

Design of Experiments:

Factors are parameters we use as input variables in our experiments. The method fractional factorial design of experiments uses the term factor [1, 2]. To make DoE easier and more attractive to industry, Dr. Taguchi developed the Taguchi methods [4]. Taguchi intended these methods as cost-effective methods to improve the performance of a product by reducing its variability in customer usage conditions. Since the intention is to improve companies' competitive position, these methods have attracted the attention of many industries and academic communities across the globe [3].

These methods use Orthogonal Arrays (OAs) to minimize the number of test runs (or combinations) needed for an experiment. In Taguchi methods, the key role of OAs is to permit engineers to evaluate a system design with respect to robustness against noise and cost involved by changing settings of control factors (system design parameters). An OA is an inspection device to prevent a "poor design" from going "downstream". Arrays can have factors with many value levels, the most commonly encountered are two and three level factors. OAs should be used to help us discover experimental failures when interactions exist [3]. The amount of tests we should perform is highly dependent on the Degrees of Freedom (DoF) for the factors, being the amount of distinct information we seek [4].

Regression:

Multiple linear regression, MLR, is the classical method that combines a set of several x-variables in linear combinations, which correlate as closely as possible to the corresponding single y-vector. The MLR model equation can be written as equation 1 [2].

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_ix_i + f$$

We wish to find the vector of regression coefficients (b) so that the error term (f) is the smallest possible. To do this, one uses the least squares criterion on the squared error terms: find b so that $f^T(T)^*f$ is minimized [5].

Logistic regression is one of the most common forms of general linear models and is suited for cases where the response distribution is binomial [2]. Equation 2 shows the logistic regression equation, where the sigmoid function maps the predicted response to a range between 0 and 1.

$$\log\left(\frac{y}{1-y}\right) = b_0 + b_1x_1 + b_2x_2 + \dots + b_ix_i + f$$
$$y = 1/(1 + e^{-(b_0 + b_1x_1 + b_2x_2 + \dots + b_ix_i + f)})$$

Surface regression is a form of regression where we aim to create a surface that describes the variation in our data, being how the dependent output variable varies as a function of the independent input variables. We can use a first order model approximation to fit the surface to our data, as in equation 1. If there is a curvature in the system performance, we must use a polynomial of higher degree, like the second order model in equation 3 [2].

$$y = b_0 + \sum_{i=1}^k b_ix_i + \sum_{i=1}^k b_{ii}x_i^2 + \sum_{i < j} b_{ij}x_ix_j + f$$

In the context of this work, the probability of detection is binomially distributed with probability of success being 'y'.

[1] K. Dunn, Process Improvements Using Data, 2021.

[2] D. C. Montgomery, Design and Analysis of Experiments, 8th ed. Wiley, 2017.

[3] G. Taguchi, R. Jugulum, and S. Taguchi, Computer-based robust engineering : essentials for DFSS, Milwaukee, Wisconsin: ASQ Quality Press, 2004.

[4] R. K. Roy, E. J. Kehoe, Ed. A Primer on the Taguchi Method, 2nd ed. Michigan, USA: Society of Manufacturing Engineers (in English), 2010, p. 304.

[5] K. H. Esbensen, D. Guyot, F. Westad, L. P. Houmøller, and A. S. A. Camo, Multivariate data analysis - in practice : an introduction to multivariate data analysis and experimental design, 5th ed. Oslo: Camo, 2001.

Logistic regression:

Let x_k denotes the coded variable (values between -1 and 1) corresponding to factor k . Further, let $p(y = 1|x)$ denote the probability of locating a distressed vessel given the values of the explanatory variable; x . It follows by Bayes theorem that,

$$p(y = 1|x) = \frac{p(x|y = 1)p(y = 1)}{p(x|y = -1)p(y = -1) + p(x|y = 1)p(y = 1)}$$

$$\Downarrow$$
$$p(y = 1|x) = \frac{\frac{p(x|y = 1)p(y = 1)}{p(x|y = -1)p(y = -1)}}{1 + \frac{p(x|y = 1)p(y = 1)}{p(x|y = -1)p(y = -1)}}$$
$$\Downarrow$$

$$p(y = 1|x) = \frac{1}{1 + \frac{p(y = -1|x)}{p(y = 1|x)}}$$

For ease of notation let,

Then

$$q(x) = p(y = 1|x)$$
$$1 - q(x) = p(y = -1|x)$$

That is,

$$p(y = 1|x) = \frac{1}{1 + \frac{1 - q(x)}{q(x)}}$$

The logistic model is defined as follows:

$$\ln\left(\frac{q(x)}{1 - q(x)}\right) = f(x; \alpha, \beta) = \alpha + \beta_1x_1 + \dots + \beta_7x_7$$

This implies that,

$$q(x; \alpha, \beta) = \frac{1}{1 + \exp(-f(x; \alpha, \beta))}$$

This establishes a relation between probability of successfully detection the distressed vessel and the explanatory variables. The next step is to estimate the parameters α and β .

It follows by application of the Bayes theorem that,

$$p(\alpha, \beta | \{(x_i, y_i)\}) \propto \prod_{i=1}^n q(x_i; \alpha, \beta)^{y_i} (1 - q(x_i; \alpha, \beta))^{1-y_i} \times p(\alpha, \beta)$$

Note that n is the number of runs and not the number of variables! The vector y has 7 components, i.e.,

the number of coded variables. The index i corresponds to the run number.

In the above equation the likelihood term is

$$L(y; x, \alpha, \beta) \propto \prod_{i=1}^n q(x_i; \alpha, \beta)^{y_i} (1 - q(x_i; \alpha, \beta))^{1-y_i}$$

and the last term is the prior. In the maximum likelihood approach, α and β assumes that the prior is uniform over the region of interest in the parameter space and tries to find that maximizes the likelihood function. This is equivalent to finding the maximum of the log-likelihood function. The log-likelihood is,

$$\ln[L(y; x, \alpha, \beta)] = Constant + \sum_{i=1}^n [y_i \ln[q(x_i; \alpha, \beta)] + (1 - y_i) \ln[1 - q(x_i; \alpha, \beta)]]$$

The maximization is done with respect to α and β . The software like Minitab numerically maximizes the above equation to find an estimate for α and β .