Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://amstat.tandfonline.com/loi/uasa20

Quasi-Likelihood Estimation in Semiparametric Models

Thomas A. Severini & Joan G. Staniswalis

To cite this article: Thomas A. Severini & Joan G. Staniswalis (1994) Quasi-Likelihood Estimation in Semiparametric Models, Journal of the American Statistical Association, 89:426, 501-511, DOI: 10.1080/01621459.1994.10476774

To link to this article: https://doi.org/10.1080/01621459.1994.10476774

Published online: 27 Feb 2012.

Quasi-likelihood Estimation in Semiparametric Models

Thomas A. SEVERINI and Joan G. STANISWALIS*

Suppose the expected value of a response variable Y may be written $h(\mathbf{X}\beta + \gamma(\mathbf{T}))$ where $\mathbf{X}$ and $\mathbf{T}$ are covariates, each of which may be vector-valued, β is an unknown parameter vector, γ is an unknown smooth function, and h is a known function. In this article, we outline a method for estimating the parameter β, γ of this type of semiparametric model using a quasi-likelihood function. Algorithms for computing the estimates are given and the asymptotic distribution theory for the estimators is developed. The generalization of this approach to the case in which Y is a multivariate response is also considered. The methodology is illustrated on two data sets and the results of a small Monte Carlo study are presented.

KEY WORDS: Generalized linear models: Multivariate regression; Nonparametric regression; Partially linear models; Smoothing.

1. INTRODUCTION

Consider a model for the relationship between a dependent variable Y and an independent variable $\mathbf{Z}$, which may be multidimensional. Suppose that the covariate vector $\mathbf{Z}$ is written as $\mathbf{Z} = (\mathbf{X}, \mathbf{T})$ and suppose that the conditional expected value of Y given $\mathbf{Z}$ is written as $h(\mathbf{X}\beta + \gamma(\mathbf{T}))$, where β is an unknown parameter vector, γ is an unknown smooth function, and h is a known function. Hence there is a parametric model for the relationship between Y and $\mathbf{X}$, but for the relationship between Y and $\mathbf{T}$ a nonparametric relationship is used; the resulting model is then semiparametric.

The goal of this article is to present a method of estimation for this type of semiparametric model based on a quasi-likelihood function. A quasi-likelihood function has properties similar to those of a likelihood function, but requires only specification of the second-moment properties of Y, rather than the entire distribution (see, for example, McCullagh and Nelder 1989, chap. 9). Quasi-likelihood methods are useful because they can be used in cases where exact distributional information is not available and, because only second-moment assumptions are required, quasi-likelihood methods enjoy a certain robustness of validity. Hence the proposed method allows precise estimation of the relationship between the response and the covariate $\mathbf{X}$ without requiring a parametric specification for the relationship between the response and $\mathbf{T}$ and without requiring exact distributional assumptions.

For univariate Y, there have been several methods of estimation proposed for semiparametric generalized linear models, which are closely related to the models considered here. Robinson (1988) and Speckman (1988) considered estimation in models that are special cases of the model considered here. Green and Yandell (1985) considered estimation of both the parametric and nonparametric components of the model by maximizing a penalized likelihood function (see also Green 1987). Hastie and Tibshirani (1990, secs. 5.3 and 6.7) also considered this approach, along with methods of estimation based on their "backfitting algorithm," which has been successful in estimation of purely nonparametric generalized additive models. Hunsberger (1990) considered estimation of the nonparametric and parametric components of the model by maximizing a weighted likelihood function. Härdle and Stoker (1989) considered estimation in a class of semiparametric models using "average derivative estimation" (ADE). Although ADE is relatively easy to apply, it has several drawbacks; in particular, the parameter β is estimated only up to a scale normalization factor, and the method does not apply if any of the covariates in X are discrete or are functionally related. On the other hand, ADE can be applied when the function h is unknown, a situation not handled by the methods considered here. Related models have also been studied in the econometric literature (see, for example, Ichimura and Lee 1991, Powell, Stock, and Stoker 1989, and the references therein).

The estimation method proposed here is based on the concept of a "generalized profile likelihood," introduced by Severini and Wong (1992). In this method, first the parametric component of the model is held fixed and an estimate of the nonparametric component of the model is obtained using some type of smoothing method; this estimate depends on the value at which the parametric component was held fixed. This estimate is then used to construct a "profile likelihood" for the parametric component, using either a likelihood or quasi-likelihood function. This profile likelihood function is then used to obtain an estimator of the parametric component of the model, using a method such as maximum likelihood. The advantage of this estimation method is that it effectively separates the estimation problem into two parts, so that the nonparametric component is estimated using a nonparametric method, whereas the parametric component is estimated using a parametric method. Furthermore, the resulting estimator of the parametric component converges to the true parameter value at a $\sqrt{n}$ rate when the usual optimal rate for the smoothing parameter is used.

This article is organized as follows. Quasi-likelihood estimation in parametric models is reviewed in Section 2 and adapted to semiparametric models in Section 3. The large-sample theory is considered in Section 4. Estimation of the asymptotic covariance matrix of the estimator is discussed in Section 5, and computation of the estimates is addressed in Section 6. An application of the methodology, as well as some Monte Carlo results on the small-sample behavior of the estimator for the case of a univariate response, are pre-

* Thomas A. Severini is Assistant Professor, Department of Statistics, Northwestern University, Evanston, IL 60208. Joan G. Staniswalis is Associate Professor, Department of Mathematical Sciences, University of Texas at El Paso, El Paso, TX 79968. The work of Staniswalis was supported by National Science Foundation Grant DMS 9121555 and National Institutes of Health Grant S06GM08012-22. The authors thank the referees and associate editor for useful comments.

Journal of the American Statistical Association
June 1994, Vol. 89, No. 426, Theory and Methods

sented in Section 7. In Section 8 the methodology developed in Sections 1–7 for a univariate response Y is extended to the case of a multivariate Y. An application of the methodology for the case of a multivariate response is presented in Section 9. Finally, technical proofs and regularity conditions are given in the Appendix.

2. QUASI-LIKELIHOOD ESTIMATION

Let Y denote a random variable with distribution depending on a scalar parameter θ, and let $\mu = \mu(\theta)$ denote $E\{Y; \theta\}$. If the distribution of Y forms a one-dimensional exponential family with density $f(y; \theta) = \exp\{y\theta - b(\theta) + c(y)\}$ for some functions b, c, then given a random sample $Y_1, \ldots, Y_n$ from the distribution of Y, the maximum likelihood estimate (MLE) $\hat{\theta}$ of θ solves $\sum (Y_j - b'(\theta)) = 0$. Because $\mu(\theta) = b'(\theta)$ and $\text{var}\{Y; \theta\} = b''(\theta) \equiv V(\mu(\theta))$, $\hat{\theta}$ may also be obtained by solving

$$\sum \frac{(Y_j - \mu(\theta))\mu'(\theta)}{V(\mu(\theta))} = 0. \tag{1}$$

Alternatively, an estimate of θ may be obtained by simply assuming functional forms of μ and V and solving (1) without specifying the distribution of Y. This is the idea behind quasi-likelihood estimation. Note that only the first two moments of the distribution of Y need to be specified. For the special case in which the distribution of Y is from a one-parameter exponential family, quasi-likelihood and maximum likelihood estimation coincide, because the exponential family distribution is specified by its first two moments.

Because (1) plays the role of the first derivative of the log-likelihood function, the quasi-likelihood function is defined by

$$Q(\mu; y) = \int_{\mu}^{y} (s - y)/V(s)\, ds.$$

Under some regularity conditions, $\sum Q(\mu; Y_j)$ behaves like a log-likelihood function for μ based on $Y_1, \ldots, Y_n$ and $Q(\mu; y)$ behaves like the logarithm of a density function for Y. For instance, using integration by parts, $Q(\mu; y)$ may be taken to be

$$(y - \mu)\Omega(\mu) + \int^{\mu} \Omega(s)\, ds, \qquad \Omega(\mu) = \int^{\mu} 1/V(s)\, ds.$$

Let $Q^*(\mu) = E\{Q(\mu; Y); \mu_0\}$, where μ_0 denotes the true parameter value. It follows immediately that $Q^*(\mu)$ has a unique maximum at $\mu = \mu_0$, and hence, under standard regularity conditions, the quasi-likelihood estimator $\hat{\mu}$ of μ is a consistent estimator of μ_0 and $\sqrt{n}(\hat{\mu} - \mu_0)$ is asymptotically normally distributed (see McCullagh 1983; Moore 1986).

Now consider the regression setting. Let $(\mathbf{T}_j, \mathbf{X}_j, Y_j)$, $j = 1, \ldots, n$, denote independent copies of $(\mathbf{T}, \mathbf{X}, Y)$, where Y is a scalar response variable, $\mathbf{X}$ is a $1 \times p$ vector of covariates, and $\mathbf{T}$ is a $1 \times q$ vector of covariates. Let $E\{Y|\mathbf{X}, \mathbf{T}\} = \mu(\mathbf{X}, \mathbf{T})$ and $\text{var}\{Y|\mathbf{X}, \mathbf{T}\} = \sigma^2 V\{\mu(\mathbf{X}, \mathbf{T})\}$, where σ^2 is an unknown parameter, μ is an unknown function, and V is a known function. Suppose further that $\mu(\mathbf{X}, \mathbf{T}) = h(\mathbf{T}\alpha + \mathbf{X}\beta)$, where h is a known monotone function and $\alpha \equiv (\alpha_1, \ldots, \alpha_q)'$ and $\beta \equiv (\beta_1, \ldots, \beta_p)'$ are unknown parameter vectors. Note that if the conditional distribution of Y given $\mathbf{X}, \mathbf{T}$ is an exponential family distribution, then the MLE's of α and β can be obtained by maximizing $\sum Q(h(\mathbf{T}_j\alpha + \mathbf{X}_j\beta); Y_j)$ with respect to α and β. This maximization may be carried out by solving the likelihood equations

$$\frac{\partial}{\partial \alpha_k} \sum Q(h(\mathbf{T}_j\alpha + \mathbf{X}_j\beta); Y_j) = 0, \qquad k = 1, \ldots, q \tag{3}$$

and

$$\frac{\partial}{\partial \beta_k} \sum Q(h(\mathbf{T}_j\alpha + \mathbf{X}_j\beta); Y_j) = 0, \qquad k = 1, \ldots, p \tag{4}$$

for α, β. Using the quasi-likelihood approach, we may choose to solve (3) and (4) by specifying the functions h and V without making any other assumptions about the distribution of Y.

Note that one approach to maximizing the quasi-likelihood function is to first solve (3) for fixed β to obtain $\hat{\alpha}_\beta$ and then maximize $\sum Q(h(\mathbf{T}_j\hat{\alpha}_\beta + \mathbf{X}_j\beta); Y_j)$ with respect to β. Hence instead of solving (4), we solve

$$\frac{\partial}{\partial \beta_k} \sum Q(h(\mathbf{T}_j\hat{\alpha}_\beta + \mathbf{X}_j\beta); Y_j) = 0, \qquad k = 1, \ldots, p \tag{5}$$

to obtain $\hat{\beta}$. When (3) and (4) represent the likelihood equations for α and β, then the resulting estimators of α and β are simply the MLE's calculated by a two-stage procedure; similarly, in the quasi-likelihood case the resulting estimates are the quasi-likelihood estimates of α and β.

3. QUASI-LIKELIHOOD ESTIMATION IN SEMIPARAMETRIC MODELS

Now suppose that $E\{Y|\mathbf{X}, \mathbf{T}\} = \mu(\mathbf{X}, \mathbf{T})$ and $\text{var}\{Y|\mathbf{X}, \mathbf{T}\} = \sigma^2 V(\mu)$, where $\mu(\mathbf{X}, \mathbf{T}) = h(\gamma(\mathbf{T}) + \mathbf{X}\beta)$, γ is an unknown smooth function from $\Re^q$ to $\Re$, and β is an unknown $p \times 1$ parameter vector. Assume that $\mathbf{T}$ takes values in $\mathcal{T}$, a closed rectangle in $\Re^q$. Hence the model is semiparametric; typically, β is the parameter of interest and γ plays the role of an (infinite-dimensional) nuisance parameter. To estimate β and γ, we use the approach of the previous section. First, for fixed β we estimate γ as a function of β to obtain $\hat{\gamma}_\beta$; note that this is a nonparametric estimation problem. Then, setting $\gamma = \hat{\gamma}_\beta$, we estimate the parametric component β; note that this is a parametric problem. Hence this approach to estimation in the semiparametric model effectively separates the estimation problem into parametric and nonparametric components (see Severini and Wong 1992 for further details).

To estimate γ for fixed β, we generalize the approach of Staniswalis (1989) for likelihood-based estimation and use what might be called a weighted quasi-likelihood. Let W denote a kernel on $\Re^q$ that is a direct product of a kernel w on the real line, and let $b > 0$ denote a sequence of bandwidths depending on n. For each fixed $\mathbf{t}$ and β, let $\hat{\gamma}_\beta(\mathbf{t})$ denote the solution in η of

$$\sum W\left(\frac{\mathbf{t} - \mathbf{T}_j}{b}\right) \frac{\partial}{\partial \eta} Q(h(\eta + \mathbf{X}_j\beta); Y_j) = 0; \tag{6}$$

note that this equation is analogous to Equation (3). Given the estimator $\hat{\gamma}_\beta(\mathbf{t})$, an estimate of β, $\hat{\beta}$ is then obtained by solving

$$\frac{\partial}{\partial \beta_k} \sum Q(h(\hat{\gamma}_\beta(\mathbf{T}_j) + \mathbf{X}_j\beta); Y_j) = 0, \qquad k = 1, \ldots, p, \quad (7)$$

which are analogous to Equation (5).

Although (6) may be used to obtain $\hat{\gamma}_\beta(\mathbf{t})$, the bias of the resulting estimator is higher for $\mathbf{t}$ near the boundary of $\mathcal{T}$ than for $\mathbf{t}$ away from the boundary; this is undesirable for both practical and theoretical reasons. There are several ways to reduce this bias near the boundary. In the univariate case, when $q = 1$, either a boundary-corrected kernel estimator (Rice 1984) or locally linear kernel estimator (Fan 1992) may be used. Although either of these methods may be extended to the multivariate case, the resulting technical details for the development of the asymptotic theory become cumbersome. Another approach is to use "trimming" in which in forming the sum in (7), only those observations with $\mathbf{T}_i$ away from the boundary are used. Although this approach is less desirable from a practical point of view, it has the advantage that the theoretical results can be presented in a clear form. Hence here we will use trimming in describing the estimation procedure and in presenting the asymptotic theory of Section 4.

Hence let $\mathcal{T}_0$ denote a compact subset of $\operatorname{int}(\mathcal{T})$, and let $I_j = 1$ if $\mathbf{T}_j \in \mathcal{T}_0$ and 0 otherwise. Then, instead of (7), we will solve

$$\frac{\partial}{\partial \beta_k} \sum I_j Q(h(\hat{\gamma}_\beta(\mathbf{T}_j) + \mathbf{X}_j\beta); Y_j)) = 0, \qquad k = 1, \ldots, p. \quad (8)$$

In some cases, (6) or (8) may be solved explicitly to obtain closed-form solutions for the estimates; algorithms for computing the estimates in the general case will be considered in detail in Section 6. We now consider some examples.

Example 1. Let $h(s) = s$ and $V(\mu) = 1$; the resulting quasi-likelihood function is the log-likelihood function for normally distributed data. In this case, Equation (6) may be solved explicitly to obtain

$$\hat{\gamma}_\beta(\mathbf{t}) = \frac{\sum_j W\left(\frac{\mathbf{t} - \mathbf{T}_j}{b}\right)(Y_j - \mathbf{X}_j\beta)}{\sum W\left(\frac{t - T_j}{b}\right)}. \quad (9)$$

Equation (8) may also be solved explicitly to obtain

$$\hat{\beta} = [(\mathbf{X} - \hat{\mathbf{X}})'\mathbf{D}(\mathbf{X} - \hat{\mathbf{X}})]^{-1}(\mathbf{X} - \hat{\mathbf{X}})'\mathbf{D}(\mathbf{Y} - \hat{\mathbf{Y}}), \quad (10)$$

where $\mathbf{X}$ represents the matrix with ith row $\mathbf{X}_i$, $\hat{\mathbf{X}}$ represents the matrix with ith row

$$\hat{\mathbf{X}}_i = \frac{\sum_j W\left(\frac{\mathbf{T}_i - \mathbf{T}_j}{b}\right)\mathbf{X}_j}{\sum W\left(\frac{\mathbf{T}_i - \mathbf{T}_j}{b}\right)},$$

$\mathbf{Y}$ represents the column vector with ith element Y_i, $\hat{\mathbf{Y}}$ is defined analogously to $\hat{\mathbf{X}}$, and $\mathbf{D}$ is an $n \times n$ diagonal matrix with jth diagonal element I_j. The estimator $\hat{\beta}$ is essentially the estimator studied by Robinson (1988) and Speckman (1988).

Example 2. Let $h(s) = \exp(s)$ and $V(\mu) = \mu^2$, $\mu > 0$; the resulting quasi-likelihood function is the log-likelihood function for data distributed according to a gamma distribution. In this case, Equation (6) may be solved explicitly to obtain

$$\hat{\gamma}_\beta(t) = \log\left(\frac{\sum W\left(\frac{\mathbf{t} - \mathbf{T}_j}{b}\right)\exp\{-\mathbf{X}_j\beta\}Y_j}{\sum W\left(\frac{\mathbf{t} - \mathbf{T}_j}{b}\right)}\right).$$

Here an explicit solution to Equation (8) is not available, so an iterative approach to obtaining $\hat{\beta}$ is needed.

Example 3. Let $h(s) = \exp\{s\}/(1 + \exp\{s\})$ and $V(\mu) = \mu(1 - \mu)$, $0 < \mu < 1$. For $\sigma = 1$, the resulting quasi-likelihood function is the log-likelihood function for data distributed according to a binomial distribution; otherwise, σ^2 represents an overdispersion parameter. Neither Equation (6) nor Equation (8) can be solved explicitly.

To estimate σ^2, we use an estimator analogous to the one used in standard parametric quasi-likelihood estimation. Let $\hat{h}_j = h(\mathbf{X}_j\hat{\beta} + \hat{\gamma}_{\hat{\beta}}(\mathbf{T}_j))$. Then σ^2 can be estimated by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_j (Y_j - \hat{h}_j)^2/V(\hat{h}_j).$$

4. ASYMPTOTIC PROPERTIES OF THE ESTIMATORS

Throughout the remainder of the article we assume that the regularity conditions given in the Appendix are in effect. Let γ_0, β_0, and σ_0^2 denote the true parameter values so that $E_0\{Y|\mathbf{X}, \mathbf{T}\} = h(\gamma_0(\mathbf{T}) + \mathbf{X}\beta_0)$ and $\operatorname{var}_0\{Y|\mathbf{X}, \mathbf{T}\} = \sigma_0^2 V(h(\gamma_0(\mathbf{T}) + X\beta_0))$, where E_0 and var_0 denote expected value and variance under the true model. Define

$$M(\eta; \beta, \mathbf{t}) = E_0\left\{\frac{\partial}{\partial \eta} Q(h(\eta + \mathbf{X}_1\beta); Y_1)|\mathbf{T} = \mathbf{t}\right\}$$

and let $\gamma_\beta(\mathbf{t})$ denote the solution in η of $M(\eta; \beta, \mathbf{t}) = 0$ for each fixed β, $\mathbf{t}$. Note that $\hat{\gamma}_\beta(\mathbf{t})$ is an estimator of $\gamma_\beta(\mathbf{t})$, and hence $\gamma_\beta(\mathbf{t})$ plays a central role in the large-sample properties of $\hat{\beta}$. Let $\mathbf{\Sigma}_0$ denote the $p \times p$ matrix such that $\mathbf{\Sigma}_0^{-1}$ has (i, j)th value

$$E_0\left\{I_1 \frac{\partial^2}{\partial \beta_i \partial \beta_j} Q(h(\hat{\gamma}_\beta(\mathbf{T}_1) + \mathbf{X}_1\beta); Y_1))\right\}.$$

The proof of the following proposition is discussed in the Appendix.

Proposition 1. Let $N(\mathbf{0}, \mathbf{\Sigma})$ represent a p-dimensional random vector with mean vector $(\mathbf{0}, \ldots, \mathbf{0})'$ and covariance matrix $\mathbf{\Sigma}$. Then

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \sigma_0^2\mathbf{\Sigma}_0) \quad \text{as} \quad n \to \infty.$$

Proposition 2. Let $\|g\| = \sup_{s\in\mathcal{T}_0}|g(\mathbf{s})|$. Then

$$\|\hat{\gamma}_{\hat{\beta}} - \gamma_0\| = o_p(n^{-1/4}) \quad \text{and}$$
$$\hat{\sigma}^2 = \sigma_0^2 + o_p(1) \quad \text{as} \quad n \to \infty.$$

The proof follows immediately from condition (7) of the Appendix along with the fact that $\hat{\beta} = \beta_0 + O_p(n^{-1/2})$.

5. ESTIMATION OF THE ASYMPTOTIC COVARIANCE MATRIX OF $\hat{\beta}$

For inference about β based on the asymptotic distribution of $\hat{\beta}$, an estimate of the covariance matrix $\mathbf{\Sigma}_0$ is needed. Let $G(\cdot) = h'(\cdot)/V(h(\cdot))$, and let $\gamma'_{j\beta_0}(\mathbf{t}) = d\gamma_{\beta_0}(\mathbf{t})/d\beta_j$. Then $\mathbf{\Sigma}_0^{-1}$ has (i,j)th element

$$E_0\{I_1 G(\gamma_0(\mathbf{T}_1) + \mathbf{X}_1\beta_0)(X_{1i} + \gamma'_{i\beta_0}(\mathbf{T}_1)) \times (X_{1j} + \gamma'_{j\beta_0}(\mathbf{T}_1))h'(\gamma_0(\mathbf{T}_1) + \mathbf{X}_1\beta_0)\};$$

here $\mathbf{X}_1 = (X_{11}, \ldots, X_{1p})$. Let $\hat{\mathbf{\Sigma}}_0$ denote the $p \times p$ matrix such that $\hat{\mathbf{\Sigma}}_0^{-1}$ has (i,j)th element

$$\frac{1}{n}\sum_k I_k G(\hat{h}_k)(X_{ki} - \hat{X}_{ki})(X_{kj} - \hat{X}_{kj})h'(\hat{\gamma}_{\hat{\beta}}(\mathbf{T}_k) + \mathbf{X}_k\hat{\beta}),$$

where $\hat{X}_{ki} = -\hat{\gamma}'_{i\hat{\beta}}(\mathbf{T}_k) - d\hat{\gamma}_{\beta}(\mathbf{T}_k)/d\beta_i$. It follows, using standard arguments, that $\hat{\mathbf{\Sigma}}_0 \xrightarrow{p} \mathbf{\Sigma}_0$ as $n \to \infty$.

The covariance matrix $\mathbf{\Sigma}_0$ and its estimator $\hat{\mathbf{\Sigma}}_0$ are calculated under the assumption that the variance function of the model, $V(\cdot)$, is correctly specified. Because in practice strong prior information about the variance function is often not available, it is useful to have an estimator of the asymptotic covariance matrix of $\hat{\beta}$ that does not require this assumption. Let $V_0(\cdot)$ denote the unknown true variance function so that $\text{var}\{Y|\mathbf{X},\mathbf{T}\} = V_0(\mu)$, where $\mu \equiv \mu(\mathbf{X},\mathbf{T}) = E(Y|\mathbf{X},\mathbf{T}) = h(\gamma(\mathbf{T}) + X\beta)$; here the overdispersion parameter may be taken to be 1. But in forming the quasi-likelihood function, the function V is used in place of V_0. The matrices $\mathbf{\Sigma}_0$ and $\hat{\mathbf{\Sigma}}_0$ are based on the assumption that $V = V_0$. We now consider the case in which this assumption does not necessarily hold.

Let $\mathbf{\Sigma}_1$ denote the $p \times p$ matrix such that $\mathbf{\Sigma}_1^{-1}$ has (i,j)th element

$$E_0\{I_1 G(\gamma_0(\mathbf{T}_1) + \mathbf{X}_1\beta_0)^2(X_{1i} + \gamma'_{i\beta_0}(\mathbf{T}_1)) \times (X_{1j} + \gamma'_{j\beta_0}(\mathbf{T}_1))V_0(\gamma_0(\mathbf{T}_1) + \mathbf{X}_1\beta_0)\}.$$

A generalization of Proposition 1 shows that $\hat{\beta}$ has asymptotic covariance matrix $\mathbf{\Sigma} \equiv \mathbf{\Sigma}_0\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_0$; this result does not assume that the variance function V used in forming the quasi-likelihood function is correctly specified. If V is correctly specified so that $\sigma_0^2 V = V_0$, then $\mathbf{\Sigma}_1 = \sigma_0^{-2}\mathbf{\Sigma}_0$ so that $\mathbf{\Sigma} = \sigma_0^2\mathbf{\Sigma}_0$—the result given in Section 4.

To estimate $\mathbf{\Sigma}$, an estimator of $\mathbf{\Sigma}_1$ is needed. Let $\hat{\mathbf{\Sigma}}_1$ denote the $p \times p$ matrix such that the (i,j)th element of $\hat{\mathbf{\Sigma}}_1^{-1}$ is given by

$$\frac{1}{n}\sum\{I_k G(\gamma_0(\mathbf{T}_k) + \mathbf{X}_k\beta_0)^2(X_{ki} + \gamma'_{i\beta_0}(\mathbf{T}_k)) \times (X_{kj} + \gamma'_{j\beta_0}(\mathbf{T}_k))(Y_k - \hat{h}_k)^2\},$$

where $\hat{h}_k = h(\mathbf{X}_k\hat{\beta} + \hat{\gamma}_{\hat{\beta}}(\mathbf{T}_k))$. Let $\hat{\mathbf{\Sigma}} = \hat{\mathbf{\Sigma}}_0\hat{\mathbf{\Sigma}}_1^{-1}\hat{\mathbf{\Sigma}}_0$; it is straightforward to show that $\hat{\mathbf{\Sigma}}$ is a consistent estimator of $\mathbf{\Sigma}$.

6. COMPUTATION OF THE ESTIMATES

We now consider an algorithm for computing the estimates. Let

$$\psi_1(\eta;\beta,\mathbf{t}) = \sum_j W\left(\frac{\mathbf{t}-\mathbf{T}_j}{b}\right)G(\eta + \mathbf{X}_j\beta)(Y_j - h(\eta + \mathbf{X}_j\beta))$$

and

$$\psi_{2k}(\beta,\varphi_\beta(\cdot)) = \sum_j I_j G(\varphi_\beta(\mathbf{T}_j) + \mathbf{X}_j\beta)(X_{jk} + \varphi'_{k\beta}(\mathbf{T}_j)) \times (Y_j - h(\varphi_\beta(\mathbf{T}_j) + \mathbf{X}_j\beta)), \quad k = 1, \ldots, p;$$

here $\varphi_\beta(\mathbf{t})$ is an arbitrary function from $\Re^q$ to $\Re$ for each β and a differentiable function of β for each $\mathbf{t}$. We are writing $\varphi'_{k\beta} = \partial\varphi_\beta/\partial\beta_k$. Then β and γ are estimated by the following procedure:

a. For each $\mathbf{t}, \beta$, calculate $\hat{\gamma}_\beta(\mathbf{t})$ by solving $\psi_1(\eta;\beta,\mathbf{t}) = 0$ for η. This corresponds to (6).

b. Estimate β by solving $\psi_{2k}(\beta;\hat{\gamma}_\beta) = 0$, $k = 1, \ldots, p$ for β. This corresponds to (8).

c. Estimate γ by $\hat{\gamma}_{\hat{\beta}}$.

Suppose that $\hat{\gamma}_\beta(\mathbf{t})$ and $\hat{\gamma}'_{k\beta}(\mathbf{t})$ are available for each $\mathbf{t}, \beta$; consider step b, calculation of $\hat{\beta}$. Let $\mathbf{A}$ denote the $p \times p$ matrix with the (i,j)th element given by

$$\begin{aligned}\mathbf{A}_{ij} &= E_0\left\{\frac{d}{d\beta_j}\psi_{2i}(\beta;\gamma_\beta)|_{\beta=\beta_0}|\mathbf{T}_1, \ldots, \mathbf{T}_n; \mathbf{X}_1, \ldots, \mathbf{X}_n\right\} \\ &= -\sum_k I_k G(\gamma_0(\mathbf{T}_k) + \mathbf{X}_k\beta_0)(X_{ki} + \gamma'_{i\beta_0}(\mathbf{T}_k)) \\ &\qquad \times (X_{kj} + \gamma'_{j\beta_0}(\mathbf{T}_k))h'(\gamma_0(\mathbf{T}_k) + \mathbf{X}_k\beta_0), \end{aligned} \tag{12}$$

and let $\hat{\mathbf{A}}(\beta)$ denote the $p \times p$ matrix with (i,j)th element given by (11) with β_0 replaced by β and γ_0 replaced by $\hat{\gamma}_\beta$. Let $\hat{\mathbf{B}}$ denote the $p \times 1$ vector with kth element given by $\psi_{2k}(\beta;\hat{\gamma}_\beta)$. Then, using Fisher's scoring method (McCullagh and Nelder 1989, sec. 2.5), an initial estimate $\tilde{\beta}$ can be updated to $\tilde{\beta}^*$ using

$$\tilde{\beta}^* = \tilde{\beta} - \hat{\mathbf{A}}(\tilde{\beta})^{-1}\hat{\mathbf{B}}(\tilde{\beta}). \tag{12}$$

This iteration can be continued until convergence. Note that the estimated asymptotic covariance matrix of $\hat{\beta}$ can now be easily determined using $\hat{\sigma}^2\hat{\mathbf{\Sigma}}_0 = -n\hat{\sigma}^2\hat{\mathbf{A}}(\hat{\beta})^{-1}$; this assumes that V is correctly specified.

To obtain a starting value for the iteration, let $C_i = h^{-1}(Y_i)$ and consider the model with $E\{C_i|\mathbf{X}_i,\mathbf{T}_i\} = \gamma(\mathbf{T}_i) + \mathbf{X}_i\beta$ and $\text{var}\{C_i|\mathbf{X}_i,\mathbf{T}_i\} = \sigma^2$; this is the model considered in Example 1. By (10), using quasi-likelihood estimation for this model leads to

$$\hat{\beta} = [(\mathbf{X}-\hat{\mathbf{X}})'\mathbf{D}(\mathbf{X}-\hat{\mathbf{X}})]^{-1}(\mathbf{X}-\hat{\mathbf{X}})'\mathbf{D}(\mathbf{C}-\hat{\mathbf{C}}),$$

where $\mathbf{C}$ represents the column vector with ith element C_i, $\hat{\mathbf{C}}$ represents the column vector with ith element

$$\hat{C}_i = \frac{\sum_j W\left(\frac{\mathbf{T}_i-\mathbf{T}_j}{b}\right)C_j}{\sum_j W\left(\frac{\mathbf{T}_i-\mathbf{T}_j}{b}\right)},$$

and $\mathbf{D}$ is an $n \times n$ diagonal matrix with jth diagonal element I_j. Note that average derivative estimation (Härdle and Stoker 1989) could be used on the model for the C_i to obtain the starting values for β, provided that the covariates in $\mathbf{X}$ are not discrete or functionally related.

In some cases, $\psi_1(\eta; \beta, \mathbf{t}) = 0$ can be solved explicitly for η, for each β, $\mathbf{t}$, to obtain closed-form expressions for $\hat{\gamma}_\beta(\mathbf{t})$ and $\hat{\gamma}'_{k\beta}(\mathbf{t})$; in these cases, (12) can be used with those expressions to obtain $\hat{\beta}$.

Example 2 (continued). An expression for $\hat{\gamma}_\beta(\mathbf{t})$ was given in Section 3. Let $\hat{\mathbf{X}}(\beta)$ denote the $n \times p$ matrix with ith row

$$\hat{\mathbf{X}}_i = \frac{\sum_j W\left(\frac{\mathbf{T}_i - \mathbf{T}_j}{b}\right)\exp\{-\mathbf{X}_j\beta\}\mathbf{X}_j Y_j}{\sum W\left(\frac{\mathbf{T}_i - \mathbf{T}_j}{b}\right)\exp\{-\mathbf{X}_j\beta\}Y_j}.$$

Then $\hat{\mathbf{A}}(\beta) = -(\mathbf{X} - \hat{\mathbf{X}}(\beta))'\mathbf{D}(\mathbf{X} - \hat{\mathbf{X}}(\beta))$, and $\hat{\mathbf{B}}(\beta)$ has ith element

$$\sum_j I_j \exp\{-(\hat{\gamma}_\beta(\mathbf{T}_j) + \mathbf{X}_j\beta)\}(X_{ji} - \hat{X}_{ji}) \times (Y_j - \exp\{\hat{\gamma}_\beta(\mathbf{T}_j) + \mathbf{X}_j\beta\}).$$

$\hat{\beta}$ can then be calculated using the iteration (12).

In general, iterative methods are also needed to calculate $\hat{\gamma}_\beta(\mathbf{t})$ and $\hat{\gamma}'_{k\beta}(\mathbf{t})$. Fix β and $\mathbf{t}$ and let $\eta_0 = \gamma_\beta(\mathbf{t})$. Consider solving $\psi_1(\eta; \beta, \mathbf{t}) = 0$ for $\eta = \hat{\gamma}_\beta(\mathbf{t})$. Because

$$E_0\left\{\frac{d}{d\eta}\psi_1(\eta; \beta, \mathbf{t})|_{\eta=\eta_0}|\mathbf{X}_1, \ldots, \mathbf{X}_n\right\} = -\sum_j W\left(\frac{\mathbf{t} - \mathbf{T}_j}{b}\right)G(\eta_0 + \mathbf{X}_j\beta)h'(\eta_0 + \mathbf{X}_j\beta),$$

an initial estimate $\tilde{\eta}$ can be updated to $\tilde{\eta}^*$ using Fisher's scoring method:

$$\tilde{\eta}^* = \tilde{\eta} + \frac{\sum_j W\left(\frac{\mathbf{t} - \mathbf{T}_j}{b}\right)G(\tilde{\eta} + \mathbf{X}_j\beta)(Y_j - h(\tilde{\eta} + X_j\beta))}{\sum_j W\left(\frac{\mathbf{t} - \mathbf{T}_j}{b}\right)G(\tilde{\eta} + X_j\beta)h'(\tilde{\eta} + X_j\beta)}. \tag{13}$$

To obtain a starting value for the iteration, we adapt the approach of McCullagh and Nelder (1989, sec. 2.5). Hence suppose that for fixed β, $\tilde{\eta}_0$ satisfies $\tilde{\eta}_0 + \mathbf{X}_j\beta = h^{-1}(Y_j)$, $j = 1, \ldots, n$; note that the existence of such an $\tilde{\eta}_0$ is entirely hypothetical and is used only as a device to determine a starting value. Equation (13) then may be written as

$$\tilde{\eta}^* = \frac{\sum_j W\left(\frac{\mathbf{t} - \mathbf{T}_j}{b}\right)G(\tilde{\eta} + \mathbf{X}_j\beta)h'(\tilde{\eta} + \mathbf{X}_j\beta) \times \left[\tilde{\eta} + \frac{Y_j - h(\tilde{\eta} + \mathbf{X}_j\beta)}{h'(\tilde{\eta} + \mathbf{X}_j\beta)}\right]}{\sum_j W\left(\frac{\mathbf{t} - \mathbf{T}_j}{b}\right)G(\tilde{\eta} + \mathbf{X}_j\beta)h'(\tilde{\eta} + X_j\beta)};$$

hence $\tilde{\eta}_0$ may be updated to

$$\tilde{\eta} = \frac{\sum W\left(\frac{\mathbf{t} - \mathbf{T}_j}{b}\right)G(C_j)h'(C_j)(C_j - \mathbf{X}_j\beta)}{\sum W\left(\frac{\mathbf{t} - \mathbf{T}_j}{b}\right)G(C_j)h'(C_j)}, \tag{14}$$

which can be used as a starting value; recall that $C_j = h^{-1}(Y_j)$.

To calculate $\hat{\beta}$, $\hat{\gamma}'_{k\beta}(\mathbf{t}) = d\hat{\gamma}_\beta(\mathbf{t})/d\beta_k$, $k = 1, \ldots, p$, are also needed. Because $\hat{\gamma}_\beta(\mathbf{t})$ satisfies $\psi_1(\hat{\gamma}_\beta(\mathbf{t}); \beta, \mathbf{t}) = 0$ for all β, $\mathbf{t}$ it follows that $d\psi_1(\hat{\gamma}_\beta(\mathbf{t}); \beta, \mathbf{t})/d\beta = 0$ for all β, $\mathbf{t}$. Solving this for $\hat{\gamma}'_{k\beta}(\mathbf{t})$ yields

$$\hat{\gamma}'_{k\beta}(t) = -\frac{\sum_j W\left(\frac{\mathbf{t} - \mathbf{T}_j}{b}\right)Z_j(\mathbf{t}, \beta)\mathbf{X}_{jk}}{\sum_j W\left(\frac{\mathbf{t} - \mathbf{T}_j}{b}\right)Z_j(\mathbf{t}, \beta)}, \tag{15}$$

$$Z_j(\mathbf{t}, \beta) = G'(\hat{\gamma}_\beta(t) + \mathbf{X}_j\beta)(Y_j - h(\hat{\gamma}_\beta(\mathbf{t}) + \mathbf{X}_j\beta)) - G(\hat{\gamma}_\beta(\mathbf{t}) + \mathbf{X}_j\beta)h'(\hat{\gamma}_\beta(\mathbf{t}) + \mathbf{X}_j\beta).$$

Example 3 (continued). Fix β, $\mathbf{t}$; consider the calculation of $\hat{\eta} = \hat{\gamma}_\beta(\mathbf{t})$. For this choice of h and V, the iteration (13) becomes

$$\tilde{\eta}^* = \tilde{\eta} + \frac{\sum_j W\left(\frac{\mathbf{t} - \mathbf{T}_j}{b}\right)(Y_j - h(\tilde{\eta} + \mathbf{X}_j\beta))}{\sum W\left(\frac{\mathbf{t} - \mathbf{T}_j}{b}\right)h(\tilde{\eta} + \mathbf{X}_j\beta)/(1 + \exp\{\tilde{\eta} + \mathbf{X}_j\beta\})}.$$

A starting value for the iteration is given by

$$\tilde{\eta} = \frac{\sum W\left(\frac{\mathbf{t} - \mathbf{T}_j}{b}\right)Y_j(1 - Y_j)\left[\log\left(\frac{Y_j}{1 - Y_j}\right) - \mathbf{X}_j\beta\right]}{\sum W\left(\frac{\mathbf{t} - \mathbf{T}_j}{b}\right)Y_j(1 - Y_j)};$$

a small adjustment is needed in calculating $\log(Y_j/(1 - Y_j))$ for $Y_j = 0, 1$. Here $Z_j(\mathbf{t}, \beta) = -\exp\{\hat{\gamma}_\beta(\mathbf{t}) + \mathbf{X}_j\beta\}/(1 + \exp\{\hat{\gamma}_\beta(\mathbf{t}) + \mathbf{X}_j\beta\})^2$, which can be used to calculate $\hat{\gamma}'_{k\beta}(\mathbf{t})$ given $\hat{\gamma}_\beta(\mathbf{t})$. Let $\hat{\mathbf{X}}(\beta)$ denote the $n \times p$ matrix with ith row

$$\hat{\mathbf{X}}_i(\beta) = \frac{\sum_j W\left(\frac{\mathbf{T}_i - \mathbf{T}_j}{b}\right)Z_j(\mathbf{T}_i, \beta)\mathbf{X}_j}{\sum_j W\left(\frac{\mathbf{T}_i - \mathbf{T}_j}{b}\right)\mathbf{Z}_j(\mathbf{T}_i, \beta)},$$

and let $\mathbf{D}(\beta)$ denote the $n \times n$ diagonal matrix with jth diagonal element $I_j Z_j(\mathbf{T}_j, \beta)$. Then

$$\hat{\mathbf{A}}(\beta) = (\mathbf{X} - \hat{\mathbf{X}}(\beta))'\mathbf{D}(\beta)(\mathbf{X} - \hat{\mathbf{X}}(\beta)),$$

and $\hat{\mathbf{B}}(\beta)$ is the $p \times 1$ vector with ith element

$$\sum_j (X_{ji} - \hat{X}_{ji})(Y_j - h(\hat{\gamma}_\beta(\mathbf{T}_j) + \mathbf{X}_j\beta)).$$

$\hat{\beta}$ can now be calculated using (12).

The remaining computational issue is the selection of the bandwidth, b, used in the calculation of $\hat{\gamma}_\beta$. Consider the linear regression of $C_1, \ldots, C_n$ on $\mathbf{X}_1, \ldots, \mathbf{X}_n$ without an intercept term, and let $R_1, \ldots, R_n$ denote the residuals from

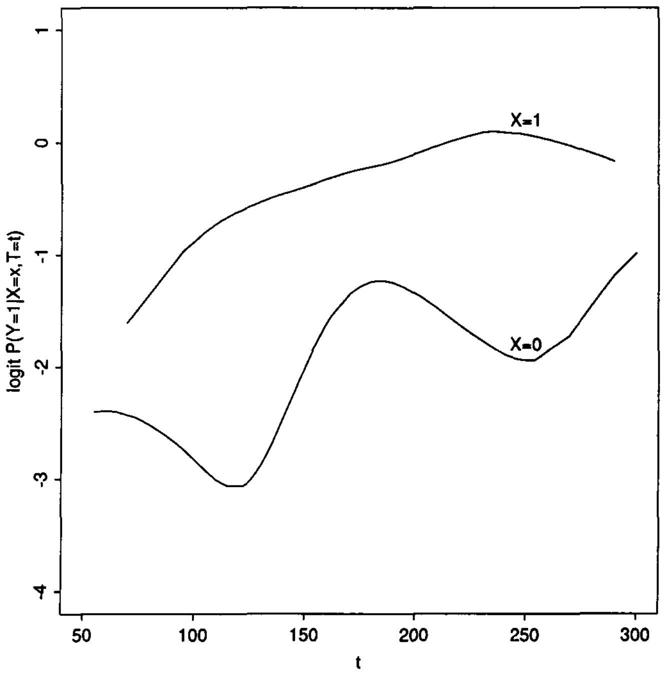

Figure 1. Nonparametric Estimate of Logit P(Y = 1 | X = x, T = T).

that regression. Under the semiparametric model, we expect $E(R_j|\mathbf{T}_j)$ to have smoothness properties similar to those of $\gamma(\mathbf{t})$. Hence we may choose the degree of smoothing for the semiparametric model by choosing the appropriate degree of smoothing in the nonparametric regression of R_j on $\mathbf{T}_j$. Any method of selecting b in a nonparametric regression problem may now be used; in the application of Section 6, cross-validation was used and gave good results. However, the theoretical properties of bandwidth selection procedures in semiparametric models are not yet known.

7. UNIVARIATE APPLICATIONS

7.1 Torrey Yucca Data

The data analyzed in this section were collected by the Environmental Management Office in Fort Bliss, Texas. Between 2,000 and 3,000 acres of the Dona Ana Firing Range on the Fort Bliss Military Reservation were burned during March 23-25, 1988. The data consist of measurements made on the Torrey yucca plants located in the burn site and in an unburned area directly west of the study sites. Of interest is whether Torrey yucca plants in the burned area are more likely to reproduce by root sprouts than are those in the unburned area. Roughly half of the $n = 208$ Torrey yucca plants sampled were unburned.

Let Y denote an indicator random variable that takes the value 1 if the randomly selected Torrey yucca plant has a root sprout and 0 otherwise. Let X take the value 1 if the Torrey yucca plant is in the burn site and 0 otherwise, and let T denote the height of the plant, which varies between 50 and 300 cm. We modeled $\mu(X, T) = P(Y = 1 | X, T)$ as in Example 3, using $h(s) = \exp\{s\}/(1 + \exp\{s\})$ and $V(\mu) = \mu(1 - \mu)$.

Exploratory plots of the data are given in Figures 1 and 2. Figure 1 is a nonparametric estimate of γ computed separately for the burned and unburned plants using the boundary-corrected smoother $\text{logit}(\sum w_jY_j/\sum w_j)$, where $w_j = W((t - T_j)/b)$; in each case, the bandwidth was selected by cross-validation. Note that the two curves are increasing at about the same rate, but that the shape of the estimate based on the data from the unburned site has peaks and valleys around $t = 125, 175, 250$ that are not present in the other estimate. Figure 2 contains plots of the conditional density of T given X. Recall that the bias of the Nadaraya–Watson kernel estimator depends on both the shape of the regression function and the shape of the density of T (Chu and Marron 1991). Note that the density estimate based on the data from the unburned site has peaks and valleys around $t = 100, 150, 200$ that are not present in the other density estimate; these differences in the density estimates help explain the differences in the shape of the estimates given in Figure 1. Hence we were willing to assume that the two underlying curves had the same shape but were just shifted by a constant; the biologist that collected the data was comfortable with this assumption.

The kernel $W(v) = \frac{15}{16}(1 - v^2)^2 I_{[-1,1]}(v)$ with the boundary modification of Rice (1984) was used in the estimation procedure. Trimming of the 3 lowest and 11 highest values of T in Equation (8) was needed for this data, to avoid overflow errors (Härdle and Stoker 1989). The programs required to carry out the computations were written in SAS and FORTRAN and were executed on a Sun SPARCstation 1+.

Prior to performing the iterations needed for computing $\hat\beta$, the bandwidth b and an initial estimate $\hat\beta$ of β were computed as follows. The bandwidth was selected using cross-validation. For $j = 1, \ldots, n$, let $C_j = \log(.1/.9)$ if $Y_j = 0$ and $\log(.9/.1)$ otherwise. Compute the residuals from the simpler linear regression of $C_1, \ldots, C_n$ on $X_1, \ldots, X_n$ for a model without an intercept term. The bandwidth was de-

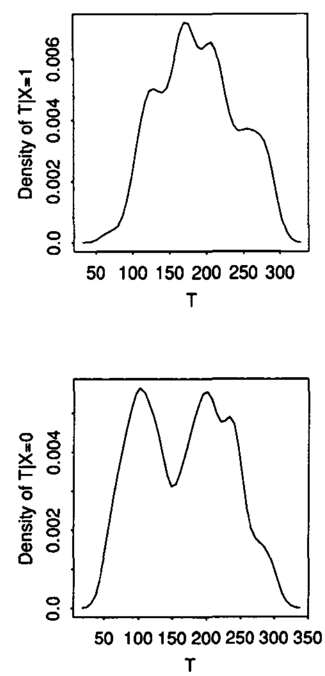

Figure 2. Nonparametric Estimate of the Conditional Density of T Given X = x.

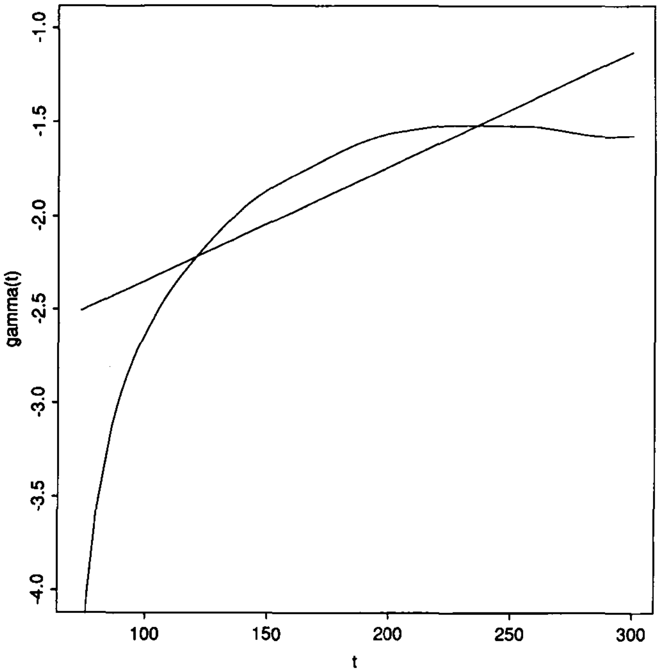

Figure 3. Estimates of Gamma for the Torrey Yucca Data.

termined as the one that minimizes the cross-validation sum of squares for the nonparametric regression of these residuals on $T_1, \ldots, T_n$. This procedure led to the choice $b = .5$. The starting value $\tilde{\beta}$ was determined by regressing $C_i - \hat{C}_i$, $i = 1, \ldots, n$ on $X_i - \hat{X}_i$, $i = 1, \ldots, n$; here $\hat{C}_i$ and $\hat{X}_i$ refer to the fitted value from the nonparametric regression of $C_1, \ldots, C_n$ and $X_1, \ldots, X_n$ on $T_1, \ldots, T_n$. The starting value was determined to be 1.234.

The iterative procedures described in Section 6 [Equations (12)–(15)] were used to compute $\hat{\gamma}_\beta$ and $\hat{\beta}$; both procedures were continued until a change in the estimate of less than 10^{-5} was achieved. The final estimate of β is $\hat{\beta} = 1.467$ with a standard error of .377, computed according to the asymptotic theory of Section 4; the nonparametric standard error is .366. The final estimate of σ^2 is 1.05. The estimates of β and σ^2 are relatively insensitive to the choice of b. For example, if $b = .4$ is used, then the estimate of β is 1.468 with a standard error of .396; the estimate of σ^2 is 1.13. A plot of $\hat{\gamma}_{\hat{\beta}}$ is given in Figure 3, together with the linear estimate of γ. Once the bandwidth was chosen and the starting value $\tilde{\beta}$ was computed, the CPU time used to compute $\hat{\beta}$ and $\hat{\gamma}_{\hat{\beta}}$ for this data set was .4 seconds, and the real time was 2 minutes.

A test of the null hypothesis that $\beta = 0$ may be conducted using the asymptotic normality of $\hat{\beta}$; clearly, the null hypothesis is rejected. Another approach to testing this hypothesis is to use the scaled deviance function (McCullagh and Nelder 1989, chap. 9), calculated under the full model and under the reduced model with $\beta = 0$. Asymptotic theory for the parametric case suggests that if the reduced model holds, then the difference of these deviances will be approximately distributed as a χ_1^2 random variable. For these data this difference is 16.92, which agrees closely with the result obtained by comparing $\hat{\beta}$ to its standard error.

If instead a fully parametric logistic model linear in X and T is used, then we obtain $\hat{\beta} = 1.59$ with a standard error of .354 using the SAS procedure PROC CATMOD on the full data set. The linear estimate of γ obtained from this procedure is $.00778T - 3.3135$.

7.2 Simulated Data

A small simulation study was conducted to study the finite sample properties of the $\hat{\beta}$ in the semiparametric model used in Section 7.1. The binomial random number generator RANBIN in SAS was used to generate 100 data sets consisting of independent pseudorandom indicator random variables according to the model $P(Y = 1 \mid X, T) = \beta_0 X + \gamma_0(T)$; the values of X and T used were those observed in the Torrey yucca data. For all of the generated data sets, $\sigma^2 = 1$ and $\beta_0 = 1.56$ were used. Two choices of γ_0 were used: a linear function $\gamma_0(t) = -3.32 + .008t$ and a nonlinear function $\gamma_0(t) = -3.32 + .008t + f(t)$, where $f(t) = 2.5[1 + (t - 175)^2/2{,}500]^{-1}$. The linear function is approximately the linear function of T fit to the original Torrey yucca data with PROC CATMOD. For each choice of γ_0, β was estimated under both the semiparametric model used in Section 7.1 and a fully parametric logistic model linear in X and T.

The results of this study are presented in Table 1. Along with the average values of $\hat{\beta}$ and $\hat{\sigma}^2$ over the 100 generated data sets, a Monte Carlo estimate of the standard error of $\hat{\beta}$ is given, as well as the average theoretical standard error of $\hat{\beta}$ calculated using both the model-based and nonparametric methods described in Section 5; the numbers in parentheses are the standard deviations over the 100 sets of data. For the semiparametric estimates for the linear function for gamma, the results are based on only 95 replications, because 5 cases had overflow/underflow errors. These results, together with normal probability plots of $\hat{\beta}$, indicate that the asymptotic distribution theory given in Proposition 1 pro-

Table 1. Results of the Monte Carlo Study

	Linear γ_0		*Nonlinear γ_0*	
	Semiparametric	*Parametric*	*Semiparametric*	*Parametric*
$\hat{\beta}$	1.570	1.571	1.602	1.622
$\hat{\sigma}^2$	1.01 (.07)		.95 (.03)	
Monte Carlo SE of $\hat{\beta}$	.358	.351	.361	.332
Theoretical SE of $\hat{\beta}$				
Model-based	.390 (.03)		.330 (.01)	
Nonparametric	.389 (.03)		.330 (.01)	

vides a accurate approximation to actual sampling distribution of $\hat{\beta}$ for the sample size and covariates used in the Torrey yucca example. Figure 4 contains estimates of $\gamma(t)$ from a single Monte Carlo realization; the curved solid line is γ_0, the dashed line is a nonparametric estimate of γ using $b = .5$, the dotted line is a nonparametric estimate of γ using $b = .3$, and the straight solid line is a parametric estimate of γ. These results indicate that the nonparametric estimates of γ accurately capture the important features of γ_0.

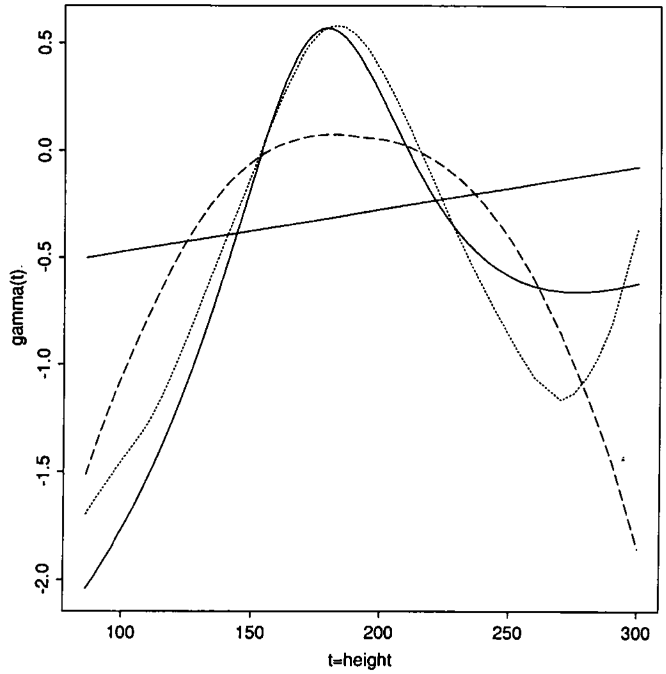

Figure 4. Monte Carlo Estimates of Gamma.

8. GENERALIZATION TO A MULTIVARIATE RESPONSE

In this section we extend the methodology developed in Sections 1–7 for a univariate response to the analysis of multivariate data using the class of models considered by Liang and Zeger (1986) and Zeger and Liang (1986). The extension to the case where the response $\mathbf{Y}$ is an $r \times 1$ random vector follows easily on observing that the n independent univariate observations considered in Sections 1–7 can be described as a single multivariate response ($r = n$) with a diagonal covariance matrix. The generalization involves relaxing the assumption of a diagonal covariance matrix and assuming that we have independent observations of the multivariate response. To simplify the presentation, trimming is not used here.

Let $\mathbf{Y}_u = (Y_{u1}, \ldots, Y_{ur})'$ denote the $r \times 1$ vector-valued response for the uth experimental unit, let $\mathbf{X}_u = (\mathbf{X}'_{u1}, \ldots, \mathbf{X}'_{ur})'$ denote the corresponding $r \times p$ matrix of covariate values to be included in the regression model parametrically, and let $\mathbf{T}_u = (\mathbf{T}'_{u1}, \ldots, \mathbf{T}'_{ur})'$ denote the corresponding $r \times q$ matrix of covariate values to be included in the regression model nonparametrically. It is assumed that $(\mathbf{T}_u, \mathbf{X}_u, \mathbf{Y}_u)$, $u = 1, \ldots, n$, are independent and identically distributed. Let $\mathbf{R}$ denote the common correlation matrix of $\mathbf{Y}_u | \mathbf{X}_u, \mathbf{T}_u$. Further, it is assumed that $(\mathbf{T}_{ui}, \mathbf{X}_{ui}, Y_{ui})$ for $i = 1, \ldots, r$, satisfy $E(Y_{ui}|\mathbf{X}_{ui}, \mathbf{T}_{ui}) = \mu(\mathbf{X}_{ui}, \mathbf{T}_{ui})$, $\mu(\mathbf{X}_{ui}, \mathbf{T}_{ui}) = h(\gamma(\mathbf{T}_{ui}) + \mathbf{X}_{ui}\beta)$, and $\mathrm{var}(Y_{ui}|\mathbf{X}_{ui}, \mathbf{T}_{ui}) = \sigma^2 V(\mu(\mathbf{X}_{ui}, \mathbf{T}_{ui}))$ as before.

Recall that for the case of a univariate response, the quasi-likelihood estimate of β solves Equation (7), $\hat{\mathbf{B}}(\beta) = \hat{\mathbf{D}}'\hat{\mathbf{V}}^{-1}\hat{\mathbf{S}} = 0$, where $\hat{\mathbf{S}} = \hat{\mathbf{S}}(\beta) = (Y_1 - \hat{\mu}_1, \ldots, Y_n - \hat{\mu}_n)'$, $\hat{\mu}_i = h(\mathbf{X}_i\beta + \hat{\gamma}_\beta(\mathbf{T}_i))$, $\hat{\mathbf{D}} = \hat{\mathbf{D}}(\beta) = \partial\hat{\mu}/\partial\beta$, and $\hat{\mathbf{V}} = \hat{\mathbf{V}}(\beta) = \mathrm{diag}\{V(\hat{\mu}_1), \ldots, V(\hat{\mu}_n)\}$. The generalization to multivariate Y uses as its estimate of β the solution to

$$\sum_{u=1}^{n} \hat{\mathbf{B}}_u(\beta; \mathbf{R}) = \sum_{u=1}^{n} \hat{\mathbf{D}}'_u\hat{\boldsymbol{\Sigma}}_u^{-1}\hat{\mathbf{S}}_u = 0. \tag{17}$$

The uth experimental unit contributes a term of the form $\hat{\mathbf{B}}_u(\beta; \mathbf{R}) = \hat{\mathbf{D}}'_u\hat{\boldsymbol{\Sigma}}_u^{-1}\hat{\mathbf{S}}_u$ to the gradient of the profile quasi-likelihood, where $\hat{\mathbf{D}}_u$, $\hat{\mathbf{S}}_u$, and $\hat{\mathbf{V}}_u$ are defined as in the univariate case although now applying only to the uth experimental unit and $\hat{\boldsymbol{\Sigma}}_u = \hat{\mathbf{V}}_u^{1/2}\mathbf{R}\hat{\mathbf{V}}_u^{1/2}$, which allows for correlation among the r measurements according to $\mathbf{R}$. The matrix $\mathbf{R}$ is a "working" correlation matrix and does not necessarily represent the true correlation matrix.

For each fixed $\mathbf{t}$ and β, let $\hat{\gamma}_\beta(\mathbf{t})$ denote the solution in $\lambda = \gamma(\mathbf{t})$ of

$$\sum_{u=1}^{n} \tilde{\mathbf{D}}'_u\tilde{\boldsymbol{\Sigma}}_u^{-1}\mathbf{K}_u\tilde{\mathbf{S}}_u = 0, \tag{18}$$

where $\tilde{\boldsymbol{\Sigma}}_u = \tilde{\mathbf{V}}_u^{1/2}\mathbf{R}\tilde{\mathbf{V}}_u^{1/2}$, $\tilde{\mathbf{D}}_u$ and $\tilde{\mathbf{V}}_u$ are defined as $\hat{\mathbf{D}}_u$ and $\hat{\mathbf{V}}_u$ except that they are evaluated at $\gamma(\mathbf{t})$ rather than at $\hat{\gamma}_\beta(\mathbf{t})$, and $\mathbf{K}_u$ is an $r \times r$ diagonal matrix with jth diagonal element $W((\mathbf{t} - \mathbf{T}_{uj})/b)$. The estimators can be calculated using a generalization of the algorithm in Section 6.

The following result follows from a minor modification of the proofs for Propositions 1 and 2, which hold for $r = 1$. The proof is omitted.

Proposition 3. Let β_0 and $\gamma_0 = \gamma_0(\mathbf{t})$ denote the true parameter values. Let $\hat{\beta}$ and $\hat{\gamma}_\beta$ denote the solutions to Equations (17) and (18). Under the regularity conditions for Propositions 1 and 2 given in the Appendix,

$$n^{1/2}(\hat{\beta} - \beta_0) \xrightarrow{\mathcal{D}} N_p(0, \boldsymbol{\Gamma}),$$

$$\|\hat{\gamma}_{\hat{\beta}} - \gamma_0\| = o_p(1) \quad \text{as} \quad n \to \infty,$$

where $\boldsymbol{\Gamma} = \mathrm{plim}_{n\to\infty} \boldsymbol{\Gamma}_n(\beta)|_{\beta=\hat{\beta}}$ and

$$\boldsymbol{\Gamma}_n(\beta) = \left\{\sum_{u=1}^{n} \hat{\mathbf{A}}_u(\beta; \mathbf{R})/n\right\}^{-1} \left\{\sum_{u=1}^{n} \hat{\mathbf{D}}'_u\hat{\boldsymbol{\Sigma}}_u^{-1}\mathrm{cov}(\hat{\mathbf{S}}_u)\hat{\boldsymbol{\Sigma}}_u^{-1}\hat{\mathbf{D}}_u/n\right\} \times \left\{\sum_{u=1}^{n} \hat{\mathbf{A}}_u(\beta; \mathbf{R})/n\right\}^{-1}.$$

Note that $\boldsymbol{\Gamma}$ may be consistently estimated by replacing $\mathrm{cov}(\hat{\mathbf{S}}_u)$ with $\hat{\mathbf{S}}_u\hat{\mathbf{S}}'_u$ in $\boldsymbol{\Gamma}_n(\hat{\beta})$, yielding

$$\hat{\boldsymbol{\Gamma}}_n = n\left\{\sum_{u=1}^{n} \hat{\mathbf{A}}_u(\hat{\beta}; \mathbf{R})\right\}^{-1} \left\{\sum_{u=1}^{n} \hat{\mathbf{B}}_u(\hat{\beta}; \mathbf{R})\hat{\mathbf{B}}_u(\hat{\beta}; \mathbf{R})'\right\} \times \left\{\sum_{u=1}^{n} \hat{\mathbf{A}}_u(\hat{\beta}; \mathbf{R})\right\}^{-1}.$$

Corollary 1. If the working correlation matrix $\mathbf{R} = \mathbf{R}_0$ and the variance function $V(\cdot) = V_0(\cdot)$ are correctly specified, then $\boldsymbol{\Gamma} = \operatorname{plim}_{n\to\infty} \sigma^2 n\{\sum_{u=1}^{n} \hat{\mathbf{A}}_u(\boldsymbol{\beta}; \mathbf{R}_0)\}^{-1}$.

The scale parameter σ^2 can be estimated consistently by $\hat{\sigma}^2 = \sigma^2(\hat{\boldsymbol{\beta}}, \hat{\gamma}_{\hat{\boldsymbol{\beta}}})$, where

$$\sigma^2(\boldsymbol{\beta}, \gamma) = \frac{1}{rn} \sum_{k=1}^{r} \sum_{u=1}^{n} (Y_{ur} - h_{ur})/V(h_{ur})$$

and $h_{ur} = h(\mathbf{X}_{ur}\boldsymbol{\beta} + \gamma(\mathbf{T}_{ur}))$. Note that estimation of σ^2 is not needed for estimation of $\boldsymbol{\beta}$ nor for computation of $\hat{\boldsymbol{\Gamma}}_n$.

To apply these methods, the correlation matrix $\mathbf{R}$ must be specified. Let $\mathbf{R}_0$ denote the true correlation matrix. An important property of these quasi-likelihood estimators is that $\mathbf{R} = \mathbf{R}_0$ is not needed to ensure consistency of the estimator of $\boldsymbol{\beta}$. Furthermore, Liang and Zeger (1986) showed that there is not a significant loss of efficiency if $\mathbf{R} = \mathbf{I}$ is used and the r responses are only slightly correlated. Corollary 2 provides a sufficient condition for no loss of efficiency in estimation of $\boldsymbol{\beta}$ when $\mathbf{R} \neq \mathbf{R}_0$.

Corollary 2. Suppose that $\mathbf{F} = \operatorname{plim}_{n\to\infty} \mathbf{V}^{-1/2}\hat{\mathbf{D}}_u$ is a constant matrix and the variance function $V(\cdot) = V_0(\cdot)$ is correctly specified.

a. Then

$$\boldsymbol{\Gamma} = \sigma^2(\mathbf{F}'\mathbf{R}^{-1}\mathbf{F})^{-1}(\mathbf{F}'\mathbf{R}^{-1}\mathbf{R}_0\mathbf{R}^{-1}\mathbf{F})(\mathbf{F}'\mathbf{R}^{-}\mathbf{F})^{-1},$$

and if the working correlation matrix $\mathbf{R} = \mathbf{R}_0$ is correctly specified, then

$$\boldsymbol{\Gamma} = \sigma^2(\mathbf{F}'\mathbf{R}_0^{-1}\mathbf{F})^{-1}.$$

b. Let the true correlation matrix $\mathbf{R}_0 = \mathbf{R}(\alpha_0)$, where α_0 is an unknown vector of parameters and $\mathbf{R}(\cdot)$ is a parametric specification of the correlation structure. Let the working correlation matrix $\mathbf{R} = \mathbf{R}(\alpha_1)$ for some arbitrarily chosen value α_1 in the parameter space.

If $\mathbf{R}(\alpha)\mathbf{F} = c(\alpha)\mathbf{F}$ for some constant $c(\alpha) \neq 0$ for all α in the parameter space, then $\mathbf{R}(\alpha)^{-1}\mathbf{F} = c(\alpha)^{-1}\mathbf{F}$ and there is no loss of efficiency in using $\mathbf{R} = \mathbf{R}(\alpha_1)$ as the working correlation matrix.

Corollary 2 states that if the columns of $\mathbf{F}$ are eigenvectors of $\mathbf{R}(\alpha)$ with corresponding eigenvalue $c(\alpha)$ and the variance function has been specified correctly, then there is no loss of efficiency from misspecification of α in the working correlation matrix. This observation has implications for our analysis of data in Section 9. Note that these results concerning efficiency of the estimators under misspecification of $\mathbf{R}$ also apply to parametric regression models.

9. MULTIVARIATE APPLICATIONS

The data analyzed in this section were collected during 1990 by Huanmin Lu, a Research Fellow in the Department of Mechanical and Industrial Engineering, for Dr. Andrew Swift, Project Director of the El Paso Solar Pond Project, University of Texas at El Paso. The brine in a salinity gradient solar pond is dripped across the surface of nets to evaporate the water and increase the brine's salt concentration. The

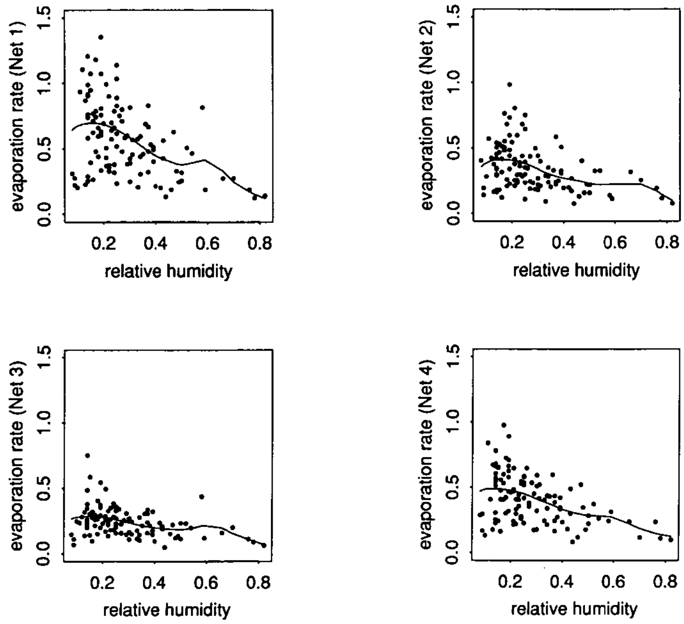

Figure 5. Kernel Estimates of Evaporation Rate Versus Relative Humidity for Each Net.

high-salt concentration brine is pumped back to the depths of the solar pond, allowing for recycling of the salt in the solar pond. One purpose of the experiment was to study the effect on evaporation rate of the four net materials. Over a period of several months, weather and brine salinity measurements were collected along with the evaporation rate for each net, for a total of 111 observations on each of the four nets. A plot of the data, along with a nonparametric kernel estimate, is given in Figure 5 for each of the four nets. These exploratory plots use a bandwidth of $b = .2$, which was arbitrarily chosen by trial and error; the quartic kernel is used along with the boundary modification of Rice (1984).

Let $\mathbf{Y}' = (Y_1, \ldots, Y_4)$ be the vector of evaporation rates (pounds per square foot/hour) for the four nets, and let T denote the relative humidity as a proportion. The relative humidity is measured once each time that an observation $\mathbf{Y}$ is recorded. Let $\mathbf{X} = (\mathbf{X}_1', \ldots, \mathbf{X}_4')'$, where $\mathbf{X}_i$ is a 3×1 row vector that takes on the values (1, 0, 0) for net 1, (0, 1, 0) for net 2, (0, 0, 1) for net 3, and (0, 0, 0) for net 4. Here $r = 4$, $p = 3$, $q = 1$, and $n = 111$.

Based on graphical exploration of the data, the evaporation rate was modeled as a function of relative humidity using $h(s) = e^s$ and $V(\mu) = \mu^2$. To simplify the computations, $\mathbf{R} = \mathbf{I}$ is used. For this choice of $\mathbf{R}$, the computation of $\hat{\gamma}_{\boldsymbol{\beta}}(t)$ is as in Example 2 of Section 3. Trimming of the observations with the three highest levels of relative humidity was needed to avoid taking the logarithm of a negative number; this is a consequence of the sparsity of the data for large values of T and the fact that the boundary corrected kernel has negative lobes. The final parameter estimates are $\hat{\boldsymbol{\beta}}' = (.3442, -.1760, -.5146)$ and $\hat{\sigma}^2 = .2307$; the bandwidth $b = .41$ was used for the nonparametric estimation.

A parametric quasi-likelihood fit to the trimmed data set using the GLM function in Splus with $\mathbf{R} = \mathbf{I}$ and a model linear in $\mathbf{X}$ and T yields the parameter estimates $\hat{\gamma}(t) = -.5250 - 1.3786t$ and $\hat{\beta}' = (.3494, -.1810, -.5114)$, whereas the fit to the entire data set yields $\hat{\gamma}(t) = -.4751 - 1.5837t$ and $\hat{\beta}' = (.3436, -.1795, -.5096)$. Figure 6 contains a plot of $\hat{\gamma}_{\beta}(t)$ for both the parametric and semiparametric models; there is evidence that $\gamma(t)$ is nonlinear.

The true correlation structure $\mathbf{R}_0$ is likely to be an exchangeable correlation structure given by $\mathbf{R}(\alpha_0)$, where $\mathbf{R}(\alpha)$ has (i, j)th element α for $i \neq j$, because many factors affecting the rate of evaporation will affect all four net materials. For this data set, $\text{plim}_{n\to\infty} \mathbf{V}^{-1/2}\hat{\mathbf{D}}_u$ is a 4×3 matrix $\mathbf{F} = \{\mathbf{F}_{ij}\}$, where $F_{ij} = \frac{3}{4}$ if $i = j$ and $-\frac{1}{4}$ otherwise. It is straightforward to verify that $\mathbf{R}(\alpha)\mathbf{F} = (1 - \alpha)\mathbf{F}$. Hence by Corollary 2b, if the data have an exchangeable correlation structure and the variance function has been specified correctly, then there is no loss of efficiency in using $\mathbf{R} = \mathbf{I}$ as the working correlation matrix. If the data have an exchangeable correlation structure and the variance function $V(\cdot)$ is correctly specified, then the asymptotic covariance matrix is $(1 - \alpha)\sigma^2(\mathbf{F}'\mathbf{F})^{-1}/n$ (Corollary 2a).

Assuming that $\mathbf{R} = \mathbf{R}_0 = I$ and that the variance function has been correctly specified, the estimated covariance matrix $\hat{\sigma}^2(\mathbf{F}'\mathbf{F})^{-1}/n$ (Corollary 2a) is a 3×3 matrix with diagonal elements .0043 and off-diagonal elements .0021. On the other hand, if we do not assume that $\mathbf{R} = \mathbf{R}_0 = \mathbf{I}$ or that the variance function has been correctly specified, then the asymptotic covariance matrix is estimated nonparametrically with $n^{-1}\hat{\Gamma}_n$ (Proposition 3), which is equal to

$$\begin{pmatrix} .00169 & .00035 & .00073 \\ .00035 & .00138 & .00046 \\ .00073 & .00046 & .00148 \end{pmatrix}.$$

Comparison of the two estimated covariance matrices reveals that they differ by about a factor of $\frac{1}{4}$. Hence there is support for our choice of $V(\mu) = \mu^2$ as the variance function and an exchangeable correlation structure with $\alpha = \frac{3}{4}$.

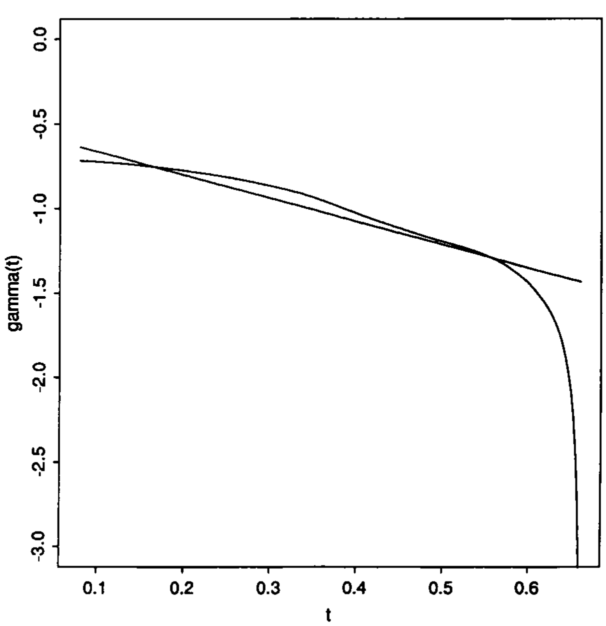

Figure 6. Estimates of Gamma for the Solar Pond Data.

An asymptotically valid test of H_0: $\beta = 0$ may be based on $X^2 = n\hat{\beta}'\hat{\Gamma}_n^{-1}\hat{\beta}$, which has an asymptotic chi-squared distribution with 3 degrees of freedom under the reduced model. The observed value of the X^2 test statistic for testing H_0: $\beta = 0$ using the nonparametric estimate of the covariance matrix is 449.37 and is statistically significant at the .05 level. Therefore, the fabric used in the net does appear to affect the rate of evaporation rate of the brine. A test of $\beta = 0$ using the deviance might be constructed along the lines suggested by Harris (1992).

APPENDIX: TECHNICAL DETAILS

The following regularity conditions are assumed to hold:

1. $\mathbf{T}$ takes values in a compact set $\mathcal{T} \subset \Re^q$, $\mathbf{X}$ takes values in a compact set $\mathcal{X} \subset \Re^p$, and σ takes values in a compact subset of $(0, \infty)$.
2. The parameter β takes values in a compact set $\mathcal{B} \subset \Re^p$. The parameter γ takes values in the set $\Gamma = \{g \in C^2(\mathcal{T}) : \|g\| \leq C\}$ for some sufficiently large constant C.
3. Let $\mathcal{M}$ denote a compact subset of $\Re$ such that $\gamma(\mathbf{t}) + \mathbf{X}\beta \in \mathcal{M}$ for all $\mathbf{t} \in \mathcal{T}$, $\mathbf{X} \in \mathcal{X}$, $\gamma \in \Gamma$, and $\beta \in \mathcal{B}$, and let $\mathcal{H} = h(\mathcal{M})$. Then $\sup_{\mu\in\mathcal{H}} V(\mu) < \infty$, $\inf_{\mu\in\mathcal{M}} V(\mu) > 0$, $\sup_{\mu\in\mathcal{H}} \Omega(\mu) < \infty$, and $\sup_{\mu\in\mathcal{H}} \int^{\mu} \Omega(s)\, ds < \infty$.
4. For $r = 1, 2, 3$, the derivative $\partial^r V(\mu)/\partial\mu^r$ exists for all $\mu \in \mathcal{H}$ and is bounded for $\mu \in \mathcal{H}$. For $r = 1, 2, 3$, the derivative $\partial^r h(m)/\partial m^r$ exists for all $m \in \mathcal{M}$ and is bounded for $m \in \mathcal{M}$.
5. $|\mathbf{\Sigma}_0| > 0$ and $|\mathbf{\Sigma}_1| > 0$.

In addition to Conditions 1 through 5, conditions are needed on the nonparametric estimate of γ. These conditions are given here as Condition 7 followed by a discussion of sufficient conditions for the kernel estimate of γ defined in Section 3 to satisfy these conditions.

7. For each $\mathbf{t} \in \mathcal{T}_0$ and each $\beta \in \mathcal{B}$, $\hat{\gamma}_\beta(\mathbf{t})$ converges in probability to some constant as $n \to \infty$; denote that constant by $\tilde{\gamma}_\beta(\mathbf{t})$.
   a. For each $\beta \in \mathcal{B}$, $\tilde{\gamma}_\beta \in \Gamma$ and for all $i, j = 0, 1, 2, i + j \leq 2$, $\partial^{i+j}\tilde{\gamma}_\beta(\mathbf{t})/\partial\mathbf{t}^i\partial\beta^j$, and $\partial^{i+j}\hat{\gamma}_\beta(\mathbf{t})/\partial\mathbf{t}^i\partial\beta^j$ exist. Here differentiation with respect to $\mathbf{t}$ refers to the array of partial derivatives with respect to $(t_1, \ldots, t_d)$ of the specified order.
   b. Let $\tilde{\gamma}_{k\beta}(\mathbf{t}) = \partial\tilde{\gamma}_\beta/\partial\beta_k$. For some $\alpha_1 \geq \alpha_2 > 0$, $\alpha_1 + \alpha_2 \geq \frac{1}{2}$.
   $$\|\hat{\gamma}_\beta - \tilde{\gamma}_\beta\| = o_p(n^{-\alpha_1}) \quad \text{and} \quad \max_k \|\hat{\gamma}'_{k\beta} - \tilde{\gamma}'_{k\beta}\|$$
   $$= o_p(n^{-\alpha_2}) \quad \text{as} \quad n \to \infty \quad \text{for} \quad \beta = \beta_0.$$
   c. $\sup_{\beta\in\mathcal{B}} \|\hat{\gamma}_\beta - \tilde{\gamma}_\beta\|$, $\sup_{\beta\in\mathcal{B}} \max_k \|\hat{\gamma}'_{k\beta} - \tilde{\gamma}_{k\beta}\|$ and $\sup_{\beta\in\mathcal{B}} \max_{i,j} \|(\partial^2/\partial\beta_i\partial\beta_j)(\hat{\gamma}_\beta - \tilde{\gamma}_\beta)\|$ are all of order $o_p(1)$ as $n \to \infty$.
   d. For some $\delta > 0$,
   $$\max_j \left\| \frac{\partial}{\partial t_j}(\hat{\gamma}_\beta - \tilde{\gamma}_\beta) \right\| = o_p(n^{-\delta}) \quad \text{as} \quad n \to \infty, \qquad \text{for} \quad \beta = \beta_0$$
   and
   $$\max_{j,k} \left\| \frac{\partial}{\partial t_j}(\hat{\gamma}'_{k\beta} - \tilde{\gamma}'_{k\beta}) \right\| = o_p(n^{-\delta}) \quad \text{as} \quad n \to \infty, \qquad \text{for} \quad \beta = \beta_0.$$
   e. The function $\tilde{\gamma}_\beta(\mathbf{t})$ satisfies $\tilde{\gamma}_\beta = \gamma_0$ and $\tilde{\gamma}'_{k\beta} = \gamma'_{k0}(\mathbf{t})$ when $\beta = \beta_0$, where $\gamma_{k0} = \partial\gamma_\beta(\mathbf{t})/\partial\beta_k|_{\beta=\beta_0}$ and γ_β is as defined in Section 3.

We now consider sufficient conditions under which the estimator $\hat{\gamma}_\beta$ defined in Section 3 satisfies Condition 7; this result is a straightforward generalization of lemma 5 of Severini and Wong (1992). It should be emphasized, however, that for specific models, Condition 7 often holds under weaker conditions. In particular, if a closed-form expression for $\hat{\gamma}_\beta$ exists, as in Examples 1 and 2, then Condition 7 can be shown to hold under considerably weaker conditions.

A. The kernel W is the direct product of a kernel $w: \Re \mapsto \Re$ of order $k > 3d/2$ with support $[-1, 1]$. For $r = 0, \ldots, k+2$, the derivative $w^{(r)}(u)$ exists and is bounded for $u \in [-1, 1]$.
B. For $r = 1, \ldots, 10 + k$, the derivative $\partial^r V(\mu)/\partial\mu^r$ exists for all $\mu \in \mathscr{H}$ and is bounded for $\mu \in \mathscr{H}$. For $r = 1, \ldots, 10 + k$, the derivative $\partial^r h(m)/\partial m^r$ exists for all $m \in \mathcal{M}$ and is bounded for $m \in \mathcal{M}$.
C. For some even integer q satisfying $q > (1 + p + d)(1 + 2d + 2k)/(2k - 3d)$, $E_0\{|Y|^q\} < \infty$.
D. Let f denote the marginal density of T, and let $f(\cdot|\mathbf{t}; \mu)$ denote the conditional density function of Y given $\mathbf{T} = \mathbf{t}$. Then $0 < \inf_{\mathbf{t}\in\mathcal{T}} f(\mathbf{t}) \le \sup_{\mathbf{t}\in\mathcal{T}} f(\mathbf{t}) < \infty$,

$$\sup_{\mathbf{t}\in\mathcal{T}} \left| \frac{\partial^{j_1+\cdots+j_d} f}{\partial t_1^{j_1}, \ldots, \partial t_d^{j_d}}(\mathbf{t}) \right| < \infty,$$

and

$$\sup_{\mathbf{t}\in\mathcal{T}} \sup_{y} \sup_{\mu\in\mathcal{M}} \left| \frac{\partial^{j_1+\cdots+j_d} f}{\partial t_1^{j_1}, \ldots, \partial t_d^{j_d}}(y|\mathbf{t}; \mu) \right| < \infty$$

for all nonnegative integers $j_1, \ldots, j_d$ satisfying $j_1 + \cdots j_d \le k + 2$. Here $\mathbf{t} = (t_1, \ldots, t_d)$ denotes a generic point in $\mathcal{T}$.
E. For each fixed β, $\mathbf{t}$ the solution to $M(\eta; \beta, \mathbf{t}) = 0$, $\gamma_\beta(\mathbf{t})$, is unique, and for any $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$\sup_\beta \sup_{\mathbf{t}\in\mathcal{T}_0} |\gamma_\beta(\mathbf{t}) - \gamma(\mathbf{t})| < \varepsilon$$

$$\text{whenever} \quad \sup_\beta \sup_{\mathbf{t}\in\mathcal{T}_0} |M(\gamma(\mathbf{t}); \beta, \mathbf{t})| < \delta.$$

F. The sequence of bandwidths b satisfies $b = O(n^{-\alpha})$, where

$$\frac{1}{4k} < \alpha < \frac{1}{4d}\frac{1-c}{1 + c(d+1)/d},$$

$c = (1 + p + d)/q$, k satisfies Condition A, and q satisfies Condition C.

Note that the "optimal" bandwidth satisfies Condition F and that if the moment-generating function of Y exists, then c may be taken to be 0.

Proof of Proposition 1

Under the regularity conditions in effect, Proposition 1 follows immediately from propositions 1 and 2 of Severini and Wong (1992), provided that the quasi-likelihood function used in estimating γ and β has the properties of log-likelihood function. For simplicity, we consider the case $p = 1$ and $q = 1$.

It follows from Condition 3 that there exists a function $\rho(\cdot)$ such that for each β,

$$\frac{1}{n}\sum I_j Q(\gamma(T_j) + X_j\beta; Y_j) \xrightarrow{p} \rho(\beta) \quad \text{as} \quad n \to \infty.$$

Let Y denote a random variable with mean μ_0, and let $Q^*(\mu) = E\{Q(\mu; Y)\}$; then Q^* has a unique maximum at $\mu = \mu_0$. It follows that $\rho(\beta)$ has a unique maximum at $\beta = \beta_0$. Consistency of $\hat{\beta}$ now follows as in proposition 1 of Severini and Wong (1992).

Proposition 2 of Severini and Wong (1992) holds, provided that $\gamma_\beta(t)$ is a "least-favorable curve." Let $H(s) = (h'(s))^2/V(s)$. Then, following Severini and Wong (1992, lemma 1), the "least-favorable direction" is given by

$$v^*(t) = -\frac{E_0\{H(\gamma_0(T) + X\beta)X \mid T = t\}}{E_0\{H(\gamma_0(T) + X_\beta)\}}.$$

Hence the result follows, provided that

$$\frac{d}{d\beta}\gamma_\beta(t)|_{\beta=\beta_0} = v^*(t) \quad \forall\, t. \tag{A.1}$$

Recall that γ_β satisfies

$$E_0\{\mathbf{G}(\gamma_\beta(T) + X\beta)(h(\gamma_0(T) + X\beta_0) - h(\gamma_\beta(T) + X\beta)) \mid T = t\} = 0 \quad \text{for each } t.$$

Differentiating this expression with respect to β, it follows that

$$E_0\{\mathbf{G}(\gamma_\beta(T) + X\beta)h'(\gamma_\beta(T) + X\beta)(\gamma'_\beta(T) + X) \mid T = t\} = 0 \quad \text{for each } t.$$

Solving for $\gamma'_\beta(t)$ and using the fact that $H(s) = \mathbf{G}(s)h'(s)$ yields (A.1). Proposition 1 now follows from proposition 2 of Severini and Wong (1992).

[*Received February 1992. Revised August 1993.*]

REFERENCES

Chu, C. K., and Marron, J. S. (1991), "Choosing a Kernel Regression Estimator" (with discussion), *Statistical Science,* 6, 404–436.

Fan, J. (1993), "Local Linear Regression Smoothers and Their Minimax Efficiency," *The Annals of Statistics,* 21, 196–216.

Green, P. J. (1987), "Penalized Likelihood for General Semi-Parametric Regression Models," *International Statistical Review,* 55, 245–259.

Green, P. J., and Yandell, B. S. (1985), "Semi-Parametric Generalized Linear Models," in *Proceedings 2nd International GLIM Conference,* Lecture Notes in Statistics, Vol. 32, New York: Springer-Verlag, pp. 245–259.

Härdle, W., and Stoker, T. M. (1989), "Investigating Smooth Multiple Regression by the Method of Average Derivatives," *Journal of the American Statistical Association,* 84, 986–995.

Harris, I. R. (1993), "Quasi-Likelihood for Mixed Models," Technical Report 93-01, University of Texas, Center for Statistical Sciences.

Hastie, T. J., and Tibshirani, R. J. (1990), *Generalized Additive Models,* London: Chapman and Hall.

Hunsberger, S. (1990), "Semiparametric Regression in Likelihood-Based Models," technical report, Medical College of Virginia, Virginia Commonwealth University, Dept. of Biostatistics.

Ichimura, H., and Lee, L. F. (1991), "Semiparametric Least Squares Estimation of Multiple Index Models: Single Equation Estimation," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics,* eds. W. A. Barnett et al., Cambridge, U.K.: Cambridge University Press, pp. 3–49.

Liang, K. Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika,* 73, 13–22.

McCullagh, P. (1983), "Quasi-Likelihood Functions," *The Annals of Statistics,* 11, 59–67.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman and Hall.

Moore, D. F. (1986), "Asymptotic Properties of Moment Estimators for Overdispersed Counts and Proportions," *Biometrika,* 73, 283–288.

Powell, J. L., Stock, J. H., and Stoker, T. M. (1989), "Semiparametric Estimation of Index Coefficients," *Econometrica,* 57, 1403–1430.

Rice, J. A. (1984), "Boundary Modification for Kernel Regression," *Communications in Statistics, Part A—Theory and Methods,* 13, 893–900.

Robinson, P. M. (1988), "Root n-Consistent Semiparametric Regression," *Econometrica,* 56, 931–954.

Severini, T. A., and Wong, W. H. (1992), "Generalized Profile Likelihood and Conditionally Parametric Models," *The Annals of Statistics,* 20, 1768–1802.

Speckman, P. E. (1988), "Regression Analysis for Partially Linear Models," *Journal of the Royal Statistical Society,* Ser. B, 50, 413–436.

Staniswalis, J. G. (1989), "On the Kernel Estimate of a Regression Function in Likelihood-Based Models," *Journal of the American Statistical Association,* 84, 276–283.

Zeger, S. L., and Liang, K. Y. (1986), "Longitudinal Data Analysis for Discrete and Continuous Outcomes," *Biometrics,* 42, 121–130.