NBER WORKING PAPER SERIES

INFERRING WELFARE MAXIMIZING TREATMENT ASSIGNMENT UNDER BUDGET CONSTRAINTS

Debopam Bhattacharya
Pascaline Dupas

Working Paper 14447
http://www.nber.org/papers/w14447

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2008

Previously circulated under the title: "Nonparametric Inference on Efficient Treatment Assignment under Budget Constraints." We thank seminar participants at Dartmouth, Oxford, UCL and the World Bank for useful comments. Dupas gratefully acknowledges funding from the Acumen Fund. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

Inferring Welfare Maximizing Treatment Assignment under Budget Constraints
Debopam Bhattacharya and Pascaline Dupas
NBER Working Paper No. 14447
October 2008, Revised July 2011
JEL No. C01,C14,I38

ABSTRACT

This paper concerns the problem of allocating a binary treatment among a target population based on observed covariates. The goal is to (i) maximize the mean social welfare arising from an eventual outcome distribution, when a budget constraint limits what fraction of the population can be treated and (ii) to infer the dual value, i.e. the minimum resources needed to attain a specific level of mean welfare via efficient treatment assignment. We consider a treatment allocation procedure based on sample data from randomized treatment assignment and derive asymptotic frequentist confidence interval for the welfare generated from it. We propose choosing the conditioning covariates through cross-validation. The methodology is applied to the efficient provision of anti-malaria bed net subsidies, using data from a randomized experiment conducted in Western Kenya. We find that subsidy allocation based on wealth, presence of children and possession of bank account can lead to a rise in subsidy use by about 9 percentage points compared to allocation based on wealth only, and by 17 percentage points compared to a purely random allocation.

Debopam Bhattacharya
Oxford University
debopam.bhattacharya@st-hildas.ox.ac.uk

Pascaline Dupas
Department of Economics
UCLA
8283 Bunche Hall
Los Angeles, CA 90095
and NBER
pdupas@econ.ucla.edu

1 Introduction

Governments of developing countries often subsidize access to key health and educational resources for the most vulnerable sections of their populations. However, such subsidizing efforts are typically constrained by binding budget ceilings. When budgets are such that only a small fraction of a target population can receive a given subsidy, the eligibility rule used to decide who will receive the subsidy can have an important effect on the overall benefit arising from the subsidy program. In this paper, we consider the problem of allocating a fixed amount of treatment resources to a target population with the aim of maximizing the mean population outcome, and the dual problem of estimating the minimum cost of achieving a given mean outcome in the population by efficient targeting of a treatment. We set-up an econometric framework for studying this problem and apply it to design welfare-maximizing allocation of subsidies for an effective malaria control tool – insecticide-treated bed nets (ITNs) – among households in a malaria-endemic region of Kenya.

Our paper contributes to a recent but steadily expanding literature in statistics and economics on how experimental evidence on treatment effect heterogeneity may be used to maximize gains from social programs. This problem has been analyzed in the literature as a statistical decision problem with finite samples under ambiguity but with no budget constraint (c.f., Dehejia (2001), Manski (2004, 2005), Stoye (2009), Hirano and Porter (2009) and Tetenov (2008)). Our substantive contribution is to study such problems in the presence of aggregate budget constraints – an extremely common situation in real life but largely ignored in the treatment choice literature.[1] The constraint makes the treatment assignment problem substantively realistic, especially for developing countries, and raises a set of unique substantive and technical issues that are absent in the unconstrained case.

On the empirical side, treatment targeting has been investigated in job-training contexts, c.f., Berger, Black and Smith (2001), Frolich (2008), Behnke et al (2008), Lechner and Smith (2007). See also the papers in Eberts et al (ed., 2002) and the special issue of the Economic Journal (Nov, 2006) on profiling. A few recent studies have used experimental data to estimate the parameters of dynamic structural models of behavior and utilized the estimates to simulate the effects of counterfactual policy interventions (c.f.

[1] Manski (2005) studies planning problems which satisfy 'separability' and specifically mentions (page 10–11) budget constraints as a situation where separability is violated and, consequently, not studied by him.

Attanasio, Meghir and Santiago (2011) and Duflo et al (2007)). Mahajan et al (2009) discuss identification and estimation of a static structural model of ITN adoption using observational data alone and use the estimated parameters to perform counterfactual policy analysis. Todd and Wolpin (2006, 2007) discuss the estimation of structural models of behavior using pre-program data and compare predictions of their estimated model with subsequent experimental data.

The rest of the paper is organized as follows. Section 2 describes the treatment assignment problem under budget constraints. Section 3 describes our sample-based approach. Section 4 introduces the estimators; section 5 develops the relevant distribution theory and section 6 discusses the covariate choice problem when assignment is based on parametric models of treatment effects. Section 7 presents the application to the welfare-maximizing allocation of bed nets in Kenya and presents a Monte Carlo exercise to inform the covariate choice problem. Section 8 concludes with directions for future research. The appendix contains technical proofs.

In the text, the symbol $:=$ will mean equal by definition, the symbols ϕ and Φ will refer to the standard normal density and c.d.f., respectively and $F_{W|V}(w|v)$ will denote the c.d.f. of the random variable W at w, conditional on $V = v$.

2 Set-up

The assignment problem: To fix ideas, let Y denote a household-level outcome and let D denote a binary treatment whose value can be affected directly by policy. Let X denote observed covariates which includes both discrete and continuous components. In the bednet example, analyzed below in detail, the population of interest is rural households of Western Kenya. We have a simple random sample drawn from two districts in Western Kenya. Each household is an observation. Y is a binary outcome which equals 1 if the household owns and uses a bed net. X is the presence of a child under 10, the wealth per capita and ownership of a bank account. The treatment, $D = 1$, is offering a highly subsidized bed net to the household. Y_1 and Y_0 are the value of the potential outcome Y with and without the treatment, respectively: $Y = DY_1 + (1 - D)Y_0$. Let $\beta_d(x)$ denote the expected outcome for households with $X = x$ when their treatment status D is set externally at $d \in \{0, 1\}$, i.e., $\beta_d(x) = E(Y_d|X = x)$. If the observed D is independent of (Y_1, Y_0) conditional on X, as in a randomized trial (the case studied here), then a

nonparametric regression of Y on X for households with $D = d$ in the sample can be used to recover this function.

First, consider an idealized version of a social planner's problem where $\beta_d(x)$ and the marginal distribution of X with support $\bar{\mathcal{X}}$ are known to the planner. The planner faces a budget constraint that the fraction of households who can be treated is at most c. We define the planner's idealized problem as the choice of a map $p : \bar{\mathcal{X}} \to [0,1]$ where $p(x)$ equals the probability that a household with $X = x$ is assigned to $D = 1$ with probability $p(x)$. We will assume that the planner wants to maximize mean outcome.[2] Then the planner's problem is

$$\max_{p(\cdot)} \int_{x \in \bar{\mathcal{X}}} \left[\beta_1(x) p(x) + \beta_0(x) \{1 - p(x)\}\right] dF_X(x)$$

subject to

$$c = \int_{x \in \bar{\mathcal{X}}} p(x) \, dF_X(x). \tag{1}$$

Define $\beta(x)$ to be $\beta_1(x) - \beta_0(x)$, i.e., the conditional average treatment effect (CATE). We allow for negative program impacts (i.e., $\Pr[\beta(X) < 0] \geq 0$), but to keep the present problem substantively interesting and analytically tractable, we will maintain the following condition throughout the paper:

Assumption Maintained (AM): (i) For some $\delta \geq 0$, we have $\Pr(\beta(X) > \delta) > c$ (the constraint binds); (ii) $\beta(X)$ has bounded support with Lebesgue density bounded away from zero.

Condition AM (i) lets us ignore some technical complications due to (additional) nondifferentiability which would arise if $\Pr(\beta(X) > 0)$ is close to c, relative to the sample size (c.f., Andrews (1999)). It also makes the problem more realistic in a developing country setting by making the budget constraint bind. condition AM (ii) is not strictly necessary for us but it simplifies the exposition and proofs.[3]

It is straightforward to show that under condition AM(i) and AM (ii), the solution to the problem subject to (1) is of the form $p(x) = 1\{\beta(x) \geq \gamma\}$ where γ satisfies: $c = \int_{x \in \bar{\mathcal{X}}} 1(\beta(x) \geq \gamma) \, dF_X(x)$. That is, γ is the $(1-c)$th quantile in the marginal distribution

[2] More generally, if the planner is interested in maximizing (a possibly covariate weighted) outcome utility, then $\beta_d(x)$ represents the expected value of the planner's utility defined on outcomes for individuals with $X = x$ and $D = d$.

[3] In the appendix part 3, we (a) show how a simple model of household consumption would imply AM(ii) and (b) briefly discuss how the analysis would proceed if we were to drop AM.

of the random variable $\beta(X)$ which is unique under assumption AM(ii). Define the resulting "ideal" welfare as

$$\rho = \int_{x \in \bar{\mathcal{X}}} \left[\beta_1(x) 1(\beta(x) \geq \gamma) + \beta_0(x) 1(\beta(x) < \gamma) \right] dF_X(x), \tag{2}$$

which would be attainable if $\beta_1(\cdot)$, $\beta_0(\cdot)$ and the distribution of X were known exactly to the planner. Note that γ also equals $\rho'(c)$, the shadow cost of the budget constraint, i.e., the expected treatment effect on the marginal household made eligible for treatment under our budget-constrained rationing rule.

Minimizing expenditure: The dual formulation of the problem is where the planner's objective is to achieve an expected outcome equal to b by allocating treatment based on covariates. The parameter of interest is the minimum amount of funds necessary to achieve b. This problem can be represented as

$$\min_{p(\cdot)} \int_{x \in \bar{\mathcal{X}}} p(x) dF_X(x) \tag{3}$$

subject to

$$\int_{x \in \bar{\mathcal{X}}} \left[\beta_1(x) p(x) + \beta_0(x) \{1 - p(x)\} \right] dF_X(x) = b. \tag{4}$$

It is not hard to show that the optimal $p(\cdot)$ will again be of the form $p(x) = 1\{\beta(x) \geq \gamma(b)\}$ where $\gamma(b)$ is such that $p(\cdot)$ satisfies (4). Note that by duality, the minimum value of (3) is simply $\rho^{-1}(b)$ where $\rho(\cdot)$ is defined in (2) and $\rho^{-1}(b) = \inf\{c : \rho(c) \geq b\}$. Notice that $\rho(\cdot)$ is monotone increasing and so $\rho^{-1}(\cdot)$ is well-defined.

Some extensions: While we focus the current paper on the mean utility of outcome as the objective, the idea applies in principle to any functional of the overall outcome distribution. Let $F_1(\cdot|x)$ and $F_0(\cdot|x)$ denote the marginal C.D.F. of the outcome, conditional on X, under treatment and no treatment respectively. These marginals can be identified using experimental data from randomized treatment allocation. If we fix the treatment probability as $p(\cdot)$, then the C.D.F. of the overall outcome corresponding to the choice $p(\cdot)$ is

$$G(\cdot; p) = \int_{x \in \bar{\mathcal{X}}} \left[F_1(\cdot|x) \times p(x) + F_0(\cdot|x) \times \{1 - p(x)\} \right] dF_X(x).$$

If the planner wishes to maximize a functional $\mathcal{F}(\cdot)$ of the C.D.F. $G(\cdot, p)$, then the optimization problem reduces to

$$\max_{p(\cdot)} \mathcal{F}(G(\cdot; p)) \text{ s.t. } c = \int_{x \in \bar{\mathcal{X}}} p(x) dF_X(x).$$

For the mean utility case studied in this paper, $\mathcal{F}(G(\cdot,p)) \equiv \int_{\mathcal{X}} U(w)\, dG(w,p)$, with $U(w) \equiv w$ denoting the mean outcome case. Similarly, $\mathcal{F}(G(\cdot;p)) \equiv G^{-1}(0.5;p)$ denotes the median maximization case. The form of the solution and the related distribution theory will of course change depending on the choice of $\mathcal{F}(\cdot)$.

A simple extension can also be made to the case where the treatment cost varies by x. Let $h(x)$ denote the per capita additional cost of treating type x. This is estimable from the experimental data when cost data are available. The original problem is now modified to

$$\begin{aligned} & \max_{p(\cdot)} \int_{x\in\bar{\mathcal{X}}} \left[\beta_1(x)\, p(x) + \beta_0(x)\left\{1 - p(x)\right\}\right] dF_X(x), \\ \text{s.t. } c &= \int_{x\in\bar{\mathcal{X}}} h(x)\, p(x)\, dF_X(x). \end{aligned}$$

The solution is of the cost-benefit form: treat if and only if $\beta(x) - h(x) \geq \gamma$, with

$$\int_{x\in\bar{\mathcal{X}}} h(x)\, 1\left(\beta(x) - h(x) \geq \gamma\right) dF_X(x) = c.$$

3 Sample-based rule

The rule $1\left\{\beta(x) \geq \gamma\right\}$ is infeasible because $\beta(\cdot)$ (and usually also $F(\cdot)$) are unknown. But we can approximate it with sample data from an experiment where the treatment was randomly assigned. In particular, we consider Manski (2004)'s Conditional Empirical Success (CES, henceforth)-type rule which is the sample analog of the rule in the previous display, viz.,

$$p(x) = 1\left\{x : \hat{\beta}(x) \geq \hat{\gamma}\right\} \text{ with } c = \frac{1}{n}\sum_{i=1}^{n} 1\left(\hat{\beta}(x_i) \geq \hat{\gamma}\right).$$

Here $\hat{\beta}(x)$ an $\hat{\gamma}$ are consistent estimates of $\beta(x)$ and γ.[4] One can then define the expected welfare of the feasible rule as

$$\rho_n := \int_{\mathcal{X}} \left[\beta_1(x) \Pr\left\{\hat{\beta}(x) \geq \hat{\gamma}\right\} + \beta_0(x) \Pr\left\{\hat{\beta}(x) < \hat{\gamma}\right\}\right] dF_X(x), \tag{5}$$

where $F_X(\cdot)$ represents the population distribution of covariate X, based on which the treatment will be allocated to new applicants. Under regularity conditions (e.g., uniform

[4]It is worth investigating whether the CES rule itself is optimal in a Hirano-Porter (2009) sense here but this question is outside the scope of the present paper. It would require extension of their methods to allow for unknown parameters, such as γ, which cannot be estimated at the $\sqrt{n}$-rate.

convergence of $\hat{\beta}(\cdot)$ to $\beta(\cdot)$, detailed in the appendix) ρ_n will converge to ρ defined in (2) as $n \to \infty$.

The condition $c = \frac{1}{n}\sum_{i=1}^{n} 1\left(\hat{\beta}(x_i) \geq \hat{\gamma}\right)$ is one that holds in the sample and will hold in the population only asymptotically. In order to have a $\hat{\gamma}$ that will make the budget constraint hold exactly in the population, the planner needs to know the marginal distribution of X, whence $\hat{\gamma}$ may be obtained by solving

$$c = \int_{\mathcal{X}} 1\left(\hat{\beta}(x) \geq \hat{\gamma}\right) dF_X(x).$$

In our empirical application, we do not know the marginal distribution of X and so we stick to the sample budget constraint formulation with the understanding that the resulting choice of $\hat{\gamma}$ will satisfy the budget constraint approximately in the population.

The technical results below concern the estimation of ρ_n, its inverse and the large sample properties of these estimators. This distribution theory, which is significantly more interesting than in the unconstrained case where γ is known to be zero, is used to (i) construct asymptotically valid confidence intervals for expected welfare ρ_n and its dual (introduced below) and (ii) address the issue of covariate choice in treatment allocation. Interestingly, it turns out here that the estimator $\hat{\rho}_n$ of ρ_n (and thus of ρ in large samples) has the parametric rate of convergence but the corresponding $\hat{\gamma}$ is not $\sqrt{n}$-consistently estimable in general. To see why, consider the special case where X is a scalar continuous covariate with c.d.f. $F_X(\cdot)$ and $\beta(\cdot)$ is strictly increasing but otherwise unknown. Then

$$1 - c = \Pr(\beta(X) < \gamma) = \Pr\left(X < \beta^{-1}(\gamma)\right) = F_X\left(\beta^{-1}(\gamma)\right).$$

Inverting this, we get that

$$\gamma = \beta\left(F_X^{-1}(1-c)\right) = E\left(Y_1 - Y_0 | X = F_X^{-1}(1-c)\right).$$

Thus γ is the nonparametric conditional expectation evaluated for the marginal treatment recipient and is therefore not $\sqrt{n}$-consistent even when the marginal of X is known. Although the distribution theory for $\hat{\gamma}$ is not of primary interest for the rest of the paper (it is used to get the distribution of $\hat{\rho}$), our result is somewhat interesting from a theoretical perspective. It shows that when $\beta(\cdot)$ is an infinite-dimensional unknown parameter and cannot be estimated in a $\sqrt{n}$-consistent way, then a quantile of the scalar random variable $\beta(X)$ is generally not $\sqrt{n}$-consistently estimable although the mean of $\beta(X)$ is. To our knowledge, this result is new and appears to us to be somewhat interesting because in most cases of interest, quantiles and means have the same rates of convergence.

Note that we can define the relevant population quantities, viz., γ, $\beta(\cdot)$ and ρ as solutions to the (semiparametric) moment conditions

$$\begin{aligned} E(Y_1 - Y_0 - \beta(X)|X) &= 0, \\ E[1(\beta(X) \leq \gamma) - 1 + c] &= 0, \\ E[\beta(X) \times 1(\beta(X) \geq \gamma) - \rho] &= 0. \end{aligned}$$

Of these, the second moment condition is differentiable in the scalar γ if $\beta(X)$ has a density but, in general, not functionally differentiable in $\beta(\cdot)$, owing to the presence of the indicator. This makes it infeasible to directly apply the methods of e.g., Chen, Linton and Van Keilegom (2003) which requires (Hadamard) differentiability of all the *population* moment conditions with respect to both the finite and the infinite dimensional parameters.

4 Estimation

We now formally define our estimators corresponding to the CES approach defined above. Suppose $\hat{\beta}(x)$ denotes an estimator of $\beta(x)$ and $\hat{\beta}_1(x)$ an estimator of $\beta_1(x)$. In order to circumvent the nonsmoothness of the population moment for γ discussed above, we use extra smoothing to construct our estimators. Suppose that $\beta(X)$ is bounded between $[-M, M]$ on the support of X. Then choose a symmetric (about zero) kernel $L(\cdot)$ with bounded support, w.l.o.g. $[-1, 1]$, the corresponding C.D.F. kernel $\bar{L}(t) = \int_{-1}^{t} L(a)\, da$ for each $t \in [-1, 1]$ and a sequence of bandwidths h_n converging (slowly) to zero as $n \to \infty$. The C.D.F. kernel simply converts the indicator function $1\{\beta(x) \leq \gamma\}$ to a function that changes smoothly from 0 to 1 as $\beta(x) - \gamma$ changes sign from positive to negative in finite samples but approaches the indicator as $n \to \infty$.

Now define $\hat{\gamma}$, and $\hat{\rho}$ by

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\left\{\bar{L}\left(\frac{\hat{\gamma} - \hat{\beta}(X_i)}{h_n}\right) - \{1 - c\}\right\} &= 0, \\ \hat{\rho} - \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\beta}_1(X_i) - \hat{\beta}(X_i)\,\bar{L}\left(\frac{\hat{\gamma} - \hat{\beta}(X_i)}{h_n}\right)\right] &= 0. \end{aligned} \tag{6}$$

Notice that the solution to the first equation in the previous display is unique. Indeed, the derivative of the LHS of the first equation w.r.t. $\hat{\gamma}$ is positive since $\bar{L}(\cdot)$ is the integral of a (positive) kernel. Given that $\hat{\gamma}$ can be uniquely solved, the solution to the second

equation must necessarily be unique because $\hat{\rho}$ is *defined* as the average of something involving $\hat{\gamma}$ and other observed quantities. The smoothing applied in (6) is similar in spirit to Horowitz's (1992) analysis of the smoothed maximum score (see also Hall et al (1999)). But in that problem, the finite-dimensional parameter of interest does not explicitly depend on any infinite-dimensional underlying parameter. In contrast, here the key parameters of interest, γ and ρ, are based on the infinite-dimensional component $\beta(\cdot)$ through population moments that are not smooth in $\beta(\cdot)$. Thus the present estimators lie at the intersection of classical 2-step semiparametric estimators and smoothing-based estimators for countering non-differentiability. This makes both the results and the proofs substantially different from both strands of the literature.

We now consider two cases as follows.

Parametric Case: First consider the case where we have a finitely parametrized regression model for Y giving us $\beta(x) = G(x,\beta)$ and $\beta_1(x) = G_1(x,\beta)$, where the functional forms G and G_1 are known and β is a finite dimensional parameter. The efficient estimators of β can be constructed here either by a direct optimal minimum distance approach or by a one-step GMM procedure. For general non-linear models with a distributional condition, the MLE will be efficient.

Nonparametric case: Suppose $X \equiv \left(X^d, X^c\right)$ where X^d contains the discrete components of X and X^c is a p-variate vector of the continuous components of X with support $\bar{\mathcal{X}}^{\,c}$ and density $f(\cdot)$. Let $K(\cdot)$ be a density kernel and ω_n a sequence of bandwidths converging to zero at an appropriate rate, to be specified later, as $n \to \infty$. Define $\hat{\mu}_1(X_i)$, $\hat{\mu}_0(X_i)$, $\hat{\pi}_1(X_i)$, $\hat{\pi}_0(X_i)$ as the Nadaraya-Watson regression estimators of respectively DY, $(1-D)Y$, D and $(1-D)$ on X, evaluated at $X = X_i$, and calculated while leaving out the ith observation, e.g.,

$$\hat{\mu}_1(X_i) = \frac{\frac{1}{n-1}\sum_{j\neq i}\frac{y_j D_j}{\omega_n^p} K\left(\frac{X_j^c - X_i^c}{\omega_n}\right) 1\left(X_j^d = X_i^d\right)}{\frac{1}{n-1}\sum_{j\neq i}\frac{1}{\omega_n^p} K\left(\frac{X_j^c - X_i^c}{\omega_n}\right) 1\left(X_j^d = X_i^d\right)}.$$

Now $\hat{\beta}(X_i)$ can be defined in terms of the above quantities as

$$\hat{\beta}(X_i) = \frac{\hat{\mu}_1(X_i)}{\hat{\pi}_1(X_i)} - \frac{\hat{\mu}_0(X_i)}{\hat{\pi}_0(X_i)}.$$

For a large number of discrete regressors, one can also use multi-dimensional kernels (typically product of several unidimensional kernels) and smooth across the discrete values as well. As is well-known, this does not affect the asymptotic distribution theory under

the same bandwidth conditions (c.f. Bierens (1994), theorem 10.4.3, Li and Racine (2007), chapter 4). If, however, in place of a continuous regressor, we have a discrete regressor taking infinitely many values (such as a Poisson random variable), one can simply use cell-averages to estimate conditional means. Following the arguments of Delgado and Mora (1995), it follows that smooth functionals of such nonparametric estimators are $\sqrt{n}$-normal, with no bias term. In our application, all discrete variables are binary and the key continuous variable (wealth) does appear to be continuously distributed. So we work with the traditional Nadaraya-Watson type estimators.

Fixed Trimming: To avoid some technical issues related to boundary bias, we use the following "fixed trimming" modification, as in Newey (1994), theorem 4.1. Let $\bar{\mathcal{X}}$ denote the support of X and let $\mathcal{X}$ be a compact subset of the interior of $\bar{\mathcal{X}}$, such that the density of X is bounded away from zero on $\mathcal{X}$. Then to define our estimands, we will simply work within $\mathcal{X}$. Notice that in this problem, which X's and what range of X's we want to use for conditioning treatment assignment is up to us. So the fixed trimming condition can be interpreted as part of the problem description.[5]However, we will suppress the trimming term $1\left(X_i \in \mathcal{X}\right)$ in the rest of our exposition to prevent notational clutter.

Finite sample bias: Note that our estimator $\hat{\rho}$ may be expressed as

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_1\left(X_i\right) - \frac{1}{n}\sum_{i=1}^{n}\hat{\beta}\left(X_i\right) \times \bar{L}\left(\frac{\hat{\gamma}-\hat{\beta}\left(X_i\right)}{h_n}\right).$$

The second term is composed of the product of $\hat{\beta}\left(X_i\right)$ and the term $\bar{L}\left(\frac{\hat{\gamma}-\hat{\beta}(X_i)}{h_n}\right)$ and the latter is large when $\hat{\beta}\left(X_i\right)$ is small. Because of the local variance in $\hat{\beta}\left(x\right)$, there is potentially a negative correlation in the estimation errors in $\hat{\beta}\left(X_i\right)$ and $\bar{L}\left(\frac{\hat{\gamma}-\hat{\beta}(X_i)}{h_n}\right)$ which would tend to introduce a positive bias in the estimate of ρ. In order to mitigate the bias, one option is to use a split-sample method as first suggested in Geisser (1975). A simulation exercise using this method is reported in the appendix, part 2. Given that the split-sample results seem to be worse to us–probably because the effective sample size is halved– we conduct the nonparametric application using full-sample methods. We will return to split-sample methods again when we discuss covariate selection but in that case, we restrict attention to parametric specifications, as suggested by a referee.

[5]Imbens and Ridder (2009) have recently considered some alternatives to fixed trimming using a "nearest interior point" modification to NW estimates.

5 Large sample theory

In this section, we consider the asymptotic distribution theory for our estimators. Section 5.1 outlines the simpler results for the parametric case where $\beta(x)$ is known up to a finite-dimensional parameter. Section 5.2 focuses on the nonparametric case. For the latter, theorem 1 concerns asymptotic properties of $\hat{\gamma}$, theorem 2 concerns $\hat{\rho}$. The key result is that $\hat{\rho}$ but not $\hat{\gamma}$ has parametric rate of convergence under regularity conditions.

5.1 Parametric Case

To get an idea for the distribution theory, consider the set-up where $\beta(x)$ is parametrically specified as $G(x,\beta)$, with $G(\cdot)$ known; typically β (the so-called "pseudo-true value") is a slope coefficient in a linear or non-linear regression and can be estimated at parametric rates using, say, GMM or the MLE. For some specific functional forms of $G(\cdot,\cdot)$, e.g., a linear one, the function $h(\beta) = \int_{-M}^{M} 1\{G(x,\beta) \leq \gamma\} dF_X(x)$ may be differentiable in β and then no smoothing would be necessary; but smoothing-based methods (i.e., using L) are more generally applicable and so we focus on that. The key result is that both γ and ρ can be estimated at the $\sqrt{n}$-rate. We provide an outline of proof in the appendix at the end of part 1.

5.2 Nonparametric Case

The nonparametric case is far more involved and occupies the bulk of our technical proofs in the appendix. It has the obvious advantage of being robust to distributional misspecifications and is therefore our preferred method. We now present two results regarding the large sample behavior of our estimators of γ and ρ. The exact conditions under which these hold are listed in the appendix together with the proofs of the results.

The next theorem deals with the distribution of $\hat{\gamma}$ which is ultimately needed for obtaining confidence intervals for ρ_n. The key assumptions for the following theorem 1 are that for some $r \geq 2$, density of $\beta(\cdot)$ has at least $r-1$ uniformly bounded derivatives, the $(r-1)$th one denoted by $f^{(r-1)}(\cdot)$, the bandwidth sequence satisfies $nh_n^{2r+1} \to \lambda < \infty$, $L(\cdot)$ is an rth order kernel satisfying $\int_{-1}^{1} u^r L(u)\, du < \infty$, and $\sigma_j^2(x) := var(Y_j | X = x)$ are finite for $j = 0, 1$. The asymptotic bias of $\hat{\gamma}$ will be proportional to $\sqrt{\lambda}$ which can be made to go to zero by choosing h_n to be of smaller order than $n^{-\frac{1}{2r+1}}$ but at the cost of increasing the variance (see below for more on this).

Theorem 1 *Under conditions A0-A3, A4(i), conditions B3(i) and B4(i) and B5, listed in the appendix, we have that*

$$\hat{\gamma} - \gamma = o_p(1).$$

Further, under conditions A0-A4 and B3-B11, detailed in the appendix, we have that

$$\sqrt{nh_n}(\hat{\gamma} - \gamma) \xrightarrow{d} N\left(\kappa, \frac{\tau^2(\gamma) + \xi^2(\gamma)}{f_\beta(\gamma)} \int_{-1}^{1} L^2(u)\, du\right),$$

where

$$\begin{aligned}
\tau^2(\gamma) &= E\left\{\frac{\sigma_0^2(X)}{\Pr(D=0|X)} | \beta(X) = \gamma\right\} \\
\xi^2(\gamma) &= E\left\{\frac{\sigma_1^2(X)}{\Pr(D=1|X)} | \beta(X) = \gamma\right\} \\
\kappa &= (-1)^r \frac{\sqrt{\lambda}}{r!} \times f_\beta^{(r-1)}(\gamma) \int_{-1}^{1} u^r L(u)\, du.
\end{aligned}$$

Proof. Appendix ■

Finally, we state the result for $\hat{\rho}$, under somewhat stronger conditions than the ones above. The r defined above now needs to be at least 4 and the kernel L is then of a "higher-order" variety. Define

$$\begin{aligned}
\psi_{1i} &= \gamma\{F_\beta(\gamma) - 1(\beta(X_j) \leq \gamma)\}, \\
\psi_{2i} &= \beta(X_i) \times 1\{\beta(X_i) \leq \gamma\} - \zeta, \\
\psi_{3i} &= 1(\beta(X_i) \leq \gamma) \times \frac{D_i}{\Pr(D=1|X_i)}\{Y_{1i} - E(Y_1|X_i)\}, \\
\psi_{4i} &= 1(\beta(X_i) \leq \gamma) \times \frac{1-D_i}{\Pr(D=0|X_i)}\{Y_{0i} - E(Y_0|X_i)\}, \\
\psi_i &= \psi_{1i} + \psi_{2i} + \psi_{3i} - \psi_{4i}.
\end{aligned}$$

Theorem 2 *Under conditions A0-A4 and B3-B11, listed in the appendix, we have that*

$$\hat{\rho} - \rho_n = o_p(1).$$

Further, under conditions A0-A5, B3-B12, listed in the appendix,

$$\sqrt{n}(\hat{\rho} - \rho_n) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} D_i \frac{Y_{1i} - \beta_1(X_i)}{\Pr(D=1|X_i)} - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_i + o_p(1). \tag{7}$$

Proof. Appendix. ■

The proof works by using

$$\hat{\rho} = \frac{1}{n}\sum_{i=1}^{n} \hat{\beta}_1 (X_i) - \frac{1}{n}\sum_{i=1}^{n} \hat{\beta} (X_i) \times \bar{L}\left(\frac{\hat{\gamma} - \hat{\beta}(X_i)}{h_n}\right).$$

The first term is then analyzed via lemma 3 in the appendix and a Taylor expansion coupled with U-statistics results are used to show that the second term has an asymptotically linear form. The final variance can be consistently estimated using sample cross-products, under standard conditions for the WLLN. It may be noted here that the estimation error in $\hat{\beta}(\cdot)$ affects the distribution of $\hat{\rho}$ through the terms ψ_{3i} and ψ_{4i}.

Bias Removal: Our proofs of the above results use bias-removal. If $f_\beta(\cdot)$ has bounded derivatives up to order $(r-1)$, then the MSE of $(\hat{\gamma} - \gamma)$ is given by

$$h_n^{2r} \times \underbrace{\left[\frac{f_\beta^{(r-1)}(\gamma)}{r!}\int_{-1}^{1} u^r L(u)\, du\right]^2}_{C} + \frac{1}{nh_n}\underbrace{\left[\frac{\tau^2(\gamma) + \xi^2(\gamma)}{f_\beta(\gamma)}\int_{-1}^{1} L^2(u)\, du\right]}_{B},$$

implying an MSE minimizing bandwidth choice of $h_n = \lambda^* n^{-\frac{1}{2r+1}}$, where $\lambda^* = \left(\frac{C}{2rB^2}\right)^{\frac{1}{2r+1}}$. This choice of h_n does not work for getting a $\sqrt{n}$-rate for $\hat{\rho}$ because (c.f. step 6A in the proof) it implies that $\sqrt{n}h_n^r = O\left(n^{\frac{1}{2(2r+1)}}\right)$ which blows up to $+\infty$. So we need to choose h_n to be smaller than the one that is MSE-optimal for γ.

Bandwidth Conditions: The conditions in the appendix impose several restrictions on bandwidths which are related to the dimension p of X, the extent of smoothness r of the density of $\beta(X)$ (higher r implying faster convergence for $\hat{\gamma}$), smoothness q of the function $\beta(\cdot)$ (higher q implying faster convergence for $\hat{\beta}(\cdot)$) and the number of moments m of the dependent variables. They imply sufficient moment and bandwidth conditions for a rate of uniform convergence, obtained in Hansen (2008) and Newey (1994). It is useful to have a specific choice of bandwidths which imply the conditions for (7) (which requires the strongest conditions). We can choose the bandwidth ω_n for estimating $\beta(\cdot)$ as $\omega_n = n^{-\omega}$ and the bandwidth $h_n = n^{-h}$ for some $w, h > 0$ which satisfy the following sufficient restrictions: $r \geq 4$, $m \geq 4$, $q \geq \frac{r+2}{r}p$, $\frac{1}{6} > h > \frac{1}{2r}$, $\frac{5}{12} < q\omega < 1$ and $p\omega < \frac{1}{4}$. This choice will satisfy all the conditions for (7) and, in particular satisfy sufficient conditions for the result that $\sup\left|\hat{\beta}(\cdot) - \beta(\cdot)\right| = o_p\left(n^{-1/4}\right)$. In particular, for $p = 1$, one can assume $q = 3$, $r = 4$, $m = 4$ and set $h = \frac{1}{7}$ and $\frac{5}{24} < \omega < \frac{1}{4}$, which is what we use in the application.

Distribution theory for dual: Recall that the value function for the dual problem $\delta_n(b)$ represents the smallest fraction of households who have to be assigned to treatment (optimally) to guarantee that the expected mean outcome is at least b. In other words, $\rho_n[\delta_n(b)]$ equals b, where $\delta_n(b)$ plays the role of c in the primal problem. The $\delta_n(b)$ is estimated by the sample analog $\hat{\delta}(b) = \hat{\rho}^{-1}(b)$. Using a standard first-order expansion, and noting that $\rho'(c)$ equals $\gamma(c)$, one gets that

$$\sqrt{n}\left(\hat{\delta}(b) - \delta_n(b)\right) = -\frac{\sqrt{n}\left(\hat{\rho}\left(\delta(b)\right) - \rho_n\left\{\delta(b)\right\}\right)}{\gamma\left\{\delta(b)\right\}} + o_p(1),$$

from which the asymptotic normality of $\sqrt{n}\left(\hat{\delta}(b) - \delta_n(b)\right)$ follows by applying the above theorems.

6 Choosing Covariates

The above analysis has considered treatment targeting by a fixed set of covariates and experimental estimates from large samples were used to determine CES type allocation rules based on these covariates. The resulting expected welfare (ρ_n) is then approximated in a large-sample sense by an appropriate sample analog ($\hat{\rho}$). However, one practical consideration in designing treatment protocols is to decide which covariates to choose for targeting. A brief outline of the problem and a practical solution based on cross-validation is discussed below. A full-blown theoretical analysis of this question is beyond the scope of the present paper. But in appendix subsection 4, we provide an outline of a potential theoretical approach, based to local-to-zero type asymptotics, for analyzing the problem.

Basically, there are two considerations involved. One is the financial cost of gathering covariate information required for implementing the desired allocations. The second is the issue of statistical precision of covariate-conditioned treatment effects obtained in the finite experimental sample.

For the following discussion (as recommended by a referee), we concentrate on the case where $\beta(\cdot)$ is parametrically specified as $x'\beta$, $\beta \in \mathbb{R}^k$ and let $\gamma := \gamma(\beta, F_X(\cdot))$ denote the treatment threshold.

Finite sample accuracy: Suppose we want to choose one among a set of covariate combinations, indexed by $\{1, ..., J\}$ with β and $\hat{\beta}_j$ denoting the parameter and its MLE, respectively, corresponding to covariate set j. Analogously, denote the treatment threshold and its estimator by γ_j and $\hat{\gamma}_j$. The case $j = 0$ will denote the full covariate set. Let

δ denote assigning an individual to treatment and consider the loss function

$$\mathcal{L}(\delta;(\beta,F)) = \int (x'\beta - \gamma(\beta,F)) \times [1(x'\beta > \gamma(\beta,F)) - \delta]\, dF(x) \tag{8}$$

The above loss function penalizes when knowledge of the true (β, F) would recommend giving $D = 1$ (0) but δ leads to recommendation of 0 (1). When using all covariates, the CES rule is of the form $\delta = 1\left\{x'\hat{\beta} > \gamma\left(\hat{\beta}, \hat{F}\right)\right\}$, where $\left(\hat{\beta}, \hat{F}\right)$ denote estimates of (β, F). Now consider two rules, one where allocation is based on covariate set j (with respective parameters β_j and γ_j) and one where it is based on the full feasible set of covariates. Using an Edgeworth expansion, the risk difference between using the full set versus the set $j = 1$, say, up to first order is given by

$$\begin{aligned} & R((\beta,F),\sigma_0(\cdot)) - R_1((\beta,F),\sigma_1(\cdot)) \\ :\quad & = \rho_{1n} - \rho_n \\ = \quad & \int (x'\beta - \gamma(\beta,F)) \times \left[\Phi\left(\frac{x'\beta_1 - \gamma(\beta_1,F)}{\sigma_1(x)/\sqrt{n}}\right) - \Phi\left(\frac{x'\beta - \gamma(\beta,F)}{\sigma(x)/\sqrt{n}}\right)\right] dF(x), \end{aligned}$$

where $\sigma_1(x)$ and $\sigma(x)$ denote respectively the asymptotic standard deviation of $x'\hat{\beta}_1 - \hat{\gamma}_1$ and $x'\hat{\beta} - \hat{\gamma}$. This expression sheds light on the role of estimation precision. Note that whenever $x'\beta - \gamma(\beta,F)$ and $x'\beta_1 - \gamma\left(\beta^1,F\right)$ have different signs, the integrand will be negative irrespective of the values of $\sigma_1(x)$ and $\sigma(x)$. When they have the same sign, however, the integrand will tend to be positive whenever $\sigma_1(x)$ is small relative to $\sigma(x)$. This suggests that the narrower model will yield higher welfare in finite samples if covariates not included in it have small impact on the treatment effect and/or there is a large amount of variation in these covariates in the sample (leading to smaller $\sigma_1(x)$).

Financial costs: Using a larger number of conditioning covariates can yield better (at least, no worse) targeting if our estimates come from large samples. However, measuring more covariates is also financially expensive. These measurement costs are incurred for everyone in the population during the actual treatment assignment and not just for individuals getting the treatment. Furthermore, unlike statistical precision costs, the measurement costs do not disappear with increasing sample size. Suppose we are choosing whether to condition treatment on a larger set X_0 or a smaller subset X_1 where gathering information about X_j entails a per capita expenditure of v_j, $j = 1, 2$. Suppose these two conditioning sets yield expected welfare functions $\rho_{0n}(\cdot)$ and $\rho_{1n}(\cdot)$, respectively. Also suppose b is the average welfare we want to attain. Then the expected per capita cost of using X_1 is the sum of the treatment cost plus the cost of measuring X_1,

i.e., $\rho_{1n}^{-1}(b) + v_1$.[6] Similarly for X_0. Therefore, we should choose X_0 if $\rho_{1n}^{-1}(b) + v_1$ is large relative to $\rho_{0n}^{-1}(b) + v_0$.

6.1 Feasible Rules

Note, however, that since the ρ_{jn}'s are unknown, they cannot be used directly in choosing j and feasible sample-based rules are warranted. This problem is a somewhat complicated version of classic model selection problems, on which there is a large literature in statistics and time-series econometrics. The complication arises here because the underlying constrained optimization problem makes the welfare function depend on both the conditional mean of Y_1, Y_0, and the marginal distribution function of the covariates in a complicated way.[7]

In the frequentist approaches to model selection in simpler settings, one aims to estimate the risk arising from the use of estimates in place of true values, i.e., in the present context, estimate ρ_{jn}. See, for instance, Claeskens and Hjort (2008, CH hereafter) for a comprehensive textbook treatment. For reasons of brevity, we concentrate on the method cross-validation for the present paper but note that other frequentist approaches, e.g., resampling based methods and Bayesian approaches could be potentially used here, as well.

Cross-validation (CV): The CV idea is to split the sample into a training part (containing $n - n_1$ observations) and a validation part (containing n_1 observations). For a given covariate choice j, for each such split, the parameters β and γ are estimated using the training part –call these $\hat{\beta}_{jt}$ and $\hat{\gamma}_{jt}$. Then these estimates are used to make allocations for the validation part and the welfare for that particular split is calculated as:

$$\tilde{\rho}_{jn} := \frac{1}{n_1}\sum_{i=1}^{n_1}\left[\frac{Y_i D_i}{\frac{1}{n_1}\sum_{i=1}^{n_1} D_i} 1\left\{x_i'\hat{\beta}_{jt} \geq \hat{\gamma}_{jt}\right\} + \frac{Y_i(1-D_i)}{\frac{1}{n_1}\sum_{i=1}^{n_1}(1-D_i)} 1\left\{x_i'\hat{\beta}_{jt} < \hat{\gamma}_{jt}\right\}\right]. \quad (9)$$

The overall welfare is calculated by averaging the per-split welfare across all possible splits of the sample. A particularly popular split in statistics is the leave-one-out version.

[6] Notice that in this case, v and ρ^{-1} are both measured in monetary units and are thus directly comparable.

[7] Constructing Manski (2004)-style bounds on exact finite sample risk for different covariate choice is difficult here owing to the complicated nature of the probabilities which involve, among other things, finite sample Lorenz shares for the finite sample conditional mean function.

The idea here is to predict welfare from using estimates from samples of size $n-1$ to design the allocation for for the remaining observation. Averaging across the sample observations then mimics the average welfare of the CES rule across samples drawn from the population.[8] A modification of the above is provided by V-fold CV (c.f., Geisser (1975)) – where the sample is split into V disjoint subsamples randomly; then each of the V subsamples is used for validation and the result averaged across them.

In order to compare the accuracies of the cross-validated and the naive sample-based estimates of welfare in finite samples, we performed a small Monte Carlo exercise based on our sample data, which is reported in the next section. In the actual application, we complement our results based on the asymptotic approach of section 4 with a leave-one-out CV approach (which gave the best results in the Monte Carlo for small samples) to see how covariate choice in the parametric specification may be dictated by sample size.

One issue we do not address here, though, is that of post model-selection inference, which may be relevant in practice, especially when one model is chosen over many competing ones. In the appendix subsection 4, we briefly outline an (local-to-zero type) asymptotic framework that seems appropriate for the covariate choice problem for finite samples and describe how that would relate to the post-model selection inference issue.

7 Application to bednet provision

7.1 Background

We now consider a substantive application –viz, the allocation of subsidies for long-lasting insecticide-treated nets (ITNs) to households, using experimental evidence from Kenya. Our treatment of interest is making subsidized ITNs available to a section of this population.

ITNs have been shown to reduce overall child mortality by up to 38 percent in regions of Africa where malaria is the leading cause of death among children under 5.[9] In addition, ITN use can help avert some of the substantial direct costs of treatment and the indirect

[8] It is possible to show that $\tilde{\rho}_{jn}$ is first-order equivalent to the naive $\hat{\rho}_{jn}$ minus a penalty term that is related to the variance of $\hat{\beta}_j$ (see appendix part 4, last sub-section).

[9] See Lengeler (2004) for a review. Earlier estimates of ITN use on reductions in child mortality from a randomized trial in Gambia were as high as 60 percent, but most estimates from randomized trials in Africa are closer to 20 percent.

costs of malaria infection on lost income.[10] Lucas (2007) estimates that, alone, the gains to education of a malaria-free environment more than compensate for the cost of an ITN. Costing $5 - $7 a net, however, ITNs are not affordable to most families (Cohen and Dupas, 2010; Dupas, 2010). For this reason, there is a large consensus that ITNs should be fully subsidized (WHO, 2007; Sachs, 2005).

Teklehaimanot et al (2007) estimate that providing one free long-lasting ITN for every two at-risk persons in sub-Saharan Africa would amount to 2.5 billion dollars. The funds committed by governments and donor agencies for ITNs have not yet reached that amount, however. For example, the Government of Kenya estimates that around 1 million pregnant women are in need of an ITN every year, but their budget will allow them to provide only 0.5 million nets per year to pregnant women over the next 5 years (Kenya Round 7 Proposal, 2007).

Under such a budget constraint, the question of how to allocate the available ITNs among households becomes an important policy question. If the treatment effect (the health impact of getting a subsidized ITN) is exactly the same for everyone in the population, then all possible allocations will lead to the same overall gains. However, when there is heterogeneity in the treatment effect (e.g. the health impact of getting a subsidized ITN varies with observed covariates, such as socioeconomic status, presence of children in the household, etc.), the gains can be maximized by a covariate-based allocation. While the health impact of using an ITN might be homogenous, the health impact of getting a highly subsidized ITN might vary across covariates since usage rates (conditional on having a net) are likely to vary across covariates. For example, households who can afford to purchase an ITN in the absence of any subsidy (because they have access to credit or are wealthy enough) will not benefit from the treatment very much (i.e. their $\beta_0(x)$ will be large and thus for them the difference $\beta_1(x) - \beta_0(x)$ is likely to be small). Likewise, since young children are the most vulnerable to the disease, households without young children might not benefit much from the treatment (i.e. their $\beta_1(x)$ will be small and thus the difference $\beta_1(x) - \beta_0(x)$ is likely to be small). For these reasons, the treatment effect is likely to vary across observed covariates such as wealth, access to financial services, and the presence of young children. An allocation rule that takes into account such heterogeneity could potentially generate important welfare gains.

[10] Ettling et al. (1994) find that poor households in a malaria-endemic area of Malawi spend roughly 28 percent of their cash income treating malaria episodes.

7.2 Design

For this application we use data from a randomized experiment conducted with rural households in Western Kenya in 2007 (Dupas, 2009). The price at which a household could purchase an ITN varied from \$0 (a free ITN) to \$4, in steps of \$0.50. Households were randomly assigned to a price. People had three months to redeem the voucher entitling them to an ITN at the assigned price. In this application, we consider two groups: households that faced a very low (highly subsidized) price (\$0 or \$0.50) and households that faced a high price of \$2 or more. Table 1 presents summary statistics on the 1007 households that form the sample used in the analysis.

The take-up rate of the ITN subsidy was 84% in the low price group and 13% in the high price group. Conditional on take-up, the usage rate was slightly higher in the low price group than in the high price group (70% versus 58%), leading to coverage rates of 60% and 8%, respectively. In what follows, we consider the low price group as the treatment group and the high price group constitutes the control. The treatment is thus "having access to a low-price ITN".[11]

The welfare measure we consider is the fraction of households covered by an ITN (i.e., the fraction that owns and uses an ITN). While incidence of malaria, health levels, wages lost due to malaria, or even consumption foregone due to wages lost due to malaria, could be the "ultimate" outcomes of interest, we concentrate on the ITN coverage rate for two reasons. First, as discussed above, multiple trials have established that ITN coverage leads to significant reductions in morbidity and mortality, especially among young children (see Lengeler (2004) for a review), and therefore the coverage rate is a good proxy for those ultimate outcomes. Second, in our data, coverage was observed directly (through surprise home visits during which enumerators checked whether the ITN was hung above a bed), and therefore is not subject to reporting biases, in contrast with self-reported health measures.

To test for heterogeneity in the treatment effect, we run an OLS regression of ITN coverage on the treatment, three covariates, and the interactions between the treatment

[11] One may note that the short-run take-up in the low price group was not 100%, since some of the "treated" had to pay a small fee (i.e., \$0.50) to access the net. In the analysis below, however, we assume that take-up was 100% in the treatment group. In other words, we consider that those who did not take-up the subsidy *cost as much* to the government as those who took-up the subsidy but didn't use their net. This is because the take-up would potentially have reached 100% if people had had more than three months to redeem their voucher.

and the covariates. The covariates are: a binary variable equal to 1 if the household includes at least one child under 10; the natural log of the value of the household's wealth per capita; and a binary variable equal to 1 if the household owns a bank account. The first covariate (presence of a child) was chosen as an indicator of the private returns to using a bed net (since young children are the most vulnerable to malaria). The two other covariates were chosen as proxies for socioeconomic status and ability to pay. They were measured through a baseline survey administered through household visits. In particular, wealth per capita was measured as follows: households were asked to list all their assets (including animal assets) and to estimate their resale value. The combined value of all assets was then divided by household size to obtain the "wealth per capita" indicator. Such wealth is typically held in relatively illiquid assets, however. For this reason, we include "ownership of a bank account" as a second ability-to-pay measure. Savings held in a bank are perfectly liquid, and in our study area, access to formal financial services has been shown to be an important determinant of households' ability to save and acquire lumpy durables (Dupas and Robinson, 2009). Of the three covariates we use, the most costly one to measure is wealth, which typically requires a lengthy home visit by an enumerator. Conditional on making this visit, asking respondents about their household composition and their ownership of a bank account would not be onerous. One issue though is that of the reliability of such self-reported information, once people know the allocation rule. This is not a concern in our dataset since the respondents were explicitly told (through the informed consent process) that their answers would have no bearing on what treatment they would be eligible for in the future, but in an actual program, once the allocation rule is well-known, collecting reliable information on the covariates that determine eligibility might be quite costly. This cost would then need to be accounted for in choosing covariates, as discussed above.

The results are presented in Table 2. The treatment was randomized at the household-level so no clustering correction is needed. We find that the treatment effect appears significantly higher for households with a child under 10 and significantly lower for households that own a bank account. The treatment effect is also lower for those with greater wealth, but the standard error is also large and the interaction term is not significant. An F-test of the joint significance of the three interaction terms is significant at 5%.[12]

[12]Two commonly used covariates are age and education level. They do not appear to affect the treatment effect in our sample, possibly because the sample is relatively homogeneous in terms of those

7.3 Analysis

7.3.1 Non-Parametric Analysis: Choice of Kernels and Bandwidths

For bias-removal, we use the kernels[13]

$$\begin{aligned} K(t) &= 0.5 \times \left(3 - t^2\right) \times \phi(t), \\ \bar{L}(t) &= \frac{15}{32}\left(\frac{7}{5}t^5 - \frac{10}{3}t^3 + 3t + \frac{16}{15}\right) \times 1(-1 \leq t \leq 1) + 1(t > 1), \end{aligned}$$

where $\phi(\cdot)$ is the standard normal density. Two bandwidths are needed for the non-parametric estimation: the bandwidth ω_n in the estimation of the conditional ATE $\beta(X)$, and the bandwidth h_n in the smoothing correction. Figure 1 graphs how the estimated treatment threshold $\hat{\gamma}$ (Panel A) and value function $\hat{\rho}$ (Panel B) vary with h_n for a range of possible ω_n. We find that both estimates are insensitive to the choice of h_n. In Figure 2, we present $\hat{\gamma}$ and $\hat{\rho}$ for two budget constraint levels: $c = 0.5$ (Panel A) and $c = 0.25$ (Panel B). The stability of $\hat{\rho}$ over a reasonable range of bandwidths suggests that the choice of bandwidths should have little effect on the nonparametric estimates of the value function.

Figure 3 graphs a leave-one-out cross validation criterion function for $\beta(x)$. The function is plotted over the range $\omega_n \in [0.3, 0.4]$, which correspond roughly to $n^{-1/6}$ and $n^{-1/8}$, respectively. The function seems to dip around $\omega_n = 0.33$. Given the small sensitivity of our estimates of ρ and, to a certain extent, γ to the choice of ω_n, we show the results for both $\omega_n = 0.3$ and $\omega_n = 0.4$. We use $h_n = 0.35$; recall that the results seem very insensitive to the choice of h_n for a given choice of ω_n.

7.3.2 Conditional ATE

The nonparametric estimate of the CATE $\hat{\beta}(x)$ was computed corresponding to two bandwidths $\omega_n = 0.3$ and $\omega_n = 0.4$. The parametric estimator of $\beta(X)$ was computed as $\hat{\beta}(x) = (\hat{\pi}_0 + x'\hat{\pi}_{01})$, where $\hat{\pi}_0$ and $\hat{\pi}_{01}$ are OLS estimates in the regression (presented in Table 2):

$$y_i = \theta_0 + X_i'\theta_1 + \delta_0 D_i + X_i'\delta_1 \times D_i + \varepsilon_i. \tag{10}$$

two characteristics.

[13] To see how our results are affected by choice of a higher order kernel, we repeated the analysis for a standard normal kernel instead of $K(t)$. The results are numerically very similar and do not imply any substantively different conclusion (results available upon request).

Figure 4, Panel A graphs the kernel density of the conditional ATE $\beta(X)$ computed with the two proposed bandwidths, as well as the parametric estimate for comparison purposes. Observations with X such that $\beta(X)$ is below -0.2 or above 0.9 were discarded in accordance with appendix condition A2. These cutoffs were the 1 and 99 percentile values in the empirical distribution of $\beta(X)$.

Figure 4, Panel B presents the c.d.f. of the conditional ATE $\beta(X)$ computed both parametrically and nonparametrically. The stepwise shape for the c.d.f. in the parametric model is essentially due to the binary nature of two of the three covariates since the interaction of the treatment with wealth appears to be close to zero in the parametric case. Finally, Panel C of Figure 4 presents the estimates of $\beta_1(x)$ and $\beta_0(x)$ as a function of the continuous regressor (wealth), for each cell defined by the discrete regressors.

7.3.3 Unrestricted and Restricted Value Functions

In what follows, we compare the "first best" allocation (the unrestricted case, in which the allocation is based on all three covariates) with two "restricted" cases: (i) means-testing where the allocation is based only on wealth– which is extremely common in both developed and developing countries, and (ii) purely random allocation which is not covariate-based at all (this is totally uncommon as far as we know, but makes for a convenient benchmark). Notice that in the random allocation case, the estimated value function is linear in c:

$$\hat{\rho}(c) = \frac{1}{n}\sum_{i=1}^{n}\left\{c \times \hat{\beta}_1(X_i) + (1-c) \times \hat{\beta}_0(X_i)\right\}.$$

Figure 5 graphs the parametric and nonparametric estimates for the treatment threshold $\gamma(c)$ and the value function $\rho(c)$ in the unrestricted case. The nonparametric estimates seem very stable over the two choices of bandwidth. The nonparametric estimates of the unrestricted value function are higher than the parametric estimates.

7.3.4 Welfare Losses

Figure 6 combines the estimates of the value function $\rho(c)$ for all three cases: allocation based on all three covariates (called unrestricted), allocation based on wealth only, and random allocation. Panel A presents the parametric estimates and Panel B presents the nonparametric estimates. Presenting all three cases on the same graph helps visualize

the welfare loss when the optimal allocation is not implementable. In contrast to the parametric estimates, the non-parametric estimates suggest that means-testing is a clear "second best", generating a higher mean outcome than random allocation does. The standard errors of the welfare losses generated by the two suboptimal allocations are shown in Table A1 for two budget constraints ($c = 25\%$ and $c = 50\%$).

7.3.5 Dual Problem

In Table 3 we report the minimum resources needed to attain a certain expected outcome: we compute the share of the population that needs to be treated in order to achieve a given target value function by allocating treatment based on all three covariates (column 2). We then calculate the additional resources that are needed when the optimal, unrestricted allocation is not possible, and the allocation is instead based on wealth only (column 3) or the allocation is purely random (column 4). The nonparametric estimates with the bandwidth $\omega = 0.4$ suggest that, compared to the unrestricted allocation, an allocation based on wealth only requires treating an additional 6.5 percentage points of the population compared to the optimal allocation (Panel B2, column 3). The additional spending is higher when the allocation is purely random: an extra 13.7 percentage points of the population need to be treated to reach the target usage rate, compared to the optimal allocation (Panel B2, column 4).

7.4 Estimating welfare in finite samples

The application above was based on the large sample approximation to welfare ρ_n generated from using a fixed set of covariates. The approximation is provided by $\hat{\rho}_n$, the sample-based estimate of the welfare. This approximation may be inaccurate in finite samples if the number of conditioning covariates is large, relative to the sample size. In order to compare the accuracies of $\hat{\rho}_n$, the naive sample-based estimate with that of $\tilde{\rho}_n$, the cross-validated estimate of finite sample risk, we perform a small Monte Carlo exercise on a parametric specification, as follows.

7.4.1 Monte-Carlo

Making random draws from the sample of interest, we create an artificial dataset of 1008 observations and remove the outcomes. We generate outcomes Y_1 and Y_0 through the

probit equation

$$Y_T = 1\left(b_0 + b_1 T + b_2' X + b_3' T \otimes X + e > 0\right),$$

where the b coefficients are chosen equal to the estimates from a probit regression of Y on the regressors, the treatment and the interaction of the regressors and the treatment, in the original sample. The error e is chosen to be 3 times a standard normal. This is used as our population. From this population, we draw samples of varying size. For each sample, we estimate β and γ through a linear (as opposed to probit) regression, calculate the implied allocations and, by averaging over many samples draws from the population, compute for each sample size (i) the actual welfare (ρ_n), (ii) the estimated welfare ($\hat{\rho}_n$) and the cross-validated welfare ($\tilde{\rho}_n$) where we use both a leave-one-out and a 2-fold CV.[14] The results are reported in Table 4A (estimated and true welfare), B (Bias and standard deviation of welfare estimator) and C (MSE of welfare estimator). We see in table 4A that as the sample size increases, ρ_ns are generally increasing as expected but $\hat{\rho}_n$ and $\tilde{\rho}_n$ are not. Table 4A also shows that for smaller sample sizes, the welfare from using fewer covariates is larger and in this case, the leave-one-out CV criterion favours the smaller model while $\hat{\rho}_n$ always favors the larger model.

From table 4B, we see that the bias of $\hat{\rho}_n$ in estimating ρ_n is larger in almost all cases than the bias in $\tilde{\rho}_n$s and this is more pronounced as the sample size falls. This is to be expected, given that CV is generically equivalent to a higher-order bias removal. It is also instructive that the bias of $\hat{\rho}_n$ is positive, which is consistent with the point made just before section 5. In terms of the MSE, reported in table 4C, the leave-one-out CV performs the best except when the sample size is relatively large and consequently $\hat{\rho}_n$ is at least as good as a leave-one-out CV. However, given that the leave-one-out CV criterion performs best in small samples and almost as well in larger samples as the naive estimator $\hat{\rho}_n$, we propose using it to choose between alternative subsets of targeting covariates, in the application. Those results are reported in table 5 and discussed next.

7.4.2 Cross-Validation Results for Application

We consider the parametric model (10) and two subsets of covariates – (i) household wealth – the most popular targeting criterion in both developed and developing countries and (ii) the combination of wealth, number of children under 10 and possession of a bank account.

[14] Results are indistinguishable between the nonsmooth and the smoothed version of (9) with h_n varying between $n^{-1/5}$ to $n^{-1/8}$.

We calculate the leave-one-out welfare estimate using three samples of different sizes. These samples were chosen by randomly sorting the original data of 1008 observations (with a specific seed) and choosing the top 500, 800 and 1008 observations. The welfare from using different subsets of regressors and different sample sizes are displayed in table 5.

The table can be read as follows. For the first panel with c=0.50, sample size=500, the number 0.3540 was obtained as follows. Choose the first 500 households, as described above. Then (i) regress the outcome for individuals randomized into treatment on wealth only and (ii) regress the outcome for individuals randomized out of treatment on wealth, bank account and child. For each individual in the data, predict their outcomes with the subsidy (Y_1) using the first set of estimated coefficients and their outcome without the subsidy (Y_0) using the second set of estimates. Calculate the difference and its median in the sample which is $\hat{\gamma}$. Then calculate the cross-validated welfare estimate as described in the previous section. Repeat for the other sample sizes. The highlighted values indicate the highest number among the four cells. The naive estimates ($\hat{\rho}$) are reported in parentheses.

From table 5 it is apparent that CV recommends using wealth only to target the subsidy when the sample size is less than 800. For sample size 800, the two conditioning sets yield similar levels of welfare and for the full sample, it is advisable to use all three covariates for both Y_0 and Y_1. Further, for the smaller sample sizes, it is nearly as effective to predict Y_1 using wealth only and Y_0 using all three covariates as it is to use wealth only for both Y_1 and Y_0. The reverse conditioning produces significantly lower welfare. This asymmetry arises most likely because there are many more observations with no treatment and so the additional covariates –children and bank account– have relatively low explanatory power for treatment effects when the household actually receives the subsidy and relatively high explanatory power when it does not. In contrast, the naive estimate, which ignores the finite sample inaccuracy, suggests using all 3 covariates for Y_0 for all sample sizes.

8 Conclusion

In this paper, we have considered a social planner's problem of allocating a binary treatment among a target population based on observed covariates in the presence of budget

constraints. The paper proposes a simple covariate-conditioned allocation rule based on sample data from a randomized experiment. The paper then derives and uses large-sample frequentist properties of these rules to infer the expected welfare from the rule and the minimum cost of attaining a specific average welfare– i.e., the dual problem. These methods are applied to data on the provision of anti-malaria bed nets in Western Kenya. The empirical findings are that a government which can afford to distribute bed net subsidies to only 50% of its target population can, if using an allocation rule based on multiple covariates, increase actual bed-net coverage by 17 to 20 percentage points relative to random allocation. These conclusions are based on large sample approximations and we investigate, using leave-one-out and 2-fold cross-validation and in both a Monte Carlo and the actual application data, their robustness to small sample inaccuracy. Our findings suggest that under parametric specifications and when using evidence from small experimental samples (under 800 in our application), it is better to condition the subsidy solely on household wealth; however, for our full sample of size 1008, the expected subsidy use increases by about 9 percent when we condition on two discrete household characteristics, viz. presence of small children and possession of a bank account, in addition to wealth.

This paper has left several topics to future research. A formal analysis of covariate-selection and post model-selection inference in treatment choice problems is being pursued by the present authors. This analysis uses tools from the frequentist literature on model selection, paying specific attention to bias-variance trade-offs implicit in various forms of cross-validation or resampling-based techniques. A particularly challenging task is to consider this problem when treatment responses are not assumed to have specific parametric forms. On the substantive front, one possible extension is to the design of conditional cash-transfer programs, which have gained popularity in a large number of central and south American countries. A second extension would be to incorporate treatment externalities into an analysis of efficient treatment assignment. Extension to multiple treatments would also be a theoretically interesting and practically relevant exercise.

A caveat to our analysis is the implicit condition that the covariate distributions are not affected by the targeting strategy used. This may be violated if the population composition changes in response to changes in the targeting rule, e.g., switching subsidy eligibility towards families with children in a district may see an influx of families with children from neighboring districts, thereby altering the marginal distribution of covari-

ates. Such migration is plausible only when the size of the transfer is high enough relative to migration costs and thus quite unlikely at the usual scale of in-kind transfer schemes present in the developing world today. But for larger sized transfers, this caveat can potentially be an important one.

The methods proposed here have wider applicability, beyond subsidy targeting in developing countries, to nearly any situation of constrained treatment assignment such as deciding eligibility rules for access to credit under aggregate fund constraints or allocating the unemployed to subsidized job-training programs when subsidy totals are limited by the government's budget outlay.

We would like to end with the observation that in development circles, there has been a recent push for more experimental evidence on the impact of social programs, as part of a general effort to improve the effectiveness of aid (Duflo, Kremer and Glennerster, 2008). For example, the World Bank recently launched the DIME initiative, an effort to increase the number of Bank-funded projects with impact evaluation components. We believe that as randomized trials of social programs, e.g., Oportunidades (PROGRESA) in Mexico, become more common in both developed and developing countries, our methodology will become increasingly relevant in helping governments and aid-agencies roll out positive-impact programs via efficient allocation rules.

References

[1] Andrews, Donald. W. K. (1999): "Estimation When a Parameter Is on a Boundary," Econometrica, 1999, 67: 1341-1383.

[2] Attanasio, Orazio, Costas Meghir and Santiago (2011): "Education Choices in Mexico. Using a Structural Model and a Randomized Experiment to Evaluate PROGRESA," forthcoming in Review of Economic Studies.

[3] Behnke, Stephanie, Markus Frölich and Michael Lechner (2009): "Targeting Labour Market Programmes – Results from a Randomized Experiment." Swiss Journal of Economics and Statistics, 145, 221-268.

[4] Berger, Mark, Dan Black and Jeffrey Smith (2001): "Evaluating Profiling as a Means of Allocating Government Services," in Michael Lechner and Friedhelm Pfeiffer (eds.), *Econometric Evaluation of Active Labour Market Policies*, Heidelberg: Physica, 59-84.

[5] Bhattacharya, D. (2009): "Inferring Optimal Peer Allocation using Experimental Data," Journal of the American Statistical Association, Vol. 104, No. 486: 486–500.

[6] Claeskens and Hjort (2008): Model Selection and Model Averaging, Cambridge University Press.

[7] Cohen, Jessica and Dupas,Pascaline (2010): "Free distribution or cost-sharing? Evidence from a randomized malaria experiment in Kenya". *Quarterly Journal of Economics, 125 (1): 1-45.*

[8] Dehejia, Rajeev H (2005): Program Evaluation as a decision Problem, *Journal of Econometrics*, vol. 125, no. 1-2, pp. 141-73.

[9] Delgado, M and J. Mora (1995): Nonparametric and semiparametric estimation with discrete regressors, Econometrica, vol. 63, no. 6, pp. 1477-1484.

[10] Duflo Esther, Rema Hanna and Stephen Ryan (2007). "Monitoring Works: Getting Teachers to Come to School", forthcoming, American Economic Review.

[11] Duflo, Esther, Michael Kremer and Rachel Glennerster (2006). "Using Randomization in Development Economics Research: A Tool Kit", Handbook of Development Economics, Volume 4, Editors T. Paul Schultz, pages 3895-3957 (2008).

[12] Dupas, Pascaline (2009). "What matters (and what does not) in households' decision to invest in malaria prevention?" *American Economic Review Papers and Proceedings*, 99(2): 224-30.

[13] Dupas, Pascaline (2010). "Short-Run Subsidies and Long-Run Adoption of New Health Products: Evidence from a Field Experiment". mimeo, UCLA.

[14] Dupas, Pascaline, and Robinson, Jonathan (2009). "Savings Constraints and Microenterprise Development: Experimental Evidence from Kenya". NBER WP #14693.

[15] Eberts, Randall W., Christopher J. O'Leary, and Stephen A. Wandner, (eds. 2002): *Targeting Employment Services*, Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

[16] Efron, B. (1986). How biased is the apparent error rate of a prediction rule?, J. Amer. Statist. Assoc., 81(394):461–470.

[17] Efron. B (1990): The jackknife, the bootstrap, and other resampling plans". Society of Industrial and Applied Mathematics CBMS-NSF Monographs, 38.

[18] Ettling M, et al. (1994): "Economic Impact of Malaria in Malawian Households." *Tropical Medicine and Parasitology.* 45: 74-79.

[19] Frolich, M. (2008): "Statistical treatment choice: an application to active labour market programmes," Journal of the American Statistical Association, 103 (482), 547-558.

[20] Geisser, S. (1975). The predictive sample reuse method with applications. Journal of the Amer. Statist. Assoc., 70:320–328.

[21] Hahn, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects", Econometrica, Vol. 66, No. 2, pp. 315-331.

[22] Hall, Peter and Wolff, Rodney C. L. and Yao, Qiwei (1999): Methods for estimating a conditional distribution function. Journal of the American Statistical Association, 94 (445). pp. 154-163.

[23] Hansen, B. (2008): "Uniform convergence rates for kernel estimation with dependent data," *Econometric Theory*, 24, pp. 726-748.

[24] Härdle, Wolfgang; Hall, Peter; Marron, J. S. (1988): "How far are automatically chosen regression smoothing parameters from their optimum?," *Journal of the American Statistical Association* 83, no. 401, 86–101.

[25] Hawley, William A. et al. (2003): "Community-Wide Effects of Permethrin-Treated Bed Nets on Child Mortality and Malaria Morbidity in Western Kenya." *American Journal of Tropical Medicine and Hygiene*. 68(Suppl. 4): 121-127.

[26] Hirano, K. and J. Porter (2009): "Asymptotics for Statistical Treatment Rules", *Econometrica,* vol. 77(5), pages 1683-1701.

[27] Hirano, K., G. Imbens, G. Ridder (2003): "Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score," *Econometrica* 71, 1161-1189.

[28] Horowitz, J (1992): "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica*, Vol. 60, No. 3, 505-531.

[29] Imbens, G. and G. Ridder (2009): Estimation and Inference for Generalized Full and Partial Means and Average Derivatives, working paper, Harvard University.

[30] Kenya Round 7 Proposal in response to the Global Fund to fight AIDS, Tuberculosis and Malaria 7th call for proposals. Available online at: http://www.theglobalfund.org/en/files/apply/call7/notapproved/7KENM_1527_0_full.pdf

[31] Lechner, M. and J. Smith (2007): What is the value added by case workers?, Labour Economics, 14, 135-151, 2007.

[32] Lengeler, Christian (2004): "Insecticide-treated bed nets and curtains for preventing malaria". Cochrane Database Syst Rev 2004; 2:CD000363.

[33] Lucas, Adrienne (2007): "Economic Effects of Malaria Eradication: Evidence from the Malarial Periphery". Mimeo, Wellesley College.

[34] Mahajan, A, Tarozzi, A., Yoong J., Blackburn B. (2009): "Bednets, Information and Malaria in Orissa," mimeo.

[35] Manski, C. (2004): "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, vol. 72, no. 4, pp. 1221-46.

[36] Manski, C. (2005): Social choice with partial knowledge of treatment response, Princeton University Press.

[37] Newey, W. K. (1994): "Kernel Estimation of Partial Means and a General Variance Estimator." Econometric Theory 10, 233-253.

[38] (NM) Newey, W. and McFadden, D. (1994): "Large sample estimation and hypothesis testing," *Handbook of Econometrics*, vol IV, pp 2113-2245, Elsevier Science B.V. Amsterdam.

[39] Sachs, Jeffrey (2005): *The End of Poverty: Economic Possibilities for Our Time*, New York: Penguin Press.

[40] Stoye, J. (2009): "Minimax Regret Treatment Choice with Finite Samples," Journal of Econometrics 151, 70-81, 2009.

[41] Teklehaimanot, Awash, et al (2007): "Scaling Up Malaria Control in Africa: An Economic and Epidemiological Assessment," American Journal of Tropical Medicine and Hygiene, 77(Suppl 6), pp. 138–144.

[42] Tetenov, A. (2008): Statistical treatment choice based on asymmetric minmax regrest criterion, forthcoming, Journal of Econometrics.

[43] Todd, Petra and Kenneth Wolpin (2006): "Assessing the Impact of a School Subsidy Program in Mexico: Using Experimental Data to Validate a Dynamic Behavioral Model of Child Schooling and Fertility," *American Economic Review*, Vol. 96, No.5, 1384-1417.

[44] WHO (2007): WHO Global Malaria Programme: Position Statement on ITNs. http://www.who.int/malaria/docs/itn/ITNspospaperfinal.pdf

9 Appendix

9.1 Proofs of theorems

In the proofs below, CMT will denote continuous mapping theorem and DCT the Lebesgue dominated convergence theorem. The discrete regressors will not play any substantive roles in our analysis; so we will drop them in our proofs and put them back into our final results at the end. In our proofs, the notation $\tilde{\beta}(x)$ and $\tilde{\gamma}$ will be used to denote values intermediate between $\hat{\beta}(x)$ and $\beta(x)$ and $\hat{\gamma}$ and γ, respectively; c, M_1 and $M(x)$ will denote a finite number, a bounded positive constant and a uniformly bounded positive function respectively, whose actual values may be different in different places. The latter will be used in the expressions for upper bounds for various quantities which appear in the proof. We first state a set of conditions under which the following results will hold.

A0(i) (Y_i, X_i, D_i) $i = 1, 2, ...n$ is a random sample; **A0(ii)** D is randomly allocated;

A1. Conditional on every value x^d assumed by the discrete regressors, $\mathcal{X}^c$ is a p-dimensional compact subset of the support of the continuous components X^c; the density of X^c satisfies that $F_X(x) \geq \delta > 0$ for all $x \in \mathcal{X}^c$; furthermore, $\mu_1(\cdot)$, $\mu_0(\cdot)$ and thus $\beta(\cdot)$ and their qth order derivatives are uniformly bounded on $\mathcal{X}^c$ for some $q > p$. Also, $E|Y_j|^m$ and $E(|Y_j|^m | X = x) \times F_X(x)$ are bounded for some $m \geq 4$.

A2. For some $M > 0$, $\beta(x) \in [-M, M]$ for every $x \in \mathcal{X}$; **A3.** $K(\cdot)$ is a qth order p-dimensional bounded kernel, with $q > p$ and the bandwidth sequence ω_n satisfies (i) $\omega_n \to 0$ (ii) $\sqrt{n}\omega_n^q \to 0$; **A4.** The kernel $\bar{L}(\cdot)$ is uniformly bounded with a bandwidth sequence $h_n \to 0$ and $nh_n \to \infty$.

A1 and A2 are standard. $\mathcal{X}^c$ corresponds to the fixed trimming. condition A3 part (i) is standard. condition A3 part (ii) is an "undersmoothing" requirement, which is commonly used in semiparametric problems for bias removal; it is also a key condition for condition B11 below (c.f. NM, lemma 8.10).

Let $\hat{f}_\beta(u)$ and $f_\beta(u)$ denote respectively the estimated and the true density of $\beta(X)$ at u. Under conditions A1 and A2, we can apply lemma B.1 of Newey (1994) to conclude **B1.** $\sup_{x \in \mathcal{X}} \left| \hat{\beta}(x) - \beta(x) \right| = O_p\left\{ \left(\frac{\ln n}{n\omega_n^p} \right)^{1/2} + \omega_n^q \right\}$ and apply theorem 7 of Hansen (2008) specialized to bounded c_n (given fixed trimming), iid data and $r = 0$, to conclude: **B2.** $\sup_{u \in [-M,M]} \left| \hat{f}_\beta(u) - f_\beta(u) \right| = o_p(1)$. Although these are consequences, rather than primitive conditions, we refer to these as conditions **B1** and **B2** for easy reference in subsequent use.

Introduce the following additional conditions:

B3(i) The first derivative of kernel $\bar{L}(\cdot)$, denoted by L, is uniformly bounded; additionally $L(\cdot)$ satisfies a uniform Lipschitz condition: $|L(u) - L(u')| \leq |u - u'| \Lambda$ for some $\Lambda < \infty$ for all u, u'; **B4.(i)** $h_n \to 0$, $nh_n^4 \to \alpha \in (0, \infty]$ and $n^{1/4} \left\{ \left(\frac{\ln n}{n\omega_n^p} \right)^{1/2} + \omega_n^q \right\} \to 0$; **B5.** The density of $\beta(X)$ is strictly positive on an open set containing γ

Most standard kernels with polynomial structure and with bounded support will satisfy B3.

Let $\hat{F}_{\hat{\beta}}(t) \equiv \frac{1}{n} \sum_{i=1}^n \bar{L}\left(\frac{t - \hat{\beta}(X_i)}{h_n} \right)$.

Lemma 1 *Under conditions A0-A3, A4(i), B1, B2, B3(i) and B4(i),*

$$\sup_{t \in [-M,M]} \left| \hat{F}_{\hat{\beta}}(t) - F_\beta(t) \right| \xrightarrow{P} 0.$$

Proof. Observe that

$$\begin{aligned}
& \hat{F}_{\hat{\beta}}(t) - F_\beta(t) \\
= & \frac{1}{n} \sum_{i=1}^n \bar{L}\left(\frac{t - \hat{\beta}(X_i)}{h_n} \right) - \frac{1}{n} \sum_{i=1}^n \bar{L}\left(\frac{t - \beta(X_i)}{h_n} \right) + \frac{1}{n} \sum_{i=1}^n \left\{ \bar{L}\left(\frac{t - \beta(X_i)}{h_n} \right) - 1(\beta(X_i) \leq t) \right\} \\
& + \frac{1}{n} \sum_{i=1}^n \{1(\beta(X_i) \leq t) - F_\beta(t)\} \\
= & \frac{1}{nh_n} \sum_{i=1}^n L\left(\frac{t - \tilde{\beta}(X_i)}{h_n} \right) \left\{ \beta(X_i) - \hat{\beta}(X_i) \right\} + \frac{1}{n} \sum_{i=1}^n \left\{ \bar{L}\left(\frac{t - \beta(X_i)}{h_n} \right) - 1(\beta(X_i) \leq t) \right\} \\
& + \frac{1}{n} \sum_{i=1}^n \{1(\beta(X_i) \leq t) - F_\beta(t)\}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
& \sup_{t \in [-M,M]} \left| \hat{F}_{\hat{\beta}}(t) - F_\beta(t) \right| \\
\leq & \sup_{t \in [-M,M]} \left| \frac{1}{n} \sum_{i=1}^n \{1(\beta(X_i) \leq t) - F_\beta(t)\} \right| \\
& + \sup_{t \in [-M,M]} \left| \frac{1}{n} \sum_{i=1}^n \left\{ \bar{L}\left(\frac{t - \beta(X_i)}{h_n} \right) - 1(\beta(X_i) \leq t) \right\} \right| \\
& + \frac{1}{n^{1/4} h_n} \left(\frac{1}{n} \sum_{i=1}^n \sup_{t \in [-M,M]} \left| L\left(\frac{t - \tilde{\beta}(X_i)}{h_n} \right) \right| \right) \times \left\{ n^{1/4} \sup_a \left| \beta(a) - \hat{\beta}(a) \right| \right\}
\end{aligned}$$

By condition B3(i) (i.e. $L(\cdot)$ is uniformly bounded), condition B4(i) and condition B1, the third term is $o_p(1)$. The first term is $o_p(1)$ by the standard Glivenko-Cantelli theorem. Under the stated conditions on $\bar{L}$ and that $\beta(X)$ has a Lebesgue density uniformly bounded above, we can apply lemma 4 of Horowitz (1992) to conclude that the second term in the previous display is $o_p(1)$. (This is analogous to Horowitz's proof that $\lim_{\alpha\to 0}\Pr(|b'x|<\alpha)$; here we have that

$$\begin{aligned}\lim_{\alpha\to 0}\Pr(|t-\beta(X)|<\alpha) &= \lim_{\alpha\to 0}\left[F_\beta(t+\alpha)-F_\beta(t-\alpha)\right]\\ &\leq 2\lim_{\alpha\to 0}\left\{\alpha\times\sup_{t\in\mathbb{R}}\left[f_\beta(t)\right]\right\}=0,\end{aligned}$$

and the rest of the proof is identical to Horowitz lemma 4). ■

Theorem 1 (Consistency of $\hat{\gamma}$):

Proof. Fix $\varepsilon>0$. Then $F_\beta(\gamma+\varepsilon)-1+c>0$ and $1-c-F_\beta(\gamma-\varepsilon)>0$, by condition (B5). Therefore, we have that

$$\begin{aligned}\Pr(|\hat{\gamma}-\gamma|>\varepsilon) &\leq \Pr(\hat{\gamma}>\gamma+\varepsilon)+\Pr(\hat{\gamma}<\gamma-\varepsilon)\\ &\leq \Pr\left(\hat{F}_{\hat{\beta}}(\hat{\gamma})>\hat{F}_{\hat{\beta}}(\gamma+\varepsilon)\right)+\Pr\left(\hat{F}_{\hat{\beta}}(\hat{\gamma})<\hat{F}_{\hat{\beta}}(\gamma-\varepsilon)\right)\\ &= \Pr\left(1-c>\hat{F}_{\hat{\beta}}(\gamma+\varepsilon)\right)+\Pr\left(1-c<\hat{F}_{\hat{\beta}}(\gamma-\varepsilon)\right)\\ &\leq \Pr\left(F_\beta(\gamma+\varepsilon)-1+c<F_\beta(\gamma+\varepsilon)-\hat{F}_{\hat{\beta}}(\gamma+\varepsilon)\right)\\ &\quad+\Pr\left(1-c-F_\beta(\gamma-\varepsilon)<\hat{F}_{\hat{\beta}}(\gamma-\varepsilon)-F_\beta(\gamma-\varepsilon)\right)\\ &\leq \Pr\left(F_\beta(\gamma+\varepsilon)-1+c<\sup_{t\in[-M,M]}\left|\hat{F}_{\hat{\beta}}(t)-F_\beta(t)\right|\right)\\ &\quad+\Pr\left(1-c-F_\beta(\gamma-\varepsilon)<\sup_{t\in[-M,M]}\left|\hat{F}_{\hat{\beta}}(t)-F_\beta(t)\right|\right)\end{aligned}$$

both of which converge to zero by lemma 1.

■

Now, define

$$\hat{f}_{\hat{\beta}}(t)=\frac{1}{nh_n}\sum_{i=1}^{n}L\left(\frac{t-\hat{\beta}(X_i)}{h_n}\right).$$

The following lemma shows that $\hat{f}_{\hat{\beta}}(\cdot)$ converges to $f_\beta(\cdot)$ in probability, uniformly on the support of $\beta(X)$.

Lemma 2 *Under conditions A0-A4 and B3-B5,*

$$\sup_{u\in[-M,M]}\left|\hat{f}_{\hat{\beta}}(u)-f_{\beta}(u)\right|=o_p(1).$$

Proof. Observe that

$$\hat{f}_{\hat{\beta}}(u)-f_{\beta}(u)=\frac{1}{nh_n}\sum_{i=1}^{n}L\left(\frac{u-\hat{\beta}(X_i)}{h_n}\right)-f_{\beta}(u)$$

By triangle inequality,

$$\sup_{u\in[-M,M]}\left|\hat{f}_{\hat{\beta}}(u)-f_{\beta}(u)\right|\leq\sup_{u\in[-M,M]}\left|\hat{f}_{\beta}(u)-f_{\beta}(u)\right|+\sup_{u\in[-M,M]}\left|\hat{f}_{\hat{\beta}}(u)-\hat{f}_{\beta}(u)\right|.$$

The first term is $o_p(1)$ due to B2. As for the second term, notice that

$$\begin{aligned}\left|\hat{f}_{\hat{\beta}}(t)-\hat{f}_{\beta}(t)\right| &= \left|\frac{1}{nh_n}\sum_{i=1}^{n}\left\{L\left(\frac{t-\hat{\beta}(X_i)}{h_n}\right)-L\left(\frac{t-\beta(X_i)}{h_n}\right)\right\}\right| \\ &\leq \frac{\sup_x\left|\hat{\beta}(x)-\beta(x)\right|}{h_n}\Lambda, \text{ by B3,} \\ &= O_p\left(\frac{1}{h_n}\times\left\{\left(\frac{\ln n}{n\omega_n^p}\right)^{1/2}+\omega_n^q\right\}\right),\end{aligned}$$

by conclusion B2 and assumption B3. Therefore by condition B4, we get the conclusion. ■

Additional conditions:

A4 (ii) The kernel $\bar{L}(\cdot)$ has two derivatives which are also uniformly bounded. For some $r\geq 2$, we have **B7.** the density of $\beta(X)$ is $(r-1)$ times continuously differentiable, the $(r-1)$th derivative $f_{\beta}^{(r-1)}(\cdot)$ is bounded and Lipschitz in a neighborhood of γ; **B8.** $nh_n^{2r+1}\to\lambda<\infty$; **B9.** $L(\cdot)$ is symmetric around zero and has bounded support $[-1,1]$, is of order r and $\int_{-1}^{1}L^2(u)\,du<\infty$. Also assume that **B10.** $\sigma_j^2(x)=Var(Y_j|X=x)$ for $j=1,2$ are finite; **B11.** For $j=0,1$,

$$\begin{aligned}\sqrt{n}\sup_{x\in\mathcal{X}}\left\|\left\{\hat{\mu}_j(x)-\mu_j(x)\right\}\left\{\hat{\pi}_j(x)-\pi_j(x)\right\}\right\| &= o_p(1), \\ \sqrt{n}\sup_{x}\left\|\left\{\hat{\pi}_j(x)-\pi_j(x)\right\}\right\|^2 &= o_p(1).\end{aligned}$$

Condition B11 is basically the same as B4 with μ and π replacing β and hold under exactly the same conditions as B4. These are well-known requirement for $\sqrt{n}$ -normality for semiparametric estimators (c.f. Newey and McFadden (1994), section 8.3).

Theorem 1 (Distribution of $\hat{\gamma}$):

To derive the distribution theory for $\hat{\gamma}$, we will use the following first-order approximation

$$F_\beta(\gamma) = 1 - c = \hat{F}_{\hat{\beta}}(\hat{\gamma}) = \hat{F}_{\hat{\beta}}(\gamma) + (\hat{\gamma} - \gamma)\hat{f}_{\hat{\beta}}(\tilde{\gamma})$$

where $\tilde{\gamma}$ is between $\hat{\gamma}$ and γ. This gives us the following first-order expansion for $\hat{\gamma}$:

$$\begin{aligned}
&(\hat{\gamma} - \gamma) \\
= \; & \left\{\hat{f}_{\hat{\beta}}(\tilde{\gamma})\right\}^{-1} \underbrace{\left\{F_\beta(\gamma) - \frac{1}{n}\sum_{i=1}^{n} \bar{L}\left(\frac{\gamma - \beta(X_i)}{h_n}\right)\right\}}_{T_{1n}} \\
& + \left\{\hat{f}_{\hat{\beta}}(\tilde{\gamma})\right\}^{-1} \underbrace{\left[\frac{1}{n}\sum_{i=1}^{n}\left(\bar{L}\left(\frac{\gamma - \beta(X_i)}{h_n}\right) - \bar{L}\left(\frac{\gamma - \hat{\beta}(X_i)}{h_n}\right)\right)\right]}_{T_{2n}}. \qquad (11)
\end{aligned}$$

The proof will proceed in three steps: step 1 is that the multiplier $\left\{\hat{f}_{\hat{\beta}}(\tilde{\gamma})\right\}^{-1}$ converges in probability to $\{f_\beta(\gamma)\}^{-1}$. Step 2 is that the term T_{1n} will be $O_p\left(\frac{1}{\sqrt{nh_n}}\right)$. Finally in step 3 we will show, using U-statistic type decompositions, that the term T_{2n} will be $O_p\left(\frac{1}{\sqrt{nh_n}}\right)$. These will eventually lead to the result that $\sqrt{nh_n}(\hat{\gamma} - \gamma)$ converges to a normal distribution, as required.

Proof. Step 1. We first show that

$$\hat{f}_{\hat{\beta}}(\tilde{\gamma}) - f_\beta(\gamma) \xrightarrow{P} 0. \qquad (12)$$

$$\begin{aligned}
\left|\hat{f}_{\hat{\beta}}(\tilde{\gamma}) - f_\beta(\gamma)\right| &\leq \left|\hat{f}_{\hat{\beta}}(\tilde{\gamma}) - f_\beta(\tilde{\gamma})\right| + \left|f_\beta(\tilde{\gamma}) - f_\beta(\gamma)\right| \\
&\leq \underbrace{\sup_{a\in[-M,M]}\left|\hat{f}_{\hat{\beta}}(a) - f_\beta(a)\right|}_{o_p(1),\text{ by lemma 2}} + \underbrace{\left|f_\beta(\tilde{\gamma}) - f_\beta(\gamma)\right|}_{o_p(1)\text{ by CMT and theorem 1}} = o_p(1).
\end{aligned}$$

Step 2: We will show that

$$\sqrt{nh_n}\left\{F_\beta(\gamma) - \bar{L}\left(\frac{\gamma - \beta(X_i)}{h_n}\right)\right\} = \kappa + o_p(1). \qquad (13)$$

Observe that

$$\begin{aligned}
T_n &\equiv \frac{1}{n}\sum_{i=1}^{n}\left\{F_\beta(\gamma) - \bar{L}\left(\frac{\gamma - \beta(X_i)}{h_n}\right)\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\{F_\beta(\gamma) - 1(\beta(X_i) \leq \gamma)\} + \frac{1}{n}\sum_{i=1}^{n}\left\{1(\beta(X_i) \leq \gamma) - \bar{L}\left(\frac{\gamma - \beta(X_i)}{h_n}\right)\right\} \\
&\equiv T_{2n} - T_{1n}. \qquad (14)
\end{aligned}$$

Now,

$$\sqrt{nh_n}T_{2n} = \sqrt{h_n} \times \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ F_\beta(\gamma) - 1\left(\beta(X_i) \leq \gamma\right)\right\}}_{O_p(1)} = o_p(1).$$

We will show that

$$E\left(\sqrt{nh_n}T_{1n} - \kappa\right)^2 = h_n Var\left(\sqrt{n}T_{1n}\right) + \left\{E\left(\sqrt{nh_n}T_{1n} - \kappa\right)\right\}^2 \to 0 \tag{15}$$

and thus

$$\sqrt{nh_n}T_{1n} - \kappa = o_p(1). \tag{16}$$

Now,

$$\begin{aligned} & Var\left(\sqrt{n}T_{1n}\right) \\ = \; & E\left\{1\left(\beta(X_i) \leq \gamma\right) - \bar{L}\left(\frac{\gamma - \beta(X_i)}{h_n}\right)\right\}^2 - \left\{F_\beta(\gamma) - E\left\{\bar{L}\left(\frac{\gamma - \beta(X_i)}{h_n}\right)\right\}\right\}^2 \end{aligned} \tag{17}$$

Observe that

$$\begin{aligned} & E\left(1\left(\beta(X_i) \leq \gamma\right) - \bar{L}\left(\frac{\gamma - \beta(X_i)}{h_n}\right)\right)^2 \\ = \; & \int_{-M}^{\gamma} \left\{\bar{L}\left(\frac{\gamma - t}{h_n}\right) - 1\right\}^2 f_\beta(t)\, dt + \int_{\gamma}^{M} \left\{\bar{L}\left(\frac{\gamma - t}{h_n}\right)\right\}^2 f_\beta(t)\, dt \end{aligned}$$

and both of the terms in the previous display converge to zero by the DCT since $\lim_{a\to\infty} \bar{L}(a) = 1 = 1 - \lim_{a\to-\infty} \bar{L}(a)$.

Next,

$$\begin{aligned} & F_\beta(\gamma) - E\left\{\bar{L}\left(\frac{\gamma - \beta(X_i)}{h_n}\right)\right\} \\ = \; & \int_{-M}^{\gamma} \left[1\left(t \leq \gamma\right) - \bar{L}\left(\frac{\gamma - t}{h_n}\right)\right] f_\beta(t)\, dt - \int_{\gamma}^{M} \bar{L}\left(\frac{\gamma - t}{h_n}\right) f_\beta(t)\, dt \to 0, \text{ by the DCT.} \end{aligned}$$

Thus, from (17), we have that

$$Var\left(\sqrt{n}T_{1n}\right) \to 0 \text{ as } n \to \infty. \tag{18}$$

Next, consider

$$\begin{aligned} E\left(T_{1n}\right) &= E\left\{1\left(\beta\left(X_i\right)\leq\gamma\right)-\bar{L}\left(\frac{\gamma-\beta\left(X_i\right)}{h_n}\right)\right\} \\ &= \left\{F_\beta\left(\gamma\right)-\int_{-M}^{M}\bar{L}\left(\frac{\gamma-t}{h_n}\right)f_\beta\left(t\right)dt\right\} \\ &= \left\{F_\beta\left(\gamma\right)-\bar{L}\left(\frac{\gamma-t}{h_n}\right)F_\beta\left(t\right)|_{-M}^{M}-\frac{1}{h_n}\int_{-M}^{M}F_\beta\left(t\right)L\left(\frac{\gamma-t}{h_n}\right)dt\right\} \\ &= \left\{F_\beta\left(\gamma\right)-\int_{\frac{\gamma-M}{h_n}}^{\frac{\gamma+M}{h_n}}F_\beta\left(\gamma-uh_n\right)L\left(u\right)du\right\} \\ &= (-1)^{r+1}\frac{h_n^r}{r!}\times f_\beta^{(r-1)}\left(\gamma\right)\times\int_{-1}^{1}u^rL\left(u\right)du+o\left(h_n^r\right), \text{ by condition B7.} \end{aligned}$$

This implies that

$$\begin{aligned} E\left(\sqrt{nh_n}T_{1n}\right) &= (-1)^{r+1}\frac{\sqrt{n}h_n^{r+1/2}}{r!}\times f_\beta^{(r-1)}\left(\gamma\right)\times\int_{-1}^{1}u^rL\left(u\right)du+o\left(h_n^r\right) \\ &\rightarrow \kappa, \text{ by condition B7, B8.} \end{aligned} \tag{19}$$

Now, (18) and (19) imply (15) and thus (16).

Step 3: We will now analyze the second term in (11):

$$S_n=\frac{1}{n}\sum_{i=1}^{n}\left[\bar{L}\left(\frac{\gamma-\beta\left(X_i\right)}{h_n}\right)-\bar{L}\left(\frac{\gamma-\hat{\beta}\left(X_i\right)}{h_n}\right)\right]du,$$

using U-statistic type decompositions to show that

$$\begin{aligned} \sqrt{nh_n}S_n &= \frac{\sqrt{h_n}}{\sqrt{n}}\sum_{j=1}^{n}\left\{\left[\lambda_{1n}\left(Z_j\right)-E\left\{\lambda_{1n}\left(Z_j\right)\right\}\right]-\left[\lambda_{2n}\left(Z_j\right)-E\left\{\lambda_{2n}\left(Z_j\right)\right\}\right]\right\} \\ &\quad +o_p\left(1\right)\xrightarrow{d}N\left(0,\eta^2\right), \end{aligned} \tag{20}$$

where the triangular arrays $\lambda_{1n}\left(Z_j\right)$, $\lambda_{2n}\left(Z_j\right)$ and the constant $\eta^2>0$, will be specified below.

To that end observe that

$$\begin{aligned} \sqrt{nh_n}S_n &= \frac{\sqrt{h_n}}{\sqrt{n}}\sum_{i=1}^{n}\left[\bar{L}\left(\frac{\gamma-\beta\left(X_i\right)}{h_n}\right)-\bar{L}\left(\frac{\gamma-\hat{\beta}\left(X_i\right)}{h_n}\right)\right] \\ &= \frac{\sqrt{h_n}}{\sqrt{n}h_n}\sum_{i=1}^{n}\left\{\hat{\beta}\left(X_i\right)-\beta\left(X_i\right)\right\}L\left(\frac{\gamma-\beta\left(X_i\right)}{h_n}\right) \\ &\quad +\frac{\sqrt{h_n}}{2\sqrt{n}h_n^2}\sum_{i=1}^{n}\left\{\hat{\beta}\left(X_i\right)-\beta\left(X_i\right)\right\}^2L'\left(\frac{\tilde{\beta}\left(X_i\right)-\gamma}{h_n}\right). \end{aligned}$$

The second term in absolute value has an expectation which is of the order of

$$\sup_{x\in\mathcal{X}}\left|\beta(x)-\hat{\beta}(x)\right|^2\frac{\sqrt{n}}{h_n^{3/2}}\to 0, \text{ by condition B4.}$$

Next, note that

$$\hat{\beta}(X_i)-\beta(X_i)=\left\{\frac{\hat{\mu}_1(X_i)}{\hat{\pi}_1(X_i)}-\frac{\mu_1(X_i)}{\pi_1(X_i)}\right\}-\left\{\frac{\hat{\mu}_0(X_i)}{\hat{\pi}_0(X_i)}-\frac{\mu_0(X_i)}{\pi_0(X_i)}\right\} \tag{21}$$

We will simply work with the first term because the proof is exactly analogous for the second term and show that

$$\frac{1}{\sqrt{nh_n}}\sum_{i=1}^{n}\left\{\hat{\beta}(X_i)-\beta(X_i)\right\}L\left(\frac{\gamma-\beta(X_i)}{h_n}\right)=O_p(1).$$

Step 3A: Now,

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{\hat{\mu}_1(X_i)}{\hat{\pi}_1(X_i)}-\frac{\mu_1(X_i)}{\pi_1(X_i)}\right\}\frac{1}{h_n}L\left(\frac{\beta(X_i)-\gamma}{h_n}\right)\\
=\;&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{\hat{\mu}_1(X_i)-\mu_1(X_i)}{\pi(X_i)}\right\}\frac{1}{h_n}L\left(\frac{\beta(X_i)-\gamma}{h_n}\right)\\
&-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{\mu(X_i)}{\pi(X_i)}\frac{\hat{\pi}_1(X_i)-\pi_1(X_i)}{\pi_1(X_i)}\right\}\frac{1}{h_n}L\left(\frac{\beta(X_i)-\gamma}{h_n}\right)\\
&-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\{\hat{\mu}_1(X_i)-\mu(X_i)\}\{\hat{\pi}_1(X_i)-\pi_1(X_i)\}}{\pi_1(X_i)\hat{\pi}_1(X_i)}\frac{1}{h_n}L\left(\frac{\beta(X_i)-\gamma}{h_n}\right)\\
&+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\mu(X_i)\{\hat{\pi}_1(X_i)-\pi_1(X_i)\}^2}{\pi_1^2(X_i)\hat{\pi}_1(X_i)}\frac{1}{h_n}L\left(\frac{\beta(X_i)-\gamma}{h_n}\right).
\end{aligned} \tag{22}$$

The last two terms in absolute value have expectations that are bounded above by a positive scalar times $\sqrt{n}\sup_x\|\{\hat{\mu}_1(x)-\mu_1(x)\}\{\hat{\pi}_1(x)-\pi_1(x)\}\|$ and $\sqrt{n}\sup_x\|\{\hat{\pi}_1(x)-\pi(x)\}\|^2$. Condition B11 above then implies that these are both $o_p(1)$.

Now, the first two terms in (22) add up to

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\pi_1(X_i)\hat{\mu}_1(X_i)-\mu_1(X_i)\hat{\pi}_1(X_i)}{\pi_1^2(X_i)}\frac{1}{h_n}L\left(\frac{\beta(X_i)-\gamma}{h_n}\right)\\
= \;& \sqrt{n}\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\underbrace{\left[\begin{array}{c}\frac{1}{\pi_1^2(X_i)}\left\{\pi_1(X_i)Y_jD_j-\mu_1(X_i)D_j\right\}\frac{1}{\omega_n^p}K\left(\frac{X_j-X_i}{\omega_n}\right)\\ \times\frac{1}{h_n}L\left(\frac{\beta(X_i)-\gamma}{h_n}\right)\end{array}\right]}_{=w_n(Z_i,Z_j),\text{ say}}\\
= \;& \underbrace{\frac{1}{\sqrt{n}(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\left[w_n(Z_i,Z_j)-E(w_n(Z_i,Z_j)|Z_i)-E(w_n(Z_i,Z_j)|Z_j)+E(w_n(Z_i,Z_j))\right]}_{U_{1n}}\\
&+\underbrace{\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left[E(w_n(Z_i,Z_j)|Z_j)-E(w_n(Z_i,Z_j))\right]}_{U_{2n}}+\underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}E(w_n(Z_i,Z_j)|Z_i)}_{U_{3n}}. \qquad (23)
\end{aligned}$$

Step 3B: We first show that

$$U_{3n}=o_p(1). \qquad (24)$$

Notice that

$$\begin{aligned}
&E\left[\frac{\frac{1}{h_n}L\left(\frac{\beta(X_i)-\gamma}{h_n}\right)}{\pi_1^2(X_i)}\times\left\{\pi_1(X_i)Y_jD_j-\mu_1(X_i)D_j\right\}\frac{1}{\omega_n^p}K\left(\frac{X_j-X_i}{\omega_n}\right)|Z_i\right]\\
&\overset{L.I.E.}{=}E\left[\frac{\frac{1}{h_n}L\left(\frac{\beta(X_i)-\gamma}{h_n}\right)}{\pi_1^2(X_i)}\times\left\{\pi_1(X_i)\frac{\mu_1(X_j)}{f(X_j)}-\mu_1(X_i)\frac{\pi_1(X_j)}{f(X_j)}\right\}\frac{1}{\omega_n^p}K\left(\frac{X_j-X_i}{\omega_n}\right)|Z_i\right]\\
= \;& \frac{\frac{1}{h_n}L\left(\frac{\beta(X_i)-\gamma}{h_n}\right)}{\pi_1^2(X_i)}\times\int\frac{\pi_1(X_i)\mu_1(x)-\mu_1(X_i)\pi_1(x)}{f(x)}\frac{1}{\omega_n^p}K\left(\frac{x-X_i}{\omega_n}\right)f(x)\,dx\\
= \;& \frac{\frac{1}{h_n}L\left(\frac{\beta(X_i)-\gamma}{h_n}\right)}{\pi_1^2(X_i)}\times\int\left[\pi_1(X_i)\mu_1(X_i+u\omega_n)-\mu_1(X_i)\pi_1(X_i+u\omega_n)\right]K(u)\,du\\
&\overset{A1}{=}H(X_i)\times\frac{1}{h_n}L\left(\frac{\beta(X_i)-\gamma}{h_n}\right)\times O(\omega_n^q),
\end{aligned}$$

for some uniformly bounded function H by condition. Therefore,

$$U_{3n}=O(\omega_n^q)\times\frac{1}{\sqrt{n}}\sum_{i=1}^{n}H(X_i)\times\frac{1}{h_n}L\left(\frac{\beta(X_i)-\gamma}{h_n}\right)=O_p\left(\sqrt{n}\omega_n^q\right)=o_p(1),$$

by condition B4.

Step 3C: The term

$$U_{1n} = \frac{1}{\sqrt{n}\,(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \left[w_n\left(Z_i, Z_j\right) - E\left(w_n\left(Z_i, Z_j\right)|Z_i\right) - E\left(w_n\left(Z_i, Z_j\right)|Z_j\right) + E\left(w_n\left(Z_i, Z_j\right)\right) \right]$$

can be analyzed using essentially the steps of Powell, Stoker and Stock (1989), lemma 3.1, whence one can conclude that

$$E\left(U_{1n}^2\right) = o\left(1\right) \tag{25}$$

The key step is to show that

$$E\left(w_n^2\left(Z_i, Z_j\right)\right) = o\left(n\right).$$

Observe that

$$\begin{aligned}
& n^{-1} E\left(w_n^2\left(Z_i, Z_j\right)\right) \\
= \; & n^{-1} E\left\{ \frac{1}{\pi_1^4\left(X_i\right)} \left\{ \pi_1\left(X_i\right) Y_j D_j - \mu_1\left(X_i\right) D_j \right\}^2 \frac{1}{\omega_n^{2p}} K^2\left(\frac{X_j - X_i}{\omega_n}\right) \times \left[\frac{1}{h_n} L\left(\frac{\beta\left(X_i\right) - \gamma}{h_n}\right) \right]^2 \right\} \\
= \; & n^{-1} E\left\{ \begin{array}{c} \frac{1}{\pi_1^4(X_i)} \frac{1}{\omega_n^{2p}} K^2\left(\frac{X_j - X_i}{\omega_n}\right) \times \left[\frac{1}{h_n} L\left(\frac{\beta(X_i) - \gamma}{h_n}\right) \right]^2 \\ \times \left\{ \pi_1^2\left(X_i\right) E\left(Y^2 D | X_j\right) + \mu_1^2(X_i) E\left(D|X_j\right) - 2\pi_1\left(X_i\right) \mu_1\left(X_i\right) E\left(YD|X_j\right) \right\} \end{array} \right\} \\
= \; & \frac{1}{n\omega_n^p h_n^2} \int \left\{ \begin{array}{c} \frac{1}{\pi_1^4(X)} K^2\left(u\right) \times \left[L\left(\frac{\beta(x) - \gamma}{h_n}\right) \right]^2 \\ \times \left\{ \begin{array}{c} \pi_1^2\left(x\right) E\left(Y^2 D | X = x + u\omega_n\right) \\ + \mu_1^2(x) E\left(D|X = x + u\omega_n\right) \\ -2\pi_1\left(x\right) \mu_1\left(x\right) E\left(YD|X = x + u\omega_n\right) \end{array} \right\} \end{array} \right\} f_X\left(x\right) f_X\left(x + u\omega_n\right) du dx \\
= \; & O\left(\frac{1}{n\omega_n^p h_n^2}\right) \to 0 \text{ which is implied by B4.}
\end{aligned}$$

Step 3D: Now consider the term

$$U_{2n} = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left[E\left(w_n\left(Z_i, Z_j\right)|Z_j\right) - E\left(w_n\left(Z_i, Z_j\right)\right) \right]$$

Observe that

$$\begin{aligned}
& E\left(w_n\left(Z_i,Z_j\right)|Z_j\right) \\
= & E\left\{\left[\frac{1}{\pi_1^2\left(X_i\right)}\left\{\pi_1\left(X_i\right)Y_jD_j-\mu_1\left(X_i\right)D_j\right\}\frac{1}{\omega_n^p}K\left(\frac{-X_j+X_i}{\omega_n}\right)\times\left[\frac{1}{h_n}L\left(\frac{\beta\left(X_i\right)-\gamma}{h_n}\right)\right]\right]|Y_j,D_j,X_j\right\} \\
= & \int\left[\frac{1}{\pi_1^2\left(x\right)}\left\{\pi_1\left(x\right)Y_jD_j-\mu_1\left(x\right)D_j\right\}\frac{1}{\omega_n^p}K\left(\frac{-X_j+x}{\omega_n}\right)\times\left[\frac{1}{h_n}L\left(\frac{\beta\left(x\right)-\gamma}{h_n}\right)\right]\right]f\left(x\right)dx \\
= & \int\left[\frac{\pi_1\left(X_j+u\omega_n\right)Y_jD_j-\mu_1\left(X_j+u\omega_n\right)D_j}{\pi_1^2\left(X_j+u\omega_n\right)}K\left(u\right)\times\left[\frac{1}{h_n}L\left(\frac{\beta\left(X_j+u\omega_n\right)-\gamma}{h_n}\right)\right]\right]f\left(X_j+u\omega_n\right)du \\
= & \left[\frac{1}{\pi_1^2\left(X_j\right)}\left\{\pi_1\left(X_j\right)Y_jD_j-\mu_1\left(X_j\right)D_j\right\}f\left(X_j\right)\times\left[\frac{1}{h_n}L\left(\frac{\beta\left(X_j\right)-\gamma}{h_n}\right)\right]\right]\int K\left(u\right)du+O\left(\omega_n^q\right) \\
= & \left[\underbrace{\left[\frac{1}{\pi_1^2\left(X_j\right)}\left\{\pi_1\left(X_j\right)Y_jD_j-\mu_1\left(X_j\right)D_j\right\}f\left(X_j\right)\right]}_{W(Z_j)}\times\underbrace{\left[\frac{1}{h_n}L\left(\frac{\beta\left(X_j\right)-\gamma}{h_n}\right)\right]}_{V_n(\beta(X_j))}\right]+O\left(\omega_n^q\right).
\end{aligned}$$

Notice that

$$E\left\{W\left(Z_j\right)V_n\left(\beta\left(X_j\right)\right)\right\}=E\left\{V_n\left(\beta\left(X_j\right)\right)\underbrace{E\left(W\left(Z_j\right)|X_j\right)}_{=0}\right\}=0.$$

Now

$$\begin{aligned}
Var\left(U_{2n}\right) &= Var\left\{\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left[E\left(w_n\left(Z_i,Z_j\right)|Z_j\right)-E\left(w_n\left(Z_i,Z_j\right)\right)\right]\right\} \\
&= Var\left\{E\left(w_n\left(Z_i,Z_j\right)|Z_j\right)\right\} \\
&= E\left(W\left(Z_j\right)V_{1n}\left(\beta\left(X_j\right)\right)+O\left(\omega_n^q\right)\right)^2-O\left(\omega_n^{2q}\right) \\
&= E\left\{W^2\left(Z_j\right)V_{1n}^2\left(\beta\left(X_j\right)\right)\right\}+O\left(\omega_n^{2q}\right) \\
&= O\left(E\left\{W^2\left(Z_j\right)V_{1n}^2\left(\beta\left(X_j\right)\right)\right\}\right).
\end{aligned}$$

Now, let $\xi^2\left(t\right)=E\left\{W^2\left(Z_j\right)|\beta\left(X_j\right)=t\right\}$. Then

$$\begin{aligned}
& E\left\{W^2\left(Z_j\right)V_{1n}^2\left(\beta\left(X_j\right)\right)\right\} \\
= & \int_{-M}^{M}\xi^2\left(t\right)\left[\frac{1}{h_n}L\left(\frac{t-\gamma}{h_n}\right)\right]^2f_\beta\left(t\right)dt \\
= & \frac{1}{h_n}\int_{\frac{-M-\gamma}{h_n}}^{\frac{M-\gamma}{h_n}}\xi^2\left(\gamma+uh_n\right)L^2\left(u\right)f_\beta\left(\gamma+uh_n\right)du \\
= & \frac{1}{h_n}\xi^2\left(\gamma\right)f_\beta\left(\gamma\right)\int_{\frac{-M-\gamma}{h_n}}^{\frac{M-\gamma}{h_n}}L^2\left(u\right)du+\text{terms of smaller order.}
\end{aligned}$$

This implies that

$$\begin{aligned} Var\left(\sqrt{h_n}U_{2n}\right) &= Var\left(\sqrt{\frac{h_n}{n}}\sum_{j=1}^{n}\left[E\left(w_n\left(Z_i,Z_j\right)|Z_j\right)-E\left(w_n\left(Z_i,Z_j\right)\right)\right]\right) \\ &\to \xi^2\left(\gamma\right)f_\beta\left(\gamma\right)\int_{-\infty}^{\infty}L^2\left(u\right)du. \end{aligned} \tag{26}$$

Now we will apply the Liapunov condition and use the Lindeberg CLT for triangular arrays. Consider the array

$$R_{nj}=\frac{\sqrt{h_n}}{\sqrt{n}}\left[E\left(w_n\left(Z_i,Z_j\right)|Z_j\right)-E\left(w_n\left(Z_i,Z_j\right)\right)\right],$$

which is independent across j and $E\left(R_{nj}\right)=0$. Let $U_n=\sum_{j=1}^{n}R_{nj}$. Then

$$\begin{aligned} E\left(U_n^2\right) &= \sum_{j=1}^{n}E\left(R_{nj}^2\right) \\ &= \frac{h_n}{n}\sum_{j=1}^{n}E\left(E\left(w_n\left(Z_i,Z_j\right)|Z_j\right)-E\left(w_n\left(Z_i,Z_j\right)\right)\right)^2 \\ &= \frac{h_n}{n}\sum_{j=1}^{n}Var\left(W\left(Z_j\right)V_{1n}\left(\beta\left(X_j\right)\right)\right)+o\left(1\right) \\ &= \frac{h_n}{h_n}\xi^2\left(\gamma\right)f_\beta\left(\gamma\right)\int_{-\infty}^{\infty}L^2\left(u\right)du+o\left(1\right)\to\xi^2\left(\gamma\right)f_\beta\left(\gamma\right)\int_{-\infty}^{\infty}L^2\left(u\right)du, \end{aligned}$$

by (26). To apply the Liapunov condition, observe that for any $\varepsilon>0$,

$$\begin{aligned} \sum_{j=1}^{n}E\left|R_{nj}\right|^{2+\varepsilon} &= n\left(\frac{h_n}{n}\right)^{\frac{2+\varepsilon}{2}}E\left|W\left(Z_j\right)V_{1n}\left(\beta\left(X_j\right)\right)\right|^{2+\varepsilon} \\ &= O\left(n\left(\frac{h_n}{n}\right)^{\frac{2+\varepsilon}{2}}\frac{1}{h_n^{1+\varepsilon}}\right) \\ &= O\left(\left(h_nn\right)^{-\varepsilon/2}\right)\to 0. \end{aligned}$$

Thus the Liapunov condition holds and applying the Lindeberg CLT, we get that

$$U_n=\frac{\sqrt{h_n}}{\sqrt{n}}\sum_{j=1}^{n}\left[E\left(w_n\left(Z_i,Z_j\right)|Z_j\right)-E\left(w_n\left(Z_i,Z_j\right)\right)\right]\xrightarrow{d}N\left(0,\xi^2\left(\gamma\right)f_\beta\left(\gamma\right)\int_{-\infty}^{\infty}L^2\left(u\right)du\right). \tag{27}$$

Putting together (24), (25), (26) and (27), we get that

$$\begin{aligned}
&\sqrt{\frac{h_n}{n}}\sum_{i=1}^{n}\left\{\frac{\hat{\mu}_1(X_i)}{\hat{\pi}_1(X_i)}-\frac{\mu_1(X_i)}{\pi_1(X_i)}\right\}\left[\frac{1}{h_n}L\left(\frac{\beta(X_i)-\gamma}{h_n}\right)\right]\\
&= \frac{\sqrt{h_n}}{\sqrt{n}}\sum_{j=1}^{n}\lambda_{1n}(Z_j)+o_p(1)\xrightarrow{d}N\left(0,\xi^2(\gamma)f_\beta(\gamma)\int_{-\infty}^{\infty}L^2(u)\,du\right),
\end{aligned}$$

where

$$\begin{aligned}
\lambda_{1n}(Z_j) &= \left[\frac{f(X_j)}{\pi_1^2(X_j)}\left\{\pi_1(X_j)Y_jD_j-\mu_1(X_j)D_j\right\}\right]\frac{1}{h_n}L\left(\frac{\beta(X_j)-\gamma}{h_n}\right),\\
\xi^2(t) &= E\left\{\left\{\frac{\pi_1(X)YD-\mu_1(X)D}{\pi_1^2(X)}f(X)\right\}^2|\beta(X)=t\right\}. \qquad (28)
\end{aligned}$$

Similarly, we will get that

$$\begin{aligned}
&\sqrt{\frac{h_n}{n}}\sum_{i=1}^{n}\left\{\frac{\hat{\mu}_0(X_i)}{\hat{\pi}_0(X_i)}-\frac{\mu_0(X_i)}{\pi_0(X_i)}\right\}\frac{1}{h_n}L\left(\frac{\beta(X_i)-\gamma}{h_n}\right)\\
&= \frac{\sqrt{h_n}}{\sqrt{n}}\sum_{j=1}^{n}\lambda_{2n}(Z_j)+o_p(1)\xrightarrow{d}N\left(0,\tau^2(\gamma)f_\beta(\gamma)\int_{-\infty}^{\infty}L^2(u)\,du\right),
\end{aligned}$$

where

$$\begin{aligned}
\lambda_{2n}(Z_j) &= \left[\frac{f(X_j)}{\pi_0^2(X_j)}\left\{\pi_0(X_j)Y_j(1-D_j)-\nu(X_j)(1-D_j)\right\}\right]\frac{1}{h_n}L\left(\frac{\beta(X_i)-\gamma}{h_n}\right),\\
\tau^2(t) &= E\left\{\left\{\frac{\pi_0(X)Y(1-D)-\nu(X)(1-D)}{\pi_0^2(X)}f(X)\right\}^2|\beta(X)=t\right\}. \qquad (29)
\end{aligned}$$

Thus we get that

$$\sqrt{nh_n}S_n=\frac{\sqrt{h_n}}{\sqrt{n}}\sum_{j=1}^{n}\left\{\lambda_{1n}(Z_j)-\lambda_{2n}(Z_j)\right\}+o_p(1)\xrightarrow{d}N\left(0,\eta^2\right),$$

which establishes (20).

To get the expression for η^2, note by direct multiplication that since $D_j(1-D_j)=0$ for every j, $E\left\{\lambda_{1n}(Z_j)\lambda_{2n}(Z_j)\right\}=0$. Moreover, $E\left(\lambda_{1n}(Z_j)\right)=0$. Therefore, $cov\left(\lambda_{1n}(Z_j),\lambda_{2n}(Z_j)\right)=0$. This implies that

$$\eta^2=\left\{\tau^2(\gamma)+\xi^2(\gamma)\right\}\times f_\beta(\gamma)\int_{-\infty}^{\infty}L^2(u)\,du. \qquad (30)$$

Using the definitions of $\mu_1(\cdot), \pi_0(\cdot), \nu(\cdot)$ and $\pi_1(\cdot)$, the above expressions simply to

$$\begin{aligned}\tau^2(\gamma) &= E\left\{\frac{\sigma_0^2(X)}{1-\Pr(D=1|X)}|\beta(X)=\gamma\right\} \\ \xi^2(\gamma) &= E\left\{\frac{\sigma_1^2(X)}{\Pr(D=1|X)}|\beta(X)=\gamma\right\}.\end{aligned}$$

Now put together (12), (13), (20) and (30) to conclude from (11) that

$$\begin{aligned}\sqrt{nh_n}(\hat{\gamma}-\gamma) &= \kappa + \frac{1}{f_\beta(\gamma)}\frac{\sqrt{h_n}}{\sqrt{n}}\sum_{j=1}^{n}\{\lambda_{1n}(Z_j)-\lambda_{2n}(Z_j)\}+o_p(1) \\ &\xrightarrow{d} N\left(\kappa, \frac{\tau^2(\gamma)+\xi^2(\gamma)}{f_\beta(\gamma)}\int_{-\infty}^{\infty}L^2(u)\,du\right).\end{aligned}$$

■

Lemma 3 *Under conditions A0-A4 and B1-B11,*

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left[\hat{\beta}_1(X_j)-\beta_1(X_j)\right] = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}D_j\frac{Y_{1j}-\beta_1(X_j)}{\Pr(D=1|X_j)}+o_p(1).$$

Proof. Note that

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left[\hat{\beta}_1(X_j)-\beta_1(X_j)\right] \\ =\ &\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{\hat{\mu}_1(X_i)-\mu_1(X_i)}{\pi_1(X_i)}\right\}-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{\mu_1(X_i)}{\pi_1(X_i)}\frac{\hat{\pi}_1(X_i)-\pi(X_i)}{\pi_1(X_i)}\right\} \\ &+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\{\hat{\mu}_1(X_i)-\mu_1(X_i)\}\{\hat{\pi}_1(X_i)-\pi(X_i)\}}{\pi_1(X_i)\hat{\pi}_1(X_i)} \\ &-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\hat{\mu}_1(X_i)\{\hat{\pi}_1(X_i)-\pi(X_i)\}^2}{\pi_1^2(X_i)\hat{\pi}_1(X_i)}. \qquad (31)\end{aligned}$$

The last two terms are both $o_p(1)$ under condition B11 above.

Now, the first two terms in (31) add up to

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\pi_1(X_i)\hat{\mu}_1(X_i)-\mu_1(X_i)\hat{\pi}_1(X_i)\right\}\\
=\;&\sqrt{n}\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\frac{\pi_1(X_i)Y_jD_j-\mu_1(X_i)D_j}{\pi^2(X_i)}\frac{1}{\omega_n^p}K\left(\frac{X_j-X_i}{\omega_n}\right)\\
\equiv\;&\sqrt{n}\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}w(Z_i,Z_j,\omega_n)\\
=\;&\underbrace{\frac{1}{\sqrt{n}(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\left[w(Z_i,Z_j,\omega_n)-E(w(Z_i,Z_j,\omega_n)|Z_i)-E(w(Z_i,Z_j,\omega_n)|Z_j)+E(w(Z_i,Z_j,\omega_n\right.}_{U_{1n}}\\
&\underbrace{+\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left[E(w(Z_i,Z_j,\omega_n)|Z_j)-E(w(Z_i,Z_j,\omega_n))\right]}_{U_{2n}}\\
&\underbrace{+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}E(w(Z_i,Z_j,\omega_n)|Z_i)}_{U_{3n}}.
\end{aligned}$$

We will show that

$$E(U_{1n})^2 = o(1), \tag{32}$$

$$U_{2n} = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left\{E(D|X_j)\times Y_jD_j-E(DY|X_j)\times D_j\right\}+o_p(1), \tag{33}$$

$$U_{3n} = o_p(1). \tag{34}$$

Observe that

$$\begin{aligned}
&E\left[\frac{\pi_1(X_i)Y_jD_j-\mu_1(X_i)D_j}{\pi_1^2(X_i)}\frac{1}{\omega_n^p}K\left(\frac{X_j-X_i}{\omega_n}\right)|Z_i\right]\\
&\overset{L.I.E.}{=}E\left[\frac{1}{\pi_1^2(X_i)}\left\{\pi_1(X_i)\frac{\mu_1(X_j)}{f(X_j)}-\mu_1(X_i)\frac{\pi_1(X_j)}{f(X_j)}\right\}\frac{1}{\omega_n^p}K\left(\frac{X_j-X_i}{\omega_n}\right)|Z_i\right]\\
=\;&\frac{1}{\pi_1^2(X_i)}\int\left[\pi_1(X_i)\mu_1(x)-\mu_1(X_i)\pi_1(x)\right]\frac{1}{\omega_n^p}K\left(\frac{x-X_i}{\omega_n}\right)dx\\
=\;&\frac{1}{\pi_1^2(X_i)}\int\left[\pi_1(X_i)\mu_1(X_i+u\omega_n)-\mu_1(X_i)\pi_1(X_i+u\omega_n)\right]K(u)\,du\\
&\overset{A1}{=}H(X_i)\times O(\omega_n^q),
\end{aligned}$$

for some uniformly bounded function H by condition A1. Therefore, $U_{3n}=O_p(\sqrt{n}\omega_n^q)=o_p(1)$ by condition A3 and this establishes (34).

Next observe that

$$
\begin{aligned}
& E\left[\frac{1}{\pi_1^2(X_i)}\left\{\pi_1(X_i)\,Y_jD_j-\mu_1(X_i)\,D_j\right\}\frac{1}{\omega_n^p}K\left(\frac{X_j-X_i}{\omega_n}\right)|Z_j\right] \\
= & \int\frac{1}{\pi_1^2(x)}\left\{\pi_1(x)\,Y_jD_j-\mu_1(x)\,D_j\right\}\frac{1}{\omega_n^p}K\left(\frac{X_j-x}{\omega_n}\right)f(x)\,dx \\
= & \int\frac{1}{\pi_1^2(X_j+u\omega_n)}\left\{\pi_1((X_j+u\omega_n))\,Y_jD_j-\mu_1((X_j+u\omega_n))\,D_j\right\}K(u)\,f((X_j+u\omega_n))\,du \\
= & \left\{\pi_1(X_j)\,Y_jD_j-\mu_1(X_j)\,D_j\right\}f(X_j)\int\frac{1}{\pi_1^2(X_j+u\omega_n)}K(u)\,du \\
& +\omega_n\left\{\pi_1'(X_j)\,Y_jD_j-\mu_1'(X_j)\,D_j\right\}f(X_j)\int\frac{1}{\pi_1^2(X_j+u\omega_n)}K(u)\,udu \\
& +....+\omega_n^q\left\{\pi_1^{(q)}(X_j)\,Y_jD_j-\mu_1^{(q)}(X_j)\,D_j\right\}f(X_j)\int\frac{1}{\pi_1^2(X_j+u\omega_n)}K(u)\,u^qdu \\
= & \left\{\pi_1(X_j)\,Y_jD_j-\mu_1(X_j)\,D_j\right\}\frac{f(X_j)}{\pi_1^2(X_j)}+O(\omega_n^q) \\
= & \frac{1}{\left\{E(D|X_j)\right\}^2}\left\{E(D|X_j)\times Y_jD_j-E(DY|X_j)\times D_j\right\}+O(\omega_n^q),
\end{aligned}
$$

by a dominated convergence theorem, given the uniform boundedness of $\pi_1(\cdot)$. Together with condition A3, we get (33).

One can establish (32) by essentially repeating the proof of Powell, Stoker and Stock (1989) lemma 3.1.

Combining (32), (33) and (34), we get that

$$
\begin{aligned}
& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{\hat{\mu}_1(X_i)}{\hat{\pi}_1(X_i)}-\frac{\mu_1(X_i)}{\pi_1(X_i)}\right\} \\
= & \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left\{\pi_1(X_j)\,Y_jD_j-\mu_1(X_j)\,D_j\right\}f(X_j)\frac{1}{\pi_1^2(X_j)}+o_p(1) \\
= & \frac{1}{\sqrt{n}}\sum_{j=1}^{n}D_j\frac{Y_{1j}-\beta_1(X_j)}{\Pr(D=1|X_j)}+o_p(1).
\end{aligned}
$$

■

Theorem 2 (consistency of $\hat{\rho}$)

Proof. Define

$$
\zeta=E\left\{\beta(X_i)\times1(\gamma\geq\beta(X_i))\right\},\quad\hat{\zeta}=\frac{1}{n}\sum_{i=1}^{n}\hat{\beta}(X_i)\times\bar{L}\left(\frac{\hat{\gamma}-\hat{\beta}(X_i)}{h_n}\right),
$$

so that

$$
\hat{\rho}=\frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_1(X_i)-\hat{\zeta},\quad\rho=E\left\{\beta_1(X)\right\}-\zeta.
$$

Now,

$$\begin{aligned}
\hat{\zeta}-\zeta &= \frac{1}{n}\sum_{i=1}^{n}\hat{\beta}\left(X_i\right)\bar{L}\left(\frac{\hat{\gamma}-\hat{\beta}\left(X_i\right)}{h_n}\right)-\zeta \\
&= \underbrace{\frac{1}{n}\sum_{i=1}^{n}\hat{\beta}\left(X_i\right)\bar{L}\left(\frac{\hat{\gamma}-\hat{\beta}\left(X_i\right)}{h_n}\right)-\frac{1}{n}\sum_{i=1}^{n}\beta\left(X_i\right)\bar{L}\left(\frac{\gamma-\beta\left(X_i\right)}{h_n}\right)}_{T_{1n}} \\
&\quad \underbrace{+\frac{1}{n}\sum_{i=1}^{n}\beta\left(X_i\right)\left[\bar{L}\left(\frac{\gamma-\beta\left(X_i\right)}{h_n}\right)-1\left\{\beta\left(X_i\right)\leq\gamma\right\}\right]}_{T_{2n}} \\
&\quad \underbrace{+\frac{1}{n}\sum_{i=1}^{n}\left\{\beta\left(X_i\right)1\left\{\beta\left(X_i\right)\leq\gamma\right\}-\zeta\right\}}_{=o_p(1),\text{ by standard WLLN.}}.
\end{aligned}$$

Now,

$$\begin{aligned}
\left|T_{1n}\right| &= \left|\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\beta}\left(X_i\right)\bar{L}\left(\frac{\hat{\gamma}-\hat{\beta}\left(X_i\right)}{h_n}\right)-\beta\left(X_i\right)\bar{L}\left(\frac{\gamma-\beta\left(X_i\right)}{h_n}\right)\right\}\right| \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{\beta}\left(X_i\right)-\beta\left(X_i\right)}{h_n}\right|\left|h_n\bar{L}\left(\frac{\tilde{\gamma}-\tilde{\beta}\left(X_i\right)}{h_n}\right)-\tilde{\beta}\left(X_i\right)L\left(\frac{\tilde{\gamma}-\tilde{\beta}\left(X_i\right)}{h_n}\right)\right| \\
&\quad +\frac{\left(\hat{\gamma}-\gamma\right)}{h_n}\times\frac{1}{n}\sum_{i=1}^{n}\left|\tilde{\beta}\left(X_i\right)L\left(\frac{\tilde{\gamma}-\tilde{\beta}\left(X_i\right)}{h_n}\right)\right| \\
&\leq \frac{\sup\left|\hat{\beta}\left(x\right)-\beta\left(x\right)\right|}{h_n}\frac{1}{n}\sum_{i=1}^{n}\left|h_n\bar{L}\left(\frac{\tilde{\gamma}-\tilde{\beta}\left(X_i\right)}{h_n}\right)-\tilde{\beta}\left(X_i\right)L\left(\frac{\tilde{\gamma}-\tilde{\beta}\left(X_i\right)}{h_n}\right)\right| \\
&\quad +\left(\frac{nh_n\left(\hat{\gamma}-\gamma\right)^2}{nh_n^3}\right)^{1/2}\times\frac{1}{n}\sum_{i=1}^{n}\left|\tilde{\beta}\left(X_i\right)L\left(\frac{\tilde{\gamma}-\tilde{\beta}\left(X_i\right)}{h_n}\right)\right|.
\end{aligned}$$

Since $L,\bar{L}$ are uniformly bounded, the above display is of the form

$$\leq \frac{\sup_{x\in\mathcal{X}}\left|\hat{\beta}\left(x\right)-\beta\left(x\right)\right|}{h_n}\times O_p\left(1\right)+\left(\frac{nh_n\left(\hat{\gamma}-\gamma\right)^2}{nh_n^3}\right)^{1/2}\times O_p\left(1\right).$$

Now, theorem 2 implies that $nh_n\left(\hat{\gamma}-\gamma\right)^2=O_p\left(1\right)$, conclusion B1 and condition B4 (i) imply that $\frac{\sup\left|\hat{\beta}(x)-\beta(x)\right|}{h_n}=o_p\left(1\right)$ and that $nh_n^3\to\infty$. Thus we have that $T_{1n}=o_p\left(1\right)$.

As for T_{2n}, observe that since $\beta\left(\cdot\right)$ is uniformly bounded, by using steps exactly analogous to step 2 in the proof of theorem 2 (leading to (13)), we will get by the DCT that $T_{2n}=o_p\left(1\right)$.

Now combine with lemma 3 to conclude that $|\hat{\rho}-\rho|=o_p(1)$. The proof of $|\rho_n-\rho|=o(1)$ is just a simpler version of (45) which is proved below. ■

Additional conditions:

The following additional conditions will be used to prove theorem 2.

B4 (ii) $\frac{\sqrt{n}}{h_n^2}\times\left\{\left(\frac{\ln n}{n\omega_n^p}\right)^{1/2}+\omega_n^q\right\}^2\to 0$; **B12.** $nh_n^6\to\infty$, r of condition B7 is at least 4 and $nh_n^{2r}\to 0$.

Theorem 2 (distribution of $\hat{\rho}$):

Proof. We will work with the following expansion

$$\begin{aligned}
&\hat{\zeta}-\zeta\\
=\;&\underbrace{\frac{1}{n}\sum_{i=1}^{n}\beta(X_i)\,\bar{L}\left(\frac{\gamma-\beta(X_i)}{h_n}\right)-\zeta}_{T_{1n}}\\
&\underbrace{+\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\beta}(X_i)-\beta(X_i)\right\}\left\{\bar{L}\left(\frac{\gamma-\beta(X_i)}{h_n}\right)-\frac{1}{h_n}\beta(X_i)\,L\left(\frac{\gamma-\beta(X_i)}{h_n}\right)\right\}}_{T_{2n}}\\
&\underbrace{+(\hat{\gamma}-\gamma)\frac{1}{nh_n}\sum_{i=1}^{n}\beta(X_i)\,L\left(\frac{\gamma-\beta(X_i)}{h_n}\right)}_{T_{3n}}\\
&\underbrace{-\frac{1}{4nh_n^2}\sum_{i=1}^{n}\left\{\hat{\beta}(X_i)-\beta(X_i)\right\}^2\left\{-2h_nL\left(\frac{\tilde{\gamma}-\tilde{\beta}(X_i)}{h_n}\right)+\tilde{\beta}(X_i)\,L'\left(\frac{\tilde{\gamma}-\tilde{\beta}(X_i)}{h_n}\right)\right\}}_{T_{4n}}\\
&\underbrace{+(\hat{\gamma}-\gamma)^2\times\frac{1}{4nh_n^2}\sum_{i=1}^{n}\beta(X_i)\,L'\left(\frac{\tilde{\gamma}-\tilde{\beta}(X_i)}{h_n}\right)}_{T_{5n}}\\
&\underbrace{+(\hat{\gamma}-\gamma)\frac{1}{2nh_n^2}\sum_{i=1}^{n}\left\{\hat{\beta}(X_i)-\beta(X_i)\right\}\left[h_nL\left(\frac{\tilde{\gamma}-\tilde{\beta}(X_i)}{h_n}\right)-\tilde{\beta}(X_i)\,L'\left(\frac{\tilde{\gamma}-\tilde{\beta}(X_i)}{h_n}\right)\right]}_{T_{6n}} \qquad (35)
\end{aligned}$$

Step 4: Under conclusion B1 and condition B4, the fourth term in (35) will be $o_p\left(\frac{1}{\sqrt{n}}\right)$ since $L'(\cdot)$ is assumed to be uniformly bounded in absolute value. As for the fifth term, observe by the previous theorem, that $\frac{(\hat{\gamma}-\gamma)^2}{h_n^2}=O_p\left(\frac{1}{nh_n^3}\right)=o_p\left(\frac{1}{\sqrt{n}}\right)$ by condition B12. So the fifth term in (35) will be $o_p\left(\frac{1}{\sqrt{n}}\right)$. That the sixth term is $o_p\left(\frac{1}{\sqrt{n}}\right)$ follows from combining the two previous results.

Step 5: The multiplier for the third term in (35) equals

$$\begin{aligned}\frac{1}{nh_n}\sum_{i=1}^{n}\beta(X_i)\,L\left(\frac{\beta(X_i)-\gamma}{h_n}\right) &\rightarrow E\left(\beta(X_i)\,\frac{1}{h_n}\bar{L}\left(\frac{\gamma-\beta(X_i)}{h_n}\right)\right)\\ &= \gamma f_\beta(\gamma)+O(h_n^r)\rightarrow\gamma f_\beta(\gamma),\end{aligned}$$

which follows from the standard consistency proof for e.g. kernel density estimates.

Combining steps 4 and 5, and replacing the asymptotic expansion of $(\hat{\gamma}-\gamma)$ from theorem 1, we get from (35) that

$$\begin{aligned}&\sqrt{n}\left\{\hat{\zeta}-\zeta\right\}\\ =\;&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\beta(X_i)\,\bar{L}\left(\frac{\gamma-\beta(X_i)}{h_n}\right)-\zeta\right\}-\frac{\gamma}{\sqrt{n}}\sum_{i=1}^{n}\left\{1\left(\beta(X_i)\leq\gamma\right)-F_\beta(\gamma)\right\}\\ &+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\hat{\beta}(X_i)-\beta(X_i)\right\}\left\{\bar{L}\left(\frac{\gamma-\beta(X_i)}{h_n}\right)+\left\{\gamma-\beta(X_i)\right\}\frac{1}{h_n}L\left(\frac{\gamma-\beta(X_i)}{h_n}\right)\right\}\\ &+\gamma\frac{1}{2\sqrt{n}h_n^2}\sum_{i=1}^{n}\left\{\hat{\beta}(X_i)-\beta(X_i)\right\}^2L'\left(\frac{\tilde{\beta}(X_i)-\gamma}{h_n}\right)+o_p(1).\end{aligned}\tag{36}$$

The third term in (36) in absolute value is dominated by

$$\frac{\gamma}{2}\times\left\{\left(\frac{\ln n}{n\omega_n^p}\right)^{1/2}+\omega_n^q\right\}^2\times\frac{\sqrt{n}}{h_n^2}\times O_p(1)=o_p(1),\text{ by condition B4(iii).}$$

Step 6A: Consider the first term in (36)

$$\begin{aligned}=\;&\underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\beta(X_i)\,\bar{L}\left(\frac{\gamma-\beta(X_i)}{h_n}\right)-\beta(X_i)\times1\left\{\beta(X_i)<\gamma\right\}\right]}_{T_{4n}}\\ &+\underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\beta(X_i)\times1\left\{\beta(X_i)<\gamma\right\}-\zeta\right]}_{T_{5n}=O_p(1),\text{ by CLT.}}\end{aligned}\tag{37}$$

We will show that $T_{4n}=o_p(1)$ using the arguments similar to the ones used for showing (16).

Define $g_\beta(a) = af_\beta(a)$ and $G_\beta(a) = \int_{-M}^{a} g_\beta(t)\,dt$. Then

$$\begin{aligned}
E(T_{4n}) &= \sqrt{n}E\left\{\beta(X_i)\bar{L}\left(\frac{\gamma-\beta(X_i)}{h_n}\right) - G_\beta(\gamma)\right\} \\
&= \sqrt{n}\left\{\int_{-M}^{M}\bar{L}\left(\frac{\gamma-t}{h_n}\right)g_\beta(t)\,dt - G_\beta(\gamma)\right\} \\
&= \sqrt{n}\left[\left\{\bar{L}\left(\frac{\gamma-t}{h_n}\right)G_\beta(t)\right\}|_{-M}^{M} + \int_{-M}^{M}\frac{1}{h_n}L\left(\frac{\gamma-t}{h_n}\right)G_\beta(t)\,dt - G_\beta(\gamma)\right] \\
&= \sqrt{n}\left[\int_{\frac{\gamma-M}{h_n}}^{\frac{\gamma+M}{h_n}} L(u)\,G_\beta(\gamma-uh_n)\,du - G_\beta(\gamma)\right] = O\left(\sqrt{n}h_n^r\right) \to 0, \text{ by B1} \qquad (38)
\end{aligned}$$

Next, define $g_\beta(t) = t^2 f_\beta(t)$ and $G_\beta(t) = \int_{-M}^{t} g_\beta(a)\,da$. Then

$$\begin{aligned}
& E\left\{\beta(X_i)\bar{L}\left(\frac{\gamma-\beta(X_i)}{h_n}\right) - \beta(X_i)\times 1\{\beta(X_i)<\gamma\}\right\}^2 \\
= & \int_{-M}^{M}\left\{\bar{L}\left(\frac{\gamma-t}{h_n}\right) - 1\{t<\gamma\}\right\}^2 g_\beta(t)\,dt \\
= & \int_{-M}^{M}\bar{L}^2\left(\frac{\gamma-t}{h_n}\right)g_\beta(t)\,dt + \int_{-M}^{\gamma}1\{t<\gamma\}\,g_\beta(t)\,dt \\
& -2\int_{-M}^{M}\bar{L}\left(\frac{\gamma-t}{h_n}\right)\times 1\{t<\gamma\}\times g_\beta(t)\,dt
\end{aligned}$$

Using the DCT repeatedly (c.f. the steps leading to (14)), we get that

$$\begin{aligned}
& E\left\{\beta(X_i)\bar{L}\left(\frac{\gamma-\beta(X_i)}{h_n}\right) - \beta(X_i)\times 1\{\beta(X_i)<\gamma\}\right\}^2 \\
= & \int_{-M}^{M}\bar{L}^2\left(\frac{\gamma-t}{h_n}\right)g_\beta(t)\,dt + \int_{-M}^{\gamma}1\{t<\gamma\}\,g_\beta(t)\,dt \\
& -2\int_{-M}^{M}\bar{L}\left(\frac{\gamma-t}{h_n}\right)\times 1\{t<\gamma\}\times g_\beta(t)\,dt \\
\to & \; G_\beta(\gamma) + G_\beta(\gamma) - 2G_\beta(\gamma) = 0.
\end{aligned}$$

This implies that for T_{4n} defined in (37),

$$\begin{aligned}
Var(T_{4n}) &= Var\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\beta(X_i)\bar{L}\left(\frac{\gamma-\beta(X_i)}{h_n}\right) - \beta(X_i)\times 1\{\beta(X_i)<\gamma\}\right]\right) \\
&= Var\left(\beta(X_i)\times\bar{L}\left(\frac{\gamma-\beta(X_i)}{h_n}\right) - \beta(X_i)\times 1\{\beta(X_i)<\gamma\}\right) \\
&\le E\left\{\beta(X_i)\bar{L}\left(\frac{\gamma-\beta(X_i)}{h_n}\right) - \beta(X_i)\times 1\{\beta(X_i)<\gamma\}\right\}^2 \\
&\to 0. \qquad (39)
\end{aligned}$$

From (38) and (39), we get that $E\left(T_{4n}\right)^2 \to 0$ and thus $T_{4n} = o_p\left(1\right)$.

Replacing in (36), we get that

$$\begin{aligned}
& \sqrt{n}\left\{\hat{\zeta} - \zeta\right\} \\
= & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\beta\left(X_i\right) \times 1\left\{\beta\left(X_i\right) < \gamma\right\} - \zeta\right] - \frac{\gamma}{\sqrt{n}}\sum_{i=1}^{n}\left\{1\left(\beta\left(X_i\right) \leq \gamma\right) - F_\beta\left(\gamma\right)\right\} \\
& + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\hat{\beta}\left(X_i\right) - \beta\left(X_i\right)\right\}\left\{\bar{L}\left(\frac{\gamma - \beta\left(X_i\right)}{h_n}\right) + \left\{\gamma - \beta\left(X_i\right)\right\}\frac{1}{h_n}L\left(\frac{\gamma - \beta\left(X_i\right)}{h_n}\right)\right\} \\
& + o_p\left(\frac{1}{\sqrt{n}}\right). \qquad (40)
\end{aligned}$$

The final step is to analyze the third term in (40), using U-statistic type decompositions. First notice that analogous to (23) above, we have here that up to $o_p\left(1\right)$ terms:

$$\begin{aligned}
& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\hat{\beta}\left(X_i\right) - \beta\left(X_i\right)\right\}\left\{\bar{L}\left(\frac{\gamma - \beta\left(X_i\right)}{h_n}\right) + \left\{\gamma - \beta\left(X_i\right)\right\}\frac{1}{h_n}L\left(\frac{\gamma - \beta\left(X_i\right)}{h_n}\right)\right\} \\
= & \sqrt{n}\frac{1}{n\left(n-1\right)}\sum_{i=1}^{n}\sum_{j\neq i}\left[\begin{array}{c} \frac{1}{\pi^2(X_i)}\left\{\pi\left(X_i\right)Y_jD_j - \mu_1\left(X_i\right)D_j\right\}\frac{1}{\omega_n^p}K\left(\frac{X_j - X_i}{\omega_n}\right) \\ \times\left[\bar{L}\left(\frac{\gamma-\beta(X_i)}{h_n}\right) + \left\{\gamma - \beta\left(X_i\right)\right\}\frac{1}{h_n}L\left(\frac{\gamma-\beta(X_i)}{h_n}\right)\right] \end{array}\right] \\
\equiv & \sqrt{n}\frac{1}{n\left(n-1\right)}\sum_{i=1}^{n}\sum_{j\neq i}w_n\left(Z_i, Z_j\right) \\
= & \underbrace{\frac{1}{\sqrt{n}\left(n-1\right)}\sum_{i=1}^{n}\sum_{j\neq i}\left[w_n\left(Z_i, Z_j\right) - E\left(w_n\left(Z_i, Z_j\right)|Z_i\right) - E\left(w_n\left(Z_i, Z_j\right)|Z_j\right) + E\left(w_n\left(Z_i, Z_j\right)\right)\right]}_{U_{1n}} \\
& \underbrace{+\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left[E\left(w_n\left(Z_i, Z_j\right)|Z_j\right) - E\left(w_n\left(Z_i, Z_j\right)\right)\right]}_{U_{2n}} + \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}E\left(w_n\left(Z_i, Z_j\right)|Z_i\right)}_{U_{3n}}.
\end{aligned}$$

It is straightforward (replace the kernel involving terms) to verify that we will get the same conclusion as (25) and (24) here. So we only perform the analysis for U_{2n}.

Using steps similar to the case for $\hat{\gamma}$, one gets that

$$\begin{aligned}
& E\left(w_n\left(Z_i,Z_j\right)|Z_j\right) \\
= & E\left\{\left[\begin{array}{c}\frac{1}{\pi_1^2(X_i)}\left\{\pi_1\left(X_i\right)Y_jD_j-\mu_1\left(X_i\right)D_j\right\}\frac{1}{\omega_n^p}K\left(\frac{-X_j+X_i}{\omega_n}\right)\\ \times\left[\bar{L}\left(\frac{\gamma-\beta(X_i)}{h_n}\right)+\left\{\gamma-\beta\left(X_i\right)\right\}\frac{1}{h_n}L\left(\frac{\gamma-\beta(X_i)}{h_n}\right)\right]\end{array}\right]|Y_j,D_j,X_j\right\} \\
= & \int\left[\begin{array}{c}\frac{1}{\pi_1^2(X)}\left\{\pi_1\left(x\right)Y_jD_j-\mu_1\left(x\right)D_j\right\}\frac{1}{\omega_n^p}K\left(\frac{-X_j+x}{\omega_n}\right)\\ \times\left[\bar{L}\left(\frac{\gamma-\beta(X_i)}{h_n}\right)+\left\{\gamma-\beta\left(X_i\right)\right\}\frac{1}{h_n}L\left(\frac{\gamma-\beta(X_i)}{h_n}\right)\right]\end{array}\right]f\left(x\right)dx \\
= & \int\left[\begin{array}{c}\frac{1}{\pi_1^2(X_j+u\omega_n)}\left\{\pi_1\left(X_j+u\omega_n\right)Y_jD_j-\mu_1\left(X_j+u\omega_n\right)D_j\right\}K\left(u\right)\\ \times\left[\bar{L}\left(\frac{\gamma-\beta(X_j+u\omega_n)}{h_n}\right)+\left\{\gamma-\beta\left(X_j+u\omega_n\right)\right\}\frac{1}{h_n}L\left(\frac{\gamma-\beta(X_j+u\omega_n)}{h_n}\right)\right]\end{array}\right]f\left(X_j+u\omega_n\right)du \\
= & \left[\begin{array}{c}\frac{1}{\pi_1^2(X_j)}\left\{\pi_1\left(X_j\right)Y_jD_j-\mu_1\left(X_j\right)D_j\right\}f\left(X_j\right)\\ \times\left[\bar{L}\left(\frac{\gamma-\beta(X_j)}{h_n}\right)+\left\{\gamma-\beta\left(X_j\right)\right\}\frac{1}{h_n}L\left(\frac{\gamma-\beta(X_j)}{h_n}\right)\right]\end{array}\right]\int K\left(u\right)du+O\left(\omega_n^q\right) \\
= & \left[\begin{array}{c}\underbrace{\left[\frac{\left\{\pi_1\left(X_j\right)Y_jD_j-\mu_1\left(X_j\right)D_j\right\}}{\pi_1^2\left(X_j\right)}f\left(X_j\right)\right]}_{W(Z_j)}\\ \times\underbrace{\left[\bar{L}\left(\frac{\gamma-\beta\left(X_j\right)}{h_n}\right)+\left\{\gamma-\beta\left(X_j\right)\right\}\frac{1}{h_n}L\left(\frac{\gamma-\beta\left(X_j\right)}{h_n}\right)\right]}_{V_{2n}(\beta(X_j))}\end{array}\right]+O\left(\omega_n^q\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
& \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left[E\left(w_n\left(Z_i,Z_j\right)|Z_j\right)-E\left(w_n\left(Z_i,Z_j\right)\right)\right] \\
= & \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left[E\left(w_n\left(Z_i,Z_j\right)|Z_j\right)-W\left(Z_j\right)\times 1\left(\beta\left(X_j\right)\leq\gamma\right)\right] \\
& +\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left\{W\left(Z_j\right)\times 1\left(\beta\left(X_j\right)\leq\gamma\right)\right\} \\
= & \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left\{W\left(Z_j\right)\times 1\left(\beta\left(X_j\right)\leq\gamma\right)\right\} \\
& +\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\underbrace{W\left(Z_j\right)\left[\begin{array}{c}\left\{\bar{L}\left(\frac{\gamma-\beta(X_j)}{h_n}\right)-1\left(\beta\left(X_j\right)\leq\gamma\right)\right\}\\ +\left\{\gamma-\beta\left(X_j\right)\right\}\frac{1}{h_n}L\left(\frac{\gamma-\beta(X_j)}{h_n}\right)\end{array}\right]}_{T_{nj}}. \qquad (41)
\end{aligned}$$

Now, we will show that the second term in the previous display is $o_p\left(1\right)$. Letting $\omega^2\left(t\right)=$

$E\left(W^2\left(Z_j\right)|\beta\left(X_j\right)=t\right)$, we have that

$$\begin{aligned}
& E\left(T_{nj}^2\right) \\
= & \int_{-M}^{M}\omega^2\left(t\right)\left\{\bar{L}\left(\frac{\gamma-t}{h_n}\right)-1\left(t\leq\gamma\right)\right\}^2 f_\beta\left(t\right)dt \\
& +\int_{-M}^{M}\omega^2\left(t\right)\left(\frac{\gamma-t}{h_n}\right)^2 L^2\left(\frac{\gamma-t}{h_n}\right) f_\beta\left(t\right)dt \\
& +2\int_{-M}^{M}\omega^2\left(t\right)\bar{L}\left(\frac{\gamma-t}{h_n}\right)\left(\frac{\gamma-t}{h_n}\right) L\left(\frac{\gamma-t}{h_n}\right) f_\beta\left(t\right)dt. \qquad (42)
\end{aligned}$$

The first term in (42) equals

$$\begin{aligned}
& \int_{-M}^{M}\omega^2\left(t\right)\left\{\bar{L}\left(\frac{\gamma-t}{h_n}\right)-1\left(t\leq\gamma\right)\right\}^2 f_\beta\left(t\right)dt \\
= & \int_{-M}^{\gamma}\omega^2\left(t\right)\left\{\bar{L}\left(\frac{\gamma-t}{h_n}\right)-1\right\}^2 f_\beta\left(t\right)dt+\int_{\gamma}^{M}\omega^2\left(t\right)\left\{\bar{L}\left(\frac{\gamma-t}{h_n}\right)\right\}^2 f_\beta\left(t\right)dt
\end{aligned}$$

and both of the terms in the previous display converge to zero by the DCT since $\lim_{a\to\infty}\bar{L}\left(u\right)=1=1-\lim_{a\to-\infty}\bar{L}\left(u\right)$. The second integral in (42) converges to zero by the DCT since $\lim_{u\to\pm\infty}u^2L^2\left(u\right)=0$. The third integral in (42) also converges to zero by $\lim_{u\to\pm\infty}uL\left(u\right)=0$ and the DCT. This implies that $E\left(T_{nj}^2\right)\to 0$ and thus

$$0<Var\left(\frac{1}{\sqrt{n}}\sum_{j=1}^{n}T_{nj}\right)=Var\left(T_{nj}\right)\leq E\left(T_{nj}^2\right)\to 0.$$

Next,

$$\begin{aligned}
& \sqrt{n}E\left\{W\left(Z_j\right)\left[\begin{array}{c}\left\{\bar{L}\left(\frac{\gamma-\beta(X_j)}{h_n}\right)-1\left(\beta\left(X_j\right)\leq\gamma\right)\right\} \\ +\left\{\gamma-\beta\left(X_j\right)\right\}\frac{1}{h_n}L\left(\frac{\gamma-\beta(X_j)}{h_n}\right)\end{array}\right]\right\} \\
= & \sqrt{n}E_{X_j}\left\{\underset{=0}{E\left\{W\left(Z_j\right)|X_j\right\}}\times\left[\begin{array}{c}\left\{\bar{L}\left(\frac{\gamma-\beta(X_j)}{h_n}\right)-1\left(\beta\left(X_j\right)\leq\gamma\right)\right\} \\ +\left\{\gamma-\beta\left(X_j\right)\right\}\frac{1}{h_n}L\left(\frac{\gamma-\beta(X_j)}{h_n}\right)\end{array}\right]\right\}=0.
\end{aligned}$$

So it follows that the second term in (41) is $o_p\left(1\right)$.

Thus we have that

$$\begin{aligned}
& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{\hat{\mu}_1\left(X_i\right)}{\hat{\pi}_1\left(X_i\right)}-\frac{\mu_1(X_i)}{\pi\left(X_i\right)}\right\}\left\{\bar{L}\left(\frac{\gamma-\beta\left(X_i\right)}{h_n}\right)+\left\{\gamma-\beta\left(X_i\right)\right\}\frac{1}{h_n}L\left(\frac{\gamma-\beta\left(X_i\right)}{h_n}\right)\right\} \\
= & \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left\{\frac{\left\{\pi_1\left(X_j\right)Y_jD_j-\mu_1\left(X_j\right)D_j\right\}}{\pi^2\left(X_j\right)}f_X\left(X_j\right)\times 1\left(\beta\left(X_j\right)\leq\gamma\right)\right\}+o_p\left(1\right) \\
& \xrightarrow{d} N\left(0,\int_{-M}^{\gamma}\omega^2\left(t\right)f_\beta\left(t\right)dt\right),
\end{aligned}$$

where

$$\omega^2(a) = E\left\{\left[\frac{\{\pi_1(X_j)\,Y_j D_j - \mu_1(X_j)\,D_j\}}{\pi^2(X_j)} f_X(X_j)\right]^2 |\beta(X_j) = a\right\}$$

Using exactly analogous steps, we will also get that

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{\hat{\mu}_0(X_i)}{\hat{\pi}_0(X_i)} - \frac{\mu_0(X_i)}{\pi_0(X_i)}\right\}\left\{\bar{L}\left(\frac{\gamma-\beta(X_j)}{h_n}\right) + \{\gamma-\beta(X_j)\}\frac{1}{h_n}L\left(\frac{\gamma-\beta(X_j)}{h_n}\right)\right\} \\
= &\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left\{\frac{\pi_0(X_j)\,Y_j(1-D_j) - \mu_0(X_j)(1-D_j)}{\pi_0^2(X)} f_X(X_j)\times 1(\beta(X_j)\le\gamma)\right\} + o_p(1). \\
&\xrightarrow{d} N\left(0, \int_{-M}^{\gamma}\tau^2(t)\,f_\beta(t)\,dt\right),
\end{aligned}$$

where

$$\tau^2(a) = E\left\{\left\{\frac{\pi_0(X)\,Y(1-D) - \mu_0(X)(1-D)}{\pi_0^2(X)} f(X)\right\}^2 |\beta(X) = a\right\}.$$

Finally, we get that

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\hat{\beta}(X_i) - \beta(X_i)\right\}\left\{\bar{L}\left(\frac{\gamma-\beta(X_j)}{h_n}\right) + \{\gamma-\beta(X_j)\}\frac{1}{h_n}L\left(\frac{\gamma-\beta(X_j)}{h_n}\right)\right\} \\
&\xrightarrow{d} N\left(0, \int_{-M}^{\gamma}\left\{\omega^2(t) + \tau^2(t)\right\} f_\beta(t)\,dt\right), \qquad (43)
\end{aligned}$$

since the covariances will be zero (as can be easily seen from the asymptotic linear expansions because $D(1-D) = 0$).

Replacing in (40), and noticing that

$$\frac{\pi_1(X_i)\,Y_i D_i - \mu_1(X_i)\,D_i}{\pi^2(X_i)} \times f_X(X_i) = \frac{D_i}{\Pr(D=1|X_i)}\{Y_{1i} - E(Y_1|X_i)\},$$

and

$$\begin{aligned}
&\frac{\pi_0(X_i)\,Y_i(1-D_i) - \mu_0(X_i)(1-D_i)}{\pi_0^2(X_i)} \times f_X(X_i) \\
= &\frac{1-D_i}{1-\Pr(D=1|X_i)}\{Y_{0i} - E(Y_0|X_i)\},
\end{aligned}$$

we finally arrive at

$$\begin{aligned}
&\sqrt{n}\left\{\hat{\zeta}-\zeta\right\} \\
= \;& \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left\{\frac{D_i}{\Pr\left(D=1|X_i\right)}\left\{Y_{1i}-E\left(Y_1|X_i\right)\right\}\times 1\left(\beta\left(X_j\right)\leq\gamma\right)\right\} \\
& -\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left\{\frac{1-D_i}{1-\Pr\left(D=1|X_i\right)}\left\{Y_{0i}-E\left(Y_0|X_i\right)\right\}\times 1\left(\beta\left(X_j\right)\leq\gamma\right)\right\} \\
& +\gamma\times\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left\{F_\beta\left(\gamma\right)-1\left(\beta\left(X_j\right)\leq\gamma\right)\right\} \\
& +\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\beta\left(X_i\right)\times 1\left\{\beta\left(X_i\right)<\gamma\right\}-\zeta\right\}+o_p\left(1\right). \qquad (44)
\end{aligned}$$

Now, observe that

$$\begin{aligned}
&\hat{\rho}-\rho \\
= \;& \underbrace{\frac{1}{n}\sum_{j=1}^{n}\left[\hat{\beta}_1\left(X_j\right)-\beta_1\left(X_j\right)\right]}_{D_{1n}}+\underbrace{\frac{1}{n}\sum_{j=1}^{n}\left[\beta_1\left(X_j\right)-E\left\{\beta_1\left(X_j\right)\right\}\right]}_{D_{2n}}-\left\{\hat{\zeta}-\zeta\right\}.
\end{aligned}$$

Lemma 3 describes the behavior of D_{1n}, and D_{2n} is a standard empirical process which yields that $\sqrt{n}\left(\hat{\rho}-\rho\right)$ tends to a zero-mean normal. The final step is to show

$$\rho_n-\rho=o\left(\frac{1}{\sqrt{n}}\right). \qquad (45)$$

Derivation of (45): Let $\sigma\left(\cdot\right)$ denote the asymptotic variance of $n^\alpha\left\{\hat{\beta}\left(x\right)-\hat{\gamma}-\beta\left(x\right)+\gamma\right\}$, where $\alpha<\frac{1}{2}$ (by theorem 1 and standard results on asymptotic distribution of Nadaraya-Watson estimator) and let the support of $\frac{\gamma-\beta(X)}{\sigma(X)}$ be $(-A,A)$ for $A>0$. Observe that

$$\begin{aligned}
r_n \;:\; &= \int_{\mathcal{X}}\left(\gamma-\beta\left(x\right)\right)\Pr\left(\hat{\beta}\left(x\right)\geq\hat{\gamma}\right)1\left(\beta\left(x\right)<\gamma\right)dF_X\left(x\right) \\
&= \int_{\mathcal{X}}\left(\gamma-\beta\left(x\right)\right)\Pr\left(\begin{array}{c} n^\alpha\left\{\hat{\beta}\left(x\right)-\hat{\gamma}-\beta\left(x\right)+\gamma\right\} \\ \geq n^\alpha\left\{\gamma-\beta\left(x\right)\right\}\end{array}\right)1\left(\beta\left(x\right)<\gamma\right)dF_X\left(x\right) \\
&\overset{\text{Edgeworth}}{=} \int_{\mathcal{X}}\frac{\gamma-\beta\left(x\right)}{\sigma\left(x\right)}\Phi\left(-n^\alpha\frac{\gamma-\beta\left(x\right)}{\sigma\left(x\right)}\right)1\left(\beta\left(x\right)<\gamma\right)dF_X\left(x\right)+\text{smaller order terms} \\
&\leq c\int_0^A z\Phi\left(-n^\alpha z\right)f_{\frac{\gamma-\beta(X)}{\sigma(X)}}\left(z\right)dz+\text{smaller order terms} \\
&= \frac{1}{n^{2\alpha}}\underbrace{\int_0^{An^\alpha}t\Phi\left(-t\right)f_{\frac{\gamma-\beta(X)}{\sigma(X)}}\left(\frac{t}{n^\alpha}\right)dt}_{=r_{1n},\text{ say}}+\text{smaller order terms}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\sqrt{n}r_{1n} &= \frac{\sqrt{n}}{n^{2\alpha}}\int_0^{An^\alpha} t\Phi\left(-t\right) f_{\frac{\gamma-\beta(X)}{\sigma(X)}}\left(\frac{t}{n^\alpha}\right)dt \\
&\leq c\frac{1}{n^{2\alpha-1/2}}\int_0^{An^\alpha} t\Phi\left(-t\right)dt \\
\overset{\text{L'Hospital}}{\to} c\frac{A^2n^\alpha\Phi\left(-An^\alpha\right)\alpha n^{\alpha-1}}{\left(2\alpha-1/2\right)n^{2\alpha-3/2}} &= c\frac{A^2n^{1/2}\Phi\left(-An^\alpha\right)\alpha}{\left(2\alpha-1/2\right)} \to 0 \text{ as } n\to\infty.
\end{aligned}$$

Now observe that

$$\begin{aligned}
\rho_n-\rho &= \int_{\mathcal{X}}\left[\begin{pmatrix} \left\{\beta\left(x\right)-\gamma\right\}\Pr\left\{\hat{\beta}\left(x\right)\geq\hat{\gamma}\right\}1\left(\beta\left(x\right)<\gamma\right) \\ +\left\{\gamma-\beta\left(x\right)\right\}\Pr\left\{\hat{\beta}\left(x\right)<\hat{\gamma}\right\}1\left(\beta\left(x\right)\geq\gamma\right)\end{pmatrix}\right]dF_X\left(x\right) \\
&\quad +\gamma\int_{\mathcal{X}}\begin{pmatrix} \Pr\left\{\hat{\beta}\left(x\right)\geq\hat{\gamma}\right\}1\left(\beta\left(x\right)<\gamma\right) \\ -\Pr\left\{\hat{\beta}\left(x\right)<\hat{\gamma}\right\}1\left(\beta\left(x\right)\geq\gamma\right)\end{pmatrix}dF_X\left(x\right).
\end{aligned}$$

The first term is $o\left(\frac{1}{\sqrt{n}}\right)$ by the previous calculation. Finally, the 2nd term is proportional to

$$\begin{aligned}
&\int_{\mathcal{X}}\left[\begin{pmatrix} \Pr\left\{\hat{\beta}\left(x\right)\geq\hat{\gamma}\right\}1\left(\beta\left(x\right)<\gamma\right) \\ -\Pr\left\{\hat{\beta}\left(x\right)<\hat{\gamma}\right\}1\left(\beta\left(x\right)\geq\gamma\right)\end{pmatrix}\right]dF\left(x\right) \\
= &\int_{\mathcal{X}}\left[\begin{pmatrix} \Pr\left\{\hat{\beta}\left(x\right)-\hat{\gamma}-\beta\left(x\right)+\gamma\geq\gamma-\beta\left(x\right)\right\}1\left(\beta\left(x\right)<\gamma\right) \\ -\Pr\left\{\hat{\beta}\left(x\right)-\hat{\gamma}-\beta\left(x\right)+\gamma<\gamma-\beta\left(x\right)\right\}1\left(\beta\left(x\right)\geq\gamma\right)\end{pmatrix}\right]dF\left(x\right) \\
= &\int_{\mathcal{X}}\left[\begin{pmatrix} \Pr\left\{n^\alpha\left(\hat{\beta}\left(x\right)-\hat{\gamma}-\beta\left(x\right)+\gamma\right)\geq n^\alpha\left\{\gamma-\beta\left(x\right)\right\}\right\}1\left(\beta\left(x\right)<\gamma\right) \\ -\Pr\left\{n^\alpha\left(\hat{\beta}\left(x\right)-\hat{\gamma}-\beta\left(x\right)+\gamma\right)<n^\alpha\left\{\gamma-\beta\left(x\right)\right\}\right\}1\left(\beta\left(x\right)\geq\gamma\right)\end{pmatrix}\right]dF\left(x\right) \\
= &\int_{\mathcal{X}}\left[\begin{pmatrix} \left[1-\Phi\left(n^\alpha\frac{\gamma-\beta(x)}{\sigma(x)}\right)\right]1\left(\beta\left(x\right)<\gamma\right) \\ -\Phi\left\{n^\alpha\frac{\gamma-\beta(x)}{\sigma(x)}\right\}1\left(\beta\left(x\right)\geq\gamma\right)\end{pmatrix}\right]dF\left(x\right)+\text{smaller order terms.} \qquad (46)
\end{aligned}$$

■

The leading term in (46) multiplied by $\sqrt{n}$ equals

$$\sqrt{n}\int_{-A}^{A}\left[\begin{pmatrix}\left[1-\Phi\left(n^\alpha z\right)\right]1\left(z>0\right) \\ -\Phi\left\{n^\alpha z\right\}1\left(z<0\right)\end{pmatrix}\right]f_{\frac{\gamma-\beta(X)}{\sigma(X)}}\left(z\right)dz.$$

Now, observe that

$$\begin{aligned}\sqrt{n}\int_{-A}^{A}\left[1-\Phi\left(n^{\alpha}z\right)\right]1\left(z>0\right)f_{\frac{\gamma-\beta(X)}{\sigma(X)}}\left(z\right)dz &\leq c\sqrt{n}\int_{0}^{A}\Phi\left(-n^{\alpha}z\right)dz \\ &= c\frac{\sqrt{n}}{n^{\alpha}}\int_{0}^{An^{\alpha}}\Phi\left(-t\right)dt, \\ \overset{\text{L'Hospital}}{\to} c\frac{A\alpha n^{\alpha-1}\Phi\left(-An^{\alpha}\right)}{\left(\alpha-1/2\right)n^{\alpha-3/2}} &= c\sqrt{n}\Phi\left(-An^{\alpha}\right)\to 0.\end{aligned}$$

An exactly symmetric argument applies to $\sqrt{n}\int_{-A}^{A}\Phi\left\{n^{\alpha}z\right\}1\left(z<0\right)f_{\frac{\gamma-\beta(X)}{\sigma(X)}}\left(z\right)dz$.

Consequently, $\rho_n-\rho=o\left(\frac{1}{\sqrt{n}}\right)$ and hence a $\sqrt{n}$-consistent CI for ρ is also a valid $\sqrt{n}$-consistent CI for ρ_n.

Outline of proof for parametric $\beta\left(\cdot\right)$: Suppose $\beta\left(x\right)$ is parametrically specified as $G\left(x,\beta\right)$, where $G\left(\cdot\right)$ is known; typically β (the so-called "pseudo-true value") can be estimated at parametric rates using, say, GMM. For estimation of γ and ρ resulting from the plug-in approach, we will still use smoothing with the c.d.f. kernel $\bar{L}\left(\cdot\right)$ to handle the nonsmoothness, since smoothing-based methods are more generally applicable.

The key result is that both γ and ρ can be estimated at the $\sqrt{n}$-rate. To see this, recall the asymptotic expansion for $\hat{\gamma}$:

$$\begin{aligned}&\sqrt{n}\left(\hat{\gamma}-\gamma\right) \\ =\ &\left\{\hat{f}_{\hat{\beta}}\left(\tilde{\gamma}\right)\right\}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{F_{\beta}\left(\gamma\right)-\bar{L}\left(\frac{\gamma-G\left(X_i,\beta\right)}{h_n}\right)\right\} \\ &+\left\{\hat{f}_{\hat{\beta}}\left(\tilde{\gamma}\right)\right\}^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\bar{L}\left(\frac{\gamma-G\left(X_i,\beta\right)}{h_n}\right)-\bar{L}\left(\frac{\gamma-G\left(X_i,\hat{\beta}\right)}{h_n}\right)\right]\right\}.\end{aligned}$$

Using similar steps as in the proof of theorem 2 in the appendix, the first term is asymptotically normal with mean equal to

$$\lim_{n\to\infty}\sqrt{n}h_n^{r}\times\left[\frac{(-1)^{r+1}f_{\beta}^{(r-1)}\left(\gamma\right)\times\int_{-1}^{1}u^{r}L\left(u\right)du}{r!}\right]+o\left(\sqrt{n}h_n^{r}\right)$$

which is finite if $\lim_{n\to\infty}\sqrt{n}h_n^{r}<\infty$.

As for the second term, (and this is what makes $\hat{\gamma}$ a $\sqrt{n}$-consistent estimator in the

parametric case) notice that

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\bar{L}\left(\frac{\gamma-G\left(X_i,\beta\right)}{h_n}\right)-\bar{L}\left(\frac{\gamma-G\left(X_i,\hat{\beta}\right)}{h_n}\right)\right]\\
=\;&\sqrt{n}\left(\hat{\beta}-\beta\right)'\frac{1}{nh_n}\sum_{i=1}^{n}\nabla G\left(X_i,\beta\right)L\left(\frac{\gamma-G\left(X_i,\beta\right)}{h_n}\right)+T_n,
\end{aligned}$$

where

$$|T_n|\leq M\frac{n\left\|\hat{\beta}-\beta\right\|^2}{2\sqrt{n}h_n^2}\frac{1}{n}\sum_{i=1}^{n}M_1\left(X_i\right)L'\left(\frac{G\left(X_i,\tilde{\beta}\right)-\gamma}{h_n}\right),$$

with M a fixed positive constant and $M_1\left(X\right)$ a uniformly bounded function. Since $\sqrt{n}\left(\hat{\beta}-\beta\right)=O_p\left(1\right)$, by conditions B4(i) and A4 (ii), the RHS of the previous display goes to zero if $nh_n^4\to\infty$. Then we have that

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\bar{L}\left(\frac{\gamma-G\left(X_i,\beta\right)}{h_n}\right)-\bar{L}\left(\frac{\gamma-G\left(X_i,\hat{\beta}\right)}{h_n}\right)\right]\\
=\;&\left[\sqrt{n}\left(\hat{\beta}-\beta\right)'\nabla G\left(X_i,\beta\right)\right]\times f_{G(X,\beta)}\left(\gamma\right)+o_p\left(1\right).
\end{aligned}$$

This implies that $\sqrt{n}\left(\hat{\gamma}-\gamma\right)$ will converge to a zero mean normal if $nh_n^{2r}\to 0$ and $n^3h_n^4\to\infty$ and when the density of $G\left(X,\beta\right)$ has uniformly bounded derivatives up to order $(r-1)$ where $r\geq 3$. The result for $\hat{\rho}$ will follow.

9.2 Part 2: Split sample simulation

Here we describe and perform a simulation exercise, using a split sample method in order to address the potential positive bias alluded to in section 4. Divide the sample randomly into 2 parts: call the first group $j=1,...m$ and call 2nd group $i=m+1,...n$. Then for each $i=m+1,...n$, calculate $\hat{\beta}^{(1)}\left(x_i\right)$ using only observations $j=1,...m$. Then use the second half to calculate a second independent estimate $\hat{\beta}^{(2)}\left(x_i\right)$ at the same set of values x_i as before. Then $\hat{\beta}^{(1)}\left(x_i\right)$ and $\hat{\beta}^{(2)}\left(x_i\right)$ are functions of distinct Y observations which are independent. Finally, calculate $\hat{\gamma}$ and $\hat{\rho}$ via:

$$\begin{aligned}
1-c &= \frac{1}{n}\sum_{i=m+1}^{n}\bar{L}\left(\frac{\hat{\gamma}-\hat{\beta}^{(2)}\left(x_i\right)}{h_n}\right)\\
\hat{\rho} &= \frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_1\left(X_i\right)-\frac{1}{n}\sum_{i=m+1}^{n}\hat{\beta}^{(1)}\left(x_i\right)\times\bar{L}\left(\frac{\hat{\gamma}-\hat{\beta}^{(2)}\left(x_i\right)}{h_n}\right).
\end{aligned}$$

Such split-sample methods were previously suggested in Altonji and Segal (1996) to reduce finite sample bias of GMM estimates. While this split-sample method does not change the asymptotic distribution of $\hat{\rho}$ or $\hat{\gamma}$, we investigate its potential finite sample superiority in a small Monte Carlo study as follows.

Take X=wealth and take the population to be our sample. Generate $\varepsilon \tilde{} N(0,1)$, $Y_0 \tilde{} N(0,1)$, $Y_1 = Y_0 + X + X * \varepsilon$. Then $E(Y_1|x) = \beta(x) = x$. Set $c = 0.5$, find median of wealth which equals γ and then

$$\rho = \frac{1}{N} \sum_{i=1}^{N} X_i \times 1(X_i > \gamma).$$

This γ and ρ are the "true" parameter values. Finally, generate D as a random Bernoulli variable and set $Y = DY_1 + (1 - D) Y_0$.

Now we will do the simulations as follows. For each draw from this population, first estimate ρ and γ with the full sample and then split the sample into two equal halves and repeat the estimation in the split-sample way, as described above. In both cases, we vary h_n between $n^{-1/6}$ and $n^{-1/8}$ and ω_n between $n^{-5/24}$ and $n^{-6/24}$. We present the results graphically in Figure A1. From the graphs, it appears that the full-sample estimates of ρ are generally closer to the true value and the bias for $h_n = n^{-1/8}$ is in fact negative. Given these results, we use the full-sample estimation in our application.

9.3 Part 3: Discussion of condition AM

Part A: We first present a very simple model of bednet use, which would imply AM(ii). Suppose a representative household has utility function defined on consumption and bednet adoption $U(c,d)$, $d \in \{0,1\}$ where $U(\cdot,d)$ is continuous, strictly increasing and strictly concave for $d = 0, 1$. Suppose nonsubsidized price of bednet is p in terms of the consumption good and the subsidized bednet costs 0. Let ε denote unobserved heterogeneity in household preference, distributed with c.d.f. F. Ignore other covariates (or condition on them), suppose household budget is X and assume for simplicity that $\varepsilon \perp X$. Then one can show that

$$\begin{aligned}
\beta(x) &= \Pr(\text{use net with subsidy}|X = x) - \Pr(\text{use net without subsidy}|X = x) \\
&= \Pr(U(x,1) - U(x,0) > \varepsilon) - \Pr(U(x-p,1) - U(x,0) > \varepsilon) \\
&= F\{U(x,1) - U(x,0)\} - F\{U(x-p,1) - U(x,0)\}.
\end{aligned}$$

If F is uniform, we get $\beta(x) = U(x,1) - U(x-p,1)$. For fixed $p > 0$, $\beta'(x) = \frac{\partial}{\partial x}U(x,1) - \frac{\partial}{\partial x}U(x-p,1) < 0$ if U is strictly concave in its first argument. If X is continuously distributed on a bounded support, then $\beta(X)$–a strictly monotone continuous function of X–will also be continuously distributed with a bounded support. For other choices of F, we can get piecewise monotone $\beta(\cdot)$. An intuitive interpretation of $\beta' < 0$ is simply that individual demand for bednets is less price-sensitive when individuals are wealthier.

Part B: If we do not assume **AM**, then the following generalizations are appropriate. Define

$$\begin{aligned} F^{-1}_{\beta(X)}(1-c) &= \sup\left\{a : \int_{x\in\mathcal{X}} 1(\beta(x) \geq a)\, dF(x) \leq c\right\}, \\ \delta_1 &= 1\left\{\int_{x\in\mathcal{X}} 1\left(\beta(x) \geq F^{-1}_{\beta(X)}(1-c)\right) dF(x) = c\right\} \\ \gamma &= \max\left\{F^{-1}_{\beta(X)}(1-c), 0\right\} \\ \delta_2 &= 1\left\{F^{-1}_{\beta(X)}(1-c) > 0\right\}, \end{aligned}$$

and write

$$\begin{aligned} \rho &= \delta_1\delta_2 \times \left[E\{\beta_1(X)\} - \int_{x\in\mathcal{X}} [\beta(x)\, 1(\beta(x) < \gamma)]\, dF(x)\right] \\ &\quad + (1-\delta_2) \times E\{\beta_1(X)\} \\ &\quad + (1-\delta_1)\delta_2 \times [E\{\beta_1(X)\, 1\{\beta(X) > \gamma\}\} + E\{\beta_1(X)\, 1\{\beta(X)\} = \gamma\} \times (c - \Pr(\beta(X) \geq \gamma))]. \end{aligned}$$

When $\delta_1 = 1$ and $\delta_2 = 1$, then we have the case discussed in the paper. When $\delta_2 = 0$, everyone with positive treatment effect will be treated and hence the second term in the previous display. When $\delta_2 = 1$ and $\delta_1 = 0$, everyone with CATE above γ is treated but this leaves a surplus equal to $(c - \Pr(\beta(X) > \gamma))$, which is then randomly distributed among those with the "next highest" value of $\beta(X)$, which is the third term. So if $\beta(X)$ has a point mass at a with $F_{\beta(X)}(a-) > 1-c > F_{\beta(X)}(a)$, for some a, then the surplus may then be randomly distributed among the those with $\beta(X) = a$.

As is clear from the expressions, the sample analog of ρ will have discontinuities in its asymptotic distribution, depending on what values the nuisance parameters δ_1 and δ_2 take and methods analogous to Andrews and Guggenberger would be warranted for constructing uniformly valid confidence intervals for ρ. These issues are outside the scope of the present paper and we leave them to future research. condition AM guarantees that $\delta_1 = 1 = \delta_2$. Strengthening AM (i) to $\ln(n) \times \{\Pr(\beta(X) > 0) - c\} > a$ for some $a > 0$ would let us assume away AG type situations.

9.4 Part 4: Asymptotic framework for covariate choice

Suppose $X = (X_1, X_2)$ and we want to decide whether to condition allocation on X_2 in addition to conditioning on X_1. For a fixed $\beta = (\beta_1, \beta_2)$, as $n \to \infty$, the wider model is obviously better. Therefore, to keep the covariate choice problem non-trivial in the asymptotics, we consider a sequence of models P_n with

$$E\left(Y_1 - Y_0 | X = x\right) := x'\beta_n := x_1'\beta_1 + x_2'\delta/\sqrt{n},$$

for some $\delta \in R^k$. Let $\hat{\beta}$ and $\tilde{\beta}_1$ be the MLE of β in the unrestricted (labeled ur) model and of β_1 in the restricted (labeled r) model respectively with $\hat{\gamma}^{ur}$ and $\tilde{\gamma}^r$ the respective estimates of γ. Within this "limit of experiment" framework, Le Cam type convergence theorems can be invoked under appropriate regularity conditions to show that the difference in welfare functions resulting from the larger and smaller set of covariates satisfies

$$\begin{aligned} & \lim_{n\to\infty} \sqrt{n}\left(\rho_n^{ur} - \rho_n^r\right) \\ = & \lim_{n\to\infty} \int \left(\sqrt{n}x_1'\beta_1 + x_2'\delta\right) \left[\Phi\left(\frac{\sqrt{n}x_1'\beta_1 + x_2'\delta - \gamma_n^{ur}}{\sigma^{ur}\left(x\right)}\right) - \Phi\left(\frac{\sqrt{n}x_1'\beta_1 + x_1'S_{11}^{-1}S_{12}\delta - \gamma_n^r}{\sigma^r\left(x\right)}\right)\right] dF\left(x\right) \end{aligned}$$

where $\sigma^r\left(x\right)$ and $\sigma^{ur}\left(x\right)$ represent the asymptotic std deviation of $\hat{\beta}'x - \hat{\gamma}$ and $\tilde{\beta}'x - \tilde{\gamma}$ respectively, $S_{11}^{-1}S_{12}$ represents the projection matrix from regressing X_2 on X_1 and γ_n^r and γ_n^{ur} represent the true thresholds corresponding to the wider and narrower models, respectively. $S_{11}^{-1}S_{12}\delta$ corresponds to the familiar "omitted variable bias" formula.

Since these parameters are unknown, potential feasible rules may be based on:

$$\begin{aligned} dif_{1n} &: = \frac{1}{n}\sum_{i=1}^{n}\left[x_i'\hat{\beta}_{(-i)} \times \left\{1\left\{x_i'\hat{\beta}_{(-i)} \geq \hat{\gamma}_{(-i)}^{ur}\right\} - 1\left(x_{1i}'\tilde{\beta}_{(-i)} \leq \tilde{\gamma}_{(-i)}^r\right)\right\}\right], \text{ or} \\ dif_{2n} &: = \frac{1}{n}\sum_{i=1}^{n}\left[x_i'\hat{\beta}_{(-i)} \times \left\{\bar{L}\left(\frac{x_i'\hat{\beta}_{(-i)} - \hat{\gamma}_{(-i)}^{ur}}{h_n}\right) - \bar{L}\left(\frac{x_{1i}'\tilde{\beta}_{(-i)} - \tilde{\gamma}_{(-i)}^r}{h_n}\right)\right\}\right]. \end{aligned}$$

where the subscript $(-i)$ denotes the leave-one-out estimate. The formal task is to show that under the sequence of models P_n, the sequences $\sqrt{n}\left(dif_n - (\rho_n^{ur} - \rho_n^r)\right)$ converge to an appropriate limiting random variable (see below for how to handle CV based estimates in the asymptotics). The quantiles of this latter random variable can be used to decide on the tolerance threshold c for dif_n above which the wider model will be preferred. Two other issues of interest are those of post-model selection inference, i.e., estimating

$$\rho_n^{ur}\Pr\left(dif_n > \frac{c}{\sqrt{n}}\right) + \rho_n^r\Pr\left(dif_n \leq \frac{c}{\sqrt{n}}\right);$$

and generalizing the above analysis to the case where $E\left(Y_1 - Y_0 | X = x\right)$ is not necessarily linear in x and yet we are considering two linear models to make treatment allocation. In this case, $\hat{\beta}$ and $\tilde{\beta}_1$ can be interpreted as pseudo-MLEs and dif_{1n} replaced by

$$\frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{Y_i D_i}{\hat{\pi}} - \frac{Y_i\left(1-D_i\right)}{1-\hat{\pi}}\right) \times \left(1\left\{x_i'\hat{\beta}_{(i)} \geq \hat{\gamma}_{(i)}^{ur}\right\} - 1\left(x_{1i}'\tilde{\beta}_{(i)} \leq \tilde{\gamma}_{(i)}^{r}\right)\right)\right]$$

where $\hat{\pi} = \frac{1}{n}\sum_{i=1}^{n} D_i$. An appropriate local to zero asymptotic analysis in this case seems to be an interesting task worth pursuing.

Cross-validation (heuristics for the claim in footnote 8): Using the smoothed estimates and the parametric version of theorems 1 and 2 in the main text, it follows that the difference between the naive estimate and the CV estimate of welfare is given by

$$\begin{aligned}\tilde{\rho}_{jn} - \hat{\rho}_{jn} \quad &: \quad = \frac{1}{n}\sum_{i=1}^{n} x_i'\hat{\beta}_{j(-i)}\bar{L}\left\{\frac{x_i'\hat{\beta}_{j(-i)} - \hat{\gamma}_{j(-i)}}{h_n}\right\} - \frac{1}{n}\sum_{i=1}^{n} x_i'\hat{\beta}_j\bar{L}\left\{\frac{x_i'\hat{\beta}_j - \hat{\gamma}_j}{h_n}\right\} \\ &= \quad \frac{1}{n}\sum_{i=1}^{n}\upsilon\left(x_i, h_n, \hat{\beta}_j\right)'\left\{\hat{\beta}_{j(-i)} - \hat{\beta}_j\right\} + O\left(\left\{\hat{\beta}_{j(-i)} - \hat{\beta}_j\right\}^2\right),\end{aligned}$$

for some vector $\upsilon\left(x_i, h_n, \hat{\beta}_j\right)$. Now, using an influence functions approach, it can be shown (c.f., CH page 52) that when $\hat{\beta}$ is the MLE with $-J$ the Hessian and $u\left(\cdot\right)$ the score, then

$$\hat{\beta}_{j(-i)} - \hat{\beta}_j = -n^{-1}\hat{J}^{-1}u\left(z_i\right) + o\left(\frac{1}{n}\right)$$

and the term $\hat{J}^{-1}$ is related to the estimated asymptotic variance of $\hat{\beta}$. Replacing $\hat{\beta}_{j(-i)} - \hat{\beta}_j$ in the previous display, we get the feasible form of the penalization term.