Semiparametric Binary Regression Models under Shape Constraints with an Application to Indian Schooling Data

Moulinath Banerjee, Debasri Mukherjee and Santosh Mishra

University of Michigan, Western Michigan University and Oregon State University

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 ∆ 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 (∆ = 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 χ² 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 χ² 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.

Keywords: convex minorants, likelihood ratio statistic, school attendance, semiparametric binary regression.

JEL classification no: C1

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.

---

1Research supported in part by National Science Foundation grants DMS-0306235 and DMS-0705288
If $\Delta$ is the indicator of the outcome and $X$ is a set of ($d$–dimensional) covariates believed to influence $\Delta$, one can write $\tilde{g}(\mu(X)) = \beta^T X$, 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 in addition to $X$, we have an additional (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,

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

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 $\psi$ 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, based on i.i.d. observations $\{\Delta_i, X_i, Z_i\}_{i=1}^n$ from the distribution of $\Delta, X, Z$. 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) $\hat{H}_0 : \psi(z_0) = \theta_0$ for some fixed point $z_0$, using the likelihood ratio statistic (henceforth LRS).

The key contributions of our approach are two–fold. The first is the break from the more conventional smoothness assumptions on the nonparametric component $\psi$; indeed, our smoothness assumptions are minimal as we only require the function to be continuously differentiable. Rather, we impose a shape constraint on $\psi$ 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 $\beta$ and $\psi$, 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

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

2
elaborate on this in what follows.

We show that the LRS for testing $H_0$ has a limiting $\chi^2_d$ 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 $D$; "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) $D$, 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 $\beta$ (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 $\chi^2$ quantiles. In contrast, note that inferences for $\beta$ and $\psi$ (equivalently $\mu$) 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, Romano and Wolf (1999)) or the $m$ out of $n$ bootstrap (the usual Efron–type bootstrap will not work in this situation (Sen, Banerjee and Woodroofe (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, Wright and Dykstra (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 $\psi$ in our model) obtained through such regression cannot be analyzed using standard CLTs as they are highly non–linear functionals of the empirical distribution3; 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

---

2This limit distribution does not belong to the $\chi^2_1$ family but can be thought of as an analogue of the $\chi^2_1$ 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 $\psi$ at rate $n^{1/3}$ in this problem, while the rate of convergence of $\beta$ is $\sqrt{n}$.

3This 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.
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 (\( \Delta \) 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 (Section 6).

2 COMPUTING MLES AND LIKELIHOOD RATIOS

The density function of the random vector \((\Delta, X, Z)\) 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 \(\text{Leb} \times \mu\) where \(\text{Leb}\) denotes Lebesgue measure on \([0, \infty)\) and \(\mu\) is some measure defined on \(\mathbb{R}^d\). We construct the likelihood function for the data, as:
\[
L_n(\beta, \psi; \{\Delta_i, X_i, Z_i\}_{i=1}^n) = \prod_{i=1}^n \mu(X_i, Z_i)^{\Delta_i} (1 - \mu(X_i, Z_i))^{1-\Delta_i} f(Z_i, X_i).
\]

In what follows, we denote the true underlying values of the parameters \((\beta, \psi)\) by \((\beta_0, \psi_0)\). 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

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

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 \( R_i(\beta) = \beta^T X(i) \). Denote the vector \((u_1, u_2, \ldots, u_n)\) by \( u \) and define the function \( g \) as:
\[
g(\beta, u) \equiv -l_n(\beta, u) = \sum_{i=1}^n \phi(\Delta(i), R_i(\beta), u_i) \text{ where } \phi(\delta, r, u) = -\delta \log \tilde{h}(r + u) - (1 - \delta) \log(1 - \tilde{h}(r + u)).
\]

Next, for a fixed \( \beta \), let \( \hat{u}_n^{(\beta)} = \arg\min_{u:u_1 \leq u_2 \leq \ldots \leq u_n} g(\beta, u) \). Define \( \hat{\psi}_n^{(\beta)} \) to be the (unique) right–continuous increasing step function that assumes the value \( \hat{u}_{i,n} \) (the \( i \)'th component of \( \hat{u}_n \)) at the point \( Z(i) \) and has no jump points outside of the set \( \{Z(i)\}_{i=1}^n \).

2.1 The Likelihood Ratio Statistic for Testing the Value of \( \beta \)

The likelihood ratio statistic for testing \( H_0: \beta = \beta_0 \) is given by:
\[
\text{lrtbeta}_n = 2 \left( l_n(\hat{\beta}_n, \hat{\psi}_n) - l_n(\beta_0^{(\beta_0)}) \right).
\]  

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 \( g(\beta, u) \) and is based on the following proposition.

**Proposition 1:** Let \( f(\gamma_1, \gamma_2) \) be a real–valued function defined on \( \mathbb{R}^{k_1} \times \mathbb{C} \) where \( \gamma_1 \) varies in \( \mathbb{R}^{k_1} \) and \( \gamma_2 \in \mathbb{C} \), where \( \mathbb{C} \) 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 \((\gamma_1^*, \gamma_2^*)\). Consider the following updating algorithm. Start at an arbitrary point \((\gamma_1^0, \gamma_2^0)\) in \( \mathbb{R}^{k_1} \times \mathbb{C} \). Having defined \((\gamma_1^m, \gamma_2^m)\) at stage \( m \) (\( m \geq 0 \)), set \( \gamma_2^{m+1} \equiv \arg\min_{\gamma_2 \in \mathbb{C}} f(\gamma_1^m, \gamma_2) \) and \( \gamma_1^{m+1} \) as the (unique) solution to \( \partial f(\gamma_1, \gamma_2^m) = 0 \). Then, irrespective of the starting value, the sequence of points \( \{\gamma_1^m, \gamma_2^m\}_{m \geq 0} \) converges to \((\gamma_1^*, \gamma_2^*)\).

**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).  

\footnote{Additional remark: Note that the step of updating \( \gamma_1^m \) to \( \gamma_1^{m+1} \) which involves solving \( \partial f(\gamma_1, \gamma_2^{m+1}) = 0 \) is also a minimization step. Since \( f(\gamma_1, \gamma_2^{m+1}) \) is a continuously differentiable strictly convex function of \( \gamma_1 \), it is
Consider first, the computation of the unconstrained MLEs $(\hat{\beta}_n, \hat{\psi}_n)$. The function $g$ is defined on $\mathbb{R}^d \times \hat{C}$, where $\hat{C} \equiv \{ u = (u_1, u_2, \ldots, u_n) : u_1 \leq u_2 \leq \ldots \leq u_n \}$ is a closed convex cone in $\mathbb{R}^n$. Setting $f$ in Proposition 1 to be $g$, $\gamma_1 = \beta$ and $\gamma_2 = u$, it is easy to check that $g$ is a continuously differentiable and strictly convex function that attains a unique minimum at $(\hat{\beta}_n, \hat{u}_n)$. 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 $p \geq 0$ of the algorithm, current estimates $(\hat{\beta}_n^{(p)}, \hat{u}_n^{(p)})$ are available. At Stage $p + 1$, first update the second component to $\hat{u}_n^{(p+1)}$ by minimizing $g(\hat{\beta}_n^{(p)}, u)$ over $u \in \hat{C}$, using the modified iterative convex minorant algorithm (MICM) due to Jongbloed (1998). Note that $\hat{u}_n^{(p+1)}$ is precisely the vector $\{ \hat{\psi}_n^{(\beta)}(Z_i) \}_{i=1}^{n}$, for $\beta = \hat{\beta}_n^{(p)}$.

**Step 2b.** Having updated to $\hat{u}_n^{(p+1)}$, next update $\hat{\beta}_n^{(p)}$ to $\hat{\beta}_n^{(p+1)}$ by solving $(\partial / \partial \beta) g(\beta, \hat{u}_n^{(p+1)}) = 0$ 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

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

then stop and declare $(\hat{\beta}_n^{(p+1)}, \hat{u}_n^{(p+1)})$ 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 $\hat{\psi}_n^{(\hat{\beta}_n^{(p)})}$, the MLE of $\psi$ for a fixed $\beta$. In particular, recall that $\hat{\psi}_n^{(\hat{\beta}_0)}$ is the MLE of $\psi$ under $H_0 : \beta = \hat{\beta}_0$.

**Characterizing and computing $\hat{\psi}_n^{(\beta)}$:** This is characterized by the vector $\hat{u}_n^{(\beta)} = (\hat{u}_1^{(\beta)}, \hat{u}_2^{(\beta)}, \ldots, \hat{u}_n^{(\beta)})$ that minimizes $g(\beta, u)$ over all $u_1 \leq u_2 \leq \ldots \leq u_n$. Without loss of generality one can assume that $\Delta_i = 1$ and $\Delta_m = 0$. 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.
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 \( x_0 < x_1 < \ldots < x_k \), 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 \( \text{slogcm} \{(x_i, y_i)\}_{i=0}^n \).

The solution \( \hat{u}_n^{(\beta)} \) can be viewed as the slope of the greatest convex minorant (slogcm) of a random function defined in terms of \( \hat{u}_n^{(\beta)} \) 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 \( g(\beta, u) \) in the following discussion by \( \xi(u) \) (suppressing the dependence on \( \beta \)) and \( \hat{u}_n^{(\beta)} \) by \( \hat{u} = (\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n) \). For \( 1 \leq i \leq n \), set \( d_i = \nabla v_i \xi(\hat{u}) \). Define the function \( \eta \) as follows:

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

It can be shown that \( \hat{u} \) minimizes \( \eta \) subject to the constraints that \( u_1 \leq u_2 \leq \ldots \leq u_n \) (see Section 6.2 in the appendix for the details) and hence furnishes the isotonic regression of the function \( h(i) = \hat{u}_i - \nabla v_i \xi(\hat{u}) d_i^{-1} \) on the ordered set \( \{1, 2, \ldots, n\} \) with weight function \( d_i = \nabla v_i \xi(\hat{u}) \). It is well known that the solution \( \hat{u} = (\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n) \) = slogcm \( \{\sum_{j=1}^{i} d_i, \sum_{j=1}^{i} h(i) d_i\}_{i=0}^{n} \). 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 \( \hat{u} \) is unknown, an iterative scheme is resorted to. For a fixed vector \( v \equiv (v_1, v_2, \ldots, v_n) \in C \), set \( d_{v,i} = \nabla v_i \xi(\nu) \) and define the function \( \eta_v(u) \equiv \sum_{i=1}^{n} \left[u_i - (v_i - \nabla v_i \xi(\nu) d_{v,i}^{-1})\right]^2 d_{v,i} \). Pick an initial guess for \( \hat{u} \), say \( u^{(0)} \in C \), set \( v = u^{(0)} \) and compute \( u^{(1)} \) by minimizing \( \eta_v(u) \) over \( C \); then, set \( v = u^{(1)} \), obtain \( u^{(2)} \) by minimizing \( \eta_v(u) \) again, and proceed thus, until convergence. Generally, with \( v = u^{(j)} \), we have: \( u^{(j+1)} = \text{slogcm} \left\{ \sum_{i=1}^{j} d_{v,i}, \sum_{i=1}^{j} (v_i - \nabla v_i \xi(\nu) d_{v,i}^{-1}) d_{v,i}\right\}_{i=0}^{n} \).

**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 \( \xi \) are satisfied to a pre–specified degree of tolerance. For a discussion of these conditions, see Section 6.2.
2.2 The Likelihood Ratio Statistic For Testing the Value of $\psi$ 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 $\beta$ and $\psi$, 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 $\beta$, one can compute $\hat{\psi}_{n,0} = \arg\max\psi_{\psi(z_0) = \theta_0} l_n(\beta, \psi)$. Then, $\hat{\beta}_{n,0} = \arg\max_{\beta} l_n(\beta, \hat{\psi}_{n,0})$ 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:

$$lrtpsi_n = 2 \left( l_n(\hat{\beta}_n, \hat{\psi}_n) - l_n(\hat{\beta}_{n,0}, \hat{\psi}_{n,0}) \right).$$  \hspace{1cm} (2.6)

Note that the monotone function $\hat{\psi}_{n,0}^{(\beta)}$ is identifiable only up to its values at the $Z(\beta)_i$’s (and at the fixed point $z_0$ where it is required to equal $\theta_0$) and we identify this function with the vector $\hat{\psi}_{n,0}^{(\beta)} \equiv (\hat{\psi}_{1,n,0}^{(\beta)}, \hat{\psi}_{2,n,0}^{(\beta)}, \ldots, \hat{\psi}_{m,n,0}^{(\beta)})$ where $\hat{\psi}_{i,n,0}^{(\beta)} = \hat{\psi}_{n,0}^{(\beta)}(Z(i))$. 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 $\psi$ that satisfies $\tilde{H}_0$ will have: $\psi(Z(m)) \leq \theta_0 \leq \psi(Z(m+1))$. Define $\tilde{C}_0$ to be the closed convex subset of $\mathbb{R}^n$ comprising all vectors $\mathbf{u}$ with $u_1 \leq u_2 \leq \ldots \leq u_m \leq \theta_0 \leq u_{m+1} \leq \ldots \leq u_n$. If $\tilde{\mathbf{u}}_{n,0} \equiv (\tilde{u}_{1,n,0}, \tilde{u}_{2,n,0}, \ldots, \tilde{u}_{n,n,0})$ denotes the vector $\{\hat{\psi}_{n,0}(Z(i))\}_{i=1}^n$, then $(\beta_{n,0}, \tilde{\mathbf{u}}_{n,0}) = \arg\min_{\beta \in \mathbb{R}^d, \mathbf{u} \in \tilde{C}_0} g(\beta, \mathbf{u})$. Since the function $g$ is a continuously differentiable strictly convex function defined on the closed convex set $\mathbb{R}^d \times \tilde{C}_0$ and assumes a unique minimum at $(\beta_{n,0}, \tilde{\mathbf{u}}_{n,0})$, 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 $(\hat{\beta}_{n,0}^{(0)}, \tilde{\mathbf{u}}_{n,0}^{(0)})$. Also, set an initial tolerance level $\eta > 0$, small.

**Step 2a.** At Stage $p \geq 0$ of the algorithm, current estimates $(\hat{\beta}_{n,0}^{(p)}, \tilde{\mathbf{u}}_{n,0}^{(p)})$ are available. At Stage $p+1$, first update the second component to $\tilde{\mathbf{u}}_{n,0}^{(p+1)}$ by minimizing $g(\hat{\beta}_{n,0}^{(p)}, \mathbf{u})$ over $\mathbf{u} \in \tilde{C}_0$. Note that $\tilde{\mathbf{u}}_{n,0}^{(p+1)}$ 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 $(\partial/\partial \beta) g(\beta, \tilde{\mathbf{u}}_{n,0}^{(p+1)}) = 0$ using, say, the Newton–Raphson method.

**Step 3.** *(Checking convergence)* If $\left| g(\hat{\beta}_{n,0}^{(p+1)}, \tilde{\mathbf{u}}_{n,0}^{(p+1)}) - g(\hat{\beta}_{n,0}^{(p)}, \tilde{\mathbf{u}}_{n,0}^{(p)}) \right| \leq \eta$, then stop and declare $(\hat{\beta}_{n,0}^{(p+1)}, \tilde{\mathbf{u}}_{n,0}^{(p+1)})$ 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 $\beta$.  

8
Characterizing $\hat{\psi}_n^{(\beta)}$: Finding $\hat{\psi}_n^{(\beta)}$ amounts to minimizing $g(\beta, u) = \sum_{i=1}^n \phi(\Delta(i), R_i(\beta), u_i)$ over all $u_1 \leq u_2 \leq \ldots \leq u_m \leq \theta_0 \leq u_{m+1} \leq \ldots \leq u_n$. For the remainder of this discussion, we denote the minimizing vector $\hat{u}_{n,0}^{(\beta)}$ by $\hat{u}^{(0)}$. Finding $\hat{u}^{(0)}$ 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 $u_1 \leq u_2 \leq \ldots \leq u_m \leq \theta_0$, 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 $\theta_0 \leq u_{m+1} \leq u_{m+2} \leq \ldots \leq u_n$.

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 $g_1(\beta, u_1) \equiv (u_1, u_2, \ldots, u_m)$ over $\hat{C}_1$, the closed convex cone in $\mathbb{R}^m$ defined as $\{u_1 : u_1 \leq u_2 \leq \ldots \leq u_m\}$, to obtain $\hat{u}_1 \equiv (\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_m)$. Then, $(\hat{u}_1^{(0)}, \hat{u}_2^{(0)}, \ldots, \hat{u}_m^{(0)}) = (\hat{u}_1 \land \theta_0, \hat{u}_2 \land \theta_0, \ldots, \hat{u}_m \land \theta_0)$. The minimization of $g_1(\beta, \cdot)$ over $\hat{C}_1$ 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 $(\hat{u}_{m+1}^{(0)}, \hat{u}_{m+2}^{(0)}, \ldots, u_n^{(0)}) = (\tilde{u}_{m+1} \lor \theta_0, \tilde{u}_{m+2} \lor \theta_0, \ldots, \tilde{u}_n \lor \theta_0)$ where $(\tilde{u}_{m+1}, \tilde{u}_{m+2}, \ldots, \tilde{u}_n) = \text{argmin}_{u_{m+1} \leq u_{m+2} \leq \ldots \leq u_n} g_2(\beta, u_{m+1}, u_{m+2}, \ldots, u_n)$ and uses the MICM, as in [1]. Finally $\hat{u}^{(0)} = (\hat{u}_1^{(0)}, \hat{u}_2^{(0)}, \ldots, \hat{u}_n^{(0)})$.

3 ASYMPTOTIC RESULTS

In this section we present asymptotic results for the estimation of $\beta$ and $\psi$. 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 $\log \frac{\mu(X, Z)}{1 - \mu(X, Z)} = \beta^T X + \psi(Z)$. The above display is equivalent to writing:

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

(3.7)

The parameter space for $\beta$ 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 nondecreasing cadlag (i.e. right-continuous with left-hand limits) functions from $[0, \tau]$ to $[0, M]$ where $M$ is some large positive constant. Let $(\beta_0, \Lambda_0)$ denote the true model parameters (thus, $\Lambda_0 = e^{\psi_0}$). We make the following assumptions:

(A.1) The true regression parameter $\beta_0$ is an interior point of $\mathcal{B}$.

(A.2) The covariate $X$ has bounded support. Hence, there exists $x_0$ such that $P(\|X\| \leq x_0) = 1$. Also $E(\text{Var}(X | Z))$ is positive definite with probability one.

(A.3) Let $\Lambda_0(0) = 0$. 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, $\Lambda_0$ is continuously differentiable on $[\sigma, \tau]$ with derivative $\lambda_0$ bounded away from 0 and from $\infty$. 

9
(A.5) The marginal density of $Z$, which we denote by $f_Z$, is continuous and positive on $[\sigma, \tau]$. 

(A.6) The function $h^{**}$ defined below in (3.8) defined below has a version which is differentiable componentwise with each component possessing a bounded derivative on $[\sigma, \tau]$.

**Remarks:** The boundedness of $B$ 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 $\Lambda_0$ 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 $\beta$ in this model. The log density function for the vector $(\Delta, Z, X)$ is given by:

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

The ordinary score function for $\beta$ in this model is:

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

where $Q((\delta, z, x); \beta, \Lambda) = \delta (\Lambda(z))^{-1} - (e^{\beta^T x})(1 + \Lambda(z) e^{\beta^T x})^{-1}$. The score function for $\Lambda$ is a linear operator acting on the space of functions of bounded variation on $[\sigma, \tau]$ and has the form $\hat{l}_\Lambda(\beta, \Lambda)(h(\cdot))(\delta, z, x) = h(z) Q((\delta, z, x); \beta, \Lambda)$. 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 $t \geq 0$ where $h$ is a non-decreasing non-negative function on $[\sigma, \tau]$. Computing $B_{\Lambda}(h) = \frac{\partial}{\partial t} l_{\beta, \Lambda+t h}(\delta, x, z) \big|_{t=0}$, we get $B_{\Lambda}(h) = h(z) Q((\delta, z, x); \beta, \Lambda)$. The linear operator $B_{\Lambda}$ 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 $\beta$ at the true parameter values $(\beta_0, \Lambda_0)$, which we will denote by $\hat{\bar{l}}$ for brevity, is given by

$$ \hat{\bar{l}} = \hat{l}_\beta(\beta_0, \Lambda_0) - \hat{l}_\Lambda(\beta_0, \Lambda_0)h^* $$

for functions $h^* = (h_1^*, h_2^*, \ldots, h_d^*)$ of bounded variation, such that $h_i^*$ minimizes the distance $E_{\beta_0, \Lambda_0}(\hat{l}_{\beta,i}(\beta_0, \Lambda_0) - \hat{l}_{\Lambda}(\beta_0, \Lambda_0) h(\cdot))^2$, for $h$ varying in the space of functions of bounded variation on $[\sigma, \tau]$. Here, $\hat{l}_{\beta,i}(\beta_0, \Lambda_0) = x^{(i)} \Lambda(z) Q((\delta, z, x); \beta_0, \Lambda_0)$ is the $i$th component of the ordinary score function for $\beta$ (and $x^{(i)}$ is the $i$th component of $x$). It is not difficult to see that $h_i^*$ must satisfy $E[B_{\Lambda_0}(h)(\hat{l}_{\beta,i}(\beta_0, \Lambda_0) - \hat{l}_{\Lambda_0}(h^*_i))] = 0$, for all $h$. This simplifies to
where $r$ and $s$. Thus, $z$.

The assumption (A.2) that $E(Q^2((\Delta, Z, X); \beta_0, \Lambda_0) | Z) = 0$, whence

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

In vector notation we can therefore write

$$h^*(Z) = \Lambda_0(Z) h^{**}(Z) \equiv \Lambda_0(Z) \frac{E_{\beta_0, \Lambda_0}(X Q^2((\Delta, Z, X); \beta_0, \Lambda_0) | Z)}{E_{\beta_0, \Lambda_0}(Q^2((\Delta, Z, X); \beta_0, \Lambda_0) | Z)}.$$ \hspace{1cm} (3.8)

The assumption (A.2) that $E(Var(X | Z))$ is positive definite ensures that $\hat{I}$, the efficient score function for $\beta$, is not identically zero, wheres the efficient information $I_0 = \text{Disp}(\hat{I}) = E_{\beta_0, \Lambda_0}(\hat{I}^T I)$ is positive definite (Note that $E_{\beta_0, \Lambda_0}(\hat{I}) = 0$). This entails that the MLE of $\beta$ will converge at $\sqrt{n}$ rate to $\beta_0$ 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:

$$\text{lrtbeta}_0 = 2 \log \frac{\arg\max_{\Lambda(z_0)=\tilde{\theta}_0} l_n(\beta, \Lambda)}{\arg\max_{\beta=\beta_0, \Lambda(z_0)=\tilde{\theta}_0} l_n(\beta, \Lambda)},$$ \hspace{1cm} (3.9)

Thus,

$$\text{lrtbeta}_0 = 2 l_n(\hat{\beta}_{n,0}, \hat{\Lambda}_{n,0}) - 2 l_n(\beta_0, \hat{\Lambda}_{n,0}^{(\beta_0)}),$$

where $\hat{\Lambda}_{n,0} = \exp(\hat{\psi}_{n,0})$ and $\hat{\Lambda}_{n,0}^{(\beta_0)} = \exp(\hat{\psi}_{n,0}^{(\beta_0)})$. 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 $\text{lrtbeta}_n$ as defined in (2.4) and $\text{lrtbeta}_0$ above.

**Theorem 3.1** Under Conditions (A.1) – (A.7), both $\tilde{\beta}_n$ and $\tilde{\beta}_{n,0}$ are asymptotically linear in the efficient score function and have the following representation:

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

and

$$\sqrt{n}(\tilde{\beta}_{n,0} - \beta_0) = \frac{1}{\sqrt{n}} \tilde{I}_0^{-1} \sum_{i=1}^n \tilde{I}(\Delta_i, Z_i, X_i) + s_n \hspace{1cm} (3.11)$$

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

Furthermore,

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

while

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

It follows that both $\text{lrtbeta}_n$ and $\text{lrtbeta}_0$ are asymptotically distributed like $\chi^2_d$. 

11
We next state asymptotic results concerning the nonparametric component of the model. In order to do so, we introduce the following processes. For positive constants \( c \) and \( d \) define the process \( X_{c,d}(z) := cW(z) + d z^2 \), 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 \( h \leq 0 \) and \( g_{c,d,L}(h) \) denotes 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 \( g_{c,d}^0(h) = (g_{c,d,L}(h) \wedge 0)1(h \leq 0) + (g_{c,d,R}(h) \vee 0)1(h > 0) \). 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 \( L_2[-M,M] \times L_2[-M,M] \):

\[
(g_{c,d}(h), g_{c,d}^0(h)) \overset{D}{=} \left( c(d/c)^{1/3} g_{1,1} \left( (d/c)^{2/3} h \right), c(d/c)^{1/3} g_{1,1}^0 \left( (d/c)^{2/3} h \right) \right).
\]

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:

\[
U_n(h) = n^{1/3} \left( \psi_n^{(h)}(z_0 + h_n^{-1/3} - \psi_0(z_0)) \right) \quad \text{and} \quad V_n(h) = n^{1/3} \left( \psi_n^{(h)}(z_0 + h_n^{-1/3} - \psi_0(z_0)) \right).
\]

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

**Theorem 3.2** Define \( C(z_0) = \int \frac{e^{\psi^2_{0,x} + \psi_0(z_0)}}{1 + e^{\psi^2_{0,x} + \psi_0(z_0)}} f(z_0, x) \, d\mu(x) \). 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 \( \psi_0' \) is the derivative of \( \psi_0 \). 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

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

where \( Z \equiv \arg\min_{h \in \mathbb{R}} (W(h) + h^2) \) 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

\[
g_{1,1}(0) \equiv d^2 Z
\]

(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 \( Z \) 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 \( Z \) is non-Gaussian and symmetric about 0. It can, in fact, be shown that

\[
n^{1/3} (\hat{\psi}_n(z_0) - \psi_0(z_0)) \rightarrow_d (8a^2 b)^{1/3} Z
\]

where \( \hat{\psi}_n \equiv \hat{\psi}_{\beta_n} \) is the unconstrained MLE of \( \psi \). 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 \( \beta_0 \) at a faster rate than \( n^{1/3} \), the convergence rate for \( \hat{\psi}_{\beta_n} \). Since the quantiles of \( Z \) 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 \( \psi_0 \) 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, Romano and Wolf (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 \( \psi_0 \) at a pre-fixed point of interest. Our next theorem is crucial for this purpose.

**Theorem 3.3** The likelihood ratio statistic for testing \( H_0 : \psi(z_0) = \theta_0 \), as defined in (2.6), converges in distribution to \( \mathbb{D} \) where

\[
\mathbb{D} = \int (g_{1,1}(z))^2 - (g_{0,1}(z))^2 \, dz.
\]

The random variable \( \mathbb{D} \) can be considered to be a non-regular analogue of the usual \( \chi^2_1 \) random variable, in the sense that just as the \( \chi^2_1 \) 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.

Construction of confidence sets for parameters of interest via likelihood ratio
based inversion: 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 the baseline conditional probability function, \( \Delta \) based inversion:

conditions A.1 – A.6 (with \( X \) fashion as before; i.e. \( \tilde{\psi} \)). The computation of joint confidence sets is also of importance in certain applications. Thus, one sets for the conditional probabilities in the data analysis example in Section 4. (2.6), but using the covariates \( \tilde{\psi} \) 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 \( \tilde{S}_{\psi}(z_0)/(1 + e^{\tilde{S}_{\psi}(z_0)}) \). 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 \( M_n \equiv \max(\text{lrtpsi}(z_0)(\theta_0), \text{lrtpsi}(z_1)(\theta_1)) \) where \( \text{lrtpsi}(z_0)(\theta_0) \) is the likelihood ratio statistic for testing \( \psi(z_0) = \theta_0 \) and \( \text{lrtpsi}(z_1)(\theta_1) \), the likelihood ratio statistic for testing \( \psi(z_1) = \theta_1 \). It can be shown that when the null hypothesis is true, \( \text{lrtpsi}(z_0)(\theta_0) \) and \( \text{lrtpsi}(z_1)(\theta_1) \) are asymptotically independent and \( M_n \) converges in distribution to \( D(2) \equiv \max(D_1, D_2) \) where \( D_1 \) and \( D_2 \) are identical copies of \( D \). 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 \( (\theta, \theta') \) (with \( \theta \leq \theta' \) and let \( \text{lrtpsi}(z_0)(\theta) \) and \( \text{lrtpsi}(z_1)(\theta') \) denote, respectively, the likelihood ratio statistics for testing \( \psi(z_0) = \theta \) and \( \psi(z_1) = \theta' \), 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:

\[
\tilde{S}_{\psi}(z_0), \tilde{\psi}(z_1) = \{(\theta, \theta') : \theta \leq \theta', \max(\text{lrtpsi}(z_0)(\theta), \text{lrtpsi}(z_1)(\theta')) \leq q(D(2), 1 - \alpha)\}.
\]
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 $\mathcal{S}_{\tilde{\psi}_0(z_0)} = \{ \theta : \text{lrt}_{\tilde{\psi}_0(z_0)}(\theta) \leq q(D^{(2)}, 1 - \alpha) \}$ and $\mathcal{S}_{\tilde{\psi}_0(z_1)} = \{ \theta' : \text{lrt}_{\tilde{\psi}_0(z_1)}(\theta') \leq q(D^{(2)}, 1 - \alpha) \}$. Then:

$$
\tilde{S}_{\psi_0(z_0), \tilde{\psi}_0(z_1)} = (\mathcal{S}_{\psi_0(z_0)} \times \mathcal{S}_{\psi_0(z_1)}) \cap C_2
$$

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

$$
\{(e^\theta/(1 + e^\theta), e^\theta/(1 + e^\theta)) : (\theta, \theta') \in \tilde{S}_{\psi_0(z_0), \tilde{\psi}_0(z_1)}\}.
$$

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 $z_i$'s increases, owing to the finite sample dependence among the pointwise likelihood ratio statistics.

Confidence sets for the finite dimensional regression parameter $\beta_0$ can be constructed in the usual fashion as: $\{ \beta : \text{lrt}_{\text{beta}}(\beta) \leq q_{\chi^2_{d_1-1}, 1-\alpha} \}$, where lrt_{beta}(\beta) is the likelihood ratio statistic for testing the null hypothesis that the true regression parameter is $\beta$ (see (2.7)), and $q_{\chi^2_{d_1-1}, 1-\alpha}$ is the $(1 - \alpha)$th quantile of the $\chi^2_{d_1}$ 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 $\eta_{10}$. Let $d_1$ and $d_2$ denote the dimensions of $\eta_1$ and $\eta_2$ respectively. To test $\Pi_0 : \eta_1 = \eta_{10}$, the log-likelihood function $l_n(\beta, \psi)$ is maximized over all $\beta$ of the form $(\eta_{10}, \eta_2)$ (where $\eta_2$ varies freely in $\mathbb{R}^{d_2}$) and $\psi$ monotone increasing. If we identify $\psi$, as before, with the vector $u = \{ \psi(Z(i)) \}_{i=1}^n$, then, $\tilde{g}(\eta_{12}, u) \equiv -l_n((\eta_{10}, \eta_{2}), \psi)$ is a continuously differentiable strictly convex function defined on $\mathbb{R}^{d_2} \times C$ and its minimizer can be obtained using Proposition 1. If $(\hat{\eta}_{12}, \hat{u}_{i, n}^{(\eta_{10})})$ denotes the minimizer of $\tilde{g}$, then the constrained MLEs of $(\beta, \psi)$ under $\Pi_0$ are: $((\eta_{10}, \hat{\eta}_{12}), \hat{\psi}_{i, n}^{(\eta_{10})})$ where $\hat{\psi}_{i, n}^{(\eta_{10})}$ is the (unique) right–continuous increasing step function that assumes the value $\tilde{u}_{i, n}^{(\eta_{10})}$ (the $i$th component of $\hat{u}_{i, n}^{(\eta_{10})}$) at the point $Z(i)$ and has no jump points outside of the set $\{ Z(i) \}$. The likelihood ratio statistic for testing $H_0$ is then given by:

$$
2 \left[ l_n(\beta_{n, \hat{\psi}_n}) - l_n((\eta_{10}, \hat{\eta}_{12}), \hat{\psi}_{i, n}^{(\eta_{10})}) \right]
$$

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

**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 $\hat{h}$ satisfies Condition (C) described in Section 2. Thus, for any $\hat{h}$ satisfying the concavity constraints, (a) the likelihood ratio statistic for testing $\beta = \beta_0$ as described in (2.4) converges to a $\chi^2_{d_1}$ distribution and (b) the likelihood ratio statistic for testing $H_0 : \psi(z_0) = \theta_0$ as described in (2.6) converges in distribution to $D_1$ (when $H_0$ is true). Confidence sets for $\beta, \psi(z_0)$ and $\mu(x_0, z_0)$ as well as a sub-vector of $\beta$ 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 $\beta_0$ and pointwise estimation of $\psi_0$ (equivalently $\Lambda_0$) 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 $\beta_0$ (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 $\beta_0$ (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 ($\beta_0, \Lambda_0$) 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 $\beta_0$ 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 $\beta$ 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

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>$n = 500$</th>
<th>$n = 1000$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CI(\beta)</td>
<td>CI(\Lambda(z_0))</td>
</tr>
<tr>
<td>-0.5</td>
<td>94.8</td>
<td>95.9</td>
</tr>
<tr>
<td></td>
<td>0.688</td>
<td>0.454</td>
</tr>
<tr>
<td>-0.25</td>
<td>95.5</td>
<td>94.4</td>
</tr>
<tr>
<td></td>
<td>0.684</td>
<td>0.444</td>
</tr>
<tr>
<td>0.0</td>
<td>95.6</td>
<td>95.2</td>
</tr>
<tr>
<td></td>
<td>0.686</td>
<td>0.439</td>
</tr>
<tr>
<td>0.25</td>
<td>95.0</td>
<td>94.9</td>
</tr>
<tr>
<td></td>
<td>0.682</td>
<td>0.443</td>
</tr>
<tr>
<td>0.5</td>
<td>95.0</td>
<td>94.5</td>
</tr>
<tr>
<td></td>
<td>0.691</td>
<td>0.454</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>$n = 500$</th>
<th>$n = 1000$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CI(\beta)</td>
<td>CI(\Lambda(z_0))</td>
</tr>
<tr>
<td>-0.5</td>
<td>94.8</td>
<td>95.9</td>
</tr>
<tr>
<td></td>
<td>0.688</td>
<td>0.454</td>
</tr>
<tr>
<td>-0.25</td>
<td>95.5</td>
<td>94.4</td>
</tr>
<tr>
<td></td>
<td>0.684</td>
<td>0.444</td>
</tr>
<tr>
<td>0.0</td>
<td>95.6</td>
<td>95.2</td>
</tr>
<tr>
<td></td>
<td>0.686</td>
<td>0.439</td>
</tr>
<tr>
<td>0.25</td>
<td>95.0</td>
<td>94.9</td>
</tr>
<tr>
<td></td>
<td>0.682</td>
<td>0.443</td>
</tr>
<tr>
<td>0.5</td>
<td>95.0</td>
<td>94.5</td>
</tr>
<tr>
<td></td>
<td>0.691</td>
<td>0.454</td>
</tr>
</tbody>
</table>
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 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 52\(^{nd}\) round on “Participation in Education” (conducted between July 1995- June 1996). For detailed information regarding this survey, we direct the refer to the following URL site: \texttt{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 \(\Lambda\), as described in (3.7). The monotonicity assumption on \(\Lambda\) 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 govt. of India to eliminate illiteracy.

\(^7\)In the nation wide survey data there is no information on variables needed to construct such indices

\(^8\)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.
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.\(^\text{10}\) Our response variable is whether a child is attending primary school or not ($\Delta = 1$ if attending, 0 otherwise).

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 accord 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.

**TABLE 1A AROUND HERE**

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

\(^{10}\) We consider an equal representation subsample from the bigger sample.
versa.\textsuperscript{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.\textsuperscript{12} Thus, our modified model includes only standard of living, SCST and HEADLIT. Table 1B reports the results for the modified parametric model.

\textbf{TABLE 1B AROUND HERE}

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. Table 2A reports the estimated probabilities of school–attendance and their likelihood–ratio based confidence intervals associated 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.

\textbf{TABLES 2A, 2B AND 2C AROUND HERE}

Some interesting observations can be made jointly from Tables 2A, 2B and 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/nonliterate) populace for each individual group (as expected).
(iii) Non Bimaru females always significantly dominate Bimaru females for each category.
(iv) Non Bimaru males almost always dominate Bimaru males for each category.
(v) NonBimaru males dominate NonBimaru females for both non literate 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

\textsuperscript{11}For brevity we don’t report these results.
\textsuperscript{12}The headlit variable partly captures its effect.
expenditure/standard of living (parametric as well as semiparametric) are presented in Figure 1 and Figure 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 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 underprivileged 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 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 \( \Lambda \). However, since the true function is smooth, it is conceivable that a smooth monotone estimate of \( \Lambda \) 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 \( \beta \) (especially in higher dimensions) is the “inversion” itself. For one dimensional \( \beta \), the convexity of the log–likelihood ratio in \( \beta \) 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 closing, 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: \( \tilde{g}(\mu(X_1, X_2, Z)) = \beta^T X + \psi_1(Z_1) + \psi_2(Z_2) \) where \( \psi_1 \) and \( \psi_2 \) are both monotone. This can be thought of as a generalized additive (semiparametric) model under a monotonicity constraint. While the MLEs of the parameters \( (\beta, \psi_1, \psi_2) \) 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 (2006) on additive isotonic regression suggests certain possibilities. However a treatment of these additive models is outside the scope of this paper.

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, Jongbloed and Wellner (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 \( \mu \) 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.

6 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, BMM(2007).

6.1 Proofs of Theorems 3.1 and 3.2

Proof of Theorem 3.1: For simplicity, we assume that $X$ and $\beta$ are both 1–dimensional. Also, in what follows, $P_{\beta,\Lambda}$ will denote the distribution of $(\Delta, Z, X)$ under parameter value $(\beta, \Lambda)$ and $p_{\beta,\Lambda}$ the corresponding density. Also $p_0$ denotes the density under the true values $(\beta_0, \Lambda_0)$ and $P_0 \equiv P_{\beta_0,\Lambda_0}$. In what follows, we will freely use the facts that (a) $(\hat{\beta}_n, \hat{\Lambda}_n)$ converges to $(\beta_0, \Lambda_0)$ in the product topology $\tau_1 \times \tau_2$ where $\tau_1$ is the Euclidean topology on $\mathbb{R}^d$ and $\tau_2$ is the topology of uniform convergence on compact sets. Also, $\hat{\Lambda}_n, 0$ converges to $\Lambda_0$ under $\tau_2$.

We will use Theorem 3.1 of Murphy and Van der Vaart (1997) to establish that $\text{lrtbeta}_n$, the likelihood ratio statistic for testing $\beta = \beta_0$ is asymptotically $\chi^2_1$ and has the representation (3.12); in the process, we will establish the asymptotically linear representation of the MLE $\hat{\beta}$ in the efficient score function (display (3.10)). We start by constructing the “approximately least favorable” one–dimensional submodels that satisfy conditions (3.6) and (3.7) on Page 1482 of Murphy and Van der Vaart (1997). For parameter values $(\beta, \Lambda)$ we define the corresponding submodel $s_t(\beta, \Lambda) = (t, \Lambda_t(\beta, \Lambda)) \equiv (t, \Lambda + (\theta - t) \phi(\Lambda) h^{**} \circ \Lambda_0^{-1} \circ \Lambda)$.

Here $\phi$ is a function mapping $[0, M]$ into $[0, \infty)$, such that (i) $\phi(y) = y$ on $[\Lambda_0(\sigma), \Lambda_0(\tau)]$, (ii) $y \mapsto \phi(y)/y$ is Lipschitz and (iii) $\phi(y) \leq c (y \wedge M - y)$ for a sufficiently large constant $c$ that depends on $(\theta_0, \Lambda_0)$ only. By the assumption that $[\Lambda_0(\sigma), \Lambda_0(\tau)] \subset (0, M)$ such a function $\phi$ exists. It is not difficult to show that for $t$ sufficiently close to $\theta$, $(t, \Lambda_t(\theta, \Lambda))$ is a valid parameter.

We next compute the scores from these approximately least favorable submodels. Let $p(\delta, z; x; t, \Lambda, \beta)$ denote the density function under parameter value $(t, \Lambda_t(\beta, \Lambda))$ and $l(\delta, z; x; t, \Lambda, \beta)$ denote the log–density. We have $l(\delta, z; x; t, \Lambda, \beta) = \delta \log \Lambda_t(\beta, \Lambda)(z) + \delta t x - \log (1 + \Lambda_t(\beta, \Lambda)(z) e^{tx})$. Define $H(z) = -\phi(\Lambda(z)) h^{**} \circ \Lambda_0^{-1} \circ \Lambda(z)$. Straightforward computations yield that the score
\( \hat{l}(\delta, x, z; t, \Lambda, \beta) = \delta (\Lambda_t(\beta, \Lambda)(z))^{-1}H(z) + \delta x - (1 + \Lambda_t(\beta, \Lambda)(z) e^{tx})^{-1} [H(z) e^{tx} + \Lambda_t(\beta, \Lambda)(z) x e^{tx}] \).

As \((t, \beta, \Lambda)\) converges to \((\beta_0, \beta_0, \Lambda_0)\) (the convergence is taken to be with respect to the product of the Euclidean topology on \(\mathbb{R}^2\) and the topology of uniform convergence on compact subsets of \([\sigma, \tau]\)), it is easy to verify that \(\hat{l}(\delta, x, z; t, \Lambda, \beta) \rightarrow \delta(\Lambda_0(z))^{-1} [-\Lambda_0(z) h^{**}(z)] + \delta x - (1 + \Lambda_0(z) e^{\beta_0 x})^{-1} [-\Lambda_0(z) h^{**}(z) + \Lambda_0(z) x e^{\beta_0 x}]\) almost everywhere, with respect to \(P_0\), the true measure and this is precisely \(\hat{l}(\delta, x, z)\), the efficient score function (as can be checked by arranging terms).

Next, we note that the class of functions \(\hat{l}(\delta, x, z; t, \Lambda, \beta)\) with \(\beta\) and \(t\) ranging in a neighborhood of \(\beta_0\) and \(\Lambda\) ranging over the class of increasing cadlag functions on \([0, \tau]\) and taking values in \([0, M]\) is a uniformly bounded and Donsker class. Furthermore, the class

\[ \mathcal{F} = \{ g(\delta, z, x; t, \Lambda, \beta) = (1/p(\delta, z, x; t, \Lambda, \beta)) (\partial^2/\partial t^2) p(\delta, z, x; t, \Lambda, \beta) \} \]

as \((\beta, t, \Lambda)\) vary in a sufficiently small neighborhood of \((\beta_0, \beta_0, \Lambda_0)\) is a Glivenko–Cantelli class of functions (for the details of these arguments, see BMM (2007)). Furthermore \(g(\delta, z, x; t, \Lambda, \beta) \rightarrow g(\delta, z, x; \beta_0, \Lambda_0, \beta_0)\) almost surely under \(P_0\) (the true underlying distribution generating the data), as \((t, \Lambda, \beta) \rightarrow (\beta_0, \Lambda_0, \beta_0)\). It now follows from the discussion on Pages 71-77 of Banerjee (2000) that Condition (3.8) on Page 1482 of Murphy and Van der Vaart (1997) is satisfied.

Next, we need to verify the “unbiasedness condition” – this is condition (3.9) of Murphy and Van der Vaart (1997) and in this case can be written as \(\sqrt{n} \mathbb{P}_n (\hat{l}(\delta, x, z; \beta_0, \hat{\Lambda}_n^{(\beta_0)}), \beta_0) \rightarrow_p 0\). By the consistency of \(\hat{\Lambda}_n^{(\beta_0)}\) for \(\Lambda_0\) and the facts that (i) the class of score functions obtained from the least favorable submodels are Donsker and (ii) \(P_0 \hat{l} = 0\), the above condition is equivalent to \(\sqrt{n} P_0 (\hat{l}(\delta, x, z; \beta_0, \hat{\Lambda}_n^{(\beta_0)}), \beta_0) \rightarrow_p 0\). (For notational convenience, we henceforth abbreviate \(\hat{l}(\delta, x, z; \beta_0, \Lambda, \beta_0)\) to \(\hat{l}(\cdot; \beta_0, \Lambda, \beta_0)\) and \(\hat{\Lambda}_n^{(\beta_0)}\) to \(\hat{\Lambda}_0\).) The proof of this fact follows very similar arguments to those involved in verifying the unbiasedness condition for the Cox model with current status data in Murphy and van der Vaart (1997) and makes crucial use of the result that \(\int_{\sigma} (\Lambda_0(u) - \Lambda_0(u))^2 du = O_p(n^{-2/3})\); for the full details, see BMM(2007).

It only remains to show the asymptotic linearity of the MLE in the efficient score function – in other words, establishing the representation (3.10). Since \(\hat{\beta}\) maximizes the function \(t \mapsto \mathbb{P}_n \log p(t, \Lambda_t(\beta, \hat{\Lambda}))\) over \(t\), we have (a): \(\mathbb{P}_n (\hat{l}(\cdot; \beta, \hat{\Lambda}, \hat{\beta}) = 0\). Let \(\mathbb{G}_n\) denote the empirical process \(\sqrt{n} (\mathbb{P}_n - P_0)\). By the Donsker property of the class of functions \(\hat{l}(\cdot; t, \Lambda, \beta)\) and the consistency of \((\hat{\beta}, \hat{\Lambda})\) we have (b): \(\mathbb{G}_n (\hat{l}(\cdot; \hat{\beta}, \hat{\Lambda}, \hat{\beta}) - \hat{l}(\cdot; \beta_0, \Lambda_0, \beta_0)) \rightarrow_p 0\). From (a) and (b), we readily deduce (c): \(\sqrt{n} P_0 (\hat{l}(\cdot; \hat{\beta}, \hat{\Lambda}, \hat{\beta})) = \mathbb{G}_n (\hat{l}(\cdot; \beta_0, \Lambda_0, \beta_0)) + o_p(1)\). By arguments similar to those used to show that \(\sqrt{n} P_0 (\hat{l}(\cdot; \beta_0, \Lambda_0, \beta_0)) \rightarrow_p 0\) (the unbiasedness condition) it can be shown that \(\sqrt{n} P_0 (\hat{l}(\cdot; \beta_0, \Lambda_0, \beta_0)) \rightarrow_p 0\). It follows that

\[ -\sqrt{n} P_0 (\hat{l}(\cdot; \beta_0, \Lambda_0, \beta_0)) + \sqrt{n} P_0 (\hat{l}(\cdot; \beta_0, \Lambda_0, \beta_0)) = \mathbb{G}_n (\hat{l}(\cdot; \beta_0, \Lambda_0, \beta_0)) + o_p(1) \quad (6.14) \]
Now define \( \tilde{\kappa}(;t, \Lambda) = \frac{\partial}{\partial t} \hat{l}(;t, \Lambda, t) \). Then, \( P_0(\tilde{\kappa}(;t, \Lambda)) = P_0 \left[ (\partial/\partial t)\hat{l}(;t, \Lambda, t) \right] = (\partial/\partial t) P_0 \left[ \hat{l}(;t, \Lambda, t) \right] \). Denote the left side of (6.14) by \( L \). Then, by an application of the mean value theorem and interchanging integration and differentiation, we have

\[
L = -\sqrt{n} \left[ P_0 \left( \hat{l}(;\hat{\beta}, \hat{\Lambda}, \hat{\beta}) \right) - P_0 \left( \hat{l}(;\beta_0, \Lambda, \beta_0) \right) \right] = -\sqrt{n} \left[ P_0 \left( \tilde{\kappa}(;\hat{\beta}, \hat{\Lambda}) \right) \right] (\beta_n - \beta_0)
\]

where \( \hat{\beta} \) lies between \( \beta_n \) and \( \beta_0 \) and does not depend on \( (\delta, z, x) \). But \( M_n \equiv P_0 \left( \tilde{\kappa}(;\hat{\beta}, \hat{\Lambda}) \right) \to P_0(\tilde{\kappa}(;\beta_0, \Lambda_0)) \) as \( n \to \infty \) (and \( (\beta_n, \Lambda_n) \to P (\beta_0, \Lambda_0) \)), by a simple application of the DCT. A brief argument (see BMM (2007)) shows that this equals \( -I_0 \). Now, from (6.14) and the representation of \( L \) in terms of \( \tilde{\kappa} \), we have,

\[
-\sqrt{n} M_n(\tilde{\beta}_n - \beta_0) = \sqrt{n} (P_n - P_0) \hat{l}_0 + o_p(1) = \sqrt{n} \mathbb{P}_n \hat{l}_0 + o_p(1),
\]

which can be rewritten as \( \sqrt{n} (\tilde{\beta}_n - \beta_0) = -M_n^{-1} \sqrt{n} \mathbb{P}_n \hat{l}_0 - M_n^{-1} o_p(1) = I_0^{-1} \sqrt{n} \mathbb{P}_n \hat{l}_0 + o_p(1) \) by Slutsky’s theorem, completing the proof of asymptotic efficiency.

Hence, facts (3.10) and (3.12) are established. The asymptotically linear representation for \( \beta_n \) and the limiting \( \chi^2 \) distribution for \( \text{lrtbeta}_n^0 \) may be established by similar steps. Some additional care needs to be exercised, since the parameter space for \( \Lambda \) is now restricted by fixing the value at the point \( z_0 \). Roughly the intuition is the following: The unconstrained MLE of \( \beta \), is \( \sqrt{n} \)-consistent and asymptotically efficient for the given model. The unconstrained likelihood ratio statistic for testing \( \beta = \beta_0 \), which we denote by \( \text{lrtbeta}_n \) is asymptotically \( \chi^2_2 \). These properties will be preserved even when we compute the above statistics under the single (true) constraint that \( \Lambda(z_0) = \theta_0 \). In fact, the same asymptotic representations for the above statistics will continue to hold when we constrain \( \Lambda \) at finitely many points. Note however, that the limit distribution of the MLE will generally be affected under infinitely many constraints on \( \Lambda \). This is easily seen when we constrain \( \Lambda \) on the support of \( Z \). In this case \( \Lambda \) is completely known and the asymptotic variance of \( \beta \) is the inverse of the ordinary information for \( \theta \) as opposed to the efficient information.

**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 \( (g_{a,b}(h), g_{a,b}^0(h)) \). Thus, it is shown that for any \( (h_1, h_2, \ldots, h_k) \), the random vector \( \{U_n(h_i)\}_{i=1}^k, \{V_n(h_i)\}_{i=1}^k \) \( \to_d \) \( \{g_{a,b}(h_i)\}_{i=1}^k, \{g_{a,b}^0(h_i)\}_{i=1}^k \) 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 \( (\psi_n, \phi_n) \) in \( L_2[-K, K] \times L_2[-K, K] \) such that \( \psi_n \) and \( \phi_n \) are monotone functions and \( (\psi_n, \phi_n) \) converges pointwise to \( (\phi, \psi) \) (where \( (\phi, \psi) \) 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 \( (g_{a,b}(h), g_{a,b}^0(h)) \), that \( (U_n(h), V_n(h)) \to_d (g_{a,b}(h), g_{a,b}^0(h)) \) 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). □

6.2 Further Details About The Unconstrained And Constrained MLEs From Section 2

Details of the “self–consistent” characterization of \( \hat{u}_{n}^{(\beta)} \) in Section 2: As in Section 2, we denote the function \( g(\beta, u) \) in the discussion that follows by \( \xi(u) \). This is strictly convex in \( u \equiv (u_{1}, u_{2}, \ldots, u_{n}) \) and for simplicity we denote \( \hat{u}_{n}^{(\beta)} \), its minimizer over the region \( C := \{ u : u_{1} \leq u_{2} \leq \ldots \leq u_{n} \} \), by \( \hat{u} \) (suppressing the dependence on \( n \) and \( \beta \)). Let \( \nabla_{j} \xi(u) \) denote the \( j \)'th partial derivative of \( \xi \) with respect to \( u \). Using the Kuhn–Tucker theorem for optimizing a convex function over a closed convex set, we find that \( \hat{u} = (\hat{u}_{1}, \hat{u}_{2}, \ldots, \hat{u}_{n}) \) is uniquely characterized by the conditions:

\[
\sum_{j=i+1}^{n} \nabla_{j} \xi(\hat{u}) \geq 0, \quad \text{for} \quad i = 1, 2, \ldots, (n - 1) \tag{6.14}
\]

and

\[
\sum_{j=1}^{n} \nabla_{j} \xi(\hat{u}) = 0. \tag{6.15}
\]

Consider now, the following (quadratic) function \( \hat{\xi}(u) = \frac{1}{2} [u - \hat{u} + K^{-1} \nabla \xi(\hat{u})]^{T} K [u - \hat{u} + K^{-1} \nabla \xi(\hat{u})] \) where \( K \) is some positive definite matrix. Note that Hess(\( \hat{\xi} \)) = \( K \) which is positive definite; thus \( \hat{\xi} \) is a strictly convex function. It is also finite and continuously differentiable over \( \mathbb{R}^{n} \). Also, \( \nabla \hat{\xi}(u) = K (u - \hat{u} + K^{-1} \nabla \xi(\hat{u})) \). Now, consider the problem of minimizing \( \hat{\xi} \) over \( C \). If \( u^{*} \) is the (unique) global minimizer, then necessary and sufficient conditions are given by conditions (6.14) for \( i = 1, 2, \ldots, n - 1 \) and (6.15), with \( \xi \) replaced by \( \hat{\xi} \) and \( \hat{u} \) replaced by \( u^{*} \). Now, \( \nabla \hat{\xi}(\hat{u}) = \nabla \xi(\hat{u}) \), so that the vector \( \hat{u} \in C \) does indeed satisfy the conditions (6.14) for \( i = 1, 2, \ldots, n - 1 \) and (6.15), with \( \xi \) replaced by \( \hat{\xi} \). It follows that \( \hat{u} \) is the unique minimizer of \( \hat{\xi} \) over \( C \), i.e. \( u^{*} = \hat{u} \).

It now suffices to try to minimize \( \hat{\xi} \); of course the problem here is that \( \hat{u} \) is unknown and \( \hat{\xi} \) is defined in terms of \( \hat{u} \). However, an iterative scheme can be developed along the following lines. Choosing \( K \) to be a diagonal matrix with the \( i, i \)'th entry being \( d_{i} \equiv \nabla_{ii} \xi(\hat{u}) \) (\( K \) thus defined is a p.d. matrix, since the diagonal entries of the Hessian of \( \xi \) at the minimizer \( \hat{u} \), which is a positive definite matrix, are positive), we see that the above quadratic form reduces to \( \eta(u)/2 \) where \( \eta(u) = \sum_{i=1}^{n} [u_{i} - (\hat{u}_{i} - \nabla_{i} \xi(\hat{u}))d_{i}^{-1}]^{2} d_{i} \). Thus, \( \hat{u} \) minimizes \( \eta(u) \) subject to the constraints that \( u_{1} \leq u_{2} \leq \ldots \leq u_{n} \) and therefore furnishes the isotonic regression of the function \( h(i) = \hat{u}_{i} - \nabla_{i} \xi(\hat{u})d_{i}^{-1} \) 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
\( \hat{u} \equiv (\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n) = \text{slogcm} \left\{ \sum_{j=1}^i d_i, \sum_{j=1}^i h(i) d_i \right\}_{i=0}^n \). This representation leads to the MICM as outlined in Section 2.

Implications of the self–consistent/self–induced characterization of \( \hat{u}_n^{(β)} \): Recall that \( \hat{u}_n^{(β)} = (\hat{u}_{1,n}^{(β)}, \hat{u}_{2,n}^{(β)}, \ldots, \hat{u}_{n,n}^{(β)}) \). 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 \( j \in B_i, \) \( \hat{u}_{j,n}^{(β)} \) equals \( w_i \), with the common block values, the \( w_i \)'s, satisfying \( w_1 < w_2 < \ldots < w_k \). Since the \( \hat{u}_{j,n}^{(β)} \)'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 \( \hat{u}_n^{(β)} \) 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 \( \phi(\Delta(i), R_i(β), t) \) by \( \phi_{i,β}(t) \) and its first and second derivatives with respect to \( t \) by \( \phi'_{i,β}(t) \) and \( \phi''_{i,β}(t) \). Identifying the function \( \hat{ψ}_n^{(β)} \) with the vector \( \hat{u}_n^{(β)} \) in the usual fashion, we can write

\[
\hat{ψ}_n^{(β)} \equiv \text{slogcm} \left\{ \sum_{i=1}^k \phi'_{i,β}(\hat{ψ}_n^{(β)}(Z(i))) \right\}_{\left(\sum_{i=1}^k \phi''_{i,β}(\hat{ψ}_n^{(β)}(Z(i)))\right)}^{k=0}.
\]

Hence, we can write \( w_i \) as

\[
w_i = \hat{ψ}_n^{(β)}(Z(j)) = \frac{\sum_{k \in B_i} \{\hat{ψ}_n^{(β)}(Z(k)) \phi''_{k,β}(\hat{ψ}_n^{(β)}(Z(k))) - \phi'_{k,β}(\hat{ψ}_n^{(β)}(Z(k)))\}}{\sum_{k \in B_i} \phi''_{k,β}(\hat{ψ}_n^{(β)}(Z(k)))} \text{ for } j \in B_i. \tag{6.16}
\]

Further details about \( \hat{u}_{n,0}^{(β)} \): The vector \( \hat{u}_{n,0}^{(β)} \), which we identify with \( \hat{ψ}_n^{(β)}(0) \) (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 \( \hat{u}_n^{(β)} \) 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 \( g(β, u) \) to \( ξ(u) \), suppressing the dependence on \( β \). We also abbreviate \( \hat{u}_{n,0}^{(β)} \) to \( \hat{u}(0) \). For each \( i \), set \( d_i = \nabla_i ξ(\hat{u}(0)) \). Then, \( \hat{u}(0) \) minimizes

\[
A(u_1, u_2, \ldots, u_n) = \sum_{i=1}^n \left[ u_i - \left( \hat{u}_i^{(0)} - \nabla_i ξ(\hat{u}(0)) d_i^{-1} \right) \right]^2 d_i \text{ subject to the constraints that } u_1 \leq u_2 \leq \ldots \leq u_m \leq \theta_0 \leq u_{m+1} \leq \ldots \leq u_n.
\]

Let \( \hat{B}_1, \hat{B}_2, \ldots, \hat{B}_l \) denote the level blocks of \( \hat{u}_{n,0}^{(β)} \) and let \( \{\hat{w}_i\}_{i=1}^l \) denote the corresponding level values. Then, as long as \( \hat{w}_i \neq \theta_0 \), it can be written as

\[
\hat{w}_i = \hat{ψ}_{n,0}^{(β)}(Z(j)) = \frac{\sum_{k \in \hat{B}_i} \{ψ_{n,0}^{(β)}(Z(k)) \phi''_{k,β}(ψ_{n,0}^{(β)}(Z(k))) - \phi'_{k,β}(ψ_{n,0}^{(β)}(Z(k)))\}}{\sum_{k \in \hat{B}_i} \phi''_{k,β}(ψ_{n,0}^{(β)}(Z(k)))} \text{ for } j \in \hat{B}_i. \tag{6.17}
\]

This representation is once again, a direct outcome of the self–induced characterization, and will prove useful in what follows.
6.3 Proof of Theorem 3.3:

The likelihood ratio statistic of interest can be written as

\[
\text{lrt}_{i_n} = 2 \left( l_n(\hat{\beta}_n, \hat{\psi}_n) - l_n(\hat{\beta}_n, \hat{\psi}_{n,0}) \right)
\]

\[
= 2 \left( l_n(\beta_0, \hat{\psi}_{n}(\beta_0)) - l_n(\beta_0, \hat{\psi}_{n,0}(\beta_0)) \right) + 2 \left( l_n(\hat{\beta}_n, \hat{\psi}_n) - l_n(\beta_0, \hat{\psi}_{n}(\beta_0)) \right) - 2 \left( l_n(\hat{\beta}_n, \hat{\psi}_n) - l_n(\beta_0, \hat{\psi}_{n,0}(\beta_0)) \right).
\]

From Theorem 3.1 \( \hat{R}_n \equiv 2 \left( l_n(\hat{\beta}_n, \hat{\psi}_n) - l_n(\beta_0, \hat{\psi}_{n,0}(\beta_0)) \right) = o_p(1) \) whence it suffices to find the asymptotic distribution of \( C_n = 2 \left( l_n(\beta_0, \hat{\psi}_{n}(\beta_0)) - l_n(\beta_0, \hat{\psi}_{n,0}(\beta_0)) \right) \). This is precisely the likelihood ratio statistic for testing \( \psi(z_0) = \theta_0 \) holding \( \beta \) fixed at its true value \( \beta_0 \).

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 \( \phi \) 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 \( \hat{\psi}(Z(i)) \), \( \hat{\psi}_{n,0}(\beta_0)(Z(i)) \) as \( \psi_0(Z(i)) \), 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 \( \hat{\psi}(Z(i)) \) and \( \psi_0(Z(i)) \) differ is denoted by \( J_n \). Now, \( C_n = -2 T_n \) where

\[
T_n = \sum_{i=1}^{n} \phi_i(\hat{\psi}(Z(i))) - \sum_{i=1}^{n} \phi_i(\psi_0(Z(i))) = \sum_{i \in J_n} \phi_i(\hat{\psi}(Z(i))) - \sum_{i \in J_n} \phi_i(\psi_0(Z(i)))
\]

\[
= \sum_{i \in J_n} \phi_i'(\psi_0(z_0)) \left[ (\hat{\psi}(Z(i)) - \psi_0(z_0)) - (\psi_0(Z(i)) - \psi_0(z_0)) \right]
\]

\[
+ \sum_{i \in J_n} \frac{1}{2} \phi_i''(\psi_0(z_0)) \left[ (\hat{\psi}(Z(i)) - \psi_0(z_0))^2 - (\psi_0(Z(i)) - \psi_0(z_0))^2 \right] + R_n
\]

\[
= T_{n,1} + T_{n,2} + R_n \text{ by Taylor–expanding } \phi_i(t) \text{ around } \psi_0(z_0).
\]

Here, \( R_n = \sum_{i \in J_n} (1/6) [\phi_i'''(\hat{\psi}(Z(i))) \left( \hat{\psi}(Z(i)) - \psi_0(z_0) \right)^3 - \phi_i'''(\psi_0(Z(i))) \left( \psi_0(Z(i)) - \psi_0(z_0) \right)^3] \)

(where \( \hat{\psi}(Z(i)) \) is some point between \( \hat{\psi}(Z(i)) \) and \( \psi_0(z_0) \) and \( \psi_0(Z(i)) \) is some point between \( \psi_0(Z(i)) \) and \( \psi_0(z_0) \) and can be shown to converge to 0 in probability by using the facts that (a) \( \sup_{i \in J_n} \phi_i'''(\hat{\psi}(Z(i))) \) and \( \sup_{i \in J_n} \phi_i'''(\psi_0(Z(i))) \) are \( O_p(1) \), (b) \( \sup_{z \in D_n} | \hat{\psi}(z) - \psi_0(z) | \) and \( \sup_{z \in D_n} | \psi_0(z) - \psi_0(z_0) | \) are \( O_p(n^{-1/3}) \) where \( D_n \) is the set on which \( \hat{\psi} \) and \( \psi_0 \) 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

\[
T_{n,2} = \sum_{i \in J_n} \frac{1}{2} \phi_i''(\psi_0(z_0)) \left[ (\hat{\psi}(Z(i)) - \psi_0(z_0))^2 - (\psi_0(Z(i)) - \psi_0(z_0))^2 \right]
\]

\[
= \sum_{i \in J_n} \frac{1}{2} \phi_i''(\hat{\psi}(Z(i)))[\hat{\psi}(Z(i)) - \psi_0(z_0)]^2 - \sum_{i \in J_n} \frac{1}{2} \phi_i''(\psi_0(Z(i)))[\psi_0(Z(i)) - \psi_0(z_0)]^2 + o_p(1).
\]

Now consider, \( T_{n,1} = \sum_{i \in J_n} \phi_i'(\psi_0(z_0))[\hat{\psi}(Z(i)) - \psi_0(z_0)] - \sum_{i \in J_n} \phi_i'(\psi_0(z_0))[\psi_0(Z(i)) - \psi_0(z_0)] = S_1 - S_2 \). Consider the term \( S_2 \). Note that for each \( i \in J_n \), we can write:

\[
\phi_i'(\psi_0(z_0)) = \phi_i'(\psi_0(Z(i))) + (\psi_0(z_0) - \psi_0(Z(i))) \phi_i''(\psi_0(Z(i))) + \frac{1}{2} \phi_i'''(\psi_0(Z(i))) \psi_0(z_0) - \psi_0(Z(i))^2
\]

28
where $\tilde{\psi}_0(Z(\cdot))^{\ast\ast}$ is a point between $\tilde{\psi}_0(Z(\cdot))$ and $\psi(z_0)$. We then have,

$$S_2 = \sum_{i \in J_n} \left[ \phi_i'(\tilde{\psi}_0(Z(\cdot))) + (\psi(z_0) - \tilde{\psi}_0(Z(\cdot))) \phi_i''(\tilde{\psi}_0(Z(\cdot))) + \frac{1}{2} \phi_i'''(\tilde{\psi}_0(Z(\cdot))^{\ast\ast})(\psi(z_0) - \tilde{\psi}_0(Z(\cdot)))^2 \right]$$

$$= \sum_{i \in J_n} \left[ \phi_i'(\tilde{\psi}_0(Z(\cdot))) + (\psi(z_0) - \tilde{\psi}_0(Z(\cdot))) \phi_i''(\tilde{\psi}_0(Z(\cdot))) \right] (\tilde{\psi}_0(Z(\cdot)) - \psi(z_0)) + o_p(1)$$

$$= -\sum_{i \in J_n} \phi_i''(\tilde{\psi}_0(Z(\cdot))) \left[ \tilde{\psi}_0(Z(\cdot)) - \frac{\phi_i'(\tilde{\psi}_0(Z(\cdot)))}{\phi_i''(\tilde{\psi}_0(Z(\cdot)))} \psi(z_0) \right] (\tilde{\psi}_0(Z(\cdot)) - \psi(z_0)) + o_p(1),$$

where the fact that the term involving $\phi_i'''$ is $o_p(1)$ is deduced by arguments similar to those needed to show that $R_n$ is $o_p(1)$. Now, let $B_0, B_1, \ldots, B_r$ denote the level blocks for $\tilde{\psi}_0(Z(\cdot))$ that constitute $J_n$, with level values $w_1^0, w_2^0, \ldots, w_r^0$ and suppose that $w_i^0 = \psi(z_0) \equiv \theta_0$. Then,

$$S_2 + o_p(1) = \sum_{j=1}^r \sum_{i \in B_j} \left[ \phi_i''(\tilde{\psi}_0(Z(\cdot))) \left( \tilde{\psi}_0(Z(\cdot)) - \frac{\phi_i'(\tilde{\psi}_0(Z(\cdot)))}{\phi_i''(\tilde{\psi}_0(Z(\cdot)))} \psi(z_0) \right) - \psi(z_0) \phi_i''(\tilde{\psi}_0(Z(\cdot))) \right] (w_i^0 - \psi(z_0))$$

$$= -\sum_{j=1}^r \sum_{i \in B_j} \phi_i''(w_i^0) \left( w_i^0 - \frac{\phi_i'(w_i^0)}{\phi_i''(w_i^0)} \psi(z_0) \right) - \psi(z_0) \phi_i''(w_i^0) \left( w_i^0 - \psi(z_0) \right)$$

$$= -\sum_{j \neq l} (w_j^0 - \psi(z_0)) \left[ \sum_{i \in B_j} \phi_i''(w_i^0) w_j^0 - \phi_i'(w_j^0) - \psi(z_0) \sum_{i \in B_j} \phi_i''(w_i^0) \right]$$

$$= -\sum_{j \neq l} (w_j^0 - \psi(z_0)) \left[ \sum_{i \in B_j} \phi_i''(w_i^0) \left( \frac{\sum_{i \in B_j} \phi_i'(w_i^0) w_j^0 - \phi_i'(w_j^0)}{\sum_{i \in B_j} \phi_i''(w_i^0)} \right) - \psi(z_0) \right]$$

$$= -\sum_{j \neq l} \sum_{i \in B_j} \phi_i''(w_i^0) \left( w_j^0 - \psi(z_0) \right)^2,$$

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 $w(0) = [\sum_{k \in B'} (w(0) f_k(w(0)) - f_k(w(0)))]/[\sum_{k \in B'} f_k(w(0))]$ provided $w(0) \neq \theta_0$. This is a direct consequence of the representation (6.17). It follows that

$$S_2 + o_p(1) = -\sum_{j \neq l} \sum_{i \in B_j} \phi_i''(w_i^0) \left( w_j^0 - \psi(z_0) \right)^2 = -\sum_{j=1}^r \sum_{i \in B_j} \phi_i''(w_i^0) \left( w_j^0 - \psi(z_0) \right)^2$$

$$= -\sum_{j=1}^r \sum_{i \in B_j} \phi_i''(\tilde{\psi}_0(Z(\cdot))) \left( \tilde{\psi}_0(Z(\cdot)) - \psi(z_0) \right)^2 = -\sum_{i \in J_n} \phi_i''(\tilde{\psi}_0(Z(\cdot))) \left( \tilde{\psi}_0(Z(\cdot)) - \psi(z_0) \right)^2.$$
\[ \psi_0(z_0)^2 + o_p(1). \] Now, on using (6.18) and the fact that \( R_n \) is \( o_p(1) \) we get

\[
T_n = T_{n,1} + T_{n,2} + o_p(1)
\]

\[
= -\frac{1}{2} \sum_{i \in J_n} \phi_i''(\tilde{\psi}(Z(i))) (\tilde{\psi}(Z(i)) - \psi_0(z_0))^2 + \frac{1}{2} \sum_{i \in J_n} \phi_i''(\tilde{\psi}_0(Z(i))) (\tilde{\psi}_0(Z(i)) - \psi_0(z_0))^2 + o_p(1),
\]

whence

\[
C_n = -2 T_n = \sum_{i \in J_n} \phi_i''(\tilde{\psi}(Z(i))) (\tilde{\psi}(Z(i)) - \psi_0(z_0))^2 - \sum_{i \in J_n} \phi_i''(\tilde{\psi}_0(Z(i))) (\tilde{\psi}_0(Z(i)) - \psi_0(z_0))^2 + o_p(1)
\]

\[
= \sum_{i \in J_n} \phi_i''(\psi_0(Z(i))) (\tilde{\psi}(Z(i)) - \psi_0(z_0))^2 - \sum_{i \in J_n} \phi_i''(\psi_0(Z(i))) (\tilde{\psi}_0(Z(i)) - \psi_0(z_0))^2 + o_p(1).
\]

Now, \( \phi_i''(\psi_0(Z(i))) = \frac{\exp(\psi_0(Z(i)) + \beta_0^T X(i))}{(1 + \exp(\psi_0(Z(i)) + \beta_0^T X(i)))^2} \), whence

\[
C_n = \sum_{i \in J_n} \frac{\exp(\psi_0(Z(i)) + \beta_0^T X(i))}{(1 + \exp(\psi_0(Z(i)) + \beta_0^T X(i)))^2} \left[ (\tilde{\psi}(Z(i)) - \psi_0(z_0))^2 - (\tilde{\psi}_0(Z(i)) - \psi_0(z_0))^2 \right] + o_p(1)
\]

\[
= n^{1/3} (\mathbb{P}_n - P) \xi_n(\delta, z, x) + n^{1/3} P \xi_n(\delta, z, x) + o_p(1)
\]

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 \( (\Delta, Z, X) \), \( \xi_n(\delta, z, x) \) is the random function given by

\[
\xi_n(\delta, z, x) = \frac{\exp(\psi_0(z) + \beta_0^T x)}{(1 + \exp(\psi_0(z) + \beta_0^T x))^2} \left[ (n^{1/3}(\psi(z) - \psi_0(z))^2) - (n^{1/3}(\tilde{\psi}_0(z) - \psi_0(z))^2) \right] 1(Z \in D_n).
\]

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

\[
n^{1/3} P \xi_n(\delta, z, x) = n^{1/3} P \left[ h(z, x) K_n(z) \right] \quad \text{where} \quad K_n(z) = \left[ (n^{1/3}(\tilde{\psi}(z) - \psi_0(z))^2) - (n^{1/3}(\tilde{\psi}_0(z) - \psi_0(z))^2) \right] 1(Z \in D_n) \quad \text{and} \quad h(z, x) = \frac{\exp(\psi_0(z) + \beta_0^T x)}{(1 + \exp(\psi_0(z) + \beta_0^T x))^2}.
\]

Thus,

\[
n^{1/3} P \xi_n(\delta, z, x) = n^{1/3} P \left[ K_n(z) h(z, x) \right] = n^{1/3} \int_{D_n} K_n(z) E(h(Z, X) \mid Z = z) f_Z(z) dz
\]

\[
= n^{1/3} \int_{D_n} K_n(z_0 + h n^{-1/3}) w(z_0 + h n^{-1/3}) f_Z(z_0 + h n^{-1/3}) dh
\]

\[30\]
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, 
\[ K_n(z_0 + h n^{-1/3}) = (U_n^2(h) - V_n^2(h))1 \quad (h \in \tilde{D}_n) \]
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 \( w(z) = \int \frac{\exp(\psi_0(z) + \beta_n^T x)}{(1 + \exp(\psi_0(z) + \beta_n^T x))^2} f(z, x) \, dz \). 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, \( n^{1/3} P \xi_n(\delta, z, x) = \int w(z) f_Z(z) (U_n^2(h) - V_n^2(h)) \, dh + o_p(1) \). 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 \( n^{1/3} P \xi_n(\delta, z, x) \rightarrow_d a^{-2} \int \left( (g_{a,b}(h))^2 - (g_{a,b}^0(h))^2 \right) \, dh \), (6.18) and the fact that \( a^{-2} \int \left( (g_{a,b}(h))^2 - (g_{a,b}^0(h))^2 \right) \, dh \equiv \int \left( (g_{1,1}(h))^2 - (g_{1,1}^0(z))^2 \right) \, dh \equiv D \) follows as a direct application of Lemma 3.1 followed by the change of variable theorem from calculus. It remains to show that
\[
\tilde{R}_n = 2 \left( l_n(\hat{\beta}_n, \hat{\psi}_n) - l_n(\beta_0, \hat{\psi}_0) \right) - 2 \left( l_n(\hat{\beta}_n, \hat{\Lambda}_{n,0}) - l_n(\beta_0, \hat{\psi}_0) \right)
\equiv 2 \left( l_n(\hat{\beta}_n, \hat{\Lambda}_n) - l_n(\beta_0, \hat{\psi}_0) \right) - 2 \left( l_n(\hat{\beta}_n, \hat{\Lambda}_{n,0}) - l_n(\beta_0, \hat{\psi}_0) \right)
\]
is \( o_p(1) \). This is precisely \( lrt\beta_n - lrt\beta_n^0 \). From Theorem 3.1 we get:
\[
lrt\beta_n - lrt\beta_n^0 = n \left( (\hat{\beta}_n - \beta_0)^T \tilde{I}_0 (\hat{\beta}_n - \beta_0) - (\hat{\beta}_n - \beta_0)^T \tilde{I}_0 (\beta_0 - \beta_0) \right) + o_p(1)
\equiv n \left( (\hat{\beta}_n - \beta_0)^T \tilde{I}_0 (\hat{\beta}_n - \beta_0) + 2 \sqrt{n} (\hat{\beta}_n - \beta_0)^T \tilde{I}_0 \sqrt{n} (\hat{\beta}_n - \beta_0) + o_p(1) \right)
\equiv \sqrt{n} (\hat{\beta}_n - \beta_0)^T \tilde{I}_0 \sqrt{n} (\hat{\beta}_n - \beta_0) + 2 \sqrt{n} (\hat{\beta}_n - \beta_0)^T \tilde{I}_0 \sqrt{n} (\hat{\beta}_n - \beta_0) + o_p(1)
\equiv I_n + I_{I_n} + o_p(1) .
\]
The fact that \( I_n \) is \( o_p(1) \) follows from the observation that \( \sqrt{n} (\hat{\beta}_n - \beta_0) = r_n - s_n \), which is \( o_p(1) \) (by Theorem 3.1). The fact that \( I_{I_n} \) is \( o_p(1) \) follows on using the facts that \( \sqrt{n} (\hat{\beta}_n - \beta_0) \) is \( o_p(1) \) and that \( \sqrt{n} (\hat{\beta}_n - \beta_0) \) is \( O_p(1) \).

References


**Table 1A: Parametric Results**

<table>
<thead>
<tr>
<th>Variables/Group</th>
<th>Bimaru Female</th>
<th>Bimaru Male</th>
<th>NonBimaru Female</th>
<th>NonBimaru Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-3.160610 (0.0946)</td>
<td>-2.526723 (0.2319)</td>
<td>-6.048611 (0.0068)</td>
<td>-4.426694 (0.0147)</td>
</tr>
<tr>
<td>Standard of Living</td>
<td>0.408983 (0.0733)</td>
<td>0.504875 (0.0591)</td>
<td>0.635662 (0.0160)</td>
<td>0.567153 (0.0142)</td>
</tr>
<tr>
<td>SCST Dummy</td>
<td>-0.833532 (0.0044)</td>
<td>-1.514270 (0.0000)</td>
<td>-0.537154 (0.0801)</td>
<td>-0.648857 (0.0409)</td>
</tr>
<tr>
<td>HeadLiteracy</td>
<td>1.010064 (0.0001)</td>
<td>0.779417 (0.0071)</td>
<td>1.245977 (0.0000)</td>
<td>0.686764 (0.0298)</td>
</tr>
<tr>
<td>Cost of Schooling</td>
<td>-0.069846 (0.7004)</td>
<td>-0.150194 (0.4060)</td>
<td>0.250532 (0.1779)</td>
<td>0.166638 (0.2669)</td>
</tr>
<tr>
<td>TLC</td>
<td>0.049153 (0.8542)</td>
<td>0.187457 (0.5386)</td>
<td>0.378737 (0.2243)</td>
<td>0.628891 (0.0438)</td>
</tr>
<tr>
<td>Midday Meal</td>
<td>0.460717 (0.1633)</td>
<td>0.536007 (0.1329)</td>
<td>0.266945 (0.3971)</td>
<td>0.061193 (0.8435)</td>
</tr>
</tbody>
</table>

Note: p-values are in the parentheses.

**Table 1B: Parametric Results with Significant Variables**

<table>
<thead>
<tr>
<th>Variables/Group</th>
<th>Bimaru Female</th>
<th>Bimaru Male</th>
<th>NonBimaru Female</th>
<th>NonBimaru Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-3.459782 0.0518</td>
<td>-2.686465 0.1824</td>
<td>-4.501767 0.0256</td>
<td>-3.499595 0.0444</td>
</tr>
<tr>
<td>Standard of Living</td>
<td>0.411713 0.0633</td>
<td>0.446572 0.0835</td>
<td>0.659107 0.0096</td>
<td>0.601905 0.0061</td>
</tr>
<tr>
<td>SCST Dummy</td>
<td>-0.743829 0.0084</td>
<td>-1.410061 0.0000</td>
<td>-0.665219 0.0235</td>
<td>-0.781412 0.0110</td>
</tr>
<tr>
<td>HeadLiteracy</td>
<td>1.018684 0.0000</td>
<td>0.780840 0.0066</td>
<td>1.208981 0.0001</td>
<td>0.784424 0.0113</td>
</tr>
</tbody>
</table>

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.
Table 2A: Nonparametric Results for Probabilities and their Confidence Interval at the First Quartile

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

Table 2B: Nonparametric Results for Probabilities and their Confidence Interval at the Median

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

Table 2C: Nonparametric Results for Probabilities and their Confidence Interval at the Third Quartile

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