Technium Vol. 24, pp.88-92 (2024)
ISSN: 2668-778X
www.techniumscience.com

Quality Assessment of Electrical Equipment using ANOVA and Linear Regression Techniques

Catalin Silviu Nutu

Constanta Maritime University, Mircea cel Batran Bd. 104, Constanta, Romania
nutu_catalin@yahoo.com

Abstract. The article concerns utilization of linear regression and analysis of variance (ANOVA) for a set of data regarding the quality of electrical equipment. More precisely, this set of data concerns the nonconformant housings of electric motors for different diameters of housings. Both statistical methods, linear regression and ANOVA are employed to analyse this data with regard to their linear dependency and to check whether this data is likely to come from groups with the same mean.

Keywords. Housing of electric motor, linear regression, analysis of variance (ANOVA), nonconformity, defect

1. General aspects regarding ANOVA and Linear Regression

In the past it was a clear distinction between regression, analysis of variance (ANOVA) and analysis of covariance (ANCOVA). So, the statistic textbooks spoke of them as they were different entities designed for different types of problems. Nowadays all three procedures are regarded as being subsumed under what is called the general linear model (GLM).

Modern statistical software still contains separate procedures for regression and analysis of variance (ANOVA). These differences should be seen in terms of convenience in use of one procedure compared to the other, that is for certain types of data the problem of fitting a general linear model can be solved more convenient using an ANOVA procedure than a regression procedure.

Fitting a general linear model means to predict one or more dependent variables as a linear combination of independent variables. The dependent variables are also called response variables, while the independent variables are called predictor or explanatory variables and the coefficients of the linear combination are called weights or partial regression coefficients.

Regarding the notation the dependent variables are denoted with Y, the independent variables are denoted with X. The weights or coefficients are denoted with β_i where i is the corresponding independent variable. The two other features of the linear model are the intercept and the prediction error. The intercept is denoted by α and represents a constant. The prediction error, also called residual is the difference between the observation and the predicted value. The prediction error is denoted by e, while the prediction (predicted value) is denoted by Y^. The assumptions underlying the general linear model are the following four: linearity, normality of the residuals, equality of residual variances and fixed independent variables measured without error.

In our case the data to be analysed are in the table comprising defect housings corresponding to four main defects for different eight diameters of the housings of electric motors. In order to analyse the data, we check two assumptions using the regression analysis and the analysis of variance.

Technium Vol. 24, pp.88-92 (2024)
ISSN: 2668-778X
www.techniumscience.com

First assumption to be tested is that between the data in the table, there is a linear relationship that is to say that the nonconformant housings (defects) are related. This fact has the meaning that there is a systematic error in the production process or the means of production used are not reliable any more, the result being in errors in products (defects).

If the first assumption fails, the second assumption to be tested is that the defects come from the same population and the assumption of equal means is to be checked using the analysis of variance (ANOVA).

2. Theoretical aspects regarding the Linear Regression and ANOVA

This section of the paper presents the theory and the steps related to the application of both statistical techniques, that is to say: Linear Regression and ANOVA

2.1 Theoretical aspects on linear regression:

Regression analysis allows one to make predictions about the numerical dependent variable based on one or more numerical independent variables.

There are two possible outcomes: if such a relation is found, it is consistent with the hypothesis of a causal influence between the dependent and independent variables, if instead no relation is found then the result suggests that there is no causal influence between the dependent and independent variables. However, in both cases the presence or the absence of the causal influence is not proved.

In the regression analysis one has to check whether there is a linear relationship between the output and input, or between the dependent and independent variable or between the response and explanatory.

The regression model has the form:

$$y = \beta_0 + \beta_1 x + e \qquad (1)$$

where β_1 is called the regression coefficient or the slope of the line and e is the error term.

The parameters β_0 and β_1 need to be estimated and their estimations will be denoted as b_0 and b_1 and hence the equation is

$$y\hat{} = b_0 + b_1 x \qquad (2)$$

b_0 and b_1 are to be calculated by minimizing the sum of square errors

$$SSE=\sum_1^n e_i^2=\sum_1^n (y_i - y\hat{}_i)^2=\sum_1^n (y_i - b_0 - b_1 x_i)^2 \qquad (3)$$

Now, in order to test whether there is a linear relationship between two vectors, one must partition the total variation:

$$y_i - mean(y) = \left(y\hat{}_i - mean(y)\right) + (y_i - y\hat{}_i) \text{for i=1,2...n} \qquad (4)$$

By squaring the above relationship and summing from i=1 to n it follows:

$$\sum_{i=1}^n ((y_i - mean(y))^2 = \sum_{i=1}^n \left(y\hat{}_i - mean(y)\right)^2 + \sum_{i=1}^n (y_i - y\hat{}_i)^2 \qquad (5)$$

or in an equivalent form:

$$SS_{total} = SS_{regression} + SS_{residuals}$$

Which tells us that the total variation can be explained by the variation due to regression and the variation due to the fact that the observation do not lie on the regression line.

The alternatives to be tested in the case of linear regression are following:

(H_0): $\beta_1=0$ against (H_1): $\beta_1\neq 0$

The calculated test statistic is

$$f = MS_{regression}/MS_{residuals}$$

Technium Vol. 24, pp.88-92 (2024)
ISSN: 2668-778X
www.techniumscience.com

Where,

$$MS_{regression} = SS_{regression}/1,$$

(1 being the degree of freedom for regression)

$$MS_{residuals} = SS_{residuals}/(n-2)\ ,$$

(n-2 being the degree of freedom for residuals, where n is the number of observations)

The test statistic has Fisher distribution f with 1 respectively (n-2) degrees of freedom and depending on its value, one can conclude whether there is a linear relationship between y and x, or not.

If the value of calculated f statistic is greater than the value of the quantile (inverse of CDF) at x=0.99 the null hypothesis must be rejected and the alternative hypothesis accepted

2.2 Theoretical aspects on ANOVA:

In the analysis of variance (ANOVA), the goal is to check whether the groups have the same mean under the assumption that the groups come from the same normal population.

In order to achieve that goal, the total sum of squares is partitioned in the sum of squares between the groups and the sum of squares within the groups:

Thus:

$$SS_{total} = SS_{between\ groups} + SS_{within\ groups}$$

Taking into account of degrees of freedom the mean squares are calculated:

$$MS_{between} = SS_{between}/(k-1)$$

where k is the number of groups

$$MS_{within} = SS_{within}/(n-k)$$

where n is total amount of data

The concept of degrees of freedom is summarized by two points. First, the degree of freedom associated with a term equals the number of independent values of that term. Second, the degree of freedom is used as correction factor for the sum of squares (SS) associated with that term.

The sum of squares divided by the degrees of freedom gives the value of the mean squares (MS) and the MS is corrected in the sense that is an average amount of variation for each of the estimated independent parameters.

The f statistic is calculated as:

$$f = MS_{between}/MS_{within}$$

The f statistic is also known under the name of f ratio or f observed.

By comparing the obtained f statistic with the value coming from Fisher distribution with (k-1) and (n-k) degrees of freedom for a certain significance level one can conclude about the rejecting or not-rejecting of the null hypothesis (groups have the same mean).

If the value of f statistic is large, the null hypothesis can be rejected and concluded that there are more than chance differences between the means of the groups analysed.

How large the value of f statistic must be in order to reject the null hypothesis depends of the significance level chosen in order to perform the test. One must compare the value of the f observed with the value of f critical for a certain significance level. If f observed is greater than f critical the null hypothesis (means of the groups are equal) can be rejected in favour of the alternative hypothesis (means of the groups are not equal).

Technium Vol. 24, pp.88-92 (2024)
ISSN: 2668-778X
www.techniumscience.com

3. Linear regression and ANOVA techniques applied to the set of data

The following table summarizes the defect housings of electric motors, taking into account the type of defect and the diameter of the housing. The number of defects for each cell in the table corresponds to the same number of the housings of motors observed. Thus, each cell number comes from a sample of same number of observed housings.

Diameter	Defect1	Defect2	Defect3	Defect4
112	151	153	145	156
132	157	156	147	152
160	149	147	154	158
180	167	163	157	149
200	145	147	146	160
225	163	150	160	153
250	164	161	162	157
315	150	155	155	163

Table 1. Defects of four types, per diameter, for a population of housings of electric motors

The first assumption to be verified is that the four types of defects are somehow dependent, fact that can lead to the conclusion that a systematic error lies in the production process or the means of production used are not reliable any more.

In order to check whether there is linear dependency between the types of defects, one checks whether between each column vectors corresponding to Defect2, Defect3 and Defect4 and column vector corresponding to Deffect1 exists such a relationship.

For the first check (of linear dependency between Deffect2 and Deffect1), following results are obtained:
Calculated regression line: y=x*0.54565+69.015
SS tot=250
SS reg=138.60
SS res=111.40
Correlation coefficient between y and x: corr(x,y)=sqrt(SS reg/SS tot)=0.74457
Value of calculated f statistic is f=7.4644
Value of the 0.99 quantile for the F distribution with 1 and n-2=6 degrees of freedom is 13.745
Since 7.4644<13.745 the null hypothesis must not be rejected. Therefore at 0.99 significance level there is no linear relationship between y and x.

For the second check (of linear dependency between Defect3 and Defect1), following results are obtained:
Calculated regression line: y=x*0.55317+67.094
SS tot=299.50
SS reg=142.44
SS res=157.06
Correlation coefficient between y and x: corr(x,y)=sqrt(SS reg/SS tot)=0.68963
Value of calculated f statistic is f=5.4416
Value of the 0.99 quantile for the F distribution with 1 and n-2=6 degrees of freedom is 13.745
Since 5.4416<13.745 the null hypothesis must not be rejected. Therefore at 0.99 significance level there is no linear relationship between y and x.

For the third check (of linear dependency between Deffect4 and Deffect1), following results are obtained:
Calculated regression line: y=x*(-0.41676)+220.91
SS tot=144
SS reg=80.851
SS res=63.149
Correlation coefficient between y and x: corr(x,y)=sqrt(SS reg/SS tot)= -0.74931
Value of calculated f statistic is f=7.6819

Value of the 0.99 quantile for the F distribution with 1 and n-2=6 degrees of freedom is 13.745
Since 7.6819<13.745 the null hypothesis must not be rejected. Therefore at 0.99 significance level there is no linear relationship between y and x.

Now, performing analysis of variance for the data in the table (ANOVA) we obtain following results:
SS between groups=43 at 4-1=3 degrees of freedom
SS within groups=1159 at 32-4=28 degrees of freedom
MS between groups=43/3=14.3333
MS within groups=1159/28=41.3929
The calculated value of the f statistic is f=0.3463
Since f is smaller than the value of the 0.99 quantile of F distribution with 3 and 28 degrees of freedom which is 4.5681, the null hypothesis must not be rejected.
Therefore, at a significance level of 0.99 the four defects have the same mean.

4. Conclusions

Both tests used in this paper, the test for linear regression and the test for equal means as well, are subsumed to the general linear model.

The similarities of both tests are obvious and they regard, on one hand, the using of the F test and on the other hand the concept of partitioning the sum of squares, on the other.

Although the F test is used to assess different hypothesis the test is used in the same sense, that is to say the null hypothesis is rejected when the value of f statistic (f ratio or f observed) is greater than the value of the f critical corresponding to the F distribution with the corresponding degrees of freedom in both cases.

The concept of partitioning is the main idea behind the both – checking of regression and checking of equal means). The term partitioning is used with the meaning of "breaking into components". By breaking into components, the total sum of squares (SS tot) in the both cases and using those components by dividing them by the corresponding degrees of freedom the values of f statistic are obtained.
Using the results of the analysis one can conclude about the fact that there is no linear dependency between the defects at the significance level of 0.99.

Performing of the ANOVA on the same set of data it results that the four types of defects are very probable to come from groups with the same mean.

References

[1] K. Erwin, "Statistische Methoden Und Ihre Anwendungen", 7. Auflage, Vandenhoeck&Ruprecht

[2] C. Walck, "Handbook On Statistical Distributions For Experimentalists", University of Stockholm 2007

[3] M. A. Shayib, "Applied Statistics", 1st Edition 2013, Downloaded free from Bookboon.Com

[4] J. Miller, P. Haden, „Statistical Analysis with The General Linear Model", 2006

[5] W. Haerdle, L. Simar, "Applied Multivariate Statistical Analysis", Springer 2003