



Application of linear regression analysis to acoustic fatigue data

Associated software: ESDUpac M1012



Endorsed by
The Royal Aeronautical Society

ESDU DATA ITEMS

Data Items provide validated information in engineering design and analysis for use by, or under the supervision of, professionally qualified engineers. The data are founded on an evaluation of all the relevant information, both published and unpublished, and are invariably supported by original work of ESDU staff engineers or consultants. The whole process is subject to independent review for which crucial support is provided by industrial companies, government research laboratories, universities and others from around the world through the participation of some of their leading experts on ESDU Technical Committees. This process ensures that the results of much valuable work (theoretical, experimental and operational), which may not be widely available or in a readily usable form, can be communicated concisely and accurately to the engineering community.

We are constantly striving to develop new work and review data already issued. Any comments arising out of your use of our data, or any suggestions for new topics or information that might lead to improvements, will help us to provide a better service.

THE PREPARATION OF THIS DATA ITEM

The work on this particular Data Item was monitored and guided by the Vibration and Acoustic Fatigue Committee. This Committee first met in 1966 and now has the following membership:

Chairman	
Prof R.D. Adams	— University of Bristol
Members	
Dr N.S. Ferguson	— University of Southampton
Mr S. Georgiou	— SELEX Sensors and Airborne Systems Ltd, Luton
Dr P. Green*	— BAE Systems, Preston, Lancashire
Dr A. Grillenbeck*	— IABG Space Test Center, Germany
Mr A.R. Harris*	— Northrop Grumman, USA
Dr Hilmi Kurt-Elli	— Rolls Royce plc, Derby
Mr D.M.A. Millar*	— Bombardier Aerospace Short Brothers plc, Belfast
Dr M. Nash	— QinetiQ, Farnborough
Dr S.A. Rizzi*	— NASA, Langley Research Center, USA
Dr J. Santiago Prowald*	— European Space Agency (ESTEC), The Netherlands
Mr E. Smyth	— Bombardier Aerospace Short Brothers plc, Belfast
Prof R.G. White	— University of Southampton (Member Emeritus)
Prof J.J. Wijker	— Dutch Space, Netherlands

* Corresponding Member

The construction and subsequent development of the Data Item were undertaken by

Dr J. Anderson	— Senior Engineer
Dr M. Bratos-Anderson	— Consultant.

The person with overall responsibility for the work in this subject area is Dr C.B. Chinoy, Head of the Aircraft Noise and Structural Dynamics Group.

APPLICATION OF LINEAR REGRESSION ANALYSIS TO ACOUSTIC FATIGUE DATA

CONTENTS

	Page
1. NOTATION	1
2. INTRODUCTION	4
3. LINEAR REGRESSION	4
3.1 Properties of the Least Squares Fit Method	8
3.2 Confidence Intervals	10
3.2.1 Confidence intervals for A and B	10
3.3 Confidence Intervals for Population Regression Line $\mu_{y x} = A + Bx$	11
3.3.1 Confidence intervals based on Student's t distribution	11
3.3.2 Confidence intervals based on F distribution	12
4. HYPOTHESIS TEST	13
5. QUADRATIC REGRESSION	15
6. OUTLIERS	15
7. SAMPLING AND REPLICATION	18
7.1 Sample Size	18
7.2 Replication	18
8. EXAMPLE	18
9. COMPUTER PROGRAM	22
10. REFERENCES	24

APPLICATION OF LINEAR REGRESSION ANALYSIS TO ACOUSTIC FATIGUE DATA

1. NOTATION

A	constant (intercept)
B	constant (slope)
b	material constant
C	constant
$\mathbf{e}$	vector with residuals as components
$\tilde{e}_i$	standardized residual
$(\tilde{e}_n)_i$	maximum normed residual
$(\tilde{e}_{stud})_i$	studentized residual
$(\tilde{e}_t)_i$	R - Student's value
e_i	residual
F	random variable associated with F probability density function
F_0	value of test statistic
f	value of test statistic (ratio of estimations of variance $\sigma_{\bar{y}_i}^2$)
$\mathbf{H}$	hat matrix
h_{ii}	diagonal elements of hat matrix
K_{rms}	material constant
k	number of data points in sample
l	number of measurement levels of x
m_i	number of replicate measurements of y for a given x_i
N	number of cycles to failure
n	degrees of freedom for Student's t distribution
n_1	degrees of freedom for F_{n_1, n_2} distribution

Issued December 2010 – 28 pages

n_2	degrees of freedom for F_{n_1, n_2} distribution
P	probability
p	probability density function
Q	sum of the squared deviation of observed values y_i from mean value $\mu_{y x_i}$
r	sample correlation coefficient
r^2	coefficient of determination
S	stress
S_{rms}	root mean-squared value of stress
s_{rl}^2	estimator of variance $\sigma_{\bar{y}_i}^2$
s_{exp}^2	estimator of variance $\sigma_{\bar{y}_i}^2$
t	random variable of Student's t probability density function
(t', t'')	interval of confidence
t_0	value of test statistic
x_i	sample values of independent random variable x
y_i	sample values of dependent random variable y
$\mathbf{y}$	vector representation of sample values of dependent random variable
$\hat{\mathbf{y}}$	vector representation of fitted mean values of dependent random variable
$\hat{y}$	fitted mean value of y at x (estimate of $\mu_{y x}$)
$\bar{y}_i$	sample average at x_i
Y_i	random variable with expected value $\mu_{y x_i}$ and variance $\sigma_{y x_i}^2$
α	level of significance
$1 - \alpha$	confidence level
ϵ	model error (unknown measurement error), Equations (3.2)
ε	strain
θ	parameter in Equation (3.25)

μ	mean value
$\mu_{y x}$	mean (expected) value of y at given x
$\hat{\sigma}$	standard deviation estimate for error ϵ
$\hat{\sigma}_{\hat{A}}$	standard deviation estimate for $\hat{A}$
$\hat{\sigma}_{\hat{y}}$	standard deviation estimate for $\hat{y}$
$\hat{\sigma}^2$	error variance estimator
ξ	random variable
$\hat{}$	above symbol indicates estimator
$\bar{}$	above symbol indicates average value
subscript rl	refers to regression line
subscript exp	refers to experimental results
$E[]$	is expected value

Abbreviations

GLARE	glass/aluminium reinforced laminate
GUI	graphical user interface
HCF	high cycle fatigue
LCF	low cycle fatigue
MLE	maximum likelihood estimation
$S - N$	stress versus number of cycles to failure
$\epsilon - N$	strain versus number of cycles to failure

2. INTRODUCTION

The Item describes linear regression analysis, the statistical method used for samples of random variables which are interrelated. The method is applicable to obtain stress-life ($S-N$), as well as strain-life ($\epsilon-N$) curves from experimental fatigue data. The strain-life ($\epsilon-N$) curves are particularly used for composites and polymers. For some materials, the ($S-N$) curves are characterised by three distinct zones, each of which can be approximated by straight lines, which refer to LCF (low cycle fatigue), HCF (high cycle fatigue) and the fatigue limit regions. Acoustic fatigue of structural components is high cycle fatigue, caused by the random loading of the component. With random loading the stress or strain is specified in terms of the root mean-squared values. During loading, essentially elastic deformation occurs in the material. In the HCF region the $S_{rms}-N$ curves in logarithmic form can be presented as a straight line,

$$\log N + b \log S_{rms} = \log K_{rms}, \quad (2.1)$$

where N denotes the number of cycles to failure, S_{rms} is the root mean-squared value of stress and both b and K_{rms} are material constants.

Data for the prediction of the life of structures which undergo acoustic loading are obtained by subjecting small coupons or specimens to a random force which causes the coupons to vibrate (Reference 9). The coupons are often in the form of small cantilevers where the strain may be measured near the root. Excitation of the coupon is by a narrow band random signal, usually centred on the fundamental frequency of the coupon. For a given level of strain (or stress) the life of the coupon, or the number of cycles to failure, is measured. The aim of the test is to determine a relationship between stress (or strain) and the number of cycles to failure in the form of an $S-N$ graph (References 4, 5 and 20). In this Item a procedure for testing the assumption of the linear model in the context of $S-N$ curves is provided. Similarly, a procedure for the detection of outliers is introduced. Samples of experimental data with runouts are not considered.

3. LINEAR REGRESSION

A sample of k measurement variable pairs (x, y) , represented by k points in (x, y) space is considered: $\{(x_1, y_1) \dots (x_k, y_k)\}$. It is assumed that the dependent variable, y , has much more uncertainty than the variable, x , controlled in the experiment. The uncertainty of variable x can be ignored, hence, only the dependent variable y is random. If a perfect linear correlation exists between both variables, then the relationship between any pair of experimental values (x_i, y_i) would be given by

$$y_i = A + Bx_i, \quad i = 1 \dots k, \quad (3.1)$$

where A and B are constants, which represent the intercept and the slope of the straight line in (x, y) space.

However, a perfect linear relationship between experimental data never occurs in practice. Scatter, caused by random effects, always exists. Thus, the relationship between the two measurable variables (x, y) can be expressed by a population regression model (Reference 17)

$$y = A + Bx + \epsilon \quad (3.2a)$$

or, in terms of pairs of measurement data (x_i, y_i) ,

$$y_i = A + Bx_i + \epsilon_i, \quad (3.2b)$$

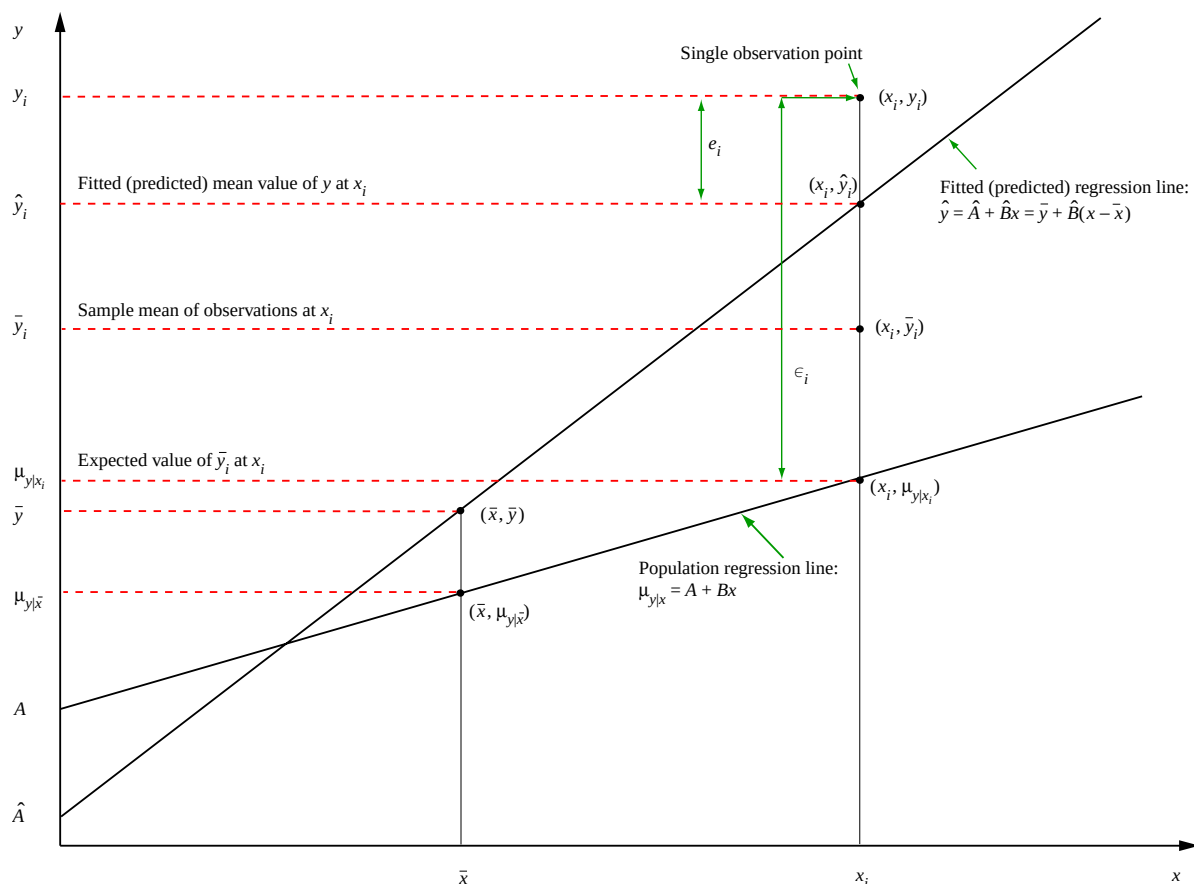
where ϵ_i is the unobserved measurement error by which the data value, y_i , differs from the value corresponding to the linear relationship (3.1).

It is assumed that the errors are random and independent, normally distributed with a mean value of zero ($E[\epsilon_{y|x_i}] = 0, i = 1 \dots k$) and the same variance, σ^2 . The data value, y_i , appearing in the pair (x_i, y_i) of sample data is the value of the random variable, Y_i , which corresponds to a fixed value of x_i (denoted by $Y_i = y|x_i$) with expected (mean) value $\mu_{y|x_i}$ and variance, $\text{Var}(y|x_i) = \sigma_{y|x_i}^2$. The assumptions, referring to errors, combined with either Equation (3.2a) or Equation (3.2b), leads to the conclusion that the random variables $Y_i (i = 1 \dots k)$ follow the normal distribution with mean $\mu_{y|x_i}$ and unknown variance $\text{Var}(y|x_i) = \sigma_{y|x_i}^2 = \sigma^2$.

The independent controlled variable, x , in Equations (3.2) is called the regressor and the dependent random variable, y , is described often as the response (Reference 7). Since in Equations (3.2) there is only one regressor variable, Equations (3.2) can be addressed as a simple linear regression model (Reference 7). For the linear model the mean or expected response is linearly related to the regressor, x ,

$$\mu_{y|x} = E[y|x] = A + Bx, \quad (3.3)$$

where A and B are the regression coefficients (Reference 7). See Sketch 3.1.



Sketch 3.1 Population and fitted regression lines

Equation (3.3) is described as a population regression equation (Reference 7). The line $y = A + Bx$ is called the true regression line and coefficients A and B are described also as true values (Reference 19). They are constants, the values of which are unknown during the replication of the experiment. The constants, A and B , can be estimated from the sample data using the least squares fit method. The estimates $\hat{A}$ and $\hat{B}$ of the constants A and B are random variables, the values of which vary from sample to sample. The aim of the method is to find the least squares line,

$$\hat{y} = \hat{A} + \hat{B}x, \quad (3.4)$$

described as a fitted or predicted line (References 6, 7 and 19) with the best estimators of the coefficients A and B . The best estimation of A and B is achieved through the minimalization of the sum, Q , of the squared deviations of the experimental values of y_i from the mean value, $\mu_{y|x_i} = A + Bx_i$ at x_i , i.e.,

$$Q = \sum_{i=1}^k (y_i - A - Bx_i)^2. \quad (3.5)$$

The procedure is called linear regression or the method of least squares fit (References 1, 2, 6, 12, 15, 17 and 19). The least squares criterion, minimalization of Q , leads to the system of the two equations:

$$\left. \frac{\partial Q}{\partial A} \right|_{\hat{A}, \hat{B}} = 0 \quad \text{and} \quad \left. \frac{\partial Q}{\partial B} \right|_{\hat{A}, \hat{B}} = 0, \quad (3.6)$$

or

$$\hat{A}k + \hat{B} \sum_{i=1}^k x_i = \sum_{i=1}^k y_i \quad (3.7)$$

and

$$\hat{A} \sum_{i=1}^k x_i + \hat{B} \sum_{i=1}^k x_i^2 = \sum_{i=1}^k x_i y_i. \quad (3.8)$$

Equations (3.7) and (3.8) are called the normal equations and their solution provides the estimators (indicated by $\hat{}$) of the constants, A and B ,

$$\hat{A} = \bar{y} - \hat{B}\bar{x} \quad (3.9)$$

and

$$\hat{B} = \frac{\sum_{i=1}^k (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^k (x_i - \bar{x})^2}, \quad (3.10)$$

where $\bar{x} = \frac{\sum_{i=1}^k x_i}{k}$, $\bar{y} = \frac{\sum_{i=1}^k y_i}{k}$ and k denotes the sample size. The resulting line, Equation (3.4), is called the fitted or predicted linear regression line for y on x . The estimators, $\hat{A}$ and $\hat{B}$ are normally distributed with expected values A and B , respectively. They are linear unbiased estimators with minimum variance, hence, they are described as 'the best linear estimators' (References 2 and 17). The variable, $\hat{y}$ is the predicted mean value of the random variable, y , for a given x .

Another statistic is $\hat{\sigma}^2$, the unbiased estimator of the model error variance ($\hat{\sigma}_{y|x}^2 = \hat{\sigma}^2$), given by

$$\hat{\sigma}^2 = \frac{1}{k-2} \sum_{i=1}^k (y_i - \hat{A} - \hat{B} x_i)^2 \quad (3.11)$$

or,

$$\hat{\sigma}^2 = \frac{1}{k-2} \sum_{i=1}^k e_i^2, \quad (3.12)$$

where e_i denotes the residual, associated with the point (x_i, y_i) . The residual, e_i , is the difference between the value y_i and the fitted value, $\hat{y}_i = \hat{A} + \hat{B}x_i$, at x_i . It defines the vertical distance of the point (x_i, y_i) from the fitted line. The residuals play a significant role in the investigation of the adequacy of the regression model, *e.g.*, the assumption of the normal distribution for the unobservable errors, ϵ_i , can be validated by the Anderson-Darling test for the normality of the distribution of the residuals (References 8, 14 and 17). The expression (3.12) is also called the residual mean square (Reference 17). Finally, the Equation (3.11) defines also the estimator of the standard deviation of the distribution of y_i about the fitted (predicted) regression line,

$$\hat{\sigma} = \left[\frac{1}{k-2} \sum_{i=1}^k (y_i - \hat{A} - \hat{B}x_i)^2 \right]^{1/2}. \quad (3.13)$$

The standard deviation, $\hat{\sigma}$, is a measure of uncertainty in y_i ($i = 1, \dots, k$). Knowing the standard deviation estimator, $\hat{\sigma}$, it is possible to derive estimates of the standard deviations for $\hat{A}$ and $\hat{B}$, which are, respectively, (References 3, 12 and 15)

$$\hat{\sigma}_{\hat{A}} = \hat{\sigma} \left[\frac{\sum_{i=1}^k x_i^2}{k \sum_{i=1}^k x_i^2 - \left(\sum_{i=1}^k x_i \right)^2} \right]^{1/2} = \hat{\sigma} \left[\frac{1}{k} + \frac{\bar{x}^2}{\sum_{i=1}^k (x_i - \bar{x})^2} \right]^{1/2} \quad (3.14)$$

and

$$\hat{\sigma}_{\hat{B}} = \hat{\sigma} \left[\sum_{i=1}^k (x_i - \bar{x})^2 \right]^{-1/2}. \quad (3.15)$$