ELECTROMAGNETIC SUSCEPTIBILITY DATA ANALYSIS USING MULTIVARIATE LOGISTIC REGRESSION

Cynthia A Ropiak[1], David C Stoudt[2], and Robert L Gardner[3], Consultant

[1]Directed Energy Technologies Office,
Code B20, Naval Surface Warfare Center Dahlgren Division
17320 Dahlgren Road
Dahlgren, VA 22448-5100, USA
Email: ***RopiakCA@nswc.navy.mil***

[2]Joint Program Office for Special Technology Countermeasures
Code J07, Naval Surface Warfare Center Dahlgren Division
17320 Dahlgren Road
Dahlgren, VA 22448-5100, USA
E-mail: StoudtDC@nswc.navy.mil

[3]As (1) above, but E-mail: GardnerRL@nswc.navy.mil

ABSTRACT

Electromagnetic Susceptibility data is categorical in nature and therefore lends itself naturally to a logistic regression technique. Utilization of logistic regression allows for the functionalization of experiment parameters with susceptibility data, which in turn facilitates the visualization of the outcome of tests. However, fitting susceptibility data by way of logistic regression is an empirical method that if misused can lead to erroneous predictions. This paper outlines the logistic regression technique as it is applied to the susceptibility data analysis problem and demonstrates its utility as well as some of its limitations.

INTRODUCTION

For a number of electromagnetic interference (EMI) and high power microwave (HPM) problems determination of the conditions for upset of complex electronic systems is of interest. In particular, the identification of highly sensitive configurations or parameter sets is of interest. To accomplish this goal it is necessary to develop the tools required to predict the outcome of an illumination of the complex electronic system under test. Usually, these systems are much too complex to use deterministic analysis techniques so that the addition of information from empirical approaches is forced. Due to the categorical nature of susceptibility data a choice technique for the data analysis is multivariate logistic regression.

The utilization of multivariate logistic regression in the analysis of electromagnetic susceptibility data has a number of advantages. First and foremost, it facilitates the functionalization of experimental parameters with their associated effect. Once a functional fit has been obtained it is possible to interpolate within the parameter space spanned in the experiment. The interpolation capability allows for confident predictions of the outcome of experimental parameter combinations not tested, but bounded by the data set exercised in the experiment.

A second and arguably more important result associated with the existence of a logistic regression fit is that it enables cautious predictions outside of the experimental data set. These predictions may be used by Experimenters to guide them when selecting parameter combinations to try in future tests. For example, it is often of interest to determine the conditions for maximum upset of the device under test. The logistic regression fit can provide this information. However, due to the empirical nature of the logistic regression technique, if this maximum lies outside of the data set spanned in the experiment it may be necessary to verify this prediction experimentally. But it should be noted that without the logistic regression fit, the knowledge of where to target the measurements in the next test might not exist.

A particularly important consequence of having a functionalized fit through the use of logistic regression is that it provides a vehicle for visualizing the outcome of effects tests. Results from effects tests are often complex and convoluted making it extremely challenging to form useful conclusions from the experiments. In addition, a second obstacle presents itself when an attempt is made to communicate the findings from an effects test with the appropriate degree of importance. In other words, the implications of the outcome of effects tests are not always clear. The existence of a visualization tool allows for the communication of results that may otherwise have not been possible.

Although the application of multivariate logistic regression has tremendous potential, it also has associated with it a number of subtleties, which if not addressed can potentially lead to the misuse of the technique resulting in incorrect analysis results. The subtleties alluded to are associated with such things as data sampling, choice of fit parameters and the lack of identification of correlating variables possibly due to their latent nature.

This paper outlines the application of multivariate logistic regression to the problem of susceptibility data analysis. Interpolation and extrapolation of logistic regression fits are discussed as well as the visualization capabilities the technique provides. And finally, the issues associated with the improper use of the technique are addressed by way of example.

MULTIVARIATE LOGISTIC REGRESSION

An overview of the multivariate logistic regression technique proceeds as follows. We begin by defining the conditional mean:

$$\pi(\vec{x}) = \frac{e^{g(\vec{x})}}{1+e^{g(\vec{x})}} \tag{1}$$

In a susceptibility experiment, the conditional mean corresponds to the probability of effect. The g(x) term in the expression for the conditional mean is know as the logit:

$$g(\vec{x}) = \beta_0 + \beta_1 x_1 + \beta_1 x_1 + \ldots \tag{2}$$

The x_i terms in the logit function are the variables exercised in a given experiment, or some quantity derived from the experimental variables. The β_i's are the fit parameters which are found by maximizing the likelihood function:

$$l(\vec{\beta}) = \prod_{i=1}^{n} \pi(\vec{x})^{y_i} \left[1-\pi(\vec{x})\right]^{1-y_i} \tag{3}$$

The y_i's are the outcome variables, which may be 0 or 1 indicating either a system response (effect) or not.

GENERAL APPLICATION

We begin by considering the design and implementation of a susceptibility experiment. Fig. 1 represents a notional susceptibility experiment. For the purpose of demonstration a susceptibility data set will be fabricated using a problem that is solvable analytically. The academic problem is depicted in Fig.2.

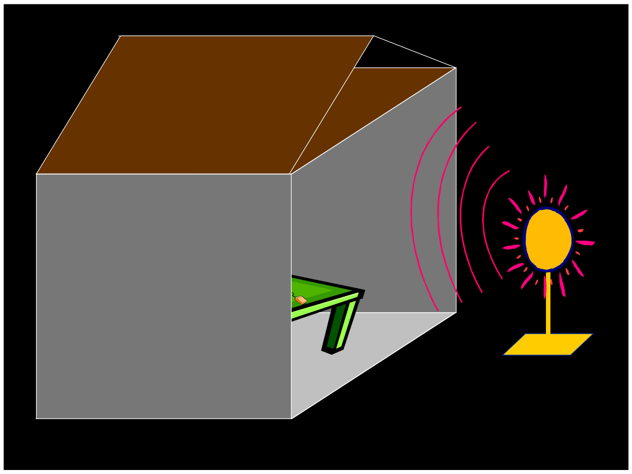

Fig. 1: Cartoon schematic of a possible susceptibility experiment on a computer

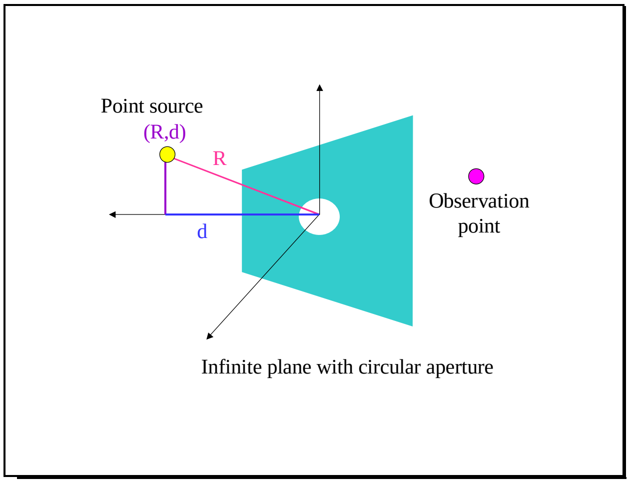

Fig. 2: Point source incident on an infinite plane with a circular aperture

The source location parameters ‘R’ and ‘d’ are varied and the electric field at the fixed observation point is calculated. Susceptibility criteria is established by defining an effects condition to occur if an electric field value exceeds some set value ‘x’. In order to emulate a measurement, random noise may be added to the electric field calculation. Sampling a Gaussian distribution normalized to 10% of the electric field threshold criteria generates the random noise. This technique may be used to generate as rich a data set as desired. The data set which consists of source parameter combinations (R,d) and corresponding effect scores of either 0 or 1 are then fed into the logistic regression model and a fit is performed. A plot of a logistic regression fit for a data set created in this manner is presented in Fig. 3.

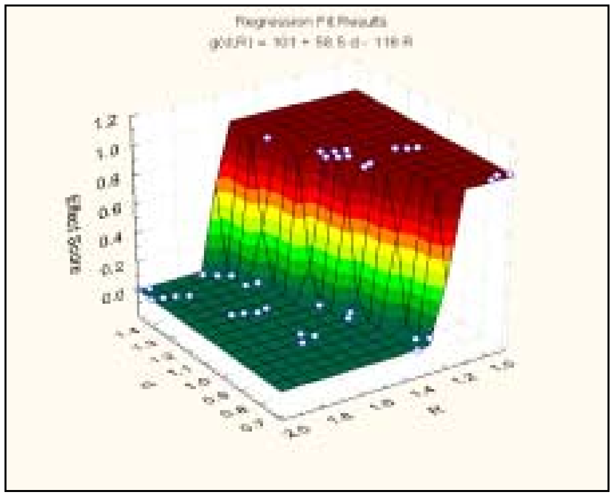

Fig. 3: Surface plot of logistic regression fit results for a fabricated susceptibility data set

Inspection of Fig. 3 reveals the true impact of having a functional fit in hand. It is now possible to predict the outcome of variable combinations not tested, but bounded by the data set. For example, the specific (R,d) combination of (1.5,1.15) was not tested, but the fit predicts an effect score of 0 (no effect) for this combination of source parameters. This is a very powerful result. In addition to interpolating within the data set, it is possible to carefully predict outside of the data set and use these predictions to guide future experiments.

EXPERIMENTAL APPLICATION

The example presented in the previous section served the purpose of illustrating the technique. While the application of logistic regression to the fabricated data set was straight forward, this will in general not be the case for experimental data sets. This section discusses some of the considerations that must be made when applying this method to a real data set.

The first step in the application of logistic regression to an experimental data set is to identify the input variables (the x_i’s) in the model. This may be accomplished by constructing the empirical logit, which is defined as:

$$Logit_{EMPIRICAL} = Log\left[\frac{P(x)}{1-P(x)}\right] \tag{4}$$

Where the ‘P’ in the empirical logit expression refers to the percent effect for the given x value. An assumption of the logistic regression model is that the logit is linear in the x_i variables. This assumption is referred to herein as the linearity assumption. A scatter plot of the empirical logit for the data will reflect whether or not the linearity assumption is valid for a particular x_i variable. The scatter plot may also facilitate the identification of the proper input variable x_i for the logistic regression model. This point is illustrated in Fig. 4 and Fig. 5. For the purpose of illustration, consider a sample data set containing a source parameter labeled ‘variable 1’. The ‘variable 1’ would be a parameter exercised during the experiment; source location for example. Fig. 4 is a scatter plot of the empirical logit versus variable 1 for our sample experiment. Note that the plot is not linear. However, if a plot of the empirical logit versus the square root of variable 1 is reasonably linear. This point is illustrated in Fig. 5. Hence, √(variable 1) is identified as the x_1 in the logistic regression model and the fitting process would proceed accordingly.

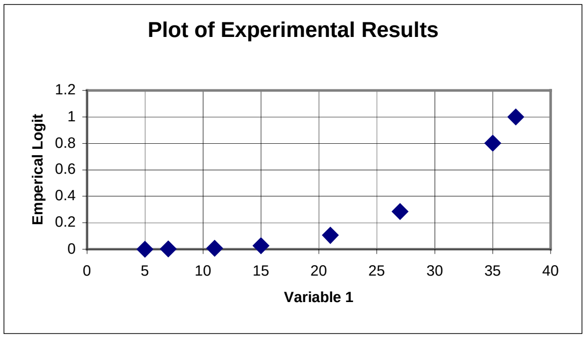

Fig. 4: Plot of the empirical logit versus experimental variable 1. Note the violation of the linearity assumption

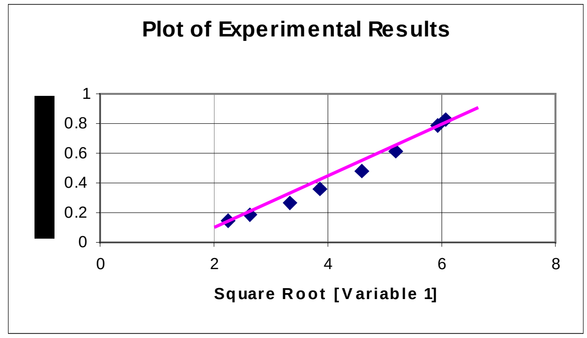

Fig. 5: Plot of the empirical logit versus the square root of the experimental variable 1. Note that the linearity assumption is valid for this plot

It should be noted that if the true physical dependence of the measured effects with variable 1 was in fact $y^{0.52}$ rather than $y^{.0.5}$ we would probably not be able to determine this. In other words, it is highly unlikely that a data set would be rich enough to bear out distinctions between functional dependencies that are so closely related. Therefore, it is necessary to use our knowledge of physics to guide us in the proper selection of the x_i variables. Without such knowledge, there is the potential for improper identification of variable dependencies that may prove to be significant for variable choices not exercised in the experiment. In other words, extrapolation of logistic regression fits outside of the data set spanned in a given experiment is dangerous due to subtleties such as the one just discussed.

One final point to note in the application of multivariate logistic regression concerns interaction terms. When fitting a data set, all interaction terms of fit variables x_i must also be considered. If the interaction terms are significant (correlate with effects) the fit statistics will reflect this information. However, if the logistic regression fit were not performed, interaction terms may not have been identified. It should be noted that the existence and identification of interaction terms may prove to be quite important.

POTENTIAL MISSUSE OF THE TECHNIQUE

It is imperative that this technique be applied with rigor if the results are to be believed with any degree of confidence. First and foremost is the need for a balanced test design. Although this technique may be applied to an imbalanced data set, the conclusions are not always as clear, and the potential for the lack of identification of an otherwise significant variable increases with increasingly imbalanced test designs.

A second and arguably more catastrophic consequence is associated with the violation of the linearity assumption. As discussed in the previous section, an assumption of the logistic regression model is that the x_i variables are linear with respect to the empirical logit. If this is not the case, the logistic regression model should not be used. However, the logistic regression model could be applied naively without checking the validity of the linearity assumption, which may in turn, result in a fit that does not appropriately reflect the functional dependence of the x_i variables with respect to the susceptibility data.

In general, errors associated with imbalanced test designs and the violation of the linearity assumption will manifest themselves in a poor fit which in turn should have associated fit statistics that reflect this fact. In other words, a poor fit could have a small chi-squared statistic, it may fail the Hosmer-Lemeshow test, it may reflect overly influential points and/or have large p-value's associated with the fit coefficients. The aforementioned tests are just a handful of criteria that must be investigated when assessing the quality of a logistic regression fit. If the fit statistics are sufficiently explored, the assessment of whether or not the technique has been successful should be straight forward. However, if care is not taken to explore all aspects of the fit, an invalid conclusion that the fit is representative of the data set, which in turn is representative of physical reality, may be reached.

CONCLUSIONS

Multivariate logistic regression provides a means for the functionalization of susceptibility data. If used with care, this can be of great use. It allows for the identification of the most significant variables exercised in an experiment as well as the identification of interaction terms. In addition, it provides a means for visualization of the outcome of an experiment. However, due to the fact that this is an empirical technique its application must proceed cautiously and even more important is that care is taken when attempting to infer information from the outcome of the regression fit.