Chapter 1

INTRODUCTION AND LITERATURE REVIEW

§1.1 Motivation

In recent years, a large portion of scientific research has been driven by environmental concerns and health-related issues. As the millennium looms on the horizon, there is growing interest in improving the quality of human life by preserving our natural surroundings and fostering new medical advances. Environmental studies typically focus on assessing the effect of pollutants on the delicate balances of nature's ecosystems while medical research centers on the eradication of serious illnesses that affect the world's populations. The toxicity study is a common thread which runs through both environmental and health-related research. Practitioners in these fields are often interested in assessing the effect of a single toxicant or group of toxicants on a particular environment. Since these studies have considerable impact on human life, proper experimental design and statistical analyses are of the utmost importance.

A common statistical technique used in toxicity studies is inverse regression. While the response is interesting in these studies, the more important quantity is the effective

concentration (EC) of the toxicant or mixture of toxicants in question. The effective concentration refers to the amount of a single toxicant or the combination of concentrations in a mixture of toxicants which elicits a particular response. For example, one scientist may be interested in the amount of a certain pesticide that is lethal to 90% of fruit flies. A researcher in a biomedical field may be interested in the combination of several drugs which alleviate symptoms in 80% of the patients in the study group. In both of these examples, the goal of the experiment is estimation of a particular EC of the substance being studied. The dichotomous response in these and many other situations suggests that logistic regression would be an appropriate technique for modeling the probability of a particular outcome (Finney, 1978). However, there is a growing class of toxicity studies involving non-dichotomous data.

This growing class of toxicity studies includes impaired reproduction studies in which the response is measured as a count (Oris and Bailer, 1993; Minkin, 1993). The purpose of an impaired reproduction study is to determine how toxicants affect the reproductive ability of organisms or cells. The EC in these studies is the amount of toxicant which causes a particular amount of impairment in reproduction or a particular percentage decrease in the cell or organism count. One such study in the Statistics Department at Virginia Tech was funded by the United States Air Force. On occasion, the Air Force discharges jet fuel into aquatic ecosystems. The goal of this study was to determine how various combinations of levels of the toxicants impair population growth through decreased organism reproduction. Impaired reproduction studies are the tools which can help identify the concentration of the various toxicants responsible for these damages.

Although these studies are quite prevalent in the biological realm, the efficient design of experiments for these studies is still in its infancy. Chiacchierini (1996) has done extensive work formulating optimal designs in the single regressor case via response surface methodology. In her dissertation, she developed optimal designs for four impaired reproduction models including a linear model with normal errors, a Poisson exponential model, a Poisson linear model, and a model known as a Poisson power model. Within each of these types of single regressor models, optimal experimental designs were found for three optimality criteria. One of the criteria addresses

estimation of model parameters, one concentrates on prediction with the final fitted model, and one addresses estimation of a particular EC (Minkin, 1993).

In this dissertation, we begin by offering an overview of design optimality in Chapter 2. Single regressor design techniques are covered in Chapters 3-6 while designs for multiple regressor models are detailed in Chapters 7-11. In Chapters 3 and 4, we address the dependency of optimal designs on the unknown parameters of the model by developing Bayesian optimal designs with priors on the parameters. Equivalence theory is adapted in Chapter 5 to verify the optimality of some of the Bayesian designs found in Chapters 3 and 4. In Chapter 6, we expand upon Chiacchierini's designs for several single regressor models by formulating Bayesian designs robust to model misspecification. Chapter 7 details both D-optimal and D_s-optimal designs for the *k*-regressor no interaction (or main effects) exponential model. Chapter 8 continues the work on the *k*-regressor model with D, D_s, and interaction optimal designs for the interaction (or full) model. In both Chapters 7 and 8, augmentations for lack-of-fit testing are covered along with designs on restricted regions. Chapter 9 contains sections on two important response surface topics for the exponential model including fractional factorials and prediction variance. Robustness to parameter (or EC) misspecification in the multiple regressor model is addressed in Chapter 10 via Bayesian design techniques. The extension of the methodologies discussed in Chapters 3-10 to the exponential model in general is demonstrated in Chapter 11. Finally, Chapter 12 contains topics for future research.

§1.2 Impaired Reproduction Studies

A typical impaired reproduction experiment consists of a number of experimental units, some of which are treated with particular concentrations of toxicant and others that are labeled as controls. Controls are experimental units which are not treated with toxicant. (The role of control observations in these studies will be discussed in the next section.) At a specific time designated by the experimenter, the organisms in each experimental unit are counted. At this point, a statistical model is fit to the count data in order to make inferences. Three models are used in this work.

§1.3 The Poisson Models

Two models with Poisson error structures are considered in this research. They are the Poisson exponential model and the Poisson linear model. Both of these are derived using general linear models (GLM) theory (McCullagh and Nelder, 1989). GLM theory allows models to be formed from any error distribution that is a member of the exponential family. It uses link functions to determine the specific form of the model. Generalized linear models are so named because they all contain a linear function often known as the " linear predictor", $\mathbf{x}_i'\boldsymbol{\beta}$ where $\mathbf{x}_i'$ is a vector of independent variables and $\boldsymbol{\beta}$ is a vector of unknown regression coefficients.

The derivation of the GLM models begins with the expression for the density function for a member of the exponential family,

$$f(y_i) = \exp\left\{r(\phi)\left[y_i\theta_i - g(\theta_i)\right] + h(y_i,\phi)\right\} \tag{1.3.1}$$

where ϕ is the scale parameter and θ_i is a natural location parameter which is related to the mean of the distribution. The development of a particular GLM model begins with the specification of a link function. In this context, a link function is a function, s, of the mean of the distribution which is modeled by the linear predictor, $\mathbf{x}_i'\boldsymbol{\beta}$ so

$$s(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta} \tag{1.3.2}$$

where s is some function which is monotonic in $\mathbf{x}_i'\boldsymbol{\beta}$. Of course, expression (1.3.2) implies that the mean of the distribution is modeled by

$$\mu_i = s^{-1}\left(\mathbf{x}_i'\boldsymbol{\beta}\right). \tag{1.3.3}$$

Many different GLM models based on the same error structure can be built by using different link functions. However, each distribution in the exponential family has a natural or canonical link function. The canonical link function is determined by θ_i, the natural location parameter, which is

a function of the mean determined by the density function. The general form for the canonical link function for any distribution is given by

$$s(\boldsymbol{\mu}_i) = \boldsymbol{\theta}_i = \mathbf{x}_i'\boldsymbol{\beta}. \tag{1.3.4}$$

Salient features resulting from the use of the canonical link include a simple form of the Fisher information matrix and sufficient statistics for the model parameters which are linear functions of the observations.

A general nonlinear model for impaired reproduction data can be written as

$$y_{ij} = f\left(\mathbf{x}_i'\boldsymbol{\beta}\right) + \boldsymbol{\varepsilon}_{ij} \tag{1.3.5}$$

where $y_{ij} \sim \text{Poisson}(\boldsymbol{\lambda}_i)$. The link function determines $f\left(\mathbf{x}_i'\boldsymbol{\beta}\right)$, the function used to model the mean of the distribution. The two models with Poisson errors considered in this work are the model derived from the log link function, which is the canonical link for the Poisson distribution, and the model derived from the identity link function.

§1.3.1 Model Development

The likelihood, in exponential family form, for a single observation from a Poisson distribution is

$$f\left(y_{ij}\right) = e^{y_i \ln \boldsymbol{\lambda}_i - \boldsymbol{\lambda}_i - \ln y_i!}. \tag{1.3.6}$$

Based on the GLM theory presented previously, it is easy to see that the canonical link for the Poisson distribution is the log link. Use of the log link results in an exponential model for the mean given by

$$\ln \lambda_i = \mathbf{x}_i' \boldsymbol{\beta} \qquad (1.3.7)$$
$$\Rightarrow \lambda_i = f\left(\mathbf{x}_i' \boldsymbol{\beta}\right) = e^{\mathbf{x}_i' \boldsymbol{\beta}}$$

for i=1, . . ., n_i.

Using the identity link rather than the log link results in the formulation of a Poisson linear model. Since $\lambda_i = \mathbf{x}_i' \boldsymbol{\beta}$ is already in terms of the mean of the distribution, the single regressor linear model is

$$y_{ij} = \mathbf{x}_i' \boldsymbol{\beta} + \boldsymbol{\varepsilon}_{ij}. \qquad (1.3.8)$$

§1.3.2 The Square Root Transformation Model

The third and final model used in this research is the square root transformation model. Linear models are often fit to count data when the range of the data is quite small. However, the possibility of using ordinary least squares estimation is eliminated because the homogeneity of variance assumption is violated. When faced with this situation for Poisson data, the square root transformation is frequently used to stabilize variance. The square root transformation model can be written as

$$\sqrt{y_{ij}} = \beta_0 + \beta_1 x_i + \varepsilon_{ij}^*. \qquad (1.3.9)$$

where the e_{ij}^* are distributed N(0,1/4). Using a Taylor series expansion, it can be verified that

$$E\left(\sqrt{y_{ij}}\right) \approx \sqrt{E\left(y_{ij}\right)} = \sqrt{\lambda_i}. \qquad (1.3.10)$$

With the models explained, details specific to impaired reproduction studies will be covered.

§1.4 Control Observations

As mentioned earlier, control observations play a paramount role in impaired reproduction studies. In order to measure the amount of impairment which occurs in an environment, one must be

familiar with the population's size in the absence of toxicant. Thus, at least one observation should be taken at the control, where no toxicant has been administered to the experimental unit. Fortunately, most of the optimal designs in this work allocate some experimental units to the control. However, it should be noted that the design criteria can be easily altered to force allocation of points at the control if necessary.

§1.5 Notes on Poisson Models

The theoretical support for a Poisson random variable is the integers on $[0,\infty)$. However, Poisson design problems require the specification of a finite maximum for the random variable of the experiment. In impaired reproduction studies, this upper limit on the value of the random variable is the amount of reproduction that occurs at the control. Thus, λ_c, the mean reproduction at the control, is the maximum possible reproduction that can occur in any experimental unit. As a result, the distribution of the data is really a truncated Poisson. In addition, this implies that the expected response at any non-control design point is a percentage or proportion of the expected response at the control. Mathematically this means that

$$E(y_{ij}) = \lambda_i = q_i \lambda_c \qquad (1.5.1)$$

where $0 \le q_i \le 1$. The quantity q_i is the "percent of maximum reproduction that occurs at design point i" or 1 - (proportion of impairment at design point i). With the exposition on the models complete, the focus now shifts to the presentation of the design optimality criteria pertinent to this work.