APPENDIX H- REGRESSION FITTING AND CORRELATION

Description

Regression is one way to fit a function to data. The technique finds the coefficients of a mathematical relationship that most closely fits to actual data by minimizing total composite error. Composite error here is defined as the square of the difference between the model output and the data value summed for each data point. Another term for this type of solution is "least squares." Standard correlation for a 2-dimensional model indicates the strength of a linear relationship between the two dimensions. An alternate method for finding a model fit is maximum likelihood estimation (MLE). Both methods agree reasonably well for large sample sizes but may disagree significantly for small sample sizes (20 occurrence data points or less).

Purpose

Curve fitting is used in establishing a continuous and relatively smooth model for interpolating or extrapolating output values that are between or outside the data values themselves. Such extensions are useful for reliability engineering purposes to make inferences about the underlying failure mechanisms and to evaluate compliance with specified reliability requirements.

A probability problem solution requires finding the probability after the model and the input values are known or given. A problem in statistics is the opposite of a probability problem. When you are trying to establish an appropriate model based on analysis of actual data values - you are performing statistical analysis. Regression fitting qualifies as statistical analysis.

The correlation coefficient quantifies the goodness-of-fit. It measures the strength of a linear relationship between the horizontal and vertical scale on a 2D plot.

Benefits

Regression provides a less-biased result than the MLE solution for smaller sample sizes, although both methods give good results when sample sizes are large. Regression is easier to explain than MLE for justification of technique used for example in presentations. The correlation coefficient is a straightforward way to evaluate the model goodness-of-fit. Establish time change criteria of components to ensure that a system continues to function at its design reliability goals in the future.

Implementation

Often more complex non-linear relationships are linearized first so that simpler regression methods can be used. For general relationships, this may lead to overemphasis of certain portions of the input/output range. However, for reliability/probability analysis, this is mostly acceptable since the area of greatest importance (the high reliability area) is the one generally emphasized during the probability plot regression process.

Process Flow

Locate the data points being analyzed on the appropriate axis or axes. For Weibull data, for example, match the times to failure on the horizontal scale with an estimated probability of occurrence on the vertical scale using either the median rank plotting positions or some other vertical plotting position estimate such as mean rank or Hazen's rank. For Crow/AMSAA analysis, the X-Y data is plotted on standard log-log scaling.

Solve for the fit line through the data with a standard regression formula, reference Abernethy (2002). Note that for regular 2-parameter Weibull analysis, the horizontal scale becomes the dependent variable and the vertical scale becomes the independent variable. For grouped data, the horizontal scale becomes the independent variable and the vertical scale becomes the dependent variable. The least squares solution formula usually found in standard reference books can be directly used for grouped data. For point-by-point data, the axes must be swapped. For Crow/AMSAA data, the standard textbook formula for regression may be used.

Example

For example, suppose results indicate failure ages of 23, 44, 52, and 18 operating hours on four bearing units during an endurance test. This is discrete event data. You want to establish a cumulative distribution function (CDF) model so that you can estimate how many hours of operation are allowable on each unit if only 10 percent of the entire bearing population can fail.

Solution:

First, put the four data values in increasing value order and assign median ranks accordingly with Y (dependent variable) and X (independent variable). The horizontal scale with the original data values is considered to be the dependent scale for point-by-point data:

Y vs. X

18x0.1591

23x0.3864

44x0.6136

52x0.8409

Formulas have been established for solutions of this type based on minimizing the sum of the squared deviations between the fit line and the actual data points. The regression solution formulas are the following:

$B = (\Sigma(x * y) - (\Sigma (x) * \Sigma (y)) / N) / (\Sigma (x \wedge 2) - \Sigma (x \wedge 2) / N)$

$A = \Sigma (x * y) - B * (\Sigma (x))$

WHERE … x and y as used here are transformed versions of the actual X and Y (2-dimensional) data values to adhere to a Weibull scaled plot. Note that the summation symbol (Σ) indicates summation over all data values.

$x = \ln(\ln(1 / (1 - F)))$

$y = \ln(Y)$ AND … the correlation coefficient (r) is given by:

$r = (\Sigma (x * y) - (\Sigma (x) * \Sigma (y)) / N) / \text{sqr}((\Sigma (x \wedge 2) - \Sigma (x \wedge 2) / N) * (\Sigma (y \wedge 2) - \Sigma (y \wedge 2) / N))$

The solution produces a Weibull Beta value of 2.064 and a Weibull characteristic life value of 39.32 hours with correlation coefficient of 0.96073 and square of correlation of 0.923 (an average fit for this sample size). The fit line crosses 10 percent occurrence at 13.215 hours, and that is your answer. You can allow just 13.215 hours of operation on all of the bearings and expect approximately 10 percent failure.

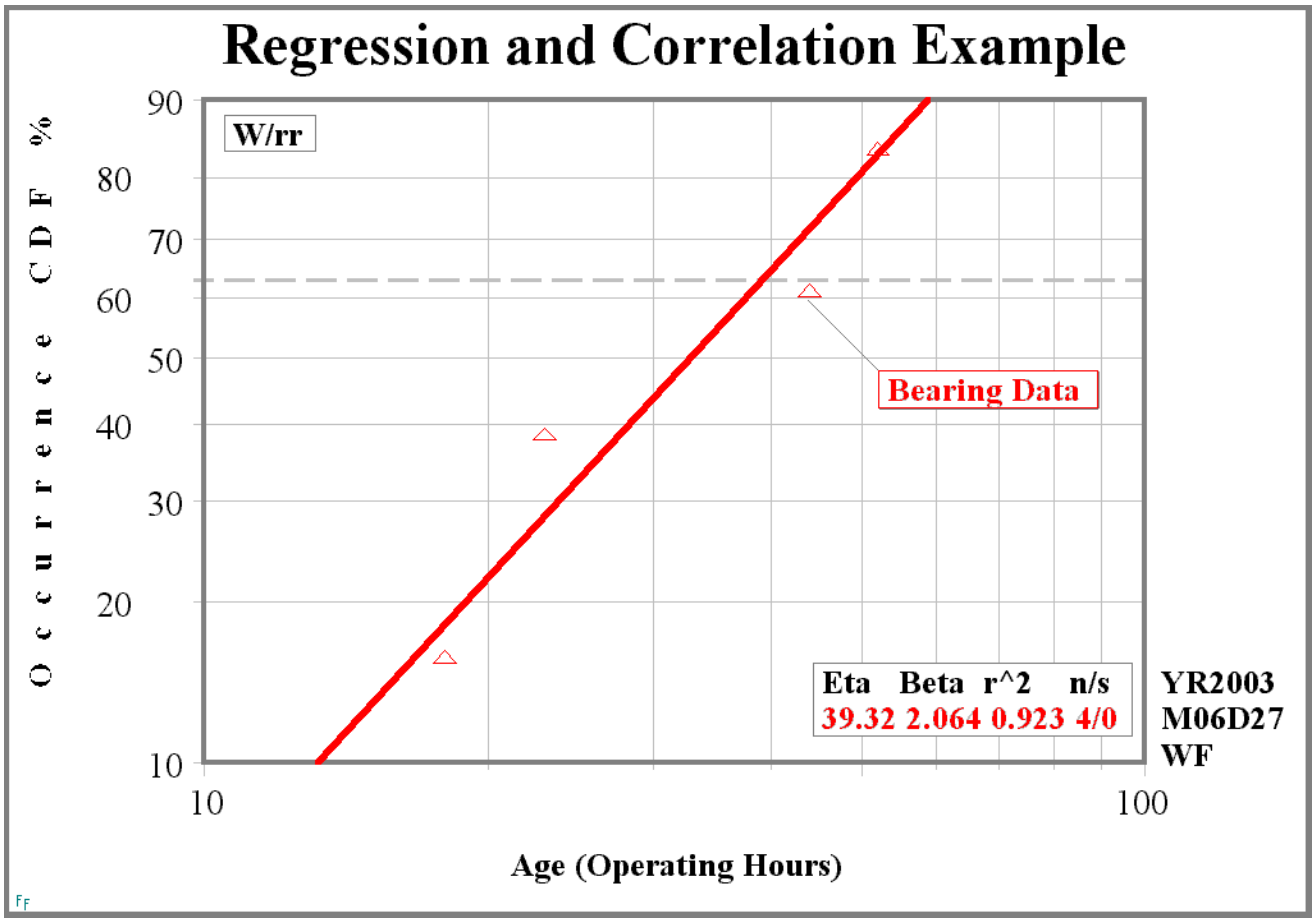

Figure H.1. Example of Regression and Correlation

General Comments

Least squares solution is fast and accurate if the data points are actual occurrences and not suspensions. Inclusion of suspensions (non-occurrences) for reliability issues may complicate the solution.

References

Abernethy, Robert B. *The New Weibull Handbook*, 4th Edition. Self-published, North Palm Beach, FL. Copyright 1993-2002.

SAE JA1000-1.

ISO 10017.