Paper CD-02 – 2015

SAS® Software as an Essential Tool in Statistical Consulting and Research

Erika Larsen and Timothy E. O'Brien
Loyola University Chicago

Abstract:

Modelling in bioassay often uses linear, generalized linear (e.g., logistic or multi-category logit) and nonlinear regression models. As such, key matrices – and specialized software such as procedures in SAS/STAT and SAS/IML – are pervasive and wide-ranging since they are essential in point and interval estimation methods for the associated model parameters as well as for hypothesis testing and prediction.

Using key representative illustrations, this paper highlights the forms of some of these important matrices in the context of estimation, hypothesis testing, and optimal experimental design. After pointing out the inadequacy of these "optimal" experimental designs to detect lack-of-fit, we demonstrate how "robust" optimal designs are obtained by working with generalizations of the model. These latter designs are thus useful for both parameter estimation and checking for goodness-of-fit. Examples are provided using models from toxicology and pharmacology.

Keywords: Goodness-of-Fit; Logistic Regression; Multinomial Regression Models; Optimal Design; Robustness

I. Introduction:

Regression modelling is one of the most useful techniques in the applied sciences to determine relationships between attributes or variables. As such, matrices, which are extensively used and key in all aspects of estimation, hypothesis testing and prediction, are the unifying theme of these methods, and are the focus of this paper. To illustrate, a researcher may be interested in determining an association between the dose of a drug and a person's blood pressure. Then, depending upon distributional assumptions, linear, nonlinear or logistic regression methods may then be used to characterize this relationship.

Before a regression model can be fit, however, an experimental design (i.e., study plan) must be established and implemented, and the data must be collected. Indeed, each stage of this process – from choosing an efficient design, to recording the ensuing data, to data analysis (and perhaps prediction) – involves the use of crucial matrices, as is underscored below.

II. **Regression Modelling**:

Using Y to denote the outcome or dependent random variable, y to denote the realization of this random variable, and x to denote the independent variable (or $x_1, x_2 \ldots x_k$ in the case of several independent variables), we use the term "model" to comprise the following components:

(a) the assumed distribution for the response variable (Y) – usually chosen from the exponential family
(b) the link function connecting $\mu = E(Y)$ with the explanatory variable(s) and the model parameters (these parameters stacked in the vector $\boldsymbol{\theta}$)
(c) the [mean] model function $\eta(x, \boldsymbol{\theta})$, which joins the explanatory variable(s) and the model parameters
(d) the variance (denoted σ^2) or variance function (perhaps depending on $\boldsymbol{\theta}$ and/or additional parameters such as σ^2)
(e) the nature of the observations, such as independent or correlated (e.g. nested) measurements

Perhaps the most common form of regression modelling is simple linear regression, in which Y is assumed to have a normal distribution with constant variance and identity link, and with model function $\eta(x, \boldsymbol{\theta}) = \beta_0 + \beta_1 x$ so the model parameter vector here is $\boldsymbol{\theta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$. In this instance, the n independent realizations (observations) of Y are stacked into the vector $\boldsymbol{y}$ and the errors (i.e., deviations between the actual responses and those predicted by the line) are stacked into the vector $\boldsymbol{\varepsilon}$, so that $\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$ and $\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$. Also, with $n \times 2$ matrix $\boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$, the entire linear model system can be written as

$$\boldsymbol{y} = \boldsymbol{\eta}(x, \boldsymbol{\theta}) + \boldsymbol{\varepsilon} = \boldsymbol{X\theta} + \boldsymbol{\varepsilon} \tag{1}$$

Parameter estimation of the model parameters in $\boldsymbol{\theta}$, based on maximum likelihood estimation – which here is equivalent to minimizing the sum of squared errors – yields the estimate $\widehat{\boldsymbol{\theta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$. This estimate in turn satisfies the so-called normal equations,

$$\boldsymbol{X^T}\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}\right) = \boldsymbol{0} \qquad i.e.\ \boldsymbol{X^T X}\widehat{\boldsymbol{\theta}} = \boldsymbol{X^T y} \tag{2}$$

As a consequence, the maximum-likelihood estimate (MLE) – and also the least-squares estimate (LSE) – for the constant variance normal linear model is

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} \tag{3}$$

The Moore-Penrose inverse is used in (3) whenever $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$ is not invertible. Note that the variance of this MLE is the $p \times p$ matrix

$$\boldsymbol{var}\left(\widehat{\boldsymbol{\theta}}\right) = \sigma^2(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1} \tag{4}$$

This result is also relevant in the context of optimal designs discussed below since certain optimal design strategies attempt to minimize model parameter variances.

The simple linear regression model is easily extended to the multiple linear regression model which includes several independent variables ($x_1, x_2 \ldots x_k$). Even though the $\boldsymbol{X}$ matrix changes (i.e., in that it is then of dimension $n \times p$ with $p = k + 1$), the model structure in (1) and MLE/LSE in (3) remain unchanged.

One situation in which a multiple linear model is useful is in the case of response surface modelling (RSM). A quadratic RSM with two explanatory variables, x_1 and x_2, again posits normal responses with constant variance but with model function $\eta(x, \boldsymbol{\theta}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2$. It is then common practice to again stack the observations into vector form but to group the linear terms $(1, x_1, x_2)$ into $n \times 3$ matrix $\boldsymbol{X_1}$ and the quadratic terms $(x_1 x_2, x_1^2, x_2^2)$ into $n \times 3$ matrix $\boldsymbol{X_2}$. In this case, with $\boldsymbol{\theta_1} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$ and $\boldsymbol{\theta_2} = \begin{pmatrix} \beta_{12} \\ \beta_{11} \\ \beta_{22} \end{pmatrix}$, equation (1) becomes

$$\boldsymbol{y} = \boldsymbol{X\theta} + \boldsymbol{\varepsilon} = [\boldsymbol{X_1}|\boldsymbol{X_2}]\begin{pmatrix} \boldsymbol{\theta_1} \\ \boldsymbol{\theta_2} \end{pmatrix} + \boldsymbol{\varepsilon} = \boldsymbol{X_1\theta_1} + \boldsymbol{X_2\theta_2} + \boldsymbol{\varepsilon} \tag{5}$$

Mindful of the key role of $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$ (and its inverse), in this case we have

$$\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X_1^T X_1} & \boldsymbol{X_1^T X_2} \\ \boldsymbol{X_2^T X_1} & \boldsymbol{X_2^T X_2} \end{bmatrix} \tag{6}$$

Interestingly, even though this RSM involves quadratic terms (in the mathematical sense), it is called a linear model since the independent variables (i.e., those in $\boldsymbol{X}$) and the model parameters (in $\boldsymbol{\theta}$) in $\boldsymbol{\eta}(x, \boldsymbol{\theta})$ separate. More precisely, we say that the model function $\eta(x, \boldsymbol{\theta})$ is a linear model function if none of the partial derivative of $\eta(x, \boldsymbol{\theta})$ with respect to the model parameters involves any model parameter(s). In the case of a normal dependent variable with constant variances, a model function for which at least one of these partial derivatives involves at least one model parameter is called a nonlinear model function. Thus, constant-variance nonlinear models are identical to (1) but with stacked nonlinear model function vector $\boldsymbol{\eta}(x, \boldsymbol{\theta})$ substituted in place of $\boldsymbol{X\theta}$. Interestingly, this means that linear models are a special case of nonlinear ones. For nonlinear models, the counterpart of the $\boldsymbol{X}$ matrix above is the $n \times p$ Jacobian matrix,

denoted $\boldsymbol{V}$. The i^{th} row of $\boldsymbol{V}$ corresponds to the i^{th} observation (and x_i), and comprises $\left[\frac{\partial\eta(x_i,\boldsymbol{\theta})}{\partial\theta_1}, \frac{\partial\eta(x_i,\boldsymbol{\theta})}{\partial\theta_2} \ldots \frac{\partial\eta(x_i,\boldsymbol{\theta})}{\partial\theta_p}\right]$. By definition, $\boldsymbol{V}$ depends upon at least one component of $\boldsymbol{\theta}$, so we note that $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\theta})$. In this instance, the analogue of the normal equations in (2) is

$$\widehat{\boldsymbol{V}}^{\boldsymbol{T}}\left(\boldsymbol{y} - \boldsymbol{\eta}\left(\widehat{\boldsymbol{\theta}}\right)\right) = \boldsymbol{0} \qquad i.e.\ \widehat{\boldsymbol{V}}^{\boldsymbol{T}}\boldsymbol{\eta}\left(\widehat{\boldsymbol{\theta}}\right) = \widehat{\boldsymbol{V}}^{\boldsymbol{T}}\boldsymbol{y} \tag{7}$$

Here, $\widehat{\boldsymbol{V}} = \boldsymbol{V}\left(\widehat{\boldsymbol{\theta}}\right)$, and in general equation (7) is a nonlinear system of p equations in p unknowns – i.e., the p parameters in $\boldsymbol{\theta}$. To illustrate with a simplistic ($p = 1$) example, for $i = 1,2, \ldots n$ and the one-parameter decay model function $(x_i, \theta) = e^{-\theta x_i}$, we have $\frac{\partial\eta(x_i,\theta)}{\partial\theta} = -x_i e^{-\theta x_i}$. It follows that the normal equation (7) here is

$$\sum_{i=1}^{n} x_i y_i e^{-\widehat{\theta} x_i} = \sum_{i=1}^{n} x_i e^{-2\widehat{\theta} x_i} \tag{8}$$

Solution of the nonlinear equation (8) for $\widehat{\theta}$ – and of equation (7) for $\boldsymbol{\theta}$ in general – typically requires iterative methods such as the Modified Gauss-Newton method discussed in Seber & Wild (1989) and Bates & Watts (2007) and used by many software packages. Note that in general, the (asymptotic) theoretical variance of $\widehat{\boldsymbol{\theta}}$ is $\boldsymbol{var}\left(\widehat{\boldsymbol{\theta}}\right) = \sigma^2(\boldsymbol{V}^{\boldsymbol{T}}\boldsymbol{V})^{-1}$ and its estimate is

$$\widehat{\boldsymbol{var}}\left(\widehat{\boldsymbol{\theta}}\right) = \hat{\sigma}^2\left(\widehat{\boldsymbol{V}}^{\boldsymbol{T}}\widehat{\boldsymbol{V}}\right)^{-1} \tag{9}$$

The term 'asymptotic' is used here to emphasize the approximate nature of these variance terms and the fact that the difference between the approximation and exact values is in general lower with a larger sample size.

Our second illustration of a nonlinear model involves two-parameters and the Fieller-Creasy model function, which can be written $\eta(x, \boldsymbol{\theta}) = \theta_1 D_1 + \theta_1\theta_2 D_2$. This situation corresponds to one in which n_1 cases (e.g., individuals) receive one treatment with mean μ_1, n_2 cases receive a second treatment with mean μ_2 (both with the same variance), $\theta_1 = \mu_1$ and $\theta_2 = \frac{\mu_2}{\mu_1}$, and where D_1 and D_2 are indicator variables associated the two treatment groups. Thus, $D_1 = 1$ for subjects receiving treatment one, $D_1 = 0$ for subjects receiving treatment two, and $D_2 = 1 - D_1$. Another way to write this model function is $\eta(x, \boldsymbol{\theta}) = \theta_1$ for treatment-one subjects (i.e., for $i = 1,2, \ldots n_1$) and $\eta(x, \boldsymbol{\theta}) = \theta_1\theta_2$ for treatment-two subjects (i.e., for $i = n_1 + 1, n_1 + 2, \ldots n_1 + n_2 = n$). The key parameter here is θ_2 since it corresponds to the ratio of the treatment means and thus provides the means to test for equality of the means by testing whether $\theta_2 = 1$.

Demonstrating that the Fieller-Creasy model function is nonlinear is established by examining the partial derivatives and noting that at least one of these involves a model

parameter: for subjects receiving treatment one, $\frac{\partial \eta(x_i,\boldsymbol{\theta})}{\partial \theta_1} = 1$ and $\frac{\partial \eta(x_i,\boldsymbol{\theta})}{\partial \theta_2} = 0$, and for subjects receiving treatment two, $\frac{\partial \eta(x_i,\boldsymbol{\theta})}{\partial \theta_1} = \theta_2$ and $\frac{\partial \eta(x_i,\boldsymbol{\theta})}{\partial \theta_2} = \theta_1$. The corresponding Jacobian matrix is then

$$\boldsymbol{V} = \begin{bmatrix} \mathbf{1}_{\boldsymbol{n_1}} & \mathbf{0}_{\boldsymbol{n_1}} \\ \theta_2 \mathbf{1}_{\boldsymbol{n_2}} & \theta_1 \mathbf{1}_{\boldsymbol{n_2}} \end{bmatrix}$$

Here, $\mathbf{1}_{\boldsymbol{n_1}}$ and $\mathbf{1}_{\boldsymbol{n_2}}$ are vectors of one's of lengths n_1 and n_2 respectively, and $\mathbf{0}_{\boldsymbol{n_1}}$ is a vector of zero's of length n_1. Using basic matrix results, we obtain

$$\begin{aligned} \boldsymbol{V}^T\boldsymbol{V} &= \begin{bmatrix} n_1 + n_2\theta_2^2 & n_2\theta_1\theta_2 \\ n_2\theta_1\theta_2 & n_2\theta_1^2 \end{bmatrix}, \\ (\boldsymbol{V}^T\boldsymbol{V})^{-1} &= \frac{1}{n_1 n_2 \theta_1^2} \begin{bmatrix} n_2\theta_1^2 & -n_2\theta_1\theta_2 \\ -n_2\theta_1\theta_2 & n_1 + n_2\theta_2^2 \end{bmatrix} \end{aligned} \tag{10}$$

To demonstrate the application of (9), from (10) we see that the estimated variance associated with $\hat{\theta}_1 = \bar{y}_1$ is $\hat{\sigma}^2 \frac{n_2\hat{\theta}_1^2}{n_1 n_2 \hat{\theta}_1^2} = \frac{\hat{\sigma}^2}{n_1}$, and the estimated variance of $\hat{\theta}_2 = \frac{\bar{y}_2}{\bar{y}_1}$ is $\hat{\sigma}^2 \frac{n_1 + n_2\hat{\theta}_2^2}{n_1 n_2 \hat{\theta}_1^2}$.

Another example of a nonlinear model involves normal, constant-variance responses and either the two-parameter log-logistic ($LL2$) or Weibull ($WEIB2$) model functions, written respectively as

$$\eta_{LL2}(x, \boldsymbol{\theta}) = \frac{1}{1 + (x/\theta_1)^{\theta_2}} \qquad \eta_{WEIB2}(x, \boldsymbol{\theta}) = e^{-(x/\theta_1)^{\theta_2}} \tag{11}$$

The graphs of these two model functions have the usual down-sloping sigmoidal shape typical of bioassay and toxicological data. In both cases $\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$ where θ_1 is a location parameter and θ_2 controls the slope. Interestingly, since both of these model functions are similar, their fits to biological data are very close, thereby providing two 2-parameter model functions which are rivals. In sections IV and V, we point out that chosen experimental designs should provide information to estimate the model parameters, but to do so even if a rival model function is more appropriate for the mechanism under study.

In addition to simple/multiple linear models and nonlinear models, another popular model used to represent bioassay and toxicology data is the binary/binomial logistic model. Whereas the above (normal, constant variance) linear and nonlinear models use the identity link function, logistic modelling (of the common variety discussed here) uses the logit-link function, $g(z) = \log\left(\frac{z}{1-z}\right)$. Nonetheless, in analogous manner to normal-

theory linear and nonlinear models, estimation and experimental design methods for the logistic model are based on the corresponding likelihood. In the case of the constant-variance normal-theory linear and nonlinear models, by (1) the log-likelihood LL is inversely related to the sum of squares, $S(\boldsymbol{\theta}) = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = \sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{n}\big(y_i - \eta(x_i,\boldsymbol{\theta})\big)^2$. Also, setting the first derivatives of LL (the so-called score vector function) to zero gives the normal equations in (2) and (7); also, the second derivative of LL (matrix) leads to the information matrix and to its inverse (and variance-covariance matrix for $\widehat{\boldsymbol{\theta}}$) given in (4) and (9).

For the binary/binomial logistic model, the x variable often can correspond to dose or concentration and the researcher selects k such dose points, viz, $x_1, x_2, \ldots x_k$ to run the experiment; this selection is indeed part of the experimental design problem discussed below. This model assumes that independently n_i subjects (experimental units) receive dose x_i, and that the number of "successes" y_i has a binomial distribution with success probability

$$\pi_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \quad i.e.\ log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i \tag{12}$$

Extensions of this expression to multiple independent variables are indeed straightforward.

As for the above simple linear regression model, the model parameters in (12) are $\boldsymbol{\theta} = \begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}$; due to the binomial and independence assumptions here, after dropping a constant it follows that the corresponding log-likelihood expression is

$$\begin{aligned} LL(\boldsymbol{\theta}) &= \sum_{i=1}^{k}\left[y_i log\frac{\pi_i}{1-\pi_i} + n_i log(1-\pi_i)\right] \\ &= \sum_{i=1}^{k}\left\{y_i(\beta_0+\beta_1 x_i) - n_i log(1+e^{\beta_0+\beta_1 x_i})\right\} \end{aligned} \tag{13}$$

In this case, letting $\boldsymbol{E}(\boldsymbol{Y})$ represent the expected value vector $\begin{pmatrix}E(Y_1)\\ E(Y_2)\\ \vdots\\ E(Y_k)\end{pmatrix} = \begin{pmatrix}n_1\pi_1\\ n_2\pi_2\\ \vdots\\ n_k\pi_k\end{pmatrix}$, differentiating (13) with respect to $\boldsymbol{\theta}$ and setting to zero gives the normal equations,

$$\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{E}(\boldsymbol{Y})) = 0 \quad \text{i.e.,} \quad \begin{cases}\sum_{i=1}^{k}(y_i - n_i\pi_i) = 0\\ \sum_{i=1}^{k} x_i(y_i - n_i\pi_i) = 0\end{cases} \tag{14}$$

As for nonlinear models, since the model parameters enter these equations in a nonlinear manner, iterative methods such as Modified Gauss-Newton method are

usually used to obtain the parameter estimates. Differentiating a second time yields the (Fisher) information matrix:

$$\boldsymbol{M}(\boldsymbol{\theta}) = -E\left[\frac{\partial^2 LL}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right] = \begin{bmatrix} \sum_{i=1}^{k} n_i\pi_i(1-\pi_i) & \sum_{i=1}^{k} n_i x_i\pi_i(1-\pi_i) \\ \sum_{i=1}^{k} n_i x_i\pi_i(1-\pi_i) & \sum_{i=1}^{k} n_i x_i^2\pi_i(1-\pi_i) \end{bmatrix} \quad (15)$$

For this logistic model, $\boldsymbol{M}(\boldsymbol{\theta}) = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$ with $\boldsymbol{W} = diag\{n_1\pi_1(1-\pi_1), n_2\pi_2(1-\pi_2), \dots, n_k\pi_k(1-\pi_k)\}$, and the dependence of $\boldsymbol{W}$ upon $\boldsymbol{\theta}$ is apparent since the π_i depend upon $\boldsymbol{\theta}$. In similar spirit to (4) and (9) and with $\widehat{\boldsymbol{W}} = \boldsymbol{W}(\widehat{\boldsymbol{\theta}})$, we have for this binomial logistic model the estimated (asymptotic) variance-covariance matrix

$$\widehat{\boldsymbol{var}}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{M^{-1}}(\widehat{\boldsymbol{\theta}}) = \left(\boldsymbol{X^T}\widehat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1} \quad (16)$$

The previous examples of simple and multiple linear models, nonlinear models, and binomial logistic model demonstrate several commonalities – yet with some subtle nuances. In each case, the likelihood is used to generate score functions and information matrices then used to give model parameter estimates and associated variances. Model parameter estimates, obtained from so-called normal equations (equating the score functions to zero), involve iterative methods for all but linear models; similarly the associated variance-covariance matrices are estimated using the model parameter estimates for all but linear models, and are thus deemed approximate.

Since most applied researchers wish to efficiently estimate parameters (including producing reliable confidence intervals for these) as well as provide some basis for the assessment of the quality of fit, we next discuss these in the context of modelling.

III. **Super-modelling and Goodness-of-Fit**:

In many instances, it is wise to connect distinct model functions. One approach to doing this is to find a generalized "super-model" that contains the original model(s) as special cases. To illustrate, in the context of Response Surface Models (RSM) with two explanatory variables discussed in the previous section, consider the following two rival model functions: $\eta_1(x, \boldsymbol{\theta}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$ and $\eta_2(x, \boldsymbol{\theta}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2$. Clearly, both of these model functions are nested in (i.e., special cases of) the super-model $\eta_{SM}(x, \boldsymbol{\theta}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2$ since η_1 results from η_{SM} when $\beta_{11} = \beta_{12} = 0$ and η_2 results from η_{SM} when $\beta_{12} = 0$.

Our focus here is on the following: when a researcher has a model function in mind to describe a certain phenomenon, it is useful to be able to fit this model function (by estimating the model parameters) and to also test for inadequacies of this assumed model function. As such, it is useful to think of the assumed model function as a special

case of a larger super-model. Thus, when a researcher feels that η_1 is the correct model function, it is useful to consider the more general η_{SM} model function, which in turn contains other useful rival model functions (such as η_2) as special cases. Clearly this process of choosing η_{SM} is not unique, so we focus on wise choices of η_{SM}, which in turn contain other useful and important rival model functions as special cases.

Generalizing linear models is generally straightforward: one can simply add higher-order terms onto the original model function. Although this process is markedly more difficult for nonlinear model functions, some important results can be given. In the previous section, it was noted that the two-parameter log-logistic (η_{LL2}) and Weibull (η_{WEIB2}) model functions provide commonly-observed sigmoidal fits to bioassay and toxicology data. A super-model which generalizes these model functions is the three-parameter Eclectic ($EC3$) model function

$$\eta_{EC3}(x, \boldsymbol{\theta}) = \frac{1}{\left(1 + \frac{(x/\theta_1)^{\theta_2}}{\theta_3}\right)^{\theta_3}} \tag{17}$$

It is important to note that the $LL2$ model function is obtained when $\theta_3 = 1$, and the $WEIB2$ model function results for $\theta_3 \to \infty$, thereby demonstrating that $EC3$ is a super-model (generalization) of both the $LL2$ and $WEIB2$ model functions. Other important generalizations of these model functions are given in O'Brien (1994); for our present purposes we focus on the $EC3$ model function in (17) to demonstrate how obtaining key super-models can be used in the context of obtaining an optimal design.

IV. **Optimal Design Theory**:

An n-point design, denoted ξ, is written

$$\xi = \begin{Bmatrix} x_1 & x_2 & \dots & x_n \\ \omega_1 & \omega_2 & \dots & \omega_n \end{Bmatrix} \tag{18}$$

The ω_i are non-negative design weights which sum to one; the x_i are design points that belong to the design space and are not necessarily distinct. For the constant-variance normal setting with linear or nonlinear model function $\eta(x, \boldsymbol{\theta})$, the $n \times p$ Jacobian matrix is $\boldsymbol{V} = \frac{\partial \eta}{\partial \theta}$ with $\boldsymbol{\Omega} = diag\{\omega_1, \omega_2, \dots, \omega_n\}$, the $p \times p$ (Fisher) information matrix is

$$\boldsymbol{M}(\xi, \boldsymbol{\theta}) = \boldsymbol{V^T \Omega V} \tag{19}$$

In the spirit of (15), this result, as well as the general case of either non-constant variance or non-normality, follows from the more general expression for the information matrix:

$$\boldsymbol{M}(\xi, \boldsymbol{\theta}) = -E\left(\frac{\partial^2 LL}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right) \tag{20}$$

For the normal and binomial models discussed here, in noting (4), (9) and (16), the (asymptotic) variance of $\widehat{\boldsymbol{\theta}}_{MLE}$ is proportional to $\boldsymbol{M}^{-1}(\xi, \boldsymbol{\theta})$, so designs are often chosen to minimize some (convex) function of $\boldsymbol{M}^{-1}(\xi, \boldsymbol{\theta})$. To illustrate, designs which minimize its determinant are called D-optimal. Since for nonlinear and logistic models, $\boldsymbol{M}$ depends upon $\boldsymbol{\theta}$, local or Bayesian designs can be obtained.

Turning from parameter estimation to prediction, the (first-order) variance of the predicted response at the value x is

$$d(x, \xi, \boldsymbol{\theta}) = \frac{\partial \eta(x, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \boldsymbol{M}^{-1}(\xi) \frac{\partial \eta(x, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = tr\{\boldsymbol{M}^{-1}(\xi)\, \boldsymbol{M}(x)\} \tag{21}$$

Designs that minimize (over ξ) the maximum (over x) of $d(x, \xi, \boldsymbol{\theta})$ in (21) are called G-optimal. As noted above, since this predicted variance depends upon $\boldsymbol{\theta}$ for logistic and nonlinear models, researchers often seek optimal designs using a "best guess" for $\boldsymbol{\theta}$ (called a local optimal design) or assuming a plausible prior distribution on $\boldsymbol{\theta}$ (called a Bayesian optimal design).

The General Equivalence Theorem (GET) of Kiefer and Wolfowitz (1960) proves that D- and G-optimal designs are equivalent. They also showed that the variance function evaluated using the D-/G-optimal design does not exceed the line $y = p$ (where p is the number of model function parameters) – but that it will exceed this line for all other designs. A corollary establishes that the maximum of the variance function is achieved for the D-/G-optimal design at the support points of this design; this result is useful and essential for establishing optimality of a given design.

A simple example involves the constant-variance, normal simple linear regression model where x is constrained to lie in the interval $\mathfrak{X} = [0,1]$. The GET and graph of the associated variance functions demonstrate that the design $\xi_1 = \begin{Bmatrix} 0 & 1 \\ 1/2 & 1/2 \end{Bmatrix}$ is D/G-optimal in this setting. This translates into recommending that half of the observations be taken at the lowest value of $\mathfrak{X}$ ($x = 0$) and the other half at the highest value of $\mathfrak{X}$ ($x = 1$); optimality notwithstanding, researchers would be apprehensive to use the design ξ_1 in practice.

The above illustration demonstrates a practical characteristic: in most practical situations, optimal designs for p-parameter model functions comprise only p support points, thereby providing no ability to test for lack of fit of the assumed model. Researchers therefore often desire near-optimal "robust" designs that have extra support points which can then be used to test for model adequacy. To illustrate, for the

above linear regression example, notice that the design $\xi_2 = \begin{Bmatrix} 0 & 1/2 & 1 \\ 1/3 & 1/3 & 1/3 \end{Bmatrix}$ is not strictly "optimal", but may be near enough to optimal (in the sense defined in the next section) and can also be used to test for lack of fit of the assumed linear model function – at least "in the direction of" a quadratic model function.

V. **Robust Near-Optimal Design:**

As noted in sections II and III the context of Response Surface Models – and also in linear models in general – the assumed model function $\boldsymbol{\eta_1 = X_1\theta_1}$ can be embedded into the larger super-model, $\boldsymbol{\eta_{SM} = X_1\theta_1 + X_2\theta_2 = [X_1|X_2]\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} = X\theta}$. Here, $\boldsymbol{\theta_1}$ is $p_1 \times 1$, $\boldsymbol{\theta_2}$ is $p_2 \times 1$, and $p_1 + p_2 = p$. Researchers commonly encounter the situation in which it is felt that η_1 is the true model function, but a design is sought to efficiently estimate $\boldsymbol{\theta_1}$ (the model parameters of η_1) and to also provide some information to detect lack of fit of η_1 in the direction of η_{SM}. This is achieved by first noting that $\boldsymbol{X_1^T X_1}$ is the (Fisher) Information matrix associated with $\boldsymbol{\theta_1}$ in the linear model with model function η_1, and one numerical measure of information is therefore the determinant $\left|\boldsymbol{X_1^T X_1}\right|$. Also, by (6), note that the information associated with the full vector $\boldsymbol{\theta} = \begin{pmatrix}\boldsymbol{\theta_1}\\ \boldsymbol{\theta_2}\end{pmatrix}$ in η_{SM} is $\boldsymbol{X^T X}$, and by using rules of determinants for partitioned matrices, its determinant is

$$\begin{aligned} |\boldsymbol{X^T X}| &= \left|\boldsymbol{X_1^T X_1}\right| \left|\boldsymbol{X_2^T X_2 - X_2^T X_1 \left(X_1^T X_1\right)^{-1} X_1^T X_2}\right| \\ &= \left|\boldsymbol{X_1^T X_1}\right| \left|\boldsymbol{X_2^T (I - P_{X_1}) X_2}\right| \end{aligned} \tag{22}$$

Here, $P_{X_1} = \boldsymbol{X_1\left(X_1^T X_1\right)^{-1} X_1^T}$ is the projection matrix/operator onto the column space of $\boldsymbol{X_1}$, so $\boldsymbol{(I - P_{X_1})X_2}$ corresponds to the column space of the projection of $\boldsymbol{X_2}$ orthogonal to the column space of $\boldsymbol{X_1}$. As such, (22) demonstrates that the information regarding the full vector $\boldsymbol{\theta}$ in η_{SM} (viz, $|\boldsymbol{X^T X}|$) is partitioned into the product of the information regarding $\boldsymbol{\theta_1}$ in the sub-model η_1 (i.e., $\left|\boldsymbol{X_1^T X_1}\right|$) multiplied by the additional information in $\boldsymbol{\theta_2}$ not in $\boldsymbol{\theta_1}$ in the larger model η_{SM} (i.e., $\left|\boldsymbol{X_2^T (I - P_{X_1}) X_2}\right|$).

We easily extend these matrix results to $\boldsymbol{M = M(\xi, \theta)}$ in (19) and (20) by noting that for

$$\boldsymbol{M} = \begin{bmatrix}\boldsymbol{M_{11}} & \boldsymbol{M_{12}} \\ \boldsymbol{M_{21}} & \boldsymbol{M_{12}}\end{bmatrix}, \tag{23}$$

we have

$$|\boldsymbol{M}| = |\boldsymbol{M_{11}}|\left|\boldsymbol{M_{22} - M_{21} M_{11}^{-1} M_{12}}\right| \tag{24}$$

Similar to the interpretation above, (24) shows that the information regarding the full vector $\boldsymbol{\theta}$ in η_{SM} (viz, $|\boldsymbol{M}|$) is partitioned into the product of the information regarding $\boldsymbol{\theta}_1$ in the sub-model η_1 (i.e., $|\boldsymbol{M_{11}}|$) multiplied by the additional information in $\boldsymbol{\theta_2}$ not in $\boldsymbol{\theta_1}$ in the larger model η_{SM} (i.e., $\left|\boldsymbol{M_{22}} - \boldsymbol{M_{21}M_{11}^{-1}M_{12}}\right|$).

These results motivate the following compound design criterion function (objective function):

$$\boldsymbol{\Phi_k(\xi, \theta)} = \frac{\boldsymbol{\kappa}}{\boldsymbol{p_1}} \boldsymbol{log}|\boldsymbol{M_{11}}| + \frac{\boldsymbol{1-\kappa}}{\boldsymbol{p_2}} \boldsymbol{log}\left|\boldsymbol{M_{22}} - \boldsymbol{M_{21}M_{11}^{-1}M_{12}}\right| \qquad (25)$$

In this expression, the sub-model contains p_1 parameters, the super-model contains p_2 additional parameters; also, κ, which lies between 0 and 1, controls the emphasis placed on the original versus the additional parameters. As noted, $|\boldsymbol{M_{11}}|$ measures the information regarding the original p_1 model parameters $\boldsymbol{\theta_1}$ and $\left|\boldsymbol{M_{22}} - \boldsymbol{M_{21}M_{11}^{-1}M_{12}}\right|$ captures the information regarding the additional p_2 model parameters $\boldsymbol{\theta_2}$ not in $\boldsymbol{\theta_1}$. Thus, for $\kappa = 1$, the criterion yields D-optimal designs for $\boldsymbol{\theta_1}$ in the sub-model η_1, and for $\kappa = p_1/p$, it gives D-optimal designs for full vector $\boldsymbol{\theta}$ in η_{SM}. Again by the General Equivalence Theorem, D-optimality is confirmed by plotting the corresponding variance function plot and noting whether the graph exceeds the relevant horizontal line.

To illustrate, consider seeking an efficient local design for the two-parameter log-logistic ($LL2$) model function in (11) with $\theta_2 = 50$ and $\theta_3 = 1$. We also allow for the possibility that perhaps the two-parameter Weibull ($WEIB2$) model function fits. As such, we nest the $LL2$ model in the three-parameter Eclectic ($EC3$) model function in (17), and find an efficient design. Using the notation above, $p_1 = 2$ and $p = 3$. In this setting, the local D-optimal design for the $LL2$ model function (denoted ξ_D) puts the weights of $\omega = ½$ at each of the points $x = 17.62$ and $x = 141.99$, but provides no means to test for lack-of-fit. Using the nesting design criterion with $\kappa = 0.94$ in (25), the optimal design (denoted ξ_R) assigns the respective weights $\omega = 0.43, 0.24, 0.33$ to the points $x = 14.87, 70.56, 211.43$. Since this latter design has an extra support point, it can be used to test for lack of fit – more precisely, lack of fit of the $LL2$ model function in the direction of the $EC3$ model function (including the $WEIB2$ model function as another special case).

We next need a measure of the degree of 'closeness' of two designs, and one such measure is the so-called D-efficiency discussed in O'Brien & Funk (2003) and Atkinson et al (2007). Applied to the current setting, this is

$$\left(\frac{|\boldsymbol{M_{11}(\xi_R)}|}{|\boldsymbol{M_{11}(\xi_D)}|}\right)^{\boldsymbol{1/2}} = \boldsymbol{0.9456} \qquad (26)$$

With a D-efficiency of 94.56%, using the robust near-optimal design $\boldsymbol{\xi_R}$ results in a modest efficiency sacrifice of only 5.44% yet an extra support point to check for goodness of fit.

VI. **Summary**:

Researchers working in statistical consulting – especially in toxicology and bioassay – often assess dependencies between variables in their work by fitting linear, nonlinear or logistic models to their data. As such, efficient experimental designs which can be used to estimate model parameters as well as to check for goodness-of-fit of the assumed model. As demonstrated here, these methods rely on key special matrices and the method of imbedding the assumed model function into a larger class which contains important special cases is a useful and imperative strategy. As evidenced by these illustrations, the key software tool in our research and consulting work are the suite of procedures in those in SAS/STAT and SAS/IML.

VII. **Acknowledgements**:

The first author acknowledges the generous support of Loyola University Chicago Graduate School in funding her 2015-16 MS studies in Applied Statistics. The second author expresses his appreciation to the J. William Fulbright Foreign Scholarship Board for ongoing grant support; and to Vietnam National University in Hanoi, Vietnam, Kathmandu University in Dhulikhel, Nepal and Gadjah Mada University and Islamic University of Indonesia in Yogyakarta, Indonesia for kind hospitality and assistance while this work was carried out.

VIII. **References**:

- Atkinson, A.C., Donev, A. N. & Tobias, R.D., 2007, *Optimum Experimental Designs, with SAS*, Oxford: New York.

- Bates, D.M. & Watts, D.G., 2007, *Nonlinear Regression Analysis and its Applications*, Wiley: New York.

- Kiefer, J. & Wolfowitz, J., 1960, The Equivalence of Two Extremum Problems, *Canad. J. Math.*, 12, 363-366.

- O'Brien, T.E. 1994, A New Robust Design Strategy for Sigmoidal Models Based on Model Nesting. In Dutter, R. and Grossmann W., eds., *CompStat 1994*, Heidelberg: Physica-Verlag, 97-102.

- O'Brien, T.E. & Funk, G.M. 2003, A Gentle Introduction to Optimal Design for Regression Models, *Amer. Statist.*, 57, 265-267.

- Seber, G.A.F. & Wild, C.J., 1989, *Nonlinear Regression*, Wiley: New York.

CONTACT INFORMATION

Contact the authors at:

Ms. Erika Larsen
Professor T.E. O'Brien
Department of Mathematics and Statistics
Loyola University Chicago
6525 N. Sheridan Road
Chicago, IL 60626 USA
Work Phone: 1-773-508-2129
Fax: 1-773-508-2123
Email: tobrie1@luc.edu
Webpage: http://webpages.math.luc.edu/~tobrien/home.html