

Correlation Measures and Their Applications in Structural Dynamics and Data Analyses

Zhigang Wei
Tenneco Automotive Co., Ltd.

Limin Luo and Shengbin Lin
Tenneco Inc

ABSTRACT

This paper reviews the correlation concepts and tools available, with the emphasis on their historical origins, mathematical properties and applications. Two of the most commonly used statistical correlation indicators, i.e., modal assurance criterion (MAC) for structural deformation pattern identification/correlation and the coefficient of determination (R^2) for data correlation are investigated. The mathematical structure of R^2 is critically examined, and the physical meanings and their implications are discussed. Based on the insights gained from these analyses, a data scatter measure and a dependency measure are proposed. The applications of the measures for both linear and nonlinear data are also discussed. Finally, several worked examples in vehicle dynamics analysis and statistical data analyses are provided to demonstrate the effectiveness of these concepts.

CITATION: Wei, Z., Luo, L., and Lin, S., "Correlation Measures and Their Applications in Structural Dynamics and Data Analyses," *SAE Int. J. Mater. Manf.* 8(1):2015, doi:10.4271/2014-01-2307.

1. INTRODUCTION

In ground vehicle dynamic analysis, such as road load data acquisition (RLDA) related analysis for complex systems or vehicles, it is often required to find the dominant modes or deformation patterns. Then, much simpler bench tests are conducted on certain components to find the more detailed durability, reliability information of the components under the dominant modes or controlling loadings. Similar applications are encountered in many areas of experimental modal analysis and analytical structural dynamics. Clearly, the use of correlation techniques is necessary to find the dominant modes, to interpret the data available, and to simplify a test procedure. Many techniques quantifying the correlation between two pairs of vectors, such as experimentally observed and theoretically predicted, are available. Modal assurance criterion (MAC) [1] is such a tool to provide a data correlation measure, and it can be very powerful when used correctly but very misleading when used incorrectly [1].

The vehicle dynamics is not the only area the concept of correlation is widely used. In statistical data analysis, such as fatigue failure data analysis, the correlation is also a critical part in evaluating the accuracy of the models developed, the goodness-of-fit of the regression models developed and many other applications [2]. Fitting data with linear or nonlinear models are commonly applied in all science and engineering applications and correlation measures are needed to

quantitatively evaluate if a data fit is good. There are several correlation measures available and the earliest and still often used one is the Pearson correlation coefficient R [2, 3]. Another commonly used measure for data correlation is the coefficient of determination R^2 . The use of R^2 is well established in classical regression analysis [4] and widely used for data correlation measure in linear regression. Although it is also well known now, in mathematical literature and some science and engineering fields, that R^2 is not an adequate measure for nonlinear regression, many scientists and engineers still use it for data correlation assessment of nonlinear statistical data [5]. Additionally, almost all of the commercially statistical software packages calculate R^2 for nonlinear data that facilitates its frequent misuse in nonlinear data analysis [5]. To make R^2 a reasonable measure for data correlation for nonlinear data, many attempts have been made. For example, Nagelkerke [6] generalized the definition of R^2 with sample size effect and used the maximum likelihood as a criterion for fit. Cameron and Windmeijer [7] constructed R^2 for some particular models including logit, probit, Poisson, geometric, gamma and exponential.

Ever worse, have not been realized by most of the practitioners, R and R^2 are actually not good data correlation indicators even for linear data as indicated in [2] and no single measure including R^2 and R is an adequate indicator of the usefulness of the regression selection. Therefore, the meaning

of R^2 need to be fully understood so that the practitioners know what does R^2 stand for and how to interpret their observations based on the indicators.

In this paper, the correlation concepts used for structural dynamics and for statistical data correlation are treated in a unified way. The historical origins and mathematical structure of the MAC and R^2 measures are critically examined, with the emphasis on their physical interpretations.

2. CORRELATION MEASURE AND CAUCHY-SCHWARZ INEQUALITY

For two variables (vectors or random numbers) U and V , it is of significant interest to measure the relationship between them: their similarity, dependence and correlation. Correlation is the degree to which two or more quantities are associated. It would be desirable to have a correlation measure or definition that has a unique mathematical property, with which limits can be set, e.g., a range (Lower Bound, Upper Bound), for the values of a correlation measure. Ideally, the Lower Bound and Upper Bound are constants and represent two extreme situations, e.g. worst possible and best possible correlations. Furthermore, normalized values are preferred, i.e. Lower Bound=0 and Upper Bound=1. A natural mathematical tool applicable in this case is inequality, which describes a relationship that holds between two values when they are different. The Cauchy-Schwarz Inequality provides such as ideal language for the correlation measure [8].

The Cauchy-Schwarz Inequality is believed to be one of the most widely used and most important inequalities in all of mathematics [8]. The Cauchy-Schwarz Inequality states that for all vectors u and v of an inner product space it is true that $|\langle u, v \rangle|^2 \leq \langle u, u \rangle \cdot \langle v, v \rangle$. Moreover, the two sides are equal if and only if u and v are linearly dependent. In Euclidean space with the standard inner product, the Cauchy-Schwarz Inequality is

$$\left(\sum_{i=1}^n u_i v_i \right)^2 \leq \sum_{i=1}^n u_i^2 \sum_{i=1}^n v_i^2 \quad (1)$$

where n is the number of the data points.

The Inequality was published by Cauchy in 1821, proved by Bunyakovsky and Schwarz in 1859 and 1888, respectively [8]. With the elegant property of Cauchy-Schwarz Inequality, it would be of interest to come up with a single correlation parameter to measure the degree of correlation which has lower and upper bounds. A natural correlation measure derived from the Cauchy-Schwarz Inequality is Eq.(2)

$$r^2 = \frac{\left(\sum_{i=1}^n u_i v_i \right)^2}{\sum_{i=1}^n u_i^2 \sum_{i=1}^n v_i^2} \quad (2)$$

Clearly, $0 \leq r^2 \leq 1$. Eq.(2) has several interesting properties, and two of the most important properties [8] are:

- A. the correlation coefficient is independent of both origin and scale, mathematically, i.e.,

$$\begin{aligned} r^2(u, v) &= r^2(x, y) \\ u &= (x - x_0)/h \\ v &= (y - y_0)/k, \quad (h > 0 \text{ and } k > 0). \end{aligned} \quad (3)$$

- B. the correlation is symmetric, i.e.

$$r^2(u, v) = r^2(v, u) \quad (4)$$

These just the sort of nice behavior a measure of relationship are needed. Eq.(2) has many variants and two of the most important special cases are listed below:

1. MAC -Formula [1]

$$MAC = \frac{\left(\sum_{i=1}^n x_i y_i \right)^2}{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2} \quad (5)$$

for $x_0 = 0$, $y_0 = 0$ and $h = k = 1$ in Eq.(3).

2. Coefficient of determination [2]

$$R^2 = \frac{\left[\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \right]^2}{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2} \quad (6)$$

for $x_0 = \bar{x}$, $y_0 = \bar{y}$ and $h = k = 1$ in Eq.(3). Where, $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ and $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ are the respective means of x_i and y_i . Clearly, $0 \leq R^2 \leq 1$.

The MAC formula shown in Eq.(5) [1] is a statistical indicator that is most sensitive to large differences and relatively insensitive to small differences in the mode shapes [1], with 1 indicating fully consistent mode shapes and value near 0 indicates that the modes are not consistent. Actually, there is a number of similar assurance criteria available used in the area of experimental and analytical structural dynamics [1]. Nevertheless, the original MAC formula shown in Eq.(5) is the mostly used and proven to be an efficient measure of correlation and is widely applied in structural dynamics.

The coefficient of determination is a commonly used measure in statistical data analysis, but its use is still a controversial issue and the details will be discussed and addressed in the following two sections.

3. CORRELATION COEFFICIENT AND COEFFICIENT OF DETERMINATION

Historically, a similar variant of Eq.(6) was developed first from statistical field. In 1885, Sir Francis Galton, Charles Darwin's cousin, first defined the term "regression" and completed the theory of bivariate normal correlation and used on the relationship between physical characteristics of parent and offspring seeds [3]. In 1895, Karl Pearson developed the correlation efficient, Eq.(7), which is still used to measure correlation [3].

$$R = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (7)$$

Clearly, R in Eq.(7) is the square root of R^2 in Eq.(6), which can be directly derived from the Cauchy-Schwarz Inequality.

Different from R^2 , the range of R is $-1 \leq R \leq 1$ and this can be described as follows: $(x_i - \bar{x})(y_i - \bar{y})$ or R is positive if and only if x_i and y_i lie on the same side of their respective means. R is negative if and only if x_i and y_i tend to lie on opposite side of their respective means. Therefore, minus and plus signs indicate the increasing and decreasing trends, respectively.

The interpretation of R depends on the context and purposes, and 13 different formulae have been presented for R [3], and each of which suggests a different way of thinking about R , from algebraic to trigonometric settings, e.g. as a special type of mean, a special type of variance, the slope of a line, the cosine of an angle, and the tangent to an ellipse [3].

R is sometimes criticized as having no obvious interpretation for its units and this criticism is mitigated by using R^2 [3], which is more often used in applications to describe the degree of correlation. Eq.(6) is often expressed in the form shown in Eq.(8)

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = \frac{SS_{REG}}{SS_{TOT}} \quad (8)$$

where $SS_{TOT} = \sum_{i=1}^n (y_i - \bar{y})^2$, $SS_{RES} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ and $SS_{REG} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$, $\bar{y}$ is the mean of y_i , $\hat{y}_i$ is regression

value at point i . SS_{TOT} is called the total sum of squares or total variation. The greater is SS_{TOT} , the greater is the variation among the y observations. SS_{REG} is the regression sum of squares. Each deviation is simply the difference between the fitted value on the regression line and $\bar{y}$. The larger the SS_{REG} in relation to SS_{TOT} , the greater is the effect of the regression relation in accounting for variation in the y observations. SS_{RES} is the sum of residuals, and the larger the SS_{RES} , the greater is the variation of the observations around the fitted regression line. For a linear model, $y = a + bx$, it can be proven that $SS_{TOT} = SS_{REG} + SS_{RES}$, which is called liner partitioning in linear regression [2]. Two limiting values of R^2 are [2]:

1. If all observations fall on the fitted regression line, then $SS_{RES} = 0$ and $R^2 = 1$. This case indicates that the independent variable x accounts for all variation in the observation of y_i . In other words, the contribution to SS_{TOT} purely comes from the slope change of the curve.
2. If the slope of the fitted regression line is zero, then $SS_{TOT} = SS_{REG}$ and $R^2 = 0$. This case indicates that there is no linear association between x and y in the sample data, and the independent variable x is of no help in reducing the variation in the observations with linear regression.

It should be noted that for nonlinear curve fitting $SS_{TOT} \neq SS_{REG} + SS_{RES}$, therefore, R^2 is not an optimal choice for nonlinear data analysis and the use of R^2 in nonlinear data analysis is problematic [5].

4. THE NEW INTERPRETATION OF THE COEFFICIENT OF DETERMINATION

To fully understand the meaning of R^2 and properly utilize it in applications, the mathematical structure of R^2 must be understood first. For a linear data analysis, $SS_{TOT} = SS_{REG} + SS_{RES}$, we have $R^2 = SS_{REG} / (SS_{RES} + SS_{REG})$ as can be seen from Eq.(8). Therefore, the value of R^2 is controlled only by two parts: SS_{REG} and SS_{RES} . The contributions from each part can be schematically represented in Fig.1 [2].

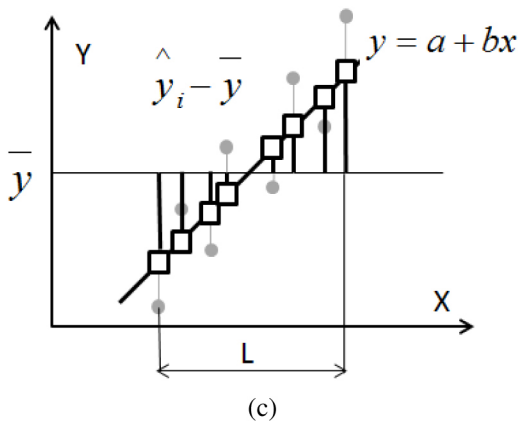
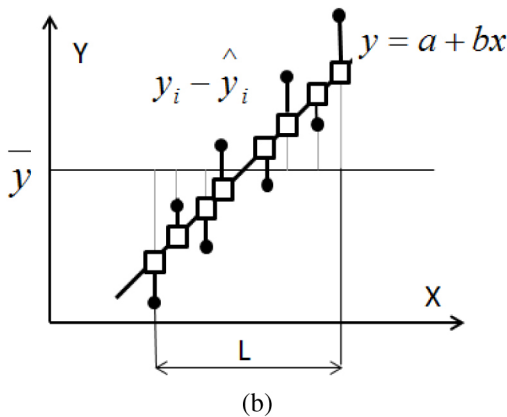
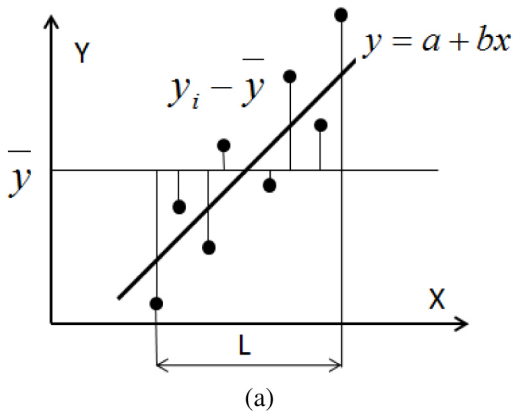


Fig. 1. Partitioning of total deviations and the interpretation of the coefficient of determination, (a) SS_{TOT} or $y_i - \bar{y}$, (b) SS_{RES} or $y_i - \hat{y}_i$, (c) SS_{REG} or $\hat{y}_i - \bar{y}$ [2]

Fig.1 reveals that SS_{TOT} can be decomposed into two distinct parts: (1) the part with symmetric characteristics, SS_{RES} , which is roughly symmetric with respect to the regression line $y = a + bx$; (2) the part with anti-symmetric characteristics, SS_{REG} , which is anti-symmetric with respect to the regression line $y = a + bx$. This is analogous to the concept of structural stress, which can be decomposed into two parts: uniform membrane stress and the anti-symmetric bending stress [9]. These two parts are fundamental elements for interpreting the meaning of R^2 and the key for developing possible new useful parameters to interpret the intrinsic data characteristics and patterns.

Clearly, due to its symmetric characteristics, the length scale must not be involved in the symmetric part SS_{RES} . By contrast, for anti-symmetric part SS_{REG} , the length scale must be an intrinsic part of a meaningful parameter and should be considered in interpreting the data pattern. Based on the observations, two distinct measures for characterizing data structures can be defined:

1. Data scatter measure, $\hat{\sigma} = \sqrt{MSE}$, which describes how big a data scatter is and how well a model correlates the test data. Where MSE is error mean square.
2. Dependency measure, $\hat{\rho} = \sqrt{MSS}/L$, which measures how strong the dependency is between the independent variable and the dependent variable. Where MSS can be called regression mean square and L is a characteristic length.

First, clearly, SS_{RES} is related to the scatter of data. For a given range of an independent variable, such as stress range in fatigue data analysis, the value of SS_{RES} is sample size (or data point number) dependent, e.g., the larger the number of data points, the larger the value of SS_{RES} . To objectively evaluate the good-of-fit, a sample size insensitive measure is required, and a simple way to do it is to normalize SS_{RES} with mean square $MSE = SS_{RES}/(n - 2)$, which is also called error mean square or residual mean square [2]. In fact, MSE is an unbiased estimator of variance for a regression model [2].

Therefore, the estimated standard deviation $\hat{\sigma} = \sqrt{MSE}$ can be used as a measure for data scatter and correlation. It reflects how big a scatter band is for a given data, and it is insensitive to the range selected. Additionally, since $\hat{\sigma}$ is length scale insensitive it can be eventually used for assessing the data correlation of nonlinear data. It should be noted that the definition $\hat{\sigma}$ is based on the assumption that the data follow the normal distribution with respect to the regression line. For other distributions, other measures can be used, e.g., Kolmogorov-Smirnov test and Anderson-Darling test are very commonly used for parametrical studies [2].

Second, for a given certain range of the independent variable, the value of SS_{REG} is also sample size (or data point number) dependent, i.e., the larger is the number of data points, the higher is the value of SS_{REG} . Similarly, to objectively evaluate the related parameters, the sample size insensitive measure is required. SS_{REG} can be normalized with mean square $MSS = SS_{REG}/(n - 2)$. Eventually, the mean square $\sqrt{MSS}$ can be used to characterize the average amplitude of the change for a given range. Additionally, since SS_{REG} and $\sqrt{MSS}$ are anti-symmetric measures and length scale dependent, the length scale must be used to characterize the gradient with respect to the length scale. A natural definition of the dependency parameter is $\hat{\rho} = \sqrt{MSS}/L$. It reflects how fast a dependent variable changes as a function of the independent variable, and it is insensitive to the range selected. It should be