The Annals of Statistics
1991, Vol. 19, No. 2, 797–816

LARGE SAMPLE THEORY OF ESTIMATION IN BIASED SAMPLING REGRESSION MODELS. I[1]

BY PETER J. BICKEL[2] AND J. RITOV

University of California, Berkeley and The Hebrew University of Jerusalem

Biased sampling regression models were introduced by Jewell, generalizing the truncated regression model studied by Bhattacharya, Chernoff and Yang. If the independent variable takes on only a finite number of values (as does the stratum variable), we show:

1. That if the slope of the underlying regression model is assumed known, then the nonparametric maximum likelihood estimates of the distribution of the independent and dependent variables (a) can be calculated from ordinary M estimates; (b) are asymptotically efficient.
2. How to construct M estimates of the slope which are always $\sqrt{n}$ consistent, asymptotically Gaussian and are efficient locally, for example, if the error distribution is Gaussian.

We support our asymptotics with a small simulation.

1. Introduction. One of the most commonly used models in statistics postulates that an observation (X, Y) drawn from an infinite population follows a linear regression model,

$$Y = \beta^T X + \varepsilon, \tag{1.0}$$

when $\varepsilon \sim G$ and $X_{p\times 1} \sim H$ are independent. Here $\beta_{p\times 1}$, G and H are assumed unknown with H not concentrated on a hyperplane. If we sample with replacement to obtain (X_k, Y_k), $k = 1, \dots, n$, classical estimates of these parameters are available whose large sample theory is well-understood. Recently, atteııtion has focused on models in which the sampling from the population is biased. The most important example is truncated regression discussed recently in Bhattacharya, Chernoff and Yang (1983), Woodroofe (1985) and Tsui, Jewell and Wu (1987). Here (X, Y) is observed only if Y is above (or below) a threshold y_0. Jewell (1985) and Jewell and Quesenberry (1986) following work of Coslett (1981), Manski and Lerman (1977) and Vardi (1985) propose a generalization of the truncated regression model giving a variety of interesting examples. These′ are the biased sampling regression

Received September 1987; revised May 1990.

[1] Research partially supported by ONR contract N00014-80-C-0163.

[2] This paper was completed while the author was visiting AT & T Bell Telephone Labs and the Courant Institute.

AMS 1980 *subject classifications*. Primary 62G05; secondary 62G20.

Key words and phrases. Regression, biased sampling, nonparametric maximum likelihood, M estimates.

models which we study. They are described as follows. We consider a population of (X, Y) pairs following the distribution (1.0).

A set of possibly overlapping nonexhaustive strata $S_1, \dots, S_S$ is defined. Stratum S_i is sampled with probability λ_i^*, $\Sigma_{i=1}^{S} \lambda_i^* = 1$. Then we sample biasedly from the stratum with probability proportional to $w(i, x, y)$ for item (x, y). In many applications $w(i, x, y) = I((x, y) \in S_i)$. If I denotes the stratum variable and G, H have densities g, h with respect to Lebesgue measure and μ, respectively, then (I, X, Y) has density with respect to counting measure × μ × Lebesgue measure

$$p(i, x, y) = \lambda_i^* w(i, x, y) \frac{g(y - \beta^T x) h(x)}{W_i(G, H)} \tag{1.1}$$

for $i = 1, \dots, S$, where $w \geq 0$,

$$W_i(G, H) = \int\int w(i, x, y) g(y - \beta^T x) h(x)\, dy\, d\mu(x)$$

and W_i are assumed finite and positive. The λ_i^* are assumed to all be positive. For truncated regression, $s = 1$, $S_1 = \{y\colon y \leq y_0\}$, $w(1, x, y) = 1(y \leq y_0)$.

We propose to study estimation of β, G, H in model (1.1) under the broad assumption that X has known finite support $\{x_1, \dots, x_K\}$. Additional conditions are specified in Section 2. In a second paper we intend to study situations in which X does not have finite support. We proceed as follows:

1. We derive and give the asymptotic theory for the nonparametric maximum likelihood estimates of G, H if β is known and show that these estimates are efficient.
2. We propose a class of estimates for β, G, H in the general model which are $\sqrt{n}$ consistent. We give the asymptotic theory of the procedures and show how members which are efficient at submodels can be selected.

The key observation needed for point 1 is the exponential family representation of (1.1) given in (2.1). Then, the asymptotics of the case β known follows as in Vardi (1982) by noting that the MLE of G can be written as a weighted sum of the empirical distributions of Y given I and X, where the weights are estimated through a solution of $(K - 1) \times (S - 1)$ M-equations in the unknowns $\lambda_i W_i$ and $h(x_i)$. The asymptotic distribution of the estimator now follows from the well-established theory of M-estimation. When β is also unknown, we add another equation to this set to take care of β. This equation can be obtained from any M-equation that may be used for the estimation of the slope without biased sampling. Together we get $(K - 1) \times (S - 1) + 1$ equations and the result follows again from the theory of M-estimation.

The paper is organized as follows. In Section 2 we define the procedures, state our main results and relate our work to that of other authors. Section 3 has the proof of the results and further discussion. A small simulation study is given in Section 4.

2. The main results.

2.1. Suppose (I_k, X_k, Y_k), $k = 1, \dots, n$ are i.i.d. with common density (1.1). Consider first the case β known ($\beta = 0$ without loss of generality). The nonparametric maximum likelihood estimates of H and G are easily seen to concentrate on $\{X_1, \dots, X_n\}$, $\{Y_1, \dots, Y_n\}$, respectively. Suppose that the λ_i are also free. As far as maximum likelihood estimation of G and H goes, this makes no difference since these estimates, if they exist, are the same as estimates based on the conditional likelihood given $I_1, \dots, I_n$. Identify the distinct values of the X_i with $\{1, \dots, n_x\}$ and those of the Y_i with $\{y_1, \dots, y_{n_y}\}$ and redefine w appropriately. We may reparametrize the model (1.1) with $\beta = 0$ as

$$P[I = i, X = j, Y = y_l] = g_l^* e^{\nu_i + \mu_j - b(y_l, \nu, \mu)} w(i, j, y_l), \tag{2.1}$$

where $g^* = (g_1^*, \dots, g_{n_y}^*)$, $\nu = (\nu_1, \dots, \nu_{n_s})$ with n_s the number of distinct observed strata and $\mu = (\mu_1, \dots, \mu_{n_x})$. Here g^* varies over the interior of the unit simplex, ν and μ vary freely and

$$b(y_l, \nu, \mu) = \log \sum_{i,j} e^{\nu_i + \mu_j} w(i, j, y_l).$$

To ensure identifiability, we can require $\sum_i e^{\nu_i} = 1$ as well as $\sum_j e^{\mu_j} = 1$. With this restriction, ν, μ and g^* are related to the original parameters by

$$\begin{aligned} g_l^* &= \left[\sum_{i,j} \frac{\lambda_i h_j}{W_i(G, H)} w(i, j, y_l) \right] dG(y_l), \\ \nu_i &= \log \lambda_i, \quad \text{where } \lambda_i = \frac{\lambda_i^*}{W_i(G, H)} \Big/ \sum_i \frac{\lambda_i^*}{W_i(G, H)}, \\ \mu_j &= \log h_j. \end{aligned} \tag{2.2}$$

This choice of reparametrization is dictated by the assumption we make later that X takes on only a finite number of values. It corresponds to the reparametrization (2), (3) given by Mallows (1985) in his discussion of Vardi's model. In this form, the likelihood is a concave function of $n_s + n_x$ parameters, in fact an exponential family, so the analysis is more transparent.

Maximum likelihood estimates of h and λ may not exist or if they exist may not be unique as was noted by Mallows (1985). Suppose, for instance, all $X_i = x_0$. Then, model (2.1) is just Vardi's (1985) biased sampling model for which these difficulties already appear. We begin with a necessary and sufficient condition for unicity and existence of maximum likelihood estimates in model (2.1).

If a, b, c are functions from $\{1, \dots, n_s\}$, $\{1, \dots, n_x\}$, $\{y_1, \dots, y_{n_y}\}$ to R, respectively, and

$$w(i, j, y_l)\{a(i) + b(j) + c(y_l)\} = 0 \tag{2.3}$$

for all i, j and l then a, b, c are constant.

This condition is a corollary of the following:

THEOREM 1. *The nonparametric maximum likelihood estimator* (*NPMLE*) *of P exists and is unique. The MLE of* (ν, μ) *for the model* (2.1) *exists and is unique up to a vector in the following subspace of* $R^{n_s+n_x}$,

$$A_n = \{(a, b) : \exists\, c \text{ such that } w(i, j, y_l)(a(i) + b(j) + c(y_l)) \equiv 0\}.$$

In particular, if the observed data are such that

$$A_n = \{(a, b) : a_1 = \cdots = a_{n_x}, b_1 = \cdots = b_{n_y}\},$$

then the nonparametric MLE of λ, h, and G exists and is unique.

If the existence and unicity condition holds, then by the exponential family structure, the MLE's $\hat{\lambda}_i$, $1 \le i \le n_s - 1$, $\hat{h}_j$, $1 \le j \le n_x - 1$, uniquely satisfy the $n_x + n_s - 2$ nonlinear equations,

$$\hat{h}_j^* = \hat{h}_j \sum_i \hat{\lambda}_i \sum_l \left\{ \hat{g}_l^* w(i, j, y_l) \left(\sum_{a,b} \hat{\lambda}_a \hat{h}_b w(a, b, y_l) \right)^{-1} \right\}, \tag{2.4}$$

$$\hat{\lambda}_i^* = \hat{\lambda}_i \sum_j \hat{h}_j \sum_l \left\{ \hat{g}_l^* w(i, j, y_l) \left(\sum_{a,b} \hat{\lambda}_a \hat{h}_b w(a, b, y_l) \right)^{-1} \right\}, \tag{2.5}$$

where

$$\begin{aligned} \hat{h}_j^* &= \#\{X_k = j\}/n, \\ \hat{\lambda}_i^* &= \#\{I_k = i\}/n, \qquad i = 1, \dots, n_s, \\ \hat{g}_l^* &= \#\{Y_k = y_l\}/n, \\ \hat{h}_{n_x} &= 1 - \sum_{j=1}^{n_x-1} \hat{h}_j, \\ \hat{\lambda}_{n_s} &= 1 - \sum_{i=1}^{n_s-1} \hat{\lambda}_i. \end{aligned}$$

Then, by (2.2), if $\hat{W}_i$ is the NPMLE of W_i,

$$\hat{g}_l = \hat{g}_l^* \left\{ \sum_j \hat{h}_j \sum_i \frac{\hat{\lambda}_i}{\hat{W}_i} w(i, j, y_l) \right\}^{-1},$$

and $\hat{W}_i$ is proportional to $\hat{\lambda}_i^* / \hat{\lambda}_i$. Hence

$$\hat{g}_l = \hat{g}_l^* \left\{ \sum_j \hat{h}_j \sum_i \frac{\hat{\lambda}_i^2}{\hat{\lambda}_i^*} w(i, j, y_l) \right\}^{-1} \Big/ \hat{c}, \tag{2.6}$$

where

$$\hat{c} = \sum_l \hat{g}_l^* \left\{ \sum_j \hat{h}_j \sum_i \frac{\hat{\lambda}_i^2}{\hat{\lambda}_i^*} w(i, j, y_l) \right\}^{-1}.$$

The NPMLE of G is given by

$$\hat{G}(y) = \sum \{\hat{g}_l : y_l \le y\}. \tag{2.7}$$

It is easy to check that $\hat{g}$, $\hat{h}$ (with $\hat{\lambda}$ defined before) are also the maximum likelihood estimates if the λ_i^* are assumed known.

The equations (2.4) and (2.5) are the likelihood equations for an exponentional family so that a variety of algorithms are available. Here is a simple and rapid (for n_s, n_x moderate) algorithm proposed by Wang, which is certain to converge if $\hat{h}, \hat{\lambda}$ exist and are unique.

0. Initialize: $\hat{g} = \hat{g}^*$, $\hat{h} = \hat{h}^*$, $\hat{\lambda} = \hat{\lambda}^*$.
1. Solve for $\hat{h}^{\mathrm{NEW}}$ in Vardi's algorithm for

$$p(i,j,\hat{g}) = \hat{\lambda}_i^* h_j \frac{\nu(i,j,\hat{g})}{\sum_t \nu(i,t,\hat{g})},$$

where

$$\nu(i,j,\hat{g}) = \sum_l w(i,j,y_l)\hat{g}_l.$$

2. Let for $i = 1,\dots,n_s$,

$$A(i,\hat{g}) = \sum_j \nu(i,j,\hat{g})\hat{h}_j.$$

Set

$$\hat{\lambda}^{\mathrm{NEW}} = \frac{\hat{\lambda}_i^*}{A(i,\hat{g})}\left(\sum_a \frac{\hat{\lambda}_a^*}{A(a,\hat{g})}\right)^{-1},$$

$$\hat{c}^{\mathrm{NEW}} = \sum_l \hat{g}_l^* \left[\sum_{i,j} \hat{h}_j^{\mathrm{NEW}} \hat{\lambda}_i^{\mathrm{NEW}} w(i,j,l)\right]^{-1}.$$

For $l = 1,\dots,n_y$,

$$\hat{g}_l^{\mathrm{NEW}} = \hat{g}_l^* \left[\sum_{i,j} \hat{h}_j^{\mathrm{NEW}} \hat{\lambda}_i^{\mathrm{NEW}} w(i,j,l)\right]^{-1} \Big/ \hat{c}^{\mathrm{NEW}}.$$

3. $\hat{g} = \hat{g}^{\mathrm{NEW}}$, $\hat{h} = \hat{h}^{\mathrm{NEW}}$, $\hat{\lambda} = \hat{\lambda}^{\mathrm{NEW}}$.
4. Return to 1 until convergence.

In special cases, simpler approaches work. If $S = 1$, $\hat{\lambda}_1 = 1$ and the algorithm reduces to Vardi's for model (2.7) as was noted by Jewell and Quesenberry (1986). In this case also the large sample theory we give later can be deduced from Gill, Vardi and Wellner (1988). If $S = 1$ and $w(x,y) = 1(y \le ax + b)$, we have the truncated regression model. In this case, $\hat{G}$ can be calculated explicitly even if X is continuous. The asymptotic theory of $\hat{G}$ and $\hat{H}$ is well-understood [see Woodroofe (1985)].

Let P_0 correspond to (λ_0, H_0, G_0). More generally, use the subscript 0 for quantities calculated under P_0.

A1. Suppose H_0 concentrates on a finite number of points $\{x_1, \dots, x_K\}$ and G_0, H_0, λ_0 are such that

$$P_0[a(I) + b(X) + c(Y) = 0] = 1$$

implies that a, b, c are all constant with probability 1.

A2. $\sum \lambda_{0i} h_{0j} w(i, x_j, y)$ is bounded and bounded away from zero.

Let $\hat{\lambda} = (\hat{\lambda}_1, \dots, \hat{\lambda}_{S-1})$, $\hat{h} = (\hat{h}_1, \dots, \hat{h}_{K-1})$ (where we identify $x_1, \dots, x_K$ with $1, \dots, K$ and $\hat{\lambda}_i, \hat{h}_j = 0$ if $\hat{\lambda}_i^*, \hat{h}_j^* = 0$). Let $\hat{\Lambda}^*, \hat{H}^*, \hat{G}^*$ be the empirical marginal distributions of I, X, Y, respectively. Define $S + K - 2$ dimensional vector functions ψ_1, ψ_2, ψ_3 on $\{1, \dots, S\}$, $\{1, \dots, x_K\}$ and R by

$$\psi_{1m}(i, \lambda, h) = \begin{cases} \delta_{im}, & 1 \le m \le S - 1, \\ 0, & S \le m \le S + K - 2, \end{cases}$$

$$\psi_{2m}(x_j, \lambda, h) = \begin{cases} 0, & 1 \le m \le S - 1, \\ \delta_{j(m-S+1)}, & S \le m \le S + K - 2, \end{cases}$$

$$\psi_{3m}(y, \lambda, h) = \begin{cases} -\lambda_m \sum_j h_j w(m, x_j, y) A^{-1}(y, \lambda, h), & 1 \le m \le S - 1, \\ -h_{m-S+1} \sum_i \lambda_i w(i, x_m - S + 1, y) A^{-1}(y, \lambda, h), & \\ & S \le m \le S + K - 2, \end{cases} \tag{2.8}$$

where

$$A(y, \lambda, h) = \sum_{i,j} \lambda_i h_j w(i, x_j, y)$$

and δ_{ij} is the Kronecker delta. Let

$$I(\lambda_0, h_0, G_0) = \mathrm{Var}_0(\psi_1(I, \lambda_0, h_0) + \psi_2(X, \lambda_0, h_0) + \psi_3(Y, \lambda_0, h_0))$$

be the $(S + K - 2) \times (S + K - 2)$ covariance matrix of the random vector, where (λ_0, h_0, G_0) are the population values. Without loss of generality, suppose $\lambda_{i0} > 0$, $h_{j0} > 0$ all i, j.

THEOREM 2. *If* A1 *holds, then*

(a) *With probability tending to* 1, *the existence and unicity condition holds and* $\hat{\lambda}, \hat{h}$ *are consistent. Further,*

$$\begin{pmatrix} \hat{\lambda} \\ \hat{h} \end{pmatrix} = \begin{pmatrix} \lambda \\ h \end{pmatrix} + I^{-1} \Big\{ \int \psi_1(i, \lambda_0, h_0)\, d\hat{\Lambda}^*(i) + \int \psi_2(x, \lambda_0, h_0)\, d\hat{H}^*(x) + \int \psi_3(y, \lambda_0, h_0)\, d\hat{G}^*(y) \Big\} + o_p(n^{-1/2}). \tag{2.9}$$

(b) *If* A2 *holds as well and* $\hat{G}$ *is given by* (2.7), *then*

$$\mathbf{L}\big(\sqrt{n}\,(\hat{G} - G_0)\big) \to \mathbf{L}(V), \tag{2.10}$$

where V *is a Gaussian process discussed in Section* 3.

(c) *The estimates* $\hat{\lambda}, \hat{h}, \hat{G}$ *are efficient.*

NOTES. (i) The covariance structure of V is determined by the expansions of Theorem 2(b) and (3.4). Its exact form does not appear to be particularly insightful.

(ii) The estimates $\hat{h}, \hat{\lambda}, \hat{G}$ are efficient in various senses. The simplest is that any finite linear combination of them (with $\hat{G}$ evaluated at a finite set of time points) has asymptotic variance which is no larger than that of any other locally regular asymptotically Gaussian estimate of the same linear combination. The definition of local regularity, which is needed to exclude superefficiency, and other aspects of efficiency are discussed in Bickel, Klaassen, Ritov and Wellner (1992), which we refer to as BKRW from now on.

2.2. If β is unknown, it is easy to see that, as in the case of ordinary regression, maximum likelihood applied to G, H, β leads to all values of β as being possible estimates. As in the ordinary case this is in part explained by the fact that the efficient influence function of β involves g'/g, whereas the maximum likelihood estimate of G for β fixed is discrete.

Following Jewell, we consider the simpler problem of constructing estimates which are locally regular Gaussian and hence $\sqrt{n}$ consistent for all G, H but are fully efficient only on a particular submodel, for example, G Gaussian.

More generally, let $\{x_1, \dots, x_K\}$ be the support of H and $\gamma(\cdot, \cdot)$ be a function from $\{x_1, \dots, x_K\} \times R$ to R^p on which we shall put conditions later. The function $\gamma^*(x, y; \beta, \Lambda, G, H)$ obtained by subtracting the expectation of γ from γ is given by

$$(x, y) \to \gamma(x, y - \beta^T x) - \frac{\int \gamma(x, u) \Sigma_{i,j} h_j \lambda_i w\big(i, x_j, u + \beta^T x_j\big)\, dG(u)}{\Sigma_{i,j} \lambda_i h_j \int w(i, x_j, u)\, dG(u)}$$

and satisfies

$$\int \gamma^*(x, y, \beta, G, H)\, dP_{(\beta, G, H)} = 0 \tag{2.11}$$

for all $P_{(\beta, \Lambda, G, H)}$ given by (1.1). So we expect the M estimate $\hat{\beta}$ solving

$$\sum_{i=1}^{n} \gamma^*\big(X_i, Y_i, \hat{\beta}, G, \Lambda, H\big) = 0,$$

where $\Lambda = (\lambda_1, \dots, \lambda_s)$, to be $\sqrt{n}$ consistent. Unfortunately, we do not know G, H, Λ. Following essentially Buckley and James (1979), we use $\gamma^*(\cdot, \beta, \hat{G}_\beta, \hat{\Lambda}_\beta, \hat{H}_\beta)$, where $\hat{G}_\beta$, $\hat{\Lambda}_\beta$, $\hat{H}_\beta$ are the MLE's of G, Λ, H for fixed β. (Buckley and James take out the conditional expectation of γ given X rather than the expectation but this is inessential.) For simplicity, we take β, x real. Extension to the general case is straightforward.

Writing $\hat{H}_\beta(x) = \Sigma\{\hat{h}_{j\beta}: x_j \le x\}$ we see from (2.4) and (2.5) that $\hat{G}_\beta$ and $\hat{H}_\beta$ are obtained by solving the following $S + K - 2$ equations for $\hat{h}_\beta, \hat{\lambda}_\beta$:

$$\hat{h}_j^* = \hat{h}_{j\beta} \sum_i \hat{\lambda}_{i\beta} \int w(i, x_j, u + \beta x_j) \times \Big(\sum_{a,b} \hat{\lambda}_{a\beta} \hat{h}_{b\beta} w(a, x_b, u + \beta x_b)\Big)^{-1} d\hat{G}_\beta^*(u), \tag{2.12}$$

$$\hat{\lambda}_i = \hat{\lambda}_{i\beta} \sum_j \hat{h}_{j\beta} \int w(i, x_j, u + \beta x_j) \times \Big(\sum_{a,b} \hat{\lambda}_{a\beta} \hat{h}_{b\beta} w(a, x_b, u + \beta x_b)\Big)^{-1} d\hat{G}_\beta^*(u), \tag{2.13}$$

where $\hat{G}_\beta^*$ is the empirical df of $Y_i - \beta X_i$, $i = 1, \dots, n$ and

$$d\hat{G}_\beta(u) = d\hat{G}_\beta^*(u) \Big(\sum_{i,j} \hat{\lambda}_{i\beta} \hat{h}_{j\beta} w(i, x_j, u + \beta x_j)\Big)^{-1} \hat{C}_\beta^{-1}, \tag{2.14}$$

where

$$\hat{C}_\beta = \int \Big(\sum_{i,j} \hat{\lambda}_{i\beta} \hat{h}_{j\beta} w(i, x_j, u + \beta x_j)\Big)^{-1} d\hat{G}_\beta^*(u). \tag{2.15}$$

Then (2.11) becomes

$$\int \gamma(x, y - \beta x)\, d\hat{P}^*(x, y) - \sum_{i,j} \hat{h}_{j\beta} \hat{\lambda}_{i\beta} \int \gamma(x_j, t) w(i, x_j, t + \beta x_j) \times \Big(\sum_{i,j} \hat{\lambda}_{i\beta} \hat{h}_{j\beta} w(i, x_j, t + \beta x_j)\Big)^{-1} d\hat{G}_\beta^*(t) = 0. \tag{2.16}$$

Now (2.12), (2.13) and (2.16) are M equations in the $S + K - 1$ unknowns $(h_{1\beta}, \dots, h_{(K-1)\beta}, \lambda_{1\beta}, \dots, \lambda_{(S-1)\beta}, \beta)$, given by

$$W(\lambda, h, \beta, \hat{P}) = 0, \tag{2.17}$$

where $\hat{P}$ is the empirical distribution of the data. The mth coordinate of $W(\lambda, h, \beta, P)$ is, for a given P on $\{1, \dots, S\} \times \{x_1, \dots, x_K\} \times R$, given by

$$W_m(\lambda, h, \beta, P) = \int \tilde{\psi}_m(i, x, y, \lambda, h, \beta)\, dP,$$

$1 \le m \le S + K - 2$, where

$$\tilde{\psi}_m(i, x, y, \lambda, h, \beta) = \tilde{\psi}_{1m}(i, \lambda, h) + \tilde{\psi}_{2m}(x, \lambda, h) + \tilde{\psi}_{3m}(x, y - \beta x, \lambda, h, \beta).$$

The functions $\tilde{\psi}_1, \tilde{\psi}_2$ are obtained by substituting $w(i, x, y + \beta x)$ for $w(i, x, y)$ in (2.8) and

$$\begin{aligned}\tilde{\psi}_{3m}&(u, \lambda, h, \beta)\\ &= -\lambda_m \sum_j h_j w(m, x_j, u + \beta x_j) A^{-1}(u, \lambda, h, \beta), \qquad \text{for } 1 \le m \le S - 1,\\ &= -h_{m-S+1} \sum_i \lambda_i w(i, x_{m-S+1}, u + \beta x_{m-S+1}) A^{-1}(u, \lambda, h, \beta),\\ &\qquad\qquad\qquad \text{for } S \le m \le S + K - 2,\end{aligned}$$

where

$$A(u, \lambda, h, \beta) = \sum_{a,b} \lambda_a h_b w(a, x_b, u + \beta x_b).$$

Further

$$\begin{aligned}W_{S+K-1}&(\lambda, h, \beta, P)\\ &= \int\Bigg[\gamma(x, y - \beta x) - \Big(\sum_{i,j} \lambda_i h_j \gamma(x_j, y - \beta x) w\big(i, x_j, y - \beta(x - x_j)\big)\Big)\\ &\qquad \times A^{-1}(y - \beta x, \lambda, h, \beta)\Bigg] dP.\end{aligned}$$

Let

$$\begin{aligned}&\tilde{\psi}_{S+K-1}(x, y, \lambda, h, \beta)\\ (2.18)\qquad &= \gamma(x, y - \beta x)\\ &\quad - \sum_{i,j} \gamma(x_j, y - \beta x) w\big(i, x_j, y - \beta(x - x_j)\big) A^{-1}(y - \beta x, \lambda, h, \beta).\end{aligned}$$

Let P_0 correspond to $(\lambda_0, h_0, G_0, \beta_0)$. Here are some further conditions.

A3. G_0 has an absolutely continuous density g_0 with finite Fisher information $\int((g_0')^2/g_0)(x)\,dx$ and $\gamma(X, Y - \beta X) \in L_2(P_0)$. Let

$$\tilde{\gamma}(I, X, Y) = \gamma(X, Y - \beta_0 X) - a_0(I) - b_0(X) - c_0(Y - \beta_0 X),$$

where $a_0(I) + b_0(X) + c_0(Y - \beta_0 X)$ is the orthogonal projection of $\gamma(X, Y - \beta_0 X)$ in $L_2(P_0)$ on the space of all variables $a(I) + b(X) + c(Y - \beta_0 X)$. Then, assume

$$(2.19)\qquad E_0\left(\tilde{\gamma}(I, X, Y) X \frac{g_0'}{g_0}(Y - \beta_0 X)\right) \ne 0.$$

A4. Let

$$V_n(\tau) = \sqrt{n}\,\big(W(\tau, \hat{P}) - W(\tau, P)\big),$$

where $\tau = (\lambda, h, \beta)$. Then

(a) $$\sup\left\{\frac{|V_n(\tau) - V_n(\tau_0)|}{1 + \sqrt{n}\,|\tau - \tau_0|} : |\tau - \tau_0| \le \varepsilon_n\right\} = o_p(1) \quad \text{for any } \varepsilon_n \downarrow 0.$$

(b) $W(\cdot\,, P_0)$ is differentiable and the matrix $\left[\dfrac{\partial W_i}{\partial \tau_j}(\tau_0, P_0)\right]$ is nonsingular.

We have the following result. For an example, see Section 4.

THEOREM 3. *Suppose* A1 *and* A2 *hold with Y replaced by $Y - \beta_0 X$ and $w(i, x, u)$ by $w(i, x, u + \beta_0 x)$, along with* A3 *and* A4. *Suppose a consistent solution $\hat{\tau} = (\hat{\lambda}, \hat{h}, \hat{\beta})$ of* (2.17) *exists. Write τ for (λ, h, β). Then*

(a) *The matrix*

$$M \equiv E_0\big[\tilde{\psi}(I, X, Y, \tau_0)\nabla l(I, X, Y, \tau_0, G_0)\big],$$

where ∇l is the gradient of the log-likelihood with respect to

$$(\lambda_1, \dots, \lambda_{S-1}, h_1, \dots, h_{K-1}, \beta),$$

is nonsingular and

$$\hat{\tau} = \tau_0 + n^{-1}M^{-1}\sum_{i=1}^{n}\tilde{\psi}(I_i, X_i, Y_i, \lambda_0, h_0, \beta_0) + o_p(n^{-1/2}).$$

(b) *If $\gamma(x, y) = -x(g_0'/g_0)(y)$, then the estimates $\hat{\lambda}, \hat{h}, \hat{\beta}$ are efficient at P_0.*

NOTE (a) CALCULATION OF $\hat{\tau}$. The following algorithm B should work. We have no assurance of its convergence

```
OUTER LOOP:
Nonlinear equation solution algorithm for $W_{S+K-1}(\hat{\lambda}_\beta, \hat{h}_\beta, \beta, \hat{P}) = 0$.
INNER LOOP:
(1) Get $\hat{\lambda}_\beta, \hat{h}_\beta$ from algorithm A with
        $w = w(i, x, y + \beta x)$.
(2) Compute $W_{S+K-1}(\hat{\lambda}_\beta, \hat{h}_\beta, \beta, \hat{P}^*)$ from (2.18).
END INNER LOOP
END
```

NOTE (b) CHECKING A4. Condition A4 is readily satisfied if in addition to A1–A2, we suppose $(\partial\gamma/\partial u)(x, u)$ and $(\partial w/\partial y)(i, x, y)$ exist and are bounded. Unfortunately, w is typically discontinuous. To assure the crucial

$$\sup\left\{\frac{\sqrt{n}\,|W(\tau, \hat{P}^*) - W(\tau_0, P_0)|}{1 + n^{1/2}|\tau - \tau_0|} : |\tau - \tau_0| \le \varepsilon_n\right\} = o_{P_0}(1),$$

we can apply arguments such as those of Pollard (1985), Section 5. For instance, as we shall show in the next section, the following conditions which cover truncated regression imply A3.

C1. For some $\alpha > 0$ and all $j = 1, \dots, n_x$,

$$E_0\{\gamma(x_j, Y - \beta x_j) - \gamma(x_j, Y - \beta' x_j)\}^2 = O(|\beta - \beta'|^{\alpha})$$

uniformly in β is some neighborhood of β_0.

C2. A2 holds.

C3. $w(i, x_j, y) = \sum_{k=1}^{K} W_{ij}(k) 1_{\{y \in I_{ij}(k)\}}$, where $\{I_{ij}(k)\}$ are intervals.

NOTE (c) ESTIMATION OF G. The natural estimate here is $\hat{G}_\beta$. A1, A2 and either boundedness of $(\partial w/\partial u)(i, x, u)$ or C1, C3 imply that

$$\mathbf{L}\left(\sqrt{n}\left(\hat{G}_\beta - G_0\right)\right) \to \mathbf{L}(W)$$

in the usual sense, where W is a Gaussian process. We give the proof of this claim and a discussion of the structure of W in the next section.

NOTE (d) ESTIMATION OF THE ASYMPTOTIC VARIANCE MATRIX OF $\hat{\tau}$. The variance matrix of $\sqrt{n}(\hat{\tau} - \tau_0)$ can be estimated easily if $\tilde{\psi}$ is smooth enough as a function of τ so that

$$M = -E(H(I, X, Y, \tau)),$$

where H is the derivative matrix of $\tilde{\psi}$. The usual estimate

$$\hat{M}^{-1} n^{-1} \sum_{i=1}^{n} \hat{\psi}\hat{\psi}^T(I_i, X_i, Y_i)(\hat{M}^{-1})^T, \tag{2.20}$$

where

$$\hat{M} = n^{-1} \sum_{i=1}^{n} H(I_i, X_i, Y_i, \hat{\tau})$$

and $\hat{\psi} \equiv \tilde{\psi}(\cdot, \cdot, \cdot, \hat{\lambda}, \hat{h}, \hat{\beta})$, will work under the usual conditions permitting approximation of $\hat{M}$ by $n^{-1}\sum_{i=1}^{n} H(I_i, X_i, Y_i, \lambda_0, h_0, \beta_0)$ and the corresponding approximation for the inner term of (2.20).

Unfortunately, $\tilde{\psi}$ involves $w(i, x, y + \beta x)$ which is often discontinuous, for example, in truncated regression. In that case, we believe careful argument will show that under mild conditions one can estimate M by

$$\hat{M} = n^{-1} \sum_{i=1}^{n} \hat{H}(I_i, X_i, Y_i, \hat{\tau}),$$

where $\hat{H}$ is obtained by taking finite differences of $\tilde{\psi}$ at step lengths of order appropriate to the assumed smoothness of g. (Derivatives can be taken for all parameters other than β.) We have not examined these questions in any detail. As usual an alternative would be to use the bootstrap.

NOTE (e) ONE STEP ESTIMATION. In the absence of results on existence and consistency of $\hat{\tau}$, if the conditions of Theorem 2 hold and a $\sqrt{n}$ consistent estimate $\tilde{\tau}$ of τ is available and $\hat{M}$ evaluated at $\tilde{\tau}$ is consistent, we can construct $\hat{\tau}$ satisfying Theorem 2(b) noniteratively by taking one Newton–Raphson step from $\tilde{\tau}$. Such $\tilde{\tau}$ are readily available if, say, we have a stratum with $w = 1$ or in special cases such as truncated regression where the structure of w is relatively simple.

NOTE (f) EFFICIENCY. If $\gamma = -xg_0'/g_0(y)$, it follows easily that $\hat{\tau}$ is efficient at all points of the submodel $\{P_{(\lambda, h, \beta, G_0)}\}$ but not at all points of $\mathbf{P}$. It is in principle possible to obtain estimates efficient at every G_0 by estimating g_0'/g_0. We do not pursue this here.

NOTE (g) ESTIMATION OF β WHEN G, H ARE UNIDENTIFIABLE. Suppose we wish to estimate β and (λ, h) are considered as nuisance parameters. Then A1 may be too strong. It is actually not needed. Note that $W_m(\lambda, h, \beta, \hat{P}) = 0$, $m = 1, \dots, S + K - 2$ define merely the MLE $\hat{\lambda}_\beta, \hat{h}_\beta$. If A1 is not satisfied, these equations are linearly dependent and can be reduced to $q < s + K - 2$ equations for q parameters defined as the coefficients of a basis for (say) the orthocomplement of $\hat{A}_n$ given in Theorem 1. Note that $\hat{\lambda}_\beta$ and $\hat{h}_\beta$ appear in the equation defining β, $W_{S+K-1}(\hat{\lambda}_\beta, \hat{h}_\beta, \beta, \hat{P}) = 0$. However, it follows from the discussion preceding (2.11) that $W_{S+K-1}(\hat{\lambda}_\beta, \hat{h}_\beta, \beta, \hat{P})$ can be rewritten as

$$\begin{aligned}&\int \tilde{\psi}_{S+K-1}(x, y, \lambda_\beta, h_\beta, \beta)\, d\hat{P} \\ &\quad = \int \gamma(x, y - \beta x)\, d\hat{P} - \int \gamma(x, y - \beta x)\, d\hat{P}_\beta,\end{aligned} \tag{2.21}$$

where $\hat{P}_\beta$ (with some abuse of notation) is the MLE of the joint distribution of (I, X, Y) assuming that β is the true slope. It was proved in Theorem 1 that $\hat{P}_\beta$ always exists and hence the estimating equation $W_{S+K-1}(\hat{\lambda}, \hat{h}_\beta, \beta, \hat{P}) = 0$ is well-defined. Smoothness conditions on this function of β or C1–C3 and A2 will guarantee asymptotic normality of a consistent root of this equation, even if A1 does not hold.

It is true, however, that the inner loop of the algorithm suggested in (a) may fail to converge (in terms of λ and h) if A1 does not hold. Yet one can use this algorithm, by stopping it when the expression in the RHS of (2.21) converges.

3. Proofs and additional discussion.

DEFINITION. Let G_S^W be the graph with vertices $\{1, \dots, n_s\}$ and edges $i \leftrightarrow_w i'$ iff $\int \{\Sigma_j w(i, j, y)\hat{h}_j^* \Sigma_j w(i', j, y)\hat{h}_j^*\}\, d\hat{G}^*(y) > 0$. Define similarly a graph G_X^W with vertices $\{1, \dots, n_x\}$.

NECESSARY CONDITION FOR EXISTENCE AND UNIQUENESS. The graphs G_S^W and G_x^W are connected.

To prove this, suppose $\hat{\lambda}, \hat{h}$ exist and are unique. Note that the graphs do not depend on the actual values of $\hat{h}_j^*$ (and $\hat{\lambda}_i^*$, respectively) but only on the pattern of positive and zero values. In particular, replace $\hat{h}_j^*$ by $\hat{h}_j$, $\hat{\lambda}_i^*$ by $\hat{\lambda}_i$, respectively.

The connectedness of the graphs is [see Vardi (1985)] now implied by existence and uniqueness of solutions of the likelihood equations in the models in which h and λ, respectively, are assumed known and equal to $\hat{h}, \hat{\lambda}$, respectively. But $\hat{\lambda}, \hat{h}$ (respectively) which are assumed to exist are precisely such solutions. The condition is not sufficient. For instance, suppose $w(1,1,1) = w(2,2,1) = 1$ and $w = 0$ otherwise and that 2 points are observed. Then the graphs are connected but the necessary and sufficient condition for existence and unicity (2.3) fails. [Take $a(1) = -b(1)$, $a(2) = -b(2)$ and $c(1) = 0$].

DEFINITION. Let G_s^S be the graph with vertices $\{1, \dots, n_s\}$ and edges $i \leftrightarrow_s i'$ iff $\int \Sigma_j \hat{h}_j^* w(i,j,l) w(i',j,t)\, d\hat{G}^*(t) > 0$. Define similarly the graph G_x^S with vertices $\{1, \dots, n_x\}$.

SUFFICIENT CONDITION FOR EXISTENCE AND UNICITY. The graphs G_s^S and G_x^S are connected.

To prove this, suppose that the graphs are connected and without loss of generality that (2.3) holds for a, b, c such that $c(Y) = 0$ and a takes on two or more values. Let $V = \{i\colon a(i) = a_0\}$, $V^c = \{i\colon a(i) \neq a_0\}$. Then V and V^c are connected in G_s^S. Then there exists $i \in V$, $i' \in V^c jt$ such that $w(i,j,t)$ and $w(i',j,t)$ are both positive. Hence $a(i) = a(i')$ a contradiction and sufficiency follows. The condition is not necessary. Take $w(1,1,1) = w(2,2,1) = w(1,2,2) = w(2,1,2) = 1$, $w = 0$ otherwise. Then the graphs are not strongly connected yet (2.3) is satisfied since $a(1) + b(1) = a(2) + b(2)$, $a(1) + b(2) = a(2) + b(1)$ implies $b(1) = b(2)$ and $a(1) = a(2)$.

If the sufficient condition is not satisfied and the necessary condition is satisfied, one should proceed to check the necessary and sufficient condition for the existence of a unique MLE. In most cases of interest, this should not be hard. In particular, if $n_s \times n_x$ is not large, checking the condition is equivalent to investigating the solution sub space of at most $n_x^2 n_x^2$ equations in $n_s + n_x$ unknowns [of the form $a(i) + b(j) = a(i') + b(j')$ for any i, j, i', j' such that $\Sigma_l w(i,j,l) w(i,j,l') > 0$].

PROOF OF THEOREM 1. Since the model is an exponential family, the proof is quite standard [cf. Brown (1987), Theorem 5.5, page 148]. Suppose that, as happens with probability tending to 1, $n_x = K$, $n_s = S$. Fix any (μ, ν) and

(a, b) and consider

$$f(\alpha) \equiv L_n(\nu - \alpha a, \mu - \alpha b) - L_n(\nu, \mu)$$
$$= \frac{1}{n}\sum_{k=1}^{n}\left[-\alpha\{a(I_k) + b(X_k)\} - \log\frac{\sum_{ij} e^{\mu_i + \nu_j - \alpha\{a(i)+b(j)\}} w(i, x_j, Y_k)}{\sum_{ij} e^{\mu_i + \nu_j} w(i, x_j, Y_k)}\right],$$

where $L_n(\nu, \mu)$ is n^{-1} times the log-likelihood of the sample at (ν, μ). Then

$$\left.\frac{\partial^2 f}{\partial \alpha^2}\right|_{\alpha=0} = -\sum_{k=1}^{n}\left[\frac{\sum_{ij}\{a(i) + b(j)\}^2 e^{\mu_i+\nu_j} w(i, x_j, Y_k)}{\sum_{ij} e^{\mu_i+\nu_j} w(i, x_j, Y_k)} - \left(\frac{\sum_{ij}\{a(i) + b(j)\} e^{\mu_i+\nu_j} w(i, x_j, Y_k)}{\sum_{ij} e^{\mu_i+\nu_j} w(i, x_j, Y_k)}\right)^2\right] \leq 0,$$

since the kth summand is the conditional variance of $a(I_k) + b(X_k)$, given Y_k. We obtain equality if and only if each term is zero or $(a(i) + b(j) + c(l))w(i, j, Y_l) \equiv 0$ for some vector c. Let A_n^c be such that $A_n \cap A_n^c = \{0\}$ and $A_n \oplus A_n^c = R^{n_s + n_x}$. We see that $L_n(\nu - \alpha a, \mu - \alpha b)$ is strictly concave in α for $(a, b) \in A_n^c$.

$$\begin{aligned} &L_n(\nu - \alpha b, \mu - \alpha a) - L_n(\nu, \mu) \\ &\quad = \frac{1}{n}\sum_{k=1}^{n}\Big[-\alpha(a(I_k) + b(X_k)) \\ &\qquad + \min_{i,j}\{a(i) + b(j) \colon w(i, j, Y_k) > 0\} + O(1)\Big], \end{aligned} \tag{3.1}$$

where $O(1)$ does not depend on the data. Since $(a, b) \notin A_n$, we must have

$$\lim_{\alpha \to \infty} L_n(\nu - \alpha b, \mu - \alpha a) = -\infty.$$

We conclude that $L_n(\mu, \nu)$ is strictly concave on A_n^c and approaches $-\infty$ as (μ, ν) approaches the boundary of A_n^c. Hence it has a unique maximum and the maximizing value corresponds to an MLE of P. To establish uniqueness and the second part of the theorem, we have to show that $P_{(\mu,\nu)} = P_{(\mu - a, \nu - b)}$, if $w(i, j, l)(a(i) + b(j) + c(l)) \equiv 0$. But the density of $P_{(\mu - a, \nu - b)}$ given by (2.1) is

$$\begin{aligned} g_l^* \frac{e^{\mu_i - a_i + \nu_j - b_j} w(i, j, Y_l)}{\sum_{k,m} e^{\mu_k - a_k + \nu_m - b_m} w(k, m, Y_l)} &= g_l^* \frac{e^{\mu_i + \nu_j + c(l)} w(i, j, Y_l)}{\sum_{k,m} e^{\mu_k + \nu_m + c(l)} w(i, j, Y_l)} \\ &= g_l^* \frac{e^{\mu_i + \nu_j} w(i, j, Y_l)}{\sum_{k,m} e^{\mu_k + \nu_m} w(k, m, Y_l)}. \qquad \square \end{aligned}$$

Proof of Theorem 2. Without loss of generality, assume either $S > 1$ or $K > 1$. Let μ_0, ν_0 correspond to the population values via (2.2). Note that $\hat{\lambda}$, $\hat{h}$

exist if and only if the concave function

$$(3.2)\quad \begin{aligned} L_n(\nu,\mu) - L_n(\nu_0,\mu_0) = &-\sum_l \hat{g}_l^*(b(Y_l,\nu,\mu) - b(Y_l,\nu_0,\mu_0)) \\ &+\sum_i (\nu_i - \nu_{i0})\hat{\lambda}_i^* + \sum_j (\mu_j - \mu_{j0})\hat{h}_j^*, \end{aligned}$$

where

$$\nu_s = \log\left(1 - \sum_{i+1}^{S-1} e^{\nu_i}\right), \qquad \mu_K = \log\left(1 - \sum_{j=1}^{K-1} e^{\mu_j}\right)$$

achieves its maximum as a function of $S + K - 2$ variables. Let G_0^* be the marginal distribution of Y and h_0^* that of X. Then, as $n \to \infty$,

$$(3.3)\quad \begin{aligned} &\sup\{|L_n(\nu,\mu) - L_n(\nu_0,\mu_0) - L(\nu,\mu) + L(\nu_0,\mu_0)|: \\ &\qquad \nu,\mu \text{ in a compact neighbourhood of } (\nu_0,\mu_0)\} \to 0, \end{aligned}$$

where

$$L(\nu,\mu) = -\int b(y,\nu,\mu)\,dG_0^*(y) + \sum_i \nu_i\lambda_{0i}^* + \sum_j \mu_j h_{0j}^*.$$

Uniformity of convergence follows since for $|\nu - \nu_0|, |\mu - \nu_0|$ sufficiently small,

$$b(y,\nu,\mu) - b(y,\nu_0,\mu_0) = \log\left(\frac{\sum_{i,j} w(i,j,y)\lambda_i h_i}{\sum_{i,j} w(i,j,y)\lambda_{i0}h_{j0}}\right)$$

is uniformly bounded and equicontinuous in λ, h. But L is concave and has as its Hessian

$$H = -E_0(\mathrm{Var}(1(I=i), 1(X=j)): 1 \le i \le s-1, 1 \le j \le K-1|Y).$$

Since L is strictly concave and

$$\nabla L(\nu_0,\mu_0) = 0,$$

L is maximized uniquely at (μ_0, ν_0). The concavity of L_n and (3.3) now imply that, with probability tending to 1, L_n is maximized in the interior of any neighborhood of (ν_0, μ_0) and consistency of the MLE follows. This result can also be obtained in a less self-contained fashion but more directly using Brown (1985) and Ritov (1987).

(b) The equations (2.4) and (2.5) are just an ordinary set of M equations, $\int\psi(i,x,y,\lambda,h)\,d\hat{P}(i,x,y) = 0$, where

$$\psi(i,x,y,\lambda,h) = \psi_1(i,\lambda,h) + \psi_2(x,\lambda,h) + \psi_2(y,\lambda,h)$$

and $\hat{P}$ is the empirical distribution of (I, X, Y). Then (b) follows from standard results; see, for example, Huber (1967).

(c) From (2.6),

$$\hat{G}(y) = \hat{C}^{-1}\int_{-\infty}^{y}\Big(\sum_{i,j}\hat{\lambda}_i\hat{h}_j w(i, x_j, u)\Big)^{-1} d\hat{G}^*(u),$$

where

$$\hat{C} = \int\Big(\sum_{i,j}\hat{\lambda}_i\hat{h}_j w(i, x_j, u)\Big)^{-1} d\hat{G}^*(u).$$

So, write

$$\hat{G}(y) = G_0(y) + \int_{-\infty}^{\infty}\frac{g_0}{g_0^*}(u)(1(u \le y) - G_0(y))\, d\big(\hat{G}^* - G_0^*\big)(u)$$

$$(3.4) \qquad -\sum_j\big(\hat{h}_j - h_{j0}\big)\int_{-\infty}^{y}\bigg(\nu_1(u,j)\frac{g_0}{g_0^*}(u) - \int_{-\infty}^{\infty}\nu_1(t,j)\frac{g_0}{g_0^*}(t)\, dG_0(t)\bigg) dG_0(u) + R_n(y),$$

where g_0, g_0^* are the densities of G_0, G_0^* and where

$$\nu_1(u,j) = \sum_i \lambda_{i0} w(i, x_j, u), \qquad \nu_2(u,i) = \sum h_{j0} w(i, x_j, u).$$

The remainder $R_n(y)$ is easily seen to be

$$O_p\big((|\hat{\lambda} - \lambda_0|^2 + |\hat{h} - h_0|^2)\big),$$

since

$$\sup_u\bigg\{\sum_{i,j}\hat{\lambda}_i\hat{h}_j w(i, x_j, u) - \frac{g_0^*}{g_0}(u)\bigg\} \to_p 0$$

and g_0^*/g_0 is bounded away from 0 by A2.

Substitute in the approximations to $\hat{\lambda} - \lambda_0, \hat{h} - h_0$ from (b) to derive the result.

(d) The influence functions of $\hat{\lambda}$, $\hat{h}$ and $\hat{G}$ all have the structure $a(I) + b(X) + c(Y)$. But the tangent space of the model (1.1) when $\beta = 0$ consists precisely of all such functions; see BKRW, Chapter 4.5. □

PROOF OF THEOREM 3. Since the estimator of (λ, h, β) is a simple M-estimator, part (a) follows from Huber (1967), see also Theorem 2.2.5 of BKRW once we have shown that M is nonsingular. But if M is singular, then there is a vector $d \in R^{S+K-1}$, $d \neq 0$, such that

$$d^T\tilde{\psi} \perp \operatorname{span}\left\{\frac{\partial l}{\partial \lambda_i}(I, X, Y, \tau_0, G_0), \frac{\partial l}{\partial h_j}(I, X, Y, \tau_0 G_0), \right. \tag{3.5}$$

$$\left. \frac{\partial l}{\partial \beta}(I, X, Y, \tau_0, G_0), 1 \leq i \leq S-1, 1 \leq j \leq K-1\right\},$$

where l is the log-likelihood of (I, X, Y). But $\operatorname{span}\{\partial l/\partial \lambda_i\}$, $1 \leq i \leq S-1$, is clearly a subset of all functions of I and $\operatorname{span}\{\partial l/\partial h_j\}$, $1 \leq j \leq K-1$, is a subset of all functions of X. Actually, since $\partial l/\partial \lambda_i$, $\partial l/\partial h_j$, $1 \leq i \leq S-1$, $1 \leq j \leq K-1$ are linearly independent vectors in $L_2(P_0)$.

$$\operatorname{span}\left\{\frac{\partial l}{\partial \lambda_i}, \frac{\partial l}{\partial h_j}, 1 \leq i \leq S-1, 1 \leq j \leq K-1\right\} = \{\text{all } a(I) + b(X)\}.$$

Hence (3.5) implies that $d^T\tilde{\psi} \perp \{\text{all } a(I) + b(X)\} \oplus \{\partial l/\partial \beta\}$. Moreover

$$\tilde{\psi} \perp c(Y - \beta_0 X) - Ec(Y - \beta_0 X|I) \quad \text{for all } c(Y - \beta_0 X) \in L_2(P_0). \tag{3.6}$$

To see this note that

$$\int \tilde{\psi}(x, y, h, \beta)\, dP_{(\lambda, h, \beta, G)}(x, y) = 0$$

for all G. Let

$$\frac{dG_\eta(y)}{dG_0} = g_0 \frac{(1 + \eta c(y))^2}{2} \frac{\left(1 + \eta^2 \|c(Y)\|^2\right)^{-1}}{4}$$

and $|c|$ be bounded away from 0 and ∞. The interchange of integration and differentiation in

$$\frac{\partial}{\partial \eta} \int \tilde{\psi}(x, y, \lambda, h, \beta)\, dP_{(\lambda h, \beta, G_\eta)}(x, y)$$

is then easily justified by A3 and (3.6) follows. Therefore, $d^T\tilde{\psi} \perp \operatorname{span}\{c(Y - \beta_0 X) + a(I) + b(X)\}$.

Let $\partial \tilde{l}/\partial \beta$ be the projector of $\partial l/\partial \beta$ on the orthocomplement of $\Gamma \equiv \{\text{all } a(I) + b(X) + C(Y - \beta_0 X), c(Y - \beta_0 X) \in L_2(P_0)\}$. We have obtained that (3.5) implies

$$d^T\tilde{\psi} \perp \Gamma \oplus \left\{\frac{\partial l}{\partial \beta}\right\} = \Gamma \oplus \left\{\frac{\partial \tilde{l}}{\partial \beta}\right\}.$$

Since $\tilde{\psi}_1, \dots, \tilde{\psi}_{S+K-2} \in \Gamma$ and $d \neq 0$, we must have $d_{S+K-1} \neq 0$. But then, if

$\langle \cdot , \cdot \rangle$ denotes the $L_2(P_0)$ inner product

$$0 = \left\langle d^T \tilde{\psi}, \frac{\partial \tilde{l}}{\partial \beta} \right\rangle = \left\langle d_{S+K-1} \tilde{\psi}_{S+K-1}, \frac{\partial \tilde{l}}{\partial \beta} \right\rangle = d_{S+K-1} \left\langle \gamma, \frac{\partial \tilde{l}}{\partial \beta} \right\rangle .$$

Since $\partial \tilde{l} / \partial \beta \perp \Gamma$,

$$d_{S+K-1} \left\langle \tilde{\gamma}, \frac{\partial l}{\partial \beta} \right\rangle \neq 0,$$

since $\tilde{\gamma}$ is the projection defined in A3. We get a contradiction. Hence M is nonsingular and part (a) follows.

(b) $\tilde{\psi}$ and hence the influence functions of $\hat{\lambda}, \hat{h}, \hat{\beta}, \hat{G}_\beta(t)$ are all of the form $a(I) + b(X) + c(Y - \beta_0 X) + eX(g_0'/g_0)(Y - \beta_0 X)$ (where e is scalar). But by BKRW, this is precisely the tangent space of model (1.1) and efficiency follows by Theorem 3.3.1 of BKRW. □

C1–C3 IMPLY A4(a) AND WEAK CONVERGENCE OF $\sqrt{n}(\hat{G}_{\hat{\beta}} - G_0)$. To establish A4(a) it is enough [cf. Pollard (1985)] to show that $\int_0^1 x^{-1/2} \log N(x)\, dx < \infty$ where for any $\varepsilon > 0$, $N(\varepsilon)$ is the smallest cardinality of a class $\mathbf{F}_\varepsilon$ with the following property. For each of λ, h, β in a neighborhood of λ_0, h_0, β_0, there are functions $\bar{f}_\varepsilon$ and $\underline{f}_\varepsilon$ such that $\underline{f}_\varepsilon \leq f(x, y, \lambda, h, \beta) \leq \bar{f}^\varepsilon$ and $E(\bar{f}_\varepsilon - f_\varepsilon)^2 \leq \varepsilon$, where

$$f(x, y, \lambda, h, \beta) = \frac{\Sigma_{ij} \lambda_i h_j p(x, y - \beta x) w\big(i, j, y - \beta(x - x_j)\big)}{\Sigma_{ij} \lambda_i h_j w\big(i, j, y - \beta(x - x_j)\big)} .$$

Now A2 ensures that the denominator of f is bounded away from zero and hence the dependence of f on λ and h is simple. The dependence of γ on β is controlled by C1, while C3 controls the dependence of f on β through the weight functions. [Note the weight function at β' is equal to the weight function at β everywhere except on a finite number of intervals, each of length $O(|\beta - \beta'|)$]. C1–C3 taken together ensure that

$$E\Big(f(X, Y, \lambda, h, \beta) - \inf_{|\beta' - \beta| < \delta} f(X, Y, \lambda, h, \beta') \Big)^2 = O(\delta^{1 \vee \alpha}).$$

$N(\varepsilon)$ is, therefore $O(\varepsilon^{\{S+K+1/2(\alpha \vee 1)\}})$ and (3.1) is satisfied.

Moreover, C1–C3 guarantee by standard fluctuation inequality argument as in Billingsley (1968) that the processes $n^{-1/2} \Sigma_{k=1}^n \{a(X_k, Y_k; u, \lambda, h, \beta) - Ea(X_k, Y_k; u, \lambda, h, \beta)\}$ are tight where

$$a(x, y; u, \lambda, h, \beta) = \frac{1_{\{y - \beta x \leq u\}}}{\Sigma_{i,j} \lambda_i h_j (w(i, j))} .$$

These processes then converge weakly to a Gaussian process with the same covariance structure and continuous sample functions. Using this and stan-

TABLE 1
Performance of $\sqrt{n}\,\hat{\beta}$

	Bias	SD	$I^{-1/2}$
1a	−0.1	1.3	1.23
b	−0.0	1.4	1.43
2a	0.2	1.2	1.19
b	0.1	1.8	1.25

dard Taylor expansion arguments, we deduce that

$$(3.7)\qquad \sqrt{n}\left(\hat{G}_{\hat{\beta}} - G_0\right) = \sqrt{n}\left(\hat{G}_{\beta_0} - G_0\right) + \left.\frac{\partial G_\beta}{\partial \beta}\right|_{\beta_0} \sqrt{n}\left(\hat{\beta} - \beta_0\right) + o_P(1)$$

and the weak convergence of $\sqrt{n}(\hat{G}_\beta - G_0)$ to W follows. The covariance structure may be deduced from Theorem 3(a) and Theorem 2(c) since the two terms in (3.7) are independent.

4. Simulation. The formulae in this situation and the simulation are the work cf Yonghua Wang.

We consider four situations: In all cases, $I = 2$, $\lambda_1 = 0.2$, $K = 2$, $x_1 = -1$, $x_2 = 1$ and $n = 100$. If $\beta = 1$, $h_1 = p = 1 - q$, the four situations are

1. $G_0 = \mathbf{N}(0,1)$	2. $G_0 =$ Logistic $(0, 0.55)$
a. $p = 0.5$	a. $p = 0.5$
b. $p = 0.75$	b. $p = 0.75$.

G_0 in 2 is the logistic distribution with mean 0 and variance 1.

For each of these situations we performed 100 simulations and obtained Monte Carlo estimates of the mean and standard deviation of $\sqrt{n}\,\hat{\beta}$ corresponding to $\gamma(x, y) = xy$ which we expect to be efficient if G_0 is Gaussian. We also computed using the theory developed in BKRW, Section 4.4, the theoretical information bounds for estimation of β.

The results using other measures of the center and spread of $\hat{\beta}$ such as the median and interquantile range are consistent with these.

The agreement in the Gaussian case is excellent. The asymptotic variance appears as usual to be approached from below. The difference in case 2(b) presumably reflects not only the difference between $n = 100$ and $n = \infty$ but also the Monte Carlo error.

REFERENCES

BEGUN, J., HALL, W. J., HUANG, W. and WELLNER, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.* **11** 432–452.

BHATTACHARYA, P. K., CHERNOFF, H. and YANG, S. S. (1983). Nonparametric estimation of the slope of a truncated regression. *Ann. Statist.* **11** 505–514.

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. (1992). Efficient and adaptive estimation in non- and semi-parametric models. Johns Hopkins Univ. Press. To appear.

BILLINGSLEY, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

BROWN, B. M. (1985). Multiparameter linearization theorems. *J. Roy Statist. Soc. Ser. B* **47** 323–331.

BROWN, L. D. (1987). *Fundamentals of Statistical Exponential Families*. IMS, Hayward, Calif.

BUCKLEY, J. and JAMES, I. (1979). Linear regression with censored data. *Biometrika* **66** 429–436.

COSLETT, S. R. (1981). Maximum likelihood estimators for choice-based samples. *Econometrica* **49** 1289–1316.

GILL, R. D., VARDI, Y. and WELLNER, J. A. (1988). Large sample theory of empirical distributions in biased sampling models. *Ann. Statist.* **16** 1069–1112.

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **1** 221–233. Univ. California Press, Berkeley.

JEWELL, N. (1985). Regression from stratified samples of dependent variable. *Biometrika* **72** 11–21.

JEWELL, N. and QUESENBERRY, C. P. (1986). Regression analysis based on stratified samples. *Biometrika* **73** 605–614.

MALLOWS, C. (1985). Discussion of "Empirical distributions in selection bias models" by Y. Vardi. *Ann. Statist.* **13** 204–205.

MANSKI, C. and LERMAN, S. (1977). The estimation of choice probabilities from choice-based samples. *Econometrica* **45** 1977–1988.

POLLARD, D. (1985). New ways to prove central limit theorems. *Econometric Theory* **1** 295–314.

RITOV, Y. (1987). Tightness of monotone random fields. *J. Roy. Statist. Soc. Ser. B* **49** 331–333.

TSUI, K., JEWELL, N. and WU. C. F. (1988). A nonparametric approach to the truncated regression problem. *J. Amer. Statist. Assoc.* **83** 785–792.

VARDI, Y. (1982). Nonparametric estimation in the presence of length bias. *Ann. Statist.* **10** 616–620.

VARDI, Y. (1985). Empirical distributions in selection bias models. *Ann. Statist.* **13** 118–203.

WOODROOFE, M. (1985). Estimating a distribution function with truncated data. *Ann. Statist.* **13** 163–177.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
STATISTICAL LABORATORY
BERKELEY, CALIFORNIA 94720

DEPARTMENT OF STATISTICS
HEBREW UNIVERSITY
JERUSALEM
ISRAEL