Department of Mechanical Engineering - University of Coimbra

Degree in Industrial Engineering and Management

Academic year 2021/2022

STATISTICS APPLIED TO ENGINEERING

Group 12

Lucas Jesus da Rocha - 2020112852

Introduction

As part of the Applied Statistics for Engineering course, we were asked to carry out a project using R programming. The data used to solve the proposed exercises was imported from Excel for group 12.

To analyze the behavior of each variable we used various tests, defining 5% as the significance level and 95% for the confidence intervals. As these variables are real data, they cannot have negative values. It is therefore necessary to observe the variables in order to eliminate these values so that their analysis is the most correct.

The process will be described in 3 different phases: filling, labeling and packaging. Our report is organized according to the statement provided by the teacher of the subject, so we will proceed to solve it without transposing the statement.

Questions

Question 1: How can Velo's behavior be characterized probabilistically?

In this question, we are asked for the probabilistic behavior of the variable *Velo,* its value being described by one of the following distributions: Normal, Chi-square or Uniform. It is expected that this variable, which represents speed, will be as constant as possible and that it will not have any extreme values.

In a first approach, before carrying out the adjustment tests for the laws requested, in order to identify the nature that this variable follows, we began by constructing and analyzing the histogram that corresponds to the given sample of the *Velo* variable.

As we can see in figure 1, the histogram coincides with a normal distribution. However, we can't say for sure that the *Velo* variable is described by this distribution without carrying out the other goodness-of-fit tests, since it is subjective. Therefore, in addition to carrying out the Anderson-Darling e Kolmogorov-Smirnoff, we will also perform the Shapiro-Wilk test

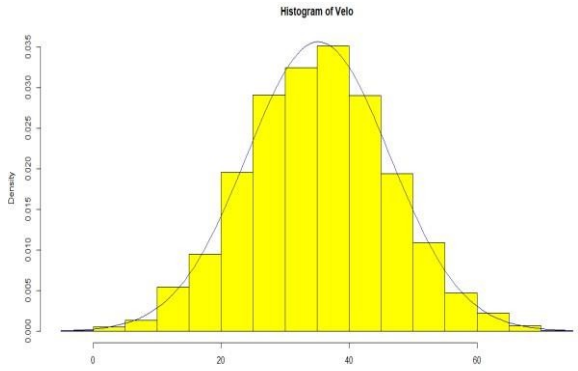

Figure 1- Histogram of the Velo variable

for the normal distribution, since this is the most powerful. Since the Chi-Square has low power, we decided that it was not necessary to perform it for the Uniform law and the Chi-Square law and only performed it for the Normal law.

We will decide whether or not to accept the law in each test, according to the comparison between the significance level mentioned above (5%) and the p-value presented in the output.

In order to check that the variable in question does in fact follow a normal distribution, we proceeded to test it. First, the chi-square test, where we began by defining a partition with 4 intervals: [0,20],]20,40],]40,60],]60, +∞ [, suggested by the previous histogram. Thus, we have come to the conclusion that as the p-value is 0.8898 and is above the significance level, its test statistic does not belong to the critical region, so we will accept that the *Velo* variable follows a normal distribution according to the mean and standard deviation of the sample. In the case of the Kolmogorov-Smirnoff test and the Anderson-Darling test, based on their p-values: 0.9677 and 0.9958, respectively, we conclude that in this case the variable also follows a normal law.

We then carried out the Kolmogorov-Smirnoff and Anderson-Darling tests for the other laws, Uniform and Chi-Square, and we saw from their p-values that they were lower than the significance level, so we concluded that *Velo does* indeed follow a normal law.

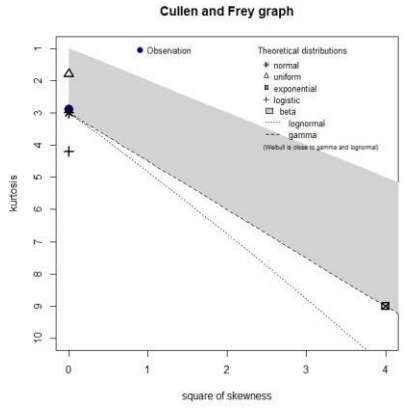

Figure 2- Analytical chart of the different laws

Below is a graph where there is only one possible value for the asymmetry and kurtosis of the distributions (normal, uniform, logistic, exponential). So, when we compare these parameters with our observation, we see that it is clear that the distribution is very close to normal.

Question 2: The ideal behavior for Press should be normal with a mean of 350 and a standard deviation of 121. Are the operating conditions suitable?

To find out whether the behavior of the *Press* variable, with a mean of 350 and a standard deviation of 121, is the most suitable for normal distribution, we will perform the Chi-Square, Kolmogorov-Smirnoff, Anderson-Darling and Shapiro-Wilk tests. The *Press* variable is the pressure with which the machine fills, which is very important for determining the final characteristics of the packaged product.

For the analytical resolution, the following hypotheses were tested:

$$\begin{cases} H_0: Press \sim \mathcal{N}(350, 121^2) \\ H_1: Press \nsim \mathcal{N}(350, 121^2) \end{cases}$$

In this way, a histogram was drawn up for the variable in question, to see if its behavior is close to a normal distribution. From the histogram we obtained the parameters needed to carry out the chi-squared test: nk and pk.

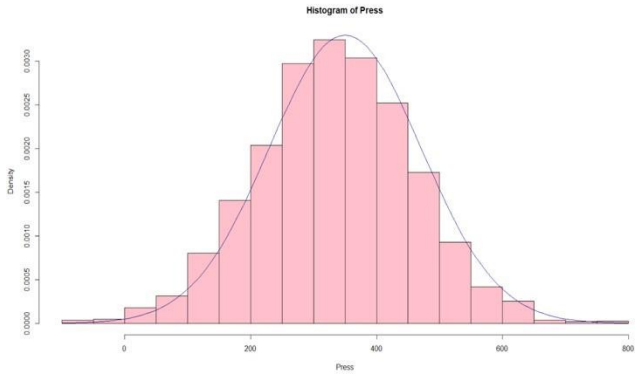

Figure 3-Histogram of the Press variable

The Shapiro-Wilk test shows a p-value of >5%, but it does not evaluate the parameters to be evaluated, it only indicates whether the variable describes a normal law or not. After carrying out the Chi-Square, Kolmogorov-Smirnoff and Anderson-Darling tests, we obtained the following p-values: 9.616×10^{-5} , 1.36×10^{-6} and 2.715×10^{-7} , respectively. We found that these were lower than 5%, and so we did not accept the H0 hypothesis at the 5% significance level, concluding that *Press* does not follow the ideal behavior described.

Question 3: Some retailers complain that they receive consignments of lubricant packaging in which the ratio between types of packaging doesn't seem to correspond to what was requested. Are these complaints justified?

As mentioned in the statement, there are three types of packaging for this lubricant: A, B and M. They are used in a ratio of 30:60:10, depending on their capacity.

This question asked us to verify the basis of the resellers' complaints, since they complained that the proportion delivered did not correspond to that requested. In order to understand the veracity of the salespeople's claim, we first checked the rules of hypothesis testing and then calculated the direct chi-squared test, where we obtained a p-value of

$=2.2x10^{-16}$. As the p-value is less than 5%, we reject H0, i.e. the retailers are right to complain about the poorly distributed proportions.

Question 4: Is there evidence of a linear explanation of the Time variable at the expense of Velo, Espe and Resis? Are these three explanatory variables really necessary?

To understand the influence of the *Velo*, *Espe* and *Resis* variables on the time variable, we began by calculating the value of the coefficient of determination (R^2) in the complete multilinear regression. The R^2 is a measure of the quality of the fit of the data, and the closer it is to 1, the better the fit. In this case, as shown in the R output, the value of R^2 is 0.6366, which is considered a reasonable fit, and the model presented is therefore accepted.

Next, we constructed a graph representing linear dependence. By analyzing it, we concluded that only *Time* and *Velo* have simultaneous linearity, which was not identified between *Time* and the other variables.

In order to see if these 3 explanatory variables are necessary, we carried out individual nullity tests. By removing the *Espe* and *Resis* variables, we obtained the following results R^2 , 0.6365 and 0.6359, respectively, and by removing *Velo* the value of R^2 was 0.0008436. Looking at these results, we see that the values of *Espe* and *Resis* are close to the value of R^2 obtained for the complete model, unlike *Velo, which* cannot be removed.

Finally, we carried out the simultaneous nullity test for *Espe* and *Resis*. By calculating the value of the test statistic and the critical region, we decided to accept the H0 hypothesis ($ß2=ß3=0$) at the 5% significance level and thus concluded that the *Time* variable can only be defined at the expense of *Velo*.

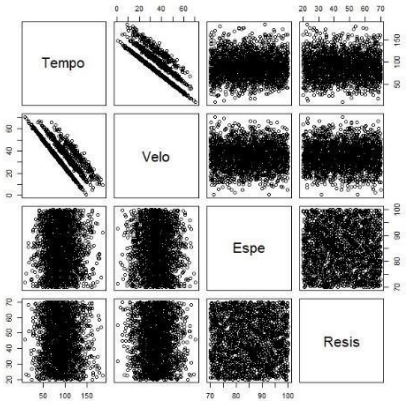

Figure 4- Scatter

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 160.22392    3.98567  40.200   <2e-16 ***
Velo         -1.99130    0.03205 -62.124   <2e-16 ***
Espe          0.03079    0.04239   0.726   0.4676
Resis        -0.05046    0.02493  -2.024   0.0431 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.82 on 2206 degrees of freedom
Multiple R-squared:  0.6366,    Adjusted R-squared:  0.6361
F-statistic:  1288 on 3 and 2206 DF,  p-value: < 2.2e-16
```

plotFigure 5-Output R

Question 5: Does it make sense to believe in an approximately linear relationship between the Cost variable and six other variables? Are these six variables really necessary?

The *Cost* variable is related to the final production cost of a certain type of car lubricant, so let's analyze the linear approximation of the variables

which can affect production i n any way, are *Velo, Press*, *Volu*, *Espe*, *Resis* and *Time*. To do this, we began with a graphical analysis, using a scatter plot, to see if the variables show any dependence on each other.

Looking at the graph in figure 6, we can easily see that not all the variables show a linear relationship with the *Cost* variable. The only ones that appear to have a linear relationship are *Velo*, *Press* and *Tempo,* while the others have no significant influence.

However, in order to obtain more accurate results, since graphical analysis is subjective, we used the coefficient of determination through the multilinear regression model for the complete model. We obtained $R^2 = 0.9466$, which is very close to 1, indicating that the model has a good fit, so there may be a close linear relationship between the *Cost* variable and the others.

To answer the second part of the question, we looked at whether all six variables were needed to describe *cost*. In order to obtain the precise values, we carried out individual nullity tests for each of the variables, thus obtaining the coefficients of determination and, by comparing them with the R^2 of the complete model, we could see which variables could possibly be removed.

Below are the coefficient of determination values for the linear regression model: without *Resis* $R^2 = 0.9466$, without *Espe* $R^2 = 0.9466$, without *Volu* $R^2 = 0.9465$, without *Velo* $R^2 = 0.9376$, without *Press* $R^2 = 0.2575$ and, finally, without *Temp* $R^2 = 0.8958$. Looking at these values, we quickly see that *Resis*, *Espe* and *Volu* have coefficients of determination that are very close to or equal to the full model. We therefore consider that we can remove these three variables individually, however, we will still test their simultaneous nullity and see if they can be removed from the model at the same time.

$$\begin{cases} H0: \beta_{Resis} = \beta_{Volu} = \beta_{Espe} = 0 \\ H1: \text{caso contrário} \end{cases}$$

When we carried out the simultaneous null test, we obtained a p-value = 0.342, which is greater than 5%, i.e. we accepted H0. In addition to the p-value, we calculated the test statistic, f0=1.113, and the critical region = [2.609, $+\infty$ [, so, in agreement with the previous statement, we conclude that f0 does not belong to the critical region and therefore we also accept H0 at the 5% significance level, so *Resis*, *Volu* and *Espe* are not necessary to define *Cost, which is* defined only by *Time* and *Velo and Press*.

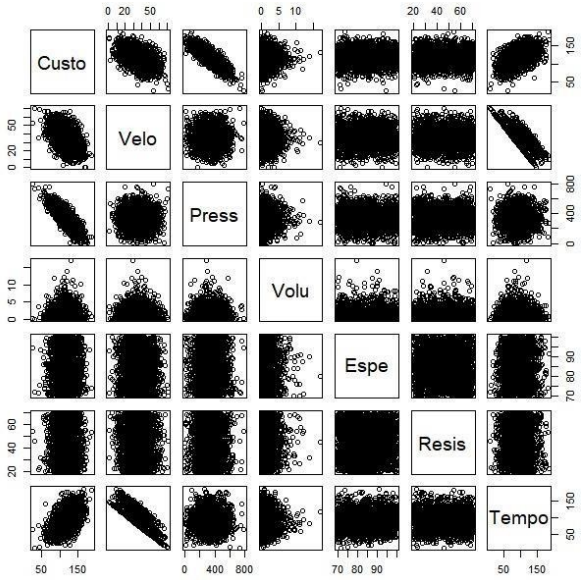

Figure 6- Scatter plot of the complete model

Question 6: Let's now try to describe the cost of production using type M packaging, the one with the smallest capacity.

a) The use of a linear approximation as a function of *Velo is* justified, *Press, Volu* and *Tempo*?

In order to check whether the use of a linear approximation is justified for the *VeloM*, *PressM*, *VoluM* and *TempoM* variables, we relied essentially on two parameters: the coefficient of determination and the scatter plot, where we can see the relationships between the variables in question in relation to the *Cost* variable.

Looking at the graph, we can see that only *VeloM* and *TempoM* show simultaneous linearity.

In order to reinforce the above idea, we ran the multilinear regression model. Looking at the p-values shown in figure 8, we can easily conclude that the *VeloM* and *TempoM* variables have values greater than 5%, showing dependence on the *Cost* variable.

As we can see in the R output, the value of the coefficient of determination (R^2) is 0.9454, very close to 1. That said, we can conclude that the multilinear regression model is a good fit.

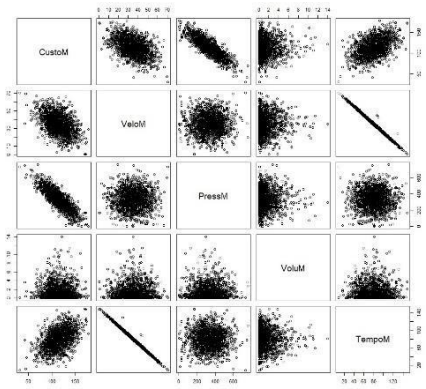

Figure 7-Scatter

```
> pairs(~CustoM+VeloM+PressM+VoluM+TempoM)
> #modelo completo
> modelo.completo<- lm(CustoM~VeloM+PressM+VoluM+TempoM)
> summary(modelo.completo)

Call:
lm(formula = CustoM ~ VeloM + PressM + VoluM + TempoM)

Residuals:
     Min       1Q   Median       3Q      Max
-19.1587  -3.3547  -0.1099   3.2653  14.5858

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept) 142.874906  12.635832   11.307  < 2e-16 ***
VeloM        -0.197132   0.168671   -1.169    0.243
PressM       -0.149943   0.001079 -138.965  < 2e-16 ***
VoluM        -0.061255   0.074519   -0.822    0.411
TempoM        0.348687   0.084076    4.147 3.56e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.989 on 1458 degrees of freedom
Multiple R-squared:  0.9454,    Adjusted R-squared:  0.9452
F-statistic:  6306 on 4 and 1458 DF,  p-value: < 2.2e-16
```

plotFigure 8-Output R

b) The Velo variable seems to be significant in explaining the final value of the Cost?

To test the dependence of cost on *VeloM*, we will analyze the behavior of the model when this variable is removed. To do this, we carried out a null test with the following hypotheses:

$$\begin{cases} \beta_1 = 0 \\ \beta_1 \neq 0 \end{cases}$$

The p-value of the *VeloM* variable, already calculated in a), is greater than 0.05, so we accept hypothesis H0 at the 5% significance level, i.e. we can remove VeloM without the model changing.

In order to confirm this, we ran a new multilinear regression test without the *VeloM* variable, and obtained an R value^2 of 0.9453, again very close to 1, so there was no maximization of the error made. Thus, VeloM is not significant in explaining the final value of *Cost and* can be removed.

c) Indicate a 95% confidence interval for Cost, depending on the explanatory variables you include in the model. How do you justify your expression?

The first thing we had to do in order to answer the question was to define the variables we were going to use in the model, out of the 4 we have available (*VeloM*, *PressM, VoluM* and *TempoM*). As we mentioned in the previous questions, the variables *VeloM* and *TempoM* have p-values greater than 0.05 and so we carried out the simultaneous null test to see if the model can only be dependent on *VoluM* and *PressM*. For this test, we found that the test statistic is equal to 2945.596 and the critical region is [3.0019, +∞ [. We can see that f0 belongs to the critical region and therefore reject H0, i.e. *VeloM* and *TempoM* cannot be excluded from the model at the same time.

So let's define the variable Cost as a function of PressM, VoluM and TimeM, and the 95% confidence interval for this variable. To define this interval we used the following formula:

$$\frac{\mathrm{x}^t\widehat{\beta} - Y(\mathrm{x})}{\sqrt{\frac{S_R}{n-k-1}}\sqrt{1+\mathrm{x}^t(\mathrm{X}^t\mathrm{X})^{-1}\mathrm{x}}} \sim t_{n-k-1}$$

Figure 7-Formula for IC

However, after carrying out the necessary calculations to obtain the requested CI, we realized that we didn't have enough data to arrive at a concrete value. So we just wrote down the expression as a function of the *CostM* variable.

CostM = 128.124 - 0.1499 *PressM* - 0.0613 *VoluM* + 0.4467 *TimeM*

d) Does the existence of a defect in the labeling seem to be indicative of a different behavior in the Cost?

In order to find out whether the existence of defects in labeling can imply changes in cost behavior, we ran a new model in which this new variable was included.

In this model we obtained a coefficient of determination of 0.9454, which is the same as the R^2 that we obtained in point a), in the model that included *VeloM*, *PressM*, *VoluM* and *TempoM*, both being close to 1. In view of this, we consider that we continue to have a good fit, so the behavior of the *Cost is* not influenced by the existence of defects.

Question 7: A key variable in this process is undoubtedly cost. Do you think the average cost is similar for the three types of packaging?

In order to understand whether the average cost is similar for the three types of packaging (A, B and M), we developed a graph of extremes and quartiles. When comparing the averages using graphical analysis alone, they appear to be similar. However, as graphical analyses always turn out to be more subjective than numerical ones, we carried out the analysis of variance test, where we obtained a p-value=$2\text{x}10^{-16}$ which is less than 5%, so we rejected H0, i.e. the averages are not similar, at the significance level tested.

Since it is not possible for the averages to be equal, we carried out a pairwise comparison of the averages to see if there are any similarities between them. However, the results obtained from the p-values are all less than 5% and therefore there is no equality between the means.

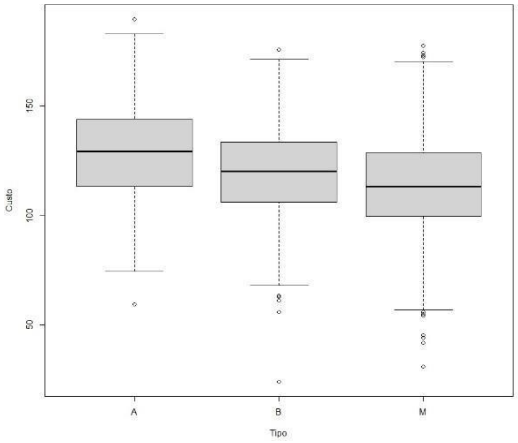

Figure 8- Graph of extremes and quartiles