MEASUREMENT ERRORS IN GENERALIZED LINEAR MODEL EXPLANATORY VARIABLES

by

Leonard Stefanski
Department of Statistics
North Carolina State University

and

Department of Biostatistics
Harvard School of Public Health

Paper Presented
at the
Third International Workshop on Statistical Modelling
Vienna July 4-8, 1988

SUMMARY

Under the assumption that response and explanatory variables follow a generalized linear model, estimating equations are derived for the case in which the explanatory variables are measured with error. Although the estimating equations are shown to have multiple solutions, a procedure is suggested for uniquely identifying the appropriate root. A by-product of the proposed computational methods is an informative plot, called the measurement error trace, which graphically illustrates the effect of measurement error on estimated parameters.

The first half of the paper reviews the material in Stefanski and Carroll (1987); the latter half focuses on computational issues.

0. INTRODUCTION

This paper studies the problem of fitting generalized linear models to data when explanatory variables are measured with error. Assuming that measurement error is normally distributed and independent of both the true explanatory and response variable, unbiased estimating equations for the generalized linear model parameters are obtained by conditioning on certain sufficient statistics. The estimating equations are suitable for both the functional and structural versions of the measurement error model. For the structural semi-parametric version of the generalized linear measurement-error model, efficient estimating equations are identified.

Definitions and statement of the modelling assumptions are given in Section 1. Estimating equations for functional models are derived in Section 2, and Section 3 contains material relevant to structural measurement-error models. To a great extent Sections 1-3 constitute a review of the recent paper by Stefanski and Carroll (1987).

A major obstacle to the application of the theory in Sections 1-3 is the nonuniqueness of solutions to the proposed estimating equations. This problem was mentioned in Stefanski and Carroll (1987), although a satisfactory solution to the problem was not given. The latter half of this paper addresses the uniqueness problem and certain computational issues which arise in applications. Sections 4 and 5 explain the manner in which multiple solutions to the estimating equations arise and a criterion for identifying the desired solution is proposed. Section 6 presents a method of computation which in turn

suggests a graphical technique, called the measurement-error trace, for displaying the effect of measurement error on parameter estimates.

1. GENERALIZED LINEAR MEASUREMENT ERROR MODELS

1.1 Generalized linear models in canonical form

Throughout this paper attention will be restricted to generalized linear models in canonical form (McCullagh & Nelder, 1983, Ch. 2). That is, given a p-vector explanatory variable U=u, it is assumed that the response variable y has the density

$$h_Y(y;\theta,u) = \exp\{\frac{y(\alpha+\beta^T u)-b(\alpha+\beta^T u)}{a(\phi)} + c(y,\phi)\} \qquad (1.1)$$

with respect to a sigma-finite measure $m(\cdot)$. In (1.1), $\theta^T=(\alpha,\beta^T,\phi)$; $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions; and the dominating measure $m(\cdot)$ does not depend on θ or u. Table 1.1 gives choices of $a(\cdot)$, $b(\cdot)$ and $m(\cdot)$ for some common nonlinear models.

Table 1. Choices of $a(\cdot)$, $b(\cdot)$ and $m(\cdot)$ for some common generalized linear models in canonical form.

	Poisson	Logistic	Gamma	Inverse Gaussian
$a(\phi)$	1	1	ϕ	ϕ
$b(\eta)$	e^{η}	$\log(1+e^{\eta})$	$\log(-1/\eta)$ $(\eta<0)$	$-(-2\eta)^{1/2}$ $(\eta<0)$
$m(\cdot)$	Counting measure on $\{0,1,\ldots\}$	Counting measure on $\{0,1\}$	Lebesque measure on $(0,\infty)$	Lebesque measure on $(0,\infty)$

The key feature these models have in common is a natural sufficient statistic for u, when u is regarded as a parameter and all other

parameters are assumed known. This is crucial to the theory presented later and thus the necessity of the restriction to canonical models. Unfortunately, some canonical models entail restriction on $\alpha+\beta^T u$, e.g. Gamma and Inverse Gaussian, and thus they are less desireable, from a modelling viewpoint, than certain noncanonical models.

1.2 The measurement error model

In a generalized linear measurement error model a proxy, X, is observed in place of U. It is assumed that conditioned on U=u, the observable random variable X has the normal density

$$h_X(x;\theta,u) = \frac{|\bar{\Omega}|^{-1/2}}{(2\pi)^{-p/2}} \exp\{-(1/2)(x-u)^T\bar{\Omega}^{-1}(x-u)\} \qquad (1.2)$$

where $\bar{\Omega}$ is the covariance matrix of the measurement error vector, x-u.

Note that like (1.1), (1.2) possesses a natural sufficient statistic for u when the other parameters are assumed known.

A generalized linear measurement error model is obtained by combining (1.1) and (1.2) under the assumption that Y and X are conditionally independent given U. The resulting density of (Y,X) conditioned on U=u is then

$$h_{Y,X}(y,x;\theta,u) = h_Y(y;\theta,u)h_X(x;\theta,u). \qquad (1.3)$$

For an independent sequence of random variables (Y_i,X_i) (i=1,..,n), let (U_i) (i=1,..,n) denote the corresponding sequence of unobserved covariables. Depending on the nature of the sampling and the inferences to be drawn, it may be appropriate to regard (U_i)

either as a sequence of constants or as a sequence of independent and identically distributed random variables. In the former case a functional model is obtained while the latter case is termed a structural model. Structural models can be further characterized as parametric or semiparametric depending on whether the distribution of U is specified parametrically or nonparametrically. Functional models and nonparametric structural models are studied in this paper.

Not all of the parameters for all versions of model (1.3) are identifiable and thus some additional information is required. It will be assumed that

$$\overline{\Omega}/a(\phi) = \Omega, \tag{1.4}$$

where Ω is known. In simple linear regression (1.4) reduces to the common identifiability assumption that the ratio of measurement-error variance to the equation error variance is known. In models for which $a(\phi)\equiv 1$, for example, logistic and Poisson regression, (1.4) requires that the measurement-error covariance matrix is known.

2. FUNCTIONAL MODELS

2.1 The functional likelihood

Under the assumptions of the functional model there are n+p+2 unknown parameters, α, β^T, ϕ and $u_1,..,u_n$. Given data (Y_i, X_i) $(i=1,..,n)$ the functional log likelihood is

$$L(\theta, u_1,..,u_n) = \sum_{i=1}^{n} \log\{h_{Y,X}(Y_i, X_i; \theta, u_i)\}. \tag{2.1}$$

For the normal linear model it is known that maximization of (2.1) with respect to $(\alpha, \beta^T, \phi, u_1, \ldots, u_n)$ results in consistent estimators of the regression coefficients α and β (Gleser, 1981).

However, for nonlinear models maximization of (2.1) is neither computationally attractive nor is it guaranteed to yield consistent estimators. In logistic regression it is known that the functional maximum likelihood estimator of (α,β^T) is not consistent (Stefanski and Carroll, 1985). This is a classic example of the failure of the method of maximum likelihood in the presence of an increasing number of nuisance parameters (Neyman and Scott, 1948).

2.2 Unbiased estimating equations

Note that (1.3) can be written

$$h_{Y,X}(y,x;\theta,u) = q(\delta,\theta,u)v(y,x,\theta) \qquad (2.2)$$

where

$$q(\delta,\theta,u) = \exp\{\frac{u^T\Omega^{-1}\delta}{a(\phi)} - \frac{u^T\Omega^{-1}u+2b(\alpha+\beta^Tu)}{2a(\phi)}\};$$
$$v(y,x,\theta) = \exp\{\frac{2\alpha y-x^T\Omega^{-1}x}{2a(\phi)} + C^*(y,\phi)\}$$
$$\delta = \delta(y,x,\theta) = x+y\Omega\beta$$
$$C^*(y,\phi) = c(y,\phi) - (1/2)\log[\{2\pi a(\phi)\}^p|\Omega|].$$

Thus viewing u as a parameter and α,β and ϕ as fixed, the statistic

$$\Delta = \Delta(Y,X,\theta) = X+Y\Omega\beta \qquad (2.3)$$

is sufficient for u. As a consequence, the distribution of $Y|\Delta$ does not depend on u, and this fact can be exploited to derive unbiased estimating equations for θ which are independent of u.

Write $h_{Y|\Delta}(y|\delta;\theta)$ for the conditional distribution of $Y|\Delta=\delta$. A routine derivation establishes that

$$h_{Y|\Delta}(y|\delta;\theta) = \exp[y\eta - (1/2)y^2\beta^T\Omega\beta/a(\phi) + c(y,\phi) - \log\{S(\eta,\beta,\phi)\}], \quad (2.4)$$

where $\eta = (\alpha+\beta^T\delta)/a(\phi)$ and $S(\cdot,\cdot,\cdot)$ is determined by the requirement that

$$\int h_{Y|\Delta}(y|\delta;\theta)\, dm(y) = 1. \quad (2.5)$$

Since $m(\cdot)$ does not depend on θ it follows from (2.5) that

$$\int \dot{h}_{Y|\Delta}(y|\delta;\theta)\, dm(y) = 0, \quad (2.6)$$

where $\dot{h}_{Y|\Delta}(y|\delta;\theta) = (\partial/\partial\theta)h_{Y|\Delta}(y|\delta;\theta)$.

Define $\psi_S(y,x,\theta) = \dot{h}_{Y|\Delta}(y|x+y\Omega\beta;\theta)$, then it follows from (2.6) that $E_\theta\{\psi_S(Y,X,\theta)\} = E_\theta[E_\theta\{\psi_S(Y,X,\theta)|\Delta\}] = 0$.

Any estimator $\hat{\theta}_S$ solving

$$\sum_{i=1}^{n} \psi_S(Y_i,X_i,\hat{\theta}_S) = 0 \quad (2.7)$$

will be called a sufficiency estimator. The score, ψ_S, is given by

$$\psi_s(y,x,\theta) = \begin{bmatrix} \{y - E(Y|\Delta=\delta)\}/a(\phi) \\ \{y - E(Y|\Delta=\delta)\}/a(\phi) - \{y^2 - E(Y^2|\Delta=\delta)\}\Omega\beta/a(\phi) \\ r(y,x,\theta) - E\{r(Y,X,\theta)|\Delta=\delta\} \end{bmatrix}_{\delta=x+y\Omega\beta} \tag{2.8}$$

A second unbiased estimating equation is found by adopting an approach due to Lindsay (1980,1982,1983). The conditional score, ψ_c, is defined via

$$\psi_c(y,x,\theta) = \begin{bmatrix} \{y-E(Y|\Delta=\delta)\}/a(\phi) \\ \{y-E(Y|\Delta=\delta)\}t(\delta)/a(\phi) \\ r(y,x,\theta)-E\{r(Y,X,\theta)|\Delta=\delta\} \end{bmatrix}_{\delta=x+y\Omega\beta}, \tag{2.9}$$

where $t(\cdot)$ is a p-vector-valued function not depending on (Y,X). With this restriction on $t(\cdot)$ note that

$$E[\{Y-E(Y|\Delta)\}t(\Delta)] = E(t(\Delta)E[\{Y-E(Y|\Delta)\}|\Delta]) = 0$$

and thus ψ_c is unbiased.

An estimator satisfying

$$\sum_{i=1}^{n} \psi_c(Y_i,X_i,\hat{\theta}_c) = 0 \tag{2.10}$$

will be called a conditional estimator.

The conditional score depends on $t(\cdot)$ which must be specified. Ideally $t(\cdot)$ would be chosen to minimize the asymptotic variance of $\hat{\theta}_c$. However, in a functional model the optimal choice $t(\cdot)$ depends on the particular sequence, $\{u_i\}$, of covariates (in fact no choice of $t(\cdot)$ is better than $t(\Delta_i)=u_i$), and thus no globally optimal solution to this problem exists. A related problem is noted by Cox and Hinkley (1974, p. 146) in connection with hypothesis testing.

The fact that $t(\Delta_i)=u_i$ is optimal, and thus so to is any one-to-one linear function of u_i, suggests choosing $t(\cdot)$ so that $E\{t(\Delta_i)\}$ is a one-to-one linear function of u_i. Thus simply taking $t(\Delta)=\Delta$ is suggested ($E(\Delta)=u+E(y)\Omega\beta$). Another possibility is suggested by the facts that X is unbiased for u and Δ is sufficient for u (assuming θ fixed) and thus $t(\Delta)=E(X|\Delta)$ is a uniformly minimum variance unbiased estimator of u (again assuming θ fixed). Note that

$$E(X|\Delta) = \Delta - E(Y|\Delta)\Omega\beta. \tag{2.11}$$

3. STRUCTURAL MODELS

3.1 The structural likelihood

Consider the nonparametric structural model defined in Section 1.2. The joint density of (Y,X) is given by

$$f_{Y,X}(y,x;\theta,g) = \int h_{Y,X}(y,x;\theta,u)g(u)d\nu(u) \tag{3.1}$$

where $h_{Y,X}$ is defined in (1.3). This constitutes a semiparametric model with parametric component θ and nonparametric component g. The density g is assumed to be an element of G, a family of densities with respect to Lebesgue measure, denoted $\nu(\cdot)$.

Let $\ell(y,x,\theta,g) = \log f_{Y,X}(y,x;\theta,g)$ and $\dot{\ell}(y,x,\theta,g) = (\partial/\partial\theta)\ell(y,x,\theta,g)$. If $g(\cdot)$ were known then $\dot{\ell}$ would be the efficient score for θ.

Assuming that differentiation and integration can be interchanged in (3.1), a useful expression for $\dot{\ell}$ is obtained by noting that

$$\dot{\ell}(y,x,\theta,g) = \frac{\int(\partial/\partial\theta)\log(h)hg d\nu}{\int hg d\nu}$$
$$= E\{(\partial/\partial\theta)\log\, h(y,x;\theta,U)\,|\,Y=y,X=x\}.$$

Thus if $g(\cdot)$ is viewed as a prior for u, then $\dot{\ell}$ has the interpretation as the posterior expectation of the functional maximum likelihood θ-score. Furthermore, since $\Delta=X+Y\Omega\beta$ is sufficient for u in the conditional model, (1.3), the conditional distribution of $U|Y,X$ is the same as that of $U|\Delta$. Therefore,

$$\dot{\ell}(y,x,\theta,g) = E\{(\partial/\partial\theta)\log\, h(y,x;\theta,U)\,|\,\Delta=x+y\Omega\beta\}. \qquad (3.2)$$

3.2 Efficient estimating equations

In model (3.1) interest lies primarily in estimation of θ, i.e., $g(\cdot)$ is a nuisance function. The conditional and sufficiency scores of Section 2 are appropriate for the structural model of this section in the sense of being unbiased but they are generally not efficient.

In this section the efficient θ-score is derived and is shown to be equal to a conditional score (2.9) with $t(\delta)=E(U|\Delta=\delta)$. Efficiency is defined in the sense of Pfanzagl (1982, Ch. 14), Begun et al. (1983) and Lindsay (1983, 1985).

Assume that the family of densities $\{h_Y(y;\eta)\}$, obtained by setting $\eta=\alpha+\beta^T u$ and fixing ϕ in the right hand side of (1.1), is a regular exponential family for $\eta \in H$ where H is one of the three open intervals $(-\infty,0)$, $(0,\infty)$, $(-\infty,\infty)$. Let $\theta = (\alpha,\beta^T,\phi)$ be an element in $\Theta = R\times R^P\times R^+$ and g an element of G. Write $\tau = (\theta,g)$ and with supp(g) denoting the support of g, define $T=\{\tau:\alpha+\beta^T u \in H \text{ for } u \in \text{supp}(g)\}$. The parameter space for the nonparametric structural model is T.

Let S be the class of estimating equations, ψ, satisfying for all τ in T:

$$\text{(i)} \quad E_\tau\{\psi(Y,X,\theta)\} = 0,$$

$$\text{(ii)} \quad E_\tau\{(\partial/\partial\theta)\psi(Y,X,\theta)\} = -E_\tau\{\psi(Y,X,\theta)\dot{\ell}^T(Y,X,\theta)\},$$

$$\text{(iii)} \quad E_\tau\{||\psi(Y,X,\theta)||^2\} < \infty.$$

If G is complete in the sense defined below, then it transpires that every score in S must be conditionally unbiased with respect to Δ (Theorem 3.1), i.e., if ψ is in S then

$$E_\tau\{\psi(Y,X,\theta|\Delta)\} = 0, \text{ for all } \tau \text{ in T.} \qquad (3.3)$$

This allows an easy derivation (Corollary 3.1) of the efficient estimating equation for θ.

Definition. A collection of functions, H, is said to be complete with respect to a measure μ if a necessary condition for

$$\int t(s)h(s)d\mu(s) = 0,$$

for all $h \in H$, is $t(\cdot)=0$ μ-almost surely.

For a fixed $\theta \in \Theta$ let $G_\theta=\{g \in G:(\theta,g) \in T\}$ and let ν_θ be Lebesgue measure on $\{u \in R^P:\alpha+\beta^T u \in H\}$.

Theorem 3.1. Assume that for each fixed $\theta \in \Theta$, G_θ is complete with respect to ν_θ. Then if $\psi \in S$, $E_\tau\{\psi(Y,X,\theta)|\Delta\}=0$ for all $\tau \in T$.

The positive-definite matrix $V_\psi = \{E(\psi\dot{\ell}^T)\}^{-1}E(\psi\psi^T)\{E(\dot{\ell}\psi^T)\}^{-1}$ measures the efficiency of ψ as an estimating equation. Under regularity conditions V_ψ is the asymptotic covariance matrix of $n^{1/2}(\hat{\theta}-\theta)$ when $\hat{\theta}$ is a consistent estimator of θ satisfying $\Sigma\psi(Y_i, X_i, \hat{\theta})=0$.

Let

$$\psi^* = \dot{\ell}(y,x,\theta,g) - E\{\dot{\ell}(Y,X,\theta,g) \mid \Delta = x+y\Omega\beta\} \qquad (3.4)$$

and let V_{ψ^*} be the associated covariance matrix. The following result states that ψ^* is the efficient θ-score.

Corollary 3.1. $V_\psi \geq V_{\psi^*}$ for all $\psi \in S$.

Proofs of Theorem 3.1 and its corollary can be found in Stefanski and Carroll (1987) and will not be given here.

To find ψ^* note that Y and U are conditionally uncorrelated given Δ. This follows from the facts that the σ-field generated by Δ is contained in the σ-field generated by (Y,X), and the conditional distributions of $U|(Y=y,X=x)$ and $U|\Delta=x+y\Omega\beta$ are identical. Thus $E(YU|\Delta) = E\{E(YU|Y,X)|\Delta\} = E\{YE(U|\Delta)|\Delta\} = E(Y|\Delta)E(U|\Delta)$. This fact and (3.2) imply that

$$\psi^* = \dot{\ell} - E(\dot{\ell}|\Delta) = E(\dot{h}/h|Y,X) - E\{E(\dot{h}/h|Y,X)|\Delta\}$$
$$= E(\dot{h}/h|Y,X) - E(\dot{h}/h|\Delta)$$

from which one finds

$$\psi^*(y,x,\theta) = \begin{bmatrix} \{y - E(Y|\Delta=\delta)\}/a(\phi) \\ \{y - E(Y|\Delta=\delta)\}E(U|\Delta=\delta)/a(\phi) \\ r(y,x,\theta) - E\{r(Y,X,\theta)|\Delta=\delta\} \end{bmatrix}_{\delta=x+y\Omega\beta}. \tag{3.5}$$

Comparison with (2.9) shows that ψ^* is a conditional score with $t(\delta) = E(U|\Delta=\delta)$. Note that $\Delta=X+Y\Omega\beta=U+Z+Y\Omega\beta$ where Z has a $N\{0,a(\phi)\Omega\}$ distribution. This implies that

$$E(U|\Delta) = \Delta - E(Y|\Delta)\Omega\beta - E(Z|\Delta).$$

But by (2.11), $\Delta - E(Y|\Delta)\Omega\beta = E(X|\Delta)$; also it can be shown that

$$E(Z|\Delta=\delta) = -a(\phi)\Omega\dot{f}_\Delta(\delta)/f_\Delta(\delta)$$

where $f_\Delta(\delta)$ is the density of Δ and $\dot{f}_\Delta(\delta) = (\partial/\partial\delta)f_\Delta(\delta)$. Thus

$$E(U|\Delta=\delta) = E(X|\Delta=\delta) + a(\phi)\Omega\dot{f}_\Delta(\delta)/f_\Delta(\delta). \tag{3.6}$$

Fully efficient estimators of (α,β^T) in the linear model have been given by Bickel and Ritov (1986).

4. LOGISTIC REGRESSION

In this section the logistic model is used as a means of illustrating and comparing the various estimating equations. The logistic model assumes that

$$pr_\theta(Y=1|U=u) = F(\alpha+\beta^T u)$$

where $F(t) = 1/(1+e^{-t})$. For this model $a(\phi) \equiv 1$ and $m(\cdot)$ is counting measure on $\{0,1\}$. The conditional distribution of $Y|\Delta=\delta$ is given by

$$pr_\theta(Y=1|\Delta=\delta) = F\{\alpha+(\delta-(1/2)\Omega\beta)^T\beta\}. \tag{4.1}$$

The sufficiency score is, by (2.8),

$$\psi_s(y,x,\Theta) = [y - F\{\alpha+(\delta-(1/2)\Omega\beta)^T\beta\}]\begin{pmatrix}1\\ t(\delta)\end{pmatrix}\Big|_{\delta=x+y\Omega\beta,} \qquad (4.2)$$

with $t(\delta) = \delta-\Omega\beta$.

Note that if Δ^* is defined as $\Delta^*=\Delta-(1/2)\Omega\beta$, with a similar notation for $\delta^*=\delta-(1/2)\Omega\beta$, then

$$pr_\Theta(Y=1|\Delta^*=\delta^*) = F(\alpha+\beta^T\delta^*); \qquad (4.3)$$

i.e., conditioned on $\Delta^*=\delta^*$, Y follows a logistic model. This closure property (equality of the conditional distributions of $Y|U$ and $Y|\Delta^*$ where Δ^* is some function of Δ) seems only to hold for the normal and logistic models.

The sufficiency score is a conditional score for the particular choice $t(\delta)=\delta-\Omega\beta$. In Section 3 it was suggested that taking $t(\delta) = E(X|\Delta=\delta)$ should lead to promising estimating equations. For the logistic model

$$E(X|\Delta=\delta) + \delta - F\{\alpha+\beta^T(\delta-(1/2)\Omega\beta)\}\Omega\beta. \qquad (4.4)$$

Fully efficient estimation requires that $t(\delta) = E(X|\Delta=\delta) + \Omega\dot{f}_\Delta(\delta)/f_\Delta(\delta)$, see equation (3.6).

All of the scores for the logistic model share a common problem; the associated estimating equations may possess multiple roots. Consider, for example, the conditional score with $t(\delta) = \delta-(1/2)\Omega\beta$ and let

$$G_n(\alpha,\beta) = \sum_{i=1}^{n} \psi_c(Y_i,X_i,\theta)$$

$$= \sum_{i=1}^{n} [Y_i - F\{\alpha+\beta^T X_i+(Y_i-1/2)\beta^T\Omega\beta\}] \begin{pmatrix} 1 \\ X_i+(Y_i-1/2)\Omega\beta \end{pmatrix}. \quad (4.5)$$

Note that if $Y_i=1$,

$$\psi_c(Y_i,X_i,\theta) = \{1 - F(\alpha+\beta^T X_i+(1/2)\beta^T\Omega\beta)\}\begin{pmatrix} 1 \\ X_i+\Omega\beta/2 \end{pmatrix}$$

and if $\beta^T\Omega\beta \to \infty$, $\psi_c(Y_i,X_i,\theta) \to 0$; also if $Y_i=0$,

$$\psi_c(Y_i,X_i,\theta) = -F(\alpha+\beta^T X_i-(1/2)\beta^T\Omega\beta)\begin{pmatrix} 1 \\ X_i-\Omega\beta/2 \end{pmatrix}$$

and again $\psi_c(Y_i,X_i,\theta) \to 0$ as $\beta^T\Omega\beta \to \infty$. Thus if $||\beta|| \to \infty$ in such a way that $\beta^T\Omega\beta \to \infty$, $G_n(\alpha,\beta) \to 0$. The manner in which ψ_c depends on β through $\beta^T\Omega\beta$ makes the score behave similar to "redescending" scores which find applications in robust statistics. A consequence of this behavior is the fact that $G_n(\cdot,\cdot)$ may have multiple roots not all of which lead to consistent sequences of estimators.

Figure 1 displays a graph of the second component of $G_n(0,\beta)$ vs β for $\beta \in [0.25,10.25]$. Sample size, n, was set at 100; the (U_i) are distributed as standard normal random variates; the (X_i) were generated according to the model

$$X_i = U_i + \sqrt{\Omega}\, Z_i$$

where (Z_i) are standard normal random variables independent of the (U_i); $\Omega=1$; and Y_i were generated according to model (1.1) with $\alpha=0$ and $\beta=1$. It is evident that $G_n(0,\beta)$ contains multiple roots in the interval $[0.25,10.25]$.

The problem of multiple roots is not specific to logistic regression. It results from the fact that β enters model (2.4)

quadratically through the term $\beta^T \Omega \beta$. The next section discusses the problem of multiple roots and suggests a strategy for locating the appropriate root.

5. THE PROBLEM OF MULTIPLE ROOTS

It is often the case that (2.7) and (2.10) have multiple roots. This is not a finite-sample problem; it persists asymptotically. Thus a strategy is required for selecting the correct solution to equations (2.7) and (2.10).

As a means of motivation, the problem of multiple roots will be discussed first in the context of the simple linear errors-in-variables version of model (1.3). The insights gained from this investigation are then generalized to nonlinear models and illustrated in the context of logistic regression.

Let β_o be a scalar and suppose that given U=u, Y has a normal distribution with mean $\alpha+\beta_o u$ and variance σ^2. The conditional distribution of $Y|\Delta=\delta$ is normal with variance $\sigma^2/(1+\beta_o^T \Omega \beta_o)$ and mean $(\alpha+\beta_o \delta)/(1+\beta_o^T \Omega \beta_o)$. It can be shown that for this model the estimating equations (2.7) imply that

$$\hat{\alpha}_s = \bar{Y} - \hat{\beta}_s \bar{X}$$

and $\hat{\beta}_s$ solves

$$-\hat{\beta}_s^2 \Omega S_{YX} + (S_{YY} \Omega - S_{XX}) \hat{\beta}_s + S_{YX} = 0, \qquad (5.1)$$

where

$$S_{YX} = \sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X}); \quad S_{XX} = \sum_{i=1}^{n} (X_i - \bar{X})^2;$$

$$S_{YY} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

This quadratic equation has two real roots, $\hat{\beta}_{s,1}$ and $\hat{\beta}_{s,2}$ (Kendall and Stuart 1979, Ch. 29), converging to β_o and $-1/\Omega\beta_o$, ($\beta_o \neq 0$). A number of root-selection criteria can be formulated for this model but unfortunately they do not generalize to nonlinear multiple-regression models. For example, the correct root, $\hat{\beta}_{s,1}$, has, at least asymptotically, the same sign as the so-called naive estimator, $\hat{\beta}_N = S_{YX}/S_{XX}$.

Consider the function

$$f(\beta,\tau) = -\beta^2\tau\Omega S_{YX} + (S_{YY}\tau\Omega - S_{XX})\beta + S_{YX} , \qquad (5.2)$$

obtained by replacing Ω with $\tau\Omega$ in (5.1). Note that when $\tau=0$ the equation

$$f(\beta,\tau) = 0 \qquad (5.3)$$

has the unique solution $\hat{\beta}_N = S_{YX}/S_{XX}$; while for $\tau=1$, (5.3) and (5.1) are identical. Since (5.3) has a unique solution at $\tau=0$, the implicit function theorem guarantees the existence of a unique function, $\hat{\beta}(\tau)$, solving (5.3) for all τ in some neighborhood of $\tau=0$. It transpires that $\hat{\beta}(\tau)$ exists uniquely for all $\tau \in [0,1]$ and that $\hat{\beta}(1) = \hat{\beta}_{s,1}$. That is the "correct" root of (5.1) is determined by the fact that it lies on the locus of solutions, $\{\hat{\beta}(\tau): 0 \leq \tau \leq 1)\}$ to (5.3), which in turn is uniquely determined by the condition that $\hat{\beta}(0) = S_{YX}/S_{XX}$. These ideas are illustrated in Figure 2.

Now consider logistic regression and the estimating equations given in (4.5). Define $G_n(\alpha,\beta,\tau)$ for $0 \leq \tau \leq 1$ by

$$G_n(\alpha,\beta,\tau) = n^{-1}\sum_{i=1}^{n}[Y_i - F\{\alpha+\beta^T X_i+(Y_i-1/2)\beta^T\Omega\beta\tau\}]\begin{pmatrix}1\\ X_i+(Y_i-1/2)\tau\Omega\beta\end{pmatrix}. \quad (5.4)$$

Note that when $\tau=0$, $G_n(\alpha,\beta,0)$ is the gradient of an ordinary logistic log likelihood. Consequently, except in cases of quasi- or complete separation (Santner and Duffy, 1986), the equation

$$G_n(\alpha,\beta,0) = 0$$

possesses a unique finite solution, $\hat{\theta}_N = (\hat{\alpha}_N, \hat{\beta}_N^T)^T$, which is easily found using a Newton-Raphson iteration; $\hat{\theta}_N$ is the so-called naive estimator. As in the linear model the implicit function theorem guarantees the existence of a unique family of solutions, $\hat{\theta}(\tau)$, to the equation

$$G_n(\alpha,\beta,\tau) = 0,$$

in some neighborhood of $\tau=0$ such that $\hat{\theta}(0)=\hat{\theta}_N$. If $\hat{\theta}(\tau)$ exists uniquely for all $\tau \in [0,1]$ then the asymptotic arguments in Appendix I suggest that $\hat{\theta}(1)$ is the consistent estimator we seek.

A simple example illustrates the preceeding argument. Consider the problem of fitting a no-intercept ($\alpha=0$) logistic model to the data described at the end of Section 4. Figure 1 indicates that the estimating equation (4.5) has at least three roots. In Figure 3 is plotted the second component of $G(\alpha,\beta,\tau)$ for $\tau=0.00$, 0.25, 0.50, 0.75 and 1.00. The figure clearly shows which root is connected continuously to the naive estimator. Figure 4 contains plots of the same functions depicted in Figure 3 but for a data set of size n=1000. This figure makes it clear that the problem of multiple roots persists asymptotically and that the suggested root-finding procedure locates the correct root in this case.

For models other than logistic or linear regression the root-finding argument proceeds similarly. Let $\psi(Y,X,\theta,\tau)$ denote either (2.8) or (2.9) with $\tau\Omega$ in place of Ω. Then $\hat{\theta}(\tau)$ solves

$$0 = G_n(\theta,\tau) = n^{-1} \sum_{i=1}^{n} \psi(Y_i,X_i,\theta,\tau). \qquad (5.5)$$

6. THE MEASUREMENT ERROR TRACE

Let $\hat{\theta}(\tau)$, $0\leq\tau\leq1$ be the solution locus discussed in the previous section. Generally $\hat{\theta}(0)$ is easily obtained using standard computational methods; this is the naive estimator. For $\tau>0$, $\hat{\theta}(\tau)$ can be found by employing a Newton-Raphson interation of (5.5) starting from $\hat{\theta}(0)$. This iteration will converge to the desired solution provided τ is sufficiently close to zero. The solution locus, $\hat{\theta}(\tau)$,$(0\leq\tau\leq1)$, can be generated on a grid of τ values $\{0=\tau_0<\tau_1<..<\tau_k=1\}$ successively by using a Newton-Raphson iteration starting at $\hat{\theta}(\tau_i)$ to compute $\hat{\theta}(\tau_{i+1})$. The iteration scheme converges to the desired solution provided the grid mesh is sufficiently small.

Note that $\hat{\theta}(\tau)$ is the estimator one would obtain if the measurement error covariance had been assumed to be proportional to $\tau\Omega$ instead of Ω. It is often the case that a measurement error model is fit to data primarily for examining the effects of measurement error on estimated parameters. In these cases Ω may not be known, but represents a best guess or crude estimate of the measurement error covariance matrix. A plot of $\hat{\theta}(\tau)$ versus τ illustrates the nature of the dependence of the estimated parameters on the magnitude of the (assumed) error covariance. Since it is similar in design and intent to a ridge trace, the plot of $\hat{\theta}(\tau)$ versus τ is called a measurement-error trace. Provided $\hat{\theta}(\tau)$,

($0 \leq \tau \leq 1$), has been computed by the procedure suggested in the preceeding paragraph, the measurement error trace is easily constructed.

Figures 5-8 display examples of the measurement error trace for some different logistic regression measurement-error models. In each example one thousand observations were generated according to the logistic model with $\alpha=1$, $\beta^T=(0.00,0.25,0.50)$, $U \sim N(0,I_3)$, $Z \sim N(0,I_3)$ and $X=U+\Omega^{1/2}Z$. Figures 5-8 differ only with respect to the choice of Ω. For Figure 5, $\Omega=\Omega_5=I_3$; for Figure 6

$$\Omega = \Omega_6 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix};$$

for Figure 7,

$$\Omega = \Omega_7 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 2^{-1/2} \\ 0 & 2^{-1/2} & 1 \end{pmatrix};$$

and for Figure 8,

$$\Omega = \Omega_8 = \begin{pmatrix} 1 & 2^{-1/2} & 0 \\ 2^{-1/2} & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Only the estimated slope coefficients, $\hat{\beta}(\tau)$, are plotted in Figures 5-8. Recall that $\hat{\beta}(0)$ is the naive estimate, i.e., the estimate obtained by fitting a logistic model to the observed data ignoring measurement error; $\hat{\beta}(1)$ is the errors-in-variables estimate. Figures 5-8 clearly illustrate those parameters which are affected by the covariable measurement error.

ACKNOWLEDGEMENTS

Support for this work was provided in part by a Cooperative Agreement between Harvard University, SIMS, and the Environmental Protection Agency, and the National Science Foundation.

REFERENCES

Begun, J.M., Hall, W.J., Hwang, W.M. & Wellner, J.A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. Ann. Statist. 11, 432-52.

Bickel, P.J. & Ritov, Y. (1987). Efficient estimation in the errors-in-variables model. Ann. Statist. 15, 513-40.

Cox, D.R. & Hinkley, D.V. (1974). Theoretical Statistics. London: Chapman and Hall.

Gleser, L.J. (1981). Estimation in a multivariate 'errors-in-variables' regression model: large sample results. Ann. Statist. 9, 24-44.

Kendall, M.G. & Stuart, A. (1979). The Advanced Theory of Statistics, 2. London: Griffin.

Lindsay, B.G. (1980). Nuisance parameters, mixture models, and the efficiency of partial likelihood estimators. Philos. Trans. Roy. Soc. London Ser. A 296, 639-65.

Lindsay, B.G. (1982). Conditional score functions: some optimality results. Biometrika 69, 503-12.

Lindsay, B.G. (1983). Efficiency of the conditional score in a mixture setting. Ann. Statist. 11, 486-97.

Lindsay, B.G. (1985). Using empirical partially Bayes inference for increased efficiency. Ann. Statist. 13, 914-32.

McCullagh, P. & Nelder, J.A. (1983). Generalized Linear Models. London: Chapman and Hall.

Neyman, J. & Scott, E.L. (1948). Consistent estimates based on partially consistent observations. Econometrica 16, 1-32.

Pfanzagl, J. (1982). Contributions to a General Asymptotic Statistical Theory. New York: Springer-Verlag.

Santner, T.J. & Duffy, D.E. (1986). A note on A. Albert and J.A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. Biometrika 73, 755-58.

Stefanski, L.A. & Carroll, R.J. (1985). Covariate measurement error in logistic regression. Ann. Statist. 13, 1335-51.

Stefanski, L. A. & Carroll, R.J. (1987). Conditional scores and optimal scores for generalized linear measurement error models. Biometrika 74, 703-16.

APPENDIX 1

An argument is given which suggests that the root selection procedure presented in Section 5 works asymptotically.

Let λ denote an element in R^s and τ an element in $[0,1]$. Let $\{G(\cdot,\cdot), G_1(\cdot,\cdot), G_2(\cdot,\cdot),....\}$ be a sequence of functions defined on $R^s x[0,1]$, taking values in R^s such that $G_n(\cdot,\cdot)$ converges to $G(\cdot,\cdot)$ uniformly on compact sets in $R^s x[0,1]$.

Assume that $G(\lambda,0)$ has a unique root, i.e., the equation

$$G(\lambda,0) = 0 \qquad (A.1)$$

has a unique solution, λ_o. Also assume that there exists a unique element of $C^s[0,1]$, θ, such that

$$G\{\theta(\tau),\tau\} = 0 \text{ for all } \tau \in [0,1]. \qquad (A.2)$$

Note that $\theta(0)=\lambda_o$.

Finally, make the assumption that for each fixed τ, $\theta(\tau)$ is an isolated root of $G(\theta,\tau)=0$ uniformly in τ. More specifically it is assumed:

There exists an $\eta>0$, not depending on τ, such that

$$G(\lambda,\tau)=0 \text{ and } ||\lambda-\theta(\tau)|| \leq \eta \text{ imply } \lambda=\theta(\tau), \text{ for all } \tau. \qquad (A.3)$$

Under these conditions it is possible to prove the following result.

Proposition A.1: If for each n, there exists $\theta_n(\cdot) \in C^s[0,1]$ such that

$$G_n\{\theta_n(\tau),\tau\} = 0 \text{ for all } \tau \in [0,1]$$

and $\theta_n(0) \to \lambda_o$, then

$$\lim_{n} \sup_{\tau} ||\theta_n(\tau) - \theta(\tau)|| = 0.$$

Proof: It is sufficient to show that given any $\delta > 0$,

$$\limsup_{n} \sup_{\tau} ||\theta_n(\tau)-\theta(\tau)|| \leq \delta. \qquad (A.4)$$

Note that it is also sufficient to consider only those δ for which $o<\delta<\eta$, where η is defined in (A.3).

Suppose there exists some δ, $0\leq\delta\leq\eta$, for which (A.4) does not hold. It is shown that this leads to a contradiction.

Define $D_n(\cdot)$, α_n, and τ_n by the equations

$$D_n(\tau) = ||\theta_n(\tau) - \theta(\tau)||;$$

$$\alpha_n = \sup_{\tau} D_n(\tau) = D_n(\tau_n).$$

If (A.4) does not hold then it is possible to find a subsequence $\{n_k\}$, such that $\alpha_{n_k}>\delta$ for all n_k. Since $D_{n_k}(0) \to 0$, $D_{n_k}(\tau_{n_k}) = \alpha_{n_k}>\delta$, and $D_n(\cdot)$ is a continuous function, it follows that for n_k large enough, $D_{n_k}(\tau)=\delta$ for some τ, call it $\tau^*_{n_k}$. Note that the sequences $\{\tau^*_{n_k}\}$ and $\{\theta_{n_k}(\tau^*_{n_k})\}$ are both contained in compact sets. Thus a further subsequence $\{n_j\}$ can be found along which $\tau^*_{n_j} \to \tau^*$, $\theta_{n_j}(\tau^*_{n_j}) \to \lambda^*$ and

$$\delta = D_{n_j}(\tau^*_{n_j}) \to ||\lambda^*-\theta(\tau^*)||.$$

Let $\lambda_{n_j} =\theta_{n_j}(\tau^*_{n_j})$;

since $G_n \to G$ uniformly on compact sets

$$G_n(\lambda_{n_j},\tau^*_{n_j}) - G(\lambda_{n_j},\tau^*_{n_j}) \to 0.$$

By continuity of G,

$$G(\lambda_{n_j},\tau^*_{n_j}) - G(\lambda^*,\tau^*) \to 0;$$

and thus

$$G_n(\lambda_{n_j},\tau^*_{n_j}) - G(\lambda^*,\tau^*) \to 0.$$

But $G_n(\lambda_{n_j},\tau^*_{n_j}) \equiv 0$ and this means that $G(\lambda^*,\tau^*) = 0$.
So it has been shown that $G(\lambda^*,\tau^*)=0$, and $||\lambda^*-\theta(\tau^*)|| = \delta \leq \eta$ and thus (A.3) implies that $\lambda^*=\theta(\tau^*)$. Since $||\lambda^*-\theta(\tau^*)||=\delta$, this contradicts the assumption that $\delta>0$.

In the application of the proposition to the root selection problem, say for example in structural logistic regression, G_n is given by (5.4) and

$$G(\alpha,\beta,\tau) = E_{\theta_o}\{G_n(\alpha,\beta,\tau)\}.$$

Pointwise convergence of G_n to G, either almost surely or in probability, follows from the corresponding law of large numbers. Smoothness conditions on ψ_c and regularity conditions on the joint density of (Y,X) will generally guarantee that the convergence is uniform on compact sets.

The existence of a unique solution to the equation $G(\alpha,\beta,0)=0$ follows from the fact, again under regularity conditions, that $G(\alpha,\beta,0)$ is the expected gradient of a convex likelihood function; λ_o is just the limit of the so-called naive estimator. Note also that by design, $G(\alpha_o,\beta_o,1)=0$.

Now provided that $\partial G(\alpha,\beta,\tau)/\partial(\alpha,\beta)$ is nonsingular, $0\leq\tau\leq1$ (at $\tau=0,1$ this follows trivially) the Implicit Function Theorem guarantees the existence of a unique solution to $G(\alpha(\tau),\beta(\tau),\tau)=0$ for all τ in some neighborhood of zero (or one) and thus (A.2) and (A.3) simply

rule out exceptional behavior of $G(\alpha,\beta,\tau)$ for $0\leq\tau\leq1$ and thus will generally hold under sufficient regularity conditions.

Finally concavity of the usual logistic likelihood insures that $G_n(\alpha,\beta,0)=0$ has, for n large enough, a unique convergent solution. Thus provided $[\alpha_n(\tau),\beta_n^T(\tau)]$ solves $G_n\{\alpha_n(\tau),\beta_n(\tau),\tau\} = 0$, for all $\tau \in [0,1]$, the proposition establishes the convergence of $\alpha_n(\tau),\beta_n^T(\tau)$ for all τ, and hence that of $(\alpha_n(1),\beta_n^T(1))$ to (α_o,β_o^T).

Figure 1. Plot of $G_n(0,\beta)$ vs. β, $(0.25<\beta<10.25)$; model, logistic regression; sample size, 100; (Score = $G_n(0,\beta)$).

Figure 1

Figure 2. Plots of $f(\beta,\tau)$, $\tau=0(0.25)1$, vs. β; model, simple linear regression; sample size, 100; (Score = $f(\beta,\tau)$).

Figure 2

Figure 3. Plot of $G_n(0,\beta,\tau)$, $\tau=0(0.25)1$, vs. β, $(0.25<\beta<10.25)$; model, logistic regression; sample size, 100; (Score = $G_n(0,\beta,\tau)$).

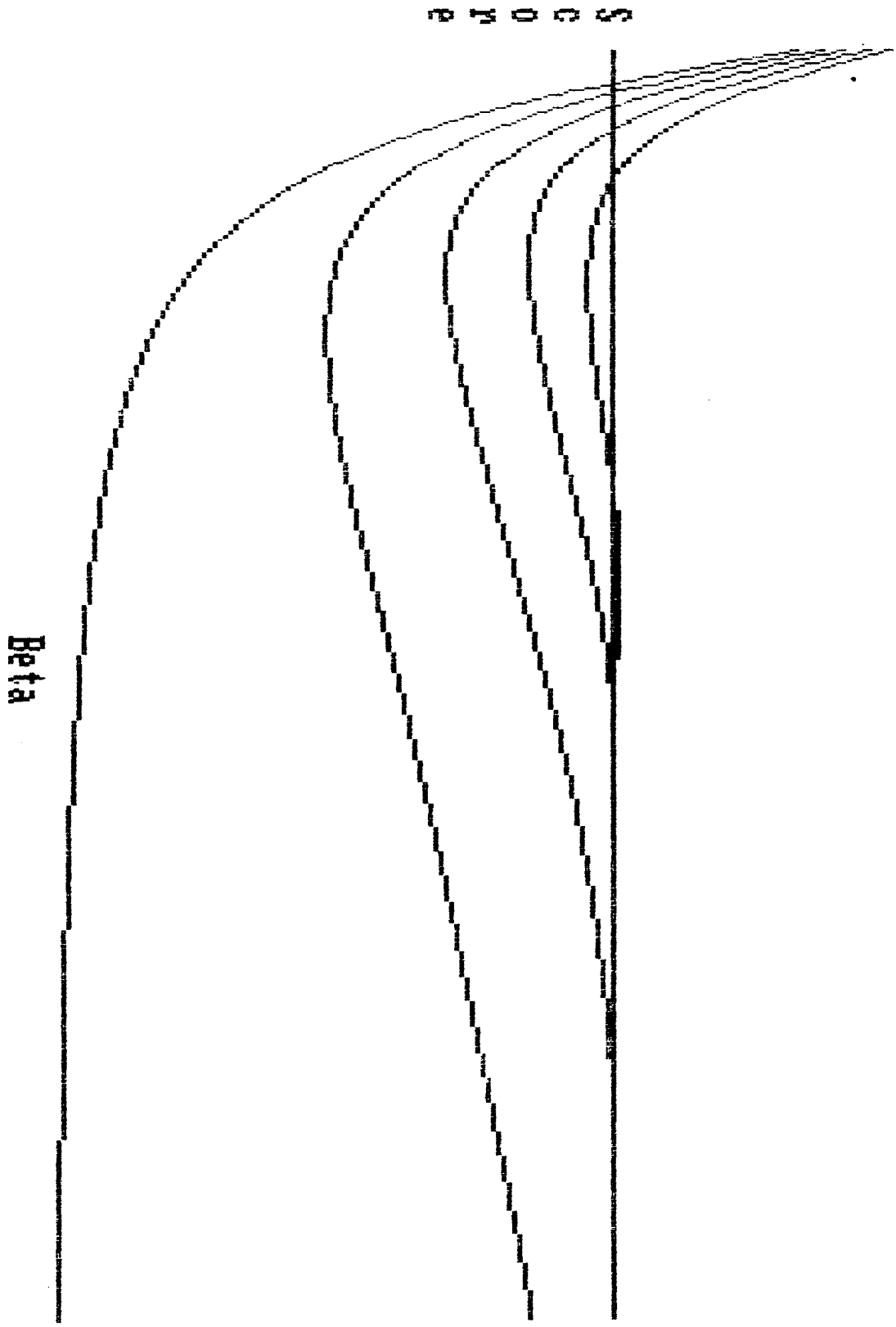

Figure 3

Figure 4. Plot of $G_n(0,\beta,\tau)$, $\tau=0(0.25)1$, vs. β, $(0.25 < \beta < 10.25)$; model, logistic regression; sample size, 1000; (Score = $G_n(0,\beta,\tau)$).

Figure 4

Figure 5. Plot of $\hat{\beta}(\tau)$ vs. τ, $0 < \tau < 1$; model, logistic regression, $(\alpha,\beta_1,\beta_2,\beta_3) = (1,0.00,0.25,0.50)$, $U \sim N(0,I_3)$, $Z \sim N(0,I_3)$, $X = U+\Omega^{1/2}Z$, $\Omega = \Omega_5$ (see Section 6); sample size=1000.

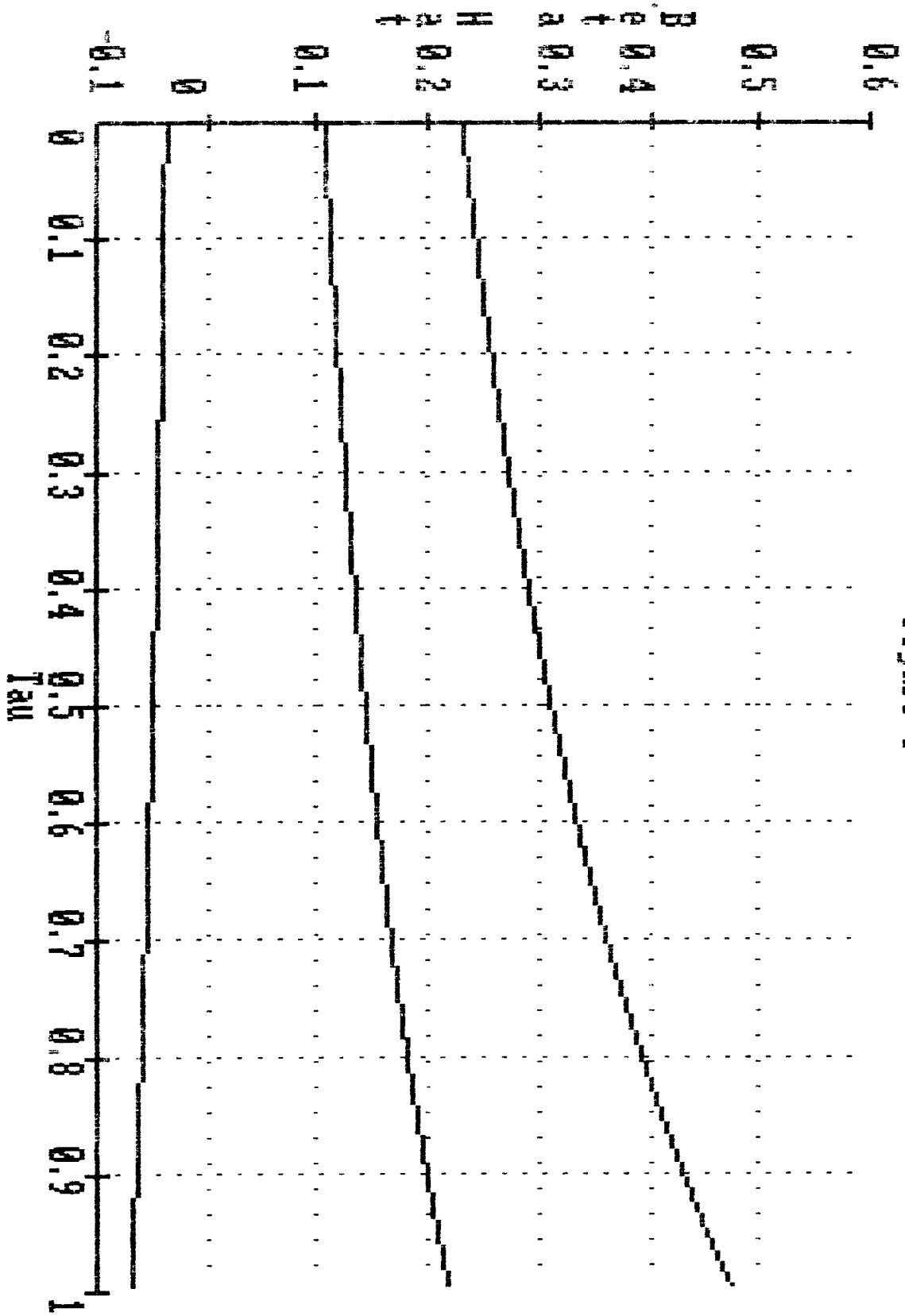

Figure 5

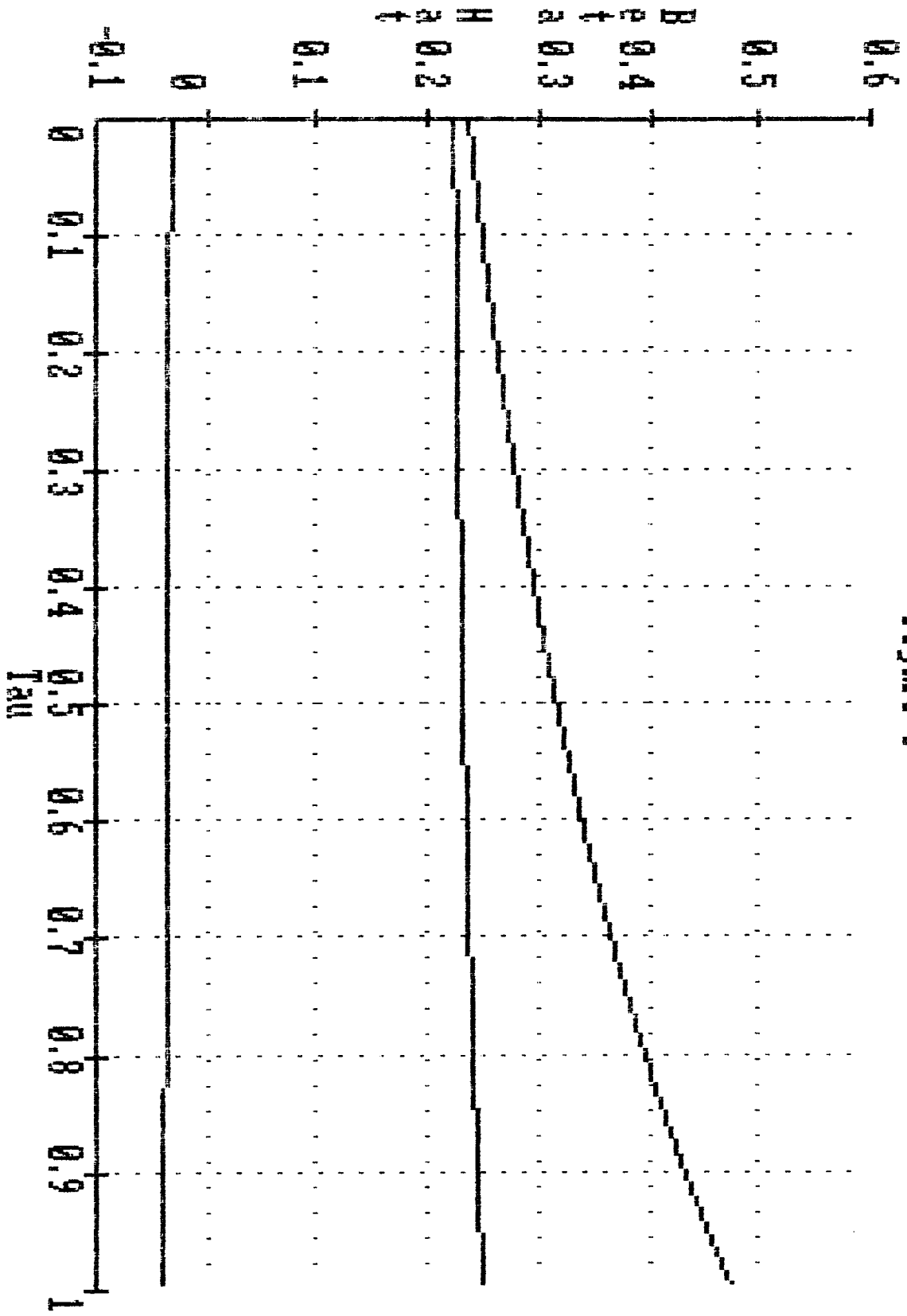

Figure 6

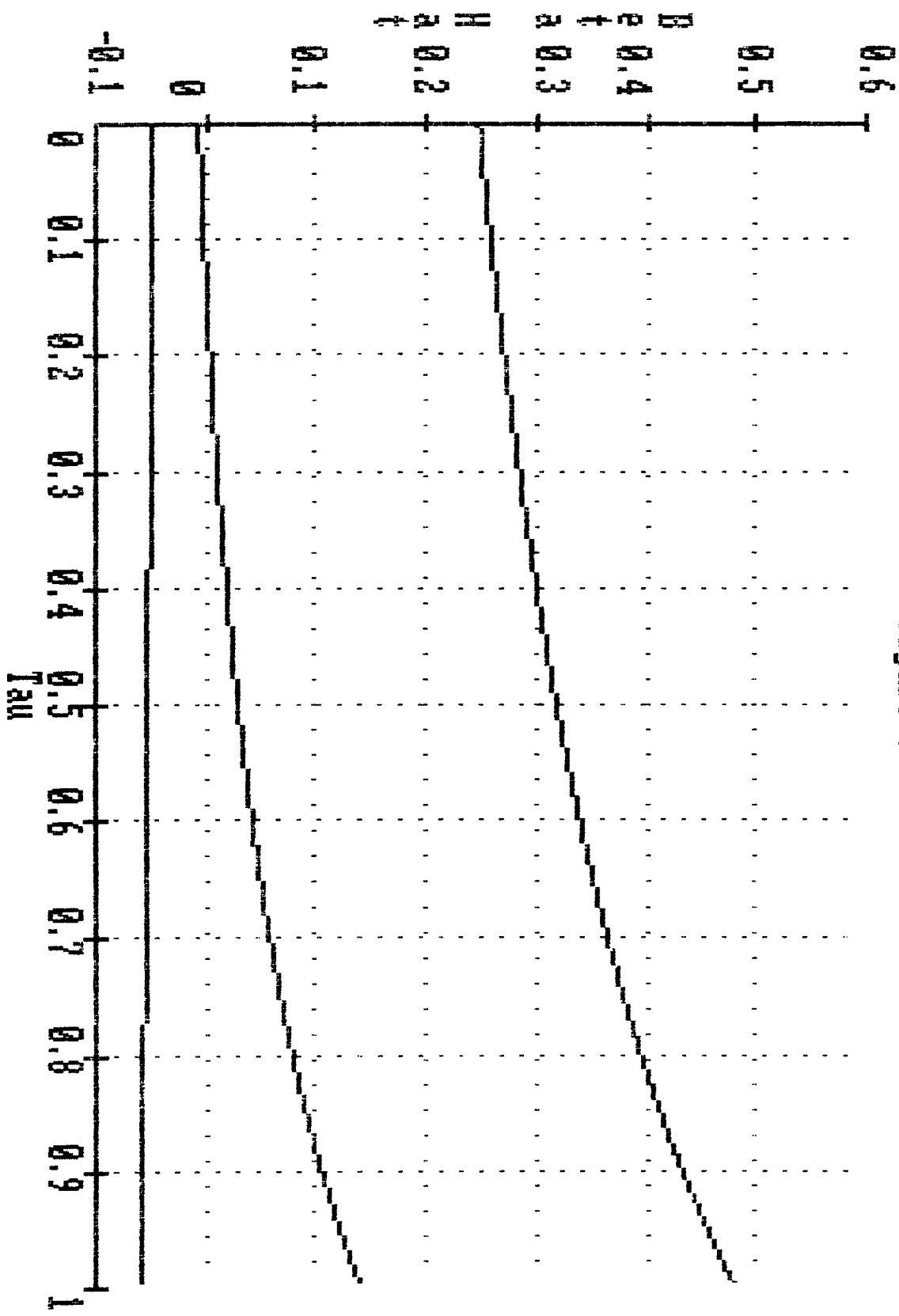

Figure 7

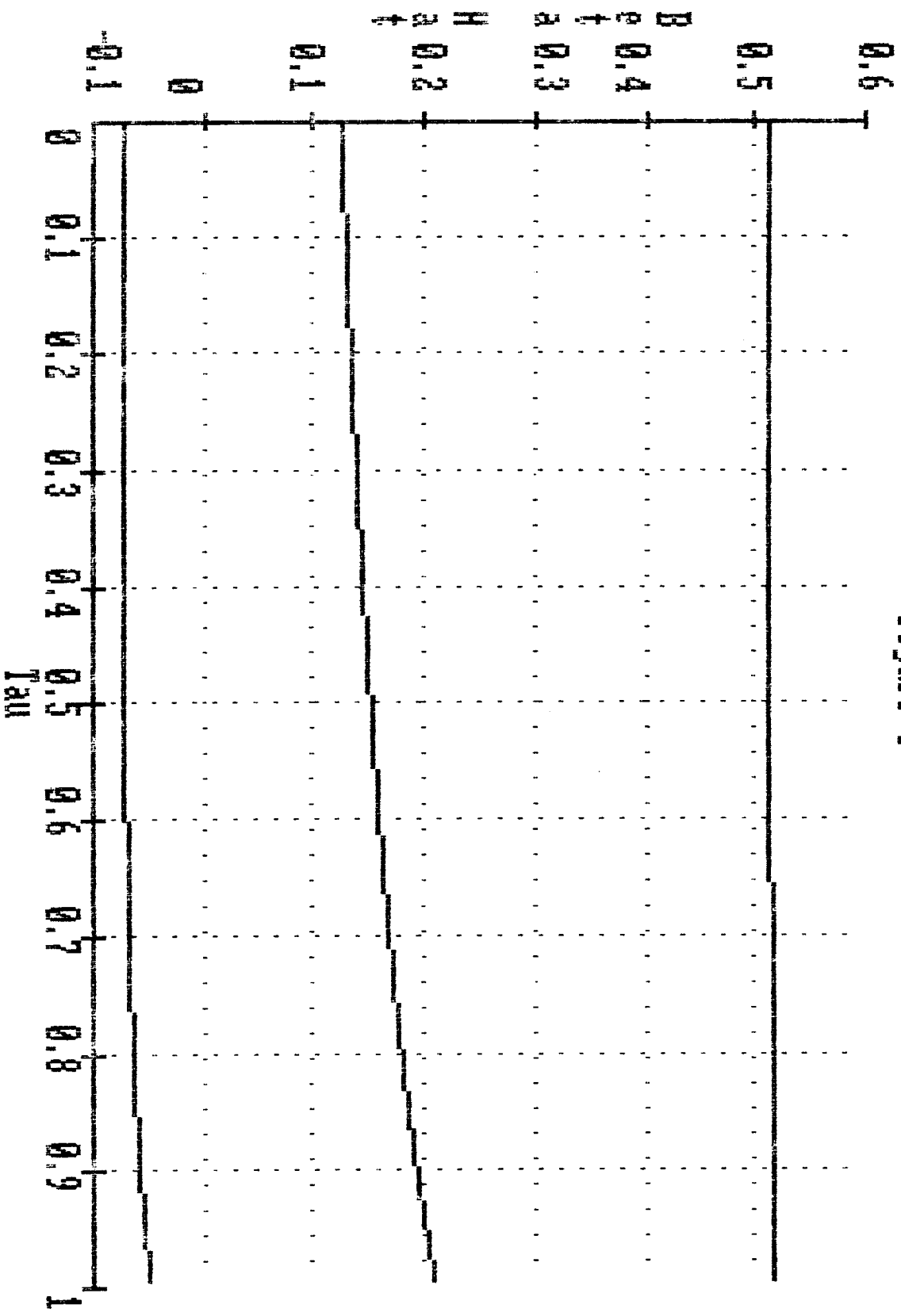

Figure 8