Hochschule
für nachhaltige Entwicklung
Eberswalde

Eberswalde University for Sustainable Development

Faculty of Forest and Environment

Written Exam Report

For

Course: Environmental spatial data analysis

Group: A

Submitted to:

Dr. Evelyn Wallor

Submitted by:

Kazi Jahidur Rahaman

Date of Submission:

February 09, 2022

Winter 2022

1. **From systematic observations in a tree nursery it is known that the germination rate of a certain apple tree species is 62 %.**

a) Which probability function / frequency distribution is appropriate to describe the above statement? Name the properties of the respective frequency distribution!

Answer: The description mentions 2 discrete probabilities (either germinate or not germinate). Therefore the probability function is discrete random probability function. It has the following has the following properties.

i. $P(X = x_i) \geq 0 \; for \; all \; x_i$

ii. $\sum P(X = x_i) = 1$

b) What is the probability that from 10 apple tree seeds at least 6 will germinate? Provide the respective parameters and parameter values and write the result in an answer sentence!

Answer: Given that,

Number of trials, n = 10

Expected number of success, k= 6:10

Probability of success, p = 62% = 0.62

The probability that from 10 apple tree seeds at least 6 will germinate would be the sum of probabilities of germinating 6, 7, 8, 9 and 10 seeds.

Using the R function sum (dbinom (k, n, p)) we get, P ($6 \leq X \leq 10$) = 0.682313

```
> n<-10
> p<-0.62
> k<-6:10
> dbinom(k,n,p)
[1] 0.248716054 0.231885644 0.141877401 0.051440929 0.008392994
> sum(dbinom(k,n,p))
[1] 0.682313
```

2. **In a certain case study area the variable soil organic carbon content is normally distributed with the following parameters: μ = 1520 mg kg^{-1} soil and σ = 830 mg kg^{-1} soil.**

   a) **What is the corresponding z-value of the z-standard normal distribution for an observed soil organic carbon content of 1200 mg kg^{-1} soil?**

   Answer:

   Here,

   μ = 1520 mg kg^{-1},

   σ = 830 mg kg^{-1} soil

   x = 1200 mg kg^{-1} soil

   According to the equation of z-conversion, $z = \frac{x-\mu}{\sigma} = \frac{1200\ mg\ kg^{-1} - 1520\ mg\ kg^{-1}}{830\ \text{mg kg}-1} = -0.39 \approx 0.39$

   Therefore, corresponding z-value is 0.39.

   b) **What is the probability that soil organic carbon content is between 1200 and 1800 mg kg^{-1} soil?**

   Answer:

   Let us assume, x_1 = 1200 mg kg^{-1} soil

   X_2 = 1800 mg kg^{-1} soil

   The probability of soil carbon content being between 1200 and 1800 mg kg^{-1} soil is, The cumulative probability of soil carbon content between (μ, 1800) – cumulative probability between (μ, 1200). We can find the cumulative probability with the R function *pnorm()*. Calculating the values from R we get,

   Probability of soil carbon content between 1200 and 1800 mg kg^{-1} soil =

   $P(z \leq 0.34) - P(z \leq 0.39) = 0.6320732 - 0.3499179 = 0.2821553 = 28\%$

```
> x1 <- 1200
> x2 <- 1800
> mean <- 1520
> sd <- 830
> pnorm(x2, mean, sd)-pnorm(x1, mean, sd)
[1] 0.2821554
```

   c) **Which soil organic nitrogen content value defines the threshold of the upper quantile of 75 %?**

Answer: The soil organic content value of 2079.826 mg ka^{-1} soil defines the upper quantile of the 75%. It can be found using the R function *qnorm(p=0.75,mean,sd)*.

3. This task is related to the data provided in the text file *DataA1.txt*. It contains measurement values of the diameter at breast height (*BHD*) [cm] for the tree species *Common douglas fir* measured at two different forest districts (BAR = Barnim, UM = Uckermark).

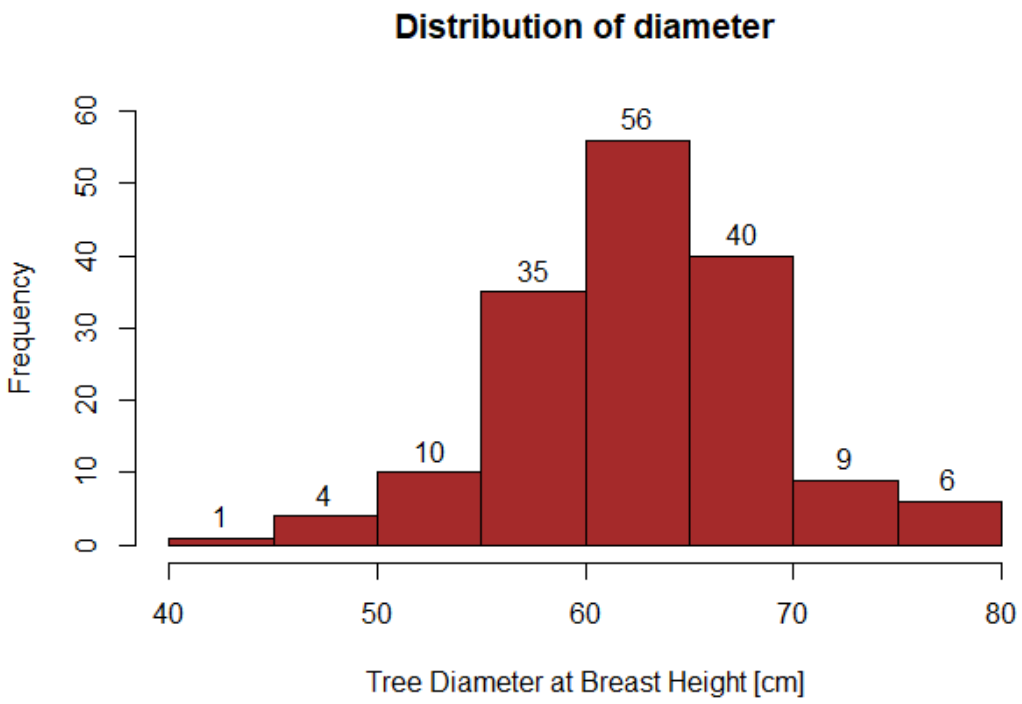
a) Name the data scales of the features *BHD, Species,* and *ForestDistrict*!

Answer: The data scales of the mentioned features are as follows-

a. BHD : Quantitative, Metrical, Continuous
b. Species: Qualitative, Nominal, Discrete
c. ForestDistrict: Qualitative, Nominal, Discrete

b) Provide a graph that illustrates the distribution of the measured feature *BHD*. Which graph is used and what are the class ranges of the lowest and uppermost *BHD* class?

Answer:



The above histogram graph illustrates the diameter distribution for the given dataset. It is seen that the dataset contains highest number (56) of trees of diameter class 60-65 cm range on the other hand, there is lowest number of trees with DBH of 40-45 (1) cm.

c) Apply an appropriate statistical test to proof if *BHD* values are normally distributed. Name the selected test, define the hypothesis, interpret the result, and conclude!

Answer: There are several tests to test the normality of a given data set. For this task, we consider Shapiro-Wilk's test to check the normality of the *BHD* values of the given dataset.

We assume,

The Null hypothesis, H_0 = The *BHD* distribution is normal and

H_A = The *BHD* distribution is not normal, and

Significance level, $\alpha = 0.05$

The output from the statistical testing software (RStudio) is:
W = 0.99106, p-value = 0.4102

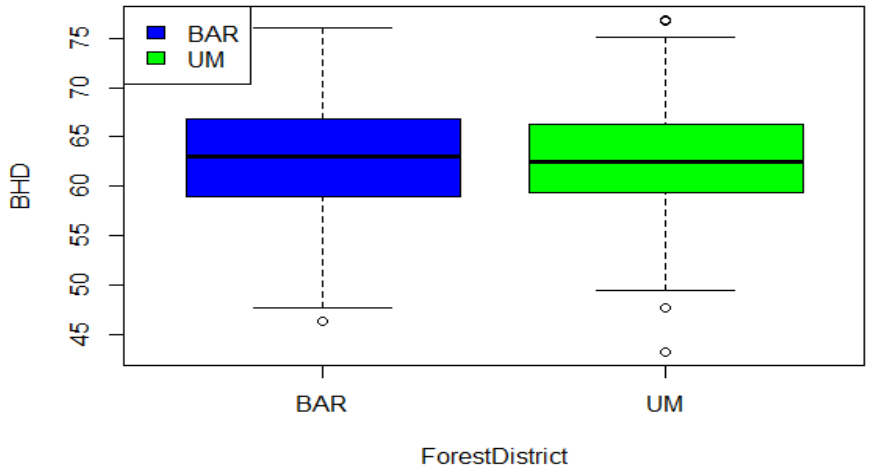
```
        Shapiro-Wilk normality test

data:  ed1$BHD
W = 0.99106, p-value = 0.4102
```

Here, the p-value = 0.0412 is greater that significance value 0.05. p-value > 0.05 is significant.
Therefore, we accept the null hypothesis H_0. That means, the distribution of *BHD* is normal.

However, since the Shapiro-Wilk test is sensitive to small sample size, it is recommended to visual confirmation too. The histogram of the distribution presented in the answer to question no (c), visually confirms our null hypotheses to be correct.

d) Provide a box-whisker plot of measured *BHD* values for each forest district. Which statistical parameters of location are illustrated in a box-whisker plot? What are the respective *BHD* values for forest district Uckermark?

Answer:



The box-whisker plot of the BHD values for each of the forest districts is presented in the above graph where, BAR = Forest district Barnim and UM = Forest district Uckermark.
The plot gives us the information of the Minimum, Maximum, and Median; as well as how diverse the BHD range is for each of the districts.
For instance, the box-whisker plot of the tree BHD's for the forest district Uckermark shows, the

- Minimum *BHD* of *Common douglas fir* trees in the area is around 49.4cm
- Maximum is around 75 cm. There are few outliers which says, few trees has heights less than 49.4 cm and more than 75 cm, but these values are unusual so considered as outliers.
- The lower quartile, median, and the upper quartile value of the heights are around 59, 62 and 66 cm which mean the data has less diversity and more consistency among the heights than that of forest district Barnim.

4. **This task is again related to the data provided in the text file *DataA1.txt*. It is known, that On average, *Common douglas fir* stands in Northern Europe, and of similar age as the trees observed in Barnim and Uckermark, show a *BHD* in spring of 72 cm.**

   a) **Which statistical test is appropriate to estimate if the trees in the provided data match with the above mentioned average value? Name it precisely and explain the purpose of the test!**

   Answer: The given case mentions a reference value to match with the provided dataset, in this case One-sample t-test is an appropriate option for evaluation. The purpose of this test is to compare the difference between the reference value and the sample values (mean of sample value).

   b) **Select one sample of *BHD* values (Barnim or Uckermark) and apply the test mentioned under a).**
   Formulate the hypothesis and conclude from the test result!

   Answer: We choose the BHD values for the district Barnim to apply the one-sample t-test mentioned in a).

   Variable = Tree *BHD* [cm] of the *Common douglas fir* trees measured at forest district Barnim.
   Reference value = 72 cm
   Significance level, $\alpha = 0.05$
   H_0 = the mean of *BHD* of sample is not statistically different from the reference value. ($\mu = 72$)
   H_A = the mean of *BHD* of sample is statistically different from the reference value. ($\mu \neq 72$)
   The output from the statistical testing software (RStudio) is:

```
        One Sample t-test

data:  ed1barnim$BHD
t = -13.856, df = 82, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 72
95 percent confidence interval:
 61.40226 64.06328
sample estimates:
mean of x
 62.73277
```

   Here, p-value 0.00000000000000022 < α.
   Therefore, we reject the H_0. Hence, accept the alternative hypothesis H_A, that is, the mean of BHD sample is statistically different from the reference value.

c) Which statistical test is appropriate to estimate if the trees' *BHD* values differ significantly between the two forest districts Barnim and Uckermark? Name it precisely and explain the purpose of the test!

Answer: Here, if we want to compare the BHD values for 2 different forest districts, we get 2 unpaired different samples of BHD values; one for the district Barnim, another fo Uckermark.

The difference between 2 unpaired samples can be tested by two-sample t-test.

The purpose of the two-sample t-test is to find the difference among the different samples by comparing their corresponding mean values. We find the μ_1 and μ_2 for the both samples and compare whether $\mu_1 - \mu_2 = 0$.

d) Proceed the test under c). Formulate the hypothesis and conclude from the test result!

Answer: Let us assume,

H_0 = *BHD* values do not differ significantly between districts Barnim and Uckermark ($\mu_1 = \mu_2$)

H_A = *BHD* values differ significantly between districts Barnim and Uckermark. ($\mu_1 \neq \mu_2$)

Significance level, $\alpha = 0.05$

By applying two-sample t-test in R, (the statistical program R chooses the test method automatically) we get,

```
        Welch Two Sample t-test

data:  ed1barnim$BHD and ed1uckermark$BHD
t = 0.39245, df = 157.45, p-value = 0.6953
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.550362  2.319238
sample estimates:
mean of x mean of y
 62.73277  62.34833
```

Here, the p-value 0.6953 > 0.05 and also shows that, means are not equal ($\mu_1 \neq \mu_2$).

Therefore, we reject the null hypothesis and accept the alternative hypothesis that is BHD values differ significantly between districts Barnim and Uckermark.

5. Explain the differences and similarities between the paired sample t-test and the Wilcoxon signed-rank test with respect to purpose, preconditions, and data properties!

Answer:

Purpose: The Paired-sample t-test and Wilcoxon signed-rank test both use paired samples. However, the paired-sample t-test looks for differences between the sample means, but Wilcoxon signed-rank test looks for the equality of the central tendencies (tendency to the median value) of the underlying connected population.

For applying Paired-sample t-test, the sample data is needed to be normally distributed, but for Wilcoxon signed-rank test data are not required to be normal.

Data properties: Although both uses unrelated paired samples, the t-test is applicable for the parametric values, while the Wilcoxon test is applicable for the categorical variables only.

6. This task is related to the data provided in the text file *DataA2.txt*. It contains *volume* measurements of a certain tree species after harvest in [m^3]. Before harvest, the *health state* of each tree was noted (scale from 5 = healthy to 10 = strongly affected).

a) Name the data scales of the features *Volume* and *health state*!

Answer: The data scales of the features are-

Volume: Quantitative, Metrical, and Continuous

health state: Qualitative, Ordinal, Discrete

b) Which statistical approach is appropriate to look for differences in the *volume* values between the recorded *health states* of harvested trees?
Name the respective approach precisely, explain the purpose, and the required assumptions!

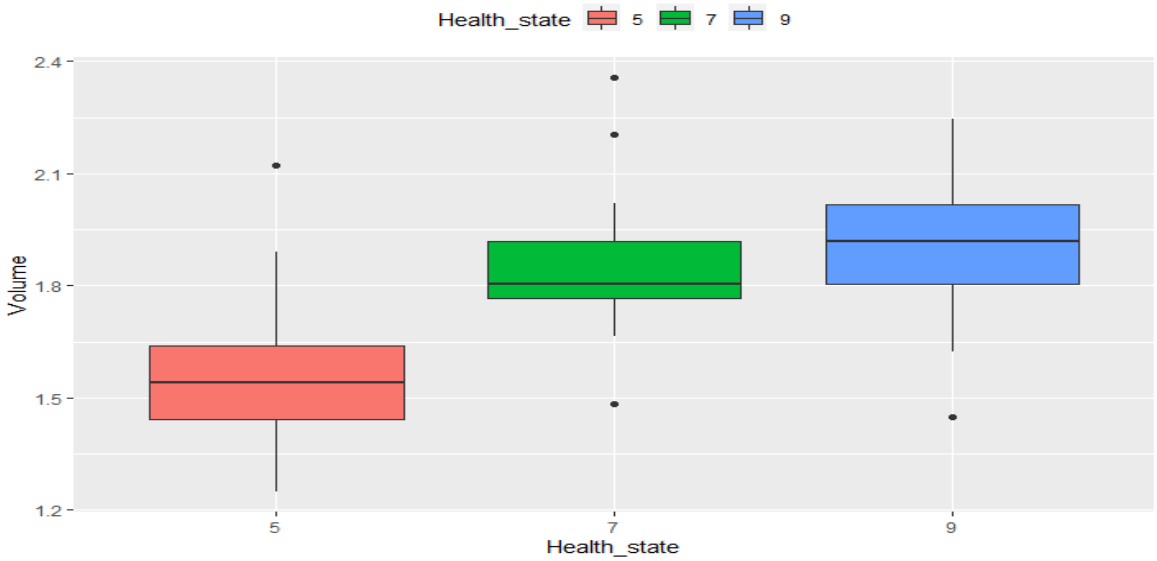
Answer: In this scenario, our aim is looking for the differences in a continuous attribute for 3 (more than 2) categorical sample groups. Analysis of Variance (ANOVA) is an appropriate test method for this case. Since we have 1 independent variable (health status) (1 independent variable with 3 small groups) we should follow the One-way ANOVA test.
The purpose of the ANOVA test is to find weather there is significant differences among the means of the dependent variable *volume* among the individual *heath sate* groups (5, 7, and 9)
For applying the ANOVA test we should consider the following assumptions-

i. Volumes are normally distributed within the health state groups 5, 7, and 9
ii. The variances within each of the groups are similar, or the health state groups are homoscedastic.
iii. The samples are independent.
iv. Group 5, 7, and 9 must have sufficient and equal sample sizes.

c) Proceed with the approach defined under b) including necessary statistical tests to proof the assumption. Formulate the hypothesis of the statistical approach mentioned under b).
With respect to the result, formulate the conclusion!

Answer: It is mentioned in answer to question b) that, the One-way ANOVA needs to consider various assumptions. We first check whether the given dataset meets the assumptions.

i. Distributions are normal.



The groups are almost normal except the group 7 shows a bit of left skewness but that could be ignored as minor but there are variances among the distributions.

ii. The variances are homogeneous and can be proved by Levene's test. Let us assume,
The null hypothesis H_0 = Variances are homogeneous.
H_A = Variances are not homogeneous.
F critical value = 0.05

From applying Levene's test, we get

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  2  0.2644 0.7685
      60
```

The F value 0.26 > F critical value 0.05. Therefore, we accept the H_0, that is the variances are homogeneous.

iii. Finally, since our sample data meets the assumptions, we can proceed to conduct the ANOVA test of one-way variant.
We assume,
H_0: The means of all groups are equal.
H_A: The means of all groups are not equal.

From applying the one-way ANOVA with R, we get-

```
             Df Sum Sq Mean Sq F value   Pr(>F)
Health_state  2  1.404  0.7020   18.26 6.39e-07 ***
Residuals    60  2.307  0.0384
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The P-value for the corresponding F value is 0.000000639 which is less than the corresponding significance value 0.

Therefore, we reject the H_0 and accept the H_A, that is the mean values are not equal.

d) Based on the result under c), which additional statistical procedure would you recommend?

Answer: It is found from the one-way ANOVA test conducted in answer to the question no c) that the means of the variances among the health state groups 5, 7 and 9 are not uniform. That means there is at least one pair of groups with differing means.

We might try to find out which pair has differing means. We can find that out by conducting the post-hoc test. There are different types of post-hoc analysis, i.e.: Tukey's HSD post-hoc, Bonferroni's post-hoc analysis, etc. Since we are dealing with the parametric data types, I would recommend Tukey's HSD post-hoc test.

```
> TukeyHSD(res.aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Volume ~ Health_state, data = ed2)

$Health_state
          diff         lwr       upr     p adj
7-5 0.29038095  0.14496637 0.4357955 0.0000324
9-5 0.33766667  0.19225208 0.4830813 0.0000018
9-7 0.04728571 -0.09812887 0.1927003 0.7157768
```

Applying the Tukey's post-hoc HSD test, we find the Health state groups 7-5 and 9-5 are significantly different in 95% confidence level.

7. This task is related to the text file *projectdata.txt* provided to students during the group work (cf. moodle learning room, zipped folder on main page). Among others, the data consist of diameter (*DBH* [cm]) and height (*Height* [m]) measurements for several tree species growing on a marteloscope site close to the HNEE forest campus.

a) Develop a simple regression model that describes the relationship between *Height* ~ *DBH* for the tree species *Copper beech*.
Provide the full equation of the resulting regression model including the estimated coefficients!

Answer: The equation for our simple linear model is

$$height, y_i = b_0 + b_1 x_i + \varepsilon_i$$

After fitting the values of DBH and Height from the dataset with R, we get,

```
Call:
lm(formula = Height ~ DBH, data = epdata_cb)

Residuals:
    Min      1Q  Median      3Q     Max
-5.3712 -0.3283  0.2577  0.5179  3.0608

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.064272   0.124546   64.75   <2e-16 ***
DBH         0.624447   0.008529   73.22   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.831 on 523 degrees of freedom
Multiple R-squared:  0.9111,    Adjusted R-squared:  0.9109
F-statistic:  5360 on 1 and 523 DF,  p-value: < 2.2e-16
```

From the output we get the value of our coefficients b_0 and b_1 which are 2.52975 and 0.84843 respectively. That means, the height of a tree will be the sum of 2.52975 and 0.84843 times of the DBH value adjusted by the variation of 0.008. So, the final equation would be for our linar model is,

$$height, y_i = 8.064272 + 0.624447 * DBH \pm 0.008$$

b) Explain what the coefficient of determination (R^2) and the residuals tell us about the quality of the developed regression model. How would you assess the quality of the regression model developed under a)?

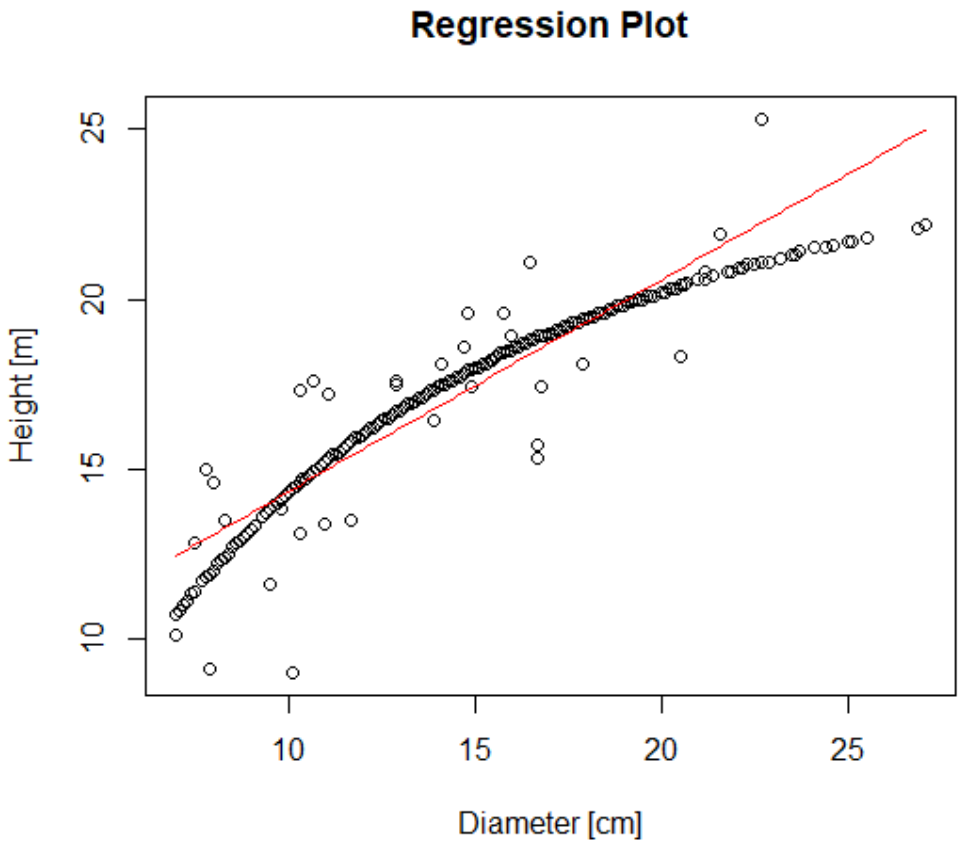
Answer: The coefficient of determination R^2 shows how well adapted the linear model is to the fitted dataset by showing the proportion of the explained variation by the model and the total variation of the model. It ranges from 0-1 and -1-0 for positive and negative values respectively. Higher the distance from 0, better the model is.

From the summary provided in answer to question a, we get the adjusted R^2 for our model is 0.9101. Therefore it could be said that, the model explains most of the variation in the data.

Sometimes, the R^2 value might not express the actual goodness of the model, in that case the Residuals of the model can also express the goodness of fitting of the model. Residuals are the difference between the actual values and the predicted value. So, the closer residuals are to 0, the good fit the model is. The residual values for the developed model says the difference of the predicted and actual values in the 1st and the last quartile is higher. That means, the model performs poorly in predicting values of those areas.

c) Visualize the estimated regression model together with the observations in a so-called regression plot!

Answer:



The above Regression plot show the regression line for the predicted values in the red. And the point values represent the actual values. It is seen that the model shows larger residual in predicting heights for the trees which have diameter <10 cm or diameter > 20 cm. More precisely saying, the data does not fit well to a linear model.

-END-