Journal of Econometrics ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Semiparametric binary regression models under shape constraints with an application to Indian schooling data

Moulinath Banerjee [a,*], Debasri Mukherjee [b], Santosh Mishra [c]

[a] University of Michigan, United States
[b] Western Michigan University, United States
[c] Oregon State University, United States

ARTICLE INFO

Article history:
Available online xxxx

JEL classification:
C1

Keywords:
Convex minorants
Likelihood ratio statistic
School attendance
Semiparametric binary regression

ABSTRACT

We consider estimation of the regression function in a semiparametric binary regression model defined through an appropriate link function (with emphasis on the logistic link) using likelihood-ratio based inversion. The dichotomous response variable Δ is influenced by a set of covariates that can be partitioned as (X, Z) where Z (real valued) is the covariate of primary interest and X (vector valued) denotes a set of control variables. For any fixed X, the conditional probability of the event of interest $(\Delta = 1)$ is assumed to be a non-decreasing function of Z. The effect of the control variables is captured by a regression parameter β. We show that the baseline conditional probability function (corresponding to $X = 0$) can be estimated by isotonic regression procedures and develop a likelihood ratio based method for constructing asymptotic confidence intervals for the conditional probability function (the regression function) that avoids the need to estimate nuisance parameters. Interestingly enough, the calibration of the likelihood ratio based confidence sets for the regression function no longer involves the usual χ^2 quantiles, but those of the distribution of a new random variable that can be characterized as a functional of convex minorants of Brownian motion with quadratic drift. Confidence sets for the regression parameter β can however be constructed using asymptotically χ^2 likelihood ratio statistics. The finite sample performance of the methods are assessed via a simulation study. The techniques of the paper are applied to data sets on primary school attendance among children belonging to different socio-economic groups in rural India.

1. Introduction

Binary regression models are used frequently to model the effects of covariates on dichotomous outcome variables. A general formulation of parametric binary regression models runs as follows. If Δ is the indicator of the outcome and X is a set of (d-dimensional) covariates believed to influence Δ, one can write $\tilde{g}(\mu(X)) = \beta^{\mathrm{T}} X$, where the regression function $\mu(X) = P(\Delta = 1 \mid X)$ and $\tilde{g}$, the link, is a smooth monotone increasing function from $(0, 1)$ to $(-\infty, \infty)$. Such models are very well-studied in the literature (see, for example, McCullagh and Nelder (1989)) from both computational and theoretical angles. Some commonly used link functions are the logit (logistic regression), the probit and the complementary log–log link. In this paper, our interest is in situations where besides X, we have another (real-valued) covariate Z whose effect on the outcome variable is known qualitatively. More specifically, larger values of Z tend to make the outcome $(\Delta = 1)$ more likely. The effect of Z in the model can be incorporated as follows. Write,

$$\tilde{g}(\mu(X, Z)) = \beta^{\mathrm{T}} X + \psi(Z) \quad (\psi \text{ increasing}). \tag{1.1}$$

Note that: (a) $\mu(X, Z)$, the conditional probability of the outcome, is increasing in Z for fixed X and (b) the nonparametric component affects the conditional probability of an outcome additively on the scale of the link function.[1] Models of this kind are useful in a variety of settings and are therefore of considerable interest. See, for example, Dunson (2003) and Dunson and Neelon (2003) where nonparametric estimation of ψ as in (1.1) above is done in a Bayesian framework; for more general treatments, where the conditional mean of the outcome variable is monotone in one of the regressors, see Manski and Tamer (2002) and Magnac and Maurin (2007) and also, related work by Manski (1988) and Magnac and Maurin (2004).

This paper treats a semiparametric binary regression model of the type described in (1.1), from the angle of likelihood inference,

* Corresponding author.
E-mail address: moulib@umich.edu (M. Banerjee).

[1] The assumption that ψ is increasing is not restrictive; if the dependence on Z is decreasing, one can use the transformed covariate $\tilde{Z} \equiv -Z$.

0304-4076/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2008.11.002

based on i.i.d. observations $\{\Delta_i, X_i, Z_i\}_{i=1}^n$ from the distribution of (Δ, X, Z). Our inference strategies are based on the maximization of the underlying likelihood function (to be described in Section 2). More specifically, we focus on testing the hypotheses: (a) $H_0 : \beta = \beta_0$ and (b) $\tilde{H}_0 : \psi(z_0) = \theta_0$ for some fixed point z_0 using the likelihood ratio statistic (henceforth LRS).

The key contributions of our approach are as follows

The first is the break from the more conventional smoothness assumptions on the nonparametric component ψ; indeed, our smoothness assumptions are minimal as we only require the function to be continuously differentiable. Rather, we impose a shape constraint on ψ that is dictated by background knowledge about the effect of Z on the outcome. One major advantage of this approach stems from the fact that shape constraints automatically "regularize" the estimation problem in the sense that the underlying likelihood function can be meaningfully maximized *without penalization or kernel smoothing*. Thus, this procedure avoids the well-known problems of choosing a penalization or smoothing parameter.

Secondly, the use of likelihood ratio statistics for making inferences on β and ψ – recall that we test the hypotheses: (a) $H_0 : \beta = \beta_0$ and (b) $\tilde{H}_0 : \psi(z_0) = \theta_0$ for some fixed point z_0, using the LRS – provides a simple but elegant way of constructing confidence sets, not only for these parameters but also for the conditional probability function/regression function ($\mu(x, z) = E(\Delta = 1 \mid X = x, Z = z)$) – a quantity of significant interest – circumventing the problem of estimating nuisance parameters. We elaborate in the next paragraph. While (a) allows us to build confidence sets for the regression parameter, (b) leads to confidence intervals for the regression function itself at different covariate profiles, as is explained in detail in Section 3. From a technical standpoint, the problem in (a) – which is dealt with in Theorem 3.1 – is similar to those posed in Murphy and Van der Vaart (1997) and can be addressed using techniques developed in that paper. But the problem in (b) is radically different from the set-up of Murphy and Van der Vaart (1997), as it involves estimating a parameter that can only be estimated at a slower ($n^{1/3}$) convergence rate than the regression parameter β (which is estimable at rate $\sqrt{n}$) and exhibits non-standard asymptotics. *The solution to this problem – the result of* Theorem 3.3 *– requires extensive use of techniques from isotonic regression and convex estimation and is the novel contribution of the current paper, at a technical level.*

We show that the LRS for testing H_0 has a limiting χ_d^2 distribution as in regular parametric problems, while that for testing $\tilde{H}_0 : \psi(z_0) = \theta_0$ converges in distribution to a "universal" random variable $\mathbb{D}$; "universal" in the sense that it does not depend on the underlying parameters of the model or the point of interest z_0.[2] This latter result is a new and powerful one as it can be used to obtain confidence sets for $\mu(x, z)$ by inverting likelihood ratio tests for testing a family of hypotheses of type (b). We emphasize that the computation of the LRS is completely objective and *does not involve smoothing/penalization* as discussed above. Furthermore, calibration of the likelihood ratio test only involves knowledge of the quantiles of (the asymptotic pivot) $\mathbb{D}$, which are well tabulated. Hence, nuisance parameters need not be estimated from the data. Of course, we reap a similar advantage while constructing confidence sets for the regression parameter β (or a sub-parameter as is discussed in Section 3) which involves inverting a family of likelihood ratio tests as in (a), calibrated via the usual χ^2 quantiles. In contrast, note that inferences for β and ψ (equivalently μ) could also be done using the limit distributions of the corresponding maximum likelihood estimates. However, we do not adopt this route as these distributions involve nuisance parameters that are difficult to estimate. One could argue that nuisance parameter estimation in this context could be obviated through the use of resampling techniques like subsampling (Politis et al., 1999) or the m out of n bootstrap (the usual Efron-type bootstrap will not work in this situation (Sen et al., 2008)). But this once again introduces the problem of choosing m, the "block size", which can be regarded as a variant of a smoothing parameter.

Thus, the likelihood ratio method is much more *automated* and *objective* than its competitors.

As far as technicalities are concerned, the methodology and asymptotic theory developed in this paper is markedly different from those used in the smoothing literature. This arises from the fact that maximum likelihood estimation under monotonicity constraints can typically be reduced to an isotonic regression problem (for a comprehensive review see Robertson et al. (1988) and more recently, Silvapulle and Sen (2004)). It is well known in the isotonic regression literature that the asymptotic behavior of estimates (like the MLEs of ψ in our model) obtained through such regression cannot be analyzed using standard CLTs as they are highly nonlinear functionals of the empirical distribution[3]; rather they are asymptotically distributed as the derivatives of convex minorants of Gaussian processes, which are non-normal. These technical details are discussed in later sections.

The techniques developed in this paper are applied to model primary school enrollment in Indian villages using a logistic regression framework. Our main variable of interest (Z) is standard of living and auxiliary covariates are SCST (the indicator variable for low caste, to be explained later), cost of going to school, some government policy measures like the midday-meal (meals will be provided if the child goes to school), TLC (total literacy campaign) and HEADLIT (a binary variable that indicates whether the head of the household is literate). The probability of attending school (Δ is the dichotomous response variable that indicates whether the child is enrolled in school) increases as the living standard of the family increases. We consider household level data separately for male and female children in both BIMARU (highly underdeveloped, to be defined later in detail) and NonBIMARU (not so backward) states. Our results indicate that despite the popular measures taken by the government, the probability of going to school is much lower for backward states and for backward castes (SCSTs) as compared to that for less backward states and higher castes. The probability of a female child going to school is much lower than that of a male child.

The rest of the paper is organized as follows. Maximum likelihood estimation and novel likelihood ratio-based inferential procedures for a general link function are discussed in Section 2. In Section 3, for concreteness, we focus primarily on the logit link, which is the most widely used link function in statistical data analysis and discuss the asymptotic results and the associated methodology for construction of confidence sets in the setting of this model. We also indicate in Section 3 that similar results continue to hold for other commonly used link functions, like the probit link or the complementary log–log link. Simulation studies and applications to real data appear in Section 4. We conclude with some discussion in Section 5. Proofs of some of the results in Section 3 are collected in the Appendix.

[2] This limit distribution does not belong to the χ_1^2 family but can be thought of as an *analogue* of the χ_1^2 distribution in nonregular statistical problems involving $n^{1/3}$ rate of convergence for maximum likelihood estimators and non-Gaussian limit distributions. Indeed, the maximum likelihood estimator $\hat{\psi}_n$ converges to the true ψ at rate $n^{1/3}$ in this problem, while the rate of convergence of $\hat{\beta}$ is $\sqrt{n}$.

[3] This is in contrast to estimates based on smoothing methods, which are essentially linear functionals of the empirical distribution, which enables the use of CLTs and leads to asymptotic normality.

Please cite this article in press as: Banerjee, M., et al., Semiparametric binary regression models under shape constraints with an application to Indian schooling data. Journal of Econometrics (2009), doi:10.1016/j.jeconom.2008.11.002

2. Computing mles and likelihood ratios

The density function of the random vector (Δ, X, Z) is given by $p(\delta, x, z) = \mu(x, z)^{\delta}(1 - \mu(x, z))^{1-\delta} f(z, x)$, where $f(z, x)$ is the joint density of (Z, X) with respect to Leb $\times \mu$ where Leb denotes Lebesgue measure on $[0, \infty)$ and μ is some measure defined on $\mathbb{R}^d$. We construct the likelihood function for the data, as:

$$L_n(\beta, \psi, \{\Delta_i, X_i, Z_i\}_{i=1}^n) = \prod_{i=1}^n \mu(X_i, Z_i)^{\Delta_i} \times (1 - \mu(X_i, Z_i))^{1-\Delta_i} f(Z_i, X_i). \quad (2.2)$$

In what follows, we denote the true underlying values of the parameters (β, ψ) by (β_0, ψ_0). Using a link function $\tilde{g}$ (satisfying **Condition C:** For $\delta = 1$ or 0, the function $v(s) := \delta \log \tilde{h}(s) + (1 - \delta) \log(1 - \tilde{h}(s))$ is concave for $s \in (-\infty, \infty)$[4]), with inverse function $\tilde{h}$, the log-likelihood function for the sample, up to an additive factor that does not involve any of the parameters of interest, is given by

$$l_n(\beta, \psi) \equiv \log L_n(\beta, \psi, \{\Delta_i, X_i, Z_i\}_{i=1}^n) = \sum_{i=1}^n l(\beta, \psi, \Delta_i, X_i, Z_i)$$

where

$$l(\beta, \psi, \delta, x, z) = \delta \log \tilde{h}(\beta^{\mathrm{T}} x + \psi(z)) + (1 - \delta) \times \log(1 - \tilde{h}(\beta^{\mathrm{T}} x + \psi(z))). \quad (2.3)$$

We next introduce some notation and define some key quantities that will be crucial to the subsequent development. Let $Z_{(1)}, Z_{(2)}, \ldots, Z_{(n)}$ denote the ordered values of the Z_i's; let $\Delta_{(i)}$ and $X_{(i)}$ denote the indicator and covariate values associated with $Z_{(i)}$. Also, let $u_i \equiv \psi(Z_{(i)})$ and $R_i(\beta) = \beta^{\mathrm{T}} X_{(i)}$. Denote the vector $(u_1, u_2, \ldots, u_n)$ by $\mathbf{u}$ and define the function g as: $g(\beta, \mathbf{u}) \equiv -l_n(\beta, \psi) = \sum_{i=1}^n \phi(\Delta_{(i)}, R_i(\beta), u_i)$ where $\phi(\delta, r, u) = -\delta \log \tilde{h}(r + u) - (1 - \delta) \log(1 - \tilde{h}(r + u))$. Note that the monotone function ψ is identifiable only up to its values at the $Z_{(i)}$'s; hence we identify ψ with the vector $\mathbf{u}$. Let $(\hat{\beta}_n, \hat{\mathbf{u}}_n) = \operatorname{argmin}_{\beta \in \mathbb{R}^d, \{\mathbf{u}: u_1 \le u_2 \le \cdots \le u_n\}} g(\beta, \mathbf{u})$. The unconstrained MLE of (β, ψ) is given by $(\hat{\beta}_n, \hat{\psi}_n)$ where $\hat{\psi}_n$ is the (unique) right-continuous increasing step function that assumes the value $\hat{u}_{i,n}$ (the ith component of $\hat{\mathbf{u}}_n$) at the point $Z_{(i)}$ and has no jump points outside of the set $\{Z_{(i)}\}_{i=1}^n$.

Next, for a fixed β, let $\hat{\mathbf{u}}_n^{(\beta)} = \operatorname{argmin}_{\{\mathbf{u}: u_1 \le u_2 \le \cdots \le u_n\}} g(\beta, \mathbf{u})$. Define $\hat{\psi}_n^{(\beta)}$ to be the (unique) right-continuous increasing step function that assumes the value $\hat{u}_{i,n}^{(\beta)}$ (the ith component of $\hat{\mathbf{u}}_n^{(\beta)}$) at the point $Z_{(i)}$ and has no jump points outside of the set $\{Z_{(i)}\}_{i=1}^n$. Then, note that: $\hat{\beta}_n = \operatorname{argmin}_{\beta} g(\beta, \hat{\mathbf{u}}_n^{(\beta)})$ and $\hat{\psi}_n = \hat{\psi}_n^{(\hat{\beta}_n)}$.

2.1. The likelihood ratio statistic for testing the value of β

The likelihood ratio statistic for testing $H_0 : \beta = \beta_0$ is given by:

$$\mathrm{lrtbeta}_n = 2(l_n(\hat{\beta}_n, \hat{\psi}_n) - l_n(\beta_0, \hat{\psi}_n^{(\beta_0)})). \quad (2.4)$$

Our computation of MLEs, $(\hat{\beta}_n, \hat{\psi}_n, \hat{\psi}_n^{(\beta_0)})$, in this semiparametric problem relies heavily on the convexity of $g(\beta, \mathbf{u})$ and is based on the following proposition.

Proposition 1. *Let $f(\gamma_1, \gamma_2)$ be a real-valued function defined on $\mathbb{R}^{k_1} \times \mathcal{C}$ where γ_1 varies in $\mathbb{R}^{k_1}$ and $\gamma_2 \in \mathcal{C}$, where $\mathcal{C}$ is a closed convex subset of $\mathbb{R}^{k_2}$. Assume that f is a continuously differentiable strictly convex function that is minimized (uniquely) at the point $(\gamma_1^\star, \gamma_2^\star)$. Consider the following updating algorithm. Start at an arbitrary point (γ_1^0, γ_2^0) in $\mathbb{R}^{k_1} \times \mathcal{C}$. Having defined (γ_1^m, γ_2^m) at stage m $(m \ge 0)$, set $\gamma_2^{m+1} \equiv \operatorname{argmin}_{\gamma_2 \in \mathcal{C}} f(\gamma_1^m, \gamma_2)$ and γ_1^{m+1} as the (unique) solution to $(\partial/\partial\, \gamma_1) f(\gamma_1, \gamma_2^{m+1}) = 0$. Then, irrespective of the starting value, the sequence of points $\{\gamma_1^m, \gamma_2^m\}_{m \ge 0}$ converges to $(\gamma_1^\star, \gamma_2^\star)$.*

Remark 1. We do not provide a proof of this proposition in this paper. The proposition follows as a direct consequence of Theorem 2.2 in Jongbloed (1998) on the convergence of an iteratively defined sequence using an algorithmic map which is adapted from a general convergence theorem (Theorem 7.2.3) from Bazaraa et al. (1993).[5]

Consider first, the computation of the unconstrained MLEs $(\hat{\beta}_n, \hat{\psi}_n)$. The function g is defined on $\mathbb{R}^d \times \tilde{\mathcal{C}}$, where $\tilde{\mathcal{C}} \equiv \{\mathbf{u} = (u_1, u_2, \ldots, u_n) : u_1 \le u_2 \le \cdots \le u_n\}$ is a closed convex cone in $\mathbb{R}^n$. Setting f in Proposition 1 to be g, $\gamma_1 = \beta$ and $\gamma_2 = \mathbf{u}$, it is easy to check that g is a continuously differentiable and strictly convex function that attains a unique minimum at $(\hat{\beta}_n, \hat{\mathbf{u}}_n)$. Thus, we are in the setting of Proposition 1 above. We next provide a step-by-step outline of the algorithm to evaluate $(\hat{\beta}_n, \hat{\psi}_n)$.

Computing the unconstrained MLEs:

Step 1. At Stage 0 of the algorithm, propose initial estimates $(\hat{\beta}_n^{(0)}, \hat{u}_n^{(0)})$. Also, set an initial tolerance level $\eta > 0$, small.

Step 2a. At Stage $p \ge 0$ of the algorithm, current estimates $(\hat{\beta}_n^{(p)}, \mathbf{u_n}^{(p)})$ are available. At Stage $p + 1$, first update the second component to $\mathbf{u_n}^{(p+1)}$ by minimizing $g(\hat{\beta}_n^{(p)}, \mathbf{u})$ over $\mathbf{u} \in \tilde{\mathcal{C}}$, using *the modified iterative convex minorant algorithm* (MICM) due to Jongbloed (1998). Note that $\mathbf{u_n}^{(p+1)}$ is precisely the vector $\{\hat{\psi}_n^{(\beta)}(Z_{(i)})\}_{i=1}^n$, for $\beta = \hat{\beta}_n^{(p)}$.

Step 2b. Having updated to $\mathbf{u_n}^{(p+1)}$, next update $\hat{\beta}_n^{(p)}$ to $\hat{\beta}_n^{(p+1)}$ by solving $(\partial/\partial\, \beta) g(\beta, \mathbf{u_n}^{(p+1)}) = 0$ using, for example, the Newton–Raphson procedure. In terms of the log-likelihood function, this amounts to solving $(\partial/\partial\, \beta) l_n(\beta, \psi) = 0$ for $\psi = \hat{\psi}_n^{(\hat{\beta}_n^{(p)})}$.

Step 3. (*Checking convergence.*) If

$$\left| \frac{g(\hat{\beta}_n^{(p+1)}, \mathbf{u}_n{}^{(p+1)}) - g(\hat{\beta}_n^{(p)}, \mathbf{u}_n{}^{(p)})}{g(\hat{\beta}_n^{(p)}, \mathbf{u}_n{}^{(p)})} \right| \le \eta$$

then stop and declare $(\hat{\beta}_n^{(p+1)}, \mathbf{u}_n{}^{(p+1)})$ as the MLEs. Otherwise, set $p = p + 1$ and return to Step 2a.

We now elaborate on Step 2a, the most involved segment of the above algorithm, that requires iterative quadratic optimization techniques under order constraints. This is precisely the problem of evaluating $\hat{\psi}_n^{(\beta)}$, the MLE of ψ for a fixed β. In particular, recall that $\hat{\psi}_n^{(\beta_0)}$ is the MLE of ψ under $H_0 : \beta = \beta_0$.

Characterizing and computing $\hat{\psi}_n^{(\beta)}$: This is characterized by the vector $\hat{\mathbf{u}}_n^{(\beta)} = (\hat{u}_{1,n}^{(\beta)} \le \hat{u}_{2,n}^{(\beta)} \cdots \le \hat{u}_{n,n}^{(\beta)})$ that minimizes $g(\beta, \mathbf{u})$ over

[4] It is easy to check that all three standard link functions used in binary regression: (a) the logit link for which $\tilde{h}(s) = e^s/(1 + e^s)$, (b) the probit link for which $\tilde{h}(s) = \Phi(s)$, Φ denoting the normal cdf and (c) the complementary log–log link for which $\tilde{h}(s) = 1 - e^{-e^s}$, satisfy this property. The concavity of v implies the convexity of the function $g(\beta, \mathbf{u})$, to be introduced soon, in its arguments and guarantees a unique minimizer that may be obtained by using standard methods from convex optimization theory.

[5] **Additional remark:** Note that the step of updating γ_1^m to γ_1^{m+1} which involves solving $(\partial/\partial\, \gamma_1) f(\gamma_1, \gamma_2^{m+1}) = 0$ is also a minimization step. Since $f(\gamma_1, \gamma_2^{m+1})$ is a continuously differentiable strictly convex function of γ_1, it is uniquely minimized at the point where its derivative is 0. Thus, we could alternatively have written: $\gamma_1^{m+1} \equiv \operatorname{argmin}_{\gamma_1 \in \mathbb{R}^{k_1}} f(\gamma_1, \gamma_2^{m+1})$.

Please cite this article in press as: Banerjee, M., et al., Semiparametric binary regression models under shape constraints with an application to Indian schooling data. Journal of Econometrics (2009), doi:10.1016/j.jeconom.2008.11.002

all $u_1 \le u_2 \le \cdots \le u_n$.[6] Before proceeding further, we introduce some notation. For points $\{(x_0, y_0), (x_1, y_1), \ldots, (x_k, y_k)\}$ where $x_0 = y_0 = 0$ and $x_0 < x_1 < \cdots < x_k$, consider the left-continuous function $P(x)$ such that $P(x_i) = y_i$ and such that $P(x)$ is constant on (x_{i-1}, x_i). We will denote the vector of slopes (left-derivatives) of the greatest convex minorant (henceforth GCM) of $P(x)$ computed at the points $(x_1, x_2, \ldots, x_n)$ by slogcm $\{(x_i, y_i)\}_{i=0}^n$.

The solution $\hat{\mathbf{u}}_n^{(\beta)}$ can be viewed as the slope of the greatest convex minorant (slogcm) of a random function defined in terms of $\hat{\mathbf{u}}_n^{(\beta)}$ itself. This *self-induced/self-consistent* characterization proves useful both for computational purposes and for the asymptotic theory. For the sake of notational convenience, we will denote $g(\beta, \mathbf{u})$ in the following discussion by $\xi(\mathbf{u})$ (suppressing the dependence on β) and $\hat{\mathbf{u}}_n^{(\beta)}$ by $\hat{\mathbf{u}} = (\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n)$. For $1 \le i \le n$, set $d_i = \nabla_{ii}\, \xi(\hat{\mathbf{u}})$. Define the function η as follows:

$$\eta(\mathbf{u}) = \sum_{i=1}^{n} \left[u_i - \hat{u}_i + \nabla_i\, \xi(\hat{\mathbf{u}})\, d_i^{-1}\right]^2 d_i$$
$$= \sum_{i=1}^{n} \left[u_i - \left(\hat{u}_i - \nabla_i\, \xi(\hat{\mathbf{u}})\, d_i^{-1}\right)\right]^2 d_i. \tag{2.5}$$

It can be shown that $\hat{\mathbf{u}}$ minimizes η subject to the constraints that $u_1 \le u_2 \le \cdots \le u_n$ (see Appendix A.2 for the details) and hence furnishes the isotonic regression of the function $h(i) = \hat{u}_i - \nabla_i\, \xi(\hat{\mathbf{u}})\, d_i^{-1}$ on the ordered set $\{1, 2, \ldots, n\}$ with weight function $d_i \equiv \nabla_{ii}\, \xi(\hat{\mathbf{u}})$. It is well known that the solution $\hat{\mathbf{u}} = (\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n) = \text{slogcm} \left\{\sum_{j=1}^{i} d_i, \sum_{j=1}^{i} h(i)\, d_i\right\}_{i=0}^{n}$. See, for example Theorem 1.2.1 of Robertson et al. (1988) and more generally, Chapter 1 of that book for an extensive discussion of isotonic regression.

Since $\hat{\mathbf{u}}$ is unknown, an iterative scheme is resorted to. For a fixed vector $\mathbf{v} \equiv (v_1, v_2, \ldots, v_n) \in \mathcal{C}$, set $d_{\mathbf{v},i} = \nabla_{ii}\, \xi(\mathbf{v})$ and define the function $\eta_{\mathbf{v}}(\mathbf{u}) \equiv \sum_{i=1}^{n} [u_i - (v_i - \nabla_i\, \xi(\mathbf{v})\, d_{\mathbf{v},i}^{-1})]^2\, d_{\mathbf{v},i}$. Pick an initial guess for $\hat{\mathbf{u}}$, say $\mathbf{u}^{(0)} \in \mathcal{C}$, set $\mathbf{v} = \mathbf{u}^{(0)}$ and compute $\mathbf{u}^{(1)}$ by minimizing $\eta_{\mathbf{v}}(\mathbf{u})$ over $\mathcal{C}$; then, set $\mathbf{v} = \mathbf{u}^{0}$, obtain $\mathbf{u}^{(2)}$ by minimizing $\eta_{\mathbf{v}}(\mathbf{u})$ again, and proceed thus, until convergence. Generally, with $\mathbf{v} = \mathbf{u}^{(j)}$, we have: $\mathbf{u}^{(j+1)} = \text{slogcm} \left\{\sum_{j=1}^{i} d_{\mathbf{v},i}, \sum_{j=1}^{i} (v_i - \nabla_i\, \xi(\mathbf{v})\, d_{\mathbf{v},i}^{-1})\, d_{\mathbf{v},i}\right\}_{i=0}^{n}$.

Remark 2. Certain convergence issues might arise with such a straightforward iterative scheme, since the algorithm could hit inadmissible regions in the search space. Jongbloed (1998) addresses this issue by using a modified iterated convex minorant (henceforth MICM) algorithm; see Section 2.4 of his paper for a discussion of the practical issues and a description of the relevant algorithm which incorporates a line search procedure to guarantee convergence to the minimizer. As this is a well-established algorithm in the isotonic regression literature, we do not discuss these subtleties any further, but refer the reader to Jongbloed's paper.

Remark 3. We have also not explicitly addressed the convergence issue. This is discussed in Jongbloed (1998). The iterations are stopped when the (necessary and sufficient) conditions that characterize the unique minimizer of ξ are satisfied to a pre-specified degree of tolerance. For a discussion of these conditions, see Appendix A.1.

[6] Without loss of generality one can assume that $\Delta_{(1)} = 1$ and $\Delta_{(n)} = 0$. If not, the effective sample size for the estimation of the parameters is $k_2 - k_1 + 1$ where k_1 is the first index i such that $\Delta_{(i)} = 1$ and k_2 is the last index such that $\Delta_{(i)} = 0$. It is not difficult to see that one can set $\hat{u}_{i,n}^{(\beta)} = -\infty$ for all $i < k_1$ and $\hat{u}_{i,n}^{(\beta)} = \infty$ for all $i > k_2$ without imposing any constraints on the other components of the minimizing vector.

2.2. The likelihood ratio statistic for testing the value of ψ at a point

We next turn our attention to the likelihood ratio test for testing $\tilde{H}_0 : \psi(z_0) = \theta_0$ with $-\infty < \theta_0 < \infty$. This requires us to compute the constrained maximizers of β and ψ, say $(\hat{\beta}_{n,0}, \hat{\psi}_{n,0})$ under $\tilde{H}_0 : \psi(z_0) = \theta_0$. As in the unconstrained case, this maximization can be achieved in two steps. For each β, one can compute $\hat{\psi}_{n,0}^{(\beta)} = \text{argmax}_{\psi:\psi(z_0)=\theta_0}\, l_n(\beta, \psi)$. Then, $\hat{\beta}_{n,0} = \text{argmax}_{\beta}\, l_n(\beta, \hat{\psi}_{n,0}^{(\beta)})$ and $\hat{\psi}_{n,0} = \hat{\psi}_{n,0}^{(\hat{\beta}_{n,0})}$. The likelihood ratio statistic for testing $\tilde{H}_0 : \psi(z_0) = \theta_0$ is given by:

$$\text{lrtpsi}_n = 2\,(l_n(\hat{\beta}_n, \hat{\psi}_n) - l_n(\hat{\beta}_{n,0}, \hat{\psi}_{n,0})). \tag{2.6}$$

Note that the monotone function $\hat{\psi}_{n,0}^{(\beta)}$ is identifiable only up to its values at the $Z_{(i)}$'s (and at the fixed point z_0 where it is required to equal θ_0) and we identify this function with the vector $\hat{\mathbf{u}}_{n,0}^{(\beta)} \equiv (\hat{u}_{1,n,0}^{(\beta)}, \hat{u}_{2,n,0}^{(\beta)}, \ldots, \hat{u}_{n,n,0}^{(\beta)})$ where $\hat{u}_{i,n,0}^{(\beta)} = \hat{\psi}_{n,0}^{(\beta)}(Z_{(i)})$. We will discuss the characterization of this vector shortly.

Before proceeding further, we introduce some notation. First, let m denote the number of Z values that are less than or equal to z_0. Then, we have $Z_{(m)} < z_0 < Z_{(m+1)}$ (with probability 1). Note that any monotone function ψ that satisfies $\tilde{H}_0$ will have: $\psi(Z_{(m)}) \le \theta_0 \le \psi(Z_{(m+1)})$. Define $\tilde{\mathcal{C}}_0$ to be the closed convex subset of $\mathbb{R}^n$ comprising all vectors $\mathbf{u}$ with $u_1 \le u_2 \le \cdots \le u_m \le \theta_0 \le u_{m+1} \le \cdots \le u_n$. If $\hat{\mathbf{u}}_{n,0} \equiv (\hat{u}_{1,n,0}, \hat{u}_{2,n,0}, \ldots, \hat{u}_{n,n,0})$ denotes the vector $\{\hat{\psi}_{n,0}(Z_{(i)})\}_{i=1}^n$, then $(\beta_{n,0}, \hat{\mathbf{u}}_{n,0}) = \text{argmin}_{\beta\in\mathbb{R}^d, \mathbf{u}\in\tilde{\mathcal{C}}_0}\, g(\beta, \mathbf{u})$. Since the function g is a continuously differentiable strictly convex function defined on the closed convex set $\mathbb{R}^d \times \tilde{\mathcal{C}}_0$ and assumes a unique minimum at $(\beta_{n,0}, \hat{\mathbf{u}}_{n,0})$, we can invoke Proposition 1 as before. The algorithm, which is similar to that in the preceding subsection, is formally presented below.

Computing the constrained MLEs under $\tilde{H}_0$:

Step 1. At Stage 0 of the algorithm, propose initial estimates $(\hat{\beta}_{n,0}^{(0)}, \hat{u}_{n,0}^{(0)})$. Also, set an initial tolerance level $\eta > 0$, small.

Step 2a. At Stage $p \ge 0$ of the algorithm, current estimates $(\hat{\beta}_{n,0}^{(p)}, \mathbf{u}_{n,0}^{(p)})$ are available. At Stage $p+1$, first update the second component to $\mathbf{u}_{\mathbf{n,0}}^{(p+1)}$ by minimizing $g(\hat{\beta}_{n,0}^{(p)}, \mathbf{u})$ over $\mathbf{u} \in \tilde{\mathcal{C}}_0$. Note that $\mathbf{u}_{n,0}^{(p+1)}$ is precisely the vector $\{\hat{\psi}_{n,0}^{\beta}(Z_{(i)})\}_{i=1}^n$, for $\beta = \hat{\beta}_{n,0}^{(p)}$.

Step 2b. Next, update $\hat{\beta}_{n,0}^{(p)}$ to $\hat{\beta}_{n,0}^{(p+1)}$ by solving $(\partial/\partial\,\beta)\, g(\beta, \mathbf{u}_{n,0}^{(p+1)}) = 0$ using, say, the Newton–Raphson method.

Step 3. (*Checking Convergence.*) If $\left|\frac{g(\hat{\beta}_{n,0}^{(p+1)}, \mathbf{u}_{n,0}^{(p+1)}) - g(\hat{\beta}_{n,0}^{(p)}, \mathbf{u}_{n,0}^{(p)})}{g(\hat{\beta}_{n,0}^{(p)}, \mathbf{u}_{n,0}^{(p)})}\right| \le \eta$, then stop and declare $(\hat{\beta}_{n,0}^{(p+1)}, \mathbf{u}_{n,0}^{(p+1)})$ as the MLEs. Otherwise, set $p = p+1$ and return to Step 2a.

It remains to elaborate on Step 2a, which involves computing $\hat{\psi}_{n,0}^{(\beta)}$ for some β.

Characterizing $\hat{\psi}_{n,0}^{(\beta)}$: Finding $\hat{\psi}_{n,0}^{(\beta)}$ amounts to minimizing $g(\beta, \mathbf{u}) = \sum_{i=1}^{n} \phi(\Delta_{(i)}, R_i(\beta), u_i)$ over all $u_1 \le u_2 \cdots \le u_m \le \theta_0 \le u_{m+1} \le \cdots \le u_n$. For the remainder of this discussion, we denote the minimizing vector $\hat{\mathbf{u}}_{n,0}^{(\beta)}$ by $\hat{\mathbf{u}}^{(0)}$. Finding $\hat{\mathbf{u}}^{(0)}$ can be reduced to solving two separate optimization problems. These are [1] Minimize $g_1(\beta, u_1, u_2, \ldots, u_m) \equiv \sum_{i=1}^{m} \phi(\Delta_{(i)}, R_i(\beta), u_i)$ over $u_1 \le u_2 \le \cdots \le u_m \le \theta_0$, and, [2] Minimize $g_2(\beta, u_{m+1}, u_{m+2}, \ldots, u_n) \equiv \sum_{i=m+1}^{n} \phi(\Delta_{(i)}, R_i(\beta), u_i)$ over $\theta_0 \le u_{m+1} \le u_{m+2} \le \cdots \le u_n$.

Consider [1] first. This is a problem that involves minimizing a smooth convex function over a convex set and one can easily write down the Kuhn–Tucker conditions characterizing the minimizer. It

is easy to see that the solution $(\hat{u}_1^{(0)}, \hat{u}_2^{(0)}, \ldots, \hat{u}_m^{(0)})$ can be obtained as follows: Minimize $g_1(\beta, \mathbf{u}_1)$ where $\mathbf{u}_1 \equiv (u_1, u_2, \ldots, u_m)$ over $\mathcal{C}_1$, the closed convex cone in $\mathbb{R}^m$ defined as $\{\mathbf{u}_1 : u_1 \le u_2 \le \cdots \le u_m\}$, to obtain $\tilde{\mathbf{u}}_1 \equiv (\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_m)$. Then, $(\hat{u}_1^{(0)}, \hat{u}_2^{(0)}, \ldots, \hat{u}_m^{(0)}) = (\tilde{u}_1 \wedge \theta_0, \tilde{u}_2 \wedge \theta_0, \ldots, \tilde{u}_m \wedge \theta_0)$. The minimization of $g_1(\beta, \cdot)$ over $\mathcal{C}_1$ requires use of the MICM and follows the same technique as described in the preceding subsection in connection with estimating $\hat{\psi}_n^{(\beta)}$. On the other hand, the solution vector to [2], say $(\hat{u}_{m+1}^{(0)}, \hat{u}_{m+2}^{(0)}, \ldots, \hat{u}_n^{(0)})$, is given by $(\hat{u}_{m+1}^{(0)}, \hat{u}_{m+2}^{(0)}, \ldots, \hat{u}_n^{(0)}) = (\tilde{u}_{m+1} \vee \theta_0, \tilde{u}_{m+2} \vee \theta_0, \ldots, \tilde{u}_n \vee \theta_0)$ where $(\tilde{u}_{m+1}, \tilde{u}_{m+2}, \ldots, \tilde{u}_n) = \operatorname{argmin}_{u_{m+1} \le u_{m+2} \le \cdots \le u_n} g_2(\beta, u_{m+1}, u_{m+2}, \ldots, u_n)$ and uses the MICM, as in [1]. Finally $\hat{\mathbf{u}}^{(0)} = (\hat{u}_1^{(0)}, \hat{u}_2^{(0)}, \ldots, \hat{u}_n^{(0)})$.

3. Asymptotic results

In this section we present asymptotic results for the estimation of β and ψ. For the sake of concreteness and ease of exposition, we present results explicitly in the setting of logistic regression. The semiparametric logistic model is given by $\log \frac{\mu(X,Z)}{1-\mu(X,Z)} = \beta^{\mathrm{T}} X + \psi(Z)$. The above display is equivalent to writing:

$$\mu(X, Z) = \frac{e^{\beta^{\mathrm{T}} X} \Lambda(Z)}{1 + e^{\beta^{\mathrm{T}} X} \Lambda(Z)}, \quad \text{where } \Lambda(Z) = e^{\psi(Z)}. \tag{3.7}$$

The parameter space for β is taken to be a bounded subset of $\mathbb{R}^d$. We denote it by $\mathcal{B}$. The parameter space for $\Lambda = e^{\psi}$ is the space of all non-decreasing cadlag (i.e. right-continuous with left-hand limits) functions from $[0, \tau]$ to $[0, M]$ where M is some large positive constant. Let (β_0, Λ_0) denote the true model parameters (thus, $\Lambda_0 = e^{\psi_0}$). We make the following assumptions:

(A.1) The true regression parameter β_0 is an interior point of $\mathcal{B}$.
(A.2) The covariate X has bounded support. Hence, there exists x_0 such that $P(\|X\| \le x_0) = 1$. Also $E(\mathrm{Var}(X \mid Z))$ is positive definite with probability one.
(A.3) Let $\tau_{\Lambda_0} = \inf\{z : \Lambda_0(z) = \infty\}$. The support of Z is an interval $[\sigma, \tau]$ with $0 < \sigma < \tau < \tau_{\Lambda_0}$.
(A.4) We assume that $0 < \Lambda_0(\sigma-) < \Lambda_0(\tau) < M$. Also, Λ_0 is continuously differentiable on $[\sigma, \tau]$ with derivative λ_0 bounded away from 0 and from ∞.
(A.5) The marginal density of Z, which we denote by f_Z, is continuous and positive on $[\sigma, \tau]$.
(A.6) The function $h^{\star\star}$ defined below in (3.8) has a version which is differentiable componentwise with each component possessing a bounded derivative on $[\sigma, \tau]$.

Remarks. The boundedness of $\mathcal{B}$ along with assumptions (A.1)–(A.3) are needed to deduce the consistency and rates of convergence of the maximum likelihood estimators. In particular, the boundedness of the covariate X does not cause a problem with applications. The utility of the assumption that the conditional dispersion of X given Z is positive definite is explained below. (A.4) and (A.5) are fairly weak regularity conditions on Λ_0 and the distribution of Z. The assumption (A.6) is a technical assumption and is required to ensure that one can define appropriate approximately *least favorable submodels*; these are finite-dimensional submodels of the given semiparametric model, with the property that the efficient score function for the semiparametric model at the true parameter values can be approximated by the usual score functions from these submodels. They turn out to be crucial for deriving the limit distribution of the likelihood ratio statistic for testing the regression parameter.

We now introduce the efficient score function for β in this model. The log density function for the vector (Δ, Z, X) is given by:

$$l_{\beta,\Lambda}(\delta, z, x) = \delta\,(\log\,\Lambda(z) + \beta^{\mathrm{T}} x) - \log\,(1 + \Lambda(z)\,\exp(\beta^{\mathrm{T}} x)) + \log f(z, x).$$

The ordinary score function for β in this model is:

$$\dot{l}_\beta(\beta, \Lambda)(\delta, z, x) = (\partial/\partial\,\beta)\, l_{\beta,\Lambda}(\delta, x, z) = x\,\Lambda(z)\,Q((\delta, z, x); \beta, \Lambda),$$

where $Q((\delta, z, x); \beta, \Lambda) = \delta\,(\Lambda(z))^{-1} - (e^{\beta^{\mathrm{T}} x})(1 + \Lambda(z)\,e^{\beta^{\mathrm{T}} x})^{-1}$. The score function for Λ is a linear operator acting on the space of functions of bounded variation on $[\sigma, \tau]$ and has the form $\dot{l}_\Lambda(\beta, \Lambda)(h(\cdot))(\delta, z, x) = h(z)\,Q((\delta, z, x); \beta, \Lambda)$. Here h is a function of bounded variation on $[\sigma, \tau]$. To compute the form of this score function, we consider curves of the form $\Lambda + t\,h$ for $t \ge 0$ where h is a non-decreasing non-negative function on $[\sigma, \tau]$. Computing $B_\Lambda(h) = \frac{\partial}{\partial t}\, l_{\beta,\Lambda+t\,h}(\delta, x, z)\,|_{t=0}$, we get $B_\Lambda(h) = h(z)\,Q((\delta, z, x); \beta, \Lambda)$. The linear operator B_Λ now extends naturally to the closed linear span of all non-decreasing h's, which is precisely the space of all functions of bounded variation on $[\sigma, \tau]$.

The efficient score function for β at the true parameter values (β_0, Λ_0), which we will denote by $\tilde{l}$ for brevity, is given by

$$\tilde{l} = \dot{l}_\beta(\beta_0, \Lambda_0) - \dot{l}_\Lambda(\beta_0, \Lambda_0)h^\star$$

for functions $h^\star = (h_1^\star, h_2^\star, \ldots, h_d^\star)$ of bounded variation, such that $h_i^\star$ minimizes the distance $E_{\beta_0,\Lambda_0}(\dot{l}_{\beta,i}(\beta_0, \Lambda_0) - \dot{l}_\Lambda(\beta_0, \Lambda_0)\,h(\cdot))^2$, for h varying in the space of functions of bounded variation on $[\sigma, \tau]$. Here, $\dot{l}_{\beta,i}(\beta_0, \Lambda_0) = x^{(i)}\,\Lambda(z)\,Q((\delta, z, x); \beta_0, \Lambda_0)$ is the ith component of the ordinary score function for β (and $x^{(i)}$ is the ith component of x). It is not difficult to see that $h_i^\star$ must satisfy $E\,[B_{\Lambda_0}(h)\,(\dot{l}_{\beta,i}(\beta_0, \Lambda_0) - B_{\Lambda_0}(h_i^\star))] = 0$, for all h. This simplifies to $E\,[Q^2((\Delta, Z, X); \beta_0, \Lambda_0)h(Z)\,[X^{(i)}\,\Lambda_0(Z) - h_i^\star(Z)]] = 0$. For this to be satisfied it suffices to have, $E\,[Q^2((\Delta, Z, X); \beta_0, \Lambda_0)\,[X^{(i)}\,\Lambda_0(Z) - h_i^\star(Z)]\,|\,Z] = 0$, whence

$$h_i^\star(Z) = \Lambda_0(Z)\,\frac{E(X^{(i)}\,Q^2((\Delta, Z, X); \beta_0, \Lambda_0)\,|\,Z)}{E(Q^2((\Delta, Z, X); \beta_0, \Lambda_0)\,|\,Z)}.$$

In vector notation we can therefore write

$$h^\star(Z) = \Lambda_0(Z)\,h^{\star\star}(Z) \equiv \Lambda_0(Z)\,\frac{E_{\beta_0,\Lambda_0}(X\,Q^2((\Delta, Z, X); \beta_0, \Lambda_0)\,|\,Z)}{E_{\beta_0,\Lambda_0}(Q^2((\Delta, Z, X); \beta_0, \Lambda_0)\,|\,Z)}. \tag{3.8}$$

The assumption (A.2) that $E(\mathrm{Var}(X \mid Z))$ is positive definite ensures that $\tilde{l}$, the efficient score function for β, is not identically zero, whence the efficient information $\tilde{I}_0 = \mathrm{Disp}(\tilde{l}) \equiv E_{\beta_0,\Lambda_0}(\tilde{l}\tilde{l}^{\mathrm{T}})$ is positive definite (note that $E_{\beta_0,\Lambda_0}(\tilde{l}) = 0$). This entails that the MLE of β will converge at $\sqrt{n}$ rate to β_0 and have an asymptotically normal distribution with a finite dispersion matrix.

Let $\tilde{\theta}_0 = e^{\theta_0}$. Now, consider the problem of testing $H_0 : \beta = \beta_0$ based on our data, but under the (true) constraint that $\Lambda(z_0) = \tilde{\theta}_0$. Thus, we define:

$$\mathrm{lrtbeta}_n^0 = 2\,\log\,\frac{\underset{\Lambda(z_0)=\tilde{\theta}_0}{\operatorname{argmax}}\; l_n(\beta, \Lambda)}{\underset{\beta=\beta_0,\Lambda(z_0)=\tilde{\theta}_0}{\operatorname{argmax}}\; l_n(\beta, \Lambda)}. \tag{3.9}$$

Thus,

$$\mathrm{lrtbeta}_n^0 = 2\,l_n(\hat{\beta}_{n,0}, \hat{\Lambda}_{n,0}) - 2\,l_n(\beta_0, \hat{\Lambda}_{n,0}^{(\beta_0)}),$$

where $\hat{\Lambda}_{n,0} = \exp(\hat{\psi}_{n,0})$ and $\hat{\Lambda}_{n,0}^{(\beta_0)} = \exp(\hat{\psi}_{n,0}^{(\beta_0)})$.

Please cite this article in press as: Banerjee, M., et al., Semiparametric binary regression models under shape constraints with an application to Indian schooling data. Journal of Econometrics (2009), doi:10.1016/j.jeconom.2008.11.002

3.1. Asymptotic results concerning the estimation of β

We now state a theorem describing the asymptotic behavior of $\hat{\beta}_n$ and $\hat{\beta}_{n,0}$ (which we subsequently denote by $\tilde{\beta}_n$) and the likelihood ratio statistics lrtbeta_n as defined in (2.4) and lrtbeta_n^0 above.

Theorem 3.1. *Under conditions* (A.1)–(A.7), *both* $\hat{\beta}_n$ *and* $\tilde{\beta}_n$ *are asymptotically linear in the efficient score function and have the following representation:*

$$\sqrt{n}\,(\hat{\beta}_n - \beta_0) = \frac{1}{\sqrt{n}}\,\tilde{I}_0^{-1} \sum_{i=1}^{n} \tilde{l}(\Delta_i, Z_i, X_i) + r_n \tag{3.10}$$

and

$$\sqrt{n}\,(\tilde{\beta}_n - \beta_0) = \frac{1}{\sqrt{n}}\,\tilde{I}_0^{-1} \sum_{i=1}^{n} \tilde{l}(\Delta_i, Z_i, X_i) + s_n \tag{3.11}$$

where r_n *and* s_n *are* $o_p(1)$. *Hence both* $\sqrt{n}\,(\hat{\beta}_n - \beta_0)$ *and* $\sqrt{n}\,(\tilde{\beta}_n - \beta_0)$ *converge in distribution to* $N(0, \tilde{I}_0^{-1})$.

Furthermore,

$$\text{lrtbeta}_n = n(\hat{\beta}_n - \beta_0)^{\mathrm{T}}\,\tilde{I}_0\,(\hat{\beta}_n - \beta_0) + o_p(1), \tag{3.12}$$

while

$$\text{lrtbeta}_n^0 = n(\tilde{\beta}_n - \beta_0)^{\mathrm{T}}\,\tilde{I}_0\,(\tilde{\beta}_n - \beta_0) + o_p(1). \tag{3.13}$$

It follows that both lrtbeta_n *and* lrtbeta_n^0 *are asymptotically distributed like* χ_d^2.

3.2. Asymptotic results concerning the estimation of ψ

To this end, we introduce the following processes. For positive constants c and d define the process $X_{c,d}(z) := c\,W(z) + d\,z^2$, where $W(z)$ is standard two-sided Brownian motion starting from 0. Let $G_{c,d}(z)$ denote the GCM of $X_{c,d}(z)$. Let $g_{c,d}(z)$ be the right derivative of $G_{c,d}$. This is a non-decreasing function that can be shown to be a piecewise constant, with finitely many jumps in any compact interval. Next, let $G_{c,d,L}(h)$ denote the GCM of $X_{c,d}(h)$ restricted to the set $h \le 0$ and $g_{c,d,L}(h)$ denote its right-derivative process. For $h > 0$, let $G_{c,d,R}(h)$ denote the GCM of $X_{c,d}(h)$ restricted to the set $h > 0$ and $g_{c,d,R}(h)$ denote its right-derivative process. Define $g_{c,d}^0(h) = (g_{c,d,L}(h) \wedge 0)\,1(h \le 0) + (g_{c,d,R}(h) \vee 0)\,1(h > 0)$. Then $g_{c,d}^0(h)$, like $g_{c,d}(h)$, is a non-decreasing function that is piecewise constant, with finitely many jumps in any compact interval and differs (almost surely) from $g_{c,d}(h)$ on a finite interval containing 0. In fact, with probability 1, $g_{c,d}^0(h)$ is identically 0 in some (random) neighborhood of 0, whereas $g_{c,d}(h)$ is almost surely non-zero in some (random) neighborhood of 0. Also, the interval $D_{c,d}$ on which $g_{c,d}$ and $g_{c,d}^0$ differ is $O_p(1)$. For more detailed descriptions of the processes $g_{c,d}$ and $g_{c,d}^0$, see Banerjee (2000), Banerjee and Wellner (2001) and Wellner (2003). Thus, $g_{1,1}$ and $g_{1,1}^0$ are the unconstrained and constrained versions of the slope processes associated with the "canonical" process $X_{1,1}(z)$. By Brownian scaling, the slope processes $g_{c,d}$ and $g_{c,d}^0$ can be related in distribution to the canonical slope processes $g_{1,1}$ and $g_{1,1}^0$. This is the content of the following proposition.

Lemma 3.1. *For any* $M > 0$, *the following distributional equality holds in the space* $L_2[-M, M] \times L_2[-M, M]$:

$$\left(g_{c,d}(h), g_{c,d}^0(h)\right) \stackrel{\mathcal{D}}{=} \left(c\,(d/c)^{1/3} g_{1,1}\left((d/c)^{2/3}h\right),\right.$$
$$\left. c\,(d/c)^{1/3} g_{1,1}^0\left((d/c)^{2/3}h\right)\right).$$

Here $L_2[-M, M]$ *denotes the space of real-valued functions on* $[-M, M]$ *with finite* L_2 *norm (with respect to Lebesgue measure).*

This is proved in Banerjee (2000), Chapter 3.

Let z_0 be an interior point of the support of Z. Define the (localized) slope processes U_n and V_n as follows:

$$U_n(h) = n^{1/3}\,(\hat{\psi}_n^{(\beta_0)}(z_0 + h\,n^{-1/3}) - \psi_0(z_0)) \quad \text{and}$$
$$V_n(h) = n^{1/3}\,(\hat{\psi}_{n,0}^{(\beta_0)}(z_0 + h\,n^{-1/3}) - \psi_0(z_0)).$$

The following theorem describes the limiting distribution of the slope processes above.

Theorem 3.2. *Define* $C(z_0) = \int \frac{e^{\beta_0^{\mathrm{T}} x + \psi_0(z_0)}}{(1 + e^{\beta_0^{\mathrm{T}} x + \psi_0(z_0)})^2} f(z_0, x)\,\mathrm{d}\,\mu(x)$. *Assume that* $0 < C(z_0) < \infty$. *Let* $a = \sqrt{\frac{1}{C(z_0)}}$ *and* $b = \frac{1}{2}\,\psi_0'(z_0)$, *where* ψ_0' *is the derivative of* ψ_0. *The processes* $(U_n(h), V_n(h))$ *converge finite dimensionally to the processes* $(g_{a,b}(h), g_{a,b}^0(h))$. *Furthermore, using the monotonicity of the processes* U_n *and* V_n, *it follows that the convergence holds in the space* $L_2[-K, K] \times L_2[-K, K]$ *for any* $K > 0$.

Setting $h = 0$ in the above theorem, we find that

$$n^{1/3}\,(\hat{\psi}_n^{(\beta_0)}(z_0) - \psi_0(z_0)) \to_d g_{a,b}(0) \equiv_d a\,(b/a)^{1/3}\,g_{1,1}(0)$$
$$\equiv_d (8\,a^2\,b)^{1/3}\,\mathbb{Z},$$

where $\mathbb{Z} \equiv \text{argmin}_{h \in \mathbb{R}}\,(W(h) + h^2)$ and its distribution is referred to in the statistical literature as Chernoff's distribution. See, for example, Groeneboom and Wellner (2001) for a detailed description. The above display utilizes the result that $g_{1,1}(0) \equiv_d 2\,\mathbb{Z}$ (since this result is not used in our proposed methodology for constructing confidence sets, discussed below, we do not establish this result in our paper). The random variable $\mathbb{Z}$ arises extensively in nonparametric problems involving *cube-root asymptotics* – problems where estimates of parameters converge at rate $n^{1/3}$ and in particular, is typically found to characterize the pointwise limit distribution of maximum likelihood estimators of monotone functions in nonparametric/semiparametric models. The distribution of $\mathbb{Z}$ is non-Gaussian and symmetric about 0. It can, in fact, be shown that

$$n^{1/3}\,(\hat{\psi}_n(z_0) - \psi_0(z_0)) \to_d (8\,a^2\,b)^{1/3}\,\mathbb{Z}$$

where $\hat{\psi}_n \equiv \hat{\psi}_n^{\hat{\beta}_n}$ is the unconstrained MLE of ψ. This is not surprising in view of the fact that $\sqrt{n}(\hat{\beta}_n - \beta_0) = O_p(1)$, so that $\hat{\beta}_n$ converges to β_0 at a faster rate than $n^{1/3}$, the convergence rate for $\hat{\psi}_n^{\hat{\beta}_n}$. Since the quantiles of $\mathbb{Z}$ are well tabulated, this result can be used to construct asymptotic confidence sets of any pre-assigned level for $\psi_0(z_0)$ (equivalently $\Lambda_0(z_0)$), but the procedure requires estimating the constants a and b which turns out to be a tricky affair (one needs to estimate the joint density of the covariates that appears in the defining integral for $C(z_0)$ in addition to the derivative of ψ_0 at the point z_0, which is quite difficult, especially at modest sample sizes). Resampling techniques, like subsampling (m out of n bootstrap without replacement) as discussed in Politis et al. (1999), can circumvent the estimation of the nuisance parameters a and b, but are computationally quite intensive. To avoid these difficulties, we do not construct MLE based confidence sets for $\psi_0(z_0)$ in this paper; rather, we resort to inversion of the likelihood ratio statistic for testing the value of ψ_0 at a pre-fixed point of interest. Our next theorem is crucial for this purpose.

Theorem 3.3. *The likelihood ratio statistic for testing* $\tilde{H}_0 : \psi(z_0) = \theta_0$, *as defined in* (2.6), *converges in distribution to* $\mathbb{D}$ *where*

$$\mathbb{D} = \int \left((g_{1,1}(z))^2 - (g_{1,1}^0(z))^2\right)\,\mathrm{d}z.$$

The random variable $\mathbb{D}$ can be considered to be a nonregular analogue of the usual χ_1^2 random variable, in the sense that just as

the χ^2_1 distribution describes the limiting likelihood ratio statistic for testing a real-valued parameter in a regular parametric model, similarly, the distribution of $\mathbb{D}$ describes the limiting likelihood ratio statistic for testing the value of a monotone function at a point in *conditionally parametric models* (see Banerjee (2007)) and more generally in pointwise estimation of monotone functions.

3.3. Construction of confidence sets for parameters of interest via likelihood ratio based inversion

Confidence sets for ψ: Denote the likelihood ratio statistic for testing the null hypothesis $\psi(z_0) = \theta$ by $\text{lrtpsi}_n(\theta)$. The computation of the likelihood ratio statistic is discussed, in detail, in Section 2. By Theorem 3.3, an approximate level $1-\alpha$ confidence set for $\psi_0(z_0)$ is given by $S_{\psi_0(z_0)} \equiv \{\theta : \text{lrtpsi}_n(\theta) \le q(\mathbb{D}, 1-\alpha)\}$, where $q(\mathbb{D}, 1-\alpha)$ is the $(1-\alpha)$th quantile of the distribution of $\mathbb{D}$ (for $\alpha = 0.05$, this is approximately 2.28). Noting that $\Lambda_0(z_0) = \exp(\psi_0(z_0))$, the corresponding confidence set for $\Lambda_0(z_0)$ is simply $\exp(S_{\psi_0(z_0)})$. Furthermore, the corresponding confidence set for the baseline conditional probability function, $E(\Delta \mid X = 0, Z = z_0)$ is simply $e^{S_{\psi_0(z_0)}}/(1 + e^{S_{\psi_0(z_0)}})$.

Confidence sets for μ: Confidence sets for the regression function at values $X = x_0, Z = z_0$, i.e. $\mu(x_0, z_0) = E(\Delta \mid X = x_0, Z = z_0)$ can also be constructed in a similar fashion. This requires redefining the covariate X, so as to convert $\mu(x_0, z_0)$ to a baseline conditional probability. Set $\tilde{X} = X - x_0$. Then $\mu(x_0, z_0) = P(\Delta = 1 \mid \tilde{X} = 0, Z = z)$. Define $\tilde{\mu}(\tilde{x}, z) = E(\Delta \mid \tilde{X} = \tilde{x}, Z = z)$. We have,

$$\tilde{\mu}(\tilde{x}, z) = \mu(\tilde{x} + x_0, z) = \frac{e^{\beta_0^T(\tilde{x}+x_0)}\,\Lambda_0(z)}{1 + e^{\beta_0^T(\tilde{x}+x_0)}\,\Lambda_0(z)} = \frac{e^{\beta_0^T\tilde{x}}\,\tilde{\Lambda}_0(z)}{1 + e^{\beta_0^T\tilde{x}}\,\tilde{\Lambda}_0(z)}$$

where $\tilde{\Lambda}_0(z) = e^{\beta_0^T x_0}\Lambda_0(z)$, with $\tilde{\psi}_0(z) \equiv \log \tilde{\Lambda}_0(z) = \beta_0^T x_0 + \psi_0(z)$. This is exactly the model considered at the beginning of Section 3 in terms of new covariates $(\tilde{X}, Z)$ and satisfies the regularity conditions A.1–A.6 (with X replaced by $\tilde{X}$). Now, $\mu(x_0, z_0) = \tilde{\mu}(0, z_0) = e^{\tilde{\psi}_0(z_0)}/(1 + e^{\tilde{\psi}_0(z_0)})$. An approximate level $1-\alpha$ confidence set for $\tilde{\psi}_0(z_0)$, say $\tilde{S}_{\tilde{\psi}_0(z_0)}$ can be found in exactly the same fashion as before; i.e. $\tilde{S}_{\tilde{\psi}_0(z_0)} = \{\theta : \widetilde{\text{lrtpsi}}_n(\theta) \le q(\mathbb{D}, 1-\alpha)\}$, where $\widetilde{\text{lrtpsi}}_n(\theta)$ is the likelihood ratio statistic for testing $\tilde{\psi}(z_0) = \theta$ and is computed in exactly the same way as the statistic in (2.6), but using the covariates $\tilde{X}$ and Z, instead of X and Z. Correspondingly, the confidence set for $\tilde{\mu}(0, z_0)$ is $e^{\tilde{S}_{\tilde{\psi}_0(z_0)}}/(1 + e^{\tilde{S}_{\tilde{\psi}_0(z_0)}})$. This principle is applied extensively to construct confidence sets for the conditional probabilities in the data analysis example in Section 4.

The construction of joint confidence sets is also of importance in certain applications. Thus, one may be interested in a joint confidence set for $(\mu(x_0, z_0), \mu(x_0, z_1))$ for $z_0 < z_1$. To this end consider the hypothesis $\tilde{H}_{0,1} : \psi(z_0) = \theta_0, \psi(z_1) = \theta_1$ where $z_0 < z_1$ and $\theta_0 < \theta_1$. A natural statistic to test this hypothesis is $M_n \equiv \max(\text{lrtpsi}_n^{(z_0)}(\theta_0), \text{lrtpsi}_n^{(z_1)}(\theta_1))$ where $\text{lrtpsi}_n^{(z_0)}(\theta_0)$ is the likelihood ratio statistic for testing $\psi(z_0) = \theta_0$ and $\text{lrtpsi}_n^{(z_1)}(\theta_1)$, the likelihood ratio statistic for testing $\psi(z_1) = \theta_1$. It can be shown that when the null hypothesis is true, $\text{lrtpsi}_n^{(z_0)}(\theta_0)$ and $\text{lrtpsi}_n^{(z_1)}(\theta_1)$ are asymptotically independent and M_n converges in distribution to $\mathbb{D}^{(2)} \equiv \max(\mathbb{D}_1, \mathbb{D}_2)$ where $\mathbb{D}_1$ and $\mathbb{D}_2$ are identical copies of $\mathbb{D}$. The quantiles of this distribution are well-tabulated and joint confidence sets for $(\psi(z_0), \psi(z_1))$ are therefore readily constructed by inversion. This leads to joint confidence sets for $(\mu(x_0, z_0), \mu(x_0, z_1))$ by centering X around x_0, as in the previous paragraph. Consider the pair (θ, θ') (with $\theta \le \theta'$) and let $\widetilde{\text{lrtpsi}}_n^{(z_0)}(\theta)$ and $\widetilde{\text{lrtpsi}}_n^{(z_1)}(\theta')$ denote, respectively, the likelihood ratio statistics for testing $\tilde{\psi}(z_0) = \theta$ and $\tilde{\psi}(z_1) = \theta'$, these being computed in exactly the same way as the statistic in (2.6) but using covariates $(\tilde{X}, Z)$ instead of (X, Z). Let:

$$\tilde{S}_{\tilde{\psi}_0(z_0),\tilde{\psi}_0(z_1)} = \{(\theta, \theta') : \theta \le \theta', \max(\widetilde{\text{lrtpsi}}_n^{(z_0)}(\theta), \widetilde{\text{lrtpsi}}_n^{(z_1)}(\theta')) \le q(\mathbb{D}^{(2)}, 1-\alpha)\}.$$

This set has a simple characterization as a polygon in $\mathbb{R}^2$ (it is either a triangle or a trapezium or a pentagon). Let $\bar{S}_{\tilde{\psi}_0(z_0)} = \{\theta : \widetilde{\text{lrtpsi}}_n^{(z_0)}(\theta) \le q(\mathbb{D}^{(2)}, 1-\alpha)\}$ and $\bar{S}_{\tilde{\psi}_0(z_1)} = \{\theta' : \widetilde{\text{lrtpsi}}_n^{(z_1)}(\theta') \le q(\mathbb{D}^{(2)}, 1-\alpha)\}$. Then:

$$\tilde{S}_{\tilde{\psi}_0(z_0),\tilde{\psi}_0(z_1)} = (\bar{S}_{\tilde{\psi}_0(z_0)} \times \bar{S}_{\tilde{\psi}_0(z_1)}) \cap \mathcal{C}_2$$

where $\mathcal{C}$ is the cone given by $\{(\theta, \theta') \in \mathbb{R}^2 : \theta \le \theta'\}$. A two-dimensional joint confidence set of level $1-\alpha$ for $(\mu(x_0, z_0), \mu(x_0, z_1))$ is given by:

$$\{(e^\theta/(1+e^\theta), e^{\theta'}/(1+e^{\theta'})) : (\theta, \theta') \in \tilde{S}_{\tilde{\psi}_0(z_0),\tilde{\psi}_0(z_1)}\}.$$

This method can be extended to provide confidence sets at more than 2 points. However, for a fixed sample size, the performance of this procedure will deteriorate as the number of z_is increases, owing to the finite sample dependence among the pointwise likelihood ratio statistics.

Confidence sets for β: Confidence sets for the finite-dimensional regression parameter β_0 can be constructed in the usual fashion as: $\{\beta : \text{lrtbeta}_n(\beta) \le q_{\chi^2_d, 1-\alpha}\}$, where $\text{lrtbeta}_n(\beta)$ is the likelihood ratio statistic for testing the null hypothesis that the true regression parameter is β (see (2.7)), and $q_{\chi^2_d,1-\alpha}$ is the $(1-\alpha)$th quantile of the χ^2_d distribution. This method can be adapted to construct confidence sets for a sub-vector of the regression parameters as well. So, consider a situation where the regression parameter vector can be partitioned as $\beta = (\eta_1, \eta_2)$. Let $\beta_0 = (\eta_{10}, \eta_{20})$ denote the true parameter value and suppose that we are interested in a confidence set for η_{10}. Let d_1 and d_2 denote the dimensions of η_1 and η_2 respectively. To test $\bar{H}_0 : \eta_1 = \eta_{10}$, the log-likelihood function $l_n(\beta, \psi)$ is maximized over all β of the form (η_{10}, η_2) (where η_2 varies freely in $\mathbb{R}^{d_2}$) and ψ monotone increasing. If we identify ψ, as before, with the vector $\mathbf{u} = \{\psi(Z_{(i)})\}_{i=1}^n$, then, $\bar{g}(\eta_2, \mathbf{u}) \equiv -l_n((\eta_{10}, \eta_2), \psi)$ is a continuously differentiable strictly convex function defined on $\mathbb{R}^{d_2} \times \mathcal{C}$ and its minimizer can be obtained using Proposition 1. If $(\hat{\eta}_2, \hat{\mathbf{u}}_n^{(\eta_{10})})$ denotes the minimizer of $\bar{g}$, then the constrained MLEs of (β, ψ) under $\bar{H}_0$ are: $((\eta_{10}, \hat{\eta}_2), \hat{\psi}_n^{(\eta_{10})})$ where $\hat{\psi}_n^{(\eta_{10})}$ is the (unique) right-continuous increasing step function that assumes the value $\hat{u}_{i,n}^{(\eta_{10})}$ (the ith component of $\hat{\mathbf{u}}_n^{(\eta_{10})}$) at the point $Z_{(i)}$ and has no jump points outside of the set $\{Z_{(i)}\}_{i=1}^n$. The likelihood ratio statistic for testing $\bar{H}_0$ is then given by:

$$2\,[l_n(\hat{\beta}_n, \hat{\psi}_n) - l_n((\eta_{10}, \hat{\eta}_2), \hat{\psi}_n^{(\eta_{10})})]$$

and converges to the $\chi^2_{d_1}$ distribution. Therefore, a level $1-\alpha$ confidence set for the sub-vector η_{10} can be readily computed via inversion and calibration using $\chi^2_{d_1}$ quantiles. We skip the details.

3.4. General link functions

Under regularity conditions analogous to those described at the beginning of this section, similar results are obtained for more general link functions, so long as the inverse link $\tilde{h}$ satisfies Condition (C) described in Section 2. Thus, for any $\tilde{h}$ satisfying the concavity constraints, (a) the likelihood ratio statistic for testing

Please cite this article in press as: Banerjee, M., et al., Semiparametric binary regression models under shape constraints with an application to Indian schooling data. Journal of Econometrics (2009), doi:10.1016/j.jeconom.2008.11.002

$\beta = \beta_0$ as described in (2.4) converges to a χ^2_d distribution *and* (b) the likelihood ratio statistic for testing $\tilde{H}_0 : \psi(z_0) = \theta_0$ as described in (2.6) converges in distribution to $\mathbb{D}$ (when $\tilde{H}_0$ is true). Confidence sets for β, $\psi(z_0)$ and $\mu(x_0, z_0)$ as well as a subvector of β may be obtained by methods analogous to those used in the logistic regression framework. Once again, owing to space constraints, we skip the details.

4. Simulations and data analysis

Simulation studies: We first present results from a simulation study that illustrates the performance of the likelihood ratio based methods for estimation of β_0 and pointwise estimation of ψ_0 (equivalently Λ_0) at one and two points. Two different settings were considered. The first corresponds to a semiparametric logistic regression model with independent covariates Z and X (both real-valued); the distribution of X is truncated normal on the interval $[-2, 2]$ while Z is independent of X and follows an exponential (1) distribution truncated between 0.5 and 2.5. In the second setting, the covariates Z and X are dependent. The distribution of X remains the same as before and that of Z given X is an exponential with rate $1 + X/8$. Five different choices of β_0 (displayed in the table) were considered and $\Lambda_0(z)$ was taken to be $z/2$. For each combination of parameter values (five in all) 2000 data replicates were generated for $n = 500$ and 1000. For each n, (i) the empirical coverage and average length of the 95% likelihood ratio based confidence intervals for β_0 (ii) the empirical coverage and average length of the 95% likelihood ratio confidence intervals for $\Lambda_0(z_0)$ (with $z_0 = 1.0$), and (iii) the empirical coverage and average volume of the 95% likelihood ratio based joint confidence sets for $(\Lambda_0(z_0), \Lambda_0(z_1))$ (with $z_1 = 2.0$), were recorded (based on these 2000 replicates). The numbers are presented in tables.

The empirical coverage is seen to be fairly close to the nominal coverage. The lengths (volumes) of the confidence intervals (sets) for any particular setting of parameter values (β_0, Λ_0) dwindle with increasing sample size, as expected. For a fixed sample size, the average lengths (volumes) of the C.I.'s (confidence sets) do not show much variability as β_0 varies, especially in the case where Z and X are independent. For each sample size and at every given setting of parameter values, the confidence intervals for β are somewhat larger, on an average, in the second setting (Z and X dependent).

Coverage and average lengths/volumes of confidence intervals/sets based on simulation studies: Z and X independent

β	$z_0 = 1.0$ and $z_1 = 2.0$					
	$n = 500$			$n = 1000$		
	CI(β)	CI($\Lambda(z_0)$)	CS($\Lambda(z_0)$, $\Lambda(z_1)$)	CI(β)	CI($\Lambda(z_0)$)	CS($\Lambda(z_0)$, $\Lambda(z_1)$)
−0.5	94.8 0.688	95.9 0.454	94.6 0.318	94.5 0.523	95.1 0.314	93.4 0.189
−0.25	95.5 0.684	94.4 0.444	94.1 0.311	94.0 0.523	95.4 0.310	94.5 0.185
0.0	95.6 0.686	95.2 0.439	93.5 0.312	94.0 0.522	95.6 0.309	95.2 0.182
0.25	95.0 0.682	94.9 0.443	94.1 0.310	94.1 0.528	94.6 0.319	95.1 0.190
0.5	95.0 0.691	94.5 0.454	94.9 0.332	95.0 0.530	94.7 0.316	95.0 0.187

Coverage and average lengths/volumes of confidence intervals/sets based on simulation studies: Z and X dependent

β	$z_0 = 1.0$ and $z_1 = 2.0$					
	$n = 500$			$n = 1000$		
	CI(β)	CI($\Lambda(z_0)$)	CS($\Lambda(z_0)$, $\Lambda(z_1)$)	CI(β)	CI($\Lambda(z_0)$)	CS($\Lambda(z_0)$, $\Lambda(z_1)$)
−0.5	94.6 0.716	95.1 0.476	94.6 0.311	95.4 0.541	94.4 0.319	94.4 0.180
−0.25	94.0 0.713	96.2 0.463	92.8 0.314	95.3 0.539	94.9 0.311	93.5 0.182
0.0	94.0 0.774	96.2 0.455	92.8 0.298	95.4 0.527	95.0 0.307	94.6 0.189
0.25	94.4 0.778	94.2 0.461	93.9 0.308	95.0 0.543	95.0 0.315	94.0 0.185
0.5	95.0 0.753	94.9 0.471	94.3 0.315	94.9 0.551	94.7 0.319	92.9 0.192

Analysis of the schooling data: In this section we provide a simple example where the proposed methodology can be applied. We consider the issue of primary school attendance in Indian villages. However, the purpose of this section is not to model the determinants of schooling rigorously but to pinpoint a few important determinants of it and then to model the probability of school attendance by using our proposed methodology. It is well known that education is a pivotal contributing factor for the development of any society. At the time of the 1991 census, only about 52% of the Indian population was literate, the corresponding numbers for male and female being 65.5% and 39% respectively. From 1991 onwards there has been an urgency on the part of Indian Government to remedy the appalling primary school attendance pattern throughout the country. Total Literacy campaign (TLC) and mid-day meals are prominent components of this remedial package. The country as a whole has improved in this regard (according to the 2001 census, the overall literacy rate has increased to 65.4%) but there still exists a significant gulf between male and female literacy and lower caste and upper caste literacy. Moreover it is common knowledge that BIMARU states, (Bihar, Madhya Pradesh, Rajasthan, and Uttar Pradesh), considered to be relatively backward states with poor social indicators, have traditionally lagged behind other states in development-indicators (percentage of the populace below poverty line, infant mortality rate, primary school enrollment, to name a few).

In our present analysis we select a few important determinants of school attendance for the rural population in India. There is a clear consensus that poverty is one of the greatest impediments to school attendance. The existing literature finds that school attendance increases significantly with an increase in standard of living or income. Apart from this primary determinant, another important determinant of school attendance is parental education which is a surrogate for parental attitude/motivation towards education. Dreze and Kingdon (2001), find a strong intergenerational effect. Handa (2002) and Lavy (1996) have also found parental literacy to be significant in the context of other developing countries. Dreze and Kingdon (2001) point out that children from scheduled class and scheduled tribe (SCST) households (to be defined later) have an "intrinsic disadvantage" in terms of going to primary school. Using the PROBE survey (which contains data for five states only, including four Bimaru states) they have also considered the effect of school quality index and village quality index on primary school enrollment. Here, we focus on a broader nation wide survey data and hence are unable to consider the above two indices.[7] However, we look into the effects

[7] In the nation wide survey data there is no information on variables needed to construct such indices.

Please cite this article in press as: Banerjee, M., et al., Semiparametric binary regression models under shape constraints with an application to Indian schooling data. Journal of Econometrics (2009), doi:10.1016/j.jeconom.2008.11.002

of mid-day meal program and total literacy campaign or TLC (to be described later), in order to investigate the role of government policies.

Our empirical analysis is based on a nationwide rural household survey collected by National Sample Survey Organization (NSSO), India in its 52nd round on "Participation in Education" (conducted between July 1995 and June 1996). For detailed information regarding this survey, we direct the refer to the following URL site: http://mospi.nic.in/mospi_nsso_data.htm.

To our knowledge, this is the most recent nationwide survey, on primary school participation, by the government of India that is available on the public domain.

Although a large number of variables have been considered in the literature, none of the studies seem to have focused on the possible nonlinearity in the relationship between school attendance and household standard of living (a primary determinant of school attendance). Misspecification of such functional forms may lead to biased estimates and incorrect inference. To address this issue, we model the effect of household standard of living on primary school enrollment nonparametrically; more specifically, the effect of standard of living on the probability of school attendance is captured by a monotone increasing function Λ, as described in (3.7). The monotonicity assumption on Λ is justified by the positive impact of standard of living on probability of school attendance, an established observation in the literature.

We proxy standard of living by annual household consumption expenditure on light durables (excluding cost of schooling and food expenses). Consumption expenditure is less volatile than income, and it is the nonfood portion of the expenditure (mainly on durable goods) that measures the income available to the household over and above subsistence needs. Dreze and Kingdon (2001) have used a similar proxy for standard of living. They considered asset, measured by durable goods. Lavy (1996) and Handa (2002) have also used an expenditure variable to proxy for standard of living for some other developing countries.[8] As mentioned before, we consider several control covariates like HEADLIT (a dummy variable indicating whether the household head is literate or not), that proxies for parental attitude towards education, cost of schooling (including tuition, fees, travel expenses, books and supplies, school uniform), SCST[9] (a dummy variable describing whether the child belongs to the schedule caste or schedule tribes population), mid-day meal (a dummy variable that takes the value 1 if the nearest school provides free lunch, and 0 otherwise) and TLC (a dummy equal to one if the village has been covered by literacy campaign, a measure taken by the Government of India to eliminate illiteracy.

Children within the age group 5–12 are considered. The proportions of male and female children are 54% and 46% respectively. We segment our sample into four groups: (i) male children from NonBIMARU states, (ii) female children from NonBIMARU states, (iii) male children from BIMARU states, and (iv) female children from BIMARU states. The school enrollment statistics for these four groups are 86.01%, 75.34%, 75.4%, and 51.5% respectively. The sample size for these four groups are 280, 280, 276 and 290 respectively.[10] Our response variable is whether a child is attending primary school or not ($\Delta = 1$ if attending, 0 otherwise).

Model-fitting: First, for a comparison, we fit a simple parametric logistic model based analysis to all four groups. Table 1A summarizes the conclusion obtained from the parametric logistic model. From the p-values, we find that standard of living has a positive and statistically significant impact on probability of enrollment for all four groups. The cost variable is statistically insignificant for all four groups (at 10%). It is possible that higher cost can also capture some information about the quality of education. So even though for a given quality of education, a rise in cost should lead to a lower probability of enrollment, it is possible that higher cost associated with better quality of education may attract more students to school. So, for a meaningful analysis, it may be useful to construct a cost index adjusting for quality of education. Due to the lack of data we are not able to develop such a measure. We also find that the SCST dummy has a negative and statistically significant coefficient. This indicates that, ceteris paribus, children belonging to the lower caste (SCST) have a lower probability of going to school. This is expected from our intuition regarding the state of the lower caste populace in India. Head literacy (HEADLIT) has a strong positive impact on school attendance and this is in accordance with results obtained from other studies. The role of both TLC and mid-day meal is much weaker than what we expected. Midday-meal is not statistically significant for any of the four groups (corruption and poor implementation are possible contributing factors for this result). TLC is statistically insignificant for all groups but for NonBIMARU male children. However, policy measures might have only long run impacts (rather than short run effects) which will not be reflected in a cross sectional framework; a panel data study may be more informative for this purpose.

We now relax the linearity assumption for the log-odds and fit the model given in (3.7). First we include all the variables used in the parametric model indicated above. The argument of the $\Lambda(.)$ function is the logarithm of consumption expenditure on light durables and the rest of the variables constitute the linear part of the log-odds expression. We get similar conclusions as far as the sign and the significance of the control covariates are concerned, i.e., the impacts of cost and midday-meal are statistically insignificant (at 10% level) and HEADLIT and SCST are statistically significant. TLC is significant only for NonBIMARU male children. However, in the semiparametric framework, the inclusion of TLC drops the significance of HEADLIT and vice versa.[11] We exclude the statistically insignificant (at 10% level) elements of X variables from the final model, namely midday-meal and cost. We also exclude TLC because it is not uniformly significant for all groups.[12] Thus, our modified model includes only standard of living, SCST and HEADLIT. Table 1B reports the results for the modified parametric model.

For the modified semiparametric model, the coefficients of SCST are −0.69, −1.29, −0.68, and −0.73 for BIMARU female, BIMARU male, NonBIMARU female and NonBIMARU male respectively. The corresponding p-values are 0.08, 0.001, 0.11 and 0.10 respectively. The coefficients of HEADLIT are 0.97, 0.78, 1.19, and 0.75 for BIMARU female, BIMARU male, NonBIMARU female and NonBIMARU male respectively. The corresponding p-values are 0.007, 0.05, 0.005 and 0.09 respectively.

Calculating probabilities of school attendance: We next focus our attention on estimating the probabilities of school attendance in the four different groups. This is achieved in Tables 2A–2C. Table 2A reports the estimated probabilities of school attendance and their likelihood-ratio based confidence intervals associated

[8] We have not considered an overall wealth variable to proxy for standard of living because of the nonavailability of such data at a nationwide level.

[9] Schedule Caste and Schedule Tribes are the most backward class and they have been oppressed socially and economically for more than few hundred years. Only after 1947 have measures been taken by the government of India to improve their socio-economic situation. But they are still considered to be in an economically and socially disadvantageous position.

[10] We consider an equal representation subsample from the bigger sample.

[11] For brevity we do not report these results.

[12] The HEADLIT variable partly captures its effect.

Please cite this article in press as: Banerjee, M., et al., Semiparametric binary regression models under shape constraints with an application to Indian schooling data. Journal of Econometrics (2009), doi:10.1016/j.jeconom.2008.11.002

Table 1A
Parametric results.

Variables	Group			
	Bimaru female	Bimaru male	NonBimaru female	NonBimaru male
Constant	−3.160610	−2.526723	−6.048611	−4.426694
	(0.0946)	(0.2319)	(0.0068)	(0.0147)
Standard of living	0.408983	0.504875	0.635662	0.567153
	(0.0733)	(0.0591)	(0.0160)	(0.0142)
SCST dummy	−0.833532	−1.514270	−0.537154	−0.648857
	(0.0044)	(0.0000)	(0.0801)	(0.0409)
Head literacy	1.010064	0.779417	1.245977	0.686764
	(0.0001)	(0.0071)	(0.0000)	(0.0298)
Cost of schooling	−0.069846	−0.150194	0.250532	0.166638
	(0.7004)	(0.4060)	(0.1779)	(0.2669)
TLC	0.049153	0.187457	0.378737	0.628891
	(0.8542)	(0.5386)	(0.2243)	(0.0438)
Midday-meal	0.460717	0.536007	0.266945	0.061193
	(0.1633)	(0.1329)	(0.3971)	(0.8435)

Note: p-values are in the parentheses.

Table 1B
Parametric results with significant variables.

Variables	Group			
	Bimaru female	Bimaru male	NonBimaru female	NonBimaru male
Constant	−3.459782	−2.686465	−4.501767	−3.499595
	0.0518	0.1824	0.0256	0.0444
Standard of living	0.411713	0.446572	0.659107	0.601905
	0.0633	0.0835	0.0096	0.0061
SCST dummy	−0.743829	−1.410061	−0.665219	−0.781412
	0.0084	0.0000	0.0235	0.0110
Head literacy	1.018684	0.780840	1.208981	0.784424
	0.0000	0.0066	0.0001	0.0113

Note: p-values are in the parentheses. TLC is significant in only one case and in our nonparametric model it is not significant if head literacy is included.

with all four groups (Bimaru female, Bimaru male, NonBimaru female, NonBimaru male) for each different category (SCST non-Lit (SCST = 1, HEADLIT = 0), SCST Lit (SCST = 1, HEADLIT = 1), Non-SCST non-Lit (SCST = 0, HEADLIT = 0), Non-SCST Lit (SCST = 0, HEADLIT = 1)) at the first quartile of expenditure (standard of living). Similar results are reported in Tables 2B and 2C at the median and the third quartile of expenditure respectively.

Since the estimation of these probabilities and *the construction of confidence intervals for the same* are the main contributions of the paper, we describe explicitly how we build the likelihood ratio confidence intervals in the context of our application. For the sake of concreteness, fix attention to the cell in Table 2B that corresponds to Bimaru Female Category 1. This comprises all female children in the BIMARU states who belong to the SCST class and come from a household where the household-head is non-literate. For these individuals, the value of $X = (X_1, X_2)$ (where X_1 is the indicator of belonging to the SCST group and X_2 is the indicator of literacy of the household-head) is fixed at $x_0 = (1, 0)$, and the value of the the primary covariate Z (standard of living variable) is fixed at z_0, its median value in the Bimaru Female group. With $\mu(x, z) = P(\Delta = 1 \mid X = x, Z = z)$, the goal is to find $\mu(x_0, z_0)$. The point-estimate of this quantity using the semiparametric maximum likelihood procedure described in Section 2 is reported as .3503. To find a 95% confidence interval, we employ the likelihood ratio based procedure as described in Section 3.3. In accordance with that discussion, we set $\tilde{X} \equiv (\tilde{X}_1, \tilde{X}_2) = X - x_0 = (X_1 - 1, X_2 - 0)$.

We now regress Δ on $(\tilde{X}, Z)$ for the Bimaru Female group using the semiparametric logistic model

$$\log \frac{\tilde{\mu}(\tilde{X}, Z)}{1 - \tilde{\mu}(\tilde{X}, Z)} = \tilde{\beta}^{\mathrm{T}} \tilde{X} + \tilde{\psi}(Z)$$

where $\tilde{\psi}$ is monotone increasing and $\tilde{\mu}(\tilde{X}, Z) = P(\Delta = 1 \mid \tilde{X}, Z)$. Once again, by the discussion in Section 3.3, the quantity of interest $\mu(x_0, z_0) = \tilde{\mu}((0, 0), z_0) = e^{\tilde{\psi}(z_0)}/(1 + e^{\tilde{\psi}(z_0)})$, and a 95% C.I. for $\tilde{\psi}(z_0)$, say $[L_n, U_n]$ translates readily into a C.I. for $\mu(x_0, z_0)$ at the same level of confidence: $[e^{L_n}/(1 + e^{L_n}), e^{U_n}/(1 + e^{U_n})]$. The C.I. for $\tilde{\psi}(z_0)$ is obtained by inverting the likelihood ratio statistic for testing the value of $\tilde{\psi}(z_0)$ using the 95th percentile of its parameter-free limit distribution under the null hypothesis as a critical value. From Theorem 3.3, this is the 95th percentile of the distribution of $\mathbb{D}$ (which is approximately 2.29).[13] More specifically, this is how the inversion is done: For a number θ, denote the likelihood ratio statistic for testing $\tilde{H}_{0,\theta} : \tilde{\psi}(z_0) = \theta$ by LRT(θ). Our goal is to find the set $\{\theta : \mathrm{LRT}(\theta) \leq 2.29\}$. The function LRT($\theta$) is a bowl-shaped (convex) function with a unique minimum; thus the set of values of θ where this function lies below 2.29 is an interval, say $[L_n, U_n]$. Noting that each of L_n, U_n solves $\mathrm{LRT}(\theta) - 2.29 = 0$, these are obtained via the method of bisection (see, for example, Burden and Faires (2000)). The computation of LRT(θ) at different points, which is required to employ the bisection method, is accomplished using the procedure described in Section 2.2, replacing θ_0 there, throughout, by θ.

In a similar fashion, we obtain confidence intervals for the other cells in the table and in other tables: for example, if $x_0 = (1, 1)$, i.e. we focus attention on the case of SCST households with a literate household head, then the corresponding $\tilde{X} = (X_1 - 1, X_2 - 1)$, while with $x_0 = (0, 0)$, i.e. Non-SCST households with non-literate household head, $\tilde{X} = X$. The value of z_0, of course, is different for each of the three tables: it is the first quartile of the standard of living variable for 2A, the median for 2B, and the third quartile for 2C. Note also that though X is binary in this particular application, our procedure for constructing likelihood ratio based

[13] See, for example, Banerjee and Wellner (2005) for tabulated values of the quantiles of $\mathbb{D}$ which are more accurate numerical approximations than the initial values provided in Banerjee and Wellner (2001).

Please cite this article in press as: Banerjee, M., et al., Semiparametric binary regression models under shape constraints with an application to Indian schooling data. Journal of Econometrics (2009), doi:10.1016/j.jeconom.2008.11.002

Table 2A
Nonparametric results for probabilities and their confidence interval at the first quartile.

Groups	Category 1 SCST non-lit	Category 2 SCST lit	Category 3 Non-SCST non-lit	Category 4 Non-SCST lit
Bimaru female	0.2137 (0.1377, 0.3149)	0.4173 (0.2927, 0.5556)	0.3514 (0.2537, 0.4645)	0.5881 (0.4758, 0.6968)
Bimaru male	0.2290 (0.1413, 0.3512)	0.3936 (0.2626, 0.5418)	0.5308 (0.3931, 0.6733)	0.7118 (0.5785, 0.8182)
NonBimaru female	0.4900 (0.3453, 0.6169)	0.7608 (0.6357, 0.8528)	0.6541 (0.5093, 0.7466)	0.8622 (0.7745, 0.9175)
NonBimaru male	0.5929 (0.4442, 0.7064)	0.7550 (0.6284, 0.8479)	0.7511 (0.6235, 0.8435)	0.8646 (0.7780, 0.9185)

Table 2B
Nonparametric results for probabilities and their confidence interval at the median.

Groups	Category 1 SCST non-lit	Category 2 SCST lit	Category 3 Non-SCST non-lit	Category 4 Non-SCST lit
Bimaru female	0.3503 (0.2283, 0.4780)	0.5868 (0.4381, 0.7118)	0.5180 (0.3710, 0.6183)	0.7390 (0.6084, 0.8087)
Bimaru male	0.4149 (0.3065, 0.5300)	0.6077 (0.4707, 0.7315)	0.7298 (0.6443, 0.8039)	0.8550 (0.7891, 0.9060)
NonBimaru female	0.4900 (0.3651, 0.6241)	0.7608 (0.6357, 0.8528)	0.6541 (0.5513, 0.7567)	0.8622 (0.7846, 0.9180)
NonBimaru male	0.5929 (0.4688, 0.7107)	0.7550 (0.6363, 0.8488)	0.7511 (0.6301, 0.8447)	0.8646 (0.7878, 0.9194)

Table 2C
Nonparametric results for probabilities and their confidence interval at the third quartile.

Groups	Category 1 SCST non-lit	Category 2 SCST lit	Category 3 Non-SCST non-lit	Category 4 Non-SCST lit
Bimaru female	0.3581 (0.2435, 0.5025)	0.5951 (0.4581, 0.7281)	0.5265 (0.4214, 0.6450)	0.7455 (0.6643, 0.8279)
Bimaru male	0.4150 (0.3065, 0.5331)	0.6077 (0.4707, 0.7317)	0.7298 (0.6443, 0.8074)	0.8550 (0.7891, 0.9066)
NonBimaru female	0.5422 (0.3940, 0.6834)	0.7967 (0.6827, 0.8772)	0.6999 (0.5726, 0.8095)	0.8853 (0.8090, 0.9336)
NonBimaru male	0.7981 (0.6846, 0.8781)	0.8932 (0.8212, 0.9384)	0.8912 (0.8181, 0.9372)	0.9455 (0.9050, 0.9693)

confidence sets for the probabilities applies to any discrete or continuous X, as is clear from the relevant part of Section 3.3 (construction of confidence intervals for μ).

Some interesting observations can be made jointly from Tables 2A–2C. These are:

(i) The literate (SCST/non-SCST) populace always dominates the nonliterate (SCST/non-SCST) populace for each individual group (as expected).
(ii) The non-SCST (literate/non-literate) populace always dominates the SCST (literate/non-literate) populace for each individual group (as expected).
(iii) NonBimaru females always significantly dominate Bimaru females for each category.
(iv) NonBimaru males almost always dominate Bimaru males for each category.
(v) NonBimaru males dominate NonBimaru females for both nonliterate categories. This phenomenon is however not that pronounced for the two literate categories.
(vi) Bimaru males dominate Bimaru females except for the SCST literate category.

SCST literate seems to be an interesting category where the distinction between male and female school enrollment probabilities are much closer as compared to other categories. This merits further investigation. Investigation across different income quartiles depict that when income increase leads to a jump in school enrollment probabilities, the jump is almost always more pronounced for male children (in comparison to their female counterparts). In general, we see that Bimaru states and females are significantly lagging behind, especially when the head of the household is not literate. Plots of the conditional probability of school attendance for varying expenditure/standard of living (parametric as well as semiparametric) are presented in Figs. 1 and 2 for the worst (SCST non-Lit) and the best (Non-SCST Lit) case scenarios, and are compatible with the above conclusions. Moreover, it is evident from the plots that in some cases the parametric fit is somewhat biased. This also justifies our use of a nonparametric approach for capturing the dependence of the enrollment probability on the standard of living, for these datasets.

Since 1991, the government of India has taken measures which are especially oriented towards BIMARU states and female children. However, our analysis based on the nationwide survey conducted during 1995–1996 reflects that Bimaru states and female children are still trailing as far as school-enrollment is concerned (for most of the categories) and the discrepancy with better-off groups is not negligible either. Even plain census figures on literacy for the year 2001 attests to this observation. This suggests the need for vigorous educational measures for the socially underprevilaged population, especially for female children from Bimaru states.

5. Concluding discussion

In this paper, we have studied semiparametric binary regression models under monotonicity constraints with emphasis on logistic regression and have illustrated our method with an application to a study of patterns of primary school attendance in rural India. The effect of the main covariate of interest (standard of living surrogated by expenditure on consumer durables) on the

Please cite this article in press as: Banerjee, M., et al., Semiparametric binary regression models under shape constraints with an application to Indian schooling data. Journal of Econometrics (2009), doi:10.1016/j.jeconom.2008.11.002

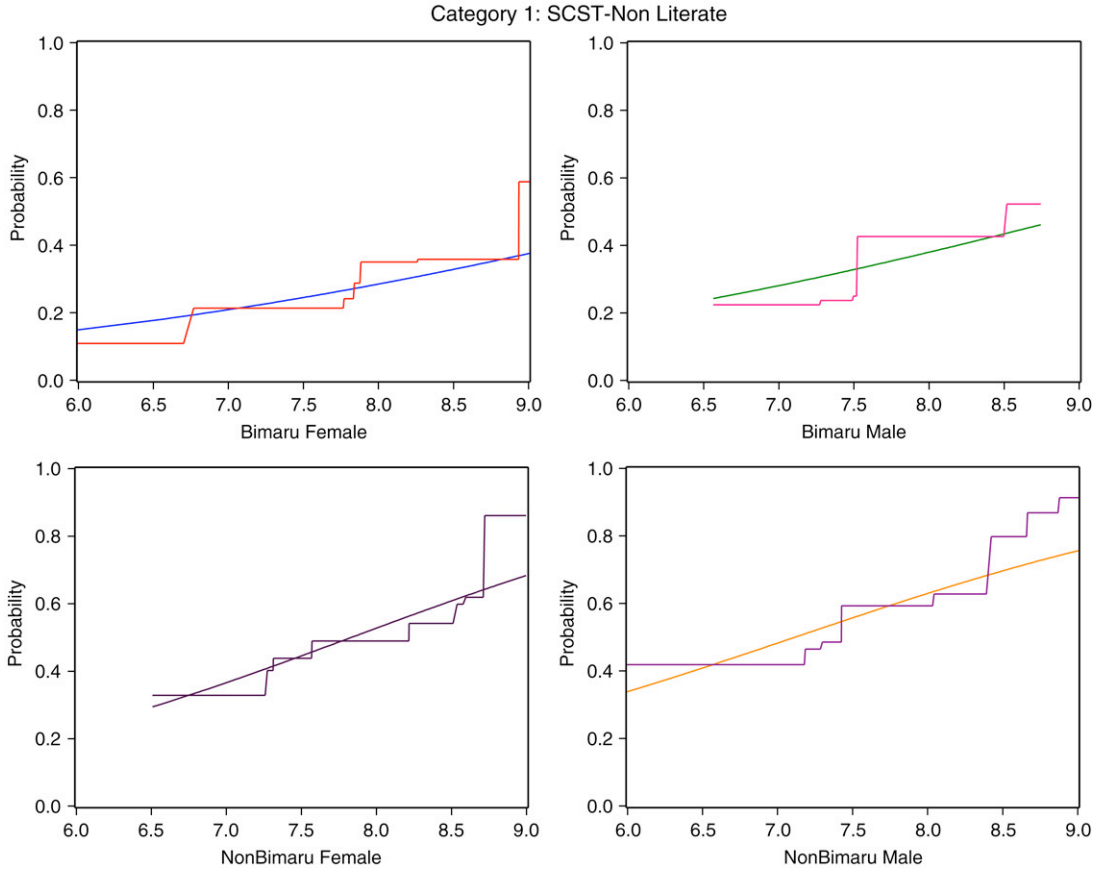

Fig. 1. Parametric vs. semiparametric baseline probability plots against standard of living.

response (the indicator of whether a child attends school) is specified through a monotone increasing function, while the effect of auxiliary covariates (caste affiliation, literacy status of the family, government policies) is captured by a finite-dimensional regression parameter. Our results indicate that (despite the popular measures taken by the government) school attendance rates, generally, continue to stay much lower for backward states and for backward castes, in comparison to relatively advanced states and higher castes. The female child, generally, seems much less likely to make it to school, as compared to the male child.

Our method extends the binary regression model used for parametric data analysis to the semiparametric domain under a monotone covariate effect. The use of likelihood ratios for estimating both the finite and infinite-dimensional components of the model and especially the regression function proves advantageous, since nuisance parameters need no longer be estimated. Some issues remain. Firstly, the likelihood (and likelihood ratio) based approach uses step estimates of the underlying monotone function Λ. However, since the true function is smooth, it is conceivable that a smooth monotone estimate of Λ may lead to better finite sample inference than the likelihood based method. Such smoothness constraints are typically imposed through penalized likelihood or penalized least squares criteria. This seems to be a direction for further research though smoothing, of course, would bring in the issue of choice of a smoothing parameter. Furthermore, it is unclear, whether smoothing under monotonicity constraints in semiparametric models would lead to asymptotic pivots (as happens with the likelihood ratio), though recent results of Pal and Banerjee (2008) in a purely nonparametric context suggests that this could be the case. Secondly, while the likelihood ratio method has natural advantages as illustrated in this paper, one problem with implementing it to construct confidence sets for the regression parameter β (especially in higher dimensions) is the "inversion" itself. For one-dimensional β, the convexity of the log-likelihood ratio in β dictates that the confidence set is an interval and a bisection method can be resorted to. For higher dimensions however, determining the level sets of the likelihood ratio can be a tricky affair. Grid search is not really an option in higher dimensions. Apart from prohibitive computational complexity, the grid-search method only gives us a grid-based approximation to the true convex set and the possibility of obtaining better approximations to the true set through advanced computational techniques suggests itself. Such techniques, if developed fairly generically, would be useful for obtaining likelihood ratio based confidence sets in a wide variety of semiparametric problems.

In conclusion, we briefly draw the reader's attention to a number of interesting questions that still need to be resolved and can provide future research directions in this area.

Firstly, the methods developed in this paper assume a one-dimensional covariate Z (in addition to the auxiliary covariates X) whose effect is modelled nonparametrically. From the point of view of applications it would be natural to incorporate multiple covariates with monotone effects. The simplest extension, for example, would be a model of the type: $\tilde{g}(\mu(X_1, X_2, Z)) = \beta^{\mathrm{T}} X + \psi_1(Z_1) + \psi_2(Z_2)$ where ψ_1 and ψ_2 are both monotone. This can be thought of as a generalized additive (semiparametric) model under a monotonicity constraint. While the MLEs of the parameters (β, ψ_1, ψ_2) can be computed using (iterative) convex optimization techniques, very little is known about the asymptotic behavior of the parameters at this point, though some recent work by Mammen and Yu (2007) on additive isotonic regression suggests certain possibilities. However a treatment of these additive models is outside the scope of this paper.

Please cite this article in press as: Banerjee, M., et al., Semiparametric binary regression models under shape constraints with an application to Indian schooling data. Journal of Econometrics (2009), doi:10.1016/j.jeconom.2008.11.002

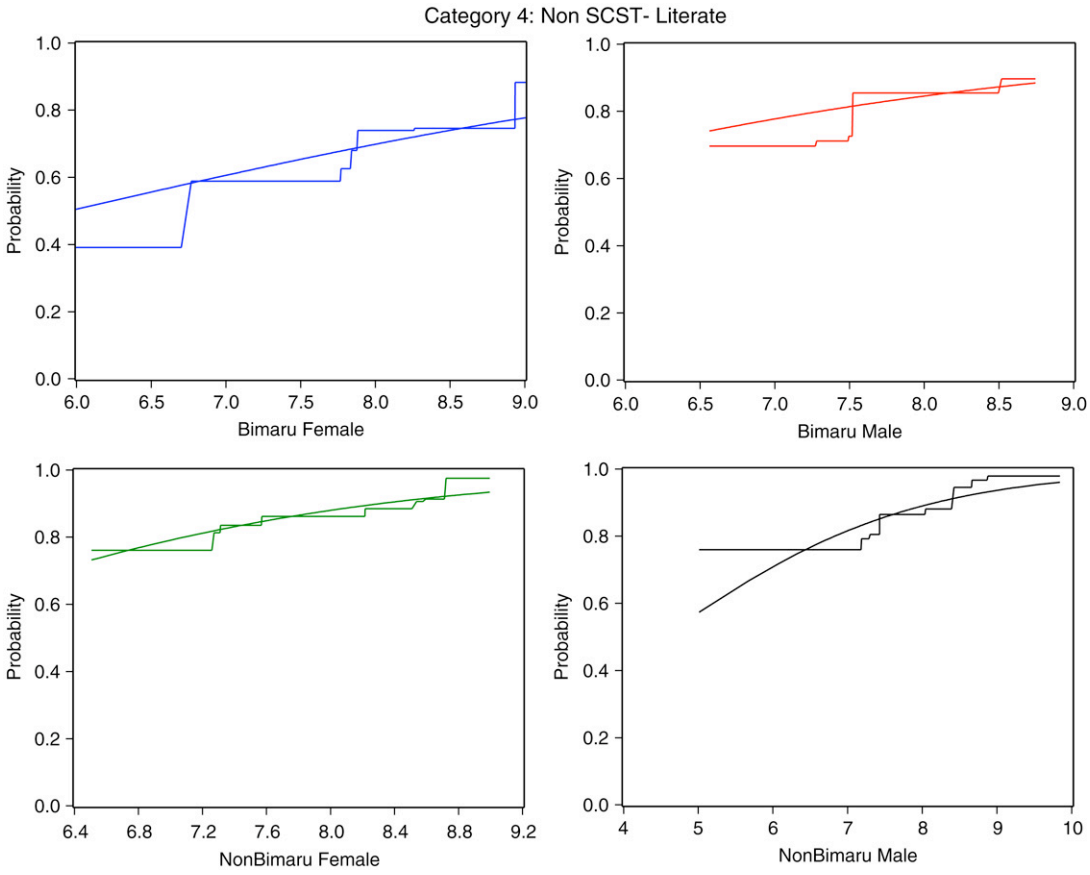

Fig. 2. Parametric vs. semiparametric baseline probability plots against standard of living.

Secondly, it is natural to ask whether methods similar to this paper can be developed for other kinds of shape constraints, in particular, constraints beyond monotonicity. A natural assumption in many settings in economics is that the underlying function is concave and increasing (for example, production functions). Such shape restrictions can also be handled on the computational front using the Kuhn–Tucker theorems but the asymptotic properties of the maximum likelihood estimates would be very different and their derivation can be expected to be a considerably difficult affair. Nonparametric likelihood estimation/least squares based estimation under concavity constraints is a relatively new area (see, in particular, the work by Groeneboom et al. (2001)) and much remains to be done still, especially in the semiparametric domain. It is however clear that the additional regularity brought in via the concavity constraint will result in an accelerated rate of convergence ($n^{2/5}$) for the maximum likelihood estimate of the nonparametric component of the model.

Finally, another pertinent question is inference on the rate of change of the response with respect to the covariates of interest. For example, one could be interested in the partial derivative of $\mu(x, z) = P(\Delta = 1 \mid X = x, Z = z)$ with respect to z. Since the estimate of μ produced by the likelihood based methods of the current paper is piecewise constant in z, a direct estimate of the derivative is unavailable. Some recent work by Groeneboom and Jongbloed (2003) suggests that kernel smoothing of the semiparametric maximum likelihood estimate using a bandwidth of order $n^{-1/7}$ will produce a consistent estimate of the derivative that is asymptotically normal with rate of convergence $n^{2/7}$, whence confidence intervals may be constructed via resampling techniques. However, a detailed study of these methods will be a fairly involved affair and is left as a topic for future research.

Acknowledgments

We would like to express our deepest thanks to the Editor, Professor Takeshi Amemiya, an anonymous AE and three anonymous referees whose perceptive comments and suggestions led to a much improved version of this paper. We also profess our gratitude to Dr. Pinaki Biswas for his programming help.

The first author's research was supported in part by National Science Foundation grants DMS-0306235 and DMS-0705288.

Appendix

Owing to space considerations and the fact that the basic techniques behind some of these proofs are available in the existing literature, we present concise versions of some of the arguments. For more detailed derivations, we refer the reader to the appendix of the companion technical report of Banerjee et al. (2007) (henceforth, BMM(2007)). The proof of Theorem 3.1, in particular, is adapted very closely from the arguments on the current status model in Murphy and Van der Vaart (1997) and is therefore omitted from the paper. See the appendix of BMM(2007) for the details.

Proof-sketch of Theorem 3.2. The first step is to establish finite-dimensional convergence of the processes $(U_n(h), V_n(h))$ to $(g_{a,b}(h), g^0_{a,b}(h))$. Thus, it is shown that for any $(h_1, h_2, \ldots, h_k)$, the random vector $\left(\{U_n(h_i)\}_{i=1}^k, \{V_n(h_i)\}_{i=1}^k\right) \to_d \left(\{g_{a,b}(h_i)\}_{i=1}^k, \{g^0_{a,b}(h_i)\}_{i=1}^k\right)$ in the space $\mathbb{R}^{2k}$. Next, to deduce the convergence in $L_2[-K, K] \times L_2[-K, K]$ note firstly that $U_n(h)$ and $V_n(h)$ are monotone functions. Now, given a sequence (ψ_n, ϕ_n) in $L_2[-K, K] \times L_2[-K, K]$ such that ψ_n and ϕ_n are monotone functions and (ϕ_n, ψ_n) converges pointwise to (ϕ, ψ) (where (ϕ, ψ) is in

$L_2[-K,K]\times L_2[-K,K]$), we can conclude that $(\psi_n,\phi_n)\to(\psi,\phi)$ in $L_2[-K,K]\times L_2[-K,K]$. It follows, in the wake of distributional convergence of all the finite-dimensional marginals of (U_n,V_n) to those of $(g_{a,b}(h),g^0_{a,b}(h))$, that $(U_n(h),V_n(h))\to_d(g_{a,b}(h),g^0_{a,b}(h))$ in $L_2[-K,K]\times L_2[-K,K]$ (this parallels the result of Corollary 2 following Theorem 3 of Huang and Zhang (1994)).

The proof of finite-dimensional convergence can be based on continuous mapping arguments for slopes-of-greatest-convex-minorant estimators, as used in the proof of Theorem 2.1 of Banerjee (2007). An alternative approach is to make use of "switching relationships" which allow us to translate the behavior of the slope of the convex minorant of a random cumulative sum diagram (this is how the estimators $\hat\psi_n^{(\beta_0)}$ and $\hat\psi_{n,0}^{(\beta_0)}$ are characterized) in terms of the minimizer of a stochastic process. The limiting behavior of the slope process can then be studied in terms of the limiting behavior of the minimizer of this stochastic process by applying continuous mapping theorems for the argmin functional. For the details, see the appendix of BMM(2007). □

A.1. Further details about the unconstrained and constrained MLEs from Section 2

Details of the "self-consistent" characterization of $\hat{\mathbf{u}}_n^{(\beta)}$ in Section 2: As in Section 2, we denote the function $g(\beta,\mathbf{u})$ in the discussion that follows by $\xi(\mathbf{u})$. This is strictly convex in $\mathbf{u}\equiv(u_1,u_2,\ldots,u_n)$ and for simplicity we denote $\hat{\mathbf{u}}_n^{(\beta)}$, its minimizer over the region $\mathcal{C}:=\{\mathbf{u}:u_1\le u_2\le\cdots\le u_n\}$, by $\hat{\mathbf{u}}$ (suppressing the dependence on n and β). Let $\nabla_j\,\xi(\mathbf{u})$ denote the j'th partial derivative of ξ with respect to $\mathbf{u}$. Using the Kuhn–Tucker theorem for optimizing a convex function over a closed convex set, we find that $\hat{\mathbf{u}}=(\hat u_1,\hat u_2,\ldots,\hat u_n)$ is uniquely characterized by the conditions:

$$\sum_{j=i+1}^{n}\nabla_j\,\xi(\hat{\mathbf{u}})\ge 0,\quad \text{for } i=1,2,\ldots,(n-1) \tag{A.14}$$

and

$$\sum_{j=1}^{n}\nabla_j\,\xi(\hat{\mathbf{u}})=0. \tag{A.15}$$

Consider now, the following (quadratic) function $\tilde\xi(\mathbf{u})=\frac{1}{2}\left[\mathbf{u}-\hat{\mathbf{u}}+\mathcal{K}^{-1}\,\nabla\,\xi(\hat{\mathbf{u}})\right]^{\mathrm{T}}\mathcal{K}\left[\mathbf{u}-\hat{\mathbf{u}}+\mathcal{K}^{-1}\,\nabla\,\xi(\hat{\mathbf{u}})\right]$ where $\mathcal{K}$ is some positive definite matrix. Note that $\mathrm{Hess}(\tilde\xi)=\mathcal{K}$ which is positive definite; thus $\tilde\xi$ is a strictly convex function. It is also finite and continuously differentiable over $\mathbb{R}^n$. Also, $\nabla\tilde\xi(\mathbf{u})=\mathcal{K}\left(\mathbf{u}-\hat{\mathbf{u}}+\mathcal{K}^{-1}\,\nabla\,\xi(\hat{\mathbf{u}})\right)$. Now, consider the problem of minimizing $\tilde\xi$ over $\mathcal{C}$. If $\mathbf{u}^\star$ is the (unique) global minimizer, then necessary and sufficient conditions are given by conditions (A.14) (for $i=1,2,\ldots,n-1$) and (A.15), with ξ replaced by $\tilde\xi$ and $\hat{\mathbf{u}}$ replaced by $\mathbf{u}^\star$. Now, $\nabla\tilde\xi(\hat{\mathbf{u}})=\nabla\xi(\hat{\mathbf{u}})$, so that the vector $\hat{\mathbf{u}}\in\mathcal{C}$ does indeed satisfy the conditions (A.14) (for $i=1,2,\ldots,n-1$) and (A.15), with ξ replaced by $\tilde\xi$. It follows that $\hat{\mathbf{u}}$ is the unique minimizer of $\tilde\xi$ over $\mathcal{C}$, i.e. $\mathbf{u}^\star=\hat{\mathbf{u}}$.

It now suffices to try to minimize $\tilde\xi$; of course the problem here is that $\hat{\mathbf{u}}$ is unknown and $\tilde\xi$ is defined in terms of $\hat{\mathbf{u}}$. However, an iterative scheme can be developed along the following lines. Choosing $\mathcal{K}$ to be a diagonal matrix with the i, ith entry being $d_i\equiv\nabla_{ii}\,\xi(\hat{\mathbf{u}})$ ($\mathcal{K}$ thus defined is a p.d. matrix, since the diagonal entries of the Hessian of ξ at the minimizer $\hat{\mathbf{u}}$, which is a positive definite matrix, are positive), we see that the above quadratic form reduces to $\eta(\mathbf{u})/2$ where $\eta(\mathbf{u})=\sum_{i=1}^n\left[u_i-\left(\hat u_i-\nabla_i\,\xi(\hat{\mathbf{u}})d_i^{-1}\right)\right]^2 d_i$. Thus, $\hat{\mathbf{u}}$ minimizes $\eta(\mathbf{u})$ subject to the constraints that $u_1\le u_2\le\cdots\le u_n$ and therefore furnishes the isotonic regression of the function $h(i)=\hat u_i-\nabla_i\,\xi(\hat{\mathbf{u}})\,d_i^{-1}$ on the ordered set $\{1,2,\ldots,n\}$ with weight function d_i. From the theory of isotonic regression, it is well known that the solution $\hat{\mathbf{u}}\equiv(\hat u_1,\hat u_2,\ldots,\hat u_n)=\mathrm{slogcm}\left\{\sum_{j=1}^i d_i,\sum_{j=1}^i h(i)\,d_i\right\}_{i=0}^n$. This representation leads to the MICM as outlined in Section 2.

Implications of the self-consistent/self-induced characterization of $\hat{\mathbf{u}}_n^{(\beta)}$: Recall that $\hat{\mathbf{u}}_n^{(\beta)}=(\hat u_{1,n}^{(\beta)},\hat u_{2,n}^{(\beta)},\ldots,\hat u_{n,n}^{(\beta)})$. Let $B_1,B_2,\ldots,B_k$ be the unique partitioning of $1,2,\ldots,n$ into ordered blocks of indices (say $B_1=\{1,2,\ldots,l_1\},B_2=\{l_1+1,l_1+2,\ldots,l_2\}$ and so on) such that, for each i, for all $j\in B_i$, $\hat u_{j,n}^{(\beta)}$ equals w_i, with the common block values, the w_i's, satisfying $w_1<w_2<\cdots<w_k$. Since the $\hat u_{j,n}^{(\beta)}$'s are increasing in j, this is possible. An important consequence of the self-consistent characterization is the fact that each w_i can be written as a weighted average of the $h(j)$'s for the j's in B_i, with the weights given by the d_j's. The B_i's are called the *level blocks* of $\hat{\mathbf{u}}_n^{(\beta)}$ and the w_i's are called the *level values*.

We now introduce some notation that will be useful in the proof of Theorem 3.3. Denote $\phi(\Delta_{(i)},R_i(\beta),t)$ by $\phi_{i,\beta}(t)$ and its first and second derivatives with respect to t by $\phi'_{i,\beta}(t)$ and $\phi''_{i,\beta}(t)$. Identifying the function $\hat\psi_n^{(\beta)}$ with the vector $\hat{\mathbf{u}}_n^{(\beta)}$ in the usual fashion, we can write

$$\hat\psi_n^{(\beta)}\equiv\mathrm{slogcm}\left\{\sum_{i=1}^k\phi''_{i,\beta}(\hat\psi_n^{(\beta)}(Z_{(i)})),\ \sum_{i=1}^k\left[\hat\psi_n^{(\beta)}(Z_{(i)})-\frac{\phi'_{i,\beta}(\hat\psi_n^{(\beta)}(Z_{(i)}))}{\phi''_{i,\beta}(\hat\psi_n^{(\beta)}(Z_{(i)}))}\right]\phi''_{i,\beta}(\hat\psi_n^{(\beta)}(Z_{(i)}))\right\}_{k=0}^n.$$

Hence, we can write w_i as

$$w_i=\hat\psi_n^{(\beta)}(Z_{(j)})=\frac{\sum_{k\in B_i}\{\hat\psi_n^{(\beta)}(Z_{(k)})\,\phi''_{k,\beta}(\hat\psi_n^{(\beta)}(Z_{(k)}))-\phi'_{k,\beta}(\hat\psi_n^{(\beta)}(Z_{(k)}))\}}{\sum_{k\in B_i}\phi''_{k,\beta}(\hat\psi_n^{(\beta)}(Z_{(k)}))}\quad\text{for } j\in B_i. \tag{A.16}$$

Further details about $\hat{\mathbf{u}}_{n,0}^{(\beta)}$: The vector $\hat{\mathbf{u}}_{n,0}^{(\beta)}$, which we identify with $\hat\psi_{n,0}^{(\beta)}$ (as explained in Section 2) also has a self-consistent/self-induced characterization in terms of the slope of the greatest convex minorant of a random function. This follows in the same way as in the case of $\hat{\mathbf{u}}_n^{(\beta)}$ by formulating a quadratic optimization problem based on the Kuhn–Tucker conditions for the corresponding minimization problem. We skip the details but give the self-consistent characterization. As before, we abbreviate $g(\beta,\mathbf{u})$ to $\xi(\mathbf{u})$, suppressing the dependence on β. We also abbreviate $\hat{\mathbf{u}}_{n,0}^{(\beta)}$ to $\hat{\mathbf{u}}^{(0)}$. For each i, set $d_i=\nabla_{ii}\,\xi(\hat{\mathbf{u}}^{(0)})$. Then, $\hat{\mathbf{u}}^{(0)}$ minimizes, $A(u_1,u_2,\ldots,u_n)=\sum_{i=1}^n\left[u_i-\left(\hat u_i^{(0)}-\nabla_i\,\xi(\hat{\mathbf{u}}^{(0)})d_i^{-1}\right)\right]^2 d_i$ subject to the constraints that $u_1\le u_2\le\cdots\le u_m\le\theta_0\le u_{m+1}\le\cdots\le u_n$. Let $\tilde B_1,\tilde B_2,\ldots,\tilde B_l$ denote the level blocks of $\hat{\mathbf{u}}_{n,0}^{(\beta)}$ and let $\{\tilde w_i\}_{i=1}^l$ denote the corresponding level values. Then, as long as $\tilde w_i\ne\theta_0$, it can be written as

$$\tilde w_i=\hat\psi_{n,0}^{(\beta)}(Z_{(j)})=\frac{\sum_{k\in\tilde B_i}\{\hat\psi_{n,0}^{(\beta)}(Z_{(k)})\,\phi''_{k,\beta}(\hat\psi_{n,0}^{(\beta)}(Z_{(k)}))-\phi'_{k,\beta}(\hat\psi_{n,0}^{(\beta)}(Z_{(k)}))\}}{\sum_{k\in\tilde B_i}\phi''_{k,\beta}(\hat\psi_{n,0}^{(\beta)}(Z_{(k)}))}\quad\text{for } j\in\tilde B_i. \tag{A.17}$$

Please cite this article in press as: Banerjee, M., et al., Semiparametric binary regression models under shape constraints with an application to Indian schooling data. Journal of Econometrics (2009), doi:10.1016/j.jeconom.2008.11.002

This representation is once again, a direct outcome of the self-induced characterization, and will prove useful in what follows.

A.2. *Proof of* Theorem 3.3

The likelihood ratio statistic of interest can be written as

$$\begin{aligned}\text{lrtpsi}_n &= 2\,(l_n(\hat\beta_n, \hat\psi_n) - l_n(\hat\beta_{n,0}, \hat\psi_{n,0})) \\ &= 2\,(l_n(\beta_0, \hat\psi_n^{(\beta_0)}) - l_n(\beta_0, \hat\psi_{n,0}^{(\beta_0)})) \\ &\quad + 2\,(l_n(\hat\beta_n, \hat\psi_n) - l_n(\beta_0, \hat\psi_n^{(\beta_0)})) \\ &\quad - 2(l_n(\hat\beta_{n,0}, \hat\psi_{n,0}) - l_n(\beta_0, \hat\psi_{n,0}^{(\beta_0)})).\end{aligned}$$

From Theorem 3.1 $\tilde R_n \equiv 2\,(l_n(\hat\beta_n, \hat\psi_n) - l_n(\beta_0, \hat\psi_n^{(\beta_0)})) - 2(l_n(\hat\beta_{n,0}, \hat\psi_{n,0}) - l_n(\beta_0, \hat\psi_{n,0}^{(\beta_0)}))$ is $o_p(1)$ whence it suffices to find the asymptotic distribution of $C_n = 2\,(l_n(\beta_0, \hat\psi_n^{(\beta_0)}) - l_n(\beta_0, \hat\psi_{n,0}^{(\beta_0)}))$. This is precisely the likelihood ratio statistic for testing $\psi(z_0) = \theta_0$ holding β fixed at its true value β_0. We can write $C_n = 2\left[\sum_{i=1}^n \phi(\Delta_{(i)}, R_i(\beta_0), \hat\psi_{n,0}^{(\beta_0)}(Z_{(i)})) - \sum_{i=1}^n \phi(\Delta_{(i)}, R_i(\beta_0), \hat\psi_n^{(\beta_0)}(Z_{(i)}))\right]$ where ϕ is as defined in Section 2. For the sake of notational compactness, in the remainder of the proof, we write $\hat\psi_n^{(\beta_0)}(Z_{(i)})$ as $\tilde\psi(Z_{(i)})$, $\hat\psi_{n,0}^{(\beta_0)}(Z_{(i)})$ as $\tilde\psi_0(Z_{(i)})$, and $\phi(\Delta_{(i)}, R_i(\beta_0), t)$ as $\phi_i(t)$. Furthermore $\partial/\partial\, t\ \phi(\Delta_{(i)}, R_i(\beta_0), t)$ will be written as $\phi_i'(t)$ and so on. The set of indices i on which $\tilde\psi(Z_{(i)})$ and $\tilde\psi_0(Z_{(i)})$ differ is denoted by J_n. Now, $C_n = -2\,T_n$ where

$$\begin{aligned}T_n &= \sum_{i=1}^n \phi_i(\tilde\psi(Z_{(i)})) - \sum_{i=1}^n \phi_i(\tilde\psi_0(Z_{(i)})) \\ &= \sum_{i\in J_n} \phi_i(\tilde\psi(Z_{(i)})) - \sum_{i\in J_n} \phi_i(\tilde\psi_0(Z_{(i)})) \\ &= \sum_{i\in J_n} \phi_i'(\psi_0(z_0))\,[(\tilde\psi(Z_{(i)}) - \psi_0(z_0)) - (\tilde\psi_0(Z_{(i)}) - \psi_0(z_0))] \\ &\quad + \sum_{i\in J_n} \frac{1}{2}\,\phi_i''(\psi_0(z_0))\left[(\tilde\psi(Z_{(i)}) - \psi_0(z_0))^2\right. \\ &\quad \left. - (\tilde\psi_0(Z_{(i)}) - \psi_0(z_0))^2\right] + R_n \\ &\equiv T_{n,1} + T_{n,2} + R_n \quad \text{by Taylor-expanding} \\ &\qquad \phi_i(t) \text{ around } \psi_0(z_0).\end{aligned}$$

Here, $R_n = \sum_{i\in J_n} (1/6)[\phi_i'''(\tilde\psi(Z_{(i)})^\star)\,(\tilde\psi(Z_{(i)}) - \psi_0(z_0))^3 - \phi_i'''(\tilde\psi_0(Z_{(i)})^\star)\,(\tilde\psi_0(Z_{(i)}) - \psi_0(z_0))^3]$ (where $\tilde\psi(Z_{(i)})^\star$ is some point between $\tilde\psi(Z_{(i)})$ and $\psi_0(z_0)$ and $\tilde\psi_0(Z_{(i)})^\star$ is some point between $\tilde\psi_0(Z_{(i)})$ and $\psi_0(z_0)$) and can be shown to converge to 0 in probability by using the facts that (a) $\sup_{i\in J_n} |\phi_i'''(\tilde\psi(Z_{(i)})^\star)|$ and $\sup_{i\in J_n} |\phi_i'''(\tilde\psi_0(Z_{(i)})^\star)|$ are $O_p(1)$, (b) $\sup_{z\in D_n} |\tilde\psi(z) - \psi_0(z_0)|$ and $\sup_{z\in D_n} |\tilde\psi_0(z) - \psi_0(z_0)|$ are $O_p(n^{-1/3})$ where D_n is the set on which $\tilde\psi$ and $\tilde\psi_0$ differ, and (c) the length of D_n is $O_p(n^{-1/3})$. Now consider $T_{n,2}$. Once again, by Taylor expansion, we have

$$\begin{aligned}T_{n,2} &\equiv \sum_{i\in J_n} \frac{1}{2}\,\phi_i''(\psi_0(z_0))\left[(\tilde\psi(Z_{(i)})\right. \\ &\quad \left. - \psi_0(z_0))^2 - (\tilde\psi_0(Z_{(i)}) - \psi_0(z_0))^2\right] \\ &= \sum_{i\in J_n} \frac{1}{2}\phi_i''(\tilde\psi(Z_{(i)}))[\tilde\psi(Z_{(i)}) - \psi_0(z_0)]^2 \\ &\quad - \sum_{i\in J_n} \frac{1}{2}\phi_i''(\tilde\psi_0(Z_{(i)}))[\tilde\psi_0(Z_{(i)}) - \psi_0(z_0)]^2 + o_p(1). \quad \text{(A.18)}\end{aligned}$$

Now consider, $T_{n,1} \equiv \sum_{i\in J_n} \phi_i'(\psi_0(z_0))(\tilde\psi(Z_{(i)}) - \psi_0(z_0)) - \sum_{i\in J_n} \phi_i'(\psi_0(z_0))\,(\tilde\psi_0(Z_{(i)}) - \psi_0(z_0)) \equiv S_1 - S_2$. Consider the term S_2. Note that for each $i \in J_n$, we can write:

$$\begin{aligned}\phi_i'(\psi_0(z_0)) &= \phi_i'(\tilde\psi_0(Z_{(i)})) + (\psi_0(z_0) - \tilde\psi_0(Z_{(i)}))\,\phi_i''(\tilde\psi_0(Z_{(i)})) \\ &\quad + \frac{1}{2}\,\phi_i'''(\tilde\psi_0(Z_{(i)})^{\star\star})(\psi_0(z_0) - \tilde\psi_0(Z_{(i)}))^2\end{aligned}$$

where $\tilde\psi_0(Z_{(i)})^{\star\star}$ is a point between $\tilde\psi_0(Z_{(i)})$ and $\psi_0(z_0)$. We then have,

$$\begin{aligned}S_2 &= \sum_{i\in J_n} \left[\phi_i'(\tilde\psi_0(Z_{(i)})) + (\psi_0(z_0) - \tilde\psi_0(Z_{(i)}))\,\phi_i''(\tilde\psi_0(Z_{(i)}))\right. \\ &\quad \left. + \frac{1}{2}\,\phi_i'''(\tilde\psi_0(Z_{(i)})^{\star\star})(\psi_0(z_0) - \tilde\psi_0(Z_{(i)}))^2\right] \\ &\quad \times(\tilde\psi_0(Z_{(i)}) - \psi_0(z_0)) \\ &= \sum_{i\in J_n} \left[\phi_i'(\tilde\psi_0(Z_{(i)})) + (\psi_0(z_0) - \tilde\psi_0(Z_{(i)}))\,\phi_i''(\tilde\psi_0(Z_{(i)}))\right] \\ &\quad \times (\tilde\psi_0(Z_{(i)}) - \psi_0(z_0)) + o_p(1) \\ &= -\sum_{i\in J_n} \phi_i''(\tilde\psi_0(Z_{(i)}))\left[\tilde\psi_0(Z_{(i)}) - \frac{\phi_i'(\tilde\psi_0(Z_{(i)}))}{\phi_i''(\tilde\psi_0(Z_{(i)}))} - \psi_0(z_0)\right] \\ &\quad \times (\tilde\psi_0(Z_{(i)}) - \psi_0(z_0)) + o_p(1),\end{aligned}$$

where the fact that the term involving ϕ_i''' is $o_p(1)$ is deduced by arguments similar to those needed to show that R_n is $o_p(1)$. Now, let $B_1^0, B_2^0, \ldots, B_r^0$ denote the level blocks for $\tilde\psi_0(Z_{(i)})$ that constitute J_n, with level values $w_1^0, w_2^0, \ldots, w_r^0$ and suppose that $w_l^0 = \psi_0(z_0) \equiv \theta_0$. Then,

$$\begin{aligned}&S_2 + o_p(1) \\ &= -\sum_{j=1}^r \sum_{i\in B_j} \left[\phi_i''(\tilde\psi_0(Z_{(i)}))\left(\tilde\psi_0(Z_{(i)}) - \frac{\phi_i'(\tilde\psi_0(Z_{(i)}))}{\phi_i''(\tilde\psi_0(Z_{(i)}))}\right)\right. \\ &\quad \left. - \psi_0(z_0)\,\phi_i''(\tilde\psi_0(Z_{(i)}))\right](\tilde\psi_0(Z_{(i)}) - \psi_0(z_0)) \\ &= -\sum_{j=1}^r \sum_{i\in B_j} \left[\phi_i''(w_j^0)\left(w_j^0 - \frac{\phi_i'(w_j^0)}{\phi_i''(w_j^0)}\right) - \psi_0(z_0)\,\phi''(w_j^0)\right] \\ &\quad \times (w_j^0 - \psi_0(z_0)) \\ &= -\sum_{j\neq l} (w_j^0 - \psi_0(z_0))\left[\sum_{i\in B_j} (\phi_i''(w_j^0)\,w_j^0 - \phi_i'(w_j^0)) - \right. \\ &\quad \left. \psi_0(z_0)\sum_{i\in B_j} \phi_i''(w_j^0)\right] \\ &= -\sum_{j\neq l} (w_j^0 - \psi_0(z_0))\left[\left(\sum_{i\in B_j} \phi_i''(w_j^0)\right)\right. \\ &\quad \left. \times \left[\frac{\sum_{i\in B_j} (\phi_i''(w_j^0)\,w_j^0 - \phi_i'(w_j^0))}{\sum_{i\in B_j} \phi_i''(w_j^0)} - \psi_0(z_0)\right]\right] \\ &= -\sum_{j\neq l} \sum_{i\in B_j} \phi_i''(w_j^0)\,(w_j^0 - \psi_0(z_0))^2,\end{aligned}$$

where this last step follows from the following observation: If B' is a level block for $\tilde\psi_0$ contained in J_n with level value $w^{(0)}$, then $w^{(0)} = [\sum_{k\in B'} (w^{(0)}\,\phi_k''(w^{(0)}) - \phi_k'(w^{(0)}))]/[\sum_{k\in B'} \phi_k''(w^{(0)})]$

Please cite this article in press as: Banerjee, M., et al., Semiparametric binary regression models under shape constraints with an application to Indian schooling data. Journal of Econometrics (2009), doi:10.1016/j.jeconom.2008.11.002

provided $w^{(0)} \neq \theta_0$. This is a direct consequence of the representation (A.17). It follows that

$$\begin{aligned} S_2 + o_p(1) &= -\sum_{j\neq l}\sum_{i\in B_j} \phi_i''(w_j^0)\,(w_j^0 - \psi_0(z_0))^2 \\ &= -\sum_{j=1}^{r}\sum_{i\in B_j} \phi_i''(w_j^0)\,(w_j^0 - \psi_0(z_0))^2 \\ &= -\sum_{j=1}^{r}\sum_{i\in B_j} \phi_i''(\tilde\psi_0(Z_{(i)}))\,(\tilde\psi_0(Z_{(i)}) - \psi_0(z_0))^2 \\ &= -\sum_{i\in J_n} \phi_i''(\tilde\psi_0(Z_{(i)}))\,(\tilde\psi_0(Z_{(i)}) - \psi_0(z_0))^2. \end{aligned}$$

It is similarly established (using (A.16)) that $S_1 + o_p(1) = -\sum_{i\in J_n} \phi_i''(\tilde\psi(Z_{(i)}))\,(\tilde\psi(Z_{(i)}) - \psi_0(z_0))^2$. It follows that $T_{n,1} = -\sum_{i\in J_n} \phi_i''(\tilde\psi(Z_{(i)}))\,(\tilde\psi(Z_{(i)}) - \psi_0(z_0))^2 + \sum_{i\in J_n} \phi_i''(\tilde\psi_0(Z_{(i)}))\,(\tilde\psi_0(Z_{(i)}) - \psi_0(z_0))^2 + o_p(1)$. Now, on using (A.18) and the fact that R_n is $o_p(1)$ we get

$$\begin{aligned} T_n &= T_{n,1} + T_{n,2} + o_p(1) \\ &= -\frac{1}{2}\sum_{i\in J_n} \phi_i''(\tilde\psi(Z_{(i)}))\,(\tilde\psi(Z_{(i)}) - \psi_0(z_0))^2 \\ &\quad + \frac{1}{2}\sum_{i\in J_n} \phi_i''(\tilde\psi_0(Z_{(i)}))\,(\tilde\psi_0(Z_{(i)}) - \psi_0(z_0))^2 + o_p(1), \end{aligned}$$

whence

$$\begin{aligned} C_n &= -2\,T_n = \sum_{i\in J_n} \phi_i''(\tilde\psi(Z_{(i)}))\,(\tilde\psi(Z_{(i)}) - \psi_0(z_0))^2 \\ &\quad - \sum_{i\in J_n} \phi_i''(\tilde\psi_0(Z_{(i)}))\,(\tilde\psi_0(Z_{(i)}) - \psi_0(z_0))^2 + o_p(1) \\ &= \sum_{i\in J_n} \phi_i''(\psi_0(Z_{(i)}))\,(\tilde\psi(Z_{(i)}) - \psi_0(z_0))^2 \\ &\quad - \sum_{i\in J_n} \phi_i''(\psi_0(Z_{(i)}))\,(\tilde\psi_0(Z_{(i)}) - \psi_0(z_0))^2 + o_p(1). \end{aligned}$$

Now, $\phi_i''(\psi_0(Z_{(i)})) = \frac{\exp(\psi_0(Z_{(i)}) + \beta_0^{\mathrm{T}} X_{(i)})}{(1+\exp(\psi_0(Z_{(i)}) + \beta_0^{\mathrm{T}} X_{(i)}))^2}$, whence

$$\begin{aligned} C_n &= \sum_{i\in J_n} \frac{\exp(\psi_0(Z_{(i)}) + \beta_0^{\mathrm{T}} X_{(i)})}{(1+\exp(\psi_0(Z_{(i)}) + \beta_0^{\mathrm{T}} X_{(i)}))^2} \\ &\quad \times \left[(\tilde\psi(Z_{(i)}) - \psi_0(z_0))^2 - (\tilde\psi_0(Z_{(i)}) - \psi_0(z_0))^2\right] + o_p(1) \\ &= n^{1/3}\,(\mathbb{P}_n - P)\,\xi_n(\delta, z, x) + n^{1/3}\,P\,\xi_n(\delta, z, x) + o_p(1) \end{aligned}$$

where $\mathbb{P}_n$ is the empirical measure of the observations $\{\Delta_i, Z_i, X_i\}_{i=1}^n$, P denotes the true underlying distribution of (Δ, Z, X), ξ_n is the random function given by

$$\begin{aligned} \xi_n(\delta, z, x) &= \frac{\exp(\psi_0(z) + \beta_0^{\mathrm{T}} x)}{(1+\exp(\psi_0(z) + \beta_0^{\mathrm{T}} x))^2} \\ &\quad \times \left[(n^{1/3}(\tilde\psi(z) - \psi_0(z_0)))^2 \right. \\ &\quad \left. - (n^{1/3}\,(\tilde\psi_0(z) - \psi_0(z_0)))^2\right] 1(z \in D_n). \end{aligned}$$

We are using operator notation here for expectations; thus $\mathbb{P}_n g$ denotes the expectation of g under the measure $\mathbb{P}_n$ and $P\,g$ denotes the expectation of g under the measure P. The function g is allowed to be a random function. Now, $n^{1/3}\,(\mathbb{P}_n - P)\,\xi_n(\delta, z, x) = n^{-1/6}\,\sqrt{n}\,(\mathbb{P}_n - P)\,\xi_n(\delta, z, x)$. Using the facts that (i) D_n is eventually contained in a set of the form $[z_0 - M\,n^{-1/3}, z_0 + M\,n^{-1/3}]$ with arbitrarily high preassigned probability (ii) the processes U_n and V_n are $O_p(1)$ on compacts and monotone increasing, along with standard preservation properties of Donsker classes of functions, it can be argued that with arbitrarily high preassigned probability, the function $\xi_n(\delta, z, x)$ lies in a Donsker class, whence it follows that $\sqrt{n}\,(\mathbb{P}_n - P)\,\xi_n(\delta, z, x)$ is $O_p(1)$; consequently $n^{1/3}\,(\mathbb{P}_n - P)\,\xi_n(\delta, z, x)$ is $O_p(n^{-1/6})$ and hence $o_p(1)$.

To find the asymptotic distribution of C_n we can therefore concentrate on the asymptotic distribution of $n^{1/3}\,P\,\xi_n(\delta, z, x) = n^{1/3}\,P\,[h(z, x)\,K_n(z)]$ where $K_n(z) = \left[(n^{1/3}(\tilde\psi(z) - \psi_0(z_0)))^2 - (n^{1/3}(\tilde\psi_0(z) - \psi_0(z_0)))^2\right] 1(Z \in D_n)$ and $h(z, x) = \frac{\exp(\psi_0(z) + \beta_0^{\mathrm{T}} x)}{(1+\exp(\psi_0(z) + \beta_0^{\mathrm{T}} x))^2}$. Thus,

$$\begin{aligned} &n^{1/3}\,P\,\xi_n(\delta, z, x) \\ &= n^{1/3}\,P\,[K_n(z)\,h(z, x)] \\ &= n^{1/3}\int_{D_n} K_n(z)\,E(h(Z, X) \mid Z = z)\,f_Z(z)\,\mathrm{d}z \\ &= n^{1/3}\int_{\tilde D_n} K_n(z_0 + h\,n^{-1/3})\,w(z_0 + h\,n^{-1/3})\,f_Z(z_0 + h\,n^{-1/3})\,\mathrm{d}h \end{aligned}$$

where $h = n^{1/3}\,(z - z_0)$, $\tilde D_n = n^{1/3}\,(D_n - z_0)$ and $w(z) = E(h(Z, X) \mid Z = z)$. Now note that, $K_n(z_0 + h\,n^{-1/3}) = (U_n^2(h) - V_n^2(h))\,1\,(h \in \tilde D_n)$ where $\tilde D_n$ is the set on which U_n and V_n differ. Also, w is continuous in z and is given by $w(z) = \int \frac{\exp(\psi_0(z) + \beta_0^{\mathrm{T}} x)}{(1+\exp(\psi_0(z) + \beta_0^{\mathrm{T}} x))^2}\,\frac{f(z,x)}{f_Z(z)}\,\mathrm{d}\mu(x)$.

On using the facts that $\tilde D_n$ is eventually contained with arbitrarily high probability in a compact set and the boundedness in probability of the processes U_n and V_n on compacts along with the continuity of the functions w and f_Z, we get, $n^{1/3}\,P\,\xi_n(\delta, z, x) = \int w(z_0)\,f_Z(z_0)\,(U_n^2(h) - V_n^2(h))\,\mathrm{d}h + o_p(1)$. But $C(z_0) = w(z_0)\,f_Z(z_0) = 1/a^2$ where a is as defined in Theorem 3.2. An application of Theorem 3.2 and Slutsky's theorem yields $n^{1/3}\,P\,\xi_n(\delta, z, x) \to_d a^{-2}\int \left((g_{a,b}(h))^2 - (g_{a,b}^0(h))^2\right)\mathrm{d}h$, and the fact that $a^{-2}\int \left((g_{a,b}(h))^2 - (g_{a,b}^0(z))^2\right)\mathrm{d}h \equiv_d \int \left((g_{1,1}(h))^2 - (g_{1,1}^0(z))^2\right)\mathrm{d}h \equiv \mathbb{D}$ follows as a direct application of Lemma 3.1 followed by the change of variable theorem from calculus.

It remains to show that

$$\begin{aligned} \tilde R_n &= 2\,(l_n(\hat\beta_n, \hat\psi_n) - l_n(\beta_0, \hat\psi_n^{\beta_0})) \\ &\quad - 2(l_n(\hat\beta_{n,0}, \hat\Lambda_{n,0}) - l_n(\beta_0, \hat\psi_{n,0}^{\beta_0})) \\ &\equiv 2\,(l_n(\hat\beta_n, \hat\Lambda_n) - l_n(\beta_0, \hat\Lambda_n^{\beta_0})) \\ &\quad - 2(l_n(\hat\beta_{n,0}, \hat\Lambda_{n,0}) - l_n(\beta_0, \hat\Lambda_{n,0}^{\beta_0})) \end{aligned}$$

is $o_p(1)$. This is precisely $\mathrm{lrtbeta}_n - \mathrm{lrtbeta}_n^0$. From Theorem 3.1 we get:

$$\begin{aligned} &\mathrm{lrtbeta}_n - \mathrm{lrtbeta}_n^0 \\ &= n\,(\hat\beta_n - \beta_0)^{\mathrm{T}}\,\tilde I_0\,(\hat\beta_n - \beta_0) - n\,(\tilde\beta_n - \beta_0)^{\mathrm{T}}\,\tilde I_0\,(\tilde\beta_n - \beta_0) + o_p(1) \\ &= n\,(\hat\beta_n - \tilde\beta_n)^{\mathrm{T}}\,\tilde I_0\,(\hat\beta_n - \tilde\beta_n) + 2\,n\,(\tilde\beta_n - \beta_0)^{\mathrm{T}}\,\tilde I_0\,(\hat\beta_n - \tilde\beta_n) + o_p(1) \\ &= \sqrt{n}\,(\hat\beta_n - \tilde\beta_n)^{\mathrm{T}}\,\tilde I_0\,\sqrt{n}\,(\hat\beta_n - \tilde\beta_n) \\ &\quad + 2\,\sqrt{n}\,(\tilde\beta_n - \beta_0)^{\mathrm{T}}\,\tilde I_0\,\sqrt{n}(\hat\beta_n - \tilde\beta_n) + o_p(1) \\ &\equiv I_n + II_n + o_p(1). \end{aligned}$$

The fact that I_n is $o_p(1)$ follows from the observation that $\sqrt{n}\,(\hat\beta_n - \tilde\beta_n) = r_n - s_n$, which is $o_p(1)$ (by Theorem 3.1). The fact that II_n is $o_p(1)$ follows on using the facts that $\sqrt{n}\,(\hat\beta_n - \tilde\beta_n)$ is $o_p(1)$ and that $\sqrt{n}\,(\tilde\beta_n - \beta_0)$ is $O_p(1)$. □

Please cite this article in press as: Banerjee, M., et al., Semiparametric binary regression models under shape constraints with an application to Indian schooling data. Journal of Econometrics (2009), doi:10.1016/j.jeconom.2008.11.002

References

Banerjee, M., 2000. Likelihood ratio inference in regular and nonregular problems. Ph.D. Dissertation. University of Washington.
Banerjee, M., Wellner, J.A., 2001. Likelihood ratio tests for monotone functions. Ann. Statist. 29, 1699–1731.
Banerjee, M., Wellner, J.A., 2005. Score statistics for current status data: Comparisons with likelihood ratio and wald statistics. Int. J. Biostatistics 1 (1), Article 3.
Banerjee, M., 2007. Likelihood based inference for monotone response models. Ann. Statist. 35, 931–956.
Banerjee, M., Mukherjee, D., Mishra, S., 2007. Semiparametric binary regression models under shape constraints. Technical Report BMM(2007). Available at: http://www.stat.lsa.umich.edu/~moulib/bmmtechrep.pdf.
Bazaraa, M.S., Sherali, H.D., Shetty, C.M., 1993. Nonlinear Programming: Theory and Applications. John Wiley & Sons.
Burden, Richard L., Faires, J. Douglas, 2000. Numerical Analysis, 7th edition. Brooks/Cole, ISBN: 978-0-534-38216-2.
Dreze, J., Kingdon, G.G., 2001. School participation in rural India. Rev. Development Econom. 5 (1), 1–24.
Dunson, D.B., 2003. Bayesian isotonic regression for discrete outcomes. Working Paper. Available at: http://ftp.isds.duke.edu/WorkingPapers/03-16.pdf.
Dunson, D.B., Neelon, B., 2003. Bayesian inference on order-constrained parameters in generalized linear models. Biometrics 59 (2), 286–295.
Groeneboom, P., Jongbloed, G., 2003. Density estimation in the uniform deconvolution model. Statist. Neerlandica 57, 136–157.
Groeneboom, P., Jongbloed, G., Wellner, J.A., 2001. Estimation of a convex function: Characterizations and asymptotic theory. Ann. Statist. 29, 1653–1698.
Groeneboom, P., Wellner, J.A., 2001. Computing Chernoff's distribution. J. Comput. Graph. Statist. 10, 388–400.
Handa, S., 2002. Raising primary school enrollment in developing countries: The relative importance of supply and demand. J. Development Econom. 69, 103–128.
Huang, Y., Zhang, C., 1994. Estimating a monotone density from censored observations. Ann. Statist. 24, 1256–1274.
Jongbloed, G., 1998. The iterative convex minorant algorithm for nonparametric estimation. J. Comput. Graph. Statist. 7, 310–321.
Lavy, V., 1996. School supply constraints and children's educational outcomes in rural Ghana. J. Development Econom. 51, 291–314.
Magnac, T., Maurin, E., 2004. Partial identification in monotone binary models: Discrete regressors and interval data. Working Paper. Available at: http://www.crest.fr/doctravail/document/2004-11.pdf.
Magnac, T., Maurin, E., 2007. Identification and information in monotone binary models. J. Econometrics 139 (1), 76–104.
Mammen, E., Yu, K., 2007. Additive isotone regression. In: Asymptotics: Particles, Processes and Inverse Problems: Festschrift for Piet Groeneboom. IMS, Beachwood, OH, USA, pp. 179–195.
Manski, C.F., 1988. Identification of binary response models. J. Amer. Statist. Assoc. 83, 729–738.
Manski, C.F., Tamer, E., 2002. Inference in regressions with interval data on a regressor or outcome. Econometrica 70, 519–546.
McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. In: Monographs on Statistics and Applied Probability, vol. 37. Chapman and Hall, London.
Murphy, S.A., Van der Vaart, A.W., 1997. Semiparametric likelihood ratio inference. Ann. Statist. 25, 1471–1509.
Pal, J., Banerjee, M., 2008. Estimation of smooth regression functions in monotone response models. J. Statist. Plann. Inference 138 (10), 3125–3143.
Politis, D.M., Romano, J.P., Wolf, M., 1999. Subsampling. Springer–Verlag, New York.
Robertson, T., Wright, F.T., Dykstra, R.L., 1988. Order Restricted Statistical Inference. Wiley, New York.
Sen, B., Banerjee, M., Woodroofe, M.B., 2008. Inconsistency of bootstrap: The Grenander estimator. Available at: http://www.stat.lsa.umich.edu/~moulib/Grenboots.pdf.
Silvapulle, M.J., Sen, P.K., 2004. Constrained Statistical Inference, in: Wiley Series in Probability and Statistics.
Wellner, J., 2003. Gaussian white noise models: Some results for monotone functions. In: Kolassa, J.E., Oakes, D. (Eds.), Crossing Boundaries: Statistical Essays in Honor of Jack Hall. In: IMS Lecture Notes-Monograph Series, vol 43. pp. 87–104.

Please cite this article in press as: Banerjee, M., et al., Semiparametric binary regression models under shape constraints with an application to Indian schooling data. Journal of Econometrics (2009), doi:10.1016/j.jeconom.2008.11.002