ISSN: 1415-7314
ISSN online: 2317-6717

ENGEVISTA
Página da revista: http://www.uff.br/engevista/seer/

Use of multiple regression analysis on the improvement plan in a beverages industry

Elenice Kall[1]

Thiago Favarini Beltrame[2]

Abstract: The growing probability of implementing an improvement plan increases the interest on the subject of organizations and various industries. This is due to the fact that it contributes, mainly, in the possibility of a significant increase in organizational performance, culture change and the increase in human capital. This plan of improvements was carried out in a company of soda, with the factory in Santa Maria and Distribution Centers in Passo Fundo and Santa Cruz do Sul (RS), primarily in three sectors: quality control, maintenance and production. In order to measure the improvement plan it is necessary to apply the statistical tool that can correlate the dependent and independent variables. Methodologically, samples of a generic product A from line 3 were analyzed for 10 days in a row, in order to demonstrate the possibility of correlation between the loss of gas (dependent variable) with the torque and sequence of days (independent variables), and through this correlation perform the multiple regression. Since torque is the force applied on the lid in its open/close system and it is essential that it is standardized and controlled so as not to lose more than 15% of gas in the lid throughout the days. It has been found that there is a negative/strong correlation with carbonation and sequence of days, and by the significant regression (ANOVA), used to adjust the regression equation, it was found that the value of F (31.93067) is greater to the sequence of days, as well as p (0.000481) value is smaller in this variable. Analyzing the adjusted R-square (0.77460933) it was found that the biggest one can be found in the variable. With this, it was found an equation which allows predicting the correlation of reduction of carbon dioxide.

Keywords: Multiple regression, correlation, improvements plan.

[1] UFSM – Universidade Federal de Santa Maria
[2] UFSM – Universidade Federal de Santa Maria

ISSN: 1415-7314
ISSN online: 2317-6717

1. Introduction

The feasibility of implementing an improvement plan stimulates the interest on the subject of organizations in various industries because it contributes not only to improve the quality of products, services and processes, but also it allows a significant increase in organizational performance, in culture change and increasing human capital (Santos and Martins, 2008).

The plan aims to add improvements to deployment of strategic planning. However, according to Sellito and Ribeiro (2004) an important part of the strategic planning of organizations is the measuring of their results. If the measurement is inconsistent with the strategic goals, these cannot be achieved. To Bittici (1995) a system to measure results must have some capabilities: (i) to form the global view, avoiding local sub optimization; (ii) to deploy strategic objectives to operational levels; (iii) to provide the full understanding of structure of goals and conflicts; (iv) to adopt a hierarchical fashion, similar to a system of information, considering the operational capacity of the organization to collect and store the data required; and (v) to consider aspects of organizational culture.

It is known that the management system is a crucial ingredient of responsiveness to changes in the environment, because it determines the way in which the administration realizes the challenges, diagnoses their impacts, decides what to do and implements its decisions (Ansoff; McDonnel, 1993). And continuous improvement, according to Caffyn (1999), can be conceptualized ad a broad process focused on incremental innovation, which involves the entire organization.

However, to be able to measure the improvement plan it is necessary to accomplish an application of statistical tool so that the variables which are studied can be correlated. In order to check whether there is a relationship between two or more variables, or to determine whether the changes for one of the variables are accompanied by changes in the other one, a multiple regression analysis will be used to check for correlation between the loss of gas bottled beverages with torque.

For this work the variables studied were the effect of torque in plastic lids (torque) and the loss of gas in beverages, plastic covers tend to lose gas – especially if misapplied in the closing process – as days pass by, due to the chemical composition of the caps, which contains polypropylene, and the fact that they have undergone elastic deformation, the soda plastic bottled loses a certain amount of carbon dioxide (CO2), both through the walls of PET (polyethylene terephthalate) and the cover.

One of the goals of this improvement plan was to reach, basically, three sectors: quality control, maintenance, and production. Occasionally, one of the main purposes is to ensure the quality of 10 essential attributes for the enterprise sector of quality control for placing the rating. The attribute approached was the torque, which is the force applied on the lid in the open/close system. This force is not standardized or when not performed preventive maintenance on equipment in the

ISSN: 1415-7314
ISSN online: 2317-6717

specified period, it will generate and application in the covers by magnetic head out of the determined specification.

The level of carbonation will vary from product to product and for each there is a good effervescence. This is due to the flavor, taste and characteristics of different beverages. In general terms, fruit drinks are carbonated at low levels, colas and alcoholic beverages at medium levels, and other beverages such as tonics at high level to enable its dissolution in the liquid component in the carbonator (Hammer and Francis, 1993; Tocchini and Nisida, 1995).

It is extremely important that, after its determination, the carbonation is maintained in established patter depending on the type of beverage and the degree of acceptance by the consumer (Tocchini and Nisida, 1995; Martins, 2012).

The loss of carbon dioxide in beverages is an important factor to be considered in quality control of a product. This control involves not only the production stage, but also the characteristics of packaging systems and used for storage, transportation and distribution (Dantas, 1999).

The study was conducted in the industry of soda, with factory located in Santa Maria and Distribution Centers in Passo Fundo and Santa Cruz do Sul. The factory operates in the food producing, marketing lines and distributing beverages, with approximately 798 employees. The company has a factory in Santa Maria with 23,000 m^2 of built area and total area of 90,000 m^2, plus distribution centers in Passo Fundo and Santa Cruz do Sul. It currently operates with 4 lines that include modern manufacturing equipment (1 line for filling cans, 2 lines for PET packaging and 1 line for returnable glass bottles), making it self-sustaining in glass containers, aluminum cans and PET. Currently there are several categories of products available, all within the field of beverages, such as soft drinks, beers, juices, teas, energy drinks, mineral water, flavored waters, chocolate beverages and sports drinks.

2. Objective

This paper aims to conduct a verification of the amount of gas loss in a particular drink, and, on this aspect, find an equation which allows predicting the correlation reduction of carbon dioxide through a statistical technique known as multiple linear regression.

3. Methodology

To carry out the analysis of these data, scatter diagram, tests of correlation coefficient and regression analysis were used. Through regression analysis it is possible to calculate the value of a quantity in relation to the others, or combination of others (Leite et al., 2006). By statistical process, it is determined an algebraic expression that relates the dependent variable (time – days) to the independent variables or explanatory of their behavior (Sanvivente and Santos, 1995). The option to perform this study with multiple regression is due to the fact that statistical analysis technique

ISSN: 1415-7314
ISSN online: 2317-6717

exists in a large number of independent variables, able (or not) to explain the variation found on the dependent variable (Hair, Jr et al., 2005), mentioned in the work, which is the loss of carbon dioxide.

In addition, multiple linear regression is useful for selecting which variables are really meaningful and that, therefore, contribute to a better adherence of the model.

Another important point is that the multiple regression equation allows you to add any number of independent variables, which can take continuous or discrete values (Samohyl, 2009).

The pairs of values of two variables may be placed in a cartesian diagram called "dispersion diagram". The advantage of building a scatter diagram is that, often a simple observation already gives a fairly good idea of how the two variables are related.

A measure of the degree and sign of the correlation is given by the covariance between two random variables X and Y, which is a numerical measure of linear association between them, and defined by (1):

$$\text{Cov(X, Y)} = \frac{1}{n}\left[\sum x.y - \frac{\sum x.\sum y}{n}\right]. \quad (1)$$

It is more convenient to use a measure of correlation, the correlation coefficient Pearson estimator as ρxy, defined by (2):

$$r_{xy} = \frac{\text{Cov}(x, y)}{\hat{\sigma}_x \hat{\sigma}_y} = \frac{S_{xy}}{\sqrt{S_{xx}.S_{yy}}} \quad (2)$$

$$r_{xy} = \frac{\sum xy - \frac{\sum x . \sum y}{n}}{\left[\left[\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right].\left[\sum y^2 - \frac{\left(\sum y\right)^2}{n}\right]\right]^{\frac{1}{2}}} = \frac{S_{xy}}{\left(S_{xx} . S_{yy}\right)^{\frac{1}{2}}} = \sqrt{\frac{S_{xy}.S_{xy}}{S_{xx}.S_{yy}}} = \sqrt{\frac{b.S_{xy}}{S_{yy}}}$$

Where: the sums of squares are (3):

$$S_{xy} = \sum x.y - \frac{\sum x.\sum y}{n}; \qquad S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n}; \qquad S_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} \quad (3)$$

n = number of observation pairs

The correlation coefficient rxy is only an estimate of the population correlation coefficient ρxy and one must not forget that the value of rxy is calculated based on the "n" data pairs.

Often the sample points can have a correlation and yet not population. In this case, there is a problem of inference, since rxy ≠ 0 is not guarantee that ρxy ≠ 0. The problem can be solved by

ISSN: 1415-7314
ISSN online: 2317-6717

applying a hypothesis test to check if the value of rxy is consistent with the sample size n, in the significance level α, that there really is linear correlation between the variables.

H_0: $\rho = 0$ (there is no correlation between X and Y)
H_1: $\rho \neq 0$ (there is correlation between X and Y).

$$t_c = \frac{r_{xy} \cdot \sqrt{n-2}}{\sqrt{1-r^2_{xy}}} = \frac{r_{xy}}{S_r} \approx \text{ Distribution "t" of student with n – 2 degrees of liberty}$$

where: $S_r = \sqrt{\frac{1-r^2}{n-2}}$, standard error of the correlation coefficient

The correlation coefficient is a measure of linear relationship or more variables, and denoted by (R) indicates the closeness of the points to the regression line and the closer R is to 1.0, the closer the points are the linear regression; the closer R is to zero, the poorer is the adjustment of the regression line to the points (Maher, 2001).

The square of R, known as the coefficient of determination or (R2) regression, aims to disclose as the independent variables explain the variation of the dependent variables, i.e., is a measure that seeks to reflect how the values of Y are related to X, varying from 0 to 1, so that the closer to 1 the better (Leite et al., 2006).

According to Milone and Angelini (1995) correlation and regression are statistical technique that are based on the concepts of sampling to know how and if two statistical variables from the same population or not, are related to each other.

According to Gujarati (2000), possible relationships between the explanatory variables of the phenomena fall into "simple regression analysis" when studying the dependence of one variable with respect to a single explanatory variable, and in "multiple regression analysis" when the study includes more than one independent variable to explain the dependent variable.

More broadly, the techniques aim to generate a regression line that best fits a set of data points representing all data on certain variables, where the resulting estimates have a broader basis (Leite et al., 2006). Regression analysis aimed to investigate how the amount of gas and number of valves (independent variable) influence the elapsed time (dependent variable). The statistical tools available in the software Statistica were used for application testing.

4. Results

For a preliminary analysis of the existence of correlation between torque, time and carbonation, Table 1 and Figure 1 function as both exemplification and representation of the tests, which showed the torque behavior in relation to gas on line 03 along time. Samples of product X were analyzed along a 10-day period. A 50-percent variation in torque and 17.32 percent of gas loss were noticed, considering the reduction from the first to the last sample. The acceptable levels of gas

ISSN: 1415-7314
ISSN online: 2317-6717

loss cannot be higher than 5-8 percent; levels above that imply gas loss through the cap, thus showing that the main cause for gas loss was the plastic deformation provoked by the action of the magnetic heads.

Table 1. Carbonation and torque tests on line 03

Sequency	Date	Torque (lb.in)	Carbonatacion (Vol)	
1º	05/04/2011	12	4,85	
2º	06/04/2011	10	4,57	
3º	07/04/2011	11	4,57	
4º	08/04/2011	8	4,51	
5º	13/04/2011	10	4,36	% Torque variation: 50%
6º	14/04/2011	12	4,28	% Gas lost: 17,32%
7º	15/04/2011	11	4,31	
8º	16/04/2011	10	4,46	
9º	18/04/2011	6	4,20	
10º	19/04/2011	6	4,01	

Figure 1. Graphic of the Carbonation and Torque Test on line 03

As the purpose of this study was to find out an equation to predict the correlation of carbonic gas reduction by considering two independent variables (days and torque), first it was necessary to check the existence of correlation between those two variables.

The first step was to build the scatter plot, according to Figure 2, which shows the correlation. The closer the coefficient is to – 1 or + 1, the stronger the correlation is; on the other hand, the closer

ISSN: 1415-7314
ISSN online: 2317-6717

the coefficient is to zero, the weaker the correlation is. In this case, there is a negative, strong correlation.

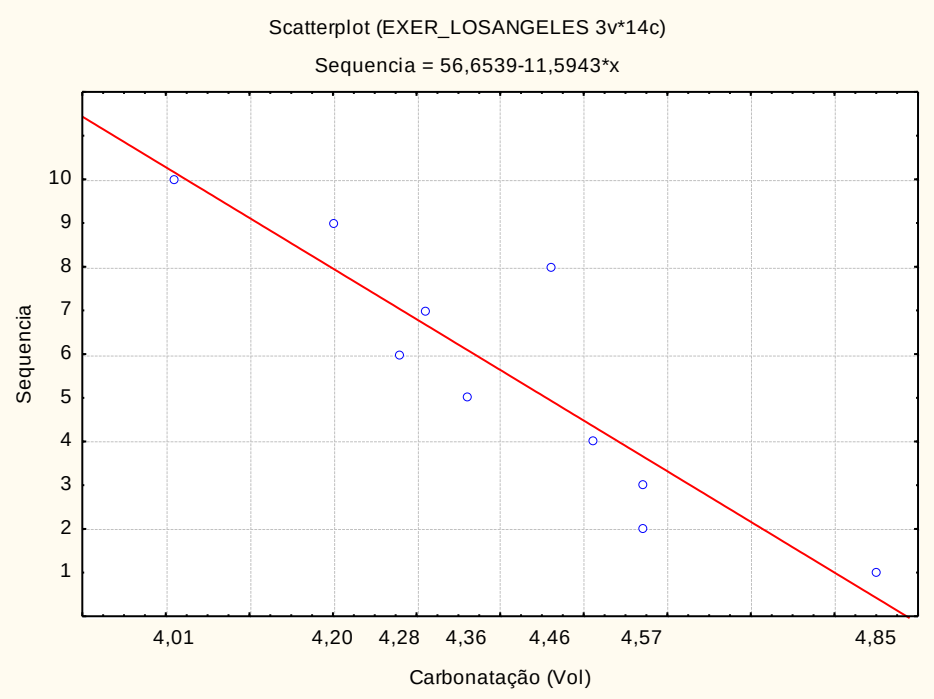

Figure 2. Graphic of dispersion

The next step was to conduct a significant regression between the variables to find out which independent variable was the most suitable to the model. The table 2 shows the significant regression analysis between variables.

Table 2. Results of the multiple regression

	Variable sequence days	***Variable days and torque sequence***	***Variable torque***
F	31,93067	14,61593	5,189359
p-value	0,000481	0,003170	0,052233
multiple R	0,89423304	0,89822040	0,62725627
R-Squared	0,79965273	0,80679988	0,39345042
Adjusted R-squared	0,77460933	0,75159985	0,31763173
comments	10	10	10

ISSN: 1415-7314
ISSN online: 2317-6717

As can be seen in Table 2, by combining the three possibilities to adjust the regression equation, F value is higher for the day sequence, and p-value is lower in this variable. By analyzing the adjusted R-squared, the highest of them is in this variable.

For the test of model significance, two possibilities were considered from the selection shown in Table 2. Table 3 presents a comparison between the two regression models designed for the two best F values and the best adjusted R-squared. X1 is the independent variable 'day sequence', X2 is the independent variable 'torque', and Y is the dependent variable 'gas loss'.

Table 3. Multiple regression models

Model I	$Y = 4{,}791333 - 0{,}894233X_1$
Model II	$Y = 4{,}652776 - 0{,}826035X_1 + 0{,}108619X_2$

The best model that could be found was model I, as defined by the values seen in Table 2. However, as the values of adjusted R-squared for both models were close, the option was to analyze the adjusted R-squared, since it must be preferred to the R-squared; F statistic, from the variance analysis, combined with its p-value, is preferred to the adjusted R-squared. The other step was the normality test (Figure 3) of the parameter defined as 'day sequence'. The independent variable 'days' was selected for this model.

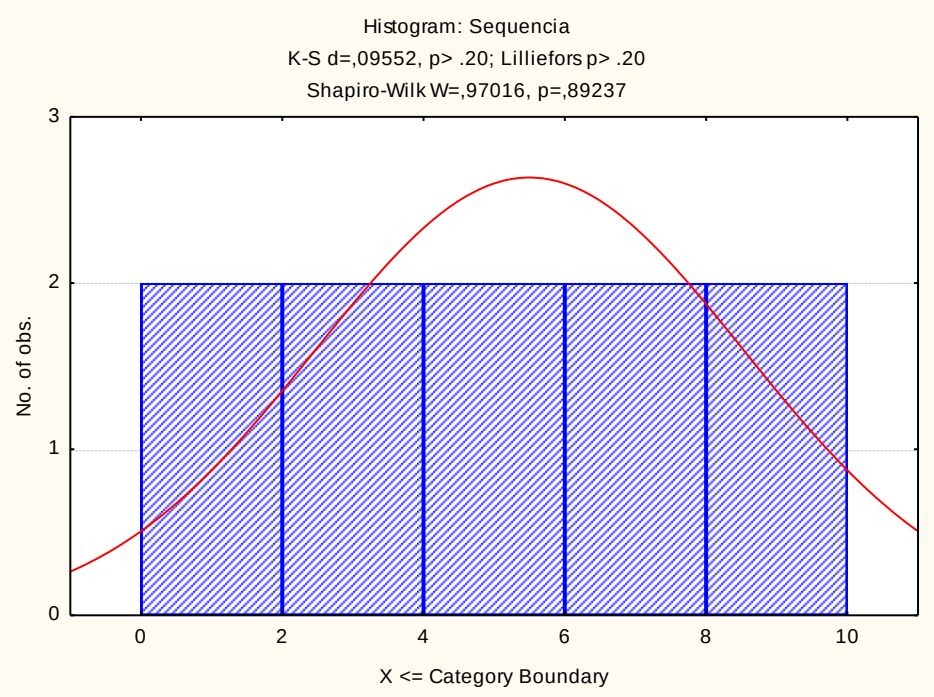

Figure 3. Analysis of normal sequence of days

It is known that the p-value is the lowest significance level that leads to the rejection of the null hypothesis Ho; the lower the p-value, the more significant the test. In this case, p = 0.89237, that is, p < 5%, so it is accepted that Ho does not follow a normal distribution. The normality condition

ISSN: 1415-7314
ISSN online: 2317-6717

is not necessary for obtaining the estimators of minimum squares, but it is fundamental for defining the confidence intervals and significance tests. This is also evidenced in Figure 4, which illustrates the distribution of the regression residuals, indicating that they do not follow a normal distribution.

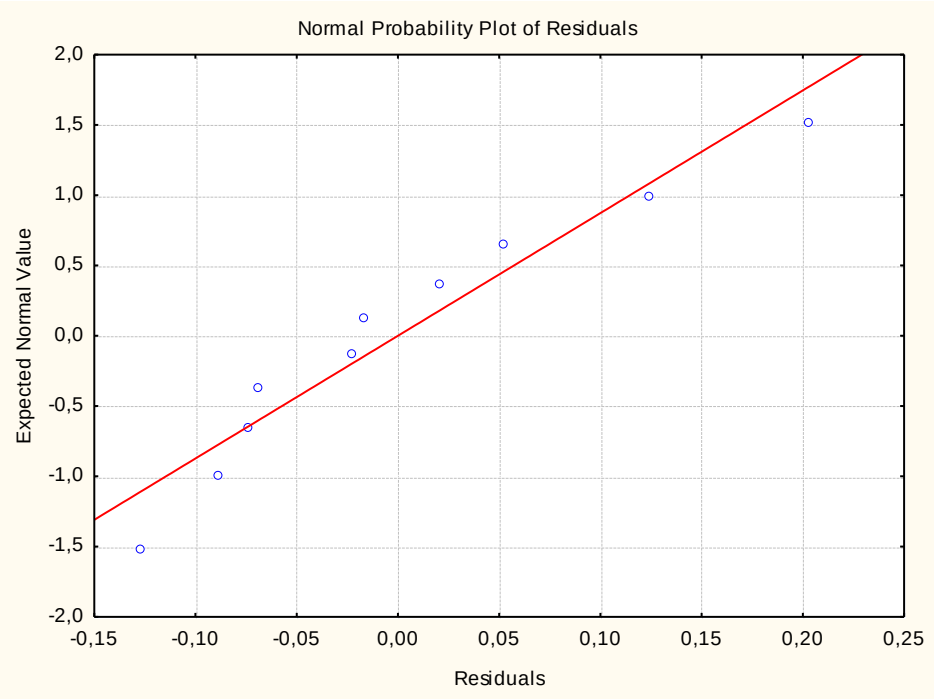

Figure 4. Analysis of normality of the regression residuals

Concerning homoscedasticity, this is the constant variance of residuals. It is desirable that errors are random, i.e. they must not be related to the characteristics of the variables under study. The dots are randomly distributed, without a definite behavior; in this case, there is homoscedasticity. But if there is some tendency (increase/decrease/oscillation), then there is heteroscedasticity. If there is heteroscedasticity, changes of variables (usually logarithmical) or other more complex solutions may be tried, and the model should be modified. The graphical analysis of the cloud of dots is simple and can be very useful. According to Figure 5, the dots are rather scattered, and this evidences that there is not an assumption of violation of the existence of homoscedasticity.

ISSN: 1415-7314
ISSN online: 2317-6717

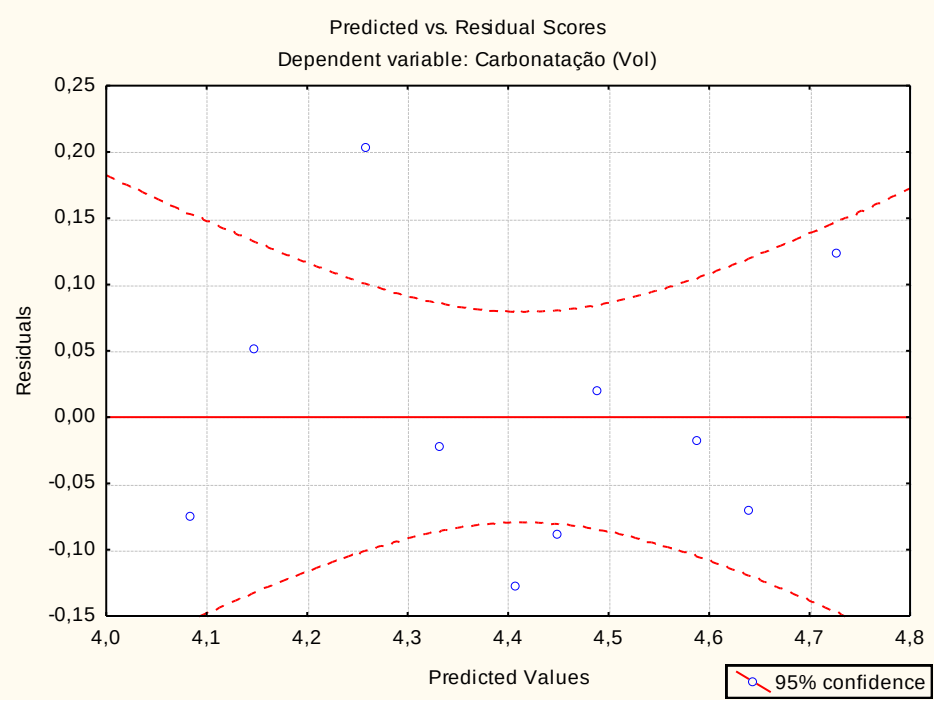

Figure 5. Analysis of homoscedasticity

For this model, outliers have not been found. Outliers generate false estimations in the model. Besides, the standard residual has a mean value equal to zero and constant variance, which show the occurrence of homoscedasticity.

Regarding the assumption of independence, the existence of autocorrelation among data should be considered, i.e. whether an observation was influenced by or influenced the observations performed both before and after it (Jordan, 2009). By using the Durbin-Watson test, it is possible to conclude that there is no independence among data; therefore, there is no autocorrelation.

The assumption of multicollinearity can be indirectly checked through the p-values of the coefficients obtained from the regression analysis. As all these values were extremely significant, even if there were multicollinearity, it would be overcome by the strength of relation existing between the variables (Samohyl, 2009). When there are more than two strongly related independent variables, there is multicollinearity. Multicollinearity significantly affects the coefficients of the regression equation by altering the value and even its sign in comparison to what would occur if this problem did not exist. The verification of existence of collinearity is performed through the examination of the correlation matrix by relating all the variables of the analysis, or by considering other criteria, such as the variance infraction vector (VIF). In this analysis, there is no problem of multicollinearity.

5. Conclusions

Finally, the main statistical assumptions of correlation and regression (normality, homoscedasticity, error independence, multicollinearity and linearity) were confirmed, thus evidencing the significance of the results obtained.

ISSN: 1415-7314
ISSN online: 2317-6717

The main limitation of this mathematical and statistical model is the fact that the equation obtained does not faithfully represent the dynamical process that occurs in the procedure, which in this case is the individual application of each plastic cap by the magnetic head. The mathematical equation found is static.

It was noticed that there is a negative, strong correlation with both the carbonation and the day sequence; by performing the significant regression (ANOVA) to adjust the regression equation, it was found that the F value (31.93067) is higher for day sequence, and the p-value (0.000481) is lower in this variable. By analyzing the adjusted R-squared (0.77460933), the highest of them is in this variable. Thus, an equation to predict the correlation of carbonic gas reduction has been found through the statistical technique known as multiple linear regression.

ISSN: 1415-7314
ISSN online: 2317-6717

6 References

ANSOFF, H. I.; MCDONNELL, E. J. 1993. *Implantando a administração estratégica*. 2 ed. São Paulo: Atlas. p. 592, (in Portuguese).

BITITCI, U. 1995. *Modelling of performance measurement systems in manufacturing enterprises*. Int. J. Production Economics, Elsevier Science B.V., v. 42, p. 137-147.

CAFFYN, S. 1999. *Development of a continuous improvement self-assessment tool*. International Journal of Operations & Production Management. v. 19. n.1.

DANTAS, S.T. 1999. *Embalagens e a sua interação com alimentos e bebidas*. Campinas: CETEA/ITAL, (in Portuguese).

FRANCIS, A.J.; HARMER, P.W. 1993. *Zumos de frutas y bebidas refrescantes*. Zaragoza: Acríbia.

GUJARATI, D.N. 2000. *Econometria básica*. São Paulo: Makron Books(in Portuguese).

HAIR Jr, J.F.; ANDERSON, R.E.; TATHAM, R.L.; BLACK, W.C. 2005. *Análise Multivariada de Dados*.5.ed. Porto Alegre: Bookman, (in Portuguese).

JORDAN, J.R. 2009. *Modelagem estatística para ensaios de resistência na indústria de celulose e papel*: Anais XXIX Encontro Nacional De Engenharia De Produção. Salvador (BA), (in Portuguese).

LEITE, R.M.; CLEMENTE, A.; GARCIA, R. 2006. *Análise de Regressão: uma ferramenta para a previsão de vendas*. UFPR, (in Portuguese).

MAHER, M. 2001. *Contabilidade de custos: criando valor para a administração*. São Paulo: Atlas, (in Portuguese).

MILONE, G.; ANGELINI, F. 1995. *Estatística Aplicada*. São Paulo: Atlas, (in Portuguese).

SAMOHYL, R. W. 2009. Controle Estatístico da qualidade. Rio de Janeiro: Elsevier, (in Portuguese).

SANTOS, A.B.; MARTINS, M.F. 2008. Modelo de referência para estruturar o Seis Sigma nas organizações. *Gestão e Produção*. São Carlos, v.15, n 1, (in Portuguese).

SANVICENTE, A.Z.; SANTOS, C. da C. *Orçamento na administração de empresas: planejamento e controle*. São Paulo: Atlas, 1995, (in Portuguese).

SELITTO, M.A.; RIBEIRO, J.L.D. 2004 Construção de indicadores para avaliação de conceitos intangíveis em sistemas produtivos. *Gestão e Produção*. Vol.11. Nº1, p.75-90, (in Portuguese).

TOCCHINI, R.P.; NISIDA, A.L.A.C. 1995. *Industrialização de refrigerantes*. Campinas: ITAL, (in Portuguese).

WHITE MARTINS. 2015. *Carbonatação*: a vida da bebida. Disponível em: <http:// www.whitemartins.com.br>. Acesso em: 2 de novembro 2015.