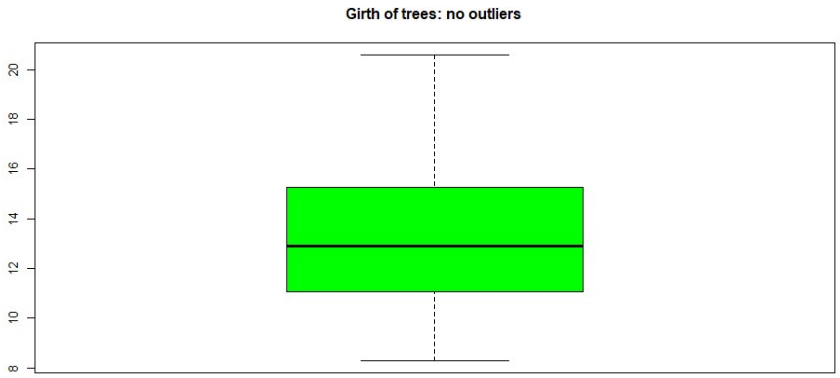

Figure 1: Boxplot - Girth

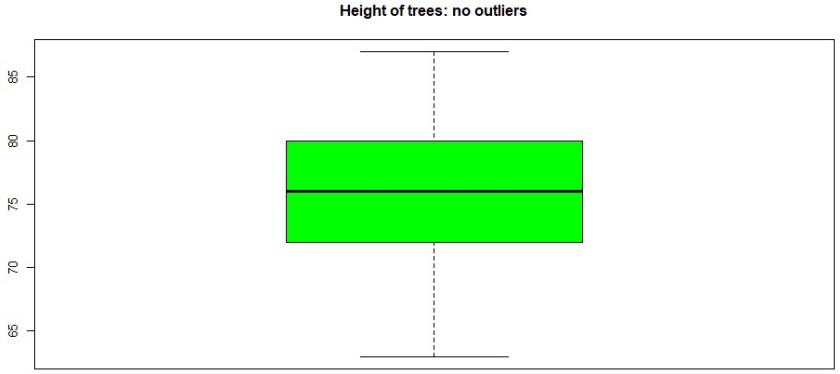

Figure 2: Boxplot - Height

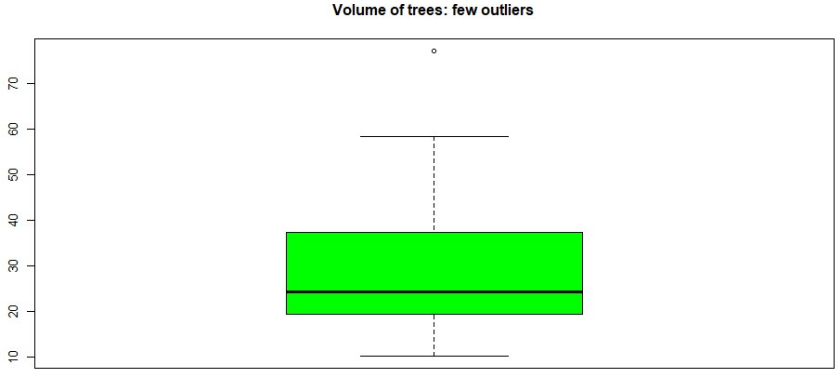

Figure 3: Boxplot - Volume

5 Simple linear regression

General description: We want to see if the girth and height of a Black Cherry Tree influence the volume of the timber that can be obtained.

1. To verify if the data set is suitable for linear regression (in other words, to decide whether we can make a predictive model), we calculate the correlation between girth and volume $\simeq 0.9671194$ (girth will be used as predictor and volume as response), as well as the correlation between height and volume $\simeq 0.5982497$. There appears to be a stronger relationship between girth and volume than the one regarding height-volume as the correlation coefficient girth - volume is closer to 1. Using the ggpairs() function from the GGally package, we create a plot matrix to better visualize how the variables relate to one another. From looking at the ggpairs() output, girth definitely seems to be related to volume: the correlation coefficient is close to 1, and the points seem to have a linear pattern.

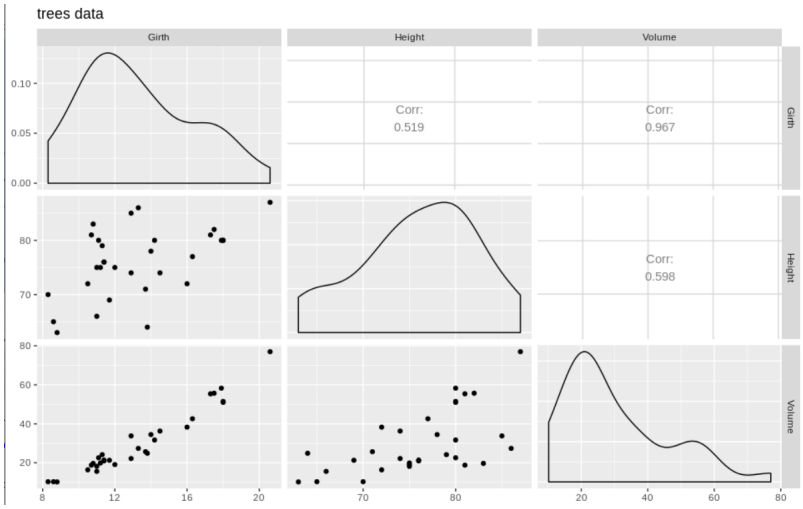

Figure 4: ggpairs() output

2. Next, we will split the data set into a 80:20 sample (training:test), then, build the model on the 80% sample and then use the model thus built to predict the dependent variable on test data. Analyzing the plot for girth and volume variables on training data, we notice an ascending tendency (the bigger the tree's girth, the greater the volume of timber we can obtain).

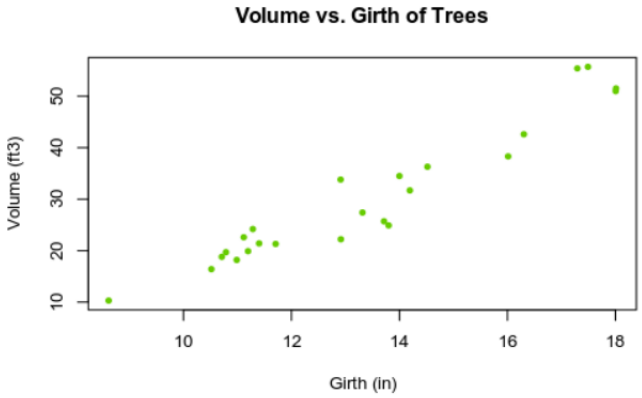

Figure 5: Individual plot for girth-volume

3. Calling lm() function on training data, we build a simple linear model where tree's volume (result) depends only on girth (predictor, independent variable). The lm() function fits a line to our data that is as close as possible to all 31 of our observations. More specifically, it fits the line in such a way that the sum of the squared difference between the points and the line is minimized; this method is known as "minimizing least squares".

```
Call:
lm(formula = Volume ~ Girth, data = trainingData)

Residuals:
   Min     1Q Median     3Q    Max
-7.301 -1.545 -0.235  1.721  6.861

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -32.2183     3.9828  -8.089 4.91e-08 ***
Girth         4.6680     0.2925  15.960 1.40e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.759 on 22 degrees of freedom
Multiple R-squared:  0.9205,    Adjusted R-squared:  0.9169
F-statistic: 254.7 on 1 and 22 DF,  p-value: 1.4e-13
```

Figure 6: Simple linear model summary

4. Then we apply summary() for the created model. From the summary we notice that the pvalue is 1.4e-13. As pvalue is less than 0.05 we conclude that the predictor girth variable is significant. Furthermore, the R^2 value is a measure of how close our data are to the linear regression model. In our case the adjusted R^2 equals 0.9169 and is close to 1 which indicates

a well-fitting model. From the regression line we notice an ascending tendency indicating that a larger volume of timber is obtained from a tree with a bigger diameter.

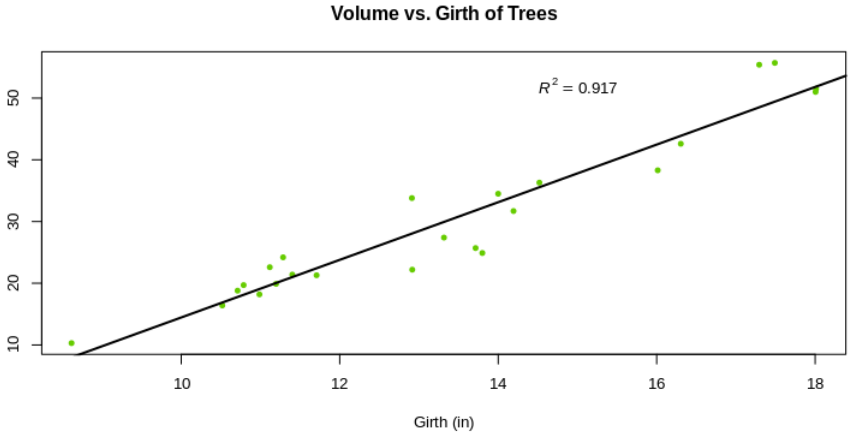

Figure 7: Simple linear regression

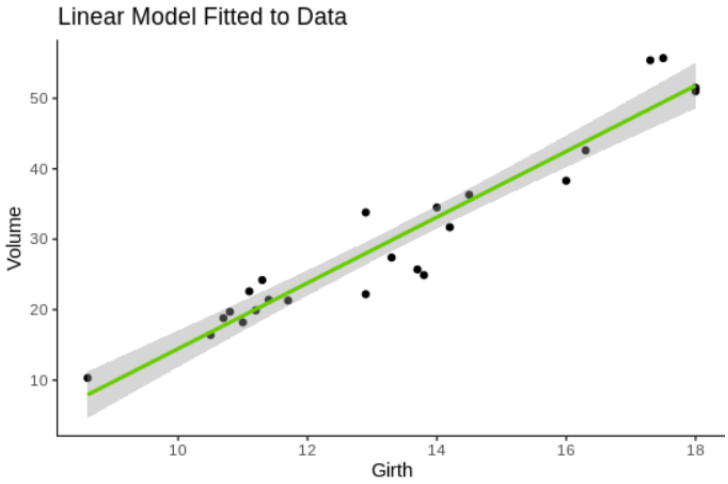

Figure 8: Simple linear regression with confidence interval

5. The gray shading around the line represents a confidence interval of 0.95, the probability that the true linear model for the girth and volume of all black cherry trees will lie within the confidence interval of the regression model fitted to our data. Thus, our data is strong enough to let us develop a useful model for making predictions.

```
> head(actuals_preds)
   actuals predicteds
1     10.3   6.526385
3     10.2   8.860404
7     15.6  19.130091
12    21.0  20.997306
15    19.1  23.798130
28    58.3  51.339562
> min_max_accuracy <- mean(apply(actuals_preds, 1, min) / apply(actuals_preds, 1, max))
> mape <- mean(abs((actuals_preds$predicteds - actuals_preds$actuals))/actuals_preds$actuals)
> min_max_accuracy
[1] 0.8330377
> mape
[1] 0.1798647
```

Figure 9: Predictions for test data

6. We'll use the predict() function, a generic R function for making predictions. predict() takes as arguments our linear regression model and the values of our test data set. By calculating accuracy measures (like min-max-accuracy $\simeq 0.8330377$ - the higher the better) and error rates (MAPE Mean absolute percentage error $\simeq 0.1798647$ - the lower the better), we can find out the prediction accuracy of the model. Since the value of the error rate is very small, and the min-max-accuracy value is close to 1, we conclude that the accuracy of the predictions is considerable.

6 Multiple linear regression

General description: We wonder if we can improve our model's predictive ability if we use all the information we have available (girth and height) to make predictions about tree volume.

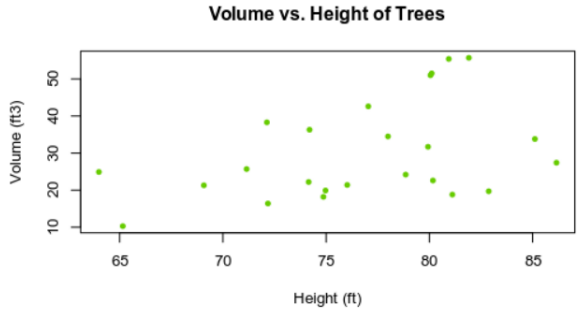

Figure 10: Individual plot for height - volume

1. As seen in the model above, there exists an ascending relationship between girth and volume. But the correlation between height and volume cannot be ignored as it is high enough to be factored in ($\simeq 0.5982497$). By interpreting the plot for height - volume we notice another increasing inclination (but a lower one this time). Thus, we suspect that, for the most part, as the height of the tree increases, so does the volume. Therefore, a better solution for the problem is to build a linear model that includes multiple predictor variables. We can do this by adding a slope coefficient for each additional independent variable of interest to our model: girth and height.

$$volume = \beta_0 + \beta_1 \star (girth) + \beta_2 \star (height)$$

where:

$$\beta_0, \beta_1, \beta_2$$

represents the intercept (= -56.9240), the slope coefficient for girth (= 4.4520) and the slope coefficient for height (= 0.3601)

2. This slope tells us how much the volume will change if the girth increases with one inch or height increases with one ft.

3. Analyzing the output of the summary() function for the newly created model, we can see that both the girth and the height are significantly related to the volume and that the model fits our data well. The value obtained for adjusted R^2 increased from that obtained for the previous model (= 0.9386). Also, the pvalue is lower than the predetermined level of statistical significance (= 7.28e-14), so we know that we have a significant statistical model. Since we have two predictor variables in this model, we need a third dimension to view the model. We can create a 3D scatter graph using the scatterplot3d package. First of all, we make a grid of values for our predictor variables (within our data limit; considering that the size of the data set is extremely small, we must also consider the possibility of our model overfitting, that is, to produce an analysis which corresponds too closely or exactly to a certain set of data and therefore that can not be adapted for additional data, thus failing to make reliable predictions). The expand.grid() function creates a data frame from all combinations of the factor variables.

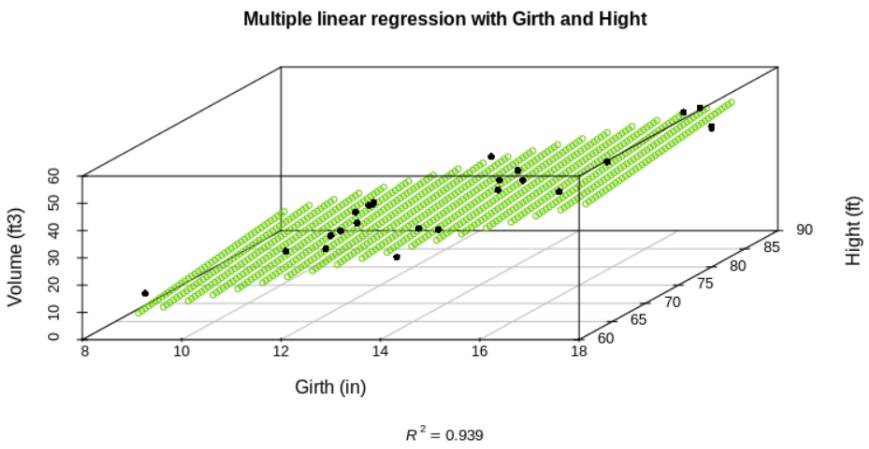

Figure 11: Multiple linear regression with girth and height

7 Multiple linear regression accounting interactions

1. Although we have made improvements, the model we just built still does not precisely reflect the reality. It assumes that the effect of tree girth on volume is independent from the effect of tree height on volume. This is clearly not the case, since we suspect that tree height and girth are related.

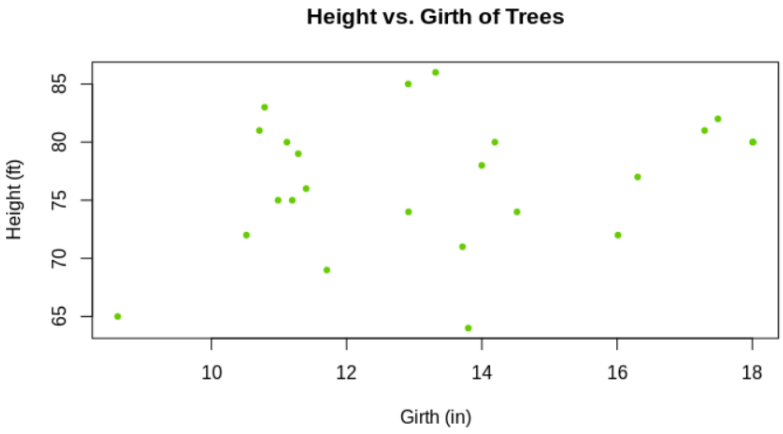

Figure 12: Individual plot for girth and height

2. As we calculate the correlation ($\simeq$0.5192801), is becoming clearer that there is a link between girth and height. Put another way, the slope for girth should increase as the slope for height increases. To account for this non-independence of predictor variables in our model, we can specify an interaction term, which is calculated as the product of the predictor variables.

$$volume = \beta_0 + \beta_1 \star (girth) + \beta_2 \star (height) + \beta_3 \star (girth \star height)$$

3. As we suspected, the interaction of girth and height is significant, suggesting that we should include the interaction term in the model we use to predict tree volume. This decision is also supported by the adjusted R^2 value closer to 1 and bigger than the R^2 from previous model (= 0.9558), the value of p (= 2.564e-14) smaller than previous one, the value of min-max-accuracy (= 0.908044) is higher than the one mentioned in previous model and the value of error rates (mape = 0.1002732) is lower than previous one; all of these suggesting that our model is the best fit for the data from the ones considered.

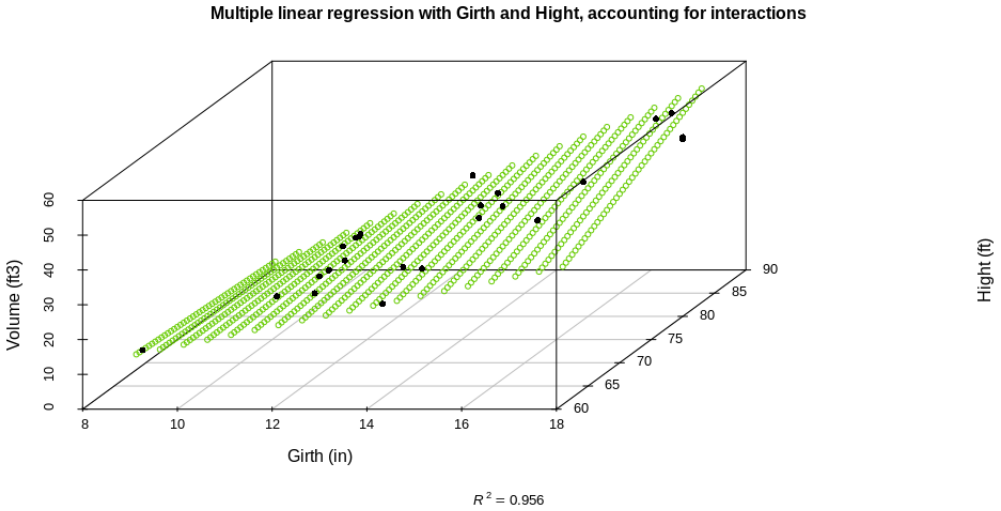

Figure 13: Multiple linear regression accounting interactions

8 Adding new data to the data frame

(a) We wonder if we can improve the predictive ability of the model by adding new data to the observations already existing in the data set. We add a new column in the data frame, a column that will contain the percentage of wood in the tree that cannot be used to obtain timber. In other words, the percentage that represents the naturally occurring defects on the surface of a black cherry tree.

(b) The appearance of defects is favored by many factors: the type and characteristics of the species, the climatic conditions, the geographical position, the destructive action of the flora, fauna or human activity and many other processes. Therefore, it all comes down to the interactions of several hidden processes on a small scale. Given that the appearance of defects results from the summation of several processes on a small scale, we suspect that their distribution is close to that of a normal distribution (specific to nature patterns). The small-scale individual fluctuations caused by each contributing process rarely follow the Gaussian curve. But by aggregating several partially uncorrelated fluctuations, each on a small scale relative to the aggregate, the sum of the fluctuations is smoothed in the Gaussian curve. Also, by combining certain defects, the probability of reducing the volume can increase significantly. Therefore, the variable chosen can significantly influence the volume of timber.

(c) With the help of the normal (rnorm) distribution, we generate a series of 31 values representing tree defects in percentages, as we

suppose they occur in nature. Since the data we need to enter into the data frame cannot be just random values (because the probability of investigations would suffer, since we do not take into account the interdependencies which occur in nature), we must establish a link between defects and at least one feature of the observations in the data frame (Height, Girth, Volume). We sense that, as the girth of a tree grows, so does the time of exposure to environmental factors that could lead to the appearance of defects. Therefore, we choose to correlate the percentage of defects with the diameters already known (we establish the values are correlated in a proportion of 30%) using an auxiliary function[1].

(d) We construct a new multiple linear regression model with interactions to predict the volume (response variable) using our girth and the percentage of defects (predictor variables). The construction process is similar to the previous case.

(e) The last linear model has the value of adjusted R^2 lower than the one obtained in the multiple regression girth - height with interactions, which is to be expected, given that we chose the correlation between defects - girth ($= 0.3$) lower than correlation between height and girth ($= 0.519$).

(f) Conclusions: The best predictive model for our set of observations, of all the ones built above, remains volume-girth-height, as the correlation height-volume ($\simeq 0.5982497$) is much higher than the defects-volume correlation ($\simeq 0.1399002$), given the fact that we started in the construction of the model choosing to correlate the defects and the height in proportion of 30%. This result checks our intuitive observations: the height influences the volume to a greater extent than the defects.

[1] `https://stats.stackexchange.com/questions/15011/generate-a-random-variable-with-a-defined-correlation-to-an-`

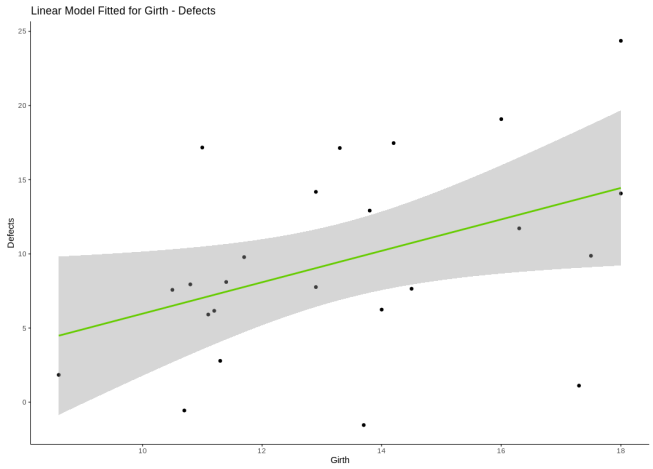

Figure 14: Date (Girth-Defects) - 30% correlation

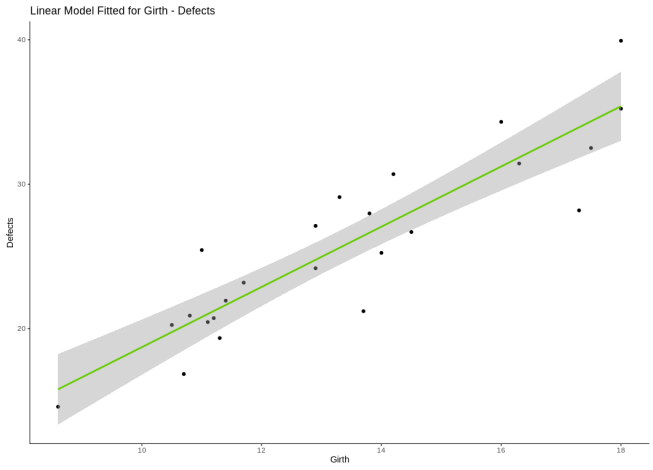

Figure 15: Date (Girth-Defects) - 90% correlation

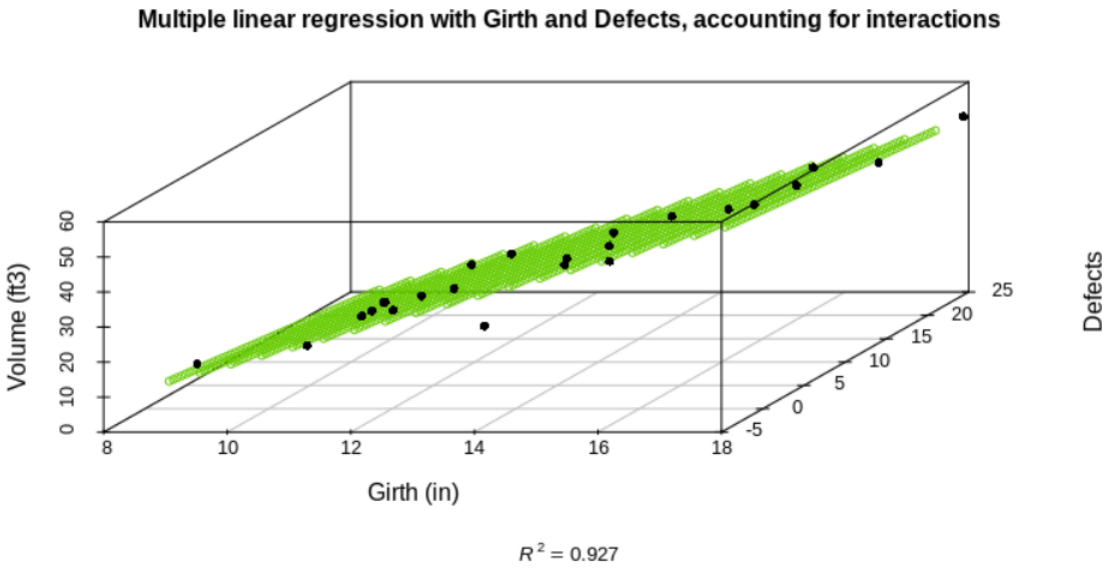

Figure 16: Multiple linear regression accounting interactions

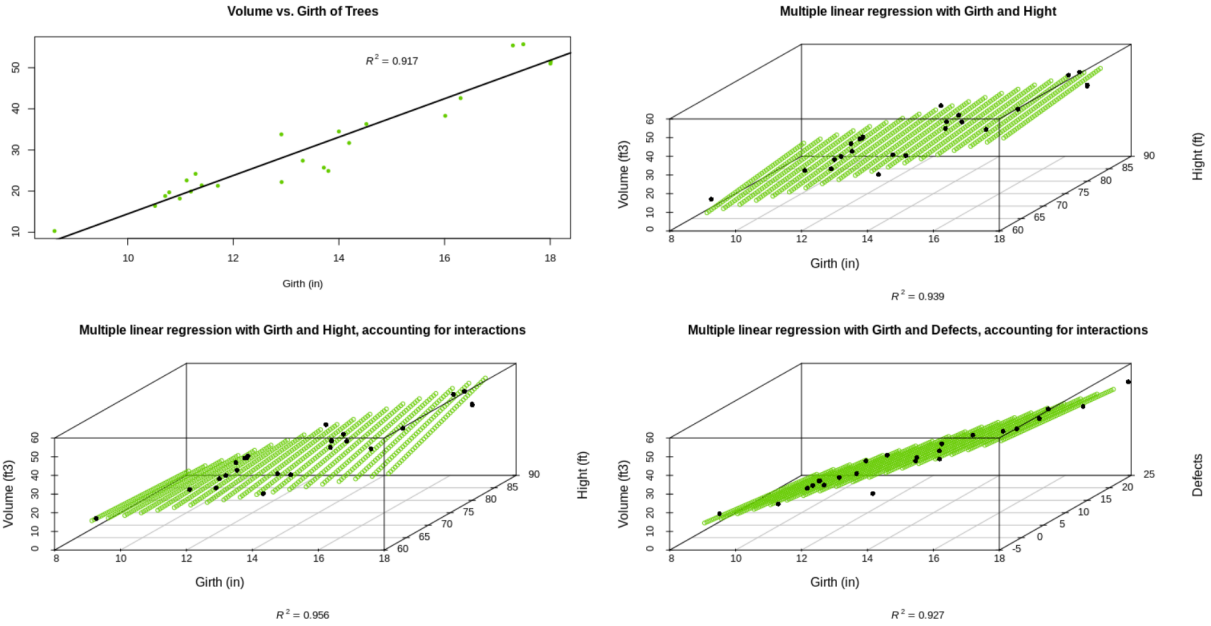

Figure 17: Linear regression summary

9 Laplace Distribution

(a) In probability theory and statistics, the Laplace distribution is a continuous probability distribution named after Pierre-Simon Laplace. It is also sometimes called the double exponential distribution, because it can be thought of as two exponential distributions (with an additional location parameter) spliced together back-to-back.[2]

(b) **Definitions and properties:** [2]

i. A random variable has a Laplace(μ,b) distribution if its probability density function is:

$$f(x \mid \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$$
$$= \frac{1}{2b} \begin{cases} \exp\left(-\frac{\mu - x}{b}\right) & \text{if } x < \mu \\ \exp\left(-\frac{x-\mu}{b}\right) & \text{if } x \geq \mu \end{cases}$$

The probability density function of the Laplace distribution is also reminiscent of the normal distribution; however, whereas the normal distribution is expressed in terms of the squared difference from the mean μ, the Laplace density is expressed in terms of the absolute difference from the mean. Consequently, the Laplace distribution has fatter tails than the normal distribution.

ii. The Laplace distribution is easy to integrate (if one distinguishes two symmetric cases) due to the use of the absolute value function. Its cumulative distribution function is as follows:

$$F(x) = \int_{-\infty}^{x} f(u)\,\mathrm{d}u = \begin{cases} \frac{1}{2}\exp\left(\frac{x-\mu}{b}\right) & \text{if } x < \mu \\ 1 - \frac{1}{2}\exp\left(-\frac{x-\mu}{b}\right) & \text{if } x \geq \mu \end{cases}$$
$$= \tfrac{1}{2} + \tfrac{1}{2}\,\mathrm{sgn}(x-\mu)\left(1 - \exp\left(-\frac{|x-\mu|}{b}\right)\right).$$

iii. Moments:

$$\mu'_r = \left(\frac{1}{2}\right)\sum_{k=0}^{r}\left[\frac{r!}{(r-k)!}b^k\mu^{(r-k)}\{1+(-1)^k\}\right] = \frac{m^{n+1}}{2b}\left(e^{m/b}E_{-n}(m/b) - e^{-m/b}E_{-n}(-m/b)\right)$$

[2] https://en.wikipedia.org/wiki/Laplace_distribution

(c) **Observations:**

i. Laplace distribution is a distribution that is symmetrical

ii. the the dispersion of the data around the mean is higher than that of a normal distribution

iii. a normal distribution has very thin tails, i.e. probability density drops very rapidly as you move further from the middle, like exp(-x^2). The Laplace distribution has moderate tails [3], going to zero like exp(-| x |).[4]

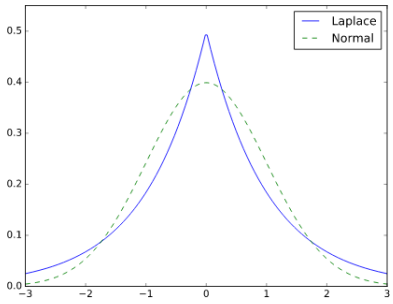

Figure 18: Laplace and Normal Distribution comparison

(d) **Applications:**
The Laplace distribution is used for modeling in signal processing, various biological processes, finance, and economics. Examples of events that may be modeled by Laplace distribution include:

i. the addition of noise drawn from a Laplacian distribution, with scaling parameter appropriate to a function's sensitivity, to the output of a statistical database query is the most common means to provide differential privacy in statistical databases

ii. in regression analysis, the least absolute deviations estimate arises as the maximum likelihood estimate if the errors have a Laplace distribution

iii. in hydrology, the Laplace distribution is applied to extreme events such as annual maximum one-day rainfalls and river discharges [5]

iv. credit risk and exotic options in financial engineering

v. insurance claims

[3] The normal distribution is the canonical example of a thin-tailed distribution, while exponential tails are conventionally the boundary between thick and thin. "Thick tailed" and "thin tailed" are often taken to mean thicker than exponential and thinner that exponential respectively.

[4] `https://www.johndcook.com/blog/2019/02/05/normal-approximation-to-Laplace-distribution`

[5] `https://en.wikipedia.org/wiki/Laplace_distribution`

vi. structural changes in switching-regime model and Kalman filter[6]

(e) **Visual reprezentations:** of the probability density function and the cumulative distribution, using x = [-100, -99, ..., 99, 100] (vector of responses), $\mu = 10$ (location parameter) and b = 7 (diversity, scale parameter)

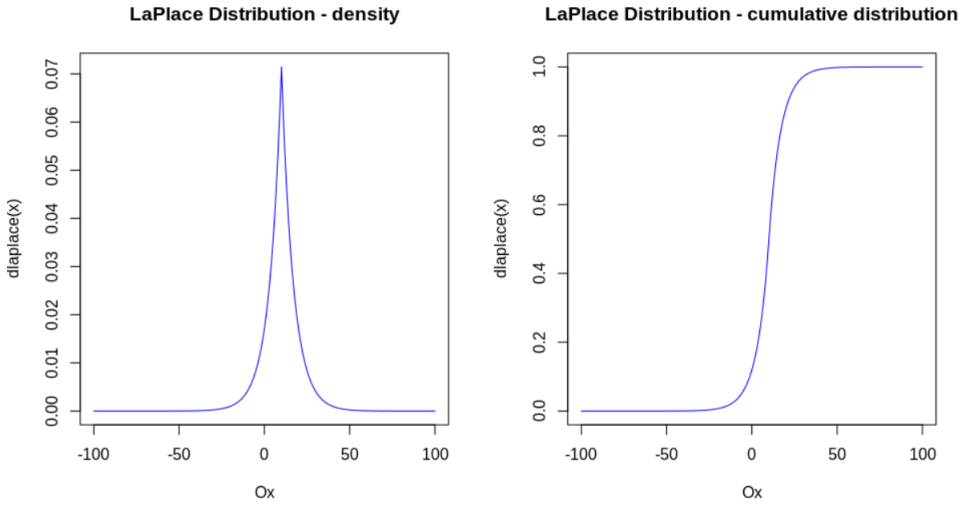

Figure 19: Laplace Distribution

(f) **Example:**
Suppose that the return of a certain stock has a Laplace distribution with $\mu = 5$ and b = 2. Compute the probability that the stock will have a return between 6 and 10.
We can compute this as follows: [6].

$$P(6 \leq X \leq 10) = \sum_{x=6}^{10} \frac{1}{2 \times 2} \exp(-\frac{|x-5|}{2}) = 0.262223$$

[6] http://wiki.stat.ucla.edu/socr/index.php/AP_Statistics_Curriculum_2007_Laplace

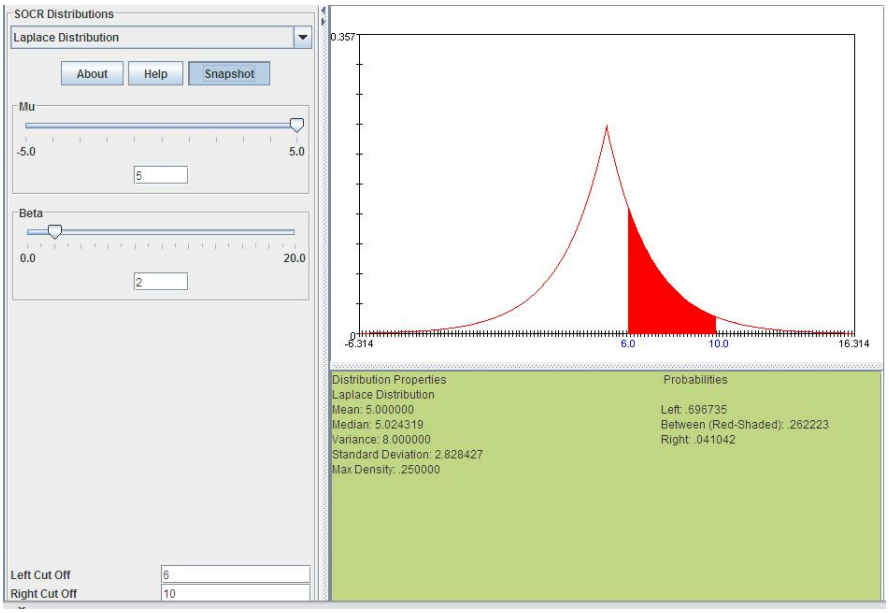

Figure 20: Laplace Distribution

10 References:

1. http://r-statistics.co/Linear-Regression.html
2. https://en.wikipedia.org/wiki/Laplace_distribution
3. http://wiki.stat.ucla.edu/socr/index.php/AP_Statistics_Curriculum_2007_Laplace
4. https://www.johndcook.com/blog/2019/02/05/normal-approximation-to-Laplace-distribution
5. https://stats.stackexchange.com/questions/15011/generate-a-random-variable-with-a-defi
6. http://rstudio-pubs-static.s3.amazonaws.com/138191_9169a18ae3d34e1492d1df67a810e5d5.html?fbclid=IwAR0I6erpHRt3YXESb8LvCExibCMDucibZa1iK-nOe3CjYb9lQCEUcuTyJUA
7. https://www.dataquest.io/blog/statistical-learning-for-predictive-modeling-r/?fbclid=IwAR0ktY7Dcnu_8HLEildpw0U--9-oh1RweyqHedrdOQKP4XkSOQbCwYpA0Aw
8. https://rpubs.com/Pun_/PredictiveModellingofVolumeofCheeryTrees?fbclid=IwAR1CWMFLwgWvafASA9S4KWcZSj1F_VI19E08t7NjSQ1XM2MYZs7j7w1B78I
9. https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/treering.html?fbclid=IwAR1YiUoAzhIg8wcHOgRdLDI-ZcVMgGtRDaz54-R-Xbf6TRmY5Wlt54gCPcQ

10. https://www.dataquest.io/blog/statistical-learning-for-predictive-modeling-r/?fbclid=IwAR0BSEgvcWC5yb8rViOOK6LhpMIMBaWGlULIpm3KflfVgrF239KiIAxo9_w

11. https://www.statmethods.net/stats/regression.html