Intercept Estimation in (Non-)Additive Semiparametric Sample Selection Models*

Wiji Arulampalam† Warwick University

Valentina Corradi ‡ Surrey University

Daniel Gutknecht§ Mannheim University

Preliminary and Incomplete
September 15, 2017

Abstract

This paper develops new estimators of the intercept for semiparametric sample selection models in the linear additive, the nonlinear additive, and the multiplicative case. In the linear additive case, we introduce a local polynomial estimator which improves over Andrews and Schafgans (1998) in various directions. First, it achieves the optimal univariate nonparametric rate regardless of the relative tail thickness of the distributions of the error terms and instrument index (the rate can be arbitrarily close to the parametric one by assuming a smoother selection bias term). Second, the estimator does not require independence of the selection error and the regressor(s). Third, it offers a way to determine the bandwidth in a data-driven manner, and, last but not least, applied researchers may implement the procedure using standard software packages. In the additive nonlinear and in the multiplicative model, estimation of the intercept has, to the best or our knowledge, not yet been addressed. We aim at filling this gap, deriving a nonlinear least squares estimator based on an objective function which is adjusted by an estimate of the selection bias term. Depending on properties of the (conditional) distribution of the instrument vector satisfying an index restriction, this estimator converges at a univariate nonparametric rate as it is based on an average of individuals with propensity score close to one. Re-visiting the famous training data of LaLonde (1986) for men and women, we find that applying our method to the non-experimental control group helps to recover the experimental average treatment effect for men, but not for women. This is in line with recent literature which finds that 'selection on unobservables' appears to play a less important role for women on welfare in the U.S..

Key-Words: Intercept Estimation, (Non-)Linear Models, Sample Selection, Boundary, Local Polynomial, Treatment Effect.

*We are grateful to seminar participants at Humboldt University Berlin and the Econometrics Study Group Meeting in Bristol for useful comments and discussions.

†Department of Economics, University of Warwick, Coventry CV4 7AL, UK. Tel. +44(0)2476 523471; email: Wiji.Arulampalam@warwick.ac.uk

‡Department of Economics, University of Surrey, School of Economics, Guildford GU2 7XH, UK, Tel. +44(0)1483 693914; email: V.Corradi@surrey.ac.uk

§Corresponding Author: Department of Economics, University of Mannheim, L7, 3-5, 68131 Mannheim, Germany. Email: Daniel.Gutknecht@gmx.de

JEL Classification: C14, C21, C24.

1 Introduction

This paper introduces novel estimators for the intercept parameter in general semiparametric selection models, covering the linear additive, as well as the nonlinear additive and the multiplicative case. To the best of our knowledge, estimation of the intercept in general (non-)additive semiparametric selection models has not yet been studied.[1]

The intercept in endogenous selection models is of fundamental importance for the evaluation of average treatment effects (Heckman, 1979, 1990): for instance, if outcomes of both selected and non selected individuals are observed, testing the equality of the intercept parameters versus the relevant one-sided alternative, allows to assess whether a specific treatment has the desired effect. This is the case of switching regression, and examples include, among others, testing for the difference in wages of unionized and non-unionized workers, or the ethnic and gender wage gap (Schafgans, 1998, 2000). Moreover, even in instances where only the outcome of the selected is observed, identification and estimation of the intercept may still be of interest: in the classical labor supply model, for instance, where wages and hours worked are only observed for employed individuals, inference about the intercept is crucial to determine the labor supply for the entire population correctly.

While the problem of identification and estimation of the intercept (and the slope coefficients) has long been resolved for the parametric case, it is well known that, when the distribution of the error term of the outcome equation remains unspecified, the intercept cannot, in general, be separately identified from the selection bias term. However, as the probability of selection approaches one, the sample bias term converges towards the unconditional mean of the outcome error, which is typically assumed to be zero. This is an example of an 'identification at infinity' argument applied to the sample selection model (Chamberlain, 1986; Lewbel, 2007). Indeed, Heckman (1990) suggested the first estimator for the intercept in a linear additive model based on the fraction of individuals having probability of being selected close to one. Andrews and Schafgans (1998) proposed a 'smoothed' version of this estimator and established its asymptotic normality. Nevertheless, the rate of convergence depends on the relative tail thickness of the error term and the instrument (index) distributions. When the latter has thicker upper tails, an (almost) parametric rate can be achieved, while equally thick tails imply at most a cubic rate of convergence (i.e., $n^{1/3}$, with n denoting the sample size). Heuristically, this is due to the fact that only individuals for which $\phi\left(z_i'\widehat{\gamma} \geq \xi_n\right) > 0$ matter, where $\phi(\cdot)$ is a smooth weighting function, ξ_n is a threshold value going to infinity, and $z_i'\widehat{\gamma}$ is an estimate of the instrument index $z_i'\gamma_0$ driving selection.[2] Very recently, Goh (2017) suggested a modification of the Andrews and Schafgans (1998) estimator, in which only individuals with $k\left(\left(\widehat{F}\left(z_i'\widehat{\gamma}\right) - 1\right)/h\right) > 0$ matter, where $k(\cdot)$ denotes a kernel function, $\widehat{F}(\cdot)$ an estimator of the marginal distribution of $z_i'\gamma_0$, and h is the bandwidth parameter. The estimator of Goh (2017) is $\sqrt{nh}-$consistent regardless of the relative thickness of tails and, in fact, he does not require explicit assumptions on the right tail of the distribution of $z_i'\gamma_0$. However, a key assumption common to all intercept estimators mentioned above, is that selection is driven by an indicator selection equation $s_i = 1\left\{z_i'\gamma_0 \geq v_i\right\}$, where the error term v_i is fully independent of z_i

[1] Kitagawa (2010) discusses the identification region of the outcome distribution (conditional on covariates) in general sample selection models. However, the focus of his paper is on a test for instrument independence rather than on estimation of the intercept.

[2] In the estimator originally proposed by Heckman (1990), $\phi(\cdot)$ is an indicator function. Consistency and asymptotic normality for the non-smooth Heckman (1990) case have been established by Schafgans and Zinde-Walsh (2002).

and so the probability of selection is determined by the marginal cumulative distribution function of $z_i'\gamma_0$. This is a rather strong assumption which is not required for the semiparametric estimation of binary choice models (see Klein and Spady, 1993) or point identification of sample selection models in general (e.g., Kitagawa, 2010). We therefore replace this independence with a weaker propensity score assumption, only requiring that the probability of being selected $p_i \equiv \mathrm{E}\left[s_i = 1|z_i\right] = \Pr\left(s_i = 1|z_i'\gamma_0\right)$ is a smooth, but not necessarily monotonic function of $z_i'\gamma_0$, and that $\mathrm{E}\left[\varepsilon_i|x_i, z_i, s_i = 1\right] = \lambda(p_i)$, the expectation of the outcome equation error ε_i given covariates x_i, instruments z_i, and selection $(s_i = 1)$ is only a function of the propensity score p_i (see also Das et al., 2003). As p_i approaches one, $\lambda(\cdot)$ converges to the unconditional mean $\mathrm{E}[\varepsilon_i]$, which is assumed to be zero in the additive. Estimating the propensity score p_i, we can therefore obtain a boundary robust estimator of the intercept via local polynomial estimation, using only those observations for which $k\left(\left(\widehat{p}_i - 1\right)/h\right) > 0$. Our estimator is $\sqrt{nh}$ consistent and asymptotically normal, and the rate can be made close to $\sqrt{n}$ by choosing higher order polynomials.[3]

While most of the literature has focused on the additive (linear) case, endogenous selection is not only a problem in these models. Count data for instance, which is typically modeled via multiplicative models, may be subject to non random sampling as well (Terza, 1998): there are several examples from the health economics literature, such as visits to the physician in a given period of (non) insured patients, where self-selection plays an important role. In the context of parametric models, estimation of these multiplicative error models with endogenous selection has been studied by Terza (1998). And only recently, Jochmans (2015) has introduced estimators of the slope parameters for semiparametric nonlinear and multiplicative models with endogenous selection.

The key difficulty in the general additive and multiplicative case is that we can no longer disentangle the intercept from the contribution of the regressors. Hence, even for individuals selected with probability close to one, we can no longer 'isolate' the intercept term. A feasible strategy, however, can be based on a correction of the outcome equation for the selection bias term $\lambda(p_i)$. That is, we show that both in the nonlinear additive as well as in the multiplicative model estimates of $\lambda(p_i)$ can be constructed in a straightforward manner.[4] More specifically, we construct $\lambda(p_i)$ as the difference (ratio) of two nonlinear, unknown functions $m\left(x_i, p_i\right)$ and $m\left(x_i, 1\right)$, which yields the bias term under the normalizations assumptions $\lambda(1) = 0$ for the additive and $\lambda(1) = 1$ for the multiplicative model, respectively. An empirical counterpart can then be constructed on the basis of higher order local polynomial estimators, and θ_0 as well as β_0 can be estimated by nonlinear least squares (NLS).[5] We establish consistency and asymptotic normality of our estimators and show that convergence occurs at the nonparametric univariate rate if $z_i'\gamma_0$ given x_i varies sufficiently strongly, which will be explained in more detail in the main body of the paper. In fact, the limiting distribution is driven by the estimation error due to $\widehat{m}\left(x_i, 1\right) - m\left(x_i, 1\right)$. Heuristically, this occurs because we are averaging only over observations with a propensity score close to one.

As extensions, we show that our estimation procedure can be extended straightforwardly to accommodate endogenous regressors in the outcome equation through a standard control function argument, and also address the issue of adaptive bandwidth choice: because of the boundary issue,

[3] Under the assumption of a symmetric joint distribution of outcome and selection error term conditional on z_i, Chen (1999) developed a $\sqrt{n}-$consistent estimator for the intercept.

[4] A similar 'bias correction' was applied in Gutknecht (2016) in the context of monotonicity testing under endogeneity.

[5] We point out that, due to the inherent boundary nature of our set-up, higher order kernel smoothing techniques typically used to reduce bias are not available in this context.

cross-validation type methods do not work well. This has already been recognized in the regression discontinuity literature, e.g. Imbens and Kalyanaraman (2012), and Calonico et al. (2014)....

The rest of the paper is organized as follows. Section 2 outlines the set-up. Section 3 studies the identification of the intercept in the three different cases.[6] Section 4 outlines the estimation of the propensity score and introduces the estimators for the additive linear, the general additive, and the multiplicative case, respectively. It also establishes the limiting distribution of the all three estimators. Section 5 discusses possibilities to determine the bandwidth in a data driven manner in our setup. Section 5.3 discusses extensions such as the one to endogenous regressors, while Section 6 .... Section 7 concludes. Appendix A contains the main identification results, while Appendix B instead contains a sequence of Lemmas, as well as the proofs of the main theorems.

INSERT ILLUSTRATIONS, 'sufficiently strongly', EVERYTHING CAN BE IMPLEMENTED USING EXISTING; STANDARD STATA ROUTINES, CORRECT SECTION 3.3; ACKNOWLEDGE ALREADY IN THE IDENTIFICATION SECTION, THAT SIMPLER IDENTIFICATION ARGUMETN POSSIBLE; BUT FOR ESTIMATION PURPOSES DO NOT PURSUE IT. ADD EMPIRICAL ILLUSTRATIONS TO ABSTRACT; INTRODUCTION, AND CONCLUSION. NEED TO BRING IN HECKMAN; ICHIMURA; SMITH AND TODD (1998). STARS FOR APPLICATION IN ILLUSTRATION.

The fact that we impose structure allows us to in principle extrapolate from the model (make policy predictions etc. at different x's outside the support, which people who do not have a model cannot do). Also, I found this article by Abadie (2003, JoE) and it seems LATE is really complicated to estimate and standard IV doesn't identify anything useful when there are X you want to condition on.

I think the only case where it really matters is if you have a model without covariates, i.e. sth like $Y = b0 \times D + U$ (where you could use the standard Wald estimator to recover LATE... and I think this only matters for cases like our LaLonde application where we have randomized assignment, but there you also do not need a continuous instrument). So I think we should simply say that because we do not impose independence, our assumptions are not nested within the standard set of LATE assumptions as shown by Vytlacil (2002)... which I think is already half-way outdated since Clement de Chaisemartin has this recent QE paper where he shows that you can identify LATE even with a weaker than monotonicity assumption. statement: Note that this also means that our set of assumptions is not nested within the standard assumption framework of the Local Average Treatment Effect literature (see Vytlacil, 2002).

But definitely no reason to fuzz too much about this.

2 Setup

To see why the intercept is a parameter of interest, consider the following simplified nonlinear model also discussed in Heckman (1990, pp. 314-315): let y_{1i} and y_{0i} denote the wages of union and of non-union workers, respectively, and x_i a vector of observable characteristics.[7] The wages of both

[6] Note that even in the linear additive case, identification does not immediately follow from Andrews and Schafgans (1998), as we do relax their independence, their Assumption 2 (see above).

[7] For simplicity, we assume the observed characteristics x_i to be the same for union and non-union workers.

groups are determined by the following restrictive, but illustrative nonlinear models:

$$y_{1i} = g\left(\theta_{01} + x_i'\beta_0\right) + \varepsilon_{1i}$$

and

$$y_{0i} = g\left(\theta_{00} + x_i'\beta_0\right) + \varepsilon_{0i},$$

where the parameter vector $\{\theta_{00}, \theta_{01}, \beta_0'\}'$ and the real-valued function $g(\cdot)$ are assumed to be known. Then, letting s_i denote the indicator for the union sector (i.e., $s_i = 1$ denotes a union worker and $s_i = 0$ denotes a non-union worker), observed wages can be written as:

$$y_i = s_i y_{1i} + (1 - s_i) y_{0i}.$$

Selection in this type of models typically arises due to the correlation of s_i and $(\varepsilon_{0i}, \varepsilon_{1i})$. That is, while the 'experimental' average treatment effect of being a union worker under random assignment is given by:

$$g\left(\theta_{01} + x_i'\beta_0\right) - g\left(\theta_{00} + x_i'\beta_0\right),$$

where x_i is typically chosen to be the skill endowment of an 'average' union worker, estimation of the model on the two subsamples ignoring the possible endogeneity of selection yields:

$$\begin{aligned} & \mathrm{E}[y_i|x_i, s_i = 1] - \mathrm{E}[y_i|x_i, s_i = 0] \\ = \; & \mathrm{E}[y_{1i}|x_i'\beta_0, s_i = 1] - \mathrm{E}[y_{0i}|x_i'\beta_0, s_i = 0] \\ = \; & g\left(\theta_{01} + x_i'\beta_0\right) - g\left(\theta_{00} + x_i'\beta_0\right) + \mathrm{E}[\varepsilon_{1i}|x_i'\beta_0, s_i = 1] - \mathrm{E}[\varepsilon_{0i}|x_i'\beta_0, s_i = 0]. \end{aligned}$$

This effect in general differs from the above treatment effect unless $\mathrm{E}[\varepsilon_{1i}|x_i'\beta_0, s_i = 1] - \mathrm{E}[\varepsilon_{0i}|x_i'\beta_0, s_i = 0] = 0$ and motivates our point of departure to provide estimators which recover the 'experimental' treatment effect $g\left(\theta_{01} + x_i'\beta_0\right) - g\left(\theta_{00} + x_i'\beta_0\right)$. Note that the above case, where the outcome of both treated and untreated individuals is observed, is an example of a switching regression model. By contrast, in the 'pure' selection model we consider below, it is typically assumed that only the outcome of (selected) individuals with $s_i = 1$ is observable, while covariates x_i are recorded for all individuals.[8] Even in this case, however, identification and estimation of the intercept might be of interest for any kind of counterfactual analysis, or policy evaluation for individuals randomly chosen from the population.

Endogenous sample selection may also arise in other types of data such as count data, which is often represented by a model with multiplicative errors. Consider for instance a standard count data model with $\widetilde{g}(\cdot) = \exp(\cdot)$, and let $\widetilde{y}_{1i}$ and $\widetilde{y}_{0i}$ denote the number of physician contacts in a given month for people with $(s_i = 1)$ and without $(s_i = 0)$ a specific insurance plan.[9] In this case, the experimental treatment effect of being under the insurance plan for randomly assigned individuals is often calculated as the relative change in expected physician contacts:

$$\left(\frac{\exp\left(\theta_{01}\right)}{\exp\left(\theta_{00}\right)} - 1\right) \times 100\%,$$

[8] For instance, wage data, where wages are only recorded for employed workers, but characteristics x_i are observed employed as well as unemployed individuals.

[9] More specifically, the models of interest are characterized by $\mathrm{E}[\widetilde{y}_{1i}|x_i'\beta_0, \widetilde{\varepsilon}_{1i}] = \exp\left(\theta_{01} + x_i'\beta_0\right)\widetilde{\varepsilon}_{1i}$ and $\mathrm{E}[\widetilde{y}_{0i}|x_i'\beta_0, \widetilde{\varepsilon}_{0i}] = \exp\left(\theta_{00} + x_i'\beta_0\right)\widetilde{\varepsilon}_{0i}$, respectively, with $\widetilde{\varepsilon}_{1i}, \widetilde{\varepsilon}_{0i} > 0$ a.s. typically being interpreted as unobserved heterogeneity.

while, under selection into insurance, the 'observable' effect would be given by (if $\mathrm{E}[\widetilde{\varepsilon}_{0i}|x_i'\beta_{00}, s_i = 0] > 0$ a.s.):

$$\left(\frac{\exp(\theta_{01})\,\mathrm{E}[\widetilde{\varepsilon}_{1i}|x_i'\beta_{01}, s_i = 1]}{\exp(\theta_{00})\,\mathrm{E}[\widetilde{\varepsilon}_{0i}|x_i'\beta_{00}, s_i = 0]} - 1\right) \times 100\%.$$

It is again clear that both expressions in general differ except for the special case where $\mathrm{E}[\widetilde{\varepsilon}_{1i}|x_i'\beta_0, s_i = 1] = \mathrm{E}[\widetilde{\varepsilon}_{0i}|x_i'\beta_0, s_i = 0]$. This motivates our interest in developing a unified method to identify and estimate the intercept parameter for nonlinear, semiparametric models with additive and multiplicative error terms.

We now introduce our set-up more formally. Note that, since the mechanics of switching regression and classical sample selection models are the same, and the former can always be viewed as consisting of two separate sample selection models for $s_i = 0$ and $s_i = 1$, we will focus on the sample selection case with $s_i = 1$ in the following (Heckman, 1979). That is, suppose we have a sample of observations where the dependent variable of interest, y_i^*, is only observed when a corresponding selection indicator variable s_i is equal to one, and not otherwise:

$$y_i = y_i^* s_i \quad \text{observed iff} \quad s_i = 1.$$

Moreover, we observe for the entire sample vectors x_i and z_i with $\dim(x_i) \geq 1$ and $\dim(z_i) \geq 1$ of covariates and instruments, respectively. Note that these two vectors may be disjoint, but z_i can also contain some (but not all) elements of x_i. Finally, we assume that the propensity to be selected is a function of the linear index $z_i'\gamma_0$:

$$\mathrm{E}[s_i|z_i] = \Pr\left(s_i = 1|z_i'\gamma_0\right) \equiv p_i.^{10} \tag{1}$$

This index assumption is a standard condition in the literature and is often used to account for discrete elements in z_i. However, unlike in the previous intercept estimation literature (Andrews and Schafgans, 1998), we do not necessarily assume that the propensity score is a monotonic function of $z_i'\gamma_0$, a consequence of the independence assumption as outlined in Section 1. Finally, note that γ_0 along with p_i will be treated as identified and thus known quantities throughout the paper since they can always be recovered by means of separate first stage regressions (e.g., Klein and Spady, 1993; Cai, 2002, see Section 4 for details).

In the following, we will examine three different cases in terms of y_i^*: we will start by considering the linear additive model:

$$y_i^* = \theta_0 + x_i'\beta_0 + \varepsilon_i, \tag{2}$$

which we address separately since there are various features which are different from the analysis of the nonlinear models (see below). Next, we will move to the more general additive nonlinear model:

$$y_i^* = g\left(\theta_0 + x_i'\beta_0\right) + \varepsilon_i, \tag{3}$$

where $g(\cdot)$ will be a known, but nonlinear function. And finally, we will examine the multiplicative model:

$$\mathrm{E}\left[y_i^*|x_i, \widetilde{\varepsilon}_i\right] = \widetilde{g}\left(\theta_0 + x_i'\beta_0\right)\widetilde{\varepsilon}_i, \tag{4}$$

[10] All (in-)equalities involving conditional expectations and/ or random variables in this paper are understood to hold almost surely, even though not explicitly stated.

with $\widetilde{g}(\cdot)$ being again a known, but nonlinear function with properties to be specified below. Note that the last case is different from the additive models as the error term typically plays the role of unobserved heterogeneity here, and by definition has to be positive valued with probability one. In all three cases, our objective is to identify and estimate the intercept θ_0 alongside β_0 in Equation (3) and (4). This is novel as estimation of θ_0 (and β_0) so far has only been studied in the additive linear set-up, under the assumption of independence between z_i and v_i.

3 Identification

3.1 Additive Linear Model

Starting with the additive linear model from Equation (2):

$$y_i^* = \theta_0 + x_i'\beta_0 + \varepsilon_i,$$

we note that similar to γ_0 and p_i above, also β_0 will be treated as a known quantity in this case since this vector can always be identified and estimated at a parametric rate using various estimators such as for instance Ahn and Powell (1993) or Newey (2009). As outlined Section 1, while the error term ε_i is assumed to satisfy the identification condition $\mathrm{E}[\varepsilon_i] = 0$, sample selection in general arises in this model because of a correlation between ε_i and the selection indicator s_i. In other words, the conditional mean (conditional on selection) of ε_i is in general non-zero. Letting E^* henceforth denote the expectation conditional on selection $s_i = 1$, we impose the following regularity condition:

Assumption I1. *It holds that* $E^*[\varepsilon_i|x_i, z_i] = E^*[\varepsilon_i|p_i] \equiv \lambda(p_i)$.

Assumption I1, which is the same as Assumption 2.1(i) in Das et al. (2003), states that, conditional on selection, the mean of ε_i depends only on the value of the propensity score p_i. This is a restriction, which is implied by other familiar conditions, such as independence of disturbances and regressors used for instance in Andrews and Schafgans (1998). From Assumption I1 and the model in Equation (2), we obtain that:

$$\mathrm{E}^*[y_i|x_i, z_i] = \theta_0 + x_i'\beta_0 + \lambda(p_i). \tag{5}$$

Equation (5) above can be used to formulate the auxiliary model:

$$y_i = \theta_0 + x_i'\beta_0 + \lambda(p_i) + u_i,$$

which satisfies $\mathrm{E}^*[u_i|x_i, p_i] = 0$ by construction and will be used to derive our identification result and the corresponding estimator. We impose the following additional regularity condition:

Assumption I2. *The image of the propensity score* p_i, Ψ_p, *is a connected, non-empty subset of the compact interval* $[0, 1]$ *and* $1 \in \Psi_p$. *The (marginal) distribution of* p_i *on* Ψ_p *is absolutely continuous with respect to Lebesgue measure with strictly positive probability density everywhere.*

Assumption I2 is rather standard in the literature using 'identification at infinity' arguments and is an assumption on the strength of the variation of the propensity score p_i. Sufficient conditions for this assumption typically require at least one element of z_i with non-zero coefficient to be continuous (conditional on the other elements), as well as the support of the error term distribution of the selection

equation, say v_i from $s_i = 1\{z_i'\gamma_0 \geq v_i\}$, to be a subset of the support of $z_i'\gamma_0$. Note that $1 \in \Psi_p$ is actually stronger than required and owed to illustration purposes to avoid technical issues such as trimming in the estimation part (see Section 4). In other words, Assumption I2 could be relaxed to accommodate error term distributions with an unbounded support (e.g., the normal distribution) by setting $\Psi_p = (c, 1)$ for some $0 \leq c < 1$.[11] Note also that the above condition on the support of p_i is substantially weaker than the corresponding assumptions in Andrews and Schafgans (1998) who require regularity on the relative behavior of the tail distributions of the unobservables and the marginal distribution of the index restriction $z_i'\gamma_0$ which are hard to verify. This fact, which has already been discussed in the Introduction, will also give rise to a 'simpler' estimation procedure in Section 4, based on a local polynomial estimator. With these two assumptions at hand, we obtain the following identification result:

Theorem I1. *Assume that p_i and β_0 are identified and that $E[\varepsilon_i] = 0$ holds. Then, under Assumptions I1 and I2, θ_0 is uniquely identified.*

3.2 General Additive Model

Next, we consider the more general additive model:

$$y_i^* = g\Big(\theta_0 + x_i'\beta_0\Big) + \varepsilon_i \tag{6}$$

where $g(\cdot)$ is some known, differentiable function. Unlike in the previous case, however, β_0 cannot be recovered in a preliminary step as in the additive linear model and will therefore have to be identified and estimated alongside θ_0.[12] As before, from Assumption I1 and Equation (6), we obtain:

$$\mathrm{E}^*[y_i|x_i, z_i] = g\left(\theta_0 + x_i'\beta_0\right) + \lambda(p_i), \tag{7}$$

which yields the auxiliary model:

$$y_i = g\left(\theta_0 + x_i'\beta_0\right) + \lambda(p_i) + u_i.$$

The error term u_i satisfies again $\mathrm{E}^*[u_i|x_i, p_i] = 0$ by construction. We now impose the following additional regularity conditions:

Assumption I3. *The image of the propensity score p_i, Ψ_p, is a connected, non-empty subset of the compact interval $[0, 1]$ and $1 \in \Psi_p$. The (marginal) distribution of p_i on Ψ_p is absolutely continuous with respect to Lebesgue measure with strictly positive probability density everywhere. Assume that either of the following two conditions holds:*

(i) For all values x in the support of x_i, it holds that $\Psi_p \subseteq \Psi_{p_x}$, where Ψ_{p_x} denotes the image of the propensity score conditional on x.

(ii) There exists (at least) a value x_0 in the interior of the support of x_i such that $\Psi_p \subseteq \Psi_{p_{x_0}}$, where $\Psi_{p_{x_0}}$ denotes the image of the propensity score conditional on x_0.

[11] Thus, if the support of the error term distribution is unbounded, so does the support of $z_i'\gamma_0$ have to be.

[12] Only in special cases where $g(\cdot)$ is separable in θ_0 and $x_i'\beta_0$, β_0 can be identified and estimated at $\sqrt{n}$ rate using recent results of Jochmans (2015). One such example is a count data model discussed in the next section.

Assumption I4. *At least one component of x_i, which has non-zero coefficient, has density with respect to Lebesgue measure that is positive everywhere on its support, conditional on the other components.*

Assumption I5. *Write $x_{1i} \equiv (1, x_i')'$ and $\mathbf{b}_0 \equiv (\theta_0, \beta_0')'$. For all $\mathbf{b} \neq \mathbf{b}_0$, it holds that:*

$$Pr\Bigg(D_{\mathbf{b}}^{\mathbf{1}} g\left(x_{1i}'\mathbf{b}_0\right) s_i\Big(y_i - g(x_{1i}'\mathbf{b}_0) - \lambda(p_i)\Big) \\ \neq E^*\Big[D_{\mathbf{b}}^{\mathbf{1}} g(x_{1i}'\mathbf{b}_0)\Big(y_i - g(x_{1i}'\mathbf{b}_0) - \lambda(p_i)\Big)\Big| x_{1i}'\mathbf{b}, \lambda(p_i)\Big]\Bigg) > 0,$$

where $D_{\mathbf{b}}^{\mathbf{1}} g(\cdot)$ denotes the vector of first order partial derivatives w.r.t. $\mathbf{b}$.

As noted in the context of Assumption I2, Assumption I3 is a condition on the variation of p_i and thus on the strength of the instrument(s). It implies that the upper boundary of the support of the error term distribution of the selection equation is contained in the support of the index $z_i'\gamma_0$. The second part of Assumption I3 (labelled (i) and (ii)) is a so called 'large support' condition: part (i) requires that the index $z_i'\gamma_0$ varies strongly at all points of the support of x_i or, in part (ii), at least at a specific, known value $x_0'\beta_0$. While identification in the nonlinear case in principle only requires conditions on the marginal distribution of p_i similar to Assumption I2 in the additive linear case, this condition is mainly driven by our estimation procedure, which is based on an explicit bias correction of a least squares criterion function using all observations in the sample rather than only the ones with propensity score 'close' to one. This allows for a possibly faster convergence rate than other procedures (see Footnote 16 in Section 4.3 for details). Thus, whether (i) or (ii) is satisfied will not impact the identification result, but will only have consequences for the convergence rate of the estimators proposed in Section 4 below. Assumptions I4 and I5 on the other hand are standard identification assumptions that would provide identification of θ_0 from the score function in more standard settings than ours.

Theorem I2. *Assume that p_i is identified and that $E[\varepsilon_i] = 0$ holds. Then, under Assumptions I1, I3(i) or I3(ii), I4, and I5, θ_0 as well as β_0 are uniquely identified.*

Theorem I2 establishes identification of the parameters θ_0 and β_0. However, as pointed out before, this requires stronger identification conditions on the (conditional) support of p_i.

3.3 General Multiplicative Model

For the multiplicative error case, we start from the model of interest (with intercept) characterized by the conditional expectation:

$$\mathrm{E}[y_i^*|x_i'\beta_0, \widetilde{\varepsilon}_i] = \widetilde{g}\Big(\theta_0 + x_{1i}'\beta_0\Big)\widetilde{\varepsilon}_i,$$

where $\widetilde{\varepsilon}_i > 0$ with probability one, and $\widetilde{g}(\cdot)$ is again a known function.[13] As pointed out before, the error term $\widetilde{\varepsilon}_i$ has a different role from the error term in the additive model and typically denotes unobserved heterogeneity. Note that in the multiplicative case, the error term satisfies the identification assumption $\mathrm{E}[\widetilde{\varepsilon}_i] = 1$. An equivalent condition to Assumption I1 is the following:

[13] As noted in Section 2, the most prominent example for count data models is $\widetilde{g}(\cdot) = \exp(\cdot)$.

Assumption I6. *It holds that* $E^*[\widetilde{\varepsilon}_i|x_i, z_i] = E^*[\widetilde{\varepsilon}_i|p_i] \equiv \widetilde{\lambda}(p_i)$.

Thus similar to the additive case, because of sample selection, our actual outcome equation is:

$$\mathrm{E}^*[y_i|x_i'\beta_0, p_i] = \widetilde{g}\Big(\theta_0 + x_i'\beta_0\Big)\widetilde{\lambda}(p_i),$$

where we assumed again that the mean of $\widetilde{\varepsilon}_i$ depends, conditional on selection, only on the value of the propensity score. As before, we construct the following auxiliary equation:

$$y_i = \widetilde{g}\Big(\theta_0 + x_i'\beta_0\Big)\widetilde{\lambda}(p_i) + \widetilde{u}_i,$$

which satisfies $\mathrm{E}^*[\widetilde{u}_i|x_i, p_i] = 0$ by construction. In addition to Assumption I1 and I3, we require the following slightly modified version of Assumption I5:

Assumption I7. *Using again the definitions of* x_{1i} *and* $\mathbf{b}_0$ *from Assumption I5, for all* $\mathbf{b} \neq \mathbf{b}_0$ *and* $\lambda(p_i) \neq 0$*, it holds that:*

$$Pr\Bigg(D^{\mathbf{1}}_{\mathbf{b}}\widetilde{g}\left(x_{1i}'\mathbf{b}_0\right) s_i\Big(y_i/\widetilde{\lambda}(p_i) - \widetilde{g}(x_{1i}'\mathbf{b}_0)\Big) \neq E^*\Big[D^{\mathbf{1}}_{\mathbf{b}}\widetilde{g}(x_{1i}'\mathbf{b}_0)\Big(y_i/\widetilde{\lambda}(p_i) - \widetilde{g}(x_{1i}'\mathbf{b}_0)\Big)\Big|x_{1i}'\mathbf{b}, \widetilde{\lambda}(p_i)\Big]\Bigg) > 0,$$

where $D^{\mathbf{1}}_{\mathbf{b}}\widetilde{g}(\cdot)$ *denotes the first order derivative of the real valued function* $\widetilde{g}(\cdot)$ *w.r.t.* $\mathbf{b}$.

Theorem I3. *Assume that* p_i *is known (identified) and that* $E[\widetilde{\varepsilon}_i] = 1$ *holds. Also, assume that* $E^*[y_i|x_i'\beta_0, p_i] \neq 0$ *a.s.. Then, under Assumptions I6, I3(i) or I3(ii), I4 and I7,* θ_0 *and* β_0 *are uniquely identified.*

4 Estimation

4.1 Propensity Score Estimation

Before proceeding to the estimation of the intercept, we introduce the estimator of the propensity score, as defined in (1). This is different from Andrews and Schafgans (1998) who, given $s_i = 1\{z_i'\gamma_0 > v_i\}$ with v_i independent of z_i and the assumption on the tails of v_i and $z_i'\gamma_0$, simply need to estimate γ_0. Thus, Assumption E7 below gives high level conditions on the expansion of $\widehat{p}_i - p_i$, and we now provide an estimator of p_i which indeed satisfies them.

As it is customary in the semiparametric selection literature, we start by estimating γ_0 using the Klein and Spady (1993) estimator, say $\widehat{\gamma}$. For the estimation of the intercept, however, it is also crucial to have an accurate estimator of the propensity score when the latter is close to the boundary. That is, recalling that $p_i = \mathrm{E}[s_i = 1|z_i'\gamma_0]$, a boundary issue occurs when $z_i'\gamma_0$ is close or in fact on the boundary of its support, which may arise when the support of $z_i'\gamma_0$ is compact, or when it grows slowly with the sample size. Still, as p_i is not necessarily a monotonic function of $z_i'\gamma_0$, there is no one-to-one mapping between $z_i'\gamma_0$ being on the right boundary and p_i being close to one. Nevertheless,

if it is indeed the case that $s_i = 1\{z_i'\gamma_0 > v_i\}$, with v_i independent of z_i, then p_i is close to one when $z_i'\gamma_0$ is close to its right boundary.

It is well known in the literature that standard local polynomial estimators of conditional distribution functions are boundary robust, but not constrained to lie between zero and one. By contrast, Hall et al. (1999) have introduced a weighted Nadaraya-Watson (WNW) estimator that shares the favorable properties of the local linear polynomial estimator, and yet is constrained to lie between zero and one. The limiting distribution of this estimator was established by Cai (2002). Hereafter, we will also use the WNW estimator to ensure robustness against possible boundary issues. Let

$$\widehat{p}_i = \frac{\frac{1}{nh_1}\sum_{j=1}^n s_j \pi_j\left(z_i'\widehat{\gamma}\right) k\left(\frac{z_j'\widehat{\gamma}-z_i'\widehat{\gamma}}{h_1}\right)}{\frac{1}{nh_1}\sum_{j=1}^n \pi_j\left(z_i'\widehat{\gamma}\right) k\left(\frac{z_j'\widehat{\gamma}-z_i'\widehat{\gamma}}{h_1}\right)}, \tag{8}$$

where

$$\pi_j\left(z_i'\widehat{\gamma}\right) = \frac{1}{n}\left(1+\eta_n\left(z_j'\widehat{\gamma}-z_i'\widehat{\gamma}\right)k\left(\frac{z_j'\widehat{\gamma}-z_i'\widehat{\gamma}}{h_1}\right)\right)^{-1}$$

and

$$\eta_n = \arg\max_{\eta} \ln\left(1+\eta\left(z_j'\widehat{\gamma}-z_i'\widehat{\gamma}\right)k\left(\frac{z_j'\widehat{\gamma}-z_i'\widehat{\gamma}}{h_1}\right)\right).$$

If the data satisfies Assumption E1 below, the kernel $k(\cdot)$ satisfies Assumption E2, and if $(\widehat{\gamma}-\gamma_0)$ satisfies Assumption E5, then the expansion for $(\widehat{p}_i - p_i)$ in Assumption E7 follows from Theorem 1 in Cai (2002). We discuss adaptive choice of h_1 in Section 5.3.

4.2 Additive Linear Model

As outlined in Section 3, in the additive linear case we can estimate the slope parameters β_0 using for instance Li and Wooldridge (2002) or Ahn and Powell (1993) in a way that $\left(\widehat{\beta}-\beta_0\right)$ satisfies the parametric convergence rate condition in Assumption E5 below. Now, recalling that

$$\mathrm{E}^*\left[y_i|x_i'\beta_0, z_i\right] = \theta_0 + x_i'\beta_0 + \lambda(p_i),$$

it follows that

$$y_i s_i = \left(\theta_0 + x_i'\beta_0 + \lambda(p_i) + u_i\right)s_i,$$

where $\mathrm{E}^*\left[u_i|x_i,p_i\right] = 0$ by construction. Since $\lambda(1) = 0$, we can estimate θ_0 by a local polynomial estimator, using $y_i - x_i'\widehat{\beta}$ as dependent variable and only those observations for which $\widehat{p}_i$ is close to one. That is, let $\widehat{b}_0(1)$ be the first element of the vector:

$$\left(\widehat{b}_0(1),\widehat{b}_1(1),...,\widehat{b}_q(1)\right) = \arg\min_{b_0,b_1,...,b_q}\frac{1}{nh_2}\sum_{i=1}^n s_i\left(y_i - x_i'\widehat{\beta}_0 - b_0 - \frac{1}{t!}\sum_{t=1}^q b_t(\widehat{p}_i-1)^t\right)^2 k\left(\frac{\widehat{p}_i-1}{h_2}\right).$$

where $\widehat{p}_i$ is defined in (8), and $k(\cdot)$ is a second order kernel satisfying Assumption E2 below. Letting $[\mathbf{a}]_1$ denote the first element of the vector $\mathbf{a}$, the estimator for the intercept, $\widehat{b}_0(1)$, can be written more compactly as:

$$\widehat{b}_0(1) = \left[\widehat{\mathbf{b}}(1)\right]_1 = \left(\frac{1}{nh_2}\sum_{i=1}^n s_i\widehat{\mathcal{X}}_i(1)\widehat{K}_i(1)\widehat{\mathcal{X}}_i'(1)\right)^{-1}\frac{1}{nh_2}\sum_{i=1}^n s_i\widehat{\mathcal{X}}_i(1)\widehat{K}_i(1)\widehat{\mathcal{Y}}_i, \tag{9}$$

where

$$\widehat{\mathcal{X}}_i(1) = \left(1, (\widehat{p}_i - 1)\frac{1}{h_2}, ..., (\widehat{p}_i - 1)^q \frac{1}{q! h_2^q}\right)'$$

and $\widehat{\mathcal{Y}}_i = y_i - x_i'\widehat{\beta}$, as well as $\widehat{K}_i(1) = k((\widehat{p}_i - 1/h_2))$. The major advantage of local polynomial over local constant estimators is that the bias at the boundary, $p = 1$ in our case, is of the same order as the bias in the interior if the polynomial order is odd (see, e.g. Fan and Gijbels, 1992; Ruppert and Wand, 1994).

If we set $b_j(1)$ for $j = 1, ..., q$ in (9) and replace $k((\widehat{p}_i - 1)/h_2)$ with a smooth function of $z_i'\widehat{\gamma} - \xi_n$ with $\xi_n \to \infty$ as $n \to \infty$, then $\widehat{b}(1)$ collapses to the Andrews and Schafgans (1998) estimator. Also, note that $b_j(1) = 0$ for $j = 1, ..., q$ in Equation (9) are in fact estimates of $\lambda(\cdot)$ and of its derivatives at the boundary $p = 1$.

Before we outline the asymptotic results, we outline the regularity conditions which are required for the additive linear case:

Assumption E1. *The data $\{y_i, x_i', z_i', s_i\}$ are a sequence of i.i.d. observations. Moreover, define $u_i = y_i - E^*[y_i|x_i, p_i]$ and assume that $\sigma^{*2}(1) \equiv E^*[u_i^2|p = 1] < \infty$.*

Assumption E2. *The kernel function $k(\cdot)$ is a bounded, symmetric function around zero, with compact support $[-1, 1]$ and satisfy $\int_{-\infty}^{\infty} k(v)dv = 1$. The functions $h_j = v^j k(v)$, for all j with $0 \leq j \leq q+1$ are Lipschitz continuous. The matrix $\mathbf{M}_1^1$, which contains moments of the function $k(\cdot)$, is defined in Equation (24) of Section B.1 in the Appendix and is non-singular.*

Assumption E3. *The marginal density function $f(p_i)$ is uniformly bounded, and bounded away from zero on its compact support. Moreover, the density function is twice continuously differentiable with uniformly bounded derivative.*

Assumption E4. *The function $\lambda(\cdot)$ is $(q+1)$ times continuously differentiable in its argument. The corresponding $(q+1)$-th order partial derivatives are Lipschitz continuous on their compact support.*

Assumption E5. *The parameter space $\mathcal{B}_0(1)$ is a compact interval and $b_0(1)$ lies in its interior. Also, there exist estimators of β_0 and γ_0 satisfying $\sqrt{n}\left(\widehat{\beta} - \beta_0\right) = O_p(1)$ and $\sqrt{n}(\widehat{\gamma} - \gamma_0) = O_p(1)$, respectively.*

Assumption E6. *Assume that (i) $\max\{h_1, h_2\} \to 0$ and (ii) $\min\{nh_1, nh_2\} \to \infty$. Moreover, it holds that (iii) $nh_2^{2(q+1)+1} \to 0$, as well as (iv) $nh_1^4 h_2 \to 0$ and (v) $nh_1^2 h_2^{-1} \to \infty$.*

Assumption E7. *(i) For the estimated $\widehat{p}_i$, it holds that $\max_{1 \leq i \leq n} |\widehat{p}_i - p_i| = o_p(1)$. (ii) The estimated $\widehat{p}_i$ admits the following representation:*

$$\widehat{p}_i - p_i = \frac{1}{nh_1}\sum_{j=1}^{n} \omega(z_i'\widehat{\gamma}, z_j'\widehat{\gamma})\psi_j + \Xi_n(z_i'\widehat{\gamma}) + o_p\left(\frac{1}{\sqrt{nh_1}} + h_1^2\right),$$

where

$$\omega(z_i'\widehat{\gamma}, z_j'\widehat{\gamma}) \equiv \frac{\pi_j(z_i'\widehat{\gamma})k\left(\frac{z_j'\widehat{\gamma} - z_i'\widehat{\gamma}}{h_1}\right)}{f(z_i'\widehat{\gamma})},$$

with $\pi_j(z_i'\widehat{\gamma})$ denoting a weight function as defined in Cai (2002) and ψ_j satisfies:

$$E[\psi_j|z_j'\gamma_0] = 0 \quad \textit{and} \quad E[\psi_j^2|z_j'\gamma_0] < \infty.$$

We assume that the estimator of $\widehat{\gamma}$ satisfies $\sqrt{n}\,(\widehat{\gamma} - \gamma_0) = O_p(1)$. Moreover, it holds that $\Xi_n(\cdot)$ is continuously differentiable with bounded derivative and $\max_{1\leq i\leq n} |\Xi_n(z_i'\gamma_0)| = O_p(h_1^2)$, $E[|\omega(z_i'\gamma_0, z_j'\gamma_0)|^2] = o(n)$ and:

$$\frac{1}{\sqrt{nh_1}} \sum_{j=1}^{n} \omega(z_i'\gamma_0, z_j'\gamma_0)\psi_j$$

satisfies a CLT. (iii) The image space $\mathcal{P}$ satisfies $Pr(\widehat{p} \in \mathcal{P}) \longrightarrow 1$ and $\int_0^\infty \sqrt{\log N(\lambda, \mathcal{P}, \|\cdot\|_\infty)}d\lambda < \infty$, where $N(\lambda, \mathcal{P}, \|\cdot\|_\infty)$ is the covering number w.r.t. the supremum norm of the class of functions $\mathcal{P}$.

Assumptions E1 through E4 are standard in the local polynomial estimation literature and ensure the existence of an asymptotic distribution. Note that Assumptions E4 and E6 deserve particular attention as they regulate the degree of smoothness of the selection bias term $\lambda(p_i)$, as well as the speed of convergence of h_1 and h_2 to zero. That is, assuming a higher order q in Assumptions E4 and E6 above, will allow a faster rate of convergence of the estimator $\widehat{b}_0(1)$, which can be made close to the parametric rate in principle. We refer the reader to Section 5 for more practical guidance on how to choose the bandwidth sequences h_1 and h_2 in applications. Finally, Assumptions E5 and E7 are standard high-level conditions on the convergence rates of the first step estimators $\widehat{\beta}$ and $\widehat{\gamma}$, as well as on the parameter space and on the estimation error of $\widehat{p}_i$. More specifically, Assumption E5 is typically satisfied when estimators such as the ones proposed by Li and Wooldridge (2002) or Ahn and Powell (1993) are used to recover β_0, and Klein and Spady (1993) for γ_0, respectively. Assumption E7 on the other hand requires $\widehat{p}_i - p_i$ to admit a certain linear expansion, which has been adopted from Theorem 1 in Cai (2002). Using the above regularity conditions, we can derive the asymptotic properties of our estimator.

Theorem E1. *Let Theorem I1 and Assumptions E1 through E7 hold. Then:*

(i)

$$\widehat{b}_0(1) \xrightarrow{p} b_0(1)$$

and

(ii)

$$\sqrt{nh_2}\left(\widehat{b}_0(1) - b_0(1)\right) \xrightarrow{d} N\left(0, \Omega_{0,1}\right),$$

with

$$\Omega_{0,1} = E\left[s_i\sigma^{*2}(1)f(1)^{-1}\right]\left[\mathbf{M}_1^{1-1}\Gamma^1\mathbf{M}_1^{1-1\prime}\right]_{0,0},$$

where $[A]_{0,0}$ denotes the upper left element of matrix A, $\sigma^{*2}(1) = E^*[u_i^2|p=1]$ and the matrices M_1^1 and Γ^1 are defined in Equations (24) and (25), respectively, of Section B.1 in the Appendix.

In the proof of Theorem E1 we show that under the bandwidth conditions stated Assumption E6, both the parametric as well as the nonparametric estimation errors are asymptotically negligible. This is due to the fact that the limiting distribution is driven by:

$$\frac{1}{nh_2}\sum_{i=1}^{n} s_i \mathcal{X}_i(1) K_i\left(1\right) u_i,$$

which converges at a univariate nonparametric rate itself ($\mathcal{X}_i(1)$ and $K_i\left(1\right)$ are the counterparts of $\widehat{\mathcal{X}}_i(1)$ and $\widehat{K}_i\left(1\right)$, respectively, with $\widehat{p}_i$ replaced by p_i). As noted before, choosing a larger polynomial order q in Assumptions E4 and E6 above can make, in principle, the convergence rate arbitrarily close to the parametric rate. For instance, Hall and Racine (2015) jointly select the bandwidth and the polynomial order, letting the latter slowly grow with the sample size at a logarithmic rate. However, it is not straightforward to extend standard asymptotic normality proofs for local polynomial estimators to the case in which the order grows with the sample size.

Finally, we point out that Theorem E1 suggests a straightforward way to construct a valid standard error for $\widehat{b}_0(1)$ on the basis of the asymptotic variance covariance matrix.[14] For instance, in the local linear case, we can construct an estimator of $\sigma^{*2}(1)$ as:

$$\widehat{\sigma}^2(1) = \frac{1}{nh_\sigma}\sum_{i=1}^{n} s_i \left(y_i - \widehat{b}_0(1) - x_i'\widehat{\beta} - \widehat{b}_1(1)\left(\widehat{p}_i - 1\right)\right)^2 k\left(\frac{\widehat{p}_i - 1}{h_\sigma}\right),$$

where $\widehat{b}_0(1)$, $\widehat{b}_1(1)$ are local linear estimators from the previous stage and the bandwidth h_σ is in general chosen larger than h_2. $f(1)$ on the other hand can be replaced with a boundary bias corrected standard kernel estimator:

$$\widehat{f}(1) = \frac{2}{nh_2}\sum_{i=1}^{n} k\left(\frac{\widehat{p}_i - 1}{h_\sigma}\right),$$

where the same bandwidth h_σ has been chosen for notational convenience. Finally, the theoretical moments of the kernel function in $\left[\mathbf{M}_1^{1-1}\Gamma^1\mathbf{M}_1^{1-1\prime}\right]_{0,0}$ can be computed analytically. For instance, if an ordinary second order Epanechnikov kernel is used, the upper left element of this matrix is approximately given by 4.498.[15] Thus, the standard error of $\widehat{b}_0(1)$ in the local linear case could be computed as:

$$\sqrt{\frac{4.498 \cdot \widehat{\sigma}^{*2}(1)}{nh_2\widehat{f}(1)}}.$$

4.3 General Additive and Multiplicative Model

Next, we introduce the estimators for the general additive and multiplicative model from Equations (3) and (4), respectively. We start with the general additive case, which differs conceptually from

[14] In Section 6, we also experiment with different ways to construct the standard error of the intercept estimator using, e.g. a method suggested by Fan and Gijbels (1996, p. 115), but find no significant difference in the results to the method suggested here.

[15] Note that due to the boundary issue all off-diagonal elements are actually non-zero which leads to a scaling constant which is quickly increasing with the polynomial order. For instance, for a third order polynomial the scaling constant already increases to 17.156.

the one in (9) since the nonlinear function $g(\cdot)$, which is non-separable in θ_0 and β_0 in general, prevents us from constructing a local polynomial estimator for observations close to $p = 1$.[16] Instead, following the identification idea of Theorem I2, we construct an estimator for $\mathbf{b}_0 = (\theta_0, \beta_0')'$, which is based on a 'bias corrected' least squares criterion function. More specifically, let $m(x_i, p_i)$ denote $\mathrm{E}^*[y_i|x_i, p_i]$ and so $m(x_i, 1) = \mathrm{E}^*[y_i|x_i, 1]$.[17] Then, as $\lambda(1) = 0$ by Theorem I2, we can compute $\lambda(p_i)$ as $\lambda(p_i) = m(x_i, p_i) - m(x_i, 1)$ in this additive case. For the empirical objective function in (11) below, this quantity can straightforwardly be replaced by suitable local polynomial estimators $\widehat{m}(x_i, \widehat{p}_i)$ and $\widehat{m}(x_i, 1)$ for $m(x_i, p_i)$ and $m(x_i, 1)$, respectively, where $\widehat{p}_i$ is again the estimator of p_i from (8). For the exact form of the local polynomial estimators see Equations (16) and (17) in Section B.1 of the Appendix. That is, we construct our estimator of $\mathbf{b}_0$ as the minimizer of:

$$\widehat{\mathbf{b}}_A = \arg\min_{\mathbf{b}\in\mathcal{B}_0} \frac{1}{n}\sum_{i=1}^{n} s_i \left(y_i - g\left(x_{1i}'\mathbf{b}\right) - \left(\widehat{m}(x_i, \widehat{p}_i) - \widehat{m}(x_i, 1)\right)\right)^2. \tag{11}$$

Note that this estimator is based on the assumption that p_i varies sufficiently with every value x in its support, and that $p = 1$ can be reached for each of these values. In other words, we require that Assumption I3(i) from Section 3 indeed holds. Since this assumption may sometimes be too strong in practice, we can also construct an estimator on the weaker Assumption I3(ii):

$$\widehat{\mathbf{b}}_A^* = \arg\min_{\mathbf{b}\in\mathcal{B}_0} \frac{1}{n}\sum_{i=1}^{n} s_i \left(y_i - g\left(x_{1i}'\mathbf{b}\right) - \left(\widehat{m}(x_0, \widehat{p}_i) - \widehat{m}(x_0, 1)\right)\right)^2. \tag{12}$$

That is, while the estimator of $\widehat{\lambda}(\widehat{p}_i)$ in (11) is constructed as an average over all x_i in the sample, (12) only uses observations in the neighborhood of a specific (known) value x_0. Since the pieces which will drive the limiting behavior of both estimators are given by $\widehat{m}(x_i, 1)$ and $\widehat{m}(x_0, 1)$, respectively, relaxing Assumption I3(i) will obviously also affect its convergence rate, which instead of $\sqrt{nh_2}$ will be of order $\sqrt{nh_2^{d_x+1}}$ with $d_x = \dim(x_i)$ in the latter case (see below).

The estimator for the multiplicative model is conceptually very similar to the one in (11). Similar to the general additive model, except for specific cases where $\widetilde{g}(\cdot)$ is separable in θ_0 and β_0 such as e.g. in count data models with $\widetilde{g}(\cdot) = \exp(\cdot)$ discussed below, the estimator is again based on the idea of bias correcting a least squares criterion function. However, due to the non-additive nature of the model, we now construct $\lambda(p_i)$ as $\lambda(p_i) = m(x_i, p_i)/m(x_i, 1)$ in line with Theorem I3. That is:

$$\widehat{\mathbf{b}}_M = \arg\min_{\mathbf{b}\in\mathcal{B}_0} \frac{1}{n}\sum_{i=1}^{n} s_i \left(y_i \left(\frac{\widehat{m}(x_i, 1)}{\widehat{m}(x_i, \widehat{p}_i)}\right) - \widetilde{g}\left(x_{1i}'\mathbf{b}\right)\right)^2. \tag{13}$$

[16] To understand the differences with the additive linear case, it is in fact instructive to consider an alternative estimator of the intercept (and the slope parameters) given as the minimizer of the following criterion function:

$$\arg\min_{\mathbf{b}\in\mathcal{B}_0} \frac{1}{n}\sum_{i=1}^{n} s_i \left(y_i - g\left(x_{1i}'\mathbf{b}\right)\right)^2 k\left(\frac{\widehat{p}_i - 1}{h_2}\right). \tag{10}$$

While this estimator only uses the marginal distribution of p_i, standard bias calculations show that, when the kernel function $k(\cdot)$ is taken to be a symmetric, non-negative function, the best possible convergence rate is at most a cubic rate (i.e., $n^{\frac{1}{3}}$) due to the boundary issue. By contrast, when $k(\cdot)$ is chosen to be a higher order boundary kernel, Equation (10) may not yield a unique minimum due the required negativity of $k(\cdot)$ for some parts of its support.

[17] If the support of $z_i'\gamma_0$ is unbounded, and z_i and v_i are independent, where v_i is the selection error from $s_i = 1\{z_i'\gamma_0 > v_i\}$, we adopt the convention that $m(x_i, 1) \equiv \lim_{p\to 1} \mathrm{E}[y_i|x_i, p]$.

As before, this estimator is based on the stronger identification Assumption I3(i). An alternative estimator requiring only the weaker Assumption I3(ii) is given by:

$$\widehat{\mathbf{b}}^*_M = \arg\min_{\mathbf{b}\in\mathcal{B}_0} \frac{1}{n}\sum_{i=1}^{n} s_i \left(y_i \left(\frac{\widehat{m}(x_0,1)}{\widehat{m}(x_0,\widehat{p}_i)} \right) - \widetilde{g}\left(x_{1i}'\mathbf{b}\right) \right)^2. \tag{14}$$

It is worth noting that an important application of the multiplicative model in (4) is to count data, where one typically assumes:

$$\mathrm{E}\left[y_i^*|x_i,\widetilde{\varepsilon}_i\right] = \exp\left(\theta_0 + x_i'\beta_0\right)\widetilde{\varepsilon}_i.$$

This model is in fact separable in θ_0 and β_0, and so as in the linear additive model β_0 can be estimated in a preliminary step at a parametric rate (see Jochmans, 2015). Thus, denoting this $\sqrt{n}$-consistent estimator by $\widehat{\beta}_0$, an alternative estimator of θ_0 only can in this case be constructed by minimizing:

$$\widehat{\theta}_{0,M} = \arg\min_{\theta\in\Theta_0} \frac{1}{n}\sum_{i=1}^{n} s_i \left(y_i \left(\frac{\widehat{m}(x_i'\widehat{\beta},1)}{\widehat{m}(x_i'\widehat{\beta},\widehat{p}_i)} \right) - \widetilde{g}\left(\theta + x_i'\widehat{\beta}\right) \right)^2$$

instead of (13) above. This estimator also converges at rate $\sqrt{nh_2}$ under Assumption I3(i), but has the advantage that the slope parameter vector β_0 can be estimated at a parametric rate in a separate first step, and that even under the weaker Assumption I3(ii) the convergence rate of $\widehat{\theta}_{0,M}$ is still at most $\sqrt{nh_2^2}$. Since count data models represent indeed a leading example of nonlinear models, this adaptability is another appealing feature of the estimation procedure outlined above.

For the general additive and multiplicative case, let the support of x_i be denoted by $\mathcal{X}$, and the image space of p_i be $\mathcal{P}$. For simplicity, we will assume that all elements of x_i are continuous, the extension to some (but not all) discrete elements is straightforward. The parameter space of $\mathbf{b}_0$ is given by $\mathcal{B}_0$. The following regularity conditions are required:

Assumption E8. *The data $\{y_i, x_i', z_i', s_i\}$ are a sequence of i.i.d. observations. Moreover, define $u_i = y_i - E^*[y_i|x_i,p_i]$ and $\sigma^{*2}(x_i,p_i) \equiv E^*[u_i^2|x_i,p_i]$ for the additive model and $\widetilde{u}_i = y_i - E^*[y_i|x_i,p_i]$ and $\widetilde{\sigma}^{*2}(x_i,p_i) \equiv E^*[\widetilde{u}_i^2|x_i,p_i]$ for the multiplicative model. Assume that:*

$$\sup_{(x,p)\in\mathcal{X}\times\mathcal{P}} E\left[s\sigma^{*2}(x,p)\right] < \infty \quad \textit{and} \quad \sup_{(x,p)\in\mathcal{X}\times\mathcal{P}} E\left[s\widetilde{\sigma}^{*2}(x,p)\right] < \infty$$

and that

$$\sup_{\overline{\mathbf{b}}\in\mathcal{B}_0} E\left[s_i D_{\mathbf{b}}^{\mathbf{1}} g\left(x_{1i}'\overline{\mathbf{b}}\right)^2 \right] < \infty \quad \textit{as well as} \quad \sup_{\overline{\mathbf{b}}\in\mathcal{B}_0} E\left[s_i D_{\mathbf{b}}^{\mathbf{1}} \widetilde{g}\left(x_{1i}'\overline{\mathbf{b}}\right)^2 \right] < \infty$$

holds.

Assumption E9. *The kernel function $k(\cdot)$ is a bounded, symmetric function around zero, with compact support $[-1,1]$ and satisfy $\int_{-\infty}^{\infty} k(v)dv = 1$. The functions $H_{\mathbf{j}} = \mathbf{v}^{\mathbf{j}}K(\mathbf{v})$, $K(\mathbf{v}) = k(v_1)\times k(v_2)$, for all $\mathbf{j}$ with $0 \le |\mathbf{j}| \le 2q+2$ are Lipschitz continuous. The matrices $\mathbf{M}$ and $\mathbf{M}_1$, multivariate moments of the function $K(\cdot)$, are defined in Equations (18) and (19) of Section B.1 in the Appendix and are non-singular.*

Assumption E10. *The joint density function $f(x_i,p_i)$ is uniformly bounded, and bounded away from zero on its compact support. Moreover, the density function is continuously differentiable in its second argument with uniformly bounded derivative.*

Assumption E11. *The function $m(\cdot,\cdot)$ is $(q+1)$ times continuously differentiable in both its arguments and any mixture of them, where $q>1$. The corresponding $(q+1)$-th order partial derivatives are Lipschitz continuous on their compact support.*

Assumption E12. *The parameter space $\mathcal{B}_0$ is compact and $\mathbf{b}_0$ lies in its interior. Also, there exists an estimator of γ_0 satisfying $\sqrt{n}\left(\widehat{\gamma}-\gamma_0\right)=O_p\left(1\right)$.*

Assumption E13. *Assume that (i) $\max\{h_1,h_2\}\longrightarrow 0$ as well as (ii) $nh_1\longrightarrow\infty$. Moreover, it holds that (iii) $nh_2^{d_x+\frac{3}{2}}\longrightarrow\infty$ as $n\longrightarrow\infty$ as well as (iv) $nh_2^{2(q+1)+1}\longrightarrow 0$, (v) $nh_2^{d_x+2}h_1\longrightarrow\infty$, and (vi) $nh_1^4h_2^{2q+1}\longrightarrow 0$.*

Assumptions E8 through E10 are straightforward extensions of the regularity conditions from the previous section, now also accounting for the fact that the slope coefficients are estimated, too. More specifically, Assumption E8 imposes second moments on the auxiliary model error terms u_i and $\widetilde{u}_i$, respectively, as well as on the first order derivatives of $g\left(\cdot\right)$ and $\widetilde{g}\left(\cdot\right)$. The conditions on the kernel function in Assumption E9 are standard regularity conditions satisfied by most commonly chosen second order kernel functions. Assumption E10 is a condition on the density and the support of the underlying conditioning variables x_i and p_i. Both parts are rather standard in the context of local polynomial estimators and could be relaxed at the cost of more complicated proofs. For instance, the compact support condition could be relaxed by introducing a random trimming function which admits a support growing with the sample size n. Assumption E11 ensures Taylor-series expansions to appropriate orders, which is common in inference problems of this type. By contrast, Assumption E12 is again an assumption on the parameter space of $\mathbf{b}_0$ and on the first stage estimator $\widehat{\gamma}$. The last condition regulates the speed of the bandwidth for a given polynomial order q. We again refer the reader to Section 5 for a discussion on how to choose h_1 and h_2 in practical applications. The next two theorems establish the asymptotic distribution of the estimators for the general additive and the general multiplicative model, respectively.

Theorem E2. *Grant Theorem I2, and in particular Assumption I3(i) therein, and let Assumptions E7 through E13 hold. Then*

(i)

$$\widehat{\mathbf{b}}_A\xrightarrow{p}\mathbf{b}_0$$

and

(ii)

$$\sqrt{nh_2}\left(\widehat{\mathbf{b}}_A-\mathbf{b}_0\right)\xrightarrow{d}N\left(0,\mathbf{V}_0\right),$$

where $\mathbf{V}_0=\boldsymbol{\Sigma}_0^{-1}\Omega_0\boldsymbol{\Sigma}_0^{-1}$ with $\boldsymbol{\Sigma}_0=E\left[s_iD_{\mathbf{b}}^{\mathbf{1}}g\left(x_{1i}'\mathbf{b}_0\right)^2\right]$ and

$$\Omega_0=E\left[s_i\sigma^{*2}(x_i,1)f(x_i,1)^{-1}\right]\left[\mathbf{M}_1^{-1}\Gamma\mathbf{M}_1^{-1\prime}\right]_{0,0}$$

with $[A]_{0,0}$ the upper left element of matrix A, $\sigma^{*2}(x_i,1)=E^*[u_i^2|x_i,1]$, and the matrices M_1 and Γ are defined in Equations (19) and (20), respectively, of Section B.1 in the Appendix.

Remark 1. *If we replace Assumption I3(i) by Assumption I3(ii), we obtain the following asymptotic result instead:*

$$\sqrt{nh_2^{d_x+1}}\left(\widehat{\mathbf{b}}_A - \mathbf{b}_0\right) \xrightarrow{d} N\left(0, \mathbf{V}_{00}\right),$$

where $\mathbf{V}_{00} = \mathbf{\Sigma}_0^{-1}\Omega_{00}\mathbf{\Sigma}_0^{-1}$ *with* $\mathbf{\Sigma}_0$ *as defined before and*

$$\Omega_{00} = E\left[s_i\sigma^{*2}(x_0,1)f(x_0,1)^{-1}\right]\left[\mathbf{M}_1^{-1}\Gamma_0\mathbf{M}_1^{-1\prime}\right]_{0,0}$$

with $\sigma^{*2}(x_0,1) = E^*[u_i^2|x_0,1]$ *and* Γ_0 *defined in Section B.1 of the Appendix.*

Theorem E3. *Grant Theorem I3, and in particular Assumption I3(i) therein, and let Assumptions E7 through E13 hold. Then*

(i)

$$\widehat{\mathbf{b}}_M \xrightarrow{p} \mathbf{b}_0$$

and

(ii)

$$\sqrt{nh_2}\left(\widehat{\mathbf{b}}_M - \mathbf{b}_0\right) \xrightarrow{d} N\left(0, \widetilde{\mathbf{V}}_0\right),$$

where $\widetilde{\mathbf{V}}_0 = \widetilde{\mathbf{\Sigma}}_0^{-1}\widetilde{\Omega}_0\widetilde{\mathbf{\Sigma}}_0^{-1}$ *with* $\widetilde{\mathbf{\Sigma}}_0 = E\left[s_i D_{\mathbf{b}}^{\mathbf{1}}\widetilde{g}\left(x_{1i}'\mathbf{b}_0\right)^2\right]$ *and*

$$\widetilde{\Omega}_0 = E\left[s_i\widetilde{\sigma}^{*2}(x_i,1)f(x_i,1)^{-1}\right]\left[\mathbf{M}_1^{-1}\Gamma\mathbf{M}_1^{-1\prime}\right]_{0,0}$$

with $[A]_{0,0}$ *the upper left element of matrix A,* $\widetilde{\sigma}^{*2}(x_i,1) = E^*[\widetilde{u}_i^2|x_i,1]$*, and the matrices* M_1 *and* Γ *are defined in Equations (19) and (20), respectively, of Section B.1 in the Appendix..*

Remark 2. *If we replace Assumption I3(i) by Assumption I3(ii), we obtain the following asymptotic result instead:*

$$\sqrt{nh_2^{d_x+1}}\left(\widehat{\mathbf{b}}_M - \mathbf{b}_0\right) \xrightarrow{d} N\left(0, \widetilde{\mathbf{V}}_{00}\right),$$

where $\widetilde{\mathbf{V}}_{00} = \widetilde{\mathbf{\Sigma}}_0^{-1}\widetilde{\Omega}_{00}\widetilde{\mathbf{\Sigma}}_0^{-1}$ *with* $\widetilde{\mathbf{\Sigma}}_0$ *defined as before and*

$$\widetilde{\Omega}_{00} = E\left[s_i\widetilde{\sigma}^{*2}(x_0,1)f(x_0,1)^{-1}\right]\left[\mathbf{M}_1^{-1}\Gamma_0\mathbf{M}_1^{-1\prime}\right]_{0,0}$$

with $\widetilde{\sigma}^{*2}(x_0,1) = E^*[\widetilde{u}_i^2|x_0,1]$ *and* Γ_0 *defined in Section B.1 of the Appendix.*

As in the additive linear case, note that standard errors can be constructed on the basis of the asymptotic variance covariance matrices Ω_0 and $\widetilde{\Omega}_0$, respectively. Moreover, as pointed out in the Remarks 1 and 2, observe that the rate of convergence depends on the identification Assumption I3, and can range from a univariate to a multivariate nonparametric rate.

5 Extensions

An important question in the semi- and nonparametric estimation literature concerns the choice of the tuning parameters. In this section we provide, without claiming any sort of optimality of this procedure, a practical guidance on how to choose the bandwidths in the additive linear as well as in the general (non-)additive case. The second part of this section on the other hand addresses the issue of endogenous regressors in the outcome equation.

5.1 Additive Linear Model

For the additive linear case we require a selection of the bandwidth for the estimation of the propensity score, h_1, as well as of the bandwidth for the estimation of the intercept, h_2, subject to the rate conditions in Assumption E6. It is immediate to see that if we set $h_1 = c_1 n^{-\frac{1}{2}+\epsilon_1}$ and $h_2 = c_2 n^{-\frac{1}{2(q+1)+1}-\epsilon_2}$ with $0 < \epsilon_1 < 1/2$ and $0 < \epsilon_2 < \frac{1}{2} - \frac{1}{2(q+1)+1}$, respectively, then all the rate conditions of Theorem E1 are satisfied.

Since theory does however not provide any guidance on the choice of the constants c_1 and c_2, we propose a more data driven manner to select the bandwidth in the following: starting with h_1, we first note that this bandwidth does not directly affect the rate of convergence in Theorem E1, and thus it suffices to choose a properly 'under-smoothed' bandwidth. In the empirical section, we will therefore set $\epsilon_1 = \frac{1}{4}$ and experiment with different values for c_1, e.g. the average or the standard deviation of s_i.[18]

Regarding the choice of h_2, we point out that the problem of a data driven choice for this bandwidth is in principle akin to the problem of adaptive bandwidth selection in a regression discontinuity design, which has received considerable attention over recent years (see Gelman and Imbens, 2014; Calonico et al., 2014). However, while the aforementioned papers propose methods to mitigate this problem through bias correction, we suggest a different route in this paper which does neither require an arbitrary choice of the constant c_2 nor the same form of bias correction as in these papers. That is, the idea we outline in this paper is to perform data driven 'under-smoothing': since the order of the bias for local polynomial estimators of odd degree is the same in the interior and at the boundary of the support (Fan and Gijbels, 1996), we generally recommend the use of a polynomial of odd degree.[19] Our approach is therefore suitable for the case of local polynomial estimation of degree larger one when the degree is odd, but can of course also be applied to polynomials of even order. That is, suppose $q = q^* \geq 2$, where q^* is the chosen polynomial order. In the first stage we then compute the cross-validated bandwidth associated with local polynomial regression with $\underline{q} = q^* - 2$ if q^* is odd.[20] The cross-validated bandwidth $h_{CV,\underline{q}}$ is such that $h_{CV,\underline{q}} \simeq n^{-\frac{1}{2(\underline{q}+1)+1}}$. Subsequently, we estimate the intercept $b_0(1)$ via local polynomial of order q^*, using the bandwidth $h_{CV,\underline{q}}$, so that the resulting bias is of order

$$h_{CV,\underline{q}}^{q^*+1} \simeq n^{-\frac{q^*+1}{2(\underline{q}+1)+1}}.$$

[18] Unfortunately, the procedure suggested by Hall et al. (1999) for kernel estimators of conditional distribution functions is not applicable in our context as the dependent variable s_i is binary.

[19] For even ordered polynomials, the bias at the boundary is of order $O(h^{q+1})$, while it is of order $O(h^{q+2})$ in the interior of the support. Therefore, even ordered polynomials may suffer from more severe boundary bias, see e.g. the local constant estimator (Fan and Gijbels, 1996).

[20] If q is even, then local polynomial of order q or $q-1$ give raise to the same order of bias at the boundary.

It is immediate to see that $nh_{CV,\underline{q}}^{2(q^*+1)+1} \simeq nn^{-\frac{2(q^*+1)+1}{2(\underline{q}+1)+1}} \to 0$ as $q^* > \underline{q}$. Although we can not claim any sort of optimality, the suggested approach has two crucial advantages as it neither requires an arbitrary choice of the 'constant' nor direct bias correction.

5.2 General Additive and Multiplicative Model

For the nonlinear case, the first step consists again of a nonparametric estimator of the propensity score. That is, h_1 can be chosen as above, setting e.g., $h_1 = c_1 n^{-\frac{1}{4}}$ with $0 < c_1 < 1$ and c_1 depending on first or second moment of s_i. Provided $q > 1$, we can set $h_2 = c_2 n^{-\frac{1}{2q+3/2}}$. It is immediate to see that all rate conditions in Assumption E13 are satisfied. Now, focusing on the general additive case for illustrative purposes, a possibility to select c_2 is to choose it as the minimizer of the following criterion

$$\arg \min_{h_2 \in \left\{ c_L n^{-\frac{1}{2q+3/2}}, c_U n^{-\frac{1}{2q+3/2}} \right\}} \frac{1}{n} \sum_{i=1}^{n} s_i \left(y_i - x_i \widehat{\beta}_n - \widehat{m}_{-i}(x_i, \widehat{p}_i) - \widehat{m}_{-i}(x_i, 1) \right)^2,$$

where $\widehat{m}_{-i}(x_i, \widehat{p}_i) - \widehat{m}_{-i}(x_i, 1)$ are the leave-one-out counterparts of $\widehat{m}(x_i, \widehat{p}_i) - \widehat{m}(x_i, 1)$. The multiplicative case follows by an analogous reasoning.

5.3 Endogenous Regressors

In many applications, it may be important to correct for sample selection and regressor endogeneity simultaneously (cf. Das et al., 2003). In this subsection, we show that identification can also be achieved when one of the regressors is endogenous. For the sake of simplicity, we will only consider the case of the additive model:

$$y_i^* = \theta_0 + x_i' \beta_0 + \varepsilon_i, \tag{15}$$

which satisfies again the normalization assumption $\mathrm{E}[\varepsilon_i] = 0$. As in Das et al. (2003), we assume that at least one element of x_i is endogenous, say $x_{i,2}$ with $x_i = (x_{i,1}', x_{i,2}')'$.[21] Then, assume that the following reduced form equation holds:

$$x_{i,2} = \mu(x_{i,1}, z_{i,2}) + v_i.$$

Also, with slight abuse of notation, denote z_i, the instrument vector from the selection equation, as $z_{i,1}$ and let $z_i = (z_{i,1}', z_{i,2}')'$.

Assumption I8. *Assume that $E[v_i | x_{i,1}, z_{i,2}] = 0$. Moreover, assume that $E^*[\varepsilon_i | z_i, v_i] = \lambda(p_i, v_i)$.*

It is clear that Assumption I8 identifies $\mu(x_{i,1}, z_{i,2})$ and thus in turn v_i. We can therefore obtain the following identification result

Theorem I4. *Assume that p_i and v_i are known (identified) and that Assumptions I1, I2, and I8 hold. Then, $\lambda(\cdot, \cdot)$ and in particular θ_0 is uniquely identified.*

[21] For the sake of simplicity, we do not elaborate on their extension that x_i^* is only observed when $s_i = 1$. This partial observability of x_i^* is motivated by their workhorse example of endogenous wages which are only observed when people work (the dependent variable is hours worked). However, paralleling their identification arguments, this extension is immediate.

6 Illustrations

In this section, we re-analyze two famous data sets to provide an illustration of the estimator for the linear additive model. The first one is female labor supply data that were originally used by Mroz (1987) (and later by Newey et al., 1990; Ahn and Powell, 1993). The second illustration is from LaLonde (1986) and was subsequently used by many other authors, e.g. Dehejia and Wahba (1999, 2002); Smith and Todd (2005). Mroz (1987) was concerned with the sensitivity of results to various parametric assumptions in an empirical model of married women's labor supply (hours of work) using observational data. In contrast, the focus of LaLonde (1986) was whether the experimental effect of a training program on subsequent earnings could be replicated by substituting the experimental control group with a 'comparison' group drawn from observational data.

6.1 Revisiting Mroz (1987)

After extensive investigations, Mroz concluded that the standard Tobit model exaggerated both income and wage effects in a labor supply model of married women. In addition, he also found that the bias due to an exogeneity assumption of wage rate and experience diminished substantially when self-selection into the labor force was accounted for using a more generalized Tobit specification. Thus, since employment self-selection is known to be an important factor in married women's labor supply, we assess the performance of our estimator to recover the intercept from an additive linear labor supply model. Recovering the constant from such a model is important as it allows to make out of sample predictions of women's labor supply. We compare our results to the estimates of various parametric models used by Mroz (1987) and Newey et al. (1990).

The data for the analysis are drawn from the 1975 Panel Survey of Income Dynamics (PSID) labor supply data. It contains characteristics on 753 married women, of whom 428 were working at the time of the survey. The dependent variable y_i is the annual hours of work, while the regressors x_i include the log (hourly) wage rate (assumed to be endogenous), family income (excluding the woman's income) in \$1,000s, age and education in years, and indicator variables for children under age six and above. The data set also contains additional variables that are assumed to be exogenous and possibly valid instruments: years of education of the mother and the father of the woman (separately), and an indicator for living in the city and whether the woman was unemployed during the year.

The selection of variables for the models we estimate are based on the specifications provided in Table X of Mroz (1987) and also Table 1 of Newey et al. (1990). We follow these papers and use eighteen variables for a model, where selection into the labor market is modeled parametrically (probit), but the distribution of the selection bias term is left unspecified and estimated using our method, and fewer variables for a model which is entirely semiparametric (cf. Newey et al., 1990; Ahn and Powell, 1993).[22] The reason for contrasting a semiparametric model with a 'hybrid' probit

[22]The eighteen variables are experience, experience squared, age (in years), age-squared, age-cubed, education (in years), education-squared, education-cubed, interactions between age, education, education-squared, and age-squared (excluding the interaction between education-squared and age-squared), an urban indicator, the local unemployment rate, mother's education, father's education, indicator variables for young kids (below age six) and older kids (above age six), and family income excluding woman's income. By contrast, the semiparametric model only uses ten variables, namely experience, age, education, an urban indicator, the local unemployment rate, mother's education, father's education, and again indicator variables for kids below and above six, as well as family income.

specification is that Newey et al. (1990) found results of the probit model to be similar to the ones of the semiparametric propensity score model.

Our estimator proceeds as described in Section 4: first, we estimate the propensity to be in the labor force, say $\widehat{p}_i$, using a standard local constant estimator with Epanechnikov kernel.[23] Second, we estimate the slope parameters β_0 using the projection method suggested by Li and Wooldridge (2002) with a cross-validated bandwidth and a second order Epanechnikov kernel. Finally, we fit a local polynomial regression of $\left(y_i - x_i'\widehat{\beta}_0\right)$ on $\widehat{p}_i$ to recover the intercept using again a cross-validated bandwidth. Note that the last two steps only use the selected sample.

The results from the various estimations are provided in Table 1 below. The first four columns report results from standard estimation methods, while columns five and six provide the results for our method using a probit (column five) and a semiparametric (column six) propensity score specification. Note that the results in columns five and six are only for the case where log wage is treated as an exogenous variable since output using predicted log wages was almost identical (w.r.t. the intercept estimate). Turning to the results, we note that accounting for self-selection, log wage becomes insignificant (cf. column two vs. columns three to six). Moreover, comparing the standard Tobit in column three (treating log wage as exogenous) with the specification in column five, where we use a probit propensity score specification, we find that these estimates are broadly similar, also for the intercept. However, unlike Newey et al. (1990), we do find model estimates to be sensitive to the specification of the propensity score as the estimate of the intercept changes from \$ 2439.6 (105.1) for the probit to \$ 2246.9 (121.1) for the semiparametric model. (cf. columns five and six). This is as the intercept differs by more that one s.e. This is in contrast to the choice of the polynomial order, which does not really affect the order of the results for the semiparametric specification, but does affect the probit specification. This may be another evidence for misspceficiation of the normal distribution in our case.[24] *** TO BE COMPLETED*** We confirm the finding of Newey et al. that the estimates are sensitive to the choice of the propensity score.

Choice of polynomial order does not seem to matter much when a flexible propensity score is used (more sensitive to this choice when probit scores are used) This hints at misspecification.

Next, we turn to the second illustration.

6.2 Revisiting LaLonde (1986)

To illustrate our method, we re-analyze the famous data set from LaLonde (1986) which contains both experimental as well as non-experimental data. The experimental data are from the National Supported Work (NSW) Demonstration, a training program that was implemented in the mid-1970s to provide work experience to certain groups of individuals who were disadvantaged in the labor market.[25] The individuals were randomly selected into the program between March 1975 and July 1977. This experimental data provided information on both the treated as well as the control group. The non-experimental data consisted of control group individuals only and stemmed from two surveys, the Panel Study of Income Dynamics (PSID) and the Current Population Survey (CPS).

[23] The bandwidth is chosen to be $\text{sd}(\widehat{p}_i) \cdot n^{-\frac{1}{4}}$. We also tried different scaling constants, e.g. $\text{mean}(\widehat{p}_i)$, without any significant changes to our results.

[24] Note the increase in standard error when choosing a higher order polynomial due to an increased scaling constant.

[25] The targeted groups were drug addicts, ex-offenders, or school dropouts among men, and women on welfare (i.e., women on receipt of 'Aid to Families with Dependent Children' (AFDC) welfare payments).

Table 1: Married Women's Labour Supply

Variables	OLS[1]	IV[2]	H2S1[3]	H2S2[4]	LP1[5]	LP2[6]
Log wage	−17.4	672.3	−71.8	121.4	−63.1	−43.1
	(54.2)	(210.2)	(51.1)	(210.5)	(52.7)	(53.1)
Non-wage family income	−4.3	−6.5	5.4	3.8	4.91	−0.4
	(3.7)	(4.3)	(4.3)	(4.2)	(3.9)	(3.6)
Indicator: kids aged < 6	−342.5	−284.4	83.2	54.1	92.7	32.8
	(100.0)	(117.9)	(128.6)	(123.7)	(121.1)	(118.3)
Indicator: kids aged ≥ 6	−115	−85.2	−92.9	−87.2	−91.3	−74.2
	(30.8)	(37.0)	(34.2)	(30.7)	(29.9)	(30.5)
Age (in years)	−7.7	−9.1	5.1	3.4	5.6	2.9
	(5.5)	(6.5)	(6.6)	(6.0)	(5.8)	(5.6)
Education (in years)	−14.5	−86.4	−72.8	−86.3	−68	−41.5
	(18.0)	(29.6)	(22.7)	(24.5)	(19.5)	(17.9)
Intercept (1st order)	2114.7	2254.9	2551.4	2545.2	2439.6	2246.9
	(340.1)	(399.0)	(389.6)	(337.0)	(105.1)	(121.1)
Intercept (3rd order)					2665.2	2265.3
					(200.3)	(236.7)

[1] OLS [Mroz Table IV row 1]

[2] IV regression treating log wage as endogenous [Mroz Table IV row 4]

[3] Heckman sample selection correction for labour-force participation using the full set of 18 instruments. Log wage is treated as exogeneous

[4] Heckman sample selection correction for labour-force participation using the full set of 18 instruments. Log wage is treated as endogenous and the equation estimated using IV [Mroz Table X row 3; Newey et al Table 1 col. 2; Ahn & Powell Table 1 col. 2]

[5] Local Polynomial: propensity score generated from a probit. Log wage is treated as exogenous

[6] Local Polynomial: propensity score generated from Klein and Spady estimator. Log wage is treated as exogenous

The main focus of LaLonde (1986) was whether the experimental estimates of the effect of this program on post-treatment earnings could be replicated using different econometric methods and various comparison groups drawn from the survey data. He concluded that it was not possible to replicate the experimental results. This sparked a discussion in the literature whether controlling for selection on observables through matching methods would lead to a different conclusion (see, for instance, Dehejia and Wahba (1999, 2002) and Smith and Todd (2005) for a re-analysis of the male data and, more recently, Calonico and Smith (2017) for an assessment of the women's data).

We re-visit the analysis of LaLonde (1986) using both the samples of men as well as of women. However, unlike previous literature, our focus is on 'selection on unobservables' in the non-experimental control group. Our findings suggest that part of the discrepancy between experimental and non-

experimental estimators for men may be due to this overlooked feature, while we cannot confirm such evidence for women. The latter result is in line with recent findings of Calonico and Smith (2017) who argue that time (in-)variant 'selection on unobservables' may not play an important role for the women of the NSW.

For the application, we use the experimental treatment groups also used by Smith and Todd (2005) and Calonico and Smith (2017) for the men and women, respectively, which restrict the original experimental treatment groups to those individuals who had two years of non-missing pre-program earnings and who had been randomized into the program at an early stage.[26] For the comparative analysis, we again use one particular non-experimental control group drawn from the PSID by Smith and Todd (2005) and by Calonico and Smith (2017).

In summary, we study whether using these control groups from the PSID in place of the experimental samples allows to replicate the experimental treatment effect. Our focus is on the estimation of the effect of training on subsequent earnings using the experimental estimates as the bench-mark figure, and comparison of this to the estimates obtained when the experimental control groups are substituted with the non-experimental control groups. However, as pointed out before, we pay special attention to potential biases caused by 'selection on unobservables'. Formally, let $y_{t,i}$ denote earnings in the periods before ($t = 1975$) and after treatment ($t = 1978$ for men and $t = 1979$ for women). Since Smith and Todd (2005) concluded that difference-in-differences based estimators performed the best as they are most adapt to eliminating certain time invariant biases, we use the following differenced outcome earnings equation:

$$(y_{78,i} - y_{75,i}) = \theta_0 \cdot s_i + (\varepsilon_{78,i} - \varepsilon_{75,i}) .$$

where s_i is the treatment group indicator. This specification does not contain any covariates due to the time invariance of the variables in the data set, and assumes a constant treatment effect. The parameter of interest is therefore θ_0. If one has a valid control group that can be used to impute the missing counterfactual outcome for program participants in the absence of treatment, the above equation can be used to consistently estimate θ_0 if the following condition will hold:

$$\mathrm{E}\left[(\varepsilon_{78,i} - \varepsilon_{75,i}) \,|s_i = 0\right] = 0.$$

However, comparing differenced earnings of individuals drawn from the PSID with those from the experimental treatment group reveals, especially for the men, substantial differences in the distributions of both groups, suggesting that individuals in the control group may indeed differ more systematically from the randomized treatment group and making it likely that $\mathrm{E}\left[(\varepsilon_{78,i} - \varepsilon_{75,i}) \,|s_i = 0\right] \neq 0$. In other words, our method aims at correcting what could be labeled as 'residual (i.e., time-varying) selection on unobservables' using pre-program (1974) earnings and employment status as instruments. Note in this context that we would only identify the treatment effect for individuals with a selection bias $\mathrm{E}\left[(\varepsilon_{78,i} - \varepsilon_{75,i}) \,|s_i = 0\right]$ close to zero if the assumption of a homogeneous treatment effect was relaxed.

HOW WAS THE BANDWIDTH CHOSEN; S.E. CALCULATED, POLYORDER... Our estimator proceeds as described in the paper before: first, we estimate the probability of belonging to the control

[26] More specifically, men were selected if randomized during the first four months of 1976, while women were sampled if randomized during 1976. See Smith and Todd (2005) and Calonico and Smith (2017) for further discussion on the reasons for the choice of these particular samples.

group ($s_i = 0$), say $1 - \widehat{p}_i$.[27] Second, we fit a local polynomial regression of $(y_{78,i} - y_{75,i})$ on $(1 - \widehat{p}_i)$ using the control group sample to estimate the intercept. The difference between the average earnings difference of the treatment group and the estimated intercept from the control group is then used as the estimated treatment effect. The results of our investigations are provided in Table 2 for both men and women. The relevant sample sizes are given in the notes to the tables.

We start with the results for the sample of men: the first two rows provide the estimates using the experimental treatment and control group where the dependent variable is earnings in 1978 and the differenced earnings, respectively.[28] The figures are \$2,749 and \$2,429, and are both significantly different from zero at conventional significance levels. Note that allowing for time invariant unobservables (as in the differenced equation) gives a slightly lower estimate. In row 3 we present the results from using the PSID data but not correcting for any selection. The estimate we obtain for the differenced earnings equation is \$2,270, which is much closer to the experimental figure than the estimate from the Heckman two-step procedure, which yields a significant estimate of \$3,671 (row 4). The last row provides the estimate using our method: using a third order polynomial, we estimate the treatment effect to be \$2,413 with a standard error of \$996. Note that these figures are fairly robust to the choice of the bandwidth, to smaller changes of the polynomial order as well as to the specification of the propensity score. For instance, choosing a range of bandwidths from .2 to .8 different from the 'optimal' one of .6 yielded estimates ranging from \$2,547 to \$2,410, respectively. Likewise, a specification for the propensity score with various interactions left the treatment effect estimate almost unchanged, while choosing a second or fourth order polynomial resulted in estimates of \$2,332 and \$2,553, respectively.[29] Given the rather drastic 'over-shooting' of the Heckman two-step procedure, the above findings suggest that 'selection on unobservables' plays a role for men, but that the normal distribution might not be a good approximation for the selection bias term in this context.[30]

Next, we turn to the results for the women data. Using the same specification as before, the figures from the second part of Table 2 display estimates of \$1,342 and \$1,247 for the experimental sample, both of which are significantly different from zero at conventional significance levels. Unlike before, however, we obtain a least squares estimate rather different from the experimental effect of the differenced equation, namely \$2,835 (with a standard error of \$438). The Heckman two step and our estimates are even higher with \$5,365 and \$4,077, respectively, both significant at conventional levels. We note that our estimate is fairly stable across different specifications: for instance, varying the bandwidth parameter from .1 (the chosen level) to .4 estimating the propensity score using various interactions and number of children as additional regressors does not substantially affect the estimate. Similarly,

A final comment... As has aptly been discussed in the literature, data such as the NSW training data suffer from choice-based sampling with unknown sampling weights (Heckman and Todd, 2009). That is, frequencies of treated individuals in the sample do not correspond to the treatment frequen-

[27] The propensity score is estimated using the following covariates: age, age squared, years of education, earnings in 1974, and indicators for (i) marital status, (ii) ethnicity (black, hispanic), (iii) high school dropout, (iv) unemployed in 1974. The descriptive statistics for our sample members can be found in column 3 of Table 1 of Smith and Todd (2005) and columns 3 and 5 of Table 1 of Calonico and Smith (2017), respectively. All variables have been recorded prior to treatment.

[28] All earnings are in real terms.

[29] We also ran a 'placebo test' using the experimental control group as treatment group. The result was an estimated effect of \$4.86 with a standard error of \$795.

[30] This is under the maintained assumption that the additive linear model is correctly specified.

cies in the underlying population of interest, which implies that the estimated propensity score is not necessarily consistent for the true population quantity of interest. While a major problem for propensity score matching, we deem this problem of lesser importance in our case for the following reason: several sensitivity checks indicate that our semi-parametrically estimated propensity score is very similar to that of a logit model, which is known to be 'robust' against the choice-based sampling problem in the sense that only the scale of the propensity score is affected when the data generating process is indeed logistic (Manski and Mc Fadden, 1981).[31] Though, as long as choice based sampling preserves the functional form of the propensity score (but not the scale), the latter does not represent a particular problem for our method.

As additional robustness check, we also ran a 'placebo test' using the experimental control sample as treatment group, which yielded an average treatment effect close to zero ($ 5 with a standard error of $ 573).

and point out that first stage estimation is rather robut.

should stress that in line with Smith and Todd, accounting for selection on unobservables matters. BUt what we add is that it is not only selection on time-invariant unobservables, but also selection on time-varying unobservables for the men. For the women, this does not play an important role. This is in line with the finding of Calonico and Smith that time-invariant selection on unobservables does not seem to play an important role.

7 Conclusion

To be inserted...

[31] This is confirmed by the fact that estimated treatment effects using a logistic propensity score are very similar to the ones we obtain when using a semi-parametric propensity score.

Table 2: Re-analysis of the LaLonde Data

Data Type	Dep. Variable	Details[1]	Est. Coeff.	S.E.
Men				
Exper.[2]	$(y_{78,i} - y_{75,i})$	No Reg.	\$ 2,429	\$ 1,099
Exper.[2]	$y_{78,i}$	No Reg.	\$ 2,748	\$ 1,005
Non-Exp.[3]	$(y_{78,i} - y_{75,i})$	No Reg.	\$ 2,270	\$ 959
Non-Exp.[3]	$(y_{78,i} - y_{75,i})$	Heckman 2-step	\$ 3,671	\$ 1,546
Non-Exp.[3]	$(y_{78,i} - y_{75,i})$	3rd Order Poly.	\$ 2,413	\$ 996
Women				
Exper.[4]	$(y_{79,i} - y_{75,i})$	No Reg.	\$ 1,324	\$ 479
Exper.[4]	$y_{79,i}$	No Reg.	\$ 1,247	\$ 466
Non-Exp.[5]	$(y_{79,i} - y_{75,i})$	No Reg.	\$ 2,835	\$ 438
Non-Exp.[5]	$(y_{79,i} - y_{75,i})$	Heckman 2-step	\$ 5,365	\$ 738
Non-Exp.[5]	$(y_{79,i} - y_{75,i})$	3rd Order Poly.	\$ 4,077	\$ 1,897

[1] This column provides details about the regression. Note that all earnings are in \$ 1982. The propensity score for the Heckman 2-step procedure and our method is estimated using age, age squared, years of education, earnings in 1974, and indicators for (i) marital status, (ii) ethnicity (black, hispanic), (iii) high school dropout, (iv) unemployed in 1974.
[2] Treatment group: 108 ind.; Control group: 142 ind.
[3] Treatment group: 108 ind.; Control group: 2,490 ind.
[4] Treatment group: 285 ind.; Control group: 279 ind.
[5] Treatment group: 285 ind.; Control group: 648 ind.

Appendix

A Identification Results

Proof of Theorem I1. Denote $a_i = y_i - x_i'\beta_0$. From Assumptions I1 and I3, we can deduce that:

$$\mathrm{E}^*[a_i|p_i] = \theta_0 + \lambda(p_i).$$

Also:

$$\mathrm{E}^*[a_i|1] = \theta_0 + \lambda(1) = \theta_0,$$

which follows from Assumption I2 noting that $\lambda(1)$ corresponds to $\mathrm{E}[\varepsilon_i]$, which is normalized to zero by Theorem I1. Since β_0 is known, the claim follows.[32] ■

Proof of Theorem I2. We prove Theorem I2 using Assumption I3(i), the one using I3(ii) follows

[32] Note that the extension to an error term distribution with unbounded support is immediate. That is, suppose that $\Psi_p = (0,1)$ or $\Psi_p = (c,1)$ for some $0 \leq c < 1$, then in fact it holds that:

$$\mathrm{E}^*[a_i|p_i] = \theta_0 + \lambda(p_i).$$

by identical arguments. As before, from Assumptions I1 and I3:

$$\mathrm{E}^*[y_i|x_i, p_i] = g\left(x_{1i}'\mathbf{b}_0\right) + \lambda(p_i)$$

and:

$$\mathrm{E}^*[y_i|x_i, 1] = g\left(x_{1i}'\mathbf{b}_0\right) + \lambda(1).$$

Also, by Assumption I3, note that $\lambda(1)$ corresponds to $\mathrm{E}[\varepsilon_i]$, which is normalized to zero by Theorem I2. Thus, recalling the definition of $m(\cdot,\cdot)$, it follows that $\lambda(p_i) = m(x_i, p_i) - m(x_i, 1)$. Using Assumption I5, θ_0 is therefore identified as the minimizer of:

$$\mathbf{b}_0 = \arg\min_{\mathbf{b}} \mathrm{E}^*\left[\left(y_i - g\left(x_{1i}'\mathbf{b}\right) + m(x_i, 1) - m(x_i, p_i)\right)^2\right].$$

■

Proof of Theorem I3.

Again, the proof of Theorem I3 uses Assumption I3(ii) only. From $\mathrm{E}[\widetilde{\varepsilon}_i] = 1$ and Assumption I3, we can deduce that:

$$\mathrm{E}^*[y_i|x_i, p_i] = \widetilde{g}\left(x_{1i}'\mathbf{b}_0\right)\widetilde{\lambda}(p_i),$$

and, by the same argument as before, that:

$$\mathrm{E}^*[y_i|x_i, 1] = \widetilde{g}\left(x_{1i}'\mathbf{b}_0\right).$$

Therefore, it holds that:

$$\widetilde{\lambda}(p_i) = \frac{m(x_i, p_i)}{m(x_i, 1)}.$$

And by Assumption I7, it follows that θ_0 is identified as the minimizer of:

$$\mathbf{b}_0 = \arg\min_{\mathbf{b}} \mathrm{E}^*\left[\left(y_i \frac{m(x_i, 1)}{m(x_i, p_i)} - g\left(x_{1i}'\mathbf{b}_0\right)\right)^2\right],$$

which follows since:

$$\mathrm{E}^*\left[\frac{\widetilde{u}_i}{\lambda(p_i)}\right] = 0$$

by iterated expectations. ■

Proof of Theorem I4. First note that as before:

$$\mathrm{E}^*[y_i|z_i] = \mathrm{E}^*[y_i|x_i, p_i, v_i] = \theta_0 + x_i'\beta_0 + \lambda(p_i, v_i).$$

Next, notice that $\lambda(1, v_i) = \mathrm{E}[\varepsilon_i|v_i]$ so that:

$$\int \mathrm{E}[\varepsilon_i|v] f(v) dv = 0,$$

---

by Assumptions I1 and I3. Moreover, by the modified version of Assumption I2, it holds that:

$$\lim_{p \to 1} \mathrm{E}^*[a_i|p] = \theta_0.$$

The conclusion again since β_0.

where $f(v)$ denotes a weight function satisfying $\int f(v)dv = 1$. Therefore, it holds that:

$$\lambda(p_i, v_i) = m(x_i, p_i, v_i) - \int m(x_i, 1, v) f(v) dv$$

with $m(x_i, p_i, v_i) \equiv \mathrm{E}^*[y_i | x_i, p_i, v_i]$. ■

B Asymptotic Results

B.1 Notation

General. Define $\widehat{w}_i \equiv (x_i', \widehat{p}_i)'$, $w_{i1} \equiv (x_i', 1)'$, as well as $w_i \equiv (x_i', p_i)'$. Moreover, with slight abuse of notation, we will sometimes use $\widehat{w}_{i1}$ to highlight that $m(w_{i1})$ is estimated using observations $\widehat{w}_j$, $j = 1, \ldots, n$, and use w_{-i} to denote x_i without p_i. Also, we will write $\sup_w$ and $\sup_{\mathbf{b}}$ to denote the suprema of the support of w_i and of $\mathcal{B}_0$ further specified in Assumptions E10 and E12). Finally, recall that $d_x = \dim(x_i)$. We will adopt the following notation introduced by Masry (1996): for the estimation of $\widehat{m}(\widehat{w}_i)$ and $\widehat{m}(w_{i1})$ the aim is to obtain the vectors b and b^* from a q-th order local polynomial regression of y_i on $\widehat{w}_i$ and w_{i1}, respectively, by minimizing:

$$Q_{m,n}(b, \widehat{w}_i) = \frac{1}{n h_2^2} \sum_{j=1}^{n} K\left(\frac{\widehat{w}_j - \widehat{w}_i}{h_2}\right) \left\{ y_i - \sum_{0 \leq |\mathbf{t}| \leq q} b_{\mathbf{t}} \left(\widehat{w}_j - \widehat{w}_i\right)^{\mathbf{t}} \right\}^2 \tag{16}$$

and

$$Q_{m_1,n}(b^*, w_{i1}) = \frac{1}{n h_2^2} \sum_{j=1}^{n} K\left(\frac{\widehat{w}_j - w_{i1}}{h_2}\right) \left\{ y_i - \sum_{0 \leq |\mathbf{t}| \leq q} b^*_{\mathbf{t}} \left(\widehat{w}_j - w_{i1}\right)^{\mathbf{t}} \right\}^2, \tag{17}$$

where b_0 and b_0^* denote the minimizing intercepts and:

$$\mathbf{t} = \frac{1}{\mathbf{t}!} \frac{\partial^{|\mathbf{t}|} m(w_i)}{\partial^{t_1} w_1 \ldots \partial^{t_{d_x}} w_{d_x} \partial^{t_{d_x+1}} w_{d+1}}$$

In line with the literature, we also use the following conventions:

$$\mathbf{t} = (t_1, \ldots, \mathbf{t}_{d_x+1})', \qquad \mathbf{t}! = \mathbf{t}_1! \times \ldots \times \mathbf{t}_{d_x+1}!, \qquad |\mathbf{t}| = \sum_{h=1}^{d_x+1} \mathbf{t}_h$$

as well as

$$a^{\mathbf{t}} = a_1^{\mathbf{t}_1} \times \ldots \times a_{d_x+1}^{\mathbf{t}_{d_x+1}}, \qquad \sum_{0 \leq |\mathbf{t}| \leq q} = \sum_{h=0}^{q} \sum_{\mathbf{t}_1 + \ldots + \mathbf{t}_{d_x+1} = h}^{h} \ldots \sum_{\mathbf{t}_{d_x+1}}^{h}.$$

Next, let $N_l = (l+h-1)!/(l!(h-l)!)$ be the number of distinct (d_x+1) tuples $\mathbf{t}$ with $\mathbf{t} = l$, where $h = l$ and $h = d_x+1$ (e.g. $N_0 = 1$, $N_1 = d_x+1$ (d_x+1 first order derivatives), $N_2 = ((d_x+1)(d_x+2)/2)$, etc.). After arranging them in the corresponding lexicographical order, we let ϕ_l^{-1} denote this one-to-one

map. Since $K(\nu) = k(\nu_1) \times \ldots \times k(\nu_{d_x+1})$, where $k(\cdot)$ is a univariate kernel function further defined in Assumption E9. For each $\mathbf{t}$ with $0 \leq \mathbf{t} \leq 2q$, let

$$\mu_{\mathbf{t}}(K) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} (\nu_1, \ldots, \nu_{d_x+1})^{\mathbf{t}} \prod_{l=1}^{d_x+1} k(\nu_l) d\nu_1 \ldots d\nu_{d_x+1} = \int_{\mathbb{R}^{d_x+1}} \nu^{\mathbf{t}} K(\nu) d\nu$$

and

$$\mu_{1,\mathbf{t}}(K) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \int_{-\infty}^{0} (\nu_1, \ldots, \nu_{d_x+1})^{\mathbf{t}} \prod_{l=1}^{d_x+1} k(\nu_l) d\nu_1 \times \ldots \times d\nu_{d_x+1}.$$

Also, let

$$\gamma_{\mathbf{t}}(K) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \int_{-\infty}^{0} (\nu_1, \ldots, \nu_{d_x+1})^{\mathbf{t}} \prod_{l=1}^{d_x} k(\nu_l) k^2(\nu_{d_x+1}) d\nu_1 \times \ldots \times d\nu_{d_x+1}.$$

Then, let $N = \sum_{l=0}^{q} N_l$ and define the $N \times N$ matrices:

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_{0,0} & \mathbf{M}_{0,1} & \cdots & \mathbf{M}_{0,q} \\ \mathbf{M}_{1,0} & \mathbf{M}_{1,1} & \cdots & \mathbf{M}_{1,q} \\ \vdots & & \ddots & \vdots \\ \mathbf{M}_{q,0} & & \cdots & \mathbf{M}_{q,q} \end{bmatrix}, \tag{18}$$

$$\mathbf{M}_1 = \begin{bmatrix} \mathbf{M}_{1;0,0} & \mathbf{M}_{1;0,1} & \cdots & \mathbf{M}_{1;0,q} \\ \mathbf{M}_{1;1,0} & \mathbf{M}_{1;1,1} & \cdots & \mathbf{M}_{1;1,q} \\ \vdots & & \ddots & \vdots \\ \mathbf{M}_{1;q,0} & & \cdots & \mathbf{M}_{1;q,q} \end{bmatrix}, \tag{19}$$

and

$$\mathbf{\Gamma} = \begin{bmatrix} \mathbf{\Gamma}_{0,0} & \mathbf{\Gamma}_{0,1} & \cdots & \mathbf{\Gamma}_{0,q} \\ \mathbf{\Gamma}_{1,0} & \mathbf{\Gamma}_{1,1} & \cdots & \mathbf{\Gamma}_{1,q} \\ \vdots & & \ddots & \vdots \\ \mathbf{\Gamma}_{q,0} & & \cdots & \mathbf{\Gamma}_{q,q} \end{bmatrix}, \tag{20}$$

as well as

$$\mathbf{B} = \begin{bmatrix} \mathbf{M}_{0,q+1} \\ \mathbf{M}_{1,q+1} \\ \vdots \\ \mathbf{M}_{q,q+1} \end{bmatrix},$$

where $\mathbf{M}_{i,j}$, $\mathbf{M}_{1;i,j}$, and $\mathbf{\Gamma}_{i,j}$ are $N_i \times N_j$ dimensional matrices with (m,l) elements (m,l) $\mu_{\phi_i(l)+\phi_j(m)}$, $\mu_{1;\phi_i(l)+\phi_j(m)}$ and $\gamma_{\phi_i(l)+\phi_j(m)}$, respectively.[33] Next, let $\mathcal{K}_j(w_i)$ denote an $N \times 1$ vector, i.e.

$$\mathcal{K}_j(w_i) = \begin{bmatrix} \mathcal{K}_{j,0}(w_i) \\ \mathcal{K}_{j,1}(w_i) \\ \vdots \\ \mathcal{K}_{j,q}(w_i) \end{bmatrix}.$$

Each element, $\mathcal{K}_{j,l}(w_i)$, is of dimension $N_l \times 1$ whose l_0-th element is given by $[\mathcal{K}_{j,l}(w_i)]_{l_0} = ((w_j - w_i)/h_2)^{\phi_l(l_0)} K((w_j - w_i)/h_2)$. For instance,

$$\mathcal{K}_{j,1}(w_i) = \begin{bmatrix} ((w_{j,1} - w_{i,1})/h_2) K((w_j - w_i)/h_2) \\ ((w_{j,2} - w_{i,2})/h_2) K((w_j - w_i)/h_2) \end{bmatrix}$$

and

$$\mathcal{K}_{j,2}(w_i) = \begin{bmatrix} ((w_{j,1} - w_{i,1})/h_2)^2 K((w_j - w_i)/h_2) \\ ((w_{j,1} - w_{i,1})/h_2)((w_{j,2} - w_{i,2})/h_2) K((w_j - w_i)/h_2) \\ ((w_{j,2} - w_{i,2})/h_2)((w_{j,1} - w_{i,1})/h_2) K((w_j - w_i)/h_2) \\ ((w_{j,2} - w_{i,2})/h_2)^2 K((w_j - w_i)/h_2) \end{bmatrix}.$$

The corresponding matrices are denoted as follows:

$$\mathbf{M}_n(w_i) = \begin{bmatrix} \mathbf{M}_{n;0,0}(w_i) & \mathbf{M}_{n;0,1}(w_i) & \cdots & \mathbf{M}_{n;0,q}(w_i) \\ \mathbf{M}_{n;1,0}(w_i) & \mathbf{M}_{n;1,1}(w_i) & \cdots & \mathbf{M}_{n;1,q}(w_i) \\ \vdots & & \ddots & \vdots \\ \mathbf{M}_{n;q,0}(w_i) & & \cdots & \mathbf{M}_{n;q,q}(w_i) \end{bmatrix}$$

where $\mathbf{M}_{n;i,j}(w_i)$ is of dimension $N_i \times N_j$ with (l, l_0) element:

$$[\mathbf{M}_{n;i,j}(w_i)]_{l,l_0} = \frac{1}{n h_2^{d_x+1}} \sum_{j=1}^{n} \left(\frac{(w_j - w_i)}{h_2} \right)^{\phi_i(l)+\phi_j(l_0)} K\left(\frac{(w_j - w_i)}{h_2} \right).$$

For instance:

$$[\mathbf{M}_{n;1,1}(w_i)]_{1,2} = \frac{1}{n h_2^{d_x+1}} \sum_{j=1}^{n} \left(\frac{(w_{j,1} - w_{i,1})}{h_2} \right) \left(\frac{(w_{j,2} - w_{i,2})}{h_2} \right) K\left(\frac{(w_j - w_i)}{h_2} \right).$$

Let $\widehat{\mathcal{K}}_j(\widehat{w}_i)$ and $\widehat{\mathbf{M}}_n(\widehat{w}_i)$ be defined in a corresponding manner, but with generated regressors $\{\widehat{w}_j\}_{j=1}^n$. Arrange the $N(r)$ elements of the derivatives:

$$D^{\mathbf{r}} m(w) \equiv \frac{\partial^r m(w)}{\partial^{r_1} w_1 \ldots \partial^{r_k} w_k} \qquad \text{for} \quad |\mathbf{r}| = r$$

[33] For the asymptotic results in Remarks 1 and 2, note that Γ_0 contains elements $\gamma_{\mathbf{t},0}(K)$ defined as:

$$\gamma_{\mathbf{t},0}(K) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \int_{-\infty}^{0} (\nu_1, \ldots, \nu_{d_x+1})^{\mathbf{t}} \prod_{l=1}^{d_x} k^2(\nu_l) k^2(\nu_{d_x+1}) d\nu_1 \times \ldots \times d\nu_{d_x+1}.$$

as the $N(r) \times 1$ column vectors $m^{(r)}(w)$ in the lexicographial order mentioned above. Also, subscripts $D_j^{\mathbf{1}}$ with $j = 1, \ldots, d_x + 1$ will be used to refer to specific elements of the first order partial derivative of $m(w)$. Finally, let $\iota_1 = (1, 0, \ldots, 0)' \in \mathbb{R}^N$. Then, using Equation (2.13) (p. 576) and Corollary 2(ii) (p. 580) in Masry (1996), we can write:

$$\widehat{m}(w_i) - m(w_i) = \iota_1' \left[\mathbf{M} f(w_i)\right]^{-1} \{1 + o_p(1)\} \left\{ \frac{1}{n h_2^{d_x+1}} \sum_{j=1}^{n} \mathcal{K}_j(w_i) \left[u_j + \sum_{|\mathbf{k}|=q+1} \frac{1}{\mathbf{k!}} D^{\mathbf{k}} m(w_i)(w_j - w_i)^{\mathbf{k}} + \gamma_n \right] \right\}, \tag{21}$$

where

$$\mu_n \equiv (q+1) \frac{1}{n h_2^{d_x+1}} \frac{1}{\mathbf{k!}} \sum_{|\mathbf{k}|=q+1} \mathcal{K}_j(w_i)(w_j - w_i)^{\mathbf{k}} \int_0^1 \left\{ D^{\mathbf{k}} m(w_i + \tau(w_j - w_i)) - D^{\mathbf{k}} m(w_i) \right\} (1-\tau)^q d\tau. \tag{22}$$

Note that $\frac{1}{n h_2^{d_x+1}} \sum_{j=1}^{n} \mathcal{K}_j(w_i)$ converges in mean squared to $\mathbf{M}_0 f(w_i)$, where $\mathbf{M}_0$ is a $N \times 1$ vector where the l_0 element of the l-th component is $\int \nu^{\phi_l(l_0)} K(\nu) d\nu$. Finally, define $\mu(w_i) \equiv \mathrm{E}[\gamma_n(w_i)]$. Then, by Proposition 2 (page 581) and by Theorem 4 (page 582) in Masry (1996), it follows that

$$\sup_w |\mu(w)| = o(h_2^{q+1})$$

and

$$\sup_w |h_2^{-(q+1)} \mu_n(w) - \mu(w)| = h_2^{-(q+1)} O_p(n^{-\frac{1}{2}} h_2^{-(d_x+1)/2} \sqrt{\ln(n)}).$$

Also, let

$$\varphi_n(w_i) \equiv \frac{1}{n h_2^{d_x+1}} \sum_{j=1}^{n} \mathcal{K}_j(w_i) \left(\sum_{|\mathbf{k}|=q+1} \frac{1}{\mathbf{k!}} D^{\mathbf{k}} m(w_i)(w_j - w_i)^{\mathbf{k}} \right), \tag{23}$$

and

$$\varphi(w_i) = \mathbf{B} m^{q+1}(w_i) f(w_i).$$

Then, by Theorem 2 (page 579) in Masry (1996), it follows that:

$$\sup_w |h_2^{-(q+1)} \varphi_n(w) - \varphi(w)| = O_p(n^{-\frac{1}{2}} h_2^{-(d_x+1)/2} \sqrt{\ln(n)}).$$

Additive Linear. Since the estimator in the additive linear case is a univariate higher order local polynomial estimator and conceptually different from the estimators in the general additive and multiplicative case, respectively, we define some additional notation tailored to this case only. That is, similar to before, for $0 \leq t \leq 2q$ let:

$$\mu_{1,t}(k) = \int_{-\infty}^{0} \nu^t k(\nu) d\nu.$$

as well as

$$\gamma_t(k) = \int_{-\infty}^{0} \nu^t k^2(\nu) d\nu,$$

and define the $q \times q$ lower dimensional matrices:

$$\mathbf{M}_1^1 = \begin{bmatrix} \mu_{1,0}(k) & \dots & \mu_{1,q}(k) \\ \vdots & \ddots & \vdots \\ \mu_{1,q}(k) & \dots & \mu_{1,2q}(k) \end{bmatrix} \tag{24}$$

as well as

$$\mathbf{\Gamma}^1 = \begin{bmatrix} \gamma_0(k) & \dots & \gamma_q(k) \\ \vdots & \ddots & \vdots \\ \gamma_q(k) & \dots & \gamma_{2q}(k) \end{bmatrix}. \tag{25}$$

Also, let

$$\mathcal{X}_i(1) = \left(1, (p_i - 1)\frac{1}{h_2}, ..., (p_i - 1)^q \frac{1}{q! h_2^q}\right)'$$

and denote $\mathcal{Y}_i = y_i - x_i'\beta_0$. Finally, with slight abuse of notation, call:

$$K_i(1) = k\left(\frac{p_i - 1}{h_2}\right).$$

B.2 Lemmas Additive Case

Lemma B.1. *Under Assumptions E8 through E13, it holds that:*

$$\sqrt{nh_2}\frac{1}{n}\sum_{i=1}^{n} s_i u_i = O_p\left(\sqrt{h_2}\right) \tag{26}$$

uniformly in $\mathbf{b}$.

Lemma B.2. *Under Assumptions E8 through E13, it holds that:*

$$\sqrt{nh_2}\frac{1}{n}\sum_{i=1}^{n} s_i(m(\widehat{w}_i) - \widehat{m}(\widehat{w}_i)) = O_p\left(\sqrt{h_2}\right) \tag{27}$$

uniformly in $\mathbf{b}$.

Lemma B.3. *Under Assumptions E8 through E13, it holds that:*

$$\sqrt{nh_2}\frac{1}{n}\sum_{i=1}^{n} s_i(m(\widehat{w}_{i1}) - \widehat{m}(\widehat{w}_{i1})) \xrightarrow{d} N(0, \Omega_0), \tag{28}$$

uniformly in $\mathbf{b}$, *where* Ω_0 *was defined in Theorem E2 above.*

Lemma B.4. *Under Assumptions E8 through E13, it holds that:*

$$\sqrt{nh_2}\frac{1}{n}\sum_{i=1}^{n} s_i(m(w_i) - m(\widehat{w}_i)) = O_p\left(\sqrt{h_2}\right) \tag{29}$$

uniformly in $\mathbf{b}$.

Lemma B.5. *Under Assumptions E8 through E13, it holds that:*

$$\sqrt{nh_2}\frac{1}{n}\sum_{i=1}^{n} s_i(m(w_{i1}) - m(\widehat{w}_{i1})) = O_p\left(\sqrt{h_2}\right) \tag{30}$$

uniformly in **b**.

B.3 Lemmas Multiplicative Case

Lemma B.6. *Under Assumptions E8 through E13 and $m(w_i) \neq 0$ a.s., it holds that:*

$$\sqrt{nh_2}\frac{1}{n}\sum_{i=1}^{n} s_i\frac{\widetilde{u}_i}{m(w_i)}(m(\widehat{w}_{i1}) - \widehat{m}(\widehat{w}_{i1})) = O_p\left(\sqrt{h_2}\right), \tag{31}$$

uniformly in **b**.

Lemma B.7. *Under Assumptions E8 through E13 and $m(w_i) \neq 0$ a.s., it holds that:*

$$\sqrt{nh_2}\frac{1}{n}\sum_{i=1}^{n} s_i\frac{\widetilde{u}_i}{m(w_i)}(m(w_{i1}) - m(\widehat{w}_{i1})) = O_p\left(\sqrt{h_2}\right) \tag{32}$$

uniformly in **b**.

Lemma B.8. *Under Assumptions E8 through E13 and $m(w_i) \neq 0$ a.s., it holds that:*

$$\sqrt{nh_2}\frac{1}{n}\sum_{i=1}^{n} s_i\widetilde{u}_i(m(\widehat{w}_i) - \widehat{m}(\widehat{w}_i))\frac{m(w_{i1})}{m(w_i)^2} = O_p\left(\sqrt{h_2}\right) \tag{33}$$

uniformly in **b**.

Lemma B.9. *Under Assumptions E8 through E13 and $m(w_i) \neq 0$ a.s., it holds that:*

$$\sqrt{nh_2}\frac{1}{n}\sum_{i=1}^{n} s_i\widetilde{u}_i(m(w_i) - m(\widehat{w}_i))\frac{m(w_{i1})}{m(w_i)^2} = O_p\left(\sqrt{h_2}\right) \tag{34}$$

uniformly in **b**.

B.4 Proofs of Main Theorems

Proof of Theorem E1. To prove (i) and (ii), define:

$$S_{n,1}(b_0, b_1, ..., b_q) \equiv \frac{1}{nh_2}\sum_{i=1}^{n} s_i\left(y_i - b_0 - x_i'\widehat{\beta} - \frac{1}{t!}\sum_{t=1}^{q} b_q(\widehat{p}_i - 1)^t\right)^2 k\left(\frac{\widehat{p}_i - 1}{h_2}\right)$$

and

$$\overline{S}_{n,1}(b_0, b_1, ..., b_q) \equiv \frac{1}{nh_2}\sum_{i=1}^{n} s_i\left(y_i - b_0 - x_i'\beta_0 - \frac{1}{t!}\sum_{t=1}^{q} b_q(p_i - 1)^t\right)^2 k\left(\frac{p_i - 1}{h_2}\right),$$

so that $\widetilde{b}(1)$ is the first element of:

$$\left(\widetilde{b}_0(1), \widetilde{b}_1(1), ..., \widetilde{b}_q(1)\right) = \arg \min_{b_0, b_1, ..., b_{q_2}} S_n\left(b_0, b_1, ..., b_q\right)$$

We study the limiting behavior of $\widetilde{b}(1)$, and subsequently analyze:

$$\sqrt{nh_2}\left(\widetilde{b}_0(1) - \widehat{b}_0(1)\right) = o_p(1).$$

Now, recalling the definitions of $\mathcal{Y}_i$, $\mathcal{X}_i$, and $K_i(1)$ from Section B.1, note that:

$$\widetilde{b}(1) = \left(\frac{1}{nh_2}\sum_{i=1}^{n} s_i \mathcal{X}_i(1) K_i\left(1\right) \mathcal{X}_i'(1)\right)^{-1} \frac{1}{nh_2}\sum_{i=1}^{n} s_i \mathcal{X}_i(1) K_i\left(1\right) \mathcal{Y}_i$$

and $\widetilde{b}_0(1) = \iota'\widetilde{b}(1)$. Now, noting that $b_0(1) = \theta_0$, and recalling that $u_i = y_i - \theta_0 - x_i'\theta_0 - \lambda(p_i)$,

$$\begin{aligned}
\widetilde{b}(1) &= \left(\frac{1}{nh_2}\sum_{i=1}^{n} s_i \mathcal{X}_i(1) K_i\left(1\right) \mathcal{X}_i'(1)\right)^{-1} \frac{1}{nh_2}\sum_{i=1}^{n} s_i \mathcal{X}_i(1) K_i\left(1\right) \left(\theta_0 + \lambda(p_i) + u_i\right) \\
&= \left(\frac{1}{nh_2}\sum_{i=1}^{n} s_i \mathcal{X}_i(1) K_i\left(1\right) \mathcal{X}_i'(1)\right)^{-1} \frac{1}{nh_2}\sum_{i=1}^{n} s_i \mathcal{X}_i(1) K_i\left(1\right) \left(\theta_0 + \frac{1}{t!}\sum_{t=1}^{q} b_t \left(p_i - 1\right)^t + r_n(1) + u_i\right) \\
&= \left(\frac{1}{nh_2}\sum_{i=1}^{n} s_i \mathcal{X}_i(1) K_i\left(1\right) \mathcal{X}_i'(1)\right)^{-1} \frac{1}{nh_2}\sum_{i=1}^{n} s_i \mathcal{X}_i(1) K_i\left(1\right) \left(\mathcal{X}_i'(1) b_0 + r_{i,n}(1) + u_i\right) \\
&= b_0 + \left(\frac{1}{nh_2}\sum_{i=1}^{n} s_i \mathcal{X}_i(1) K_i\left(1\right) \mathcal{X}_i'(1)\right)^{-1} \frac{1}{nh_2}\sum_{i=1}^{n} s_i \mathcal{X}_i(1) K_i\left(1\right) u_i \\
&\quad + \left(\frac{1}{nh_2}\sum_{i=1}^{n} s_i \mathcal{X}_i(1) K_i\left(1\right) \mathcal{X}_i'(1)\right)^{-1} \frac{1}{nh_2}\sum_{i=1}^{n} s_i \mathcal{X}_i(1) K_i\left(1\right) r_{i,n}(1)
\end{aligned}$$

where

$$r_{i,n}(1) = \lambda(p_i) - b_0(1) + \frac{1}{t!}\sum_{t=1}^{q} b_t(1)\left(p_i - 1\right)^t.$$

Since $k(\cdot)$ has support on $[-1, 1]$ by Assumption E9, we have that:

$$\frac{1}{nh_2}\sum_{i=1}^{n} s_i \mathcal{X}_i(1) \widetilde{K}_i\left(1\right) \mathcal{X}_i'(1) \xrightarrow{p} \mathbf{M}_1^1 \,\mathrm{E}[s_i] f(1),$$

where $\mathbf{M}_1^1$ was defined in Section B.1. Also, by Lindeberg-Levy CLT:

$$\frac{1}{\sqrt{nh_2}}\sum_{i=1}^{n} s_i \mathcal{X}_i(1) \widetilde{K}_i\left(1\right) u_i \xrightarrow{d} N\left(0, V\right),$$

where $V = \mathbf{\Gamma}^1 \mathrm{E}[s_i \sigma^{*2}(1)] f(1)$ and $\mathbf{\Gamma}^1$ was defined in Equation (25) of Section B.1. Now, note that

$$\begin{aligned} & \frac{1}{h_2} \int_0^1 \left(\lambda(p) - \frac{1}{t!} \sum_{t=1}^{q} b_t(1) (p-1)^t \right) k\left(\frac{p-1}{h_2} \right) f(p) dp \\ = & \int_{-1/h_2}^{0} \left(\lambda(1+vh_2) - \frac{1}{t!} \sum_{t=1}^{q} b_t^{\dagger}(1) (vh_2)^t \right) k(v) f(1+vh_2) dv \\ = & h_2^{q+1} f(1) D^{q+1} \lambda(1) \int_{-1/h_2}^{0} k(v) f(1+vh_2) dv \end{aligned}$$

and finally

$$\begin{aligned} & h_2^{-(q+1)} \left(\frac{1}{nh_2} \sum_{i=1}^{n} s_i \mathcal{X}_i(1) K_i(1) \mathcal{X}_i'(1) \right)^{-1} \frac{1}{nh_2} \sum_{i=1}^{n} s_i \mathcal{X}_i(1) K_i(1) r_{i,n}(1) \\ \xrightarrow{p} & \left[\mathbf{M}^1 f(1) \right]^{-1} \begin{pmatrix} \int_{-1}^{0} k(v) du \\ \vdots \\ \int_{-1}^{0} k(v) v^q du \end{pmatrix} f(1) \int_{-1}^{0} k(v) dv D^{q+1} \lambda(1) \end{aligned}$$

Hence,

$$\sqrt{nh_2} \left(\widetilde{b}_0(1) - b_0(1) \right) \xrightarrow{d} N(0, \Omega_{0,1})$$

with $\Omega_{0,1}$ as defined in the statement of the theorem. We now need to show that

$$\sqrt{nh_2} \left(\widetilde{b}_0(1) - \widehat{b}_0(1) \right) = o_p(1). \tag{35}$$

From Assumption E7, regardless p_i is an interior point or on the boundary, we have

$$\begin{aligned} \widehat{p}_i - p_i &= h_1^2 \Xi_n(z_i'\gamma) + \frac{1}{\sqrt{nh_1}} \sum_{j=1}^{n} \omega\left(z_j'\widehat{\gamma}, z_i'\widehat{\gamma} \right) \psi_j + o_p\left(h_1^2 + \frac{1}{\sqrt{nh}} \right) \\ &= h_1^2 \Xi_n(z_i'\gamma) + \frac{1}{\sqrt{nh_1}} \sum_{j=1}^{n} \omega\left(z_j'\gamma_0, z_i'\gamma_0 \right) \psi_j + o_p\left(h_1^2 + \frac{1}{\sqrt{nh}} \right) \end{aligned} \tag{36}$$

as $\widehat{\gamma} - \gamma_0 = O_p\left(\frac{1}{\sqrt{n}} \right)$, and $\frac{1}{\sqrt{nh_1}} \sum_{j=1}^{n} \omega\left(z_j'\widehat{\gamma}, z_i'\widehat{\gamma} \right) \psi_j$ satisfies a CLT. Furthermore, recalling that $\widehat{\beta} - \beta_0 = O_p\left(\frac{1}{\sqrt{n}} \right)$,

$$S_{n,1}(b_0(1), b_1(1), ...b_q(1)) - \overline{S}_{n,1}(b_0, b_1, ...b_q) = o_p(1)$$

uniformly in $b_0, ..., b_q$. Then, as a straightforward consequence of the uniform law of large numbers, $\widetilde{b}_0(1) - \widehat{b}_0(1) = o_p(1)$ and so $\widehat{b}_0(1) - b_0(1) = o_p(1)$.

We now need to show that

$$\frac{\partial S_{n,1}(b_0(1), b_1(1), ..., b_q(1))}{\partial b'} - \frac{\partial \overline{S}_{n,1}(b_0(1), b_1(1), ..., b_q(1))}{\partial b'} = o_p\left(\frac{1}{\sqrt{nh_2}} \right) \tag{37}$$

Without loss of generality we study only the first element of the vector in (37). Now,

$$\begin{aligned}
& \sqrt{nh_2}\left(\frac{\partial S_{n,1}\left(b_0(1),b_1(1),...,b_q(1)\right)}{\partial b_0}-\frac{\partial S_n\left(b_0(1),b_1(1),...,b_q(1)\right)}{\partial b_0}\right) \\
= & \frac{1}{\sqrt{nh_2}}\sum_{i=1}^{n}s_i\left(y_i-b_0(1)-x_{1i}'\theta_1-\frac{1}{t!}\sum_{t=1}^{q}b_t(1)(p_i-1)^t\right)k\left(\frac{p_i-1}{h_2}\right) \\
& -\frac{1}{nh_2}\sum_{i=1}^{n}s_i\left(y_i-b_0^{\dagger}(1)-x_{1i}'\widehat{\theta}_1-\frac{1}{t!}\sum_{t=1}^{q}b_t^{\dagger}(1)(\widehat{p}_i-1)^t\right)k\left(\frac{\widehat{p}_i-1}{h_2}\right)
\end{aligned}$$

and so

$$\begin{aligned}
& \sqrt{nh_2}\left(\frac{\partial S_{n,1}\left(b_0(1),b_1(1),...,b_q(1)\right)}{\partial b_0}-\frac{\partial \overline{S}_n\left(b_0(1),b_1(1),...,b_q(1)\right)}{\partial b_0}\right) \\
= & \frac{1}{\sqrt{nh_2}}\sum_{i=1}^{n}s_i\left(y_i-b_0(1)-x_i'\beta_0-\frac{1}{t!}\sum_{t=1}^{q}b_t^{\dagger}(1)(p_i-1)^t\right)\left(k\left(\frac{p_i-1}{h_2}\right)-k\left(\frac{\widehat{p}_i-1}{h_2}\right)\right) \\
& -\frac{1}{\sqrt{nh_2}}\sum_{i=1}^{n}s_ib_1(1)\left(\widehat{p}_i-p_i\right)k\left(\frac{p_i-1}{h_2}\right)\left(1+o_p(1)\right) \\
& +\frac{1}{\sqrt{nh_2}}\sum_{i=1}^{n}s_i\left(y_i-b_0(1)-x_i'\beta_0-b_1(1)\left(\widehat{p}_i-p_i\right)\right)\left(k\left(\frac{p_i-1}{h_2}\right)-k\left(\frac{\widehat{p}_i-1}{h_2}\right)\right)\left(1+o_p(1)\right) \\
= & I_n+II_n+III_n
\end{aligned}$$

where the $(1+o_p(1))$ comes from the fact that for all $j_1,j_2\geq 1$ and $j_1+j_2\geq 2$, $(1-p_i)^{j_1}\left(\widehat{p}_i-p_i\right)^{j_2}$ is of smaller order than $\left(\widehat{p}_i-p_i\right)$.

$$\begin{aligned}
& I_n \\
= & \frac{h_1^2}{\sqrt{nh_2}h_2}\sum_{i=1}^{n}s_iu_ik'\left(\frac{p_i-1}{h_2}\right)\Xi_n(z_i'\gamma_0)\left(1+o_p(1)\right) \\
& +\frac{1}{\sqrt{nh_2}h_2}\frac{1}{nh_1}\sum_{i=1}^{n}\sum_{j=1}^{n}s_iu_ik^{(1)}\left(\frac{p_i-1}{h_2}\right)\omega\left(z_j'\gamma_0,z_i'\gamma_0\right)\psi_j\left(1+o_p(1)\right) \\
= & I_{A,n}+I_{B,n},
\end{aligned}$$

where $u_i=y_i-b_0(1)-x_i'\beta_0-\frac{1}{t!}\sum_{t=1}^{q}b_t(1)(p_i-1)^t$. Since:

$$\frac{1}{nh_2^2}\sum_{i=1}^{n}s_iu_iK^{(1)}\left(\frac{p_i-1}{h_2}\right)\Xi_n(z_i'\gamma_0)$$

is of order $O_p(1)$ because of the law of large numbers, $I_{A,n}=O_p\left(\sqrt{n}h_2^{1/2}h_1^2\right)=o_p(1)$ for $nh_2h_1^4\to 0$.

As for $I_{B,n}$:

$$\begin{aligned} & I_{B,n} \\ = & \frac{1}{\sqrt{nh_2}h_2}\frac{1}{nh_1}\sum_{i=1}^{n} s_i u_i k'\left(\frac{p_i-1}{h_2}\right)\omega\left(0\right)\psi_i \\ & +\frac{1}{\sqrt{nh_2}h_2}\frac{1}{nh_1}\sum_{i=1}^{n}\sum_{j>i}^{n} s_i u_i k^{(1)}\left(\frac{p_i-1}{h_2}\right)\omega\left(z_j'\gamma_0, z_i'\gamma_0\right)\psi_j \\ & +\frac{1}{\sqrt{nh_2}h_2}\frac{1}{nh_1}\sum_{i=1}^{n}\sum_{j>i}^{n} s_j u_j k^{(1)}\left(\frac{p_j-1}{h_2}\right)\omega\left(z_j'\gamma_0, z_i'\gamma_0\right)\psi_i, \end{aligned} \tag{38}$$

the first term on the RHS of on the RHS of (38) is $O_p\left(\frac{h_2^{1/2}}{\sqrt{n}h_1}\right) = o_p(1)$ for $nh_1^2h_2^{-1} \to \infty$.

By noting that

$$\begin{aligned} & \frac{1}{h_2^{3/2}h_1}E\left(s_i\psi_j k^{(1)}\left(\frac{p_i-1}{h_2}\right)\omega\left(z_j'\gamma_0, z_i'\gamma_0\right)\left|\psi_i, p_i, z_i'\gamma_0\right.\right) \\ = & \frac{1}{h_2^{3/2}h_1}s_i k^{(1)}\left(\frac{p_i-1}{h_2}\right)E\left(psi_j\omega\left(z_j'\gamma_0, z_i'\gamma_0\right)\left|\psi_i, p_i, z_i'\gamma_0\right.\right) = 0 \end{aligned}$$

and

$$\begin{aligned} & \frac{1}{h_2^{3/2}h_1}E\left(s_j\psi_i u_j k^{(1)}\left(\frac{p_j-1}{h_2}\right)\omega\left(z_j'\gamma_0, z_i'\gamma_0\right)\left|\psi_i, p_i, z_i'\gamma_0\right.\right) \\ = & \frac{1}{h_2^{3/2}h_1}\psi_i E\left(s_j u_j k^{(1)}\left(\frac{p_j-1}{h_2}\right)\omega\left(z_j'\gamma_0, z_i'\gamma_0\right)\left|\psi_i, p_i, z_i'\gamma_0\right.\right) \\ = & O_p\left(\sqrt{h_2}\right), \end{aligned}$$

because of integration by parts and change of variable. It does follows the sum of the last two terms on the RHS of (38) is $o_p(1)$.

As for II_n,

$$\begin{aligned} & II_n \\ = & \frac{h_1^2}{\sqrt{nh_2}}\sum_{i=1}^{n} s_i b_1(1) k\left(\frac{p_i-1}{h_2}\right)\Xi_n\left(z_i'\gamma_0\right) \\ & \underbrace{+\frac{1}{\sqrt{nh_2}}\frac{1}{nh_1}\sum_{i=1}^{n}\sum_{j>i}^{n} s_i\psi_j k\left(\frac{p_i-1}{h_2}\right)\omega\left(z_j'\gamma_0, z_i'\gamma_0\right)}_{II_{B,n}} + o_p(1), \end{aligned} \tag{39}$$

where the $o_p(1)$ term just captures parametric estimation error, due to $\widehat{\gamma} - \gamma_0$. The first term on the RHS of (39) is $O_p\left(\sqrt{n}h_1^4h_2^{1/2}\right)$ and so is $o_p(1)$ for $nh_2h_1^4 \to 0$. As for the second term on the RHS of

(39),

$$
\begin{aligned}
& II_{B,n} \\
= \; & \frac{1}{\sqrt{nh_2}}\frac{1}{nh_1}\sum_{i=1}^{n} s_i\psi_i k\left(\frac{p_i-1}{h_2}\right)k(0) \\
& +\frac{1}{\sqrt{nh_2}}\frac{1}{nh_1}\sum_{i=1}^{n}\sum_{j>i} s_i\psi_j k\left(\frac{p_i-1}{h_2}\right)\omega\left(z_j'\gamma_0, z_i'\gamma_0\right) \\
& +\frac{1}{\sqrt{nh_2}}\frac{1}{nh_1}\sum_{i=1}^{n}\sum_{j>i} s_j\psi_i k\left(\frac{p_j-1}{h_2}\right)\omega\left(z_j'\gamma_0, z_i'\gamma_0\right). \qquad (40)
\end{aligned}
$$

The first term on the RHS of (40) is $O_p\left(\frac{1}{\sqrt{n}h_2^{-1/2}h_1}\right) = o_p(1)$ for $nh_1^2h_2^{-1}\to\infty$. Also,

$$
\begin{aligned}
& \frac{1}{h_1\sqrt{h_2}}E\left(s_i\psi_j k\left(\frac{p_i-1}{h_2}\right)\omega\left(z_j'\gamma_0, z_i'\gamma_0\right)\left|\psi_i, p_i, z_i'\gamma_0\right.\right) \\
= \; & \frac{1}{h_1\sqrt{h_2}}s_i k\left(\frac{p_i-1}{h_2}\right)E\left(s_i\psi_j k\left(\frac{z_j'\gamma_0 - z_i'\gamma_0}{h_1}\right)\left|\psi_i, p_i, z_i'\gamma_0\right.\right) = 0
\end{aligned}
$$

and

$$
\begin{aligned}
& \frac{1}{h_1\sqrt{h_2}}E\left(s_j\psi_i k\left(\frac{p_j-1}{h_2}\right)\omega\left(z_j'\gamma_0, z_i'\gamma_0\right)\left|\psi_i, p_i, z_i'\gamma_0\right.\right) \\
= \; & \frac{1}{h_1\sqrt{h_2}}\psi_i E\left(s_i\omega\left(z_j'\gamma_0, z_i'\gamma_0\right)k\left(\frac{p_j-1}{h_2}\right)\left|\psi_i, p_i, z_i'\gamma_0\right.\right) \\
= \; & O_p\left(\sqrt{h_2}\right),
\end{aligned}
$$

because of integration by parts and change of variable. It does follows the sum of the last two terms on the RHS of (40) is $o_p(1)$.

Finally, it is immediate to see that III_n cannot be of larger order than $\max\{I_n, II_n\}$. ■

Proof of Theorem E2. To prove (i), define:

$$
S_n(\mathbf{b}) \equiv \frac{1}{n}\sum_{i=1}^{n} s_i\left(y_i - g\left(x_{1i}'\mathbf{b}\right) - \left(\widehat{m}(\widehat{w}_i) - \widehat{m}(\widehat{w}_{i1})\right)\right)^2
$$

and:

$$
S_0(\mathbf{b}) \equiv \mathrm{E}^*\left[\left(y_i - g\left(x_{1i}'\mathbf{b}\right) - \left(\widehat{m}(\widehat{w}_i) - \widehat{m}(\widehat{w}_{i1})\right)\right)^2\right],
$$

recalling that, with slight abuse of notation, we used $\widehat{w}_{i1}$ in $\widehat{m}(\widehat{w}_{i1})$ to highlight that $m(w_{i1})$ is estimated using observations $\widehat{w}_j$, $j = 1, \ldots, n$. We prove this part of Theorem E2 verifying the assumptions of Theorem 2.1 in Newey and Mc Fadden (1994). Given the identification results from before and Assumptions E8 and E12 above, it remains to show that the right hand side (RHS) of the following inequality is of order $o_p(1)$:

$$
\sup_{\mathbf{b}}\left|S_n(\mathbf{b}) - S_0(\mathbf{b})\right| \leq \sup_{\mathbf{b}}\left|S_n(\mathbf{b}) - \overline{S}_n(\mathbf{b})\right| + \sup_{\mathbf{b}}\left|\overline{S}_n(\mathbf{b}) - S_0(\mathbf{b})\right|,
$$

where

$$\overline{S}_n(\mathbf{b}) \equiv \frac{1}{n}\sum_{i=1}^{n} s_i \left(y_i - g\left(x_{1i}'\mathbf{b}\right) - (m(w_i) - m(w_{i1}))\right)^2.$$

Recalling that $\widehat{\lambda}(\widehat{p}_i) = \widehat{m}(\widehat{w}_i) - \widehat{m}(\widehat{w}_{i1})$ and $\lambda(p_i) = m(w_i) - m(w_{i1})$, first note that:

$$\left|S_n(\mathbf{b}) - \overline{S}_n(\mathbf{b})\right| = \left|\frac{1}{n}\sum_{i=1}^{n} s_i \left\{2\left(y_i - g\Big(x_{1i}'\mathbf{b}\Big)\right) + \widehat{\lambda}(\widehat{p}_i) + \lambda(p_i)\right\}\left\{\widehat{\lambda}(\widehat{p}_i) - \lambda(p_i)\right\}\right|$$

Using Assumptions E8 through E13 and noting that $\Pr\left(\widehat{p} \in \mathcal{P}\right) \longrightarrow 1$, it is straightforward to show that $\left(\widehat{\lambda}(\widehat{p}_i) + \lambda(p_i)\right)\left(\widehat{\lambda}(\widehat{p}_i) - \lambda(p_i)\right)$ on the RHS of the last equality can be bounded by:

$$\begin{aligned} & C_n \sup_w |\widehat{m}(w) - m(w)| \times \left|\frac{1}{n}\sum_{i=1}^{n} s_i\right| \\ = \;& o_p(1)O_p(1) = o_p(1), \end{aligned}$$

where C_n denotes a generic constant. The first part can be addressed by similar arguments (noting that $\left|\frac{1}{n}\sum_{i=1}^{n} s_i 2\left(y_i - g\Big(x_{1i}'\mathbf{b}\Big)\right)\right| = O_p(1)$ uniformly in $\mathbf{b}$ by Markov's inequality) to conclude that these terms are of order $o_p(1)$ as well. Finally, since:

$$\mathrm{E}^*\left[\left|\left(y_i - g\left(x_{1i}'\mathbf{b}\right) - (m(w_i) - m(w_{i1}))\right)^2\right|^2\right] < \infty,$$

by a uniform law of large numbers (e.g., Lemma 2.4 in Newey and Mc Fadden (1994)) we have that:

$$\sup_{\mathbf{b}} \left|\overline{S}_n(\mathbf{b}) - S_0(\mathbf{b})\right| = o_p(1).$$

This establishes the first claim.

(ii) For the second part of Theorem E2, note that the first order condition is given by:

$$\frac{2}{n}\sum_{i=1}^{n} s_i \left(y_i - g\left(x_{1i}'\widehat{\mathbf{b}}_A\right) - \left(\widehat{m}(\widehat{w}_i) - \widehat{m}(\widehat{w}_{i1})\right)\right) D_{\mathbf{b}}^{\mathbf{1}} g\left(x_{1i}'\widehat{\mathbf{b}}_A\right) = 0.$$

After a Taylor series expansion of $g\left(x_{1i}'\widehat{\mathbf{b}}_A\right)$ around $\mathbf{b}_0$ and re-arrangements, we obtain:

$$\sqrt{nh_2}\left(\widehat{\mathbf{b}}_A - \mathbf{b}_0\right) = \mathbf{G}_n^{-1}\left(\frac{\sqrt{nh_2}}{n}\sum_{i=1}^{n} s_i \left(y_i - g\left(x_{1i}'\mathbf{b}_0\right) - \left(\widehat{m}(\widehat{w}_i) - \widehat{m}(\widehat{w}_{i1})\right)\right)\right),$$

where

$$\mathbf{G}_n \equiv \frac{1}{n}\sum_{i=1}^{n} s_i \left[D_{\mathbf{b}}^{\mathbf{1}} g\left(x_{1i}'\overline{\mathbf{b}}_0\right)\left(D_{\mathbf{b}}^{\mathbf{1}} g\left(x_{1i}'\widehat{\mathbf{b}}_A\right)\right)'\right]$$

Using consistency of $\widehat{\mathbf{b}}_A$, a uniform law of large numbers together with continuous mapping yields:

$$\mathbf{G}_n \xrightarrow{p} \mathbf{\Sigma}_0,$$

where

$$\mathbf{\Sigma}_0 = \mathrm{E}\left[s_i D^{\mathbf{1}}_{\mathbf{b}} g\left(x'_{1i}\mathbf{b}_0\right)\left(D^{\mathbf{1}}_{\mathbf{b}} g\left(x'_{1i}\mathbf{b}_0\right)\right)'\right].$$

Next, we examine the limiting behavior of

$$\begin{aligned}
&\frac{\sqrt{nh_2}}{n}\sum_{i=1}^{n} s_i\left(y_i - g\left(x'_{1i}\mathbf{b}_0\right) - \left(\widehat{m}(\widehat{w}_i) - \widehat{m}(\widehat{w}_{i1})\right)\right) \\
= \;& \frac{\sqrt{nh_2}}{n}\sum_{i=1}^{n} s_i\left(u_i + \left(m(w_i) - m(w_{i1})\right) - \left(\widehat{m}(\widehat{w}_i) - \widehat{m}(\widehat{w}_{i1})\right)\right) + o_p(1),
\end{aligned}$$

where the equality follows from the definition of $\lambda(p_i)$. The final line can be analyzed through Lemmas B.1, B.2, B.3, B.4, and B.5. This establishes the claim. ■

Proof of Theorem E3. (i) Starting with the first claim, that that as

$$\widetilde{S}_n(\mathbf{b}) \equiv \frac{1}{n}\sum_{i=1}^{n} s_i\left(y_i\frac{\widehat{m}(\widehat{w}_{i1})}{\widehat{m}(\widehat{w}_i)} - \widetilde{g}\left(x'_{1i}\mathbf{b}\right)\right)^2$$

and:

$$\widetilde{S}_0(\mathbf{b}) \equiv \mathrm{E}^*\left[\left(y_i\frac{m(w_{i1})}{m(w_i)} - \widetilde{g}\left(x'_{1i}\mathbf{b}\right)\right)^2\right].$$

We prove this theorem verifying the assumptions of Theorem 2.1 in Newey and Mc Fadden (1994). Given the identification result in Theorem E2 and Assumption E8 above, it remains to show that the RHS of the following inequality is of smaller order in probability:

$$\sup_{\mathbf{b}}\left|\widetilde{S}_n(\mathbf{b}) - \widetilde{S}_0(\mathbf{b})\right| \leq \sup_{\mathbf{b}}\left|\widetilde{S}_n(\theta) - \widetilde{\widetilde{S}}_n(\theta)\right| + \sup_{\mathbf{b}}\left|\widetilde{\widetilde{S}}_n(\mathbf{b}) - \widetilde{S}_0(\mathbf{b})\right|,$$

where

$$\widetilde{\widetilde{S}}_n(\mathbf{b}) \equiv \frac{1}{n}\sum_{i=1}^{n} s_i\left(y_i\frac{m(w_{i1})}{m(w_i)} - \widetilde{g}\left(x'_{1i}\mathbf{b}\right)\right)^2.$$

First note that:

$$\begin{aligned}
\left|\widetilde{S}_n(\mathbf{b}) - \widetilde{\widetilde{S}}_n(\mathbf{b})\right| \;=\; & \Bigg|\frac{1}{n}\sum_{i=1}^{n} s_i\Bigg\{\Bigg\{y_i^2\left(\left(\frac{\widehat{m}(\widehat{w}_{i1})}{\widehat{m}(\widehat{w}_i)}\right)^2 - \left(\frac{m(w_{i1})}{m(w_i)}\right)^2\right)\Bigg\} \\
& + \Bigg\{2y_i\left(\widetilde{g}\left(x'_{1i}\mathbf{b}\right)\frac{m(w_{i1})}{m(w_i)} - \widetilde{g}\left(x'_{1i}\mathbf{b}\right)\frac{\widehat{m}(\widehat{w}_{i1})}{\widehat{m}(\widehat{w}_i)}\right)\Bigg\}\Bigg\}\Bigg|.
\end{aligned} \tag{41}$$

Using Assumptions E8 and E13, the first term on the RHS can be bounded by:

$$\begin{aligned}
& \sup_w\left|\left(\left(\frac{\widehat{m}(w_1)}{\widehat{m}(w)}\right) - \left(\frac{m(w_1)}{m(w)}\right)\right)\left(\left(\frac{\widehat{m}(w_1)}{\widehat{m}(w)}\right) + \left(\frac{m(w_1)}{m(w)}\right)\right)\right| \times \left|\frac{1}{n}\sum_{i=1}^{n} s_i y_i^2\right| \\
\leq \;& C_n\frac{\sup_w|\widehat{m}(w) - m(w)|}{\inf_w|\widehat{m}(w)|} \times \left|\frac{1}{n}\sum_{i=1}^{n} s_i y_i^2\right| \\
= \;& o_p(1)O_p(1) = o_p(1),
\end{aligned}$$

where C_n is again a generic constant. For the second term on the RHS of Equation (41), similar arguments can be used to show that it is of order $o_p(1)$ as well. The remaining steps are as in the proof of Theorem E2.

(ii) For the second part, note that the first order condition is given by:

$$\frac{2}{n}\sum_{i=1}^{n} s_i \left(y_i \left(\frac{\widehat{m}(\widehat{w}_{i1})}{\widehat{m}(\widehat{w}_i)} \right) - \widetilde{g}\left(x_{1i}' \widehat{\mathbf{b}}_M \right) \right) D_{\mathbf{b}}^{\mathbf{1}} \widetilde{g}\left(x_{1i}' \widehat{\mathbf{b}}_M \right) = 0,$$

After a Taylor series expansion of $\widetilde{g}\left(x_{1i}' \widehat{\mathbf{b}}_M \right)$ around $\mathbf{b}_0$ and re-arrangements, we obtain:

$$\sqrt{nh_2}\left(\widehat{\mathbf{b}}_M - \mathbf{b}_0 \right) = \widetilde{\mathbf{G}}_n^{-1} \left(\frac{\sqrt{nh_2}}{n} \sum_{i=1}^{n} s_i \left(y_i \left(\frac{\widehat{m}(\widehat{w}_{i1})}{\widehat{m}(\widehat{w}_i)} \right) - \widetilde{g}\left(x_{1i}' \mathbf{b}_0 \right) \right) \right),$$

where

$$\widetilde{\mathbf{G}}_n \equiv \frac{1}{n}\sum_{i=1}^{n} s_i \left[D_{\mathbf{b}}^{\mathbf{1}} \widetilde{g}\left(x_{1i}' \overline{\mathbf{b}}_0 \right) \left(D_{\mathbf{b}}^{1} \widetilde{g}\left(\widehat{\mathbf{b}}_M \right) \right)' \right]$$

Using again consistency of $\widehat{\mathbf{b}}_M$, a uniform law of large numbers together with continuous mapping yields:

$$\widetilde{\mathbf{G}}_n \xrightarrow{p} \widetilde{\mathbf{\Sigma}}_0,$$

where

$$\widetilde{\mathbf{\Sigma}}_0 = \mathrm{E}\left[s_i D_{\mathbf{b}}^{\mathbf{1}} \widetilde{g}\left(x_{1i}' \mathbf{b}_0 \right) \left(D_{\mathbf{b}}^{1} \widetilde{g}\left(\theta_0 \right) \right)' \right].$$

Finally, the limiting behavior of

$$\begin{aligned} & \frac{\sqrt{nh_2}}{n} \sum_{i=1}^{n} s_i \left(y_i \left(\frac{\widehat{m}(\widehat{w}_{i1})}{\widehat{m}(\widehat{w}_i)} \right) - \widetilde{g}\left(x_{1i}' \mathbf{b}_0 \right) \right) \\ = & \frac{\sqrt{nh_2}}{n} \sum_{i=1}^{n} s_i \left(\widetilde{u}_i \frac{m(w_{i1})}{m(w_i)} + y_i \left(\frac{\widehat{m}(\widehat{w}_{i1})}{\widehat{m}(\widehat{w}_i)} - \frac{m(w_{i1})}{m(w_i)} \right) \right) + o_p(1), \end{aligned}$$

where the equality follows from the definition of $\widetilde{\lambda}(p_i) = m(w_i)/m(w_{i1})$. Now:

$$\begin{aligned} & \frac{\sqrt{nh_2}}{n} \sum_{i=1}^{n} s_i \left(\widetilde{u}_i \frac{m(w_{i1})}{m(w_i)} + y_i \left(\frac{\widehat{m}(\widehat{w}_{i1})}{\widehat{m}(\widehat{w}_i)} - \frac{m(w_{i1})}{m(w_i)} \right) \right) \\ = & \frac{\sqrt{nh_2}}{n} \sum_{i=1}^{n} s_i \left(\widetilde{u}_i \frac{m(w_{i1})}{m(w_i)} + y_i \left(\frac{\widehat{m}(\widehat{w})_{1i}) - m(w_{i1})}{\widehat{m}(\widehat{w}_i)} - \frac{\widehat{m}(\widehat{w}_i) - m(w_i)}{\widehat{m}(\widehat{w}_i)} \frac{m(w_{i1})}{m(w_i)} \right) \right) \\ = & \frac{\sqrt{nh_2}}{n} \sum_{i=1}^{n} s_i \left(\widetilde{u}_i \frac{m(w_{i1})}{m(w_i)} + y_i \left(\frac{\widehat{m}(\widehat{w}_{i1}) - m(w_{i1})}{m(w_i)} - \frac{\widehat{m}(\widehat{w}_i) - m(w_i)}{m(w_i)} \frac{m(w_{i1})}{m(w_i)} \right) \right) + o_p(1) \\ = & \frac{\sqrt{nh_2}}{n} \sum_{i=1}^{n} s_i \left(\widetilde{u}_i \frac{m(w_{i1})}{m(w_i)} + (1 + \frac{\widetilde{u}_i}{m(w_i)}) \left((\widehat{m}(\widehat{w}_{i1}) - m(w_{i1})) - (\widehat{m}(\widehat{w}_i) - m(w_i)) \frac{m(w_{i1})}{m(w_i)} \right) \right) + o_p(1), \end{aligned}$$

where the second equality follows by Assumptions E7, E9, and E12, the third equality by the uniform consistency of $\widehat{m}(\widehat{w}_i)$, and the last equality again from the definition of $\widetilde{\lambda}(p_i)$. Note that as in the proof of Theorem E2, the leading term is:

$$\frac{\sqrt{nh_2}}{n}\sum_{i=1}^{n} s_i\left(\widehat{m}(\widehat{w}_{i1}) - m(w_{i1})\right) = O_p(1),$$

which can be analyzed by means of Lemma B.3 and B.5 replacing u_i with $\widetilde{u}_i$ in the proofs. The same holds for:

$$\frac{\sqrt{nh_2}}{n}\sum_{i=1}^{n} s_i\left(\widetilde{u}_i\frac{m(w_{i1})}{m(w_i)} - \left(\widehat{m}(\widehat{w}_i) - m(w_i)\right)\frac{m(w_{i1})}{m(w_i)}\right)$$

which can be analyzed through Lemma B.1, B.2, and B.4, respectively, noting that $\mathrm{E}[\widetilde{u}_i|w_i] = 0$ and that:

$$\frac{m(w_{i1})}{m(w_i)}$$

is bounded away from zero a.s. by Assumption E12. The remaining terms are analyzed in Lemma B.6, B.7, B.8, and B.9. This establishes the claim. ■

B.5 Proofs of Auxiliary Lemmas

Proof of Lemma B.1. Note that by iterated expectations:

$$\mathrm{E}\left[s_i u_i\right] = 0.$$

Also, since $\mathrm{E}\left[s_i\,\mathrm{E}^*\left[u_i^2|w_i\right]\right] < \infty$ by Assumption E8, it holds that:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_i u_i = O_p\left(1\right),$$

which establishes the claim. ■

Proof of Lemma B.2. Using the representation from Equation (21), note that:

$$\begin{aligned}
&\frac{\sqrt{nh_2}}{n}\sum_{i=1}^{n} s_i(\widehat{m}(\widehat{w}_i) - m(\widehat{w}_i)) \qquad (42)\\
= \;& \frac{\sqrt{nh_2}}{n}\sum_{i=1}^{n} s_i\iota_1'\widehat{\mathbf{M}}_n^{-1}(\widehat{w}_i)\left\{\frac{1}{nh_2^{d_x+1}}\sum_{j=1}^{n}\widehat{\mathcal{K}}_j(\widehat{w}_i)u_j\right\} + \frac{\sqrt{nh_2}}{n}\sum_{i=1}^{n} s_i\iota_1'\widehat{\mathbf{M}}_n^{-1}(\widehat{w}_i)\left\{\frac{1}{nh_2^{d_x+1}}\sum_{j=1}^{n}\widehat{\mathcal{K}}_j(\widehat{w}_i)\widehat{\Delta}_{m,j}\right\},
\end{aligned}$$

where

$$\widehat{\Delta}_{m,j} = m(\widehat{w}_j) - \sum_{0\le|\mathbf{k}|\le q}\frac{1}{\mathbf{k}!}D^{\mathbf{k}}m(\widehat{w}_i)(\widehat{w}_j - \widehat{w}_i)^{\mathbf{k}}.$$

Now, define:

$$\widehat{V}_{m,n}(\widehat{w}_i) \equiv \frac{1}{nh_2^{d_x+1}}\sum_{j=1}^{n}\widehat{\mathcal{K}}_j(\widehat{w}_i)u_j$$

and

$$\widehat{B}_{m,n}(\widehat{w}_i) \equiv \frac{1}{nh_2^{d_x+1}} \sum_{j=1}^{n} \widehat{\mathcal{K}}_j(\widehat{w}_i)\widehat{\Delta}_{m,j}.$$

Thus, noting that for any two symmetric non-singular matrices A_1 and A_2, we have that $A_1^{-1} - A_2^{-1} = A_2^{-1}(A_2 - A_1)A_1^{-1}$, Equation (42) can be re-stated as:

$$\begin{aligned}
& \frac{\sqrt{nh_2}}{n} \sum_{i=1}^{n} s_i \iota_1' \widehat{\mathbf{M}}_n^{-1}(\widehat{w}_i) \left\{ \widehat{V}_{m,n}(\widehat{w}_i) + \widehat{B}_{m,n}(\widehat{w}_i) \right\} \\
= \quad & \frac{\sqrt{nh_2}}{n} \sum_{i=1}^{n} s_i \iota_1' \left[f(\widehat{w}_i)\mathbf{M} \right]^{-1} \widehat{V}_{m,n}(\widehat{w}_i) \\
& + \frac{\sqrt{nh_2}}{n} \sum_{i=1}^{n} s_i \iota_1' \left[f(\widehat{w}_i)\mathbf{M} \right]^{-1} \widehat{B}_{m,n}(\widehat{w}_i) \qquad (43) \\
& - \frac{\sqrt{nh_2}}{n} \sum_{i=1}^{n} s_i \iota_1' \left[\mathbf{M} f(\widehat{w}_i) \right]^{-1} \left[\widehat{\mathbf{M}}_n(\widehat{w}_i) - \mathbf{M} f(\widehat{w}_i) \right] \widehat{\mathbf{M}}_n^{-1}(\widehat{w}_i) \widehat{V}_{m,n}(\widehat{w}_i) \\
& - \frac{\sqrt{nh_2}}{n} \sum_{i=1}^{n} s_i \iota_1' \left[\mathbf{M}(\widehat{w}_i) \right]^{-1} \left[\widehat{\mathbf{M}}_n f(\widehat{w}_i) - \mathbf{M} f(w_i) \right] \widehat{\mathbf{M}}_n^{-1}(\widehat{w}_i) \widehat{B}_{m,n}(\widehat{w}_i) \\
\equiv \quad & \mathcal{T}_{m,n,1} + \mathcal{T}_{m,n,2} + \mathcal{T}_{m,n,3} + \mathcal{T}_{m,n,4},
\end{aligned}$$

Starting with $\mathcal{T}_{m,n,1}$, by Assumption E10 and the fact that $\max_{1 \le i \le n} |\widehat{p}_i - p_i| = o_p(1)$, it holds that:

$$\left[f(\widehat{w}_i)\mathbf{M} \right]^{-1} = \left[f(w_i)\mathbf{M} \right]^{-1} \{1 + o_p(1)\}$$

uniformly in $\mathbf{b}$. Then, we may re-write:

$$\widehat{V}_{m,n}(\widehat{w}_i) = \frac{1}{nh_2^{d_x+1}} \sum_{j=1}^{n} \mathcal{K}_j(w_i) u_j + \frac{1}{nh_2^{d_x+1}} \sum_{j=1}^{n} \left\{ \widehat{\mathcal{K}}_j(\widehat{w}_i) - \mathcal{K}_j(w_i) \right\} u_j$$

Inserting this into $\mathcal{T}_{m,n,1}$ yields:

$$\begin{aligned}
& \frac{\sqrt{nh_2}}{n} \sum_{i=1}^{n} s_i \iota_1' \left[f(\widehat{w}_i)\mathbf{M} \right]^{-1} V_{m,n}(w_i) \\
& + \frac{\sqrt{nh_2}}{n} \sum_{i=1}^{n} s_i \iota_1' \left[f(\widehat{w}_i)\mathbf{M} \right]^{-1} \left\{ \widehat{V}_{m,n}(\widehat{w}_i) - V_{m,n}(w_i) \right\} \\
\equiv \quad & \mathcal{T}_{m,n,11} + \mathcal{T}_{m,n,12}
\end{aligned}$$

We start with $\mathcal{T}_{m,n,11}$. First note that by the same argument as before we have that:

$$\frac{\sqrt{nh_2}}{n} \sum_{i=1}^{n} s_i \iota_1' \left[f(w_i)\mathbf{M} \right]^{-1} V_{m,n}(w_i) + o_p(1)$$

Define the symmetrized 'kernel function':

$$\Psi_{u,ij} \equiv s_i \iota_1' \left[f(w_i)\mathbf{M}\right]^{-1} \mathcal{K}_j(w_i) u_j + s_j \iota_1' \left[f(w_j)\mathbf{M}\right]^{-1} \mathcal{K}_i(w_j) u_i$$

Then, using the same argument of Ahn and Powell (1993, p.25) to show that by Assumptions E9, E10, and $nh_2^{2d_x+3} \longrightarrow \infty$, terms with $i = j$ are asymptotically bounded by:

$$\frac{1}{n^2 h_2^{d_x+1}} \sum_{i=1}^{n} \left\| s_i \iota_1' \left[f(w_i)\mathbf{M}\right]^{-1} \mathcal{K}(0) u_i \right\| = O_p\left(\frac{1}{nh_2^{d_x+1}}\right) = o_p\left(\frac{1}{\sqrt{nh_2}}\right),$$

we can derive the (appropriately rescaled) U-statistics:

$$\begin{aligned} & \frac{\sqrt{nh_2}}{n} \sum_{i=1}^{n} s_i \iota_1' \left[f(\widehat{w}_i)\mathbf{M}\right]^{-1} V_{m,n}(w_i) \\ = \; & \frac{\sqrt{nh_2}}{2h_2^{d_x+1}} \binom{n}{2}^{-1} \sum_{i=1}^{n} \sum_{j>i} \Psi_{u,ij} + o_p(1) \\ = \; & \frac{\sqrt{nh_2}}{nh_2^{d_x+1}} \sum_{i=1}^{n} \mathrm{E}[\Psi_{u,ij} | \varpi_i] + o_p(1), \end{aligned}$$

where the second equality follows from a standard Hoeffding decomposition,the fact that $\mathrm{E}[\Psi_{u,ij}] = 0$ by iterated expectations, and from Lemma 3.1 in Powell et al. (1989) since $\mathrm{E}\left[h_2^{-2(d_x+1)} \left\|\Psi_{u,ij}\right\|^2\right] = O(h_2^{-d_x-1}) = o(n)$. Moreover, after change of variables and iterated expectations noting that $\mathrm{E}^*[u_i^2 | w_i] = \sigma^{*2}(w_i)$, we obtain:

$$\mathrm{E}\left[\left\| \mathrm{E}\left[\frac{1}{h_2^2}\Psi_{u,ij} \middle| \varpi_i\right] \right\|^2\right] = \mathrm{E}\left[s_i \sigma^{*2}(w_i) f(w_i)^{-1}\right] \iota' \mathbf{M}^{-1} \mathbf{M}_0 \mathbf{M}_0' \mathbf{M}^{-1\prime} \iota_1.$$

where $\mathbf{M}_0 = \int \mathcal{K}(\nu) d\nu$ is the probability limit of $\frac{1}{nh_2^{d_x+1}} \sum_{j=1}^{n} \mathcal{K}_j(w_i)$, see Section B.1, and thus:

$$\frac{\sqrt{h_2}}{\sqrt{n} h_2^{d_x+1}} \sum_{i=1}^{n} \mathrm{E}[\Psi_{u,ij} | \varpi_i] = O_p(\sqrt{h_2}).$$

uniformly in $\mathbf{b}$ by Lindeberg-Levy CLT. Next, we examine $\mathcal{T}_{m,n,12}$ and note that with slight abuse of notation:

$$\begin{aligned} & \frac{1}{nh^{d+1}} \sum_{j=1}^{n} \left\{ \widehat{\mathcal{K}}_j(\widehat{w}_i) - \mathcal{K}_j(w_i) \right\} u_j \\ = \; & \frac{1}{nh^{d+2}} \sum_{j=1}^{n} \overline{\mathcal{K}}^{(1)}_{j,d_x+1}(\overline{w}_i) \left\{ (p_i - \widehat{p}_i) + (\widehat{p}_j - p_j) \right\} u_j, \end{aligned}$$

where $\overline{w}_i$ denotes an observation on the line segment between w_i and $\widehat{w}_i$. In what follows, we will only consider the part which involves $(p_i - \widehat{p}_i)$, the one with $(\widehat{p}_j - p_j)$ follows by analogous arguments.

Thus, using the linear representation from Assumption E7, we have for the first part of $\mathcal{T}_{m,n,12}$, say $\mathcal{T}_{m,n,121}$:

$$\begin{aligned}\mathcal{T}_{m,n,121} &= \frac{\sqrt{nh_2}}{nh_2^{d_x+2}}\sum_{i=1}^{n} s_i\iota_1'\left[f(\widehat{w}_i)\mathbf{M}\right]^{-1}\frac{1}{n}\sum_{j=1}^{n}\overline{\mathcal{K}}^{(1)}_{j,d_x+1}(\overline{w}_i)\left\{(p_i-\widehat{p}_i)u_j\right\}+o_p(1)\\ &= \frac{\sqrt{nh_2}}{n^2h_2^{d_x+2}}\sum_{i=1}^{n}\sum_{j=1}^{n} s_i\iota_1'\left[f(\widehat{w}_i)\mathbf{M}\right]^{-1}\overline{\mathcal{K}}^{(1)}_{j,d_x+1}(\overline{w}_i) \qquad (44)\\ &\quad\times\left\{\frac{1}{nh_1}\sum_{l=1}^{n}\omega(z_i'\widehat{\gamma},z_l'\widehat{\gamma})\psi_l+\Xi_n(z_i'\widehat{\gamma})+o_p\left(h_1^2+\frac{1}{\sqrt{nh_1}}\right)\right\}u_j+o_p(1)\end{aligned}$$

We will start with the term involving $\omega(z_i'\widehat{\gamma},z_l'\widehat{\gamma})\psi_l$. Since $\|\widehat{\gamma}-\gamma_0\|=O_p(n^{-\frac{1}{2}})$ by Assumption E7 and $nh_1\longrightarrow\infty$ by Assumption E13, note that:

$$\frac{1}{nh_1}\sum_{l=1}^{n}\omega(z_i'\widehat{\gamma},z_l'\widehat{\gamma})\psi_l+\Xi_n(z_i'\widehat{\gamma})+o_p\left(h_1^2+\frac{1}{\sqrt{nh_1}}\right)=\frac{1}{nh_1}\sum_{l=1}^{n}\omega(z_i'\gamma_0,z_l'\gamma_0)\psi_l+\Xi_n(z_i'\gamma_0)+o_p\left(h_1^2+\frac{1}{\sqrt{nh_1}}\right).$$

Moreover, since $f(\widehat{w}_i)\mathbf{M}=\{f(w_i)\mathbf{M}+o_p(1)\}$, and using another mean value expansion around p_i and p_j, we can write this expression as:

$$\mathcal{T}_{m,n,121}=\frac{\sqrt{nh_2}}{n^3h_2^{d_x+2}h_1}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n} s_i\iota_1'\left[f(w_i)\mathbf{M}\right]^{-1}\mathcal{K}^{(1)}_{j,d_x+1}(w_i)\omega(z_i'\gamma_0,z_l'\gamma_0)\psi_l u_j+o_p(1).$$

Also, note that asymptotically we can omit terms with $i=j$, $j=l$, and $j=l$ since e.g. terms with $i=j$ can be asymptotically bounded by:

$$\frac{1}{n^3h_2^{d_x+2}h_1}\sum_{i=1}^{n}\sum_{l\neq i}\left\|s_i\iota_1'\left[f(w_i)\mathbf{M}\right]^{-1}\mathcal{K}^{(1)}_{d_x+1}(0)\omega(z_i'\gamma_0,z_l'\gamma_0)\psi_l u_j\right\|=O_p\left(\frac{1}{nh_2^{d_x+2}}\right)=o_p\left(\frac{1}{\sqrt{nh_2}}\right),$$

where the last equality follows from Assumption E13. Next, we define the symmetrized kernel function for a third order U-statistic:

$$\Psi^V_{p,ijl}\equiv s_i\iota_1'\left[f(w_i)\mathbf{M}\right]^{-1}\mathcal{K}^{(1)}_{j,d_x+1}(w_i)\omega(z_i'\gamma_0,z_l'\gamma_0)\psi_l u_j+\ldots+s_j\iota_1'\left[f(w_j)\mathbf{M}\right]^{-1}\mathcal{K}^{(1)}_{i,d_x+1}(w_j)\omega(z_j'\gamma_0,z_l'\gamma_0)\psi_l u_i.$$

Note that by iterated expectations this kernel function has (unconditional) mean zero and

$$\mathrm{E}[h_2^{-2(d_x+2)}h_1^{-2}\|\Psi^V_{p,ijl}\|^2]=O(h_2^{-d_x-2}h_1^{-1})=o(n)$$

using change of variables and integration by parts arguments. By Lemma 3.1 in Powell et al. (1989) (see also Lemma A.3 in Ahn and Powell (1993)), this implies that the U-statistic is equal to its projection up to an approximation error of order $o_p(n^{-\frac{1}{2}})$. However, $h_2^{-d_x-2}h_1^{-1}\,\mathrm{E}\left[\Psi^V_{p,ijl}|\varpi_i\right]=0$ because $\mathrm{E}[\psi_l|z_l'\gamma_0]=0$ and $\mathrm{E}^*[u_i|w_i,z_i'\gamma_0]=\mathrm{E}^*[u_i|w_i,p_i]=0$ by iterated expectations. Therefore, it is straightforward to deduce that:

$$\frac{\sqrt{nh_2}}{3h_2^{d_x+2}h_1}\binom{n}{3}^{-1}\sum_{i=1}^{n}\sum_{j>i}\sum_{l>j}\Psi^V_{p,ijl}=o_p(\sqrt{h_2}).$$

For the second part on the right hand side (RHS) of the last equality of Equation (44), note that:

$$\begin{aligned} & \frac{\sqrt{nh_2}}{n^2 h_2^{d_x+2}} \sum_{i=1}^{n}\sum_{j=1}^{n} s_i \iota_1' \left[f(\widehat{w}_i)\mathbf{M}\right]^{-1} \overline{\mathcal{K}}^{(1)}_{j,d_x+1}(\overline{w}_i)\Xi_n(z_i'\gamma_0)u_j \\ \leq & \max_{1\leq i\leq n}\left|\Xi_n(z_i'\gamma_0)\right| \left\| \frac{\sqrt{nh_2}}{n^2 h_2^{d_x+2}} \sum_{i=1}^{n}\sum_{j=1}^{n} s_i \iota_1' \left[f(w_i)\mathbf{M}\right]^{-1} \mathcal{K}^{(1)}_{j,d_x+1}(w_i)u_j \right\| \end{aligned}$$

The first term on the RHS of the last inequality is of order $O_p(h_1^2)$ by Assumption E12. For the second part, note again that observations with $i = j$ can be neglected asymptotically, so that:

$$\begin{aligned} & \frac{\sqrt{nh_2}}{n^2 h_2^{d_x+2}} \sum_{i=1}^{n}\sum_{j=1}^{n} s_i \iota_1' \left[f(w_i)\mathbf{M}\right]^{-1} \mathcal{K}^{(1)}_{j,d_x+1}(w_i) \\ = & \frac{\sqrt{nh_2}}{2h_2^{d_x+2}} \binom{n}{2}^{-1} \sum_{i=1}^{n}\sum_{j>i} \left\{ s_i \iota_1' \left[f(w_i)\mathbf{M}\right]^{-1} \mathcal{K}^{(1)}_{j,d_x+1}(w_i)u_j + s_j \iota_1' \left[f(w_j)\mathbf{M}\right]^{-1} \mathcal{K}^{(1)}_{i,d_x+1}(w_j)u_i \right\} + o_p(1) \end{aligned}$$

Using standard Hoeffding decomposition, change of variables, and integration by parts arguments, it is not difficult to deduce that this term is of order $O_p(\sqrt{h_2})$ uniformly in $\mathbf{b}$. Given Assumption E12, we can therefore conclude that the whole term is of order $o_p(1)$.

Turning to $\mathcal{T}_{m,n,2}$ from Equation (43), define:

$$\Delta_{m,j} = m(w_i) - \sum_{0\leq|\mathbf{k}|\leq q} \frac{1}{\mathbf{k}!} D^{\mathbf{k}} m(w_i)(w_j - w_i)^{\mathbf{k}}.$$

in analogy to $\Delta_{m,j}$. Then, by Assumption E11, it holds that:

$$\Delta_{m,j} = \sum_{|\mathbf{k}|=q+1} \frac{1}{\mathbf{k}!} D^{\mathbf{k}} m(\overline{w})(w_j - w_i)^{\mathbf{k}}.$$

for some $\overline{w}$ in between w_j and w_i. Also:

$$\widehat{\Delta}_{m,j} = \sum_{|\mathbf{k}|=q+1} \frac{1}{\mathbf{k}!} D^{\mathbf{k}} m(\widehat{\overline{w}})(\widehat{w}_j - \widehat{w}_i)^{\mathbf{k}},$$

where $\widehat{\overline{w}}$ lies between $\widehat{w}_j$ and $\widehat{w}_i$. Since $|\Delta_{m,j}| = O_p(h_2^{q+1})$ uniformly in $\mathbf{b}$, i and j for $\|w_j - w_i\| \leq Ch_2$. Thus, it remains to analyze:

$$\begin{aligned} \widehat{\Delta}_{m,j} - \Delta_{m,j} = & \sum_{|\mathbf{k}|=q+1} \frac{1}{\mathbf{k}!} D^{\mathbf{k}} m(\widehat{\overline{w}}) \left((\widehat{w}_j - \widehat{w}_i)^{\mathbf{k}} - (w_j - w_i)^{\mathbf{k}} \right) \\ & + \sum_{|\mathbf{k}|=q+1} \frac{1}{\mathbf{k}!} \left[D^{\mathbf{k}} m(\widehat{\overline{w}}) - D^{\mathbf{k}} m(\overline{w}) \right] (w_j - w_i)^{\mathbf{k}}. \end{aligned}$$

By differentiability up to order $q+1$ by Assumption E11, and the fact that $\|\widehat{\overline{w}}_j - \overline{w}_j\| = o_p(1)$ for all i, j since $\sup_w \|\widehat{w} - w\| = o_p(1)$ and $\Pr(\widehat{p} \in \mathcal{P}) \longrightarrow 1$, as well as $\left((\widehat{w}_j - \widehat{w}_i)^{\mathbf{k}} - (w_j - w_i)^{\mathbf{k}}\right) \leq C|\widehat{p}_i - p_i|^{\mathbf{k}-1}$ for all $\mathbf{k} \leq q+1$, we obtain that:

$$\max_{1 \leq j \leq n} |\widehat{\Delta}_{m,j} - \Delta_{m,j}| = O_p(h_2^{q+1}).$$

Therefore, $\mathcal{T}_{m,n,2}$ can be bounded by:

$$\left\{ \max_{1 \leq j \leq n} |\Delta_{m,j}| + \max_{1 \leq j \leq n} |\widehat{\Delta}_{m,j} - \Delta_{m,j}| \right\} \left\| \frac{\sqrt{nh_2}}{n^2 h_2^{d_x+1}} \sum_{i=1}^{n} s_i \iota_1' \left[f(\widehat{w}_i) \mathbf{M} \right]^{-1} \widehat{\mathcal{K}}_j(\widehat{w}_i) \right\|$$

While the first term in brackets is of order $O_p(h_2^{q+1})$, the second term can be treated in a similar manner to before:

$$\frac{\sqrt{nh_2}}{n^2 h_2^{d_x+1}} \sum_{i=1}^{n} s_i \iota_1' \left[f(\widehat{w}_i) \mathbf{M} \right]^{-1} \mathcal{K}_j(w_i) + \frac{\sqrt{nh_2}}{n^2 h_2^{d_x+1}} \sum_{i=1}^{n} s_i \iota_1' \left[f(\widehat{w}_i) \mathbf{M} \right]^{-1} \{ \widehat{\mathcal{K}}_j(\widehat{w}_i) - \mathcal{K}_j(w_i) \}.$$

For the first term, standard U-statistic decomposition arguments similar to before yield that this expression is of order $O_p(\sqrt{nh_2})$ uniformly in $\mathbf{b}$. For the second term, a mean value expansion gives:

$$\frac{\sqrt{nh_2}}{n^2 h_2^{d_x+2}} \sum_{i=1}^{n} \sum_{j=1}^{n} s_i \iota_1' \left[f(\widehat{w}_i) \mathbf{M} \right]^{-1} \overline{\mathcal{K}}^{(1)}_{j,d_x+1}(\overline{w}_i) \{ (p_i - \widehat{p}_i) + (\widehat{p}_j - p_j) \} \tag{45}$$

As before, replacing $\widehat{\gamma}$ by γ_0 using Assumptions E7 and E13, the U-statistic with kernel function is given by:

$$\Psi^B_{p,ijl} \equiv s_i \iota_1' \left[f(w_i) \mathbf{M} \right]^{-1} \mathcal{K}^{(1)}_{j,d_x+1}(w_i) \omega(z_i' \gamma_0, z_l' \gamma_0) \psi_l + \ldots + s_j \iota_1' \left[f(w_j) \mathbf{M} \right]^{-1} \mathcal{K}^{(1)}_{i,d_x+1}(w_j) \omega(z_j' \gamma_0, z_l' \gamma_0) \psi_l$$

can be derived for the first term. Note that by iterated expectations this kernel function has (unconditional) mean zero since $\mathrm{E}[\psi_l | z_l' \gamma_0] = 0$ and $\mathrm{E}[h_2^{-2(d_x+2)} h_1^{-2} \|\Psi^B_{p,ijl}\|^2] = O(h_2^{-d_x-2} h_1^{-1}) = o(n)$ using change of variables and integration by parts arguments. By Lemma 3.1 in Powell et al. (1989) (see also Lemma A.3 in Ahn and Powell (1993)), this implies that the U-statistic is equal to its projection up to an approximation error of order $o_p(n^{-\frac{1}{2}})$. Thus:

$$\frac{\sqrt{nh_2}}{3 h_2^{d_x+2} h_1} \binom{n}{3}^{-1} \sum_{i=1}^{n} \sum_{j>i} \sum_{l>j} \Psi^B_{p,ijl} = \frac{\sqrt{nh_2}}{h_2^{d_x+1} h_1 n} \sum_{i=1}^{n} \mathrm{E}[\Psi^B_{p,ijl} | \varpi_i] + o_p(1).$$

Since $\mathrm{E}\left[h_2^{-2(d_x+1)} h_1^{-2} \|\Psi^B_{p,ijl}\|^2 \varpi_i \right] = O(1)$ by change of variables and integration by parts, this projection satisfies the Lindeberg-Levy CLT and is therefore of order $O_p(\sqrt{h_2})$. Furthermore, similar arguments to before can be used to show that the second term in Equation (45) is also of order $O_p(\sqrt{h_2})$. Given Assumption E13, it thus follows that $\mathcal{T}_{m,n,2} = o_p(\sqrt{h_2})$ uniformly in $\mathbf{b}$.

Finally, turning to $\mathcal{T}_{m,n,3}$ and $\mathcal{T}_{m,n,4}$, note that we have that:

$$|\widehat{\mathbf{M}}_n(\widehat{w}_i) - \mathbf{M} f(\widehat{w}_i)| \leq |\widehat{\mathbf{M}}_n(\widehat{w}_i) - \mathbf{M}_n(\widehat{w}_i)| + |\mathbf{M}_n(\widehat{w}_i) - \mathbf{M} f(\widehat{w}_i)| \tag{46}$$

For the first term on the RHS (RHS) of (46), note that a typical element of $\widehat{\mathbf{M}}_n(\widehat{w}_i) - \mathbf{M}_n(\widehat{w}_i)$ is given by:

$$\begin{aligned} & \left[\widehat{\mathbf{M}}_{n;i,j}(\widehat{w}_i)\right]_{l,l_0} - \left[\mathbf{M}_{n;i,j}(\widehat{w}_i)\right]_{l,l_0} \\ = & \frac{1}{nh_2^{d_x+1}} \sum_{j=1}^{n} \left[\left(\frac{\widehat{w}_j - \widehat{w}_i}{h} \right)^{\phi_i(l)+\phi_j(l_0)} K\left(\frac{\widehat{w}_j - \widehat{w}_i}{h} \right) - \left(\frac{w_j - \widehat{w}_i}{h} \right)^{\phi_i(l)+\phi_j(l_0)} K\left(\frac{w_j - \widehat{w}_i}{h} \right) \right] \end{aligned}$$

Note that similar arguments to the ones used for $\mathcal{T}_{m,n,12}$ and $\mathcal{T}_{m,n,2}$ can be used to show that:

$$\mathcal{T}_{m,n,3} = \frac{\sqrt{nh_2}}{n} \sum_{i=1}^{n} s_i \iota_1' \left[\mathbf{M}f(\widehat{w}_i)\right]^{-1} \left[\widehat{\mathbf{M}}_n(\widehat{w}_i) - \mathbf{M}_n(\widehat{w}_i)\right] \widehat{\mathbf{M}}_n^{-1}(\widehat{w}_i) \widehat{V}_{m,n}(\widehat{w}_i) = O_p(\sqrt{h_2})$$

and

$$\mathcal{T}_{m,n,4} = \frac{\sqrt{nh_2}}{n} \sum_{i=1}^{n} s_i \iota_1' \left[\mathbf{M}f(\widehat{w}_i)\right]^{-1} \left[\widehat{\mathbf{M}}_n(\widehat{w}_i) - \mathbf{M}_n(\widehat{w}_i)\right] \widehat{\mathbf{M}}_n^{-1}(\widehat{w}_i) \widehat{B}_{m,n}(\widehat{w}_i) = O_p(\sqrt{h_2})$$

uniformly in $\mathbf{b}$.

For the second term on the RHS of (46), re-inserting into e.g. $\mathcal{T}_{m,n,3}$, we obtain:

$$\begin{aligned} & \frac{\sqrt{nh_2}}{n} \sum_{i=1}^{n} s_i \iota_1' \left[\mathbf{M}f(\widehat{w}_i)\right]^{-1} \left[\mathbf{M}_n(\widehat{w}_i) - \mathbf{M}f(\widehat{w}_i)\right] \widehat{\mathbf{M}}_n^{-1}(\widehat{w}_i) \widehat{V}_{m,n}(\widehat{w}_i) \\ = & \frac{\sqrt{nh_2}}{n} \sum_{i=1}^{n} s_i \iota_1' \left[\mathbf{M}f(w_i)\right]^{-1} \left[\mathbf{M}_n(w_i) - \mathbf{M}f(w_i)\right] \mathbf{M}_n^{-1}(w_i) V_{m,n}(w_i) + o_p(1) \\ = & \frac{\sqrt{nh_2}}{n} \sum_{i=1}^{n} s_i \iota_1' \left[\mathbf{M}f(w_i)\right]^{-1} \left[\frac{1}{nh_2^{d_x+1}} \sum_{j=1}^{n} \mathcal{K}_j(w_i) - \mathbf{M}f(w_i) \right] \mathbf{M}_n^{-1}(w_i) V_{m,n}(w_i) + o_p(1) \end{aligned}$$

Next, notice that:

$$\begin{aligned} & \mathrm{E}\left[\frac{\sqrt{nh_2}}{n} \sum_{i=1}^{n} s_i \iota_1' \left[\mathbf{M}f(w_i)\right]^{-1} \left[\frac{1}{nh_2^{d_x+1}} \sum_{j=1}^{n} \mathcal{K}_j(w_i) - \mathbf{M}f(w_i) \right] \mathbf{M}_n^{-1}(w_i) V_{m,n}(w_i) \right] \\ = & \sqrt{nh_2} \int \mathrm{E}[s_i|w_i] \iota_1' \left[\mathbf{M}f(w_i)\right]^{-1} \left[\varphi_n(w_i) + \mu_n(w_i)\right] \mathbf{M}_n^{-1}(w_i) V_{m,n}(w_i) f(w_i) dw_i, \end{aligned}$$

where the equality follows from iterated expectations, change of variables, a Taylor expansion (see Equation (2.9) in Masry (1996, p. 575)), and the definition of $\varphi_n(w_i)$ and $\mu_n(w_i)$ in (22) and (23). Since

$$\sup_w \left| \varphi_n(w) h_2^{-q-1} - \varphi(w) \right| = O_p\left(\left(\frac{\ln n}{nh_2^{d_x+1}} \right)^{\frac{1}{2}} \right)$$

and

$$\sup_w \left| \mu_n(w) h_2^{-q-1} - \mu(w) \right| = O_p\left(\left(\frac{\ln n}{nh_2^{d_x+1}} \right)^{\frac{1}{2}} \right) h_2^{q+1}$$

we can therefore write:

$$\sqrt{n}h_2^{q+\frac{3}{2}} \int \mathrm{E}[s_i|w_i]\iota_1' \left[\mathbf{M}f(w_i)\right]^{-1} \left[\varphi(w_i)+\mu(w_i)\right] \mathbf{M}_n^{-1}(w_i) V_{m,n}(w_i) f(w_i) dw_i = \sqrt{n}h_2^{q+\frac{3}{2}} O(1) = o(\sqrt{h_2})$$

uniformly in $\mathbf{b}$. By Markov's inequality, it thus follows that $\mathcal{T}_{m,n,3}$ is of order $O_p(\sqrt{h_2})$ uniformly in $\mathbf{b}_0$. A similar argument can be made for $\mathcal{T}_{m,n,4}$, which establishes the claim. ■

Proof of Lemma B.3. Using the representation from Equation (21), note again that:

$$\begin{aligned}
& \frac{\sqrt{nh_2}}{n} \sum_{i=1}^{n} s_i (\widehat{m}(\widehat{w}_{i1}) - m(\widehat{w}_{i1})) \qquad (47)\\
= \; & \frac{\sqrt{nh_2}}{n} \sum_{i=1}^{n} s_i \iota_1' \widehat{\mathbf{M}}_n^{-1}(w_{i1}) \left\{ \frac{1}{nh_2^{d_x+1}} \sum_{j=1}^{n} \widehat{\mathcal{K}}_j(w_{i1}) u_j \right\} \\
& + \frac{\sqrt{nh_2}}{n} \sum_{i=1}^{n} s_i \iota_1' \widehat{\mathbf{M}}_n^{-1}(w_{i1}) \left\{ \frac{1}{nh_2^{d_x+1}} \sum_{j=1}^{n} \widehat{\mathcal{K}}_j(w_{i1}) \widehat{\Delta}_{m_1,j} \right\},
\end{aligned}$$

where

$$\widehat{\Delta}_{m_1,j} = m(w_{i1}) - \sum_{0 \le |\mathbf{k}| \le q} \frac{1}{\mathbf{k}!} D^{\mathbf{k}} m(w_{i1}) (\widehat{w}_j - w_{i1})^{\mathbf{k}}.$$

Defining $\widehat{V}_{m_1,n}(w_{i1})$ and $\widehat{B}_{m_1,n}(w_{i1})$ in a similar way to $\widehat{V}_{m,n}(\widehat{w}_i)$ and $\widehat{B}_{m,n}(\widehat{w}_i)$ in the proof of Lemma B.2, we can write:

$$\begin{aligned}
& \frac{\sqrt{nh_2}}{n} \sum_{i=1}^{n} s_i \iota_1' \widehat{\mathbf{M}}_n^{-1}(w_{i1}) \left\{ \widehat{V}_{m_1,n}(w_{i1}) + \widehat{B}_{m_1,n}(w_{i1}) \right\} \\
= \; & \frac{\sqrt{nh_2}}{n} \sum_{i=1}^{n} s_i \iota_1' \left[f(w_{i1}) \mathbf{M}_1 \{1+o_p(1)\} \right]^{-1} \widehat{V}_{m_1,n}(w_{i1}) \\
& + \frac{\sqrt{nh_2}}{n} \sum_{i=1}^{n} s_i \iota_1' \left[f(w_{i1}) \mathbf{M}_1 \{1+o_p(1)\} \right]^{-1} \widehat{B}_{m,n}(w_{i1}) \qquad (48)\\
& - \frac{\sqrt{nh_2}}{n} \sum_{i=1}^{n} s_i \iota_1' \left[\mathbf{M}_1 f(w_{i1}) \{1+o_p(1)\} \right]^{-1} \left[\widehat{\mathbf{M}}_n(w_{i1}) - \mathbf{M}_1 f(w_{i1}) \right] \widehat{\mathbf{M}}_n^{-1}(w_{i1}) \widehat{V}_{m,n}(w_{i1}) \\
& - \frac{\sqrt{nh_2}}{n} \sum_{i=1}^{n} s_i \iota_1' \left[\mathbf{M}_1 f(w_{i1}) \{1+o_p(1)\} \right]^{-1} \left[\widehat{\mathbf{M}}_n(w_{i1}) - \mathbf{M}_1 f(w_{i1}) \right] \widehat{\mathbf{M}}_n^{-1}(w_{i1}) \widehat{B}_{m,n}(w_{i1}) \\
\equiv \; & \mathcal{T}_{m1,n,1} + \mathcal{T}_{m1,n,2} + \mathcal{T}_{m1,n,3} + \mathcal{T}_{m1,n,4}.
\end{aligned}$$

Starting with $\mathcal{T}_{m1,n,1}$, we can again write:

$$\begin{aligned}
&\frac{\sqrt{nh_2}}{n}\sum_{i=1}^{n} s_i\iota_1'\left[f(w_{i1})\mathbf{M}_1\right]^{-1}V_{m_1,n}(w_{i1}) \\
&+\frac{\sqrt{nh_2}}{n}\sum_{i=1}^{n} s_i\iota_1'\left[f(w_{i1})\mathbf{M}_1\right]^{-1}\left\{\widehat{V}_{m_1,n}(\widehat{w}_{i1})-V_{m_1,n}(w_{i1})\right\}+o_p(1) \\
\equiv\ & \mathcal{T}_{m1,n,11}+\mathcal{T}_{m1,n,12}+o_p(1).
\end{aligned}$$

Consider $\mathcal{T}_{m1,n,11}$, which drives the distribution. Define the symmetrized 'kernel function':

$$\Psi_{u1,ij}\equiv s_i\iota_1'\left[f(w_{i1})\mathbf{M}_1\right]^{-1}\mathcal{K}_j(w_{i1})u_j+s_j\iota_1'\left[f(w_{1j})\mathbf{M}_1\right]^{-1}\mathcal{K}_i(w_{1j})u_i$$

Then, using the same argument of Ahn and Powell (1993, p.25) to show that terms with $i=j$ are asymptotically bounded by:

$$\frac{1}{n^2h_2^{d_x+1}}\sum_{i=1}^{n}\left\|s_i\iota_1'\left[f(w_{i1})\mathbf{M}_1\right]^{-1}\mathcal{K}_i(0,1)u_i\right\|=O_p\left(\frac{1}{nh_2^{d_x}}\right)=o_p\left(\frac{1}{\sqrt{nh_2}}\right),$$

where $\mathcal{K}_i(0,1)$ is defined as $\mathcal{K}_j(w_{i1})$, but with the first d_x elements equal to 0 and the last entry equal to (p_i-1). Then, derive the (appropriately rescaled) U-statistic:

$$\begin{aligned}
&\frac{\sqrt{nh_2}}{n}\sum_{i=1}^{n} s_i\iota_1'\left[f(w_{i1})\mathbf{M}_1\right]^{-1}V_{m_1,n}(w_i) \\
=\ & \frac{\sqrt{nh_2}}{2h_2^{d_x+1}}\binom{n}{2}^{-1}\sum_{i=1}^{n}\sum_{j>i}\Psi_{u1,ij}+o_p(1) \\
=\ & \frac{\sqrt{nh_2}}{nh_2^{d_x+1}}\sum_{i=1}^{n}\mathrm{E}[\Psi_{u1,ij}|\varpi_i]+o_p(1),
\end{aligned}$$

where the second equality follows from a standard Hoeffding decomposition,the fact that $\mathrm{E}[\Psi_{u1,ij}]=0$ by iterated expectations, and from Lemma 3.1 in Powell et al. (1989) since $\mathrm{E}\left[h_2^{-2(d_x+1)}\left\|\Psi_{u1,ij}\right\|^2\right]=O(h_2^{-(d_x+1)})=o(n)$. Moreover, after change of variables and iterated expectations noting that $\mathrm{E}^*[u_i^2|w_{i1}]=\sigma^{*2}(w_{i1})$ and $\mathrm{E}\left[h_2^{-d_x-1}\Psi_{u1,ij}\right]=O(h_2^{-1})$, we obtain:

$$\mathrm{E}\left[\left\|\mathrm{E}\left[\frac{1}{h_2^{d_x+1}}\Psi_{u1,ij}\middle|\varpi_i\right]\right\|^2\right]=\frac{1}{h_2}\mathrm{E}\left[s_i\sigma^{*2}(w_{i1})f(w_{i1})^{-1}\right]\iota_1'\mathbf{M}_1^{-1}\Gamma\mathbf{M}_1^{-1\prime}\iota_1,$$

uniformly in $\mathbf{b}$, where Γ, defined in Section B.1, contains submatrices $\Gamma_{i,j}$, which are of dimension $N_i\times N_j$ with (l,m)-elements $\gamma_{\phi_i(l)+\phi_j(m)}$. Therefore, by Lindeberg-Levy CLT, it holds that:

$$\frac{\sqrt{h_2}}{\sqrt{n}h_2^{d_x+1}}\sum_{i=1}^{n}\mathrm{E}[\Psi_{u1,ij}|\varpi_i]\xrightarrow{d}N(0,\Omega_0),$$

with Ω_0 defined in Lemma B.3. Next, consider $\mathcal{T}_{m1,n,12}$ and note again that:

$$\begin{aligned} & \frac{\sqrt{nh_2}}{n^2 h_2^{d_x+1}} \sum_{i=1}^{n}\sum_{j=1}^{n} s_i \iota_1' \left[f(w_{i1})\mathbf{M}_1\right]^{-1} \left\{\widehat{\mathcal{K}}_j(w_{i1}) - \mathcal{K}_j(w_{i1})\right\} u_j \\ = \; & \frac{\sqrt{nh_2}}{n^2 h_2^{d_x+2}} \sum_{i=1}^{n}\sum_{j=1}^{n} s_i \iota_1' \left[f(w_{i1})\mathbf{M}_1\right]^{-1} \overline{\mathcal{K}}^{(1)}_{j,d_x+1}(w_{i1}) \left\{(\widehat{p}_j - p_j)\right\} u_j \end{aligned} \tag{49}$$

First, we note that using the representation from Assumption E7 together with Assumption E13, we have:

$$\begin{aligned} \mathcal{T}_{m1,n,12} &= \frac{\sqrt{nh_2}}{nh_2^{d_x+2}} \sum_{i=1}^{n} s_i \iota_1' \left[f(w_{i1})\mathbf{M}_1\right]^{-1} \left\{ \frac{1}{n}\sum_{j=1}^{n} \overline{\mathcal{K}}^{(1)}_{j,d_x+1}(w_{i1}) \left(p_j - \widehat{p}_j\right) u_j \right\} + o_p(1) \qquad (50) \\ &= \frac{\sqrt{nh_2}}{n^2 h_2^{d_x+2}} \sum_{i=1}^{n}\sum_{j=1}^{n} s_i \iota_1' \left[f(w_{i1})\mathbf{M}_1\right]^{-1} \overline{\mathcal{K}}^{(1)}_{j,d_x+1}(w_{i1}) \left\{ \frac{1}{nh_1}\sum_{l=1}^{n} \omega(z_j'\gamma_0, z_l'\gamma_0)\psi_l + \Xi_n(z_j'\gamma_0) \right\} u_j + o_p(1) \end{aligned}$$

We will start with the term involving $\omega(z_j'\gamma_0, z_l'\gamma_0)\psi_l$. By another mean value expansion around $\mathcal{K}^{(1)}_{j,d_x+1}(w_{i1})$, we can write this expression as:

$$\frac{\sqrt{nh_2}}{n^3 h_2^{d_x+2} h_1} \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n} s_i \iota_1' \left[f(w_{i1})\mathbf{M}_1\right]^{-1} \mathcal{K}^{(1)}_{j,d_x+1}(w_{i1})\omega(z_j'\gamma_0, z_l'\gamma_0)\psi_l u_j + o_p(1).$$

Also, note that asymptotically we can omit terms with $i = j$, $j = l$, and $j = l$ since e.g., terms with $i = j$ can be asymptotically bounded by:

$$\frac{1}{n^3 h_2^{d_x+2} h_1} \sum_{i=1}^{n}\sum_{l\neq i} \left\| s_i \iota_1' \left[f(w_{i1})\mathbf{M}_1\right]^{-1} \mathcal{K}^{(1)}_{j,d_x+1}(0,1)\omega(z_i'\gamma_0, z_l'\gamma_0)\psi_l u_j \right\| = O_p\left(\frac{1}{nh_2^{d_x+1}}\right) = o_p\left(\frac{1}{\sqrt{nh_2}}\right).$$

Next, we define the symmetrized kernel function for a third order U-statistic:

$$\Psi^{V1}_{p,ijl} \equiv s_i \iota_1' \left[f(w_{i1})\mathbf{M}_1\right]^{-1} \mathcal{K}^{(1)}_{j,d_x+1}(w_{i1})\omega(z_j'\gamma_0, z_l'\gamma_0)\psi_l u_j + \ldots + s_j \iota_1' \left[f(w_{1j})\mathbf{M}_1\right]^{-1} \mathcal{K}^{(1)}_{i,d_x+1}(w_{1j})\omega(z_i'\gamma_0, z_l'\gamma_0)\psi_l u_i.$$

Note that by iterated expectations this kernel function has (unconditional) mean zero. Moreover:

$$
\begin{aligned}
& \mathrm{E}[h_2^{-2(d_x+2)} h_1^{-2} \|\Psi_{p,ijl}^{V1}\|^2] \\
= \;& 6\,\mathrm{E}\left[h_2^{-2(d_x+2)} h_1^{-2} \left\| s_i \iota_1' \left[f(w_{i1}) \mathbf{M}_1 \right]^{-1} \mathcal{K}_{j,d_x+1}^{(1)}(w_{i1}) \omega(z_j'\gamma_0, z_l'\gamma_0) \psi_l u_j \right\|^2 \right] \\
= \;& 6\,\mathrm{E}\Bigg[h_2^{-(d_x+3)} h_1^{-2} \int_{\mathbb{R}^{d_x}} \int_{-\frac{1}{h_2}}^{0} \left\| s_i \iota_1' \left[f(w_{i1}) \mathbf{M}_1 \right]^{-1} \mathcal{K}_{j,d_x+1}^{(1)}(\nu, \nu_{d_x+1}) \omega(z_j'\gamma_0, z_l'\gamma_0) \psi_l u_j \right\|^2 \\
& \times f(w_{-i} + h_2\nu, 1 + h_2\nu_{d_x+1}) d\nu d\nu_{d_x+1} \Bigg] \\
= \;& 6\,\mathrm{E}\Bigg[h_2^{-(d_x+3)} h_1^{-2} \int_{\mathbb{R}^{d_x}} \Bigg[\left\| s_i \iota_1' \left[f(w_{i1}) \mathbf{M}_1 \right]^{-1} \mathcal{K}(\nu, \nu_{d_x+1}) \omega(z_j'\gamma_0, z_l'\gamma_0) \psi_l u_j \right\|^2 \\
& \times f(w_{-i} + h_2\nu, 1 + h_2\nu_{d_x+1}) \Bigg]_{-\frac{1}{h_2}}^{0} d\nu \Bigg] \\
& -6\,\mathrm{E}\Bigg[h_2^{-(d_x+2)} h_1^{-2} \int_{\mathbb{R}^{d_x}} \int_{-\frac{1}{h_2}}^{0} \left\| s_i \iota_1' \left[f(w_{i1}) \mathbf{M} \right]^{-1} \mathcal{K}(\nu, \nu_{d_x}) \omega(z_j'\gamma_0, z_l'\gamma_0) \psi_l u_j \right\|^2 \\
& \times D_{d_x+1}^{\mathbf{1}} f(w_{-i} + h_2\nu, 1 + h_2\nu_{d_x+1}) d\nu_1 \Bigg],
\end{aligned}
\tag{51}
$$

where the second equality comes from a classical change of variables and the third equality from integration by parts noting that $D_{d_x+1}^{\mathbf{1}} f(w_{-i} + h_2\nu, 1 + h_2\nu_{d_x+1})$ denoting the first order partial derivative w.r.t. ν_{d_x+1}. The first term on the RHS of the last equality yields:

$$
\begin{aligned}
& 6\,\mathrm{E}\left[h_2^{-(d_x+3)} h_1^{-2} \int_{\mathbb{R}^{d_x}} \left\| s_i \iota_1' \left[f(w_{i1}) \mathbf{M}_1 \right]^{-1} \mathcal{K}(\nu, 0) \omega(z_j'\gamma_0, z_l'\gamma_0) \psi_l u_j \right\|^2 f(w_{-i} + h_2\nu, 1) d\nu \right] \\
& -6\,\mathrm{E}\left[h_2^{-(d_x+3)} h_1^{-2} \int_{\mathbb{R}^{d_x}} \left\| s_i \iota_1' \left[f(w_{i1}) \mathbf{M}_1 \right]^{-1} \mathcal{K}(\nu, -\frac{1}{h_2}) \omega(z_j'\gamma_0, z_l'\gamma_0) \psi_l u_j \right\|^2 f(w_{-i} + h_2\nu, 0) d\nu \right] \\
= \;& O(h_2^{-(d_x+3)} h_1^{-1}),
\end{aligned}
$$

which follows since the first term is of order $O(h_2^{-(d_x+3)} h_1^{-1})$ by Assumptions E7, E9, and E10, while the second term is zero by Assumption E9. Similar arguments and Assumption E11 yield that the second term on the RHS of the last equality of Equation (51) is of order $O(h_2^{-(d_x+2)} h_1^{-1})$. Therefore, by Assumption E13, we have that $\mathrm{E}[h_2^{-2(d_x+2)} h_1^{-2} \|\Psi_{p,ijl}^{V1}\|^2] = O(h_2^{-(d_x+3)} h_1^{-1}) = o(n)$. By Lemma 3.1 in Powell et al. (1989), this implies that the U-statistic is equal to its projection up to an approximation error of order $o_p(n^{-\frac{1}{2}})$. Moreover, we note that because $\mathrm{E}[\psi_l | z_l'\gamma_0] = 0$ and $\mathrm{E}^*[u_i | z_i'\gamma_0, x_i'\beta_0, p_i] = \mathrm{E}^*[u_i | x_i'\beta_0, p_i] = 0$ by iterated expectations, $h_2^{-(d_x+2)} h_1^{-1} \mathrm{E}\left[\Psi_{p,ijl}^{V1} | \varpi_i \right] = 0$, too. Therefore, it is

straightforward to deduce that:

$$\frac{\sqrt{nh_2}}{3h_2^{-(d_x+2)}h_1^{-1}}\binom{n}{3}^{-1}\sum_{i=1}^{n}\sum_{j>i}\sum_{l>j}\Psi_{p,ijl}^{V1}=o_p(\sqrt{h_2}).$$

For the second part on the RHS of the last equality of Equation (50), note that:

$$\begin{aligned}&\frac{\sqrt{nh_2}}{n^2h_2^{-(d_x+2)}}\sum_{i=1}^{n}\sum_{j=1}^{n}s_i\iota_1'\left[f(\widehat{w}_i)\mathbf{M}\right]^{-1}\overline{\mathcal{K}}_{j,d_x+1}^{(1)}(x_i'\widehat{\beta},1)\Xi_{i,n}u_j\\ \leq\ &\max_{1\leq i\leq n}|\Xi_{i,n}|\left\|\frac{\sqrt{nh_2}}{n^2h_2^{-(d_x+2)}}\sum_{i=1}^{n}\sum_{j=1}^{n}s_i\iota_1'\left[f(w_i)\mathbf{M}\right]^{-1}\mathcal{K}_{j,d_x+1}^{(1)}(w_{i1})u_j\right\|+o_p(1)\end{aligned}$$

The first term on the RHS of the last inequality is of order $O_p(h_1^2)$ by Assumption E7. For the second part, note again that observations with $i=j$ can be neglected asymptotically, so that:

$$\begin{aligned}&\frac{\sqrt{nh_2}}{n^2h_2^{-(d_x+2)}}\sum_{i=1}^{n}\sum_{j=1}^{n}s_i\iota_1'\left[f(w_i)\mathbf{M}_1\right]^{-1}\mathcal{K}_{j,d_x+1}^{(1)}(w_{i1})u_j\\ =\ &\frac{\sqrt{nh_2}}{2h_2^{-(d_x+2)}}\binom{n}{2}^{-1}\sum_{i=1}^{n}\sum_{j>i}\left\{s_i\iota_1'\left[f(w_i)\mathbf{M}_1\right]^{-1}\mathcal{K}_{j,d_x+1}^{(1)}(w_{i1})u_j+s_j\iota_1'\left[f(w_j)\mathbf{M}_1\right]^{-1}\mathcal{K}_{i,d_x+1}^{(1)}(w_{1j})u_i\right\}+o_p(1).\end{aligned}$$

Using standard Hoeffding decomposition arguments, change of variables, and integration by parts, it is not difficult to deduce that this term is of order $O_p(\sqrt{h_2})$ uniformly in $\mathbf{b}$. Given Assumption E13, we can therefore conclude that the whole term is of order $o_p(1)$.

Turning to $\mathcal{T}_{m1,n,2}$, similar to Lemma B.2, we again define:

$$\Delta_{m1,j}=m(w_i)-\sum_{0\leq|\mathbf{k}|\leq q}\frac{1}{\mathbf{k}!}D^{\mathbf{k}}m(w_{i1})(w_j-w_{i1})^{\mathbf{k}}.$$

in analogy to $\Delta_{m,j}$. Then, by Assumption E11, it holds that:

$$\Delta_{m1,j}=\sum_{|\mathbf{k}|=q+1}\frac{1}{\mathbf{k}!}D^{\mathbf{k}}m(\overline{w})(w_j-w_{i1})^{\mathbf{k}}.$$

for some $\overline{w}$ in between w_j and w_{i1}. Also:

$$\widehat{\Delta}_{m1,j}=\sum_{|\mathbf{k}|=q+1}\frac{1}{\mathbf{k}!}D^{\mathbf{k}}m(\widehat{\overline{w}}_1)(\widehat{w}_j-\widehat{w}_{i1})^{\mathbf{k}},$$

where $\widehat{\overline{w}}_1$ lies between $\widehat{w}_j$ and $\widehat{w}_{i1}$. Since $|\Delta_{m1,j}|=O_p(h_2^{q+1})$ uniformly in $\mathbf{b}$, i and j for $\|w_j-w_{i1}\|\leq Ch_2$. Thus, it remains to analyze:

$$\begin{aligned}\widehat{\Delta}_{m1,j}-\Delta_{m1,j}\ =\ &\sum_{|\mathbf{k}|=q+1}\frac{1}{\mathbf{k}!}D^{\mathbf{k}}m(\widehat{\overline{w}}_1)\left((\widehat{w}_j-\widehat{w}_{i1})^{\mathbf{k}}-(w_j-w_{i1})^{\mathbf{k}}\right)\\ &+\sum_{|\mathbf{k}|=q+1}\frac{1}{\mathbf{k}!}\left[D^{\mathbf{k}}m(\widehat{\overline{w}}_1)-D^{\mathbf{k}}m(\overline{w}_1)\right](w_j-w_{i1})^{\mathbf{k}}.\end{aligned}$$

By Assumption E11 up to order $q+2$, and the fact that $\|\widehat{\overline{w}}_{1i} - \overline{w}_{1i}\| = o_p(1)$ for all i, j since $\sup_w |\widehat{w} - w| = o_p(1)$ and $\Pr(\widehat{p} \in \mathcal{P}) \longrightarrow 1$, as well as $\left((\widehat{w}_j - \widehat{w}_{i1})^{\mathbf{k}} - (w_j - w_{i1})^{\mathbf{k}}\right) \leq C|\widehat{p}_i - p_i|^{\mathbf{k}-1}$ for all $\mathbf{k} \leq q+1$, we obtain that:

$$\max_{1 \leq j \leq n} |\widehat{\Delta}_{m1,j} - \Delta_{m1,j}| = O_p(h_2^{q+1}).$$

Therefore, $\mathcal{T}_{m1,n,2}$ can be bounded by:

$$\left\{ \max_{1 \leq j \leq n} |\Delta_{m1,j}| + \max_{1 \leq j \leq n} |\widehat{\Delta}_{m1,j} - \Delta_{m1,j}| \right\} \left\| \frac{\sqrt{nh_2}}{n^2 h_2^{d_x+1}} \sum_{i=1}^{n} \sum_{j=1}^{n} s_i \iota_1' \left[f(\widehat{w}_{i1}) \mathbf{M}_1 \right]^{-1} \widehat{\mathcal{K}}_j(\widehat{w}_{i1}) \right\|$$

While the first term in brackets is of order $O_p(h_2^{q+1})$, the second term can be treated in a similar manner to before:

$$\frac{\sqrt{nh_2}}{n^2 h_2^{d_x+1}} \sum_{i=1}^{n} \sum_{j=1}^{n} s_i \iota_1' \left[f(\widehat{w}_{i1}) \mathbf{M}_1 \right]^{-1} \mathcal{K}_j(w_{i1}) + \frac{\sqrt{nh_2}}{n^2 h_2^{d_x+1}} \sum_{i=1}^{n} \sum_{j=1}^{n} s_i \iota_1' \left[f(\widehat{w}_{i1}) \mathbf{M}_1 \right]^{-1} \{ \widehat{\mathcal{K}}_j(\widehat{w}_{i1}) - \mathcal{K}(w_{i1}) \}.$$

For the first term, standard U-statistic decomposition arguments similar to before yield that this expression is of order $O_p(\sqrt{nh_2})$ uniformly in $\mathbf{b}$. For the second term, a mean value expansion gives:

$$\frac{\sqrt{nh_2}}{n^2 h_2^{d_x+2}} \sum_{i=1}^{n} \sum_{j=1}^{n} s_i \iota_1' \left[f(\widehat{w}_{i1}) \mathbf{M}_1 \right]^{-1} \overline{\mathcal{K}}^{(1)}_{j,d_x+1}(w_{i1}) \{ \widehat{p}_j - p_j \} \tag{52}$$

As before, we can use Assumptions E7 and E13 together with another mean value expansion to re-write the term of Equation (52) as:

$$\frac{\sqrt{nh_2}}{n^2 h_2^{d_x+2}} \sum_{i=1}^{n} \sum_{j=1}^{n} s_i \iota_1' \left[f(\widehat{w}_i) \mathbf{M} \right]^{-1} \overline{\mathcal{K}}^{(1)}_{j,d_x+1}(w_{i1}) \left\{ \frac{1}{nh_1} \sum_{l=1}^{n} \omega(z_j' \gamma_0, z_l' \gamma_0) \psi_l + \Xi_n(z_j' \gamma_0) \right\} + o_p(1) \tag{53}$$

We start with the term involving $\omega(z_j' \gamma_0, z_l' \gamma_0) \psi_l$. This expression can again be approximated by a U-statistic with kernel function:

$$\Psi^{B1}_{p,ijl} \equiv s_i \iota_1' \left[f(w_{i1}) \mathbf{M}_1 \right]^{-1} \mathcal{K}^{(1)}_{j,d_x+1}(w_{i1}) \omega(z_j' \gamma_0, z_l' \gamma_0) \psi_l + \ldots + s_j \iota_1' \left[f(w_{1j}) \mathbf{M}_1 \right]^{-1} \mathcal{K}^{(1)}_{i,d_x+1}(w_{j1}) \omega(z_i' \gamma_0, z_l' \gamma_0) \psi_l.$$

Also, note that by iterated expectations this kernel function has again (unconditional) mean zero since

$\mathrm{E}[\psi_l|z_l'\gamma_0] = 0$ and:

$$
\begin{aligned}
& \mathrm{E}[h_2^{-2(d_x+2)} h_1^{-2} \|\Psi_{p,ijl}^{B1}\|^2] \\
= \; & 6\,\mathrm{E}\left[h_2^{-2(d_x+2)} h_1^{-2} \left\| s_i \iota_1' \left[f(w_{i1}) \mathbf{M}_1 \right]^{-1} \mathcal{K}_{j,d_x+1}^{(1)}(w_{i1}) \omega(z_j'\gamma_0, z_l'\gamma_0) \psi_l \right\|^2 \right] \\
= \; & 6\,\mathrm{E}\left[h_2^{-d_x-3} h_1^{-2} \int_{\mathbb{R}^{d_x}} \int_{-\frac{1}{h_2}}^{0} \left\| s_i \iota_1' \left[f(w_{i1}) \mathbf{M}_1 \right]^{-1} \mathcal{K}_{d_x+1}^{(1)}(\nu, \nu_{d_x+1}) \omega(z_j'\gamma_0, z_l'\gamma_0) \psi_l \right\|^2 \right. \\
& \left. \times f(w_{-i} + h_2\nu, 1 + h_2\nu_{d_x+1}) d\nu d\nu_{d_x+1} \right] \qquad (54) \\
= \; & 6\,\mathrm{E}\left[h_2^{-d_x-3} h_1^{-2} \int_{\mathbb{R}^{d_x}} \left[\left\| s_i \iota_1' \left[f(w_{i1}) \mathbf{M}_1 \right]^{-1} \mathcal{K}_{d_x+1}^{(1)}(\nu, \nu_{d_x+1}) \omega(z_j'\gamma_0, z_l'\gamma_0) \psi_l \right\|^2 \right. \right. \\
& \left. \left. \times f(w_{-i} + h_2\nu, 1 + h_2\nu_{d_x+1}) \right]_{-\frac{1}{h_2}}^{0} d\nu \right] \\
& -6\,\mathrm{E}\left[h_2^{-d_x-2} h_1^{-2} \int_{\mathbb{R}^{d_x}} \int_{-\frac{1}{h_2}}^{0} \left\| s_i \iota_1' \left[f(w_{i1}) \mathbf{M} \right]^{-1} \mathcal{K}_{d_x+1}^{(1)}(\nu, \nu_{d_x+1}) \omega(z_j'\gamma_0, z_l'\gamma_0) \psi_l \right\|^2 \right. \\
& \left. \times D_{d_x+1}^{\mathbf{1}} f(w_{-i} + h_2\nu, 1 + h_2\nu_{d_x+1}) d\nu \right],
\end{aligned}
$$

where the second equality comes from a classical change of variables and the third equality from integration by parts. The first term on the RHS of the last equality yields:

$$
\begin{aligned}
& 6\,\mathrm{E}\left[h_2^{-d_x-3} h_1^{-2} \int_{\mathbb{R}^{d_x}} \left\| s_i \iota_1' \left[f(w_{i1}) \mathbf{M}_1 \right]^{-1} \mathcal{K}_{d_x+1}^{(1)}(\nu, 0) \omega(z_j'\gamma_0, z_l'\gamma_0) \psi_l \right\|^2 f(w_{-i} + h_2\nu, 1) d\nu \right] \\
& -6\,\mathrm{E}\left[h_2^{-d_x-3} h_1^{-2} \int_{\mathbb{R}^{d_x}} \left\| s_i \iota_1' \left[f(w_{i1}) \mathbf{M}_1 \right]^{-1} \mathcal{K}(\nu, -\frac{1}{h_2}) \omega(z_j'\gamma_0, z_l'\gamma_0) \psi_l \right\|^2 f(w_{-i} + h_2\nu, 0) d\nu \right] \\
= \; & O(h_2^{-d_x-3} h_1^{-1}),
\end{aligned}
$$

which follows since the first term is of order $O(h_2^{-d_x-3}h_1^{-1})$ by Assumption E7, E9, and E10, while the second term is zero by Assumption E9. Similar arguments and Assumption E11 yield that the second term on the RHS of Equation (54) is of order $O(h_2^{-d_x-2}h_1^{-1})$. Therefore, by Assumption E13, we have that $\mathrm{E}[h_2^{-2(d_x+2)} h_1^{-2} \|\Psi_{p,ijl}^{B1}\|^2] = O(h_2^{-d_x-3}h_1^{-1}) = o(n)$. By Lemma 3.1 in Powell et al. (1989), this implies that the U-statistic is equal to its projection up to an approximation error of order $o_p(n^{-\frac{1}{2}})$. Thus:

$$
\frac{\sqrt{nh_2}}{3h_2^{-d_x-2}h_1} \binom{n}{3}^{-1} \sum_{i=1}^{n} \sum_{j>i} \sum_{l>j} \Psi_{p,ijl}^{B1} = \frac{\sqrt{nh_2}}{h_2^{-d_x-2}h_1 n} \sum_{i=1}^{n} \mathrm{E}[\Psi_{p,ijl}^{B1}|\varpi_i] + o_p(1).
$$

Note that:

$$
\begin{aligned}
& h_2^{-d_x-2}h_1^{-1}\,\mathrm{E}\Big[\Psi^{B1}_{p,ijl}\Big|\varpi_i\Big] \\
= \;& 2\,\mathrm{E}\left[h_2^{-d_x-2}h_1^{-1}s_j\iota_1'\left[f(w_{1j})\mathbf{M}_1\right]^{-1}\mathcal{K}^{(1)}_{l,d_x+1}(w_{1j})\omega(z_l'\gamma_0,z_i'\gamma_0)\right]\psi_i \qquad (5\\
= \;& 2\,\mathrm{E}\left[h_2^{-1}h_1^{-1}\int_{\mathbb{R}^{d_x}}\int_{-\frac{1}{h_2}}^{0} s_j\iota_1'\left[f(w_{1j})\mathbf{M}_1\right]^{-1}\mathcal{K}^{(1)}_{d_x+1}(\nu,\nu_{d_x+1})\omega(z_l'\gamma_0,z_i'\gamma_0)f(w_{-i}+h_2\nu,1+h_2\nu_{d_x+1})d\nu d\nu_{d_x+1}\middle| z_i\right] \\
= \;& 2\,\mathrm{E}\left[h_2^{-1}h_1^{-1}\int_{\mathbb{R}^{d_x}}\left[s_j\iota_1'\left[f(w_{1j})\mathbf{M}_1\right]^{-1}\mathcal{K}(\nu,\nu_{d_x+1})\omega(z_l'\gamma_0,z_i'\gamma_0)\times f(w_{-i}+h_2\nu,1+h_2\nu_{d_x+1})d\nu\right]_{-\frac{1}{h_2}}^{0} d\nu_1\middle| z_i\right]\psi_i \\
& -2\,\mathrm{E}\left[h_1^{-1}\int_{\mathbb{R}^{d_x}}\int_{-\frac{1}{h_2}}^{0} s_j\iota_1'\left[f(w_{1j})\mathbf{M}\right]^{-1}\mathcal{K}(\nu,\nu_{d_x+1})\omega(z_l'\gamma_0,z_i'\gamma_0)D^1_{d_x+1}f(w_{-i}+h_2\nu,1+h_2\nu_{d_x+1})d\nu\middle| z_i\right]\psi_i,
\end{aligned}
$$

Given Assumptions E7, E9, E10, E11, it is straightforward to see that the term in Equation (55) is of order $O(h_2^{-1})$. Using similar arguments it is also possible to derive that:

$$
\mathrm{E}\left[\left\|h_2^{-d_x-2}h_1^{-1}\,\mathrm{E}\Big[\Psi^{B1}_{p,ijl}\Big|\varpi_i\Big]\right\|^2\right] = O(h_2^{-1}).
$$

Therefore, by a standard CLT argument:

$$
\frac{\sqrt{nh_2}}{h_2^{-d_x-2}h_1 n}\sum_{i=1}^{n}\mathrm{E}[\Psi^{B1}_{p,ijl}|\varpi_i] = O_p(1)
$$

Next, we turn to the second term of Equation (53) involving $\Xi_n(z_j'\gamma_0)$. The first steps are as before:

$$
\begin{aligned}
& \frac{\sqrt{nh_2}}{n^2h_2^{d_x+2}}\sum_{i=1}^{n}\sum_{j=1}^{n} s_i\iota_1'\left[f(\widehat{w}_i)\mathbf{M}\right]^{-1}\overline{\mathcal{K}}^{(1)}_{j,2}(x_i'\widehat{\beta},1)\Xi_n(z_j'\gamma_0) \\
\leq \;& \max_{1\leq j\leq n}\left|\Xi_n(z_j'\gamma_0)\right|\left\|\frac{\sqrt{nh_2}}{n^2h_2^{d_x+2}}\sum_{i=1}^{n}\sum_{j=1}^{n} s_i\iota_1'\left[f(w_i)\mathbf{M}\right]^{-1}\mathcal{K}^{(1)}_{j,2}(w_{i1})\right\| + o_p(1) \qquad (56)
\end{aligned}
$$

The first term on the RHS of the last inequality is of order $O_p(h_1^2)$ by Assumption E7. For the second part, note again that observations with $i=j$ can be neglected asymptotically, so that:

$$
\begin{aligned}
& \frac{\sqrt{nh_2}}{n^2h_2^{d_x+2}}\sum_{i=1}^{n}\sum_{j=1}^{n} s_i\iota_1'\left[f(w_i)\mathbf{M}_1\right]^{-1}\mathcal{K}^{(1)}_{j,2}(w_{i1}) \\
= \;& \frac{\sqrt{nh_2}}{2h_2^{d_x+2}}\binom{n}{2}^{-1}\sum_{i=1}^{n}\sum_{j>i}\left\{s_i\iota_1'\left[f(w_i)\mathbf{M}_1\right]^{-1}\mathcal{K}^{(1)}_{j,2}(w_{i1}) + s_j\iota_1'\left[f(w_j)\mathbf{M}_1\right]^{-1}\mathcal{K}^{(1)}_{i,2}(w_{1j})\right\} + o_p(1).
\end{aligned}
$$

Standard Hoeffding decomposition arguments, change of variables, and integration by parts, yield that the lead term of the projection is of order:

$$
\frac{\sqrt{nh_2}}{h_2^{d_x+2}}\,\mathrm{E}\left[s_i\iota_1'\left[f(w_i)\mathbf{M}_1\right]^{-1}\mathcal{K}^{(1)}_{j,2}(w_{i1}) + s_j\iota_1'\left[f(w_j)\mathbf{M}_1\right]^{-1}\mathcal{K}^{(1)}_{i,2}(w_{1j})\right]
$$

is of order $= O\left(\sqrt{\frac{n}{h_2}}\right)$. Moreover, similar arguments to the ones used for $\mathcal{T}_{m1,n,122}$ in Equation (49) can be used to show that the second term in Equation (52) is of smaller order than the first term. Thus, collecting expressions after these lengthy arguments and using Assumption E13, we conclude that $\mathcal{T}_{m,n,2} = o_p(\sqrt{h_2})$ uniformly in $\mathbf{b}$.

Finally, similar arguments to the ones used in the proof of Lemma B.2 yield that $\mathcal{T}_{m1,n,3}$, and $\mathcal{T}_{m1,n,4}$ are of order $O_p(\sqrt{h_2})$ uniformly in $\mathbf{b}$. This completes the proof. ■

Proof of Lemma B.4. A standard mean value expansion yields:

$$-\frac{\sqrt{nh_2}}{n}\sum_{i=1}^{n} s_i D^{\mathbf{1}}_{d_x+1} m(\overline{w}_i)\Big\{\widehat{p}_i - p_i\Big\}.$$

Moreover, after a second mean value expansion around $D^{\mathbf{1}}_{d_x+1}m(w_i)$ and using Assumption E7, the term can be written as:

$$\frac{\sqrt{nh_2}}{n}\sum_{i=1}^{n} s_i D^{1}_{d_x+1} m(w_i)\Big\{\frac{1}{nh_1}\sum_{l=1}^{n}\omega(z_i'\gamma_0, z_l'\gamma_0)\psi_l + \Xi_n(z_i'\gamma_0)\Big\} + o_p(1).$$

The same arguments used in the proof of Lemma B.2 and B.3 together with Assumption E13 can be used to derive that also the second part is of order $O_p(\sqrt{h_2})$ uniformly in $\mathbf{b}$. This establishes the claim. ■

Proof of Lemma B.5. The proof follows using the same arguments as in the proof of Lemma B.4. ■

Proof of Lemma B.6. Using the representation from Equation (21), note again that:

$$\begin{aligned}
&\frac{\sqrt{nh_2}}{n}\sum_{i=1}^{n} s_i \frac{\widetilde{u}_i}{m(w_i)}(\widehat{m}(\widehat{w}_{i1}) - m(\widehat{w}_{i1})) \qquad (57)\\
=\ &\frac{\sqrt{nh_2}}{n}\sum_{i=1}^{n} s_i \frac{\widetilde{u}_i}{m(w_i)}\iota_1'\widehat{\mathbf{M}}_n^{-1}(w_{i1})\left\{\frac{1}{nh_2^{d_x+1}}\sum_{j=1}^{n}\widehat{\mathcal{K}}_j(w_{i1})u_j\right\}\\
&+\frac{\sqrt{nh_2}}{n}\sum_{i=1}^{n} s_i \frac{\widetilde{u}_i}{m(w_i)}\iota_1'\widehat{\mathbf{M}}_n^{-1}(w_{i1})\left\{\frac{1}{nh_2^{d_x+1}}\sum_{j=1}^{n}\widehat{\mathcal{K}}_j(w_{i1})\widehat{\Delta}_{m_1,j}\right\},
\end{aligned}$$

where

$$\widehat{\Delta}_{m_1,j} = m(w_{i1}) - \sum_{0\le|\mathbf{k}|\le q}\frac{1}{\mathbf{k}!}D^{\mathbf{k}}m(w_{i1})(\widehat{w}_j - w_{i1})^{\mathbf{k}}.$$

Defining $\widehat{V}_{m_1,n}(w_{i1})$ and $\widehat{B}_{m_1,n}(w_{i1})$ in a similar way to $\widehat{V}_{m,n}(\widehat{w}_i)$ and $\widehat{B}_{m,n}(\widehat{w}_i)$ in the proof of Lemma

B.2, we can write:

$$\begin{aligned}
&\frac{\sqrt{nh_2}}{n}\sum_{i=1}^{n} s_i \frac{\widetilde{u}_i}{m(w_i)} \iota_1' \widehat{\mathbf{M}}_n^{-1}(w_{i1}) \left\{ \widehat{V}_{m_1,n}(w_{i1}) + \widehat{B}_{m_1,n}(w_{i1}) \right\} \\
= \;& \frac{\sqrt{nh_2}}{n}\sum_{i=1}^{n} s_i \frac{\widetilde{u}_i}{m(w_i)} \iota_1' \left[f(w_{i1})\mathbf{M}_1\{1+o_p(1)\} \right]^{-1} \widehat{V}_{m_1,n}(w_{i1}) \\
&+ \frac{\sqrt{nh_2}}{n}\sum_{i=1}^{n} s_i \frac{\widetilde{u}_i}{m(w_i)} \iota_1' \left[f(w_{i1})\mathbf{M}_1\{1+o_p(1)\} \right]^{-1} \widehat{B}_{m,n}(w_{i1}) \qquad (58) \\
&- \frac{\sqrt{nh_2}}{n}\sum_{i=1}^{n} s_i \frac{\widetilde{u}_i}{m(w_i)} \iota_1' \left[\mathbf{M}_1 f(w_{i1})\{1+o_p(1)\} \right]^{-1} \left[\widehat{\mathbf{M}}_n(w_{i1}) - \mathbf{M}_1 f(w_{i1}) \right] \widehat{\mathbf{M}}_n^{-1}(w_{i1}) \widehat{V}_{m,n}(w_{i1}) \\
&- \frac{\sqrt{nh_2}}{n}\sum_{i=1}^{n} s_i \frac{\widetilde{u}_i}{m(w_i)} \iota_1' \left[\mathbf{M}_1 f(w_{i1})\{1+o_p(1)\} \right]^{-1} \left[\widehat{\mathbf{M}}_n(w_{i1}) - \mathbf{M}_1 f(w_{i1}) \right] \widehat{\mathbf{M}}_n^{-1}(w_{i1}) \widehat{B}_{m,n}(w_{i1}) \\
\equiv \;& \widetilde{\mathcal{T}}_{m1,n,1} + \widetilde{\mathcal{T}}_{m1,n,2} + \widetilde{\mathcal{T}}_{m1,n,3} + \widetilde{\mathcal{T}}_{m1,n,4}.
\end{aligned}$$

Starting with $\widetilde{\mathcal{T}}_{m1,n,1}$, we can again write:

$$\begin{aligned}
&\frac{\sqrt{nh_2}}{n}\sum_{i=1}^{n} s_i \frac{\widetilde{u}_i}{m(w_i)} \iota_1' \left[f(w_{i1})\mathbf{M}_1 \right]^{-1} V_{m_1,n}(w_{i1}) \\
&+ \frac{\sqrt{nh_2}}{n}\sum_{i=1}^{n} s_i \frac{\widetilde{u}_i}{m(w_i)} \iota_1' \left[f(w_{i1})\mathbf{M}_1 \right]^{-1} \left\{ \widehat{V}_{m_1,n}(w_{i1}) - V_{m_1,n}(w_{i1}) \right\} + o_p(1) \\
\equiv \;& \widetilde{\mathcal{T}}_{m1,n,11} + \widetilde{\mathcal{T}}_{m1,n,12} + o_p(1).
\end{aligned}$$

Consider $\widetilde{\mathcal{T}}_{m1,n,11}$, which drives the distribution. Define the symmetrized 'kernel function':

$$\widetilde{\Psi}_{u1,ij} \equiv s_i \frac{\widetilde{u}_i}{m(w_i)} \iota_1' \left[f(w_{i1})\mathbf{M}_1 \right]^{-1} \mathcal{K}_j(w_{i1})\widetilde{u}_j + s_j \frac{\widetilde{u}_j}{m(w_j)} \iota_1' \left[f(w_{1j})\mathbf{M}_1 \right]^{-1} \mathcal{K}_i(w_{1j})\widetilde{u}_i$$

Then, using the same argument of Ahn and Powell (1993, p.25) to show that terms with $i = j$ are asymptotically bounded by:

$$\frac{1}{n^2 h_2^{d_x+1}} \sum_{i=1}^{n} \left\| s_i \frac{\widetilde{u}_i}{m(w_i)} \iota_1' \left[f(w_{i1})\mathbf{M}_1 \right]^{-1} \mathcal{K}_i(0,1)\widetilde{u}_i \right\| = O_p\left(\frac{1}{nh_2^{d_x}} \right) = o_p\left(\frac{1}{\sqrt{nh_2}} \right),$$

where $\mathcal{K}_i(0,1)$ is defined as $\mathcal{K}_j(w_{i1})$, but with the first d_x elements equal to 0 and the last entry equal to $(p_i - 1)$. Then, derive the (appropriately rescaled) U-statistic:

$$\begin{aligned}
&\frac{\sqrt{nh_2}}{n}\sum_{i=1}^{n} s_i \frac{\widetilde{u}_i}{m(w_i)} \iota_1' \left[f(w_{i1})\mathbf{M}_1 \right]^{-1} V_{m_1,n}(w_{i1}) \\
= \;& \frac{\sqrt{nh_2}}{2h^{d_x+1}} \binom{n}{2}^{-1} \sum_{i=1}^{n} \sum_{j>i} \widetilde{\Psi}_{u1,ij} + o_p(1) \\
= \;& \frac{\sqrt{nh_2}}{nh^{d_x+1}} \sum_{i=1}^{n} \mathrm{E}[\widetilde{\Psi}_{u1,ij} | \varpi_i] + o_p(1),
\end{aligned}$$

where the second equality follows from a standard Hoeffding decomposition,the fact that $\mathrm{E}[\widetilde{\Psi}_{u1,ij}] = 0$ by iterated expectations, and from Lemma 3.1 in Powell et al. (1989) since $\mathrm{E}\left[h_2^{-2(d_x+1)} \left\|\widetilde{\Psi}_{u1,ij}\right\|^2\right] = O(h_2^{-(d_x+1)}) = o(n)$. Moreover, noting that also $\mathrm{E}[h_2^{-(d_x+1)}\widetilde{\Psi}_{u1,ij}|\varpi_i] = 0$ since $\widetilde{u}_i$ and $\widetilde{u}_j$ are uncorrelated, it is immediate to see that:

$$\frac{\sqrt{nh_2}}{2h_2^{d_x+1}}\binom{n}{2}^{-1}\sum_{i=1}^{n}\sum_{j>i}\widetilde{\Psi}_{u1,ij} = O_p(\sqrt{h_2}).$$

Similar arguments as in the proof of Lemma B.3 can be used to show that $\widetilde{\mathcal{T}}_{m1,n,12} = O_p(\sqrt{h_2})$. This also holds for the remaining terms, namely $\widetilde{\mathcal{T}}_{m1,n,2}$, $\widetilde{\mathcal{T}}_{m1,n,3}$, and $\widetilde{\mathcal{T}}_{m1,n,4}$. ■

Proof of Lemma B.7. A standard mean value expansion yields:

$$-\frac{\sqrt{nh_2}}{n}\sum_{i=1}^{n} s_i \frac{\widetilde{u}_i}{m(w_i)} D^{\mathbf{1}}_{d_x+1} m(\overline{w}_i)\Big\{\widehat{p}_i - p_i\Big\}.$$

Moreover, after a second mean value expansion around $D^{\mathbf{1}}_{d_x+1}m(w_i)$ and using Assumption E7, this can be written as:

$$\frac{\sqrt{nh_2}}{n}\sum_{i=1}^{n} s_i \frac{\widetilde{u}_i}{m(w_i)} D^{\mathbf{1}}_{d_x+1} m(w_i)\Big\{\frac{1}{nh_1}\sum_{l=1}^{n}\omega(z_i'\gamma_0, z_l'\gamma_0)\psi_l + \Xi_n(z_i'\gamma_0)\Big\} + o_p(1).$$

The same arguments used in the proof of Lemma B.2 and B.3 together with Assumption E13 can be used to derive that also the second part is of order $O_p(\sqrt{h_2})$ uniformly in $\mathbf{b}$. This establishes the claim. ■

Proof of Lemma B.8. The proof follows the same arguments as the proof of Lemma B.6. ■

Proof of Lemma B.9. The proof follows the same arguments as the proof of Lemma B.7. ■

References

Ahn, H. and J. Powell (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics 58*, 3–29.

Andrews, D. and M. Schafgans (1998). Semiparametric estimation of the intercept of a sample selection model. *Review of Economic Studies 65*, 497–517.

Cai, Z. (2002). Regression quantiles for time series. *Econometric Theory 18*(1), 169–192.

Calonico, S., M. Cattaneo, and R. Titiunic (2014). Robust nonparametric confidence interval for regression discontinuity design. *Econometrica 82*, 2295–2326.

Calonico, S. and J. Smith (2017). The women of the national supported work demonstration. *Journal of Labor Economics 35*(S1), S65–S97.

Chamberlain, G. (1986). Asymptotic efficiency in semi-parametric models with censoring. *Journal of Econometrics 32*, 189–218.

Chen, S. (1999). Distribution-free estimation of the random coefficient dummy endogenous variable model. *Journal of Econometrics 91*, 171–199.

Das, M., W. K. Newey, and F. Vella (2003). Nonparametric estimation of sample selection models. *Review of Economic Studies 70*(1), 33–58.

Dehejia, R. and S. Wahba (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the Americal Statistical Association 94*(448), 1053–1062.

Dehejia, R. and S. Wahba (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics 84*(1), 151–161.

Fan, J. and I. Gijbels (1992). Variable bandwidth and local linear regression smoothers. *The Annals of Statistics 20*(4), 2008–2036.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and Its Applications*. Number 66 in Monographs on Statistics and Applied Probability. Chapman and Hall/ CRC.

Gelman, A. and G. Imbens (2014). Why high-order polynomials should not be used in regression discontinuity designs. Working Paper 20405, NBER.

Goh, C. (2017). Rate-optimal estimation of the intercept in a semiparametric sample-selection model. Unpublished manuscript, University of Wisconsin-Milwaukee.

Gutknecht, D. (2016). Testing for monotonicity under endogeneity - an application to the reservation wage function. *Journal of Econometrics 190*, 100–114.

Hall, P. and J. Racine (2015). Infinite order cross-validated local polynomial regression. *Journal of Econometrics 185*, 510525.

Hall, P., R. Wolff, and Q. Yao (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association 94*(445), 154–163.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica 47*, 153–161.

Heckman, J. (1990). Variety of selection bias. *American Economic Review 80*(2), 679–694.

Heckman, J. and P. Todd (2009). A note on adapting propensity score matching and selection models to choice based samples. *Econometrics Journal 12*(21), S230–S234.

Imbens, G. and K. Kalyanaraman (2012). Optimal bandwidth choice for regression discontinuity estimator. *Review of Economic Studies 79*, 933–959.

Jochmans, K. (2015). Multiplicative-error models with sample selection. *Journal of Econometrics 184*, 315–327.

Kitagawa, T. (2010). Testing for instrument independence in the selection model. Unpublished manuscript, UCL.

Klein, R. and R. Spady (1993). An efficient semiparametric estimator for binary response models. *Econometrica 61*(2), 387–421.

LaLonde, R. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review 76*(4), 604–620.

Lewbel, A. (2007). Endogenous selection or treatment model estimation. *Econometric Theory 13*, 32–51.

Li, Q. and J. Wooldridge (2002). Semiparametric estimation of partially linear models for dependent data with generated regressors. *Economtric Theory 18*(3), 625–645.

Manski, C. and D. Mc Fadden (1981). Statistical analysis of discrete probability models. In C. Manski and D. Mc Fadden (Eds.), *Structural Analysis of Discrete Data with Econometric Applications*, pp. 2–49. Cambridge, MA, MIT Press.

Masry, E. (1996). Multivariate local polynomial regression for time series: Uniform strong consistency and rates. *Journal of Time Series Analysis 17*(6), 571–599.

Mroz, T. (1987). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica 55*(4), 765–799.

Newey, W., J. Powell, and J. Walker (1990). Semiparametric estimation of selectionmodels: Some empirical results. *American Economic Review Papers & Proceedings 80*(2), 324–328.

Newey, W. K. (2009). Two-step series estimation of sample selection models. *Econometrics Journal 12*, 217–229.

Newey, W. K. and D. Mc Fadden (1994). Large sample estimation and hypothesis testing. In R. Engle and D. Mc Fadden (Eds.), *Handbook of Econometrics*, Volume 4, Chapter 36, pp. 2111–2245. Elsevier.

Powell, J., J. Stock, and T. Stoker (1989). Semiparametric estimation of index coefficients. *Econometrica 57*, 1403–1430.

Ruppert, D. and M. Wand (1994). Multivariate locally weighted least squares regression. *The Annals of Statistics 22*(3), 1346–1370.

Schafgans, M. (1998). Ethnic wage difference in malaysia: Parametric and semiparametric estimation of the chinese-malay gap. *Journal of Applied Econometrics 13*, 481–504.

Schafgans, M. (2000). Gender wage difference in malaysia: Parametric and semiparametric estimation. *Journal of Develoment Economics 63*, 351–368.

Schafgans, M. and V. Zinde-Walsh (2002). On intercept estimation in the sample selection model. *Econometric Theory 18*, 40–50.

Smith, J. and P. Todd (2005). Does matching overcome lalonde's critique of nonexperimental estimators? *Journal of Econometrics 125*, 305–353.

Terza, J. (1998). Estimating count data models with endogenous switching: Sample selection and endogenous treatment effects. *Journal of Econometrics 84*, 129–154.