Linear Regression

Rustom D. Sutaria – Avia Intelligence 2016, Dubai

Introduction

As has been previously discussed in the paper on 'Monte Carlo Simulation', Risk analysis is an increasing part of every decision we make where aircraft maintenance planning & reliability are concerned. Further more risk tends to relate to cause & effect, not least the uncertainties associated with aircraft operations and maintenance.

In Linear regression, the cause and effect relationship is studied. the Independent variable is the cause (Ramp collision, lightning strike, failure on installation), and the dependent variable is the effect (damage to an aircraft or component, grounding, etc). Linear regression will help the reliability engineer to identify the dependent variables, thus contribute to the overall effectiveness of aircraft maintenance.

The reader should note, that this paper only deals with simple linear regression.

What is Linear Regression?

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable (independent variable or cause), and the other is considered to be a dependent variable (Effect).

For example, a reliability engineer might want to relate the individual component failures to their associated systems or 'next highest assemblies' to assess operational impact or effect using a linear regression model.

Prerequisites for Regression

Relationship

Before embarking on such a study, the reliability engineer should first determine whether or not there is a relationship between the particular component or system, and the next highest assembly or even the aircraft as a whole.

It is worth bearing in mind, that any given component or system may not necessarily imply the causes of inoperability with regard to the next highest assembly, associated system or the aircraft, but that there is some significant association between the individual component & the system or aircraft.

A scatterplot can be a helpful tool in determining the strength of the relationship between the two.

If there appears to be no association between the proposed component and the system (i.e., the scatterplot does not indicate any increasing or decreasing failure trends), then fitting a linear regression model to the data probably will not provide any useful data for interpretation.

Standard Deviation

For each value of X, the probability distribution of Y has the same standard deviation σ. When this condition is satisfied, the variability of the residuals will be relatively constant across all values of X, which is easily checked in a residual plot.

For any given value of X,

- The Y values are independent, as indicated by a random pattern on the residual plot.
- The Y values are roughly normally distributed (i.e., symmetric and unimodal).

Note:
A little skewness is ok if the sample size is large. A histogram or a dotplot will show the shape of the distribution.

The Least Squares Regression Line

Linear regression finds the straight line, called the **least squares regression line** or LSRL, that best represents observations in a bivariate data set. Suppose *Y* is a dependent variable, and *X* is an independent variable. The population regression line is: A linear regression line has an equation of the form:

$$Y = B_0 + B_1X$$

Where:

B_0 is a constant,

B_1 is the regression coefficient,

X is the value of the independent variable or cause, and

Y is the value of the dependent variable or effect.

However, we all know that aviation is an independently variable business however, knowledge and experience in the field, tends to identify known outcomes or effects. Given these circumstances, the reliability engineer most likely will apply the following population regression line which has an equation of the form:

$$\hat{y} = b_0 + b_1x$$

where:
b_0 is a constant,
b_1 is the regression coefficient,
x is the value of the independent variable, and
ŷ is the *predicted* value of the dependent variable.

How to Define a Regression Line

Normally, you will use a computational tool - a software package (e.g., Excel) or a graphing calculator - to find b0 and b1. You enter the X and Y values into your program or calculator, and the tool solves for each parameter.

In the unlikely event that you find yourself on a desert island without a computer or a graphing calculator, you can solve for b_0 and b_1 "by hand". Here are the equations.

$$b_1 = \Sigma \, [\, (x_i - xbar)(y_i - ybar) \,] \, / \, \Sigma \, [\, (x_i - xbar)^2]$$

$$b_1 = r * (s_y / s_x)$$

$$b_0 = ybar - b_1 * xbar$$

where

b_0 is the constant in the regression equation,
b_1 is the regression coefficient,
r is the correlation between x and y,
x_i is the *X* value of observation *i*,
y_i is the *Y* value of observation *i*,
x is the mean of *X*,0
y is the mean of *Y*,
s_x is the standard deviation of *X*, and
s_y is the standard deviation of *Y*