GENERALIZED LINEAR MODELS IN BIOMEDICAL

STATISTICAL APPLICATIONS

by

Pranab Kumar Sen
Department of Biostatistics
University of North Carolina

Institute of Statistics
Mimeo Series No. 2152

December 1995
GENERALIZED LINEAR MODELS IN BIOMEDICAL STATISTICAL APPLICATIONS†

PRANAB KUMAR SEN
University of North Carolina, Chapel Hill.

SUMMARY

In biological assays, clinical trials, life-testing problems, reliability and survival analysis, and a variety of applied fields ranging from agrometry to biodiversity to zodiacal sciences, generalized linear models covering usual linear, transformed linear, log-linear and nonlinear regression models, multinomial and logistic regression models, as well as some semiparametric ones, are potentially adoptable for drawing statistical conclusions from acquired data sets. Yet there may be some hidden barriers, originating from experimental as well as observational schemes, that merit careful examination in applications. Merits and demerits of generalized linear models in biomedical applications with due emphasis on their validity and robustness properties are thoroughly discussed.

1 INTRODUCTION

In a conventional experimental setup, the causal relationship of input and output variables is of prime interest, and such a picture is often smudged to a certain extent by chance fluctuations or errors. To accommodate plausible chance variations around predictable or systematic levels, in a conventional linear model it is tacitly assumed that the output (or, dependent) variable, \( Y \), is related to a set of explanatory (or, input) variables, \( x \) in a linear mode, subject to errors having some statistical (distributional) structures. For example, in a classical setup, it is assumed that

\[
Y = \beta' x + e,
\]

where \( \beta \) is an unknown vector of (regression) parameters, and the errors are generally assumed to have three basic properties: (i) normality (with zero mean and a finite variance \( \sigma^2 \)), (ii) independence (for different observations) and (iii) homoscedasticity (at all levels of \( x \)). As a result, statistical analysis schemes pertaining to such linear models remain vulnerable to plausible departures from linearity of the model as well as independence, homoscedasticity and normality of the error components. The quest for studying robustness and validity of normal theory linear models laid down the foundation of more complex statistical models which may be used as alternative ones. Depending on a particular situation at hand, one or more of these basic regularity assumptions may

---

†AMS 1990 subject classifications: 62J12, 62N05, 62P99

1Key words and phrases: Asymptotics; auxiliary variate; binning; bioassays; dilution, direct, indirect, parallel line, qualitative, quantal, slope ratio; canonical link; censoring; clinical trials; competing risks; concomitant variate; conditional model; correlated binary responses; dosage; frailty models; GEE; hazard; isotonic regression; link function; logit; log-linear model; marginal analysis; measurement errors; misclassification; missing data; mixed GLM; multiple endpoints; NN-method; nonlinear model; nonparametrics; partial likelihood; Poisson regression; probit; PHM; quasi-likelihood; relative potency; response-metameter; risk; semi-parametrics; smoothing; step down procedure; time-dependent coefficients.
need special scrutiny, and that may dictate particular alternative models and appropriate statistical analysis procedures. Since in diverse fields of applications, the setups may vary considerably, it may be more appropriate to consider some typical scenarios arising in some typical contexts, examine the drawbacks of the classical linear model approach, and to introduce alternative models which may have greater appeal on the grounds of robustness and validity. Generalized linear models (GLM) are important members of this bigger class, and they are therefore to be examined with due care, relevance and importance.

Our perception of GLM's is posed from a wider perspective than in Nelder and Wedderburn (1972) who showed how linearity in common statistical models could be exploited to extend classical linear models to more general ones and exact statistical inference tools can therefore be used. Further significant developments in this direction are reported in a unified manner in McCullagh and Nelder (1989). It follows from the current state of developments in this field that GLM's include, as special cases, linear regression and analysis of variance (ANOVA) models, log-linear models for categorical data, product multinomial response models, logit and probit (or normit) models in quantal bio-assays, and some simple semi-parametric models arising in reliability and survival analysis. In an exact treatise of the subject matter, GLM's pertain to a somewhat narrow class of statistical problems (where the relevance of the classical exponential family of densities can be justified). However, with the widening impact of computers and large scale data analyses in every sphere of life and science, there is a genuine need to look into the prospects from a large sample perspective with due emphasis on computational complexities as well as robustness properties. Our motivations in the current study are primarily driven by such undercurrents.

In Section 2, we provide a brief outline of the basic GLM (exact) methodology which permits a comprehensive access to the large sample counterpart, which is reported in Section 3. GLM's role in analysis of binary data with special emphasis on quantal bioassays is appraised in Section 4. Direct dilution assays are introduced in Section 5, and the case of quantitative indirect assays is presented in Section 6. The relevance of GLM's to nonlinear regression models is studied in Section 7 from a large sample view point, while the case of nonparametric regression is considered in Section 8. The concluding section deals with a broad review of the scope and limitations of GLM's in applications in a wider perspective.

2 GLM: EXACT METHODOLOGY

The genesis of GLM lies in the exponential family of densities. Let \( Y_1, \ldots, Y_n \) be independent (but not necessarily identically distributed) random variables, and assume that \( Y_i \) has the density function

\[
\begin{align*}
  f_i(y; \theta_i, \phi) &= c(y, \theta_i, \phi) \exp[(y \theta_i - b(\theta_i))/a(\phi)], \quad i = 1, \ldots, n, \\
  \end{align*}
\]

(2.1)

where the \( \theta_i \) are the parameters of interest, \( \phi(>0) \) is a nuisance (scale) parameter, and \( a(\cdot), b(\cdot) \) and \( c(\cdot) \) are functions of known forms. It is tacitly assumed that \( b(\theta) \) is twice (continuously) differentiable (with respect to \( \theta \)), and we denote the first and second order derivatives by \( b'(\cdot) \) and \( b''(\cdot) \) respectively. Then, it is easy to check that

\[
\begin{align*}
  \mu_i &= EY_i = \mu_i(\theta_i) = b'(\theta_i) ; \quad Var Y_i = a(\phi)b''(\theta_i), \quad i = 1, \ldots, n. \\
\end{align*}
\]

(2.2)

The last equation permits us to introduce the so called variance function \( v_i(\mu_i(\theta_i)) \) by letting

\[
\begin{align*}
  v_i(\mu_i(\theta_i)) &= \frac{\partial}{\partial \theta_i} \mu_i(\theta_i), \\
\end{align*}
\]

(2.3)

and it depends solely on \( \mu_i(\theta_i) \), for \( i = 1, \ldots, n \). The \( \mu_i(\theta_i) \) play a basic role in the formulation of GLM's. As we have assumed that the form of \( b(\theta_i) \) is the same for all \( i \), we have the form of \( \mu_i \)

2
and $v_i$ independent of $i$. As such, taking clue from the normal density case where $b''(.) = 1$, we conceive of a lower dimensional parameter vector $\beta = (\beta_1, \ldots, \beta_q)$ and a $n \times q$ specification matrix $X_n = (x_{n1}, \ldots, x_{nn})^t$ and express

$$\theta = (\theta_1, \ldots, \theta_n)^t = X_n\beta.$$  

(2.4)

This suggests that it may be possible to formulate a suitable transformation, called the link function

$$g\{\mu(\theta_i)\} = g(\mu_i), \quad i = 1, \ldots, n,$$  

(2.5)

where $g(u)$ is a monotone and differentiable function of $u$. Such that

$$G_n = \{g(\mu_1), \ldots, g(\mu_n)\}^t = X_n\beta.$$  

(2.6)

Therefore, we may write $g\{\mu(\theta_i)\} = x_{ni}^t\beta$, so that as in Nelder and Wedderburn (1972), we have

$$\theta_i = (g \circ \mu)^{-1}(x_{ni}^t\beta), \quad i = 1, \ldots, n.$$  

(2.7)

Thus, whenever $\mu$ is monotone, and we let $g = \mu^{-1}$, we obtain that $g \circ \mu$ is the identity function, so that $\theta_i = x_{ni}^t\beta$, and hence, such a $g$ is termed a canonical link function. Most of the literature on GLM is devoted to suitable characterization of such (canonical) link functions for various types of densities belonging to the exponential family, and then direct adaptations of the standard linear model methodology through the so called estimating equations (EE) which are the natural extensions of the usual normal equations which appear in the classical normal theory least squares methodology. Whenever $g(\cdot)$ is a canonical link function, the EE for the maximum likelihood estimator (MLE) $\hat{\beta}_n$ can be written equivalently as

$$\sum_{i=1}^{n}(Y_i - b'(x_{ni}^t\beta))x_{ni} = 0,$$  

(2.8)

and this immediately leads us to the solution

$$\hat{\beta}_n = \left[\sum_{i=1}^{n}x_{ni}x_{ni}^t\right]^{-1}\left[\sum_{i=1}^{n}x_{ni}Y_i\right].$$  

(2.9)

There are some additional complications when $g(\cdot)$ is not a canonical link function. Drawing a parallel with the classical weighted least squares (WLS) methodology, we may then consider some generalized estimating equations (GEE) based on the following notations. Let

$$r_n(\beta) = \{Y_1 - \mu_1(\beta), \ldots, Y_n - \mu_n(\beta)\}^t;$$

$$D_n(\beta) = \text{diag}[g'(\mu_1(\beta))v_1(\beta), \ldots, g'(\mu_n(\beta))v_n(\beta)].$$

(2.10)

Then a GEE can be put either in the form

$$X'_nD_n^{-1}(\beta)r_n(\beta)|_{\beta = \hat{\beta}_n} = 0,$$  

(2.11)

or in an explicit form as

$$\sum_{i=1}^{n}[g'(\mu_i(\hat{\beta}_n))v_i(\hat{\beta}_n)]^{-1}(Y_i - \mu_i(\hat{\beta}_n))x_{ni} = 0.$$  

(2.12)

Note that, as the weights themselves depend on the unknown parameters, the roots of the above equations may not always exist or be unique. The conditions needed for the existence and uniqueness are essentially those of concavity of the likelihood function and convexity of the parameter space; viz., Silvapulle and Burridge (1986) where other references to earlier works are also cited. It is clear
that in general we have implicit equations for the unknown $\beta$, and in this way the analogy with the usual MLE may also be pointed out. We may note further that if the response is multivariate then there are additional complications in the formulation of link functions and variance-covariance functions. It may be difficult to find out canonical links (and even so, they may not be coordinate-wise functions), and the variance-covariances may also be highly involved. For some important cases of GLMs, we will discuss these aspects of the GLM briefly in later sections.

3 GLM ASYMPTOTICS

With respect to general asymptotics, the GLM may generally encounter more complications than the classical LSE or MLE. While the main complication arises due to the implicit nature of the GEE, it is further aggravated by the complexities of the link and variance functions. We follow the general prescriptions in Sen and Singer (1993, sec. 7.4) in summarizing the general results depending on the underlying structure in the form of the following three models.

(1) Model I. This relates to (2.1) when the basic random variables $Y_i$ are independent and $n$ is large. A classical example is the Poisson regression model where the $Y_i$ are independent Poisson variables with means $\lambda_i$, and either the $\lambda_i$ or their logarithms are related linearly to a set of explanatory variables (vectors) $x_i$ by means of a finite dimensional parameter vector $\beta$; in this setup the sample size $n$ is assumed to be large. In terms of $\beta$ the log-likelihood function for the model (2.1) is given by

$$\ln L_n(\beta) = \sum_{i=1}^{n} [Y_i h(x'_i \beta) - b\{h(x'_i \beta)\}] + \text{constant},$$

where the constant term does not involve $\beta$ and

$$h(\cdot) = (g \ast \mu)^{-1} \text{ is } \mathcal{C} \text{ and differentiable.}$$

By some routine manipulations we derive the score statistic:

$$U_n(\beta) = \left( \frac{\partial}{\partial \beta} \right) \ln L_n(\beta)$$

$$= \sum_{i=1}^{n} \{Y_i - \mu_i(\beta)\} [g'(\mu_i(\beta))v_i(\beta)]^{-1} x_{ni}.$$  

Differentiating one more time with respect to $\beta$ and taking expectation, we obtain

$$\mathbb{E}_n(\beta) = \mathbb{E}\left\{ -\left( \frac{\partial^2}{\partial \beta \partial \beta'} \right) \ln L_n(\beta) \right\}$$

$$= \sum_{i=1}^{n} \{g'(\mu_i(\beta))\}^{-2}[v_i(\beta)]^{-1} x_{ni} x'_{ni}.$$  

If we let

$$w_{1i}(\beta) = [g'(\mu_i(\beta))]^{-2}[v_i(\beta)]^{-1}, \quad i = 1, \ldots, n;$$

$$w_{2i}(\beta) = \mu_i(\beta) [g''(\mu_i(\beta))][g'(\mu_i(\beta))]^{-2}$$

$$= b'' \{h(x'_i \beta)[g'(\mu_i(\beta))][v_i(\beta)]^{-3}\}, \quad i = 1, \ldots, n.$$  

4
then under two compactness conditions on the $w_{ki}(\beta), k = 1, 2; i = 1, \ldots, n$, it follows from Theorem 7.4.1 of Sen and Singer (1993) that for $\hat{\beta}_n$, the MLE of $\beta$ under this GLM, as $n$ increases,

$$\sqrt{n}(\hat{\beta}_n - \beta) \rightarrow_{D} N_0(0, \Gamma^{-1}(\beta)).$$ (3.7)

Note that the observed second-derivative matrix of the log-likelihood function can be expressed as equal to $I_n(\beta)$ plus a remainder term $R_n(\beta)$ (matrix)[ see Sen and Singer (1993), p.307 ] which has a more complicated form than in the usual likelihood function case. This feature calls for the incorporation of suitable compactness conditions under which this remainder matrix can be handled adequately. For an alternative derivation of the asymptotic normality (and consistency) of MLE in GLS's, we may refer to Fahrmeir and Kaufmann (1985). Compared to the level of complexities in their proof, the derivation in Sen and Singer (1993) based on the compactness conditions appears to be simpler, and more akin to the classical MLE case. Moreover, such compactness conditions follow from the natural continuity properties of the parametric functions in the exponential family, and are not that difficult to verify in other specific cases not belonging to this family. Hence we advocate the use of this compactness condition based approach.

(2) Model II. There are situations where the basic random variables $Y_i$ are themselves based on suitable subsamples of sizes (say) $n_i$ (which may not be all equal). A handy example is the classical logistic regression model where for each $i(=1, \ldots, k)$, for some fixed $k(\geq 1)$, $Y_i$ relates to the observed failure (or response) rates corresponding to a given combination of dose levels. Thus the $n_i Y_i$ are independent $\text{Bin}(n_i, \pi_i)$ variables where the $\pi_i$ can be expressed in terms of a deterministic function of the dose levels. In logistic regression, we express

$$\pi_i = \{1 + \exp(-\beta^t x_i)\}^{-1}, \text{ for } i = 1, \ldots, k,$$ (3.8)

so that for the logits we have

$$\theta_i = \log\{\pi_i/(1 - \pi_i)\} = \beta^t x_i, \text{ for } i = 1, \ldots, k.$$ (3.9)

Usually in this setup, $k$ is prefixed, while for each dose level, a number of subjects are administered, and guided by the convergence of the binomial law (appropriately normalized) to a normal law, it may be assumed that there exist suitable normalizing factors, say, $a_{n_i}(\theta_i)$ and $b_{n_i}(\theta_i)(> 0)$, such that for large values of $n_i$,

$$\{Y_i - a_{n_i}(\theta_i)\}/b_{n_i}(\theta_i) \rightarrow_{D} N(0, 1), \forall i(=1, \ldots, k).$$ (3.10)

Thus whenever the GLM can be adopted for the $\theta_i$, using the conjugate transformation on the $Y_i$ and the aforesaid normality result standard large sample methodology can be incorporated to study the asymptotic properties of the MLE of $\beta$. We will discuss this in detail in a later section. In passing we may note that allowing $k$ to be large in this setup does not create any additional complications; rather it accelerates the rate of convergence to asymptotic normality and thereby facilitates the adoption of the standard large sample methodology.

(3) Model III. This relates to the so called quasi-likelihood or quasi-score estimating equations, introduced by Wedderburn (1974). In this setup no specific assumptions are made concerning the form of the density in (2.1); it is assumed on the contrary that (i) the $E(Y_i) = \mu_i$ satisfy the GLM setup, and (ii) $\text{Var}(Y_i) = \sigma^2 V(\mu_i)$ where $\sigma^2$ is a possibly unknown positive scale function while the $V(\cdot)$ are completely known variance functions. The related estimating equations may be posed as

$$\sum_{i=1}^{n}(\partial/\partial \beta) \mu_i \{V_i(\mu_i)\}^{-1}(Y_i - \mu_i) = 0.$$ (3.11)
For the consistency and asymptotic normality of the quasi-likelihood estimators, we need to make assumptions parallel to that in Model I but related to the quasi-scores rather than the likelihood function; for some early work we refer to Godambe and Heyde (1987). Liang and Zeger (1986a) have considered a further generalization of the above methodology wherein they allowed the $V_i(\cdot)$ of known functional forms to depend on some additional (but finite dimensional) nuisance parameter, say $\eta$. Based on a $\sqrt{n}$-consistent estimator of $\eta$, they were able to establish the consistency and asymptotic normality of their GEE estimators. The methodology pertaining to Model I can be used with advantage for their generalization to GEE, and for some discussion we may refer to Sen and Singer (1993, ch. 7).

4 GLM IN QUANTAL BIOASSAYS

Quantal bioassays provide a motivating illustration for GLM's and also a comparative picture of the competing GEE vs GLSE (generalized least squares estimation) methodologies. We conceive of a dose-response study where the response is quantal (i.e., all or nothing), so that the outcome is binary in nature. Suppose that there are $k (\geq 2)$ levels of the dose, designated as $d_1, \ldots, d_k$, where the $d_j$ are nonnegative, prespecified and ordered, i.e.,

$$0 \leq d_1 < \cdots < d_k < \infty.$$  (4.1)

In addition to the dose-variable there may be other (design or auxilliary) variables. Moreover a transformation is often adopted on the dose and other variables so as to induce some simplification on the dose response regression. For this reason, we conceive of a set of auxiliary vectors (including the dosage or dose-transformation): $x_i, i = 1, \ldots, k$, which are not all the same. Suppose that $n$ subjects are divided into $k$ sets of $n_1, \ldots, n_k$ units, such that in the ith set the binary responses $Y_{ij}, j = 1, \ldots, n_i$ relaite to a Binomial law

$$P\{Y_{ij} = 1|x_i\} = 1 - P\{Y_{ij} = 0|x_i\} = \pi_i(x_i), j \geq 1; \ i = 1, \ldots, k.$$  (4.2)

In a simple dose-response model, we may set

$$\pi_i(d_i) = 1 - F(T_0 - \beta d_i), \ i = 1, \ldots, k,$$  (4.3)

where $T_0$ refers to a threshold value, and $F$ is a (tolerance) distribution function. Among various possibilities the popular choice of $F$ includes (i) the logistic and (ii) the normal distribution. In the first case, $\pi_i(d_i) = \{1 + \exp(-\beta_0 - \beta d_i)\}^{-1}, i = 1, \ldots, k,$ while in the second case, we may take $\pi_i(d_i) = \Phi(\beta_0 + \beta d_i), i = 1, \ldots, k,$ where $\Phi(\cdot)$ is the standard normal d.f. It is not possible to use an appropriate transformation (such as the logit or the normit/probit) on the individual $Y_{ij}$ and directly adopt the model (2.1). Rather, we take the total response at each dose level, i.e., $Y_i = \sum_{j=1}^{n_i} Y_{ij}$, which are independent Bin($n_i, \pi_i(d_i)$) variables, and use such a transformation on them. Thus we get the sample counterparts of the logit and probit as

$$Z_i = \log\left[\frac{Y_i}{n_i - Y_i}\right] \text{ or } \Phi^{-1}(Y_i/n_i), i = 1, \ldots, k.$$  (4.4)

These are then incorporated in the desired statistical modeling and analysis. Keeping this simple picture in mind, we now conceive of a comparatively more general model where corresponding to a set of design or auxilliary vectors $x_i, i = 1, \ldots, k$, we have response variates $Y_i, i = 1, \ldots, k$ which are independent binomial with parameters $(n_i, \pi_i)$,

$$\pi_i = \mu_i(x_i|\beta), \ i = 1, \ldots, k,$$  (4.5)
and the functional forms of the $\mu_i(.)$ are assumed to be known. Finney (1978) contains an excellent account of statistical methodology of the probit model, while the logit model has been extensively studied in the literature, mostly in the context of log-linear models for categorical data; we may refer, for example, to Sen and Singer (1993, ch.6 -7), where other pertinent references are also cited.

In this formulation we can directly link the methodology presented in Section 2 and carry out the appropriate statistical analysis by reference to the GLM methodology described therein. The main point of departure is, however, the basic fact that the $Y_i$ are not the elementary variables but are composed from them for each bucket separately. This makes the variance $Var(Y_i)$ dependent not only on the known scale factor $n_i^{-1}$ but also on the unknown $\beta$ through the $\pi_i\{1 - \pi_i\}$ which are of more complex nature. This feature calls for appropriate asymptotic tools for providing suitable approximations for the GEE and the derived MLE, and we will pay due attention to this development.

In view of the comment made above it seems that in this case we may not be able to make use of the estimating equation in (2.8), and we have to incorporate the GEE in (2.12) for obtaining the MLE $\hat{\beta}$ of $\beta$. This calls for an iterative procedure in which we start with an initial (consistent) estimator $\hat{\beta}^{(0)}$, estimate the variance function in (2.12), and using the scoring procedure outlined in Section 3, we can obtain the first-step estimator of $\beta$ as well as the variance function. We can iterate this procedure a number of times, and for a suitable choice of $r$ (usually 2 or 3), we may use an $r$-step procedure to approximate the MLE adequately. It is clear that in this iteration scheme there are certain elements of asymptotic theory which merit a careful study. First, if we use the logit, probit (or any other nonlinear transformation) on the $Y_i$, the computation of the exact variance of the transformed variate may run into complexities. For this reason in the logistic regression model we start with an estimating equation:

$$Q_n(\hat{\beta}, \beta) = \sum_{i=1}^{k} p_i(1 - p_i)n_i[Z_i - \beta_e - \beta'x_i]^2,$$

(4.6)

where

$$Z_i = log[p_i/1 - p_i], \quad p_i = Y_i/n_i, \text{ for } i = 1, \ldots, k.$$  

(4.7)

The basic difference between this estimating equation and the GEE in (2.12) (for a logit transformation) is that here we use $p_i(1 - p_i)n_i$ as an estimate of the reciprocal of the (asymptotic) variance of $Z_i$, so that the weighted least squares estimation theory can directly be adopted for the estimation of $\beta_e$ and $\beta$. On the other hand, the difference between the variance function in (2.12) and $n_ip_i(1 - p_i)$ can be shown to be $O_p(n_i^{-1/2})$, so that this approximation does not entail any significant discrepancy with the MLE derived from (2.12). In that sense, the GEE is a minor refinement over the logistic regression approach when the individual sample sizes $n_1, \ldots, n_k$ are all large. If these sample sizes are not large, the expression for the variance functions in (2.12) may not retain their exactness, and hence the rationality of the GEE in (2.12) may not be that clear. The situation with the Probit analysis is a bit more complex (see, for example, Finney 1978) where the estimating equations are more complex and so are the corresponding variance functions. Nevertheless, the iterative procedure with the GEE works out well for large sample sizes. The question remains open: How large should be the sample sizes in order that the asymptotics in GEE are adequate? Second, from the asymptotics point of view, we are basically in Model II of Section 3, where the emphasis on GEE entails much less of the GLM than the appropriateness of the transformation and its asymptotic normal properties. Thus, the primary force behind GLM is somewhat imperceptible in this asymptotic case. Finally, the GLM may entail a comparatively severe nonrobustness feature than some other alternative models. For example, suppose that the true tolerance distribution is $F$ and we assume a logistic model. Then, if the $x_i$ are so chosen that the resulting $\pi_i$ are not all clustered around 0.5, the estimator of $\beta$ based on this logit model may be highly nonrobust to plausible departures of
$F$ from a logistic distribution, A similar feature is shared by the probit model. Another important case relates to a mixture model in quantal response where a GLM approach has been considered by Cox (1992), and this also reveals the nonrobustness aspects of the conventional methodology. In actual applications this robustness feature therefore merits due considerations.

5 GLM IN DIRECT DILUTION ASSAYS

In bioassay, typically, a new (test) and a standard preparation are compared by means of reactions that follow their applications to some some subjects, and the relative potency of the test preparation with respect to the standard one is the focal point of the study. In a direct assay, the dose of a given preparation needed to produce a specified response is recorded, so that the dose is a nonnegative random variable (r.v.) while the response is certain. Consider a typical direct bioassay involving a standard preparation (S) and a test preparation (T). Suppose that the standard preparation is administered to $m$ subjects, and let $X_1, \ldots, X_m$ be the respective doses to yield the desired response. Thus we assume that the $X_i$ are independent and identically distributed (i.i.d.) r.v.'s with a distribution function (d.f.) $F_S$ defined on $\mathbb{R}^+ = [0, \infty)$. Similarly, let $Y_1, \ldots, Y_n$ be the doses for the $n$ subjects on which the test preparation has been administered; these are assumed to be i.i.d.r.v. with a d.f. $F_T$ defined on $\mathbb{R}^+$. In many cases we may think that the test preparation behaves, in a properly interpreted bioequivalence setup, as if it is a dilution (or concentration) of the standard one. Under this hypothesis, we have

$$F_T(x) = F_S(px), \forall x \in \mathbb{R}^+, \text{ where } p > 0. \quad (5.1)$$

Then $p$ is termed the relative potency of the test preparation with respect to the standard one, and (5.1) constitutes the fundamental assumption in a direct (-dilution) assay; we refer to Finney (1978) for an excellent coverage of this subject matter in a strict parametric setup where the $F$ are mostly assumed to be normal or log-normal d.f.'s. The very assumption that $F_S$ is normal may run into conceptual obstructions unless the mean of this d.f. is sufficiently large compared to its standard deviation so that $F_S(0)$ can be taken to be equal to 0. Moreover, in actual practice, these tolerance distributions are usually highly positively skewed, and hence suitable transformations on the dose (called the dosage) are generally recommended to induce more symmetry of the d.f. of the transformed variable. Among these, two popular ones are the following:

(i) log-dose transformation : $X^* = \log X$.

(ii) Power transformation : $X^* = X^\lambda$, for some $\lambda > 0$. Typically, $\lambda$ is chosen to be 1/2 or 1/3 depending on the nature of the dose.

If we work with the model (5.1) we have a typical scale-model, so that standard MLE of $p$ can be obtained under normality assumptions; note in this respect that the variances of $F_S$ and $F_T$ are not equal (unless $p = 1$), but their coefficients of variation are equal. In the case of log-dose transformation, we have a standard two-sample model (with equal variances), so that GLM remains highly pertinent. In the case of power-dose, the choice of $\lambda$ is crucial, and moreover the homogeneity of the variances of the dosages may not be taken for granted, unless $p = 1$. In this respect, there appears to be some ambiguity in Finney (1978) and other text books on bioassays where often the well known Fieller theorem (on the ratio of two means) has been adopted without proper safeguard. This technical difficulty does not arise in the case of log-dose transformation, but the resulting MLE (of $p$) becomes highly sensitive to the assumed normality. In a variety of situations the log-dose transformation induces a greater amount of symmetry to the original the d.f., but still it may not be fully symmetric, or even so, may differ from a normal d.f., often quite remarkably. Thus, in using the
log-dose transformation it should be kept in mind that, though the GLM formulation appears to be valid, the assumption of normality has to be judged very crucially in a given context, and robustness (for plausible nonnormality) considerations are therefore very pertinent. Therefore it may be argued that an estimator of \( \rho \) should be invariant under any (strictly) monotone transformation on the dose, so that the choice of the dosage would not affect the flavor of statistical analysis to be made. In that way, it would be possible to present the statistical conclusions to the biological scientists in a more convincing manner. Led by this basic motivation, Sen (1963) formulated distribution-free methods for the (point as well as interval) estimation of \( \rho \) and for testing the null hypothesis on the fundamental assumption in dilution direct assays. Essentially note that the ranks are the maximal invariants under such transformations, so that such inference tools are to be based on appropriate rank statistics.

Because of the invariance property mentioned above and to stick to the GLM as far as possible, we work with the log-dose dosage. Let \( X_i^* = \log X_i, i = 1, \ldots, m \), and \( Y_j^* = \log Y_j, j = 1, \ldots, n \), and let \( F_S^* \) and \( F_T^* \) be the d.f. of \( X^* \) and \( Y^* \) respectively. Then under (5.1), we have

\[
F_T^*(z) = F_T(e^z) = F_S(\rho e^z) = F_S^*(z + \log \rho), \quad z \in \mathbb{R}.
\]  

(5.2)

Thus, we have the classical two-sample problem where the d.f.'s \( F_T^* \) and \( F_S^* \) differ only in a possible shift parameter \( \theta = \log \rho \). This enables us to use the battery of R-estimators of shift in the two-sample problem. Sen (1963) considered in particular the Wilcoxon score and sign statistics and derived the corresponding point estimators and (distribution-free) confidence intervals for \( \theta \) which directly lead to the parallel versions for \( \rho \). Sen (1984) has a more recent account of these developments including the use of more general scores, and hence we omit these details. This clearly shows that using the linearity of the GLM along with invariance of rank based procedures it is possible to achieve robustness to a significant degree without sacrificing efficiency to any perceptible extent. In this development it is not even necessary to assume that \( F_T^* \) is symmetric (not to speak of its normality), and the use of the two-sample normal scores statistics leads to an asymptotic relative efficiency (with respect to the normal theory MLE) that is bounded from below by 1 for all \( F^* \); this lower bound is attained only for normal \( F^* \). Tests for the fundamental assumption in dilution direct assays based on (aligned) two-sample rank statistics have also been considered by some researchers, and for some account of these works we refer again to Sen (1984).

Let us look into the model (5.1) from a somewhat different angle and formulate an alternative GLM approach. If both \( F_T^* \) and \( F_S^* \) were negative exponential d.f.s with respective means \( \mu_S \) and \( \mu_T \), then we may write equivalently

\[
\hat{F}_T(x) = 1 - F_T(x) = \exp\{-x/\mu_T\}, \quad \hat{F}_S(x) = \exp\{-x/\mu_S\};
\]

\[
\hat{F}_T(x) = \{\hat{F}_S(x)\}^{\rho}, \quad \rho = \mu_S/\mu_T.
\]  

(5.3)

Thus defining the hazard function \( h_F(x) \) as \(-(d/dx)\log \hat{F}(x) = f(x)/\hat{F}(x)\), with \( f(.) = F' \), we have under (5.3),

\[
h_S(x) = h_{F_S}(x) = 1/\mu_S, \quad \forall x \in \mathbb{R}^+,
\]

\[
h_T(x) = h_{F_T}(x) = 1/\mu_T = \rho \cdot h_{F_S}(x), \quad \forall x.
\]  

(5.4)

Taking clue from this proportionality of the two hazard functions but dropping the assumption of a constant hazard (i.e., the exponentiality of the d.f.), we may now introduce the so-called proportional hazards model (PHM) by defining

\[
h_T(x) = \rho \cdot h_S(x), \quad \forall x \in \mathbb{R}^+.
\]  

(5.5)

Naturally, we have from the above equation,

\[
\log h_T(x) = \log \rho + \log h_S(x), \quad x \in \mathbb{R}^+,
\]  

(5.6)
and this brings the relevance of GLM through the log-hazard formulation. In terms of (5.5) or (5.6), sans exponentiality of the d.f.'s, it may be difficult to interpret or justify the "dilution" interpretation in (5.1). Nevertheless, (5.5) provides a meaningful interpretation of the relative performance of the standard and test preparation (with due emphasis on the underlying risks functions). This concept has been extended to a more general setup of hazard regression models, and we will briefly discuss the role of GLM in this context in a later section.

In the present context, we treat the hazard function formulation from a GLM perspective, and discuss their relative merits and demerits by reference to alternative approaches. The hazard linear regression model in (5.6) resembles the simple regression setup, but the main question arises as to how to estimate \( \log h_T(x) \) and \( \log h_S(x) \) over a range of \( x \) so as to apply the GLM methodology on such estimates. In a parametric setup (e.g., letting both \( F_T \) and \( F_S \) be Weibull with the same shape parameter, possibly unknown, or gamma or Pareto) these log-hazard functions can be expressed in terms of a known functional form with possibly unknown parameters, so they may be parametrically estimated in an efficient manner. This provides access to efficient statistical analysis using the GLM methodology when the assumed parametric forms of the d.f.'s \( F_T \) and \( F_S \) coincide with their true functional forms. In this framework, \( \sqrt{n} \) consistency of the estimated log-hazard rates can be established under quite general regularity assumptions. On the other hand, such estimated functionals are generally nonrobust to plausible discrepancies between the assumed and true d.f.'s. In actual biological assays, as has been stressed earlier, it may be quite unreasonable to assume the precise functional form of the tolerance distribution, and as such log-hazard functions are nonrobust to inexact distributional assumptions, such a parametric approach may not be advocated in general.

As a second alternative we may note that under (5.5), we have also

\[
- \log \tilde{F}_T(x) = \rho (- \log \tilde{F}_S(x)), \quad \forall x \in \mathcal{R}^+.
\]  

(5.7)

Therefore the GLM methodology may as well be applied to the log-survival functions, and in a parametric mold, this is usually much simpler and more robust than the log-hazard function formulation. Nevertheless, it remains sensitive to plausible departures of the true d.f.'s from the assumed ones. For this reason, we would retain the GLM structure in terms of the model-formulation, but otherwise allow the two d.f.'s \( F_T \) and \( F_S \) to be rather arbitrary to enhance robustness potentiality. This is the so-called semi-parametric approach (with a parametric link between the two d.f.'s but a non-parametric characterization of their functional forms), and in that way we have a semi-parametric GLM.

Let us denote the two sample empirical d.f.'s by \( F_{nT}(x) \) and \( F_{nS}(x) \) respectively. Then we have

\[
F_{mS}(x) = m^{-1} \sum_{i=1}^{m} I(X_i \leq x), \quad x \in \mathcal{R}^+,
\]

(5.8)

\[
F_{nT}(x) = n^{-1} \sum_{i=1}^{n} I(Y_i \leq x), \quad x \in \mathcal{R}^+.
\]

(5.9)

Then natural estimators of the two -log-survival functions are the following:

\[
B_{mS}(x) = - \log (1 - F_{mS}(x)), \quad x \in \mathcal{R}^+,
\]

(5.10)

\[
B_{nT}(x) = - \log (1 - F_{nT}(x)), \quad x \in \mathcal{R}^+.
\]

(5.11)

A ( zero intercept ) regression line of \( B_{nT}(x) \) on \( B_{mS}(x) \) can be fitted by appropriate statistical methodology, and the slope of this observed regression line provides the estimate of \( \rho \). The basic statistical question: Which, if any, is the most appropriate method of fitting such a regression line? To address this issue, we may note that both the log-survival functions in (5.9) are defined for \( x \in \mathcal{R}^+ \), they are nondecreasing with starting value 0 at \( x = 0 \) and have jump discontinuities at the respective sample order statistics which are distinct with probability one. Moreover both
are stochastic, so that we have a case where the dependent as well as the independent variates (functions) are subject to errors. We introduce two empirical processes \( \{W^{(1)}_m(x), x \in \mathcal{R}^+\} \) and \( \{W^{(2)}_n(x), x \in \mathcal{R}^+\} \) as follows:

\[
W^{(1)}_m(x) = [B_{ms}(x) + \log \bar{F}_S(x)], x \in \mathcal{R}^+,
W^{(2)}_n(x) = [B_{nT}(x) + \log \bar{F}_T(x)], x \in \mathcal{R}^+.
\] (5.12)

Borrowing the reverse martingale (process) characterization of the empirical d.f. (with respect to the index \( m \) (or \( n \)) and noting that \(-\log(1 - y)\) is a convex function of \( y \in (0, 1)\), we claim as in Sen (1973) that both the empirical processes defined by (5.10) are reverse submartingale processes (with respect to the sample size sequences) and that they are independent. Although such characterizations enable us to propose some ad hoc estimators of \( \rho \), they do not lend to any simple (and yet efficient) solution.

There appears to be a rather simple approach based on the conventional \( Q-Q \) plot. Note that under the proportional hazard assumption,

\[
\bar{F}_T(F_s^{-1}(1 - p)) = [\bar{F}_S(F_s^{-1}(1 - p))]^p = p^\rho, \forall p \in (0, 1).
\] (5.13)

Keeping this identity in mind, we may introduce the sample functions

\[
Z_{m,n}(p) = \log(\bar{F}_{nT}((F^{-1}_s(p)))), p \in (0, 1).
\] (5.14)

At any specific value of \( p : 0 < p < 1, Z_{m,n}(p)/\{\log p\} \) is a consistent estimator of \( \rho \). Therefore homogeneity of such pointwise estimates (over a set of points inside \((0,1)\)) conforms to the adequacy of the proportional hazard model, and then these may be pooled to obtain an efficient estimator of \( \rho \). In this context the choice of a set of gridpoints (for \( p \)) or \( p \) in continuum in a compact subinterval of \((0,1)\) merits a careful study, and the related asymptotic theory is planned to be considered in a subsequent communication.

6 GLM IN INDIRECT QUANTITATIVE ASSAYS

In an indirect quantitative assay, doses at several levels are prescribed to several subjects, and their responses, quantitative in nature, are recorded. Therefore, for a specified dose, say, \( d(>0) \), the stochastic response, say, \( Y_2 \) follows a d.f., say, \( F_2(x) \), and an average (mean, median or some other measure of the central tendency) define a dose-response regression. Generally, this regression line is nonlinear, and suitable transformations on the dose and/or response variates are made to reduce this sensibly to a linear form. Such transformations are termed the dosage (or dose-metameter) and response-metameter. In a parametric analysis, it is therefore assumed that the response-metameter has a linear regression on the dosage with a distribution of the errors of a specified form (involving some unknown parameters). Standard parametric procedures are then usually employed to estimate the parameters appearing in the regression form as well as the nuisance parameters associated with the tolerance d.f.'s. Under the basic assumptions that the underlying d.f.'s are all (homoscedastic) normal laws, detailed statistical analysis schemes and their properties are given in Finney (1978).

Parametric models based on transformations on the dose/response variates and a specified form of the tolerance distribution brings the direct relevance of the GLM, particularly, when this distribution belongs to the exponential family for which the exact GLM methodology sketched in Section 2
remains adoptable. In that set up, we may therefore claim some optimality properties of such GLM based solutions. Thus under the normality assumption, the normal theory model and analysis schemes considered in detail in Finney (1978) are very appropriate from statistical point of view. In the framework of earlier sections, we suppose that a standard and a test preparation are administered to a set of subjects, and we conceive of linear dose-response regressions along with a specified functional form of the error d.f. which involves only a finite number of nuisance parameters. These two dose-response regression lines are denoted by

\[ Y_S = \alpha_S + \beta_S x + e_S \quad \text{and} \quad Y_T = \alpha_T + \beta_T x + e_T. \] (6.1)

There are two popular models of indirect assays: (i) parallel line assays and (ii) slope ratio assays. The former arises commonly when a logarithmic transformation is used, while the other one is attuned to power-type transformations.

In a parallel line assay, the two regression lines are conceived to be parallel, so that we have that both \( e_S \) and \( e_T \) have a common d.f., and moreover

\[ \beta_S = \beta_T = \beta \quad (\text{unknown}); \quad \alpha_T - \alpha_S = \beta \log \rho, \] (6.2)

where \( \rho(> 0) \) is the relative potency of the test preparation with respect to the standard one. (6.2) is referred to as the fundamental assumption in this parallel line assay. Note that for the estimation of the common \( \beta \), we can adopt the standard linear model approach based on the classical least squares estimators, and this is optimal when the error distributions are assumed to be normal. The situation with \( \rho \) is somewhat different. Note that \( \log \rho \) can be expressed as the ratio \( (\alpha_T - \alpha_S)/\beta \), so that the least squares estimators of these three parameters can be used to yield a point estimator. But when it comes to derive a suitable confidence interval for \( (\log) \rho \), we encounter the problem of having an estimator of two statistics which are both linear (and normally distributed under normality of the errors). Asymptotically these two estimators are normal when properly standardized, so that a similar problem arises in the estimation of the standard error of \( \log \rho \) or in attaching a confidence interval for the parameter \( \log \rho \). In this context, the classical Fieller's theorem is generally adopted in a parametric setup. This is clear from the above discussion that such estimation procedures are generally nonrobust to possible departures from the assumed normality of the errors. Moreover generally the estimating equations are complex. However, allowing possible differences in the dose levels for the test and standard preparations and choosing a geometric progression for each, we obtain an equispaced design points for the dosage, and in this sense, symmetric 2k-points designs are popular and convenient for such assays. Since the dosage levels span suitable ranges, the issue of robustness merits a careful examination. Robust and nonparametric methods are therefore appealing in this respect. Sen (1971) considered some simple robust estimation and testing procedures for such parallel line assays, and these are further reviewed in Sen (1984).

The situation is somewhat different for the slope ratio assay. In such a case, referred to the linearized dose-response regression mentioned before, we have

\[ \alpha_T = \alpha_S = \alpha \quad (\text{unknown}) \quad \text{and} \quad \beta_T = \rho^\lambda \beta_S, \] (6.3)

where \( \lambda > 0 \). Moreover, in this case, the two errors \( e_S \) and \( e_T \) may not have the same distribution even if they are normal. Thus, the homoscedasticity condition does not hold (unless \( \rho = 1 \)). Note that here the equality of the intercept parameters \( \alpha_T, \alpha_S \) constitutes the fundamental assumption, and \( \rho = (\beta_T/\beta_S)^{1/\lambda} \) is nonlinear even if \( \lambda \) is specified. For this reason in attaching a confidence interval for the relative potency or even estimating the standard error of the estimated \( \rho \), there may not be an optimal procedure, even asymptotically. Fieller's theorem can be used with some advantage, but there is no guarantee that the roots are both real. In this setup because of the common intercept it is customary to have a \((2k + 1)\)-point design, and standard normal theory analysis schemes are
reported in detail in Finney (1978). Scope for vulnerability to non-normality and heteroscedasticity of the two error distributions is even more in slope ratio than in parallel line assays, and hence, an unconditional surrender to the standard GLM methodology is not advocated. Sen (1972a, b) contain the methodological development of suitable robust and nonparametric inference procedures which retain the basic GLM structure sans the normality or other specific forms of the error distributions; these were further unified in Sen (1984). We therefore skip these details.

In the preceding section, we have briefly outlined the PHM arising in direct dilution assays. Such PHM can also be imported for the indirect assays discussed above. If we assume that both the response distributions (for the test and standard preparations) are absolutely continuous so that their hazard functions \( h_T(y|x) \) and \( h_S(y|x) \), corresponding to a dosage level \( x \), are defined appropriately for dosage \( x \) belonging to a domain \( X \), then treating the dosage levels as design variates (i.e., nonstochastic), we may formulate a PHM as in Cox (1972). We consider an arbitrary (baseline) continuous, nonnegative hazard function \( h_b(y), y \in Y \) and express

\[
\begin{align*}
  h_S(y|x) &= h_b(y) \cdot \exp\{\alpha_S + \beta_S x\}, \\
  h_T(y|x) &= h_b(y) \cdot \exp\{\alpha_T + \beta_T x\},
\end{align*}
\]

(6.4)

for \( x \in X, y \in Y \). Note that in (6.4) if we work with the log-hazard functions we end up with a parametric linear regression part and a nonparametric component i.e., \( \log h_b(y) \) which does not depend on the concomitant variate(s). In this setup, it is also possible to introduce other design variables as concomitant ones, so that \( x \) may as well be taken as a vector and the corresponding \( \beta \)'s are to be taken as vector too. This formulation allows us to treat the baseline hazard function as arbitrary while the effects of the design variates and the test vs. standard (i.e., treatment) variate are taken as covariates in a simple parametric structure. For example, in a parallel line assay we may as before take the regression slopes to be equal and define the (log-)relative potency as the ratio of the difference of the intercepts and the common slope. A similar formulation can be made for the slope ratio assay. That way the relative potency can be defined and interpreted in terms of the hazard-regression parameters (on the covariates), treating the base line hazard as a nuisance parameter (functional). This formulation paves the way for semi-parametric approaches to statistical modeling and analysis of indirect quantitative bioassays, and in a more sophisticated manner appropriate counting processes with stochastic intensity functions can be incorporated to impart refined statistical tools. We may refer to the recent encyclopedic volume by Andersen et al. (1993) for an extensive treatment of this subject matter. The intricate relationship between the GLM and PHM (or semi-parametric) approaches becomes quite apparent if we probe into the corresponding estimating equations which resemble the GEE in a visible manner. The basic difference here is the nuisance parameter (functional) \( h_b(y), y \in Y \) which calls for a partial likelihood approach, and this leads to some loss of information resulting in some sacrifice of efficiency. On the other hand, this approach lends itself in a natural way to random censoring schemes which may arise in practice due to withdrawal or dropout of the subjects due to causes other than the failure or the primary end-point of the investigation, though the loss of efficacy can be quite severe under heavy censoring. Of course, in such censored models, almost all the procedures used in practice are subject to the same drawback, and hence, such a criticism should not be labelled to the PHM alone. Nevertheless, we take this opportunity to point out also the nonrobustness aspects of such refined statistical modeling and analysis schemes.

It follows from (6.4) that for parallel line assays, under the PHM, for every \( x \in X, y \in Y \),

\[
\frac{h_T(y|x)}{h_S(y|x)} = \exp\{\beta \log \rho\} = \rho^\beta
\]

(6.5)

which does not depend on the level \( y \). On the other hand, by the fundamental assumption of parallel line assays, we have

\[
\begin{align*}
  h_T(y|x) &= f(y - \alpha_T - \beta x)/\tilde{F}(y - \alpha_T - \beta x),
\end{align*}
\]

13
= f(y - α_s - βx - β\log ρ)/\hat{F}(y - α_s - βx - β\log ρ)
= h_s(y - β\log ρ | x), y ∈ \mathcal{Y}, x ∈ \mathcal{X}.

(6.6)

Therefore from the last two equations we conclude that in order that a PHM is appropriate to the parallel line assay model, we need that

h_T(y | x)/h_s(y | x) = ρ^β = h_s(y - β\log ρ | x)/h_s(y | x),

(6.7)

for every x ∈ \mathcal{X}, y ∈ \mathcal{Y}. This in turn implies that

h_s(y - β\log ρ | x) = ρ^β \cdot h_s(y | x), \forall x ∈ \mathcal{X}, y ∈ \mathcal{Y}.

(6.8)

Therefore, we need to have

h_s(y - \log c | x) = c \cdot h_s(y | x), \forall c > 0, y ∈ \mathcal{Y}, x ∈ \mathcal{X}.

(6.9)

In view of the fact that in a parallel line assay we have usually a log-dose and log-response transformation, the domain \mathcal{Y} is the entire real line, so that the last equation does not hold for most of the common tolerance distribution. A similar picture holds for the slope ratio assay. Therefore, in order that a PHM is deemed appropriate in an indirect quantitative bioassay, we may have to forego the usual fundamental assumptions of parallel line or slope ratio assays. Although alternative modeling may be possible, that would require a different interpretation of relative potency (in terms of relative risks), and such a definition may also entail a dependence of this risk ratio on the level of dosage. In clinical trials where such PHM's are commonly adopted, a basic requirement is that the two hazard functions do not cross each other, so that one survival function lies under the other for all y, x. Again this basic condition may not be generally true; we may refer to Sen (1994b) for some illustrative examples. If our basic goal is to investigate the relative risks of a standard and a test preparation over a domain of the dosage, then more general nonparametric modeling can be adopted to achieve this goal, although this may entail a very large sample size requirement. We will discuss this later. But, the sensitivity of the PHM in this context to model departures is quite clear from the above discussion. Even a small departure can alter the statistical picture and invalidate the derived conclusions to a greater extent.

7 GLM IN NONLINEAR MODELS

In the context of bioassays, we have seen that dose- and response-metameters often reduce the dose-response regression to a sensibly linear form, so that standard least squares theory or GLM can be adapted with some success. From robustness considerations, however, there may be some genuine concern about the distributional assumptions to be made and impact of their departures from the true (unknown) situation. This problem is much more complex in a general nonlinear model, even when the nature of the function is assumed to be of a given form (involving some unknown parameters or algebraic constants). In such a setup, we assume that the response variate Y is related to a set of auxiliary variables, designed by a vector x belonging to a suitable space \mathcal{X}, by means of a stochastic model

Y = g(x) + \epsilon, x ∈ \mathcal{X},

(7.1)

where the error component \epsilon has a given distribution F involving possibly some unknown parameters while the form of the regression function g(.) is assumed to be given, although it may also involve some unknown parameters. In this formulation we have a predominant parametric flavor. A semi-parametric formulation can be conceived of by rendering either the form of g to be arbitrary but F
to be given, or $g$ to be of a given form but $F$ is allowed to be arbitrary. A complete nonparametric modeling relates to the situation where both $g$ and $F$ are arbitrary. The nature of the postulated $g$ thus adds a second layer of nonrobustness prospects to GLM based parametrics.

To link such nonlinear models to the bioassay models treated in earlier sections we consider again the parallel line assays in a more general mold of nonlinear dose-response regression. We stick to the same notations as far as possible. Thus, we assume that at a given dosage level $x$, the d.f. of $Y_T$ and $Y_S$, denoted by $F_T$ and $F_S$ respectively, are given by

$$F_T(y|x) = F(y - m_T(x)), \quad F_S(y) = F(y - m_S(x)), \quad y \in \mathcal{Y}$$  \hspace{1cm} (7.2)

where $F$ has a specified functional form (for example, normal, logistic, Laplace etc), and the two regression functions $m_T(x)$ and $m_S(x), \ x \in \mathcal{X}$ are of given (nonlinear) forms, such that

$$m_T(x) - m_S(x) = \alpha^*, \ \forall x \in \mathcal{X}.$$  \hspace{1cm} (7.3)

This two-sample model can be regarded as a particular case of a more general model wherein we may introduce some other design (or dummy) variables (vectors) $c_i, i = 1, \ldots, n$, and for the ith response $Y_i$, we frame the regression model:

$$Y_i = g_1(c_i) + g_2(x_i) + \epsilon_i, \ i = 1, \ldots, n,$$  \hspace{1cm} (7.4)

where the errors $\epsilon_i$ are assumed to be i.i.d.r.v.'s with a d.f. It may not always be wise to impose this i.i.d. structure on the errors. Even in the case of simple linear models, heterscedasticity of the errors may have a profound effect on the optimality properties of common statistical inference tools and this can be more significant for GLMs; we refer to Carroll and Ruppert (1988) for some useful discussions on this nonrobustness properties $F$ of given form, and the functional forms of $g_1$ and $g_2$ are also assumed to be given, all involving some unknown parameters. It is also possible to combine $g_1$ and $g_2$ into a more complex function $g(c, x)$. In this context we may note that under the i.i.d. structure of the $\epsilon_i$, we may consider the total sum of squares due to error:

$$\sum_{i=1}^{n} \left\{ Y_i - g_1(c_i) - g_2(x_i) \right\}^2,$$  \hspace{1cm} (7.5)

which for given $Y_i, c_i, x_i, i = 1, \ldots, n$ becomes a function of the unknown parameters associated with the functional forms of $g_1$ and $g_2$. Thus minimizing the above norm with respect to these unknown parameters, we obtain a set of estimating equations which lead to their generalized least squares estimators. Such estimators are linear in the $Y_i$, but not generally in the $c_i$ or $x_i$, and hence may be quite cumbersome to solve algebraically. The situation may become even more complex when we combine the two functions $g_1, g_2$ into a single $g(c, x)$. Moreover depending on the nature of such $g$'s, the resulting LSE may be (often highly) nonrobust. Such nonrobustness arises due to possible departures from these assumed functional forms, and also from possible departures from the distributional assumption on the errors. In the literature usually such error distributions are taken to be normal; that leads to the heterscedasticity condition as a by-product, and hence in the above equation the unweighted sum of squares is justifiable. On the other hand, sans this normality of the errors, the heterscedasticity condition can not be taken for granted, and hence the sensitivity of these LSE to possible heterscedasticity of the errors remains of serious concern.

From the above discussion it appears that there is a genuine need for appropriate GLM methodology that can be adopted for nonlinear models. The principal difficulty stems from the fact that for general nonlinear models, particularly involving vector-valued auxiliary variables, finding an appropriate link function may be hard, and the solution for a canonical link function may even be harder. This puts the general GLM methodology in a rather awkward situation, and on the top of that
generally these estimates share nonrobustness properties to a greater extent than in the simpler cases treated in earlier sections. Quasi-likelihood estimators discussed in Section 3 [see (3.11)] have a better appeal in this case, but again their estimating equations may be too complex. Robustness concerns remain pertinent in this setup too. Judged from all these aspect we may conclude that whenever we have confidence in an assumed functional forms of the \( g_r, r = 1, 2 \), but we are not so confident about the particular distributional assumption made on the errors, a semi-parametric approach may work out better for large sample sizes. Of course, if the sample size is large enough, one may even go for a complete nonparametric solution which will have better robustness properties. We will discuss this aspect of semi-parametrics and nonparametrics in the next section, and hence, we avoid the repetition. Nevertheless, we may comment that for all such estimators, exact or finite-sample properties are generally hard to study, and hence the exact sample properties of the GLM methodology are not traceable in such an asymptotic setup. Of course, the GEE methodology is more appropriate here in an asymptotic setup.

8 GLM AND NONPARAMETRIC REGRESSION

The basic attraction of GLM's is the scope of their exact (statistical) treatment through the choice of appropriate link functions or other transformations. On the other hand such a choice usually demand some specific parametric structures, and hence, it may be desirable to check whether or not such parametric models are appropriate in the given context. For example, referring back to the bioassay problems, say in the quantal response case, we may use a logit or probit model under the assumption that the tolerance distribution is logistic or normal with a linear regression on the dosage. Thus not only we make specific distributional assumptions, but also assume linearity of regression. This may open up the avenues for lack of robustness on both of these counts. As such, it may be desirable to relax such stringent parametric conditions by less restrictive ones. Two specific types of models evolve in this connection:

1) Nonparametric regression models. No specific parametric structure is imposed on the tolerance distribution or the dose-response regression function.

2) Semi-parametric regression models. Here for a part of the model, some parametric structure is taken for granted while for the complementary part a nonparametric modeling is adopted. For example, in the quantal bioassay model, we may assume that the tolerance distribution is logistic or normal, but the dose-response regression is arbitrary (possibly monotone and smooth in a well defined sense). Alternatively, we may also take the tolerance distributions to be arbitrary (continuous) but the dose response regression to be of a specified parametric form. The classical Cox (1972) PHM also belongs to such semi-parametric family of models.

The choice of a semi-parametric or nonparametric regression model is largely governed by the confidence of the experimenter on plausible parametric structures underlying the experimental setup. From robustness considerations it may be preferable to choose a nonparametric model rather than a semi-parametric one. On the other hand, the very generality of a nonparametric approach may lead to a so-called large parameter space or a functional (i.e., infinite-dimensional) parameter space, and with the increase in the dimension of such parameter spaces, the precision of statistical conclusions attainable in a finite sample rapidly goes down. Thus, generally, a nonparametric approach may require a larger sample size to match the precision level attainable by a semi-parametric model, whenever the latter one is the true model. On the other hand, if the chosen model is not the true one, a semi-parametric model may result in significant bias and this may tilt the picture in an opposite direction.
It is not uncommon to have a set of auxiliary variables for which some parametric models may appear reasonable, while there may be others for which a nonparametric one would be more appropriate. For example, we may refer to the regression model in (7.4). Since there the design variables \( c_i \) are nonstochastic, with suitable transformations, a linear form for the regression function \( g_1(c_i) \) may appear to workable in some situations. On the other hand, for the stochastic variates \( x_i \), such a parametric regression function may not be that reasonable, and hence, we may prefer a complete nonparametric form for the regression function \( g_2(x) \). In conventional experimental designs, it is not uncommon to have some "fixed-effects" parameters and some other as "random-effects" ones. For example, in animal studies, if the blocks are formed by common mothers, the block effects are treated as random while the treatment effects are treated as fixed one. This results in a so-called mixed-effects models. In a traditional setup, the random effects are assumed to have (joint) normality with specific covariance-structures, and this allows one to use conventional normal theory inference tools for drawing statistical conclusions. But all these assumptions require careful scrutiny before one adopts such classical models in practice. The discussion made above pertains to GLM as well. In a GLM we may accept the GL part to a certain extent, and yet may not like to make specific distributional assumptions, so as to make our conclusions more robust to plausible departures from conventional parametric models. To illustrate this point, we refer to the PHM treated in (5.4)-(5.6). We deal with this model in a more general manner by incorporating nonstochastic design variates \( c_i \) and stochastic concomitant variates \( x_i \), along with the primary response variate \( Y_i \), for \( i = 1, \ldots, n \). Let \( f(y|c, x) \) be the conditional density of \( Y \), given \( c, x \), and the corresponding conditional hazard function be denoted by \( h(y|c, x), y \in \mathcal{Y}, c \in \mathcal{R}^p, x \in \mathcal{R}^q, p, q \geq 0 \). In a PHM we set

\[
h(y|c, x) = h_o(y) \cdot \exp \{ \beta' c + \gamma' x \}, \quad y \in \mathcal{Y}, \; c \in \mathcal{R}^p, \; x \in \mathcal{R}^q,
\]

where \( h_o(y) \) is the baseline hazard function and is treated in a nonparametric fashion as arbitrary, nonnegative and continuous function of \( y \) alone, and \( \beta \) and \( \gamma \) are unknown regression parameters (vectors). Thus the regression function is taken to be of a parametric form while \( h_o(\cdot) \) is nonparametric, so we have a semi-parametric model. From the above equation, we have

\[
\log h(y|c, x) = \log h_o(y) + \beta' c + \gamma' x, \quad y \in \mathcal{Y}, \; c \in \mathcal{R}^p, \; x \in \mathcal{R}^q,
\]

so that we have a GLM, albeit in a semi-parametric form. The ingenuity of Cox (1972) lies in the innovative incorporation of the partial likelihood principle which allows the baseline hazard to be treated as a nuisance parameter (functional). This results in a manageable statistical analysis with some loss of efficiency due to the nonparametric flavor of \( h_o \). This certainly has been a breakthrough in statistics, and the past twenty years have witnessed phenomenal growth of statistical research literature on such semi-parametric models where GLM and survival analysis have found a melting pot to blend harmoniously. The recent book by Andersen et al. (1993) is a classical treatise of this subject matter in a much more general setup, and the findings are full of promise for diverse applications in diverse fields. Nevertheless, the limitations of this approach in real applications should not be overlooked. With our primary emphasis on applications, we will stress these points a bit more elaborately as follows.

The very basic assumption of the proportionality of the hazards in (8.1) with adjustments for the auxiliary and concomitant variates in a parametric form has been a source of serious concern in many studies. Sen (1994b) has discussed one of the problems arising in the context of possible congestion of fatty substance in the upper aorta resulting in reduced oxygen flow in the brain area. This problem is common in old age, and medication and surgery are two popular medical devices to correct the problem to the extent possible. However, the risk of surgery and its aftermath may set the hazard function in a different form so that its proportionality with the hazard function in the medication protocol is generally not true. In general if the hazards are criss-crossing, the PHM may not appear to be that appropriate, and statistical conclusions based on such a stringent assumption

17
may not be valid or efficient. In some other problems the covariates may be time-dependent, and that may introduce additional complications. One possible way to handle such complications is to consider suitable time-dependent coefficients models where the regression parameter $\gamma$ is allowed to depend on time, so are the concomitant variates. With this generalization it appears that the assumption of PH becomes somewhat unnecessary [see, for example, Murphy and Sen (1991)], so the scope and validity of statistical analysis are enhanced. On the other hand, such analyses are generally more complex, and may also require comparatively larger sample size, a condition that may not be met universally in practice. Rank analysis of covariance incorporating progressive censoring may often provide an alternative solution where the emphasis on PHM or even the GLM component may not be that crucial. For some detail account of this methodology, we may refer to Sen (1979) or Sen (1981, ch.11). In particular, if we use the so-called log-rank statistics then the end product is very much comparable to the PHM when the PH assumption holds.

The nonparametric regression approach has a natural appeal. It is quite intuitive to define for a bivariate random vector $(Y, X)'$, the regression of $Y$ on $X$ as

$$m(x) = E\{Y|X = x\}, \quad x \in X.$$  \hspace{1cm} (8.3)

This definition extends naturally when $X$ and/or $Y$ are vector valued. For simplicity of presentation, we consider the case of real valued $Y$ and possibly vector valued $X$. Let $F(y|x)$ be the conditional d.f. of $Y$, given $X = x$. Then $m(x)$ is a functional of the conditional d.f. $F(.,|x)$. Thus whenever this conditional d.f. can be conveniently estimated by an empirical d.f., say, $\hat{F}_n(.,|x)$, we may be tempted in using the same functional for this empirical d.f. to estimate $m(x)$. There are two basic problems associated with this approach. First, unlike the case of real valued r.v.'s, there may not be a natural estimator of $F(.,|x)$ having some optimality properties. Second, the functional $m(x)$ as defined above is not so robust against plausible departures from the model based assumptions. There are at least two well adopted methods to take care of the first problem, while considerations of robustness and efficiency lead us to some other (possibly nonlinear) functionals which have some nice properties. We will discuss these further here.

For a set value $x_o$ of the auxiliary variate, and a chosen metric $d: \mathcal{R}^q \times \mathcal{R}^q \rightarrow \mathcal{R}^+$, consider the set of pseudovariables:

$$Z_i(x_o) = Z^*_i = d(X_i, x_o), \quad i = 1, \ldots, n.$$  \hspace{1cm} (8.4)

Let us denote the corresponding ordered observations by

$$Z_{n:1}^* \leq \cdots \leq Z_{n:n}^*,$$  \hspace{1cm} (8.5)

and note that if the marginal d.f. of $X$ is continuous, as we usually assume, then ties among the $Z_{n:i}^*$ can be neglected with probability one. Further we choose a sequence $\{k_n\}$ of positive integers such that $k_n$ is nondecreasing in $n$, with

$$\lim_{n \rightarrow \infty} k_n = \infty, \text{ but } \lim_{n \rightarrow \infty} n^{-1}k_n = 0.$$  \hspace{1cm} (8.6)

The specific choice of $k_n$ depends (among other things) on the dimension ($q$) of the $X_i$. For $q = 1$, we take $k_n = O(n^{q/2})$. Moreover we introduce a set of indices $S = (S_1, \ldots, S_n)$ by letting

$$Y_{S_i} = Y_j \text{ if } Z_{n:i} = Z_j, \quad \text{for } i, j = 1, \ldots, n.$$  \hspace{1cm} (8.7)

Then in a (KNN) $k$-nearest neighborhood method, we introduce the KNN-empirical d.f. as

$$\hat{F}_{n,k} (y; x_o) = k_n^{-1} \sum_{j=1}^{k_n} I(Y_{S_j} \leq y), \quad y \in \mathcal{Y}.$$  \hspace{1cm} (8.8)

18
Note that the $Y_i$ are conditionally independent but not necessarily identically distributed, so that the above KNK-empirical d.f. is not in general an unbiased estimator of the underlying conditional d.f. $F(.|x_o)$. Moreover, the mean square error of this empirical d.f. is $O(k^{-1}_n)$ which is slower than $O(n^{-1})$. Thus, the built-in bias and reduced rate of convergence are both to be taken into account in studying the properties of any estimator based on such an empirical d.f. A vast research literature relates to the specific estimator (empirical KNK-conditional mean):

$$
\hat{m}_{n,k_n}(x_o) = \int_{-\infty}^{\infty} yd\hat{F}_{n,k_n}(y,x_o), \ x_o \in X.
$$

(8.9)

This methodology extends directly to the GLM context where instead of the $Y_i$ we work with suitable link functions for them. But the main criticism as may be labelled against this estimator is its lack of robustness; it is very sensitive to outliers or error contaminations and also demands appropriate moment conditions on the underlying distribution which may not be easily verifiable in practice. Motivated by this deficiency perspective, more robust estimators have been considered in the literature. A simple formulation is the following. Let $m(x_o)$ be defined as the meadain or a specific quantile of the conditional d.f. $F(.|x_o)$. This is termed a quantile (median) regression or conditional quantile function. Thus one may consider the natural estimator, the corresponding measure of the KNK-empirical d.f. $\hat{F}_{n,k_n}$. More generally one may consider a class of measures

$$
\theta(x_o) = \theta(F(.|x_o)), \ x_o \in X,
$$

(8.10)

where $\theta$ is a suitable (smooth) functional of the conditional d.f. and a function of the argument $x_o$. We can define these as extended statistical functionals. With these notations we can consider an estimator based on the KNK-empirical d.f. as

$$
\hat{m}_{n,k_n}(x_o) = \theta(\hat{F}_{n,k_n}(.|x_o)), \ x_o \in X.
$$

(8.11)

We refer to Gangopadhyay and Sen (1993) and Sen (1994c) where the related literature has been adequately surveyed and suitable references are cited too. The asymptotic flavor of this methodology is quite evident from the above discussion, and resampling plans are often adopted to draw statistical conclusions in a specific context. An alternative to this KNK-methodology is the so-called kernel method which rests on an incorporation of a chosen (sufficiently smooth) kernel density to estimate the conditional density of $Y$, given $x_o$. In order to control bias and balance the mean squared error of derived estimators, here also one needs to choose the band-width for the kernel density small, typically converging to 0 at some power rate $n^{-\lambda}$ which parallels $n^{-1}k_n$. These two methodologies share common objectives and yield very much comparable results (with some trade-off between bias and variance functions), and hence, we would not go into further discussion of this alternative methodology.

The above discussion pertains to a simple multivariate model where there are no design (non-stochastic) variates. In the context of GLM such design variates are commonly encountered, and in addition we may have stochastic regressors which may qualify as concomitant or covariates (in the sense that their distribution remains unaffected by the variation of the design variate). Thus MANOCOVA (multivariate analysis of covariance) models are also relevant to GLM whenever the primary response variate(s) are amenable to appropriate link functions. However, in this context, specific distributional assumption on such conditional laws of the primary variates, given the concomitant variates, may not be that reasonable, and hence, from robustness considerations, a semi-parametric or nonparametric formulation merits serious considerations. For a complete non-parametric formulation involving conditional quantile functions and statistical functionals we may refer to Sen (1993) where other pertinent references are also cited. The more interesting point is that often we have a (generalized) linear structure bestowed by the design variates while a nonparametric component due to the concomitant variates. This leads us to the so-called mixed generalized
linear models (MGLM), which are largely semi-parametric in flavor. If we ignore the concomitant variate(s), we would have a classical GLM, so that the usual GEE based methodology can be used to draw statistical conclusions on the set of parameters in this GLM part. However, we need to keep in mind the following two drawbacks of this marginal GLM approach:

(i) This does not take into account the information contained in the covariates, and hence, the derived statistical conclusions may not be fully efficient.

(ii) This approach does not provide any information on the interrelationship of the primary and concomitant variates, and hence, the derived conclusions may not have the full scope.

One intuitive way of eliminating these drawbacks is to consider an iterative procedure which we may present as follows:

Step 1. Ignoring the concomitant variates consider a GLM and obtain the usual estimators of the fixed-effects parameters. Usually such estimators are $\sqrt{n}$-consistent, although they may not be fully efficient.

Step 2. Use the Step 1 estimators to align the primary variate observations. Such residuals are denoted by $\hat{Y}_i$, $i = 1, \ldots, n$. Use these residuals in a nonparametric fashion to quantify the regression on the concomitant variates.

Step 3. Incorporate the Step 1 and 2 estimators to formulate an iterative algorithm which results in a more efficient procedure.

The validity and rationality of standard statistical analysis tools may be at stake due to complications arising in this scheme, and hence, we have several points to ponder:

(a) With respect to Step 1, we need to put due emphasis on the distributional robustness aspects primarily because of the nonparametric component of this MGLM.

(b) The residuals or aligned observations in Step 2 are neither identically distributed nor independent in a strict sense. Thus the conventional i.i.d. assumption in a parametric or nonparametric model is not tenable here. Allowing such perturbations, a more refined and often nonstandard statistical analysis is needed in this context.

(c) The nonparametric component in the MGLM introduces an infinite dimension parameter space, so that optimal parametric procedures may no longer be desirable or efficient in the current context. In view of this in Step 3, a different iterative procedure may be advocated in order to enhance the precision of the fixed-parameter effects estimates retaining their $\sqrt{n}$-consistency, while attaining a slower rate of convergence for the nonparametric parameter (functional). The situation is quite comparable to a (factorial) experimental design where under (partial) confounding, a possibly different rate of convergence is achieved for the estimates of the parameters of primary interest and for others of secondary importance.

In the specific case of ANOCOVA mixed-model, dealing with M-, R- and regression rank scores estimators, such nonstandard analyses were prescribed by Sen (1995a, 1996). As long as a MGLM can be reduced to a mixed-effects MANOCOVA by the choice of suitable link functions, such techniques can be directly adopted in that context too. As such, we briefly summarize the findings as follows. In Step 1, we follow standard nonparametric or robust procedures for linear models (ignoring covariates); we may refer to Jurečková and Sen (1995, chs. 3-7) for detailed accounts of such procedures. Note that these estimators are robust, $\sqrt{n}$-consistent and efficient (at least asymptotically) with respect to the marginal model. In Step 2, using the residuals from Step 1 (which are subject to perturbations of the order $O(n^{-1/2})$), the KNN- (or kernel-) method is incorporated to estimate the regression functional on the concomitant variates in a nonparametric fashion. This follows along the lines described earlier this section. The trick is to verify that $O(n^{-1/2})$ perturbations in the residuals have no (asymptotic) effects on the properties of such nonparametric or robust functional estimators (which have a slower rate of convergence). In Step 3, the idea is to choose a compact set $C$ for the variation of $x_n$ (as can be done from extraneous and experimental considerations), to divide the set $C$ into a number (say, $r_n$) of subsets, and to choose $r_n = o(n/k_n)$. Then for each subset, we will have a (random) number of observations, on which a Step 1 procedure
can be incorporated to have estimators of the finite-dimensional fixed-effects parameters. Note that Anscombe-type uniform consistency in probability results are generally needed to approximate the random sample size situation by suitable nonrandom sample size situations, and weak convergence results play a basic role in this context. The classical weighted least squares methodology is then used to yield a pooled estimator which will be (asymptotically) at least as efficient as the original Step 1 estimator (if no such partitioning were made). Finally, for each of these compact subsets, we may use the KNN- or kernel-method in estimating the nonparametric regression functional, and these estimators can then be smoothed out if necessary by usual local smoothing techniques. We refer to Sen (1995a, 1996) for details.

There are some other smoothing techniques, such as the splines or wavelets, which are potentially applicable in the above context too. However, typically such a procedure is based on relatively more flexible model assumptions, and thereby encounter an even slower rate of convergence. This drawback generally renders them as primarily descriptive devices, and valid statistical conclusions based on such tools may therefore require an enormously large sample size, often beyond the scope of usual biomedical studies.

9 GLM: PERSPECTIVES AND CONTROVERSIES

Biomedical applications are generally characterized by some additional complications due to various plausible departures from the ideal experimental conditions. For example, in animal studies, the assumption of independence of the observable random elements may not be strictly true due to possible "litter" (or genetic) effects; in clinical trials or medical investigations, subjects may drop out (withdraw) from the study either due to migration or failure due to causes other than the primary one under study. This may lead to censoring and/or a competing risk model. Missing data may also crop up due a variety of other causes. Misclassification of states is another possibility in such studies. Multiple endpoints or multivariate responses are quite common in biomedical investigations. The primary as well as auxiliary or concomitant variates may be subject to measurement errors, and this may vitiate the usual (simple) statistical models and call for more complex analysis schemes. Finally, there are often some natural restraints on the parameter space, so that the GLM methodology should address to them in a very congenial manner. Exact treatment for such extended GLM analysis tools may become harder, and even the related asymptotics may become more complex. Motivated by these possibilities, we examine here the GLM adoptibility picture in a broader perspective, and discuss the usual controversies surrounding such applications of GLM procedures.

Competing risks models pose some interesting problems of identifiability even in most simple parametric setups, and the situation may become much more complex for GLMs in this setup. Apart from this basic issue, GLM based methodology in competing risks models may be generally highly nonrobust to possible departures from model assumptions. In biomedical applications, as we have stressed in earlier sections, such stringent parametric models may not be very appropriate, and hence, GLMs may not be that appealing in such applications too. With this remark, let us start with the most common problem caused by some type of censoring. Among these types, the most commonly ones encountered in statistical practice are the truncation (type I censoring), type II censoring, progressively censoring, random censoring and interval censoring schemes. Statistical perspectives and controversies in diverse censoring schemes with due emphasis on biomedical and reliability studies have been discussed in detail by Sen (1995b). In censoring schemes, particularly, in the context of random censoring, the Cox (1972) PHM has been extensively used with justifications
from a GLM methodology. However, there is a crucial need to examine the situation more critically with reference to the adoptibility of the GLM methodology when some of the basic model-assumptions are not strictly tenable. Apart from the basic proportionality of the (conditional) hazard functions, (given the concomitant variates), the other fundamental assumption in this context is the following:

*The censoring variable and the primary endpoint are stochastically independent, and further the distribution of the censoring variable is the same across the experimental setup, so the treatment vs. censoring interaction can be neglected.*

In practice, the above made independence assumption may not be very realistic, while the partial likelihood approach may not work out when this independence condition is not true. We have already discussed in an earlier section that when the PHM may not be that reasonable, one may consider a time-dependent coefficient model, and that may at least theoretically work out better. Technically, such a time-dependent coefficient model [see for example, Murphy and Sen (1991)] involves functional regression parameters, so the parameter space becomes infinite dimensional. Therefore a discretized parameter space is created by suitable binning of the time-interval in such a way that the dimension of the discrete parameter space is large but not so much compared to the total number of observations. This renders the estimability of the discretized parameters which in turn provide (via suitable smoothing device) estimates of the parameter functionals. This process may not only have comparable larger variability of the estimators but also nonnegligible bias due to discretization and smoothing. Often, integral transformations on the parameters provide suitable characterizations and lead to better convergence rates for the estimators. However, the sample size requirement may become too stringent in a practical context. Even with such an extended model, in the treatment of censoring the independence assumption remains crucial. The terminology informative censoring has been introduced in the statistical literature to incorporate possible statistical information in the censoring variables through suitable modifications of the usual independence assumption. We refer to Wu and Carroll (1988) and Wu and Bailey (1989) for some useful contributions in this area. Nevertheless, it may be pointed out that such alternative procedures takes away the GLM formulation to a greater extent, and hence in the current context, we would not go into their details.

An intermediate situation may arise when the regression coefficients may not be time-dependent but the concomitant variates are time-dependent to a sensible extent. Recall that in order to qualify for covariates, such concomitant variates should not have distributions dependent on the related design variates. Thus, while we relax the homogeneity of the covariates over time, we still need their homogeneity across treatment groups. Such a condition may not always hold in practice. As is usually the case, whenever we have stochastic covariates, they may be subject to observational or measurement errors. Moreover they may not be all continuous variables, and even so they may be subject to interval censoring. In the discrete case, such measurement errors may also lead to misclassification of states. A similar problem may crop up with the primary variate. For most of the simple GLMs, such measurement errors and misclassifications can disrupt the scope of exact statistical analyses, and even in the asymptotic case there may be certain complications requiring nonstandard avenues for valid statistical analyses. We illustrate this problem through the conventional Poisson and Logistic regression models.

Suppose that there are observable random vectors \((X_i, Y_i), i = 1, \ldots, n\), where conditional on \(X_i = x\), \(Y_i\) has the Poisson distribution with parameter \(\lambda(x)\), for \(i = 1, \ldots, n\). Assume further that the observed \(X_i\) are subject to measurement errors \(w_i\), which do not depend on the \(Y_i\), so that we write

\[
X_i = x_i + w_i, \quad i = 1, \ldots, n, \tag{9.1}
\]

where the \(w_i\) are independent error components while the true \(x_i\) are not observable. As in Section 2, we may consider here a link function for which we have a GLM. But the regression function for the (canonical) link function on the covariate \(X\) actually relates to the unobservable \(x\), so that we encounter a regression model where both the dependent and dependent variables are subject to
error. For this classical regression problem, even in the conventional linear model, the classical LSE are typically biased, and they may even be inconsistent whenever the measurement errors are not too small relative to the variability of \( X \). In the most simple case, suppose that the \( x_i \) are i.i.d.r.v.'s with \( P\{x = 0\} = 1 - P\{x = 1\} = \pi \) and that the \( w_i \) have the following probability structure:

\[
P\{w_i = 0|x_i = 0\} = 1 - P\{w_i = 1|x_i = 0\} = \pi_{00}; \quad P\{w_i = 0|x_i = 1\} = 1 - P\{w_i = -1|x_i = 1\} = \pi_{10},
\]

(9.2)

where \( \pi_{00} \) and \( \pi_{10} \) are unknown and they may not be equal. Suppose further that

\[
P\{x_i = 0\} = 1 - P\{x_i = 1\} = \pi_o, \quad i = 1, \ldots, n.
\]

(9.3)

Then from the last two equations we have

\[
P\{X_i = 0\} = P\{x_i = 0\} P\{w_i = 0|x_i = 0\} + P\{x_i = 1\} P\{w_i = -1|x_i = 1\} = \pi_o \pi_{00} + (1 - \pi_o)(1 - \pi_{10}).
\]

(9.4)

Similarly,

\[
P\{X_i = 1\} = 1 - \pi_o \pi_{00} - (1 - \pi_o)(1 - \pi_{10}).
\]

(9.5)

Note that the conditional distribution of \( Y_i \), given \( x_i = k \), is Poisson with the parameter \( \lambda_k \), \( k = 0, 1 \), and this provides the access to the exact GLM methodology discussed in Section 2. However, from the preceding equations it follows that the conditional distribution of \( Y_i \), given \( X_i = k \), \( k = 0, 1 \), is a mixture of Poisson distributions with parameters \( \lambda_0, \lambda_1 \) and the probabilities \( \pi_o, \pi_{00} \) and \( \pi_{10} \). This disrupts the usual exponential density structure and introduces more parameters in the model, often raising questions of identifiability of the model parameters for drawing statistical conclusions.

The situation becomes more complex when the \( X_i \) are not binary. Likewise, if the \( Y_i \) are subject to measurement (recording) errors, their (conditional) distribution, given \( x_i \), may no longer be Poisson, and hence, even if the covariates are not subject to measurement errors, the exact GLM structure may not hold. Clearly, when both \( X \) and \( Y \) are subject to measurement errors, we have a much more complex situation, and the GLM based methodology may not be that appealing any more. A very similar situation arises for the logistic regression model where \( Y \) is binary while \( x \) is quantitative, possibly continuous. As is the case in bioassays, the exact dose \((x)\) prescribed to a set of subjects may not tally with the amount actually consumed; in such a case, often, the actual consumption is less than the prescribed amount, so that the \( w_i \) are negative random variables. Moreover, the response may not be spontaneous, and hence, possible latent effects may introduce measurement errors and distort the logistic model to a certain extent. We refer to Copas (1988) and Carroll and Pederson (1993) for some useful studies in this important area of applied research. In biomedical applications mixture models crop up due possible subgroups of the experimental units (e.g., male and female subjects receiving a drug for the treatment of a common disease). Even if the components have distributions amenable to the GLM methodology, their mixture may not be so. Hence, it may be more appropriate to make use of suitable stratified approaches which permit such variations to be adjusted in a suitable manner. Heteroscedastic models are also commonly encountered in biomedical studies when the conditional distributions are not closely normal. The recent monograph of Carroll, Ruppert and Stefanski (1996) is an important addition to the current statistical literature. The limitations of the classical GLM and the need for robust nonlinear methodology are quite apparent from the nature of the current developments in biomedical applications.

Problems related to one-sided hypothesis testing and estimation under suitable parametric restraints arise in practice in various moulds. In biomedical applications, because of various uncontrollable extraneous factors and administrative requirements, such restraints may not be very simpler in structure. For example, a new drug is likely to receive adequate attention only if it performs at
least as well as an existing one, if any, designed for the same treatment, and in addition, it does not have any serious side effect. It may be questionable to use standard linear models in such situations, and even if GLM are used, parametric restraints to be considered along with are generally nonlinear, often defined by inequality constraints. The usual MLE may not retain optimality properties (even asymptotically) in such a model, and hence, restricted (R)-MLEs are advocated which have generally better performance characteristics, at least over the domain of the parameter space where such restraints hold. Nevertheless, computationally such RMLE may be quite cumbersome, and their distribution theory may also be quite complex, even asymptotically. Isotonic regression models have been introduced to extend conventional linear models to cope with some order restrictions. Likewise, isotonic generalized regression models are being used to facilitate the use of GLM under similar restraints on the parameters. While this is certainly welcome in the GLM and quasi-likelihood based statistical inference arena, there are some distinct undercurrents which merit a careful scrutiny in any practical application. First and foremost, in almost all such cases, the exact GLM flavor is missing, and at best, some quasi-likelihood approaches hinging primarily on the asymptotics can be prescribed. From robustness perspectives, it may be desirable to examine the underlying model structure and adopt the ones for which the associated quasi-likelihoods are likely to be less sensitive to small or moderate deviations from such assumed models. Second, the GEEs for such restricted GLM/quasi-likelihood models may generally be highly nonlinear and may need more complex iterative solutions. Standard EM or other algorithms may not automatically work out in such a case, and there is a need to develop other more efficient algorithms. Third, in a linear model setup with order restraints on the parameters, under the normal theory model, only in some simple cases, the exact distribution theory of related test statistics can be obtained, and that too under appropriate null hypotheses only. The nonnull distribution theory is still much less adequately studied than their null hypothesis counterparts (even asymptotically). We may refer to Paula and Sen (1995) where some of these problems have been discussed in detail. Finally, there is still room for further research work on optimality properties of RMLE or tests based on such RMLEs in the context of such GLMs under nonlinear parametric restraints.

Our discussion would remain somewhat incomplete without touching the role of GLMs in the important case of multivariate responses. In biomedical applications though one may have a primary endpoint, typically other variables enter into the picture either as auxiliary or concomitant variates. In the case of concomitant variates, one generally considers a conditional model where the GLM methodology can be adopted in a suitable manner. In the other case, the basic homogeneity property needed in the analysis of covariance approach may not be satisfied, and hence the conventional GLM methodology may not be totally applicable. There are also situations where the primary endpoint may itself be associated with multiple characteristics, resulting in the so called multiple endpoint situation. In some cases, there may be a hierarchy of these response variates, and this may be advantageously taken into account in formulating a step down procedure wherein in each step, conditional on the preceding ones, we have a conditional univariate model where an appropriate GLM can be adopted. We may term such models as quasi-multivariate GLMs. The only modification needed in this context is to have an increasing number of concomitant variates at the successive steps and to check that the conditional variances or scale factors satisfy the basic requirement of the exact GLM methodology. An illustrative example is the multivariate normal case where such conditional distributions are all normal with linear regressions satisfying the homoscedasticity condition. However, it may be noted that in biomedical applications the multinormality assumption may rarely be appropriate. In the univariate case, we have discussed some appropriate transformations which may render the normality of the transformed variate to be more tenable. In the multiresponse case, departures from normality may not only come from the marginal distributions but also from the association or dependence pattern of these responses. Therefore, a multinormality inducing transformation may depend simultaneously on all the coordinate elements and may be highly nonlinear in nature. To illustrate this point, we consider the simple case of binomial to multinomial distri-
butions which belong to the exponential family for which GLMs are sought to be appropriate. Let \( X = (X_1, \ldots, X_k)' \) be a \( k(\geq 2) \)-vector having a probability law on the \( k \)-dimensional simplex. Thus, we may write
\[
P(X = x) = \exp \{ \pi' x \}, \tag{9.6}
\]
where \( x \) can have only \( k \) possible outcomes \( (\delta_{i1}, \ldots, \delta_{ik})' \), \( i = 1, \ldots, k \), with the Kronecker delta \( \delta_{ij} \) defined as 1 or 0 according as \( i = j \) or not, and \( \pi = (\pi_1, \ldots, \pi_k)' \) is a vector of nonnegative probabilities adding up to 1. Suppose now that we have \( n \) i.i.d.r.v.'s following the above probability law, and we denote by \( n = (n_1, \ldots, n_k)' \) where \( n_j \) is the number of subjects whose \( j \)-th coordinate is equal to 1 (and others 0), for \( j = 1, \ldots, k \). Then \( n = n_1 + \cdots + n_k \), and the associated probability law is the classical multinomial:
\[
P(n = r) = \exp \{ \pi' r + c_n(r) \}, \quad r_j \geq 0, \quad j = 1, \ldots, k; \quad \sum_{j=1}^k r_j = n, \tag{9.7}
\]
and the component \( c_n(r) \) does not depend on the unknown parameter \( \pi \). We may also relax the i.d.-condition on the \( X_i \) and conceive of some concomitant/design variates \( Z_i \), so that the corresponding \( \pi_i \) are made to depend on the \( Z_i \) in a semi-parametric manner. Even in the most simple case of no covariates, the GEE corresponding to the multinomial law depends on the unknown parameters appearing in the covariance matrix as well, so that the exact GLM methodology does not work out in general. The quasi-likelihood approach is more adoptable in practice. In the particular case of \( k = 2 \), the usual log-linear transformation not only stabilizes the (asymptotic) variance but also accelerates the rate of convergence to (asymptotic) normality. Unfortunately, for \( k \geq 3 \), even if this variance stabilizing transformation is used coordinatewise, their variance terms depend on the unknown parameters in an involved manner, and hence the usual simplicity (or parametric orthogonality) of the GEE may not be tenable any more. Allowing concomitant variates along with the primary ones, for related GLMs, the GEE are generally quite complex, and its particular exact analysis appeal is largely gone.

In biomedical applications relating to multiple endpoints, correlated binary responses crop up in many situation. If there are \( p(\geq 2) \) (binary) characteristics, we have a set of \( 2^p \) possible outcomes, so that there are \( 2^p - 1 \) linearly independent (probability) parameters associated with such a multinomial law. Such parameters can be reorganized to formulate \( p \) marginal (main) effects, \( \binom{p}{2} \) first-order interactions, and so on. A complete formulation of the likelihood function, though based on an exponential model, may turn out to be quite involved, particularly when \( p \) is not so small. Moreover with the totality of such \( 2^p - 1 \) parameters, an exact treatment for related GEEs becomes prohibitively laborious, if not impracticable. Also, the related asymptotics demand a much larger sample size than in the single endpoint case. For this reason, often a representation involving a lesser number of parameters is advocated. Among such possibilities, special mention may be made of the Bahadur (1961) representation and the log-linear model approaches. Let \( i = (i_1, \ldots, i_p)' \) with \( i_j = 0, 1; \quad j = 1, \ldots, p \), and let let \( \mathcal{I} \) be the set of all possible \( (2^p) \) realizations of the \( i \). Define then
\[
P(X = i) = \pi(i), \quad i \in \mathcal{I}. \tag{9.8}
\]
For the marginal laws, we set
\[
\pi^{(j)}_{i_1} = P(X_j = i), \quad i = 0, 1; \quad j = 1, \ldots, p. \tag{9.9}
\]
Then, for every \( l : 2 \leq l \leq p \), there are \( \binom{p}{l} \) association parameters, \( \theta_l \), defined in terms of the \( \pi(i) \), such that we may equivalently write
\[
\pi(i) = \prod_{j=1}^p \pi^{(j)}_{i_j} + \sum_{2} (-1)^{i_1+i_2} \theta_{j_1j_2} \prod_{r=1}^2 \frac{\pi^{(j_r)}_{i_{j_1i_{j_2}}} \prod_{s=1}^2 \pi^{(s)}_{i_{j_1i_{j_2}}}}{\pi^{(j)}_{i_1i_2}},
\]
25
\[
 + \sum_{3} (-1)^{i_{j_1} + j_{i_2} + i_{j_3}} \cdot \theta_{j_1, j_2, j_3} \cdot \prod_{r=1}^{p} \frac{\pi_{r}^{(j_r)}}{\prod_{r=1}^{j_r} \pi_{r_{j_r}}} \cdot \prod_{r=1}^{p} \frac{\pi_{r}^{(j_r)}}{\prod_{r=1}^{j_r} \pi_{r_{j_r}}} + \cdots + 
\]

\[\phantom{\sum_{3} (-1)^{i_{j_1} + j_{i_2} + i_{j_3}} \cdot \theta_{j_1, j_2, j_3} \cdot \prod_{r=1}^{p} \frac{\pi_{r}^{(j_r)}}{\prod_{r=1}^{j_r} \pi_{r_{j_r}}} \cdot \prod_{r=1}^{p} \frac{\pi_{r}^{(j_r)}}{\prod_{r=1}^{j_r} \pi_{r_{j_r}}} + \cdots +} \tag{9.10}\]

where the summation \(\sum \) extends over all possible \(1 \leq j_1 \leq \cdots \leq j_r \leq p\), for \(r = 2, \ldots, p\). We refer to Sen (1995) for some detailed discussion of this model in the context of multivariate paired comparison models. In many cases, if \(p\) is greater than 3, one may ignore the higher-order association parameters in the above representation and express the probabilities \(\pi(i)\) in terms of only the marginal probabilities and lower-order association parameters. For such reduced models, the GEE methodology can be adopted with greater success. In the same vein [viz., Liang, Zeger and Qaqish (1992)], often a log-linear model is adopted wherein attention is focused on a subset of interpretable parameters (say, the main effects and first-order interactions only) which can be dealt with higher precision than in the case of the entire set of parameters. Although there is some relevance of GLM in this context, the flavor is largely of quasi-likelihood and GEE, and therefore, asymptotic considerations are very important in such adoptations.

In the recent literature, particularly in the context of survival analysis, GEEs relating to multivariate distributions have received due attention. Liang and Zeger (1986 a, b) considered longitudinal data analysis, for discrete and continuous outcomes, using GLMs and discussed possible ways to handle the complex covariance structure appearing in the related GEEs. Lin and Wei (1992) formulated some linear regression analysis procedures for multivariate failure time observations, while earlier, Wei, Lin and Weissfeld (1989) attacked the problem by modeling marginal distributions. We may also refer to Liang et al. (1992, 1993) for some further multivariate analysis of GLMs. Prentice and Cai (1992) considered the covariance and survival function estimation using censored multivariate failure time data. The frailty models (see for example, Clayton, 1978) have also gained popularity, and recently Murphy (1995) has studied the related asymptotic theory in a unified manner. Although most of these works related to certain applications, there remains ample room for further methodological developments which would match the applications to a greater extent. In this context, the emphasis on GLM methodology has been shifted to some alternative approaches (see for example, Sen 1994b), and hence, a blending of these diverse techniques may serve the purpose better. There are certain hidden issues involved in multivariate GLMs when a marginal approach is used, and more work is indeed needed to place the general methodology at a comparable level of generality with the univariate GLMs. The glaring limitations of the conventional GLMs in multivariate models need not be overemphasized; rather, these are to be regarded as the basic influencing factors underlying the much anticipated developments of novel methodology and tools which would serve the purpose more effectively in this active area of interdisciplinary research.

REFERENCES


man and Hall, London.


lished Ph. D. Dissertation, Johns Hopkins Univ. MD.


Pranab K. Sen
Promotions of Biostatistics
and Statistics
University of North Carolina,
Chapel Hill, NC 27599-7400.