Chapter 2

DESIGN OPTIMALITY

§2.1 Introduction

Much of the design optimality criteria in response surface methodology has centered around the linear model with homogenous variance. This has led to wide use of design optimality in industrial settings where linear models are appropriate (Myers and Montgomery, 1995). Until recent years, the concept had not been pursued with nonlinear and non-homogeneous variance models. Since the design optimality criteria for these GLM models are most often functions of the unknown parameters, one must have some knowledge of them in order to determine the optimal design. Many of the models used in biological applications are of this form. Hence, the concepts of design optimality have experienced little use in the biological realm. This trend is changing, though. With the increasing use of Bayesian designs and preliminary studies for the purpose of process evaluation, the use of optimal designs for GLM models is gaining popularity.

In order to apply these design optimality criteria to GLM models, one must first understand the basic tenets of this aspect of response surface methodology. The design criteria are often called alphabetic optimality criteria because they are identified by letters. Each one of these alphabetic criteria has an experimental goal associated with it that achieves a specific property for the final fitted regression model. Design optimality for regression models was introduced by Kiefer and Wolfowitz (1959) with the D and E design criteria. Their research led to the development of other criteria such as A, F, G, and Q optimality. This dissertation concentrates on Bayesian applications of D and F criteria for single regressor impairment models. The latter half of this work will employ the D, D_s, and interaction optimality criteria for the multiple regressor impairment model. In this chapter, D and F optimality will be addressed. D_s and interaction optimality will be presented after the introduction of the *k*-regressor model. In general, the models considered are of the form

$$y_{ij} = f(\mathbf{x}_i, \boldsymbol{\beta}) + \boldsymbol{\varepsilon}_{ij}. \qquad (2.1.1)$$

where i=1, . . ., d and d is the number of design points, j=1, . . . , n_i is the number of replicates at each design point, $f(\mathbf{x}_i, \boldsymbol{\beta})$ is a function of known form (linear or nonlinear in **b**) and **b** is a $p \times 1$ vector of unknown parameters, $\mathbf{x}_i$ is a $k \times 1$ vector of regressors, and $\boldsymbol{\varepsilon}_{ij}$ is an unknown error with an assumed distribution.

Two regions must be defined in $\Re^k$ space. These are the region of operability and the region of interest. The region of operability is defined by the feasible ranges for the k independent variables in the process. It is often defined on the basis of the ability of the process to function at certain settings. This region is also known as the design space. The region of interest is the area of $\Re^k$ in which the researcher is primarily concerned with the response. Generally, it is the region in which the practitioner wants to obtain the most accurate and precise predictions of the response. These regions are often the same. However, in some cases, they may differ. For the optimality criteria discussed in this dissertation, only one of the regions must be specified.

§2.2 D-Optimality

One of the most widely known of the alphabetic criteria is D-optimality. D-Optimality concentrates on estimating **b** as well as possible. This is done by minimizing the generalized variance of the estimator of **b**. Consequently, the volume of the (1-α)100% confidence ellipsoid around **b** is minimized (Myers and Montgomery, 1995).

A common technique for estimating **b** is the method of maximum likelihood. Define **b** as the MLE of **b.** The asymptotic variance-covariance matrix of **b** is the inverse of the Fisher information matrix given by

$$\mathbf{I}(\mathbf{X},\boldsymbol{\beta}) = -\mathrm{E}\left[\frac{\partial^2 \ln \mathrm{L}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right] \tag{2.2.1}$$

where $\mathrm{L}(\boldsymbol{\beta})$ is the joint likelihood function for the data (Lehmann, 1983). In non-linear or non-homogenous variance models, the derivatives that comprise the information matrix are usually dependent on the values of the unknown parameters. Since nearly all design criteria are a function of the information matrix, one must have knowledge of these parameters in order to formulate the optimal design. The D criterion depends on the information matrix through the determinant. Minimizing the generalized variance of **b** is equivalent to maximizing the determinant of the information matrix. Hence, the general D-optimality criterion is

$$\max_{\mathscr{D}}\left|\frac{\mathbf{I}(\mathbf{X},\boldsymbol{\beta})}{\mathrm{N}}\right| \tag{2.2.2}$$

where $\mathscr{D}$ indicates all possible designs over the region of operability and N is the number of available experimental units.

§2.3 The Two Level D-Optimal Design for the Single Regressor Exponential Model

In order to promote a better understanding of optimal designs in the impaired reproduction context, the derivation of Chiacchierini's (1996) optimal design for the single regressor Poisson exponential model is shown here. Recall the model from Chapter 1,

$$y_{ij} = e^{\beta_0+\beta_1 x_i} + \varepsilon_{ij}. \tag{2.3.1}$$

The information matrix for this model is

$$\mathbf{I}(\mathbf{X},\beta) = \begin{bmatrix} \Sigma\lambda_i & \Sigma\lambda_i x_i \\ \Sigma\lambda_i x_i & \Sigma\lambda_i x_i^2 \end{bmatrix}. \tag{2.3.2}$$

The determinant of the information matrix for this model can be written as follows by making the substitution $x_i = \dfrac{\lambda_i - \beta_0}{\beta_1}$:

$$|\mathbf{I}| = \left(\frac{1}{\beta_1^2}\right) n_1 n_2 \lambda_1 \lambda_2 \left(\ln\frac{\lambda_1}{\lambda_2}\right)^2. \tag{2.3.3}$$

This expression must be maximized in order to find the optimal design. This process is complicated because, like many of its GLM counterparts, this design criterion is dependent on its parameters. The dependency results from the term $\dfrac{1}{\beta_1^2}$ as well as the β_0 and β_1 "hidden" in the expected values, λ_1 and λ_2. This problem can be remedied by substituting $\lambda_1 = q_1\lambda_c$ and $\lambda_2 = \lambda_c$. The result is shown in equation (2.3.4),

$$|\mathbf{I}| = \left(\frac{1}{\beta_1^2}\right) n_1 n_2 q_1 \lambda_c^2 (\ln q_1)^2. \tag{2.3.4}$$

The design can then be found in terms of the effective concentrations or the EC_{100q} so no knowledge of parameters is needed.

To begin the actual maximization process, some extraneous factors can be eliminated from (2.3.4). It is obvious that $n_1 = n_2 = n$ maximizes the function with respect to the allocation of experimental units to design points. Since the quantities $\frac{1}{\beta_1^2}$, λ_c, and n are constants with respect to maximization they can be removed. Taking the derivative of (2.3.4) and setting it equal to zero yields solutions of q_1=0 and q_1=e^{-2}=0.1353. The solution of q_1=0 is invalid since it makes $\left|\mathbf{I}_E\right| = 0$. So, maximization of the determinant occurs when $q_1 = e^{-2} = 0.1353$. Thus, the optimal design places 50% of the observations at the $EC_{13.53}$ and the remaining 50% at the EC_{100}, or the control. Figure 2.3.1 gives the reader more insight about the placement of design points in relationship to the model.

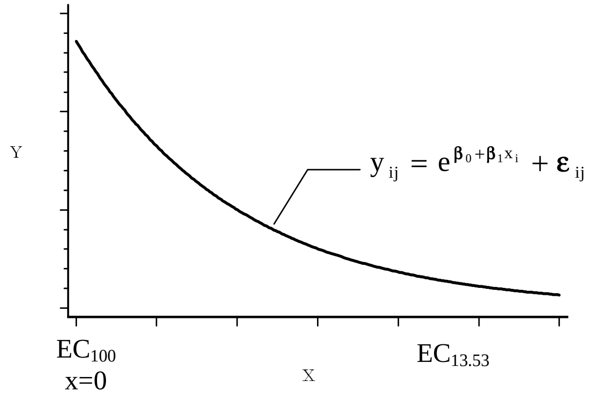

Figure 2.3.1 Characterization of the D-Optimal Single Regressor Design.

§2.4 F-Optimality

The focus of F-optimality is the estimation of a particular EC as precisely as possible. For example, a researcher in oncology may wish to develop a chemotherapy drug such that only 5% of cancer cells survive (a 95% impairment in reproduction). In this case, the particular EC of interest is the EC_5. The F-optimal design is constructed so that the length of a confidence interval around the EC of interest is minimized (Finney, 1971, 1978). This confidence interval can be derived for an EC from any model as long as the EC can be expressed as a ratio of random variables (Fieller, 1944). It is often referred to as a Fieller interval.

In contrast to the other optimality criteria, F-optimality was actually developed in the context of a non-linear model. Its most common application has been in logistic design situations (Sitter & Wu 1993; Letsinger, 1995). However, it has been applied to the Poisson exponential model in the impaired reproduction model context (Minkin, 1993). Since the criterion is largely model dependent, a general form cannot be given. However, the only model it is applied to in this work is the single regressor Poisson exponential model. In this model, the F criterion reduces to

$$\min_{\delta \in \mathcal{D}} \operatorname{var}(b_1). \tag{2.4.1}$$

It should be noted that the F-optimal designs discussed are invariant to the EC of interest. For example, the F-optimal design for the EC_{12} is also the best design to estimate the EC_{76}. This is because minimization of the $\operatorname{var}(b_1)$ does not depend on the EC. Note that the F-optimal design for this model was found by Minkin (1993). This design places 78% of the experimental units at the $EC_{7.8}$ and 22% at the EC_{100} or the control.

§2.5 Bayesian Design Optimality

Bayesian design optimality is a logical outgrowth of traditional design optimality for cases where the criterion is a function of unknown parameters. For the case of uncertainties in model parameters Chaloner and Verdinelli (1995) gives a nice historical review of Bayesian design. Chaloner and Larntz (1989) used the Bayesian method in an effort to create optimal robust designs based on the researcher’s a priori belief about the parameters for the logistic case. Letsinger (1995) extended the applications of Chaloner and Larntz. Other researchers have used this design technique with several different models including compartmental models (Atkinson, et. al., 1993) and linear models (Chaloner, 1984, Dumouchel and Jones, 1994, Neff, 1997).

The premise of the Bayesian design is that the researcher’s beliefs about the parameters can be translated into a density for the parameters. The general form of a Bayesian design criterion is given by (2.5.1).

$$\max_{\delta \in \mathcal{D}} \int_{\theta} R(\delta, \theta) \pi(\theta) d\theta \qquad (2.5.1)$$

where **q** is a vector of model parameters, $\pi(\theta)$ is the prior density of θ, $R(\delta,\theta)$ is any design optimality criterion expression of choice, and δ is any design from the set of possible designs D. This expression has its origins in Bayesian decision theory where $-R(\delta,\theta)$ represents the Bayes risk. The Bayesian optimal design minimizes the expected risk. Although (2.5.1) appears similar to expressions for Bayesian estimation, it is quite different. Unlike Bayesian estimation, $-R(\delta,\theta)$ does not represent an expected loss function. Thus, the decision, δ, to be made is a design rather than an estimator. The entire Bayesian design procedure will be employed in two different contexts in this work. In Chapters 3, 4, and 10, it will be employed in its traditional state explained above in order to find Bayesian designs based on a variety of information about the parameters. In Chapter 6, priors on the parameters will be replaced by discrete priors on a set number of models in the search for model robust Bayesian designs.