Additive Nonparametric Sample Selection Models with Endogeneity

Deniz Ozabaci*
Department of Economics
University of Connecticut

Abstract

Preliminary, please do not cite without permission.

This paper presents two additive nonparametric structural equation models to correct for sample selection bias and endogeneity. Our models can be considered as flexible version of Heckman's (1976,1979) two-step sample selection models. Similar examples exist in the literature but, to our knowledge, we are the first to look at both problems under the additive nonparametric setting, which are difficult problems to solve without parametrization. In one of our models we treat the endogenous variable as continuous, whereas in the other we assume that it is discrete. Our models enable us to keep a parameter free flexible setting, and our estimators are free from the curse of dimensionality. We provide multi-stage estimators for both models by combining series estimators with kernel regression. We show that the estimates for the smooth functions and their gradients are oracle efficient, i.e., we can estimate our smooth functions and their gradients with the precision of nonparametric regression. We also discuss a partially linear extension of our model, which may be preferable in microeconomic applications with many control variables. We test the performance of our estimators using Monte Carlo simulations. Using these simulations we show that our estimators are consistent and perform better compared to the nonparametric alternatives. Finally, we apply our model to an empirical problem and consider how maternal employment and child care affect childrens cognitive ability. We approach maternal employment decision as a sample selection problem and treat child care as endogenous. Our results show that maternal employment matters, and that effects of childcare depend on childcare type.

Keywords: Additive Regression, Sample Selection, Discrete Regressors, Endogeneity, Generated Regressors, Oracle Estimation, Nonparametric, Structural Equation
JEL Classification Codes: C14, C36, I21, J13

*Deniz Ozabaci, Department of Economics, University of Connecticut, 365 Fairfield Way, U-1063 Storrs, CT, 06269, e-mail:deniz.ozabaci@uconn.edu.

1 Introduction

In microeconomic applications, subsamples of a population are subjects of interest, such as working females, an eligible group for a particular benefit, or a subgroup that is partially observed in a survey. Estimation of such models becomes problematic when the selection of the subsample is nonrandom, or, in other words, when part of the selection process is relevant for the outcome. In this case, sample selection bias occurs if the selection process is not accounted for in the econometric methodology. If, on the other hand, the selection is random, then selection can be ignored, since the characteristics of the subsample will be similar to those of the population. Take, for example, working females; the decision to participate in the labor force may no longer be independent of the determinants of wage. Hence, sample selection should be corrected.

Another common problem in such applications is endogeneity, which is observed when a regressor is correlated with the error component. This problem may be seen in the case of omitted variables, measurement errors and simultaneity. It is also possible to observe both problems in a single economic application. If any of these three factors are seen in the analysis of working females' labor force participation, for example, the econometric model will have to correct for both endogeneity and sample selection. Hence, we are interested in a model that provides a solution to both of these problems, but we also want to keep our model flexible enough to be able to maximize the information we can get out of it.

The models we are interested in are sample selection models with endogeneity. We solve the aforementioned issues using simultaneous equation models, with all stages assumed to be additive nonparametric. We choose an additivity constraint to overcome the curse of dimensionality problem of fully nonparametric regression. An empirical researcher may think of several alternatives to additive nonparametric regression. We chose not to have a fully nonparametric model to be able to circumvent curse of dimensionality. Curse of dimensionality is the term used for the problem of slower convergence with increasing dimension size given a sample. The only theoretical solution is to increase the data size exponentially with the growing dimension. However, empirically this is usually impossible. What we see in our estimations when this problem is present are higher standard errors and unexpected signs for gradient estimates. We are interested in solving these two important empirical problems using the most flexible model possible, while keeping the estimates' efficiency high. For this reason we pick additive nonparametric regression as opposed to nonparametric regression and other semi parametric models. Another alternative to the model we adopt here may be parametric models, such as polynomials of a certain order, or nonlinear methods with certain distributional assumptions. It is well known that if either the functional form or distributional assumptions fail, the estimators become inconsistent. Any conclusion of an inconsistent model will be unreliable. If one is willing to take the risk, still these estimators may be problematic. Nonlinear methods, such as nonlinear least squares, maximum likelihood, etc., depending on the variation in the data, may be hard to estimate as well. Optimization problems are usually difficult empirical problems that people tend to overlook. The estimator we propose in here runs quickly and is easy to estimate. One can also estimate several linear models for different subgroups. Overlooking the inconsistency risk, this method may work. However, even when it is performed for the subsamples, only a limited level of heterogeneity can be presented. This is because the estimator will be giving information about the mean alone. Also, for each subsample the degrees of freedom will be effected. With our estimator we are able to present the conditional mean, conditional expectations, and the gradients for the overall sample as well as for each observation. For the type of empirical questions our model may answer, there will be

data limitations, likely a high level of heterogeneity and nonlinearities; hence we propose that additive nonparametric models may work better. Given the myriad econometric models available in the literature, we have to stress that this is our preference. One can come up with other alternatives which may work. However, we believe that we have compelling reasons given the problems we target.

Our focus will be on two models. In the first model (Model 1: Sample Selection Model with Endogenous Regressors), we assume that the endogenous variable is continuous. In the second model (Model 2: Sample Selection Model with Discrete Endogenous Regressors), we focus on a special case of Model 1 and assume that the endogenous variable is discrete over finite support.

We propose two multistage estimation procedures for these models, and both estimators combine a series estimator (smoothing splines) and kernel regression (local linear least squares regression, LLLS). The estimator for Model 1 is in line with the control function approach. For Model 2, due to the discrete nature of the endogenous variable, we propose an additive nonparametric version of the two-stage least squares regression. Both of these estimators correct for sample selection bias in addition to endogeneity. We show that our estimators for both the smooth functions of interest and their gradients exhibit oracle efficiency, which means that they can be estimated at the asymptotic accuracy of single dimensional nonparametric regression. We also show our estimators' finite sample performance using Monte-Carlo simulations. The simulations support our findings.

Finally, using our estimators we analyze the effect of maternal employment and childcare on the development of childrens cognitive ability. We take a slightly different approach to this problem than is typical in the literature. First, our analysis involves the application of a novel econometric model. Second, we analyze the effects of both maternal employment and childcare use, treating the latter as endogenous while approaching the former as a sample selection problem. We predict that working and non-working mothers will be systematically different, both in terms of decisions regarding their childs cognitive ability as well as outcomes for the child. We analyze the effects of three different child care variables over working and non-working mothers separately. Our results show that the effects of childcare change depending on maternal employment. Formal and informal child care provided by non-relatives have positive effects at the median for working mothers' children, whereas informal child care provided by relatives affect child's ability negatively. This is an interesting result when combined with the estimates for non-working mothers, for which the partial effect signs are reversed. This suggests quality effects for childcare. By the help of the high flexibility our model allows, we can also see the overall distribution of the estimated gradients. We can see that median results may be misleading since we find both negative and positive estimates across all childcare groups and types. The only exception is the partial effect of informal childcare provided by relatives to the children of working mothers. The effect is always negative and significant. We also find that the effect of formal childcare is uniformly greater compared to informal childcare provided by non-relatives. This same result is not seen for non-working mothers, however. We conclude that working mother may be able to afford better quality childcare which results in greater benefit for their children. The rest of the paper is organized as follows. In Section 2, we provide a description of our two models. Section 3 explains the estimators we propose. Section 4 discusses the asymptotic properties of our estimators, whose finite sample properties are tested and presented in Section 5. Section 6 presents the application and our results. Section 7 concludes.

2 Methodology

Similar models to the ones we propose do exist in the literature. However, our models are unique in that they are the first examples to solve both endogeneity and sample selection bias using additive nonparametric setting. Also, the estimator we propose improve upon the existing procedures by achieving oracle efficiency. Here, we adopt the triangular set-up of Newey et al. (1999), where they correct for endogeneity using a nonparametric version of the control function approach. Su and Ullah (2008) add to their model by proposing an empirically efficient kernel estimator for the triangular system, and they provide asymptotic normality proofs, as well. All of these models are fully nonparametric. As it is well known, nonparametric regression suffers from the curse of dimensionality, which means that the convergence rate is a function of the dimension of the function of interest. This causes higher variability in the estimates, unless it is paired with a substantial increase in data. Hence, this is a disadvantage for empirical applications with many variables and data limitations. One other disadvantage of such models is their high computing demand. As a solution to this, Ozabaci et al. (2014) impose an additivity constraint on the triangular set-up of Newey et. al. (1999) and Su and Ullah (2008). Moreover, the estimator proposed in Ozabaci et al. (2014) is free from the curse of dimensionality, and the conditional mean and gradient estimates are also oracle efficient and asymptotically normal.

2.1 Model 1: Sample Selection Model with Endogenous Regressors

The first model discussed in this paper can be presented as:

$$\begin{cases} Y^* = \mu_y + g_1(\mathbf{X}) + g_2(\mathbf{Z_1}) + \varepsilon, \\ \mathbf{X} = \mu_x + \pi_1(\mathbf{Z}_1) + \pi_2(\mathbf{Z}_2) + \mathbf{U}, \\ D = 1(\mu_d + m_1(\mathbf{Z}_1) + m_2(\mathbf{Z}_2) - V > 0), \\ Y = DY^*, \end{cases} \tag{2.1}$$

where Y^* is the latent variable, which is observed conditional on the factor D. D is the binary selection indicator. X is the continuously distributed endogenous variable. Z_1 is the included exogenous regressor. Z_2, on the other hand, is the excluded exogenous regressors.[1] This model is similar to the one from Ozabaci et al. (2014) since the first model also corrects for endogeneity using a similar structural equation model and assumes that the endogenous variable is continuous. However, here we also correct for sample selection bias.

For identification, and in order to reach the associated estimators, we need certain assumptions on the endogenous variable(s) and the error components:

$$\begin{aligned} &E(\mathbf{U}|\mathbf{Z}_1, \mathbf{Z}_2) = 0, \\ &E(\varepsilon|\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{U}, D = 1) = \lambda_1(\mathbf{P}) + \lambda_2(\mathbf{U}). \end{aligned} \tag{2.2}$$

(2.2) follows from Newey et al. (1999), and Das et al. (2003). P is the propensity score, where $P = E(D|\mathbf{Z}_1, \mathbf{Z}_3)$. Under these assumptions, it follows that:

$$E(Y|X, Z, D = 1) = g(\mathbf{X}, \mathbf{Z_1}, \mathbf{P}, \mathbf{U}) = \mu_y + g_1(\mathbf{X}) + g_2(\mathbf{Z_1}) + \lambda_1(\mathbf{P}) + \lambda_2(\mathbf{U})$$

[1] For clear representation we show all variables as single dimensional, however the general case can be straightforwardly obtained.

Here we have a more flexible version of Heckman's (1974,1979) model, and an additive separable nonparametric version of Newey (2009). Similar to those, we also model the sample selection term in terms of propensity score (Heckman and Robb (1996), Ahn and Powel (1993), Das et. al. (2003)). Our model also bears an affinity to that of Das et al. (2003). They correct for both sample selection and endogeneity using a structural equation model comparable to Newey et al. (1999). They assume that all the equations in their model are fully nonparametric. Our work adds to theirs by imposing an additivity constraint on all the stages of the model. Additionally, the estimator we propose differs from that found in Das et al. (2003) because our multi-stage estimator achieves oracle efficiency and is free from the curse of dimensionality.

In our model, we need the dimension restriction indicating that the dimension of Z_2 (d_{z_2}) is positive, which is the rank condition needed for identification. The rest of the identification conditions follow directly from Das et. al. (2003). We can also allow for different regressors in the first step regressions, as long as this dimension restriction holds.

2.2 Model 2: Sample Selection Model with Discrete Endogenous Regressors

Model 2 also corrects for endogeneity, but it is different from alternatives in extant literature as well as Model 1, because it allows the endogenous variable to be discrete. Due to the discrete nature of the endogenous regressor, we can no longer use the control function approach. Hence, for this model's identification, we use a transform of the model similar to the work by Das (2005). Like Das (2005), we evaluate the discrete variable over its finite support points, which allows us to reach the transform. Our Model 2 departs from that of Das (2005) because, in addition to endogeneity, it also corrects for sample selection, and all our structural equations are additive nonparametric.

The principal difference here compared to Model 1 is that the endogenous variable X is assumed to be discrete over finite support. If X is discrete, ε and Z_1 and Z_2 can no longer be independent. Due to the discontinuity in the variable, the residuals can no longer be obtained in a continuous manner. Hence, we will need new assumptions that will lead us to a different estimation scheme.

In addition to the assumptions on error components, we also need to identify the discrete variable within a nonparametric function. To achieve this, we start with the approach of Das (2005) and evaluate (2.1) over the discrete and finite support points of X. This gives us

$$\begin{aligned}
Y^* &= \mu_y + g_1(\mathbf{X}) + g_2(\mathbf{Z_1}) + \varepsilon \\
&= \mu_y + g_2(\mathbf{Z_1}) + [\beta_1 1\{X = X_1\} + \beta_2 1\{X = X_2\} + \cdots + \beta_J 1\{X = X_J\}]\,\mathbf{X} + \varepsilon \\
&= \mu_y + g_2(\mathbf{Z_1}) + \beta_1 1\{X = X_1\} + \beta_2 1\{X = X_2\} + \cdots + \beta_J \left(1 - \sum_{j=1}^{J-1} 1\{X = X_j\}\right) + \varepsilon \\
&= (\mu_y + \beta_J) + g_2(\mathbf{Z_1}) + \sum_{j=1}^{J-1} \beta_j 1\{X = X_j\},
\end{aligned}$$

where $\{x_1, x_2, \cdots, x_J\}$ are J distinct support points for X, provided that $|x_j| < \infty$ for $j = \{1, 2 \cdots J\}$. β_j (for $j = \{1, 2, J-1\}$), on the other hand, are the adjustment factors. For a binary X, Y^* can be written as:

$$Y^* = \mu + \beta\mathbf{X} + g_2(\mathbf{Z_1}) + \varepsilon \tag{2.3}$$

where $\mu = \mu_y + \beta_1$, and, following this, β can be presented as β_0, as well.

In order to identify the full model, we need to make certain assumptions on the error components, hence (2.2) becomes:

$$\begin{aligned} &E\left(\mathbf{U}|\mathbf{Z}_1, \mathbf{Z}_2\right) = 0 \\ &E\left(\varepsilon|\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3, D = 1\right) = \lambda\left(\mathbf{P}\right), \end{aligned} \tag{2.4}$$

which gives us the conditional mean as (keeping the assumption that X is binary):

$$E\left(Y|X, Z, D = 1\right) \quad = \quad \beta E\left(\mathbf{X}|Z_1, Z_2\right) + g_2\left(\mathbf{Z_1}\right) + \lambda_1\left(\mathbf{P}\right)$$

Here, we still need d_{z_2} to be positive. More specifically, we need $V\{(\mu_x + \pi_1\left(\mathbf{Z}_1\right) + \pi_2\left(\mathbf{Z}_2\right))\,|Z_1\} > 0$ to hold for identification of the conditional mean up to an additive constant.

For the remainder of this paper, including technical proofs, we will assume X to be a binary variable. This is for clear representation, and non-binary discrete X can be shown similarly.

3 Estimation

We propose a three stage estimation procedure for both of our models. The two estimators are different in that they require distinct assumptions to correct for endogeneity. For both models, we show that the additive smooth functions and their gradients can be estimated at the asymptotic accuracy of a single dimensional nonparametric regression.

We use series estimator for the first two stages of the estimators, and turn to kernel regression via one step back fitting.

Model 1: Sample Selection Model with Endogenous Regressors For the first model, we use a similar approach to Ozabaci et al. (2014). However to account for the sample selection bias, we have two first stage regressions and an additional additive smooth function at the final stage. We can summarize our estimation procedure as:

1.a Estimate the propensity score, P, via $E[D|Z_1, Z_2] \equiv m(Z_1, Z_1)$ using a series estimator, and reach $\widetilde{P}$.

1.b Estimate the conditional mean of X, via $E[X|Z_1, Z_2] \equiv \pi(Z_1, Z_2)$, again using a series estimator.

2 Obtain the residuals, $\widetilde{U}$, from Step 1.b as $X - \widetilde{\pi}(Z_1, Z_2)$, and regress Y on $\widetilde{U}$, $\widetilde{P}$, X and Z_1 using a series estimator.

3 Obtain partial residuals from the previous step and regress these partial residuals on the related variable using LLLS. For Z_1, the LLLS kernel regression will be of $\widetilde{Y}_1$ on Z_1; where $\widetilde{Y}_1 = Y - \widetilde{g}_1(X) - \widetilde{\lambda}_1(\widetilde{U}) - \widetilde{\lambda}_2(\widetilde{P})$. The remaining smooth functions can be obtained in a similar fashion.

The introduction of the third step is in line with Ozabaci et al. (2014). It ensures that the final stage estimates for the smooth functions and their gradients are free from the approximation bias and estimation error from the prior steps. The two first step regressions are similar to Das et al. (2003). We can show that the asymptotic properties of the second step estimates will be affected from the first

step. The final stage kernel regression combined with back fitting (using partial residuals) gives us oracle efficient smooth function estimates and gradients.

For the series estimators, we use smoothing splines considering their faster convergence rates compared to power series. Basis functions, $\rho(\cdot)$, are assumed to be capable of approximating any square integrable function well.

Let $\rho_j(\cdot)$ be a set of basis functions, and $\kappa_1 = \kappa_1(n)$ and $\kappa = \kappa(n)$ be integers such that $\kappa \to \infty$ as $n \to \infty$, and $\kappa_1 \leq \kappa$. We define sets of basis functions, $L_p^{\kappa_1}(\cdot)$, $L_x^{\kappa_1}(\cdot)$, and $\phi_y^{\kappa}(\cdot)$, to be able to approximate $E[D|Z_1, Z_2]$, $E[X|Z_1, Z_2]$, and $E\left(Y|X, Z, D = 1, U\right)$ respectively.

$$\begin{aligned}
L_p^{\kappa_1}(z_1, z_2) &\equiv \left[1, \rho^{\kappa_1}(z_1)', \rho^{\kappa_1}(z_2)'\right]' \\
L_x^{\kappa_1}(z_1, z_2) &\equiv \left[1, \rho^{\kappa_1}(z_1)', \rho^{\kappa_1}(z_2)'\right]' \\
\phi_y^{\kappa}(z_1, x, u, p) &\equiv \left[1, \rho^{\kappa}(z_1)', \rho^{\kappa}(x)', \rho^{\kappa}(u)', \rho^{\kappa}(p)'\right]'
\end{aligned}$$

κ and κ_1 denote knots, i.e., the number of approximating basis functions. For $Z = [Z_1', Z_2']'$ and $z = [z_1', z_2']'$; u, z, x, and p are realizations of U, Z, X, and P, over supports $\mathcal{U}$, $\mathcal{Z}$, $\mathcal{X}$, and $\mathcal{P}$, respectively. In addition to these, we have $W = [Z_1', X', U', P']'$ over support $\mathcal{W}$.

For identification, we make the following assumption for all smooth functions in the models we consider, which can be represented as: $g(0) = g(w)|_{w=0}$.

To reach the estimates, we solve the following minimization problems for the two first step and the second step series regressions:

$$\begin{aligned}
&\underset{\alpha}{min}\left\{n^{-1}\sum_{i=1}^{n}\left[D_i - L_p^{\kappa_1}\left(Z_{1i}, Z_{2i}\right)'\alpha\right]^2\right\} \\
&\underset{\psi}{min}\left\{n^{-1}\sum_{i=1}^{n}\left[X_i - L_x^{\kappa_1}\left(Z_{1i}, Z_{2i}\right)'\psi\right]^2\right\} \\
&\underset{\theta}{min}\left\{n^{-1}\sum_{i=1}^{n}\left[Y_i - \phi_y^{\kappa}\left(Z_{1i}, X_i, \widetilde{U}_i, \widetilde{P}_i\right)'\theta\right]^2\right\}
\end{aligned}$$

to be able to reach the estimates for the smooth functions and the conditional expectations at the first two step regressions, as:

$$\begin{aligned}
L_p^{\kappa_1}(z_1, z_2)'\widetilde{\alpha} &= \widetilde{\mu}_d + \widetilde{m}_1\left(\mathbf{z}_1\right) + \widetilde{m}_2\left(\mathbf{z}_2\right) \equiv \widetilde{m}\left(Z_1, Z_2\right) \\
L_x^{\kappa_1}(z_1, z_2)'\widetilde{\psi} &= \widetilde{\mu}_x + \widetilde{\pi}_1\left(\mathbf{z}_1\right) + \widetilde{\pi}_2\left(\mathbf{z}_2\right) \equiv \widetilde{\pi}\left(z_1, z_2\right) \\
\phi_y^{\kappa}(z_1, x, \widetilde{u}, \widetilde{p})'\widetilde{\theta} &= \widetilde{\mu}_y + \widetilde{g}_1\left(\mathbf{x}\right) + \widetilde{g}_2\left(\mathbf{z}_1\right) + \widetilde{\lambda}_1\left(\widetilde{\mathbf{u}}\right) + \widetilde{\lambda}_2\left(\widetilde{\mathbf{p}}\right) \equiv \widetilde{g}\left(x, z_1, \widetilde{p}, \widetilde{u}\right)
\end{aligned}$$

In the final step, we perform LLLS kernel regression between the partial residual and the related variable. We obtain estimates for both the smooth function and its partial derivative. For Step 3, $g_2(Z_1)$ can be obtained by solving:

$$\underset{a,b}{min}\left\{\sum_{i=1}^{n}\left[\widetilde{Y}_{1i} - a - \left(Z_{1i} - z_1\right)'b\right]^2 K\left(\frac{Z_{1i} - z_1}{h}\right)\right\}$$

where $\widehat{a} = \widehat{g}(z_1)$ and $\widehat{b} = \widehat{\dot{g}}(z_1)$, where $\widehat{\dot{g}}(z_1)$ is the partial derivative of $\widehat{g}(z_1)$ with respect to z_1. The remainder of the smooth functions and their gradients can be estimated accordingly. Our main interest will be on $\widehat{\dot{g}}(z_1)$ and $\widehat{g}(z_1)$ for both the estimation and the theoretical results.

Model 2: Sample Selection Model with Discrete Endogenous Regressors For the second model, we use the transform of the model presented in (2.3). The estimation process will be different from that of Model 1, since we need different assumptions. The estimator for Model 2 is an additive nonparametric version of two-stage least squares, which is adjusted for sample selection via propensity score. The two first-step regressions, as described by 1.a and 1.b, will be identical here. The main difference will be for the second step. Moreover, the final stage regression will be similar to those for Model 1. The estimation steps can be summarized as follows:

2-1.a Estimate the propensity score, P, via $E[D|Z_1, Z_2]$ using a series estimator, and reach $\widetilde{P}$.

2-1.b Estimate the conditional mean of X, via $E[X|Z_1, Z_2]$, again using a series estimator.

2-2 Obtain the conditional expectation, $E[X|Z_1, Z_2]$, from Step 2-1.b, as $\widetilde{\mu}_x + \widetilde{\pi}_1(\mathbf{Z}_1) + \widetilde{\pi}_2(\mathbf{Z}_2) \equiv \widetilde{\pi}(z_1, z_2)$. Using a series estimator run:

$$Y = \mu + \beta\left[\widetilde{\pi}(z_1, z_2)\right] + \widetilde{g}_2(Z_1) + \widetilde{\lambda}(\widetilde{P}) + \epsilon$$

2-3 Regress the partial residual on the variable of interest via LLLS kernel regression. For Z_1, the LLLS kernel regression will be of $\widetilde{Y}_1$ on Z_1; where $\widetilde{Y}_1 = Y - \widetilde{\beta}\left[\widetilde{\pi}(z_1, z_2)\right] - \widetilde{\lambda}(\widetilde{p})$. The remaining smooth functions can be obtained in a similar fashion.

Basis functions for the two first step regressions, and, following this the related minimization problems will be identical to those for the first model. The final step LLLS kernel regression will be the same, as well. The difference will be on the second step, since here the endogenous variable X is discrete and we use a transform of the first model. The set of approximation functions here will be:

$$\phi_{y,2}^{\kappa_2}(z_1, x, E[X|Z_1, Z_2], p) \quad \equiv \quad \left[1, \rho^{\kappa_2}(Z_1)', E[X|Z_1, Z_2], \rho^{\kappa_2}(P)'\right]'$$

where $\kappa_2 = \kappa_2(n)$ is an integer such that $\kappa_2 \to \infty$ as $n \to \infty$, and $\kappa_1 \leq \kappa_2$. The associated minimization problem will be the same except $\phi_y^{\kappa}(\cdot)$ needs to be replaced by $\phi_{y,2}^{\kappa_2}(\cdot)$ as follows:

$$\underset{\theta_2}{min}\left\{n^{-1}\sum_{i=1}^{n}\left[Y_i - \phi_{y,2}^{\kappa_2}\left(Z_{1i}, \widetilde{E}[X_i|Z_1, Z_2], \widetilde{P}_i\right)'\theta_2\right]^2\right\}$$

The solution to this minimization problem gives us $\widetilde{\theta}_2$, which, in turn, are used to reach series estimates for the second step in the second model.

Trimming Function: Given the binary selection variable, D, and for Model 2 with discrete X, a trimming function may be useful to bound the estimates. This helps for theoretical results, as well as potential empirical applications. Following Das et al. (2003), we introduce a trimming function. Here, though, we have additive separability; hence, our trimming function is slightly different. We therefore

use a set of trimming functions on the second step, each specific to the regressor. Our generic trimming function can be presented as:

$$\tau(s) = 1(\tau_i^0 \leq s_i \leq \tau_i^1)$$

where values τ^0 and τ^1 can be either predetermined or estimated, depending on the problem. "s" is the generic variable that should be replaced with the variable of interest. For the estimation steps 1.a and 2-1.a (i.e., regression of selection equation, or, in other words, the propensity score), we assume that τ^0 and τ^1 are between zero and one. This allows us to ensure that the propensity score is restricted between zero and one. For the rest of the regressors, we restrict them to be finite. For the second model, where the endogenous regressor is discrete, the trimming function can be adjusted accordingly. For example, for a binary X, it is again best to take zero and one as the lower and upper bounds for the trimming function.

In the presence of a trimming function, the second step set of basis functions needs to be multiplied with the associated trimming functions, and rest of the estimation will be as described.

For the theoretical results, we will focus on the untrimmed case. We know that by the nature of the trimming function,

$$\left\|\widehat{A}^* - A\right\| \leq \left\|\widehat{A} - A\right\|$$

where A represents any real symmetric matrix. $\widehat{A}$ represents the untrimmed estimator matrix of true A and $\widehat{A}^*$ represents the trimmed version. We know that these distances will be Lipschitz and the relationship will hold for L_2 norms.

Hence, if we can show our results for untrimmed version, conclusions will include trimmed version as a special case.

Selection of smoothing parameters and sieve basis functions: For our basis functions, we use B-splines of order 3 because B-splines present low multicollinearity and have computational advantages (Ozabaci et al. (2013), Powell (1981), Schumaker (2007)).

For the number of approximating terms, in our proofs we assume that they are different for each step. To be specific, we assume that the number of basis functions are the same for all the first step regressions (κ_1), but we assume that they are different for the second steps (κ, κ_2). The number of basis functions can be selected via rule of thumb method $\left(2n^{1/5}\right) + 4$, which we use this method for our simulations. Similarly, the bandwidth parameters, h, can be picked via Silverman's rule of thumb: $h = 1.06\sigma n^{-1/5}$ (this can be changed to $h = 1.06\sigma n^{-1/7}$ to estimate the gradient instead). We use this method to select bandwidth parameters for our simulations. For empirical applications, however, we advocate data driven techniques for both number of basis functions and bandwidths, such as generalized cross validation (GCV) and/or least squares cross validation (LSCV).[2]

4 Asymptotic Properties

In our theoretical proofs, we focus on the second step estimated parameter vectors and the final step estimates for the smooth functions of interest and their gradients. Mainly, we will provide four theorems.

[2]For details of these techniques please refer to Li and Racine (2007)

The first two theorems, (4.1, 4.2), are for large sample properties of θ, and $\widehat{g}(z_1)$ and $\widehat{\dot{g}}(z_1)$ from the first model, in which the endogenous variable X is continuous. The final two theorems, (4.3, 4.4), are for θ_2, and $\widehat{g}(z_1)$ and $\widehat{g}^{(1)}(z_1)$ from the second model, which has discrete endogenous regressors.

Our theoretical work combines aspects of the approaches developed in Das et al. (2003), Das (2005), and Ozabaci et al. (2014). In particular, we merge the assumptions from these works to present our results in a less restrictive fashion.

In our proofs for the first model, we add to the related work by Das et al. (2003), via our estimator's third step, by the help of which we achieve a higher efficiency. Also different from Ozabaci et al. (2014), here we allow for sample selection, which adds an additional step to their estimator. Hence, the theoretical work will need to be changed accordingly.

As for the second model, our proofs differ from Das (2005) in the following two ways. First, we allow for sample selection in addition to endogeneity, and, second, we have a third step in our estimator. Our Model 2 also differs from Ozabaci et al. (2014) because of the discrete nature of the endogenous regressor. This changes the estimator, as well as most of the theoretical proofs. However, we are still able to utilize a number of the lemmas provided in Ozabaci et al. (2014).

In our appendix, we provide detailed proofs and sketch proofs where necessary. Most importantly, our work shows that our final-stage estimates for the smooth functions and their gradients are free from the effects of the first stage, oracle-efficient, and asymptotically normal. Under both models, the asymptotic distribution of the smooth functions and their gradient estimates are equivalent to one dimensional nonparametric regression. We also show that second step estimates have asymptotic biases, which are functions of the first step estimates and the associated approximation bias.

We list the theorems below.

Theorem 4.1 *Under the Assumptions A.1-A.5:*

(i) $\widetilde{\theta}-\theta = Q_{\Phi\Phi}^{-1} n^{-1} \sum_{i=1}^{n} \Phi_i e_i + Q_{\Phi\Phi}^{-1} n^{-1} \sum_{i=1}^{n} \Phi_i \left[\bar{g}\left(X_i, Z_{1i}, U_i, P_i\right) - \Phi_i'\theta\right] - Q_{\Phi\Phi}^{-1} n^{-1} \sum_{i=1}^{n} \Phi_i \, \dot{\pi}_2\left(P_i\right)\left(\widetilde{P}_i - P_i\right) - Q_{\Phi\Phi}^{-1} n^{-1} \sum_{i=1}^{n} \Phi_i \, \dot{\pi}_1\left(U_i\right)\left(\widetilde{U}_i - U_i\right) + \mathbf{R}_{n,\theta}$;
where $\|\mathbf{R}_{n,\beta}\| = \tau_n O_P\left(\nu_n + \nu_{1n}\right)$

(ii) $\left\|\widetilde{\theta} - \theta\right\| = O_P\left(\nu_n + \nu_{1n}\right)$;

for $\nu_{1n} \equiv \kappa_1^{1/2}/n^{1/2} + \kappa_1^{-\gamma}$ *and* $\nu_n \equiv \kappa^{1/2}/n^{1/2} + \kappa^{-\gamma}$.

The result of the first theorem shows that the second step estimates for the model with continuous regressors have error components coming from first two steps. However only one term, ν_{1n} shows up in the rate for the estimated coefficient vector θ. This is partially due to the fact that we assume the number of approximating terms are equivalent for both of the first step regressions here. If we allowed them to be different, the rate would be effected. The result is also slightly different from Ozabaci et al. (2014) since we have an additional term in the bias expression, $Q_{\Phi\Phi}^{-1} n^{-1} \sum_{i=1}^{n} \Phi_i \, \dot{\pi}_2\left(P_i\right)\left(\widetilde{P}_i - P_i\right)$. This result is also in line with the findings of Das et al. (2003). However, note the difference due to their fully nonparametric specification. We continue with the third step results below.

Theorem 4.2 *Under assumptions A1-A5:*

(i) $\sqrt{nh}\left[\widehat{g}_1\left(z_1\right) - g_2\left(z_1\right) - b_{11}\left(z_1\right)\right] \xrightarrow{D} N\left(0, \Omega_{11}\left(z_1\right)\right)$, *where* $b_{11}\left(z_1\right) \equiv \frac{\upsilon_{21}}{2} h^2 \ddot{g}_1\left(z_1\right)$ *and* $\Omega_{11}\left(z_1\right) \equiv \sigma^2(z_1)/f_{Z_1}\left(z_1\right)$.

(ii) $\sqrt{nh^3}\left[\widehat{\dot{g}}_1\left(z_1\right) - \dot{g}_1\left(z_1\right) - b_{12}\left(z_1\right)\right] \xrightarrow{D} N\left(0, \Omega_{12}\left(z_1\right)\right)$, *where* $b_{12}\left(z_1\right) \equiv 0$ *and* $\Omega_{12}\left(z_1\right) \equiv \upsilon_{22}\sigma^2(z_1)/\left[\upsilon_{21}^2 f_{Z_1}\left(z_1\right)\right]$.

(iii) $\sup_{z_1 \in \mathcal{Z}_1} \|[\widehat{g}_1(z_1) - g_1(z_1)]\| = O_P\left((nh/\log n)^{-1/2} + h^2\right)$; *where* $\upsilon_{sj} \equiv \int v^s K(v)^j dv$ *for* s, $j = 0, 1, 2$; *and bias and variance terms are defined in the appendix.*

Theorem 4.2 shows the results obtained for Model 1, which has continuous endogenous regressors in addition to the sample selection problem. What we see here is that the result is equivalent to a standard one dimensional local-linear regression, similar to Ozabaci et al. (2014). Hence, we conclude that the estimator is oracle efficient.

For Model 2, which allows for discrete endogenous regressors, what we mainly show is that the coefficient obtained in the second step for the discrete regressor has $\sqrt{n}$ rate. Consequently, we reach similar results as those obtained for Model 1.

Theorem 4.3 *Under Assumptions B.1-B.5:*
(i) $\left\|\widetilde{\beta}_{\mathbf{0}} - \beta_{\mathbf{0}}\right\| = O_P\left(n^{-1/2}\right)$;
(i) $\left(\widetilde{\beta}_{\mathbf{0}} - \beta_{\mathbf{0}}\right) \xrightarrow{D} N(0, \Omega_\beta)$;
(ii) $\sup_{\mathbf{w_2} \in \mathcal{W}_2} |\widetilde{g}(\mathbf{w_2}) - g(\mathbf{w_2})| = O_P\left[\varsigma_{0\kappa_2}(\nu_{2n} + \nu_{1n})\right]$; *where* ν_{2n}, ν_{1n}, $\varsigma_{0\kappa_2}$ *and* Ω_β *are defined in the appendix.*

The main finding of Theorem 4.3 is that the coefficient of the discrete regressor has the parametric $\sqrt{n}$ rate; hence, it has faster convergence compared to the remainder of the estimates. As a result, the asymptotic properties of the conditional mean obtained at the second step will be identical to those in Ozabaci et al. (2014). Note that, compared to the first model, we avoid the approximation bias due to U, but we have to take into account the bias due to $E[X|Z_1, Z_2]$, i.e., the first step estimates for the conditional mean of discrete X.

After Theorem 4.3, reaching asymptotic properties for the third step is straightforward.

Theorem 4.4 *Under assumptions B1-B5:*
(i) $\sqrt{nh}\left[\widehat{g}_2(z_1) - g_2(z_1) - b_{21}(z_1)\right] \xrightarrow{D} N(0, \Omega_{21}(z_1))$, *where* $b_{21}(z_1) \equiv \frac{\upsilon_{21}}{2} h^2 \ddot{g}_2(z_1)$ *and* $\Omega_{21}(x) \equiv \sigma^2(z_1)/f_{Z_1}(z_1)$.
(ii) $\sqrt{nh^3}\left[\widehat{\dot{g}}_2(z_1) - \dot{g}_2(z_1) - b_{22}(z_1)\right] \xrightarrow{D} N(0, \Omega_{22}(z_1))$, *where* $b_{22}(z_1) \equiv 0$ *and* $\Omega_{22}(z_1) \equiv \upsilon_{22}\sigma^2(z_1)/\left[\upsilon_{21}^2 f_{Z_1}(z_1)\right]$.
(iii) $\sup_{z_1 \in \mathcal{Z}_1} \|[\widehat{g}_2(z_1) - g_2(z_1)]\| = O_P\left((nh/\log n)^{-1/2} + h^2\right)$.

Similar to Theorem 4.2, we reach oracle efficiency here, as well. Theorem 4.4, *(iii)*, shows the uniform convergence rate for the smooth function of Z_1, and the rate for the gradient can also be obtained.

4.1 Partially Linear Extension

For both of our models, additional linear components can be added. In terms of estimation and asymptotic approximations, our prior findings are not affected. This is because the related parameters can be estimated using series approximations. Moreover, they can be estimated at the parametric $\sqrt{n}$ rate, which is faster than our estimates of the smooth functions (and their gradients). The extension of the model can be summarized as follows:

$$\begin{cases} Y^* = \mu_y + g_1(\mathbf{X}) + g_2(\mathbf{Z_1}) + \gamma_1 \mathbf{Z}_4 + \varepsilon, \\ \mathbf{X} = \mu_x + \pi_1(\mathbf{Z}_1) + \pi_2(\mathbf{Z}_2) + \gamma_2 \mathbf{Z}_5 + \mathbf{U}, \\ D = 1\left(\mu_d + m_1(\mathbf{Z}_1) + m_2(\mathbf{Z}_3) + \gamma_3 \mathbf{Z}_6 - V > 0\right), \\ Y = DY^*, \end{cases} \tag{4.1}$$

where γ_1, γ_2 and γ_3 are parameters from the partially linear component, and Z_4, Z_5 and Z_6 are additional exogenous variables.

5 Finite Sample Properties

Using a set of Monte Carlo simulations, we analyze the finite sample properties of our estimators for both of the models we consider here. Over 1000 Monte Carlo simulations, we test our estimators for samples of sizes 100, 200, 400, 800, and 1600. We report the final stage estimates of the conditional mean, and two representative gradients (one with respect to the endogenous variable and the other with respect to one of the included exogenous variables). We report root mean square error (RMSE) to analyze convergence, variance to see the variability in our estimates over different sample sizes, and bias to show that our estimates are unbiased.

We first perform our simulations with three five dimensional DGPs (data generation processes) and present the resulting error, bias and variance measures. We design these DGPs to be able to account for the two models we presented here and another case with heteroskedastic errors. All perform well and show us consistency. Next, to be able to see the performance at a higher dimension, we take the first DGP considered in the prior group and increase the dimension to ten. We report the same results and they all support our asymptotic results. After showing the base results and higher dimensional results, we turn to performance comparisons. We estimate a nonparametric model to be able to compare it to our model and show the gain from the additivity constraint. To estimate the nonparametric model we use the first two steps of our estimator, and this approach is in line with the model by Das et. al. (2003). As expected, our model performs better than the nonparametric alternative for all measures. Finally, we want to see if there is an efficiency gain from the third step, as stated by the asymptotical results. As mentioned earlier, introduction of the third step frees the resulting estimates from approximation errors from the two first step regressions. By the help of this step, we reach oracle efficiency. Here we want to see if we can observe some of this gain in finite sample, as well. Here we can show that we have gain.

We start the discussion of our simulations with the five dimensional DGPs. All of the variables are distributed uniformly between zero and one, and error components are assumed to be normally distributed with zero mean and variance equal to one, unless stated otherwise. Similar to our model 2.1, we keep the variables denoted by the same letters. All Z's are assumed to be exogenous. The DGPs we considered are as follows:

DGP1:

$$\begin{cases} Y^* = sin\left(\mathbf{X}\right) + sin\left(\mathbf{Z_1}\right) + sin\left(\mathbf{Z_4}\right) + sin\left(\mathbf{Z_5}\right) + sin\left(\mathbf{Z_6}\right) + \varepsilon, \\ \mathbf{X} = sin\left(\mathbf{Z}_1\right) + sin\left(\mathbf{Z}_2\right) + sin\left(\mathbf{Z_7}\right) + sin\left(\mathbf{Z_8}\right) + sin\left(\mathbf{Z_9}\right) + \mathbf{U}, \\ D = 1\left(sin\left(\mathbf{Z}_1\right) + sin\left(\mathbf{Z}_3\right) + sin\left(\mathbf{Z_8}\right) + sin\left(\mathbf{Z_9}\right) + sin\left(\mathbf{Z_{10}}\right) - V > 0\right), \\ Y = DY^*. \end{cases} \tag{5.1}$$

DGP2: DGP2 is same as DGP1, except for the error component ε, which is allowed to be a function of Z's. It is assumed to be distributed normally with a variance equal to: $0.5z_1^2 + z_2^2 + z_6^4$.

DGP3: DGP3 is same as DGP1, except X, the endogenous variable, is assumed to be discrete. Here X is defined as: $X = 1\{(sin\left(\mathbf{Z}_1\right) + sin\left(\mathbf{Z}_2\right) + sin\left(\mathbf{Z_7}\right) + sin\left(\mathbf{Z_8}\right) + sin\left(\mathbf{Z_9}\right) - \mathbf{U}) > 0\}$.

As presented in Table (1), what we observe in our first simulation results is that the error and variability factors are smaller for DGP1 compared to DGP2. This is an expected result given the imposed heteroskedasticity in DGP2. Our results show that RMSE is decreasing with sample size for all three DGPs, which shows consistency. One additional result we observe is that the gradients are more precise than the conditional mean estimates. This is due to the additivity, i.e., there are a number of smooth functions in the conditional mean. Whereas, as also presented by our theoretical results, the gradients are estimated at the precision of those from a nonparametric regression with one regressor.

Next we want to see the performance at a higher dimensional setting. Hence, we take DGP1 and add more regressors to the right hand side. The new DGP we consider is as following:

DGP4:

$$\begin{aligned}
Y^* &= sin\left(\mathbf{X}\right) + sin\left(\mathbf{Z_1}\right) + sin\left(\mathbf{Z_4}\right) + sin\left(\mathbf{Z_5}\right) + sin\left(\mathbf{Z_6}\right) + \\
&\quad + sin\left(\mathbf{Z}_{14}\right) + sin\left(\mathbf{Z}_{15}\right) + sin\left(\mathbf{Z_{19}}\right) + sin\left(\mathbf{Z_{20}}\right) + sin\left(\mathbf{Z_{21}}\right) + \varepsilon, \\
\mathbf{X} &= sin\left(\mathbf{Z}_1\right) + sin\left(\mathbf{Z}_2\right) + sin\left(\mathbf{Z_7}\right) + sin\left(\mathbf{Z_8}\right) + sin\left(\mathbf{Z_9}\right) + \\
&\quad + sin\left(\mathbf{Z}_{17}\right) + sin\left(\mathbf{Z}_{18}\right) + sin\left(\mathbf{Z_{19}}\right) + sin\left(\mathbf{Z_{20}}\right) + sin\left(\mathbf{Z_{21}}\right) + \mathbf{U}, \\
D &= 1(sin\left(\mathbf{Z}_1\right) + sin\left(\mathbf{Z}_3\right) + sin\left(\mathbf{Z_8}\right) + sin\left(\mathbf{Z_9}\right) + sin\left(\mathbf{Z_{10}}\right) \\
&\quad + sin\left(\mathbf{Z}_{11}\right) + sin\left(\mathbf{Z}_{12}\right) + sin\left(\mathbf{Z_{14}}\right) + sin\left(\mathbf{Z_{15}}\right) + sin\left(\mathbf{Z_{16}}\right) - V > 0), \\
Y &= DY^*.
\end{aligned}$$

as before the error terms are normally distributed and the regressors are uniformly distributed (Table (2)). When we check the RMSE we can still see that it is decreasing with the sample size. RMSE for all sample sizes are higher compared to the results of five dimensional DGPs. This is expected result.

Next, we estimate a fully nonparametric model to be able to compare it to our results. We estimate it for both dimensions. It can clearly be seen that our model performs better, and the difference grows with the dimension. As it can be seen in Table (3), both models seem to be unbiased. The main driver for the higher RMSE here is the variance. It is clear that the nonparametric model has higher variability and relative to our estimator the difference grows with sample size. This is a finite sample verification of curse of dimensionality. Table (4) shows the same results for the higher dimensional version of the DGP.

Finally, we want to visualize any present efficiency gains from the third step in finite sample. We report the same statistical measures this time for the conditional mean estimated at the second step as well. We only present the results for DGP1, however we tried different specifications and all gave favorable results. We present our results first in Tables (5) and (6). It can clearly be seen that RMSE and variance are always lower for the third step conditional mean. In addition to Tables (5-6), in order to give a clearer picture, we draw the empirical distribution functions for the RMSE from both steps. Figures (1- 2) show the the empirical distribution functions. It is clear that the third step has a uniformly smaller RMSE compared to that of the second step. This result is important to us for two reasons. First, it is signaling us that oracle efficiency is present and our asymptotic results are verified for the finite sample. Second, by these results we show that our model performs better than the two step alternatives in the literature.

We conclude that our estimator works well in the finite sample as well. Its performance is better compared to nonparametric and two-step alternatives. Also, the simulations results support the asymptotic findings.

6 Empirical Application

Here we look at a common problem from labor economics. Our first goal in applying our models is to verify its usability and efficiency. The second goal is to analyze the effect of maternal employment on child cognitive ability development. Our model is new, so we will be the first to apply it. In addition to this, we will be the first to answer the problem using sample selection assumption. We will assume that working mothers and non-working mothers comprise two statistically and systematically different samples, and approach the problem accordingly.

We use a combined dataset from NLSY79 and Bernal and Keane (2011), and analyze the effects of childcare use on cognitive ability of children of working and non-working single mothers. We adopt Bernal and Keane's (2011) dataset because of their extensive instrument set, which includes several welfare rules and policy variables related to childcare incentives and work requirements. Their sample is restricted to single mothers, and this makes the instrument set even stronger. This is because these policy variables are shown to affect single mothers most. Ozabaci et al. (2014) utilize this instrument set and attempt the same question using a similar model. However, they solely focus on childcare use, which they also treat as endogenous. Here, we additionally analyze the difference between working and non-working mothers and their decisions regarding their children's cognitive ability development. The employment status of the mother is endogenous to the system, and we believe that the decision-making process of mothers and the effects of these decisions will change depending on this status. Hence, we believe there is not only an intercept difference but a difference in partial effects, as well.

We set a simple microeconomic model to explain the decision-making process of the mothers, and focus on estimating the child cognitive ability production function. Our results verify previous findings from the literature. We also see some interesting differences between working and non-working single mothers that were not mentioned in earlier studies. We control for three different types of childcare use and see that the signs of partial effects are reversed for these two groups of mothers. For working mothers, the two types of childcare, which are assumed to be market care, have positive signs. Non-market care, on the other hand, is found to have a negative impact on a child's cognitive ability. This is similar to the findings from Bernal and Keane (2011). For non-working mothers, however, the scenario is reversed. Market care variables are no longer positive and significant, but non-market care partial effects are. This is interpreted as tighter constraints for single mothers, and the fact that working mothers may be able to afford higher quality care. When the two groups of mothers are combined, our results are in line with the conclusions of Ozabaci et al. (2014) those can not find significant differences between different types of childcare. The majority of our results show differences between the ability outcomes of the two groups of mothers' children, which we look at separately. We also formally test this via a wild-bootstrapped version of F-test and see that the two groups are systematically different.

To present our results, we first start with the microeconomic model we design, and follow with a short discussion of the dataset that we adopt. After this, we present and discuss our estimation results.

The following model is a simplified version of those developed by Becker (1965), Desai et al. (1989), James-Burdumy (2005), Bernal (2008), and Bernal and Keane (2011). To show the mechanisms clearly, we assume for now that there is only one child, and this assumption is going to be relaxed later on. We also drop time subscripts, again for simplicity.

$$
\begin{aligned}
& U(A, L, G; z) && (6.1)\\
A &= f_A(T_m, G_c, C, \alpha) && (6.2)\\
T_m &= f_m(W, X_m) + z_m && (6.3)\\
C &= f_c(X_m, \alpha, X_c, Cost, E) + Z_c && (6.4)\\
\alpha &= f_\alpha(X_m) + \alpha_0 && (6.5)\\
G_c &= f_{G_c}(\alpha, C, I) + Z_{G_c} && (6.6)
\end{aligned}
$$

where (6.1), $U(\cdot)$, is the utility of the mother, which she maximizes by deciding on her leisure (L), market goods (G) and input quantities related to the child, through which she helps her child's cognitive development . A stands for the child's cognitive ability. Z denotes taste shifters and preference factors. T_m is the maternal time input, C is non-maternal time input (childcare), G_c stands for the goods and services purchased for the child's development, and α stands for initial ability endowment of the child. X_m are mother characteristics, X_c are child characteristics, E are exogenous factors, W shows mother's market labor factors those influence her time allocation and are not in X_m, I is income, and α_0 is the part of initial ability endowment of the child, which can not be explained by mother's characteristics.

C, T_m, and G_c are the main inputs in child's cognitive ability development function. They can be thought as both quantity and quality effects. In fact, here we assume that they are in the multiplicative form, i.e., quantity×quality. This way we can show that the quality effect of a certain input will be augmented if it is used more, and there won't be a quality effect if it is not used. In our empirical example, we will analyze C in three parts, via three different childcare types. These will also help us divide it into the market and the non-market maternal care.

Maternal time, T_m is thought to be influenced by the mother's market work and characteristics (age, education, etc.) those can affect maternal time quality, as well. Similarly, the mother's decision on non-maternal care, C, will be influenced by the same characteristics, along with the ability endowment and characteristics of the child. Here, E stands for exogenous factors that are only related to the non-market care and only influence mother's decision in this regard (example: welfare rules and policy variables from Bernal and Keane (2011)). Mother's decisions on goods and services used for the child, on the other hand, will also be influenced by the mother's characteristics, child's ability endowment, and income. Also, income is not exogenous and is determined in the system.

We test our estimator and econometric model empirically while focusing on equation (6.2). Here, we encounter three econometric problems if we use the setup above for our empirical work. First, the ability endowment of the child, which is not fully observed, will be correlated with the mother's characteristics, as non-maternal care decisions of the mother. This will create an endogeneity problem.

Second, mothers who work and who do not work (after birth) will differ in terms of the decisions they make. Their employment status will also be dependent on the same characteristics as well as the ability endowment. The mother's participation decision will depend on her characteristics (X_m), exogenous factors relevant to her participation decision (such as local demand conditions or policy variables), her child's characteristics (birthweight, gender, etc) and ability endowment, and the mother's related preferences. Important here is that the participation decision is to work or not to work after birth, which is the reason why ability endowment and child characteristics are included.

One can think of participation indicator as an endogenous variable since the unobserved ability will

likely be correlated with mother's characteristics. We agree on the correlation. However, we see the problem as a sample selection, since we also believe that the decisions of the working and non-working mothers will differ, as well as their outcomes. We mainly try to show that the partial effects of the mother's decisions on child's cognitive ability will be different for the two groups.We perform our estimations on two samples of mothers (working, non-working) separately with the correct correction for the selection bias.

The third econometric problem pertains to data. Not all factors listed in equations (6.1-6.6) are fully observed. Especially exogenous factors and input qualities are hard to measure or obtain. So, the previous solutions to this issue have been using proxies and sometimes avoiding the first two econometrics problems.

These are the reasons why the literature presents mixed results. Most examples fail to answer the first two problems simultaneously, and this is mainly due to data and econometric limitations.

James-Burdumy (2005) and Baum (2003) also try to solve the endogeneity issue and use the NLSY79 dataset. Focusing on the effects of maternal employment, they use local market conditions as instruments. Most of their related results do not show a significant effect on maternal employment. Balue and Grossberg (1992), on the other hand, use maternal employment prior to birth as an instrument and find mixed results. Bernal and Keane (2011), try to answer the same applied question, but they focus on the effect of childcare. With the help of their extensive instrument set, they find significant results. They also show that for formal childcare the effect on child's cognitive ability is not negative, while it is for informal childcare use. In on of their specifications, they further include, but fail to find a significant effect for, maternal employment. They try to include maternal employment, in one of the various specifications they try, but they can't find a significant effect for that. They assume that maternal employment is endogenous, as well. There are examples in the literate those use different datasets than NLSY79 and use better controls for childcare quality. These examples mainly show that effects depend on childcare quality, and it can indeed be positive if the quality is high (Peisner-Feinberg et al. (2001), Duncan (2003), Blau (1999)).

Here our results show that the partial effects of non-maternal care change with maternal employment and childcare types. We don't have strong controls for childcare quality, but our results in favor related conclusions from the literature. We see positive and significant effects for formal childcare use by working mothers, who are more likely to afford higher quality care. For non-working mothers, we no longer find positive results. The rest of our findings are presented below.

6.1 Data

For our analysis, we mainly use the dataset by Bernal and Keane (2011). We also take a sample from NLSY79 for the same variables and re-estimate our results, and our findings do not change drastically. The results are stronger when we use the dataset by Bernal and Keane (2011) thanks to their extensive instrument set, so we choose to present those. Here we simply list the variables that we use in our analysis in Table (7), where we also describe the instrument set generally. For details, please refer to Tables 3 and 4 in Bernal and Keane (2011).

6.2 Empirical Specification

We estimate the following empirical model:

$$
\begin{aligned}
LFP &= f_{l,1}(X_m) + f_{l,2}(X_c) + f_{l,3}(E) + \epsilon_l \\
C_s &= f_{c_s,1}(X_m) + f_{c_s,2}(X_c) + f_{c_s,2}(E) + \epsilon_{c_s} \\
I &= f_{I_1}(X_m) + f_{I_2}(X_c) + f_{I_3}(E) + \epsilon_I \\
N &= f_{n_1}(X_m) + f_{n_2}(X_c) + f_{n_3}(E) + \epsilon_n \\
Ltest &= f_1(X_m) + f_2(X_c) + f_3(C_s) + f_4(I) + f_5(N) + f_6(D) + f_7(\widehat{\epsilon}_n) + f_8(\widehat{\epsilon}_l) + f_j(\widehat{\epsilon}_{c_s}) + \lambda(\widehat{LFP}) + \epsilon_{test}
\end{aligned}
$$

where "s" stands for the childcare types: (formal, informal by relatives, informal by non-relatives). N is the number of children, which is also endogenous. Note that we have now relaxed the earlier assumption that there is just one child in the household. The mothers decisions regarding the child will depend on the number of children; both this and the mothers decisions on this number will be co-determined within the system. D stands for the two dummy variables determining the test type. All the dummy variables are modeled linearly.

There are a few additional notes we need to stress before moving on with our results. After the participation equation, the rest are estimated for separated samples with correct probability adjustment. Also, we do not need exclusion restriction for identification of sample selection models. However, in the literature it is advised for parametric models since the correction depends on the normality (or other distributional) assumptions, which may not hold. However, here we do not make any distributional assumptions, which allows us to drop the exclusion restriction.

Our model is a good fit for this application for several reasons. First, it allows for both discrete and continuous endogenous regressors. This design enables us to treat the number of children as a discrete variable while treating all others as continuous. Second, the extensive dataset provided by Bernal and Keane (2011) helps us to solve the problem of endogeneity, though this poses challenges for flexible models since in total we have over 90 regressors at the first stage. However, since our estimator is not computationally demanding and is oracle efficient, we can keep all of these regressors without worrying about the high dimension. Furthermore, since we retain the flexible setting, we can analyze heterogeneity among the mothers in detail.

6.3 Results

We first discuss our results at the median. We primarily focus on childcare use variables, but also discuss AFQT score, education, age and experience of the mother, and household income. Childcare use consists of three levels: formal childcare, informal childcare by relatives, and informal childcare by non-relatives. All of these are presented in Table (8). For working mothers, partial effects of experience, AFQT score, and education are all positive and significant at the median. Such results are expected since these characteristics have a positive effect on maternal time quality. Age, on the other hand, is found to be negative and significant. This is also expected since increasing age may have a negative quality effect through stress and amplified fatigue. Income effect is found to be insignificant. This is in line with the well-known argument in the literature that, what matters most here is lifetime income. Hence, the cumulative income for a limited period, may have limited influence. Formal and informal childcare use provided by non-relatives are found to be positive and significant at the median, though this is also not very high in magnitude. Both of these variables may be considered as market care. Informal childcare provided by relatives is found to be negative and significant at the median. It is again not very high in

magnitude. This type of childcare may be considered as the non-market care. Here, our argument is that working mothers may be able to afford high-quality care in the market, hence the positive results.

Next, we turn to non-working mothers. The results change for this group except for AFQT score and age. The rest of the mother characteristics along with income are found to be insignificant at the median. Non-working mothers, having more flexible hours, may be compensating for any differentials in this group by devoting more maternal care to the child (in line with our quality and quantity argument earlier, and their multiplicative form). On the other hand, childcare variables have different magnitudes and opposite signs. Market care variables are found to have negative partial effects, while we see positive partial effects for the non-market care. This supports our argument earlier about working mothers being able to afford better quality care. Informal childcare provided by relatives, by allowing mothers to increase their leisure time, may have a positive impact on maternal time quality through decreased stress, without affecting the mothers budgets.

We now turn to the distributions of our gradients estimates for certain variables where we see heterogeneity. The rest of the estimations, along with the first stage and participation results are also available upon request. We first present figures for the overall distributions of the estimated gradients. From mother characteristics, we only show AFQT score and age. For both mother groups, we see that after a certain score the results' variability increases, and especially for non-working mothers, we see more positive results. We conclude that what matters more is if you get a high AFQT score. When we analyze estimates of age partial effects for working mothers group, we see an upward trend in the partial effect as age increases from fifteen to twenty. In this period, mothers move from being a teenager to young adults, so this expected. Next, we also see that we observe a second peak around age thirty two. We conclude that, up to that age maternal time quality peaks, and starts falling after age thirty two due to increasing fatigue and stress. Hence, we see the two peaked partial effects of age on child cognitive ability. However, for non-working mothers, the conclusions change. The initial peak starts at a later age for this group, and we no longer observe the two-peaked distribution of estimated gradients.

When it comes to childcare variables, our findings change slightly when we see the all the estimates. For all childcare types, for all mothers, we see both positive and negative significant estimates. The signs at the median are determined by the sign of the majority of the estimated gradients. So, for example, we can say that, for non-working mothers, a majority of the children are positively affected by informal childcare provided by relatives. However, a number of them are negatively impacted, as well. The rest will be analyzed similarly. The only difference here is for the partial effect of informal childcare provided by relatives on the children with working mothers, which is always negative and significant. It is also worth noting that the results are not as strong as those presented by the medians when we check these distributions for formal childcare and informal childcare provided by non-relatives.

Next we plot the empirical distribution functions of the estimated childcare partial effects by type for each mother group. As seen in Figures (5) and (6), except for working mothers and for only formal and informal childcare provided by relatives, we do not find any first order dominance. Hence, we can not say one estimate is uniformly greater than the other. The noted exception is that, for working mothers, formal childcare had a uniformly greater partial effect compared to informal childcare provided by relatives. Since we do not find the same result for non-working mothers, here we conclude that, provided that the quality is high, formal childcare may be beneficial for the child.

Next we perform diagnostic checks on for our results. We first investigate whether working and non-working mothers should indeed be treated differently. We perform a bootstrapped version of the F.test

here, which is similar to the Chow test; only critical values are obtained via wild-bootstrap. We reject the hypothesis that the two groups are the same, hence finding evidence in favor of our earlier assumption (p value=0.001). In addition, we test whether the smooth functions of the residuals are jointly significant or not, and again find joint significance (p=0.00097).

7 Conclusion

In this paper, we considered additive nonparametric structural equations models under which we correct for both sample selection bias and endogeneity. We proposed two different models, one with a continuous endogenous variable, and another where this variable is assumed to be discrete over finite support. We provided multi-stage estimators for both of the models, using a combination of series estimators and kernel regression. In our models, the estimates of the smooth functions in the outcome equation and their gradients were found to be oracle efficient and asymptotically normal. Using a set of Monte-Carlo simulations we further showed that our estimators perform well over finite samples.

We also applied our estimators to an empirical problem from labor economics literature, analyzing the effects of maternal employment and childcare use on the cognitive ability development of children. This example was a good fit for our econometric methodology. With our estimators, we maintained a flexible design without worrying about the high dimension or computing demand. We approached this question by assuming that childcare use is endogenous while treating maternal employment as a sample selection problem. We found that the difference between working and non-working mothers is significant. Also, we saw that formal childcare had a positive impact on children of working mothers, who are more likely to be able to afford higher quality childcare.

Our models and proposed estimators thus solve the two important empirical problems of sample selection bias and endogeneity. Moreover, they do so at a very high efficiency level (oracle efficiency) without worrying about the numbers of regressors (curse of dimensionality). They also capture the overall potential heterogeneity without losing degrees of freedom. In light of the nature of the empirical problems we target, the desirable properties we presented, and the easy implementability of our estimators (low computing demand), we believe that our models are good fits for empirical research.

References

Ai, C., and Chen, X. (2003), "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica,* 71, 1795-1843.

Becker, Gary. (1965), "A theory of the allocation of time,," *Economic Journal,* 75, 299, 493517.

Bernstein, D. S. (2005), *Matrix Mathematics: Theory, Facts and Formulas with Application to Linear Systems Theory,* Princeton: Princeton University Press.

Bernal, R. (2008), "The Effect of Maternal Employment and Child Care on Children's Cognitive Development," *International Economic Review,* 49, 1173-209.

Bernal, R., and Keane, M.P. (2011), "Child Care Choices and Children's Cognitive Achievement: the Case of Single Mothers," *Journal of Labor Economics,* 29, 459-12.

Blau, D. (1999), "The Effects of Child Care Characteristics on Child Development," *The Journal of Human Resources*, XXXIV, 4.

Blau, F. D., and Grossberg, A. J. (1992), "Maternal Labor Supply and Children's Cognitive Development," *Review of Economics and Statistics,* 74, 474-1.

Chen, X. (2007), "Large Sample Sieve Estimation of Semi-Nonparametric Models," In J. J. Heckman and E. Leamer (eds.), *Handbook of Econometrics*, 6B (Chapter 76), 5549-5632. New York: Elsevier Science.

Chen, X., and Pouzo, D. (2012), "Estimation of Nonparametric Conditional Moment Models with Possibly Nonsmooth Generalized Residuals," *Econometrica,* 80, 277-321.

Darolles, S., Fan, Y., Florens, J. P., and Renault, E. (2011), "Nonparametric Instrumental Regression," *Econometrica,* 79, 1541-1565.

Desai, S., L. Chase-Lansdale and M. Robert. (1989), "Mother or Market? Effects of Maternal Employment on the Intellectual Ability of 4-year Old Children," *Demography*, 26, 4.

Duncan, G. and NICHD Early Child Care Research Network. (2003), "Modeling the Impacts of Child Care Quality on Childrens Preschool Cognitive Development," *manuscript Northwestern University.*

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications,* Chapman & Hall, London.

Hahn, J., and Ridder, G. (2013), "Asymptotic Variance of Semiparametric Estimators with Generated Regressors," *Econometrica,* 81, 315-340.

Hall, P., and Horowitz, J. L. (2005), "Nonparametric Methods for Inference in the Presence of Instrumental Variables," *Annals of Statistics,* 33, 2904-2929.

Hansen, B. E. (2008), "Uniform Convergence Rates for Kernel Estimation with Dependent Data," *Econometric Theory,* 24, 726-748.

Horowitz, J. L. (2013), "Nonparametric Additive Models," in A. Ullah, J. Racine and L. Su (eds.), *The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*, forthcoming, Oxford University Press, Oxford.

Horowitz, J. L., and Mammen, E. (2004), "Nonparametric Estimation of an Additive Model with a Link Function," *Annals of Statistics,* 32, 2412-2443.

James-Burdumy, S. (2005), "The Effect of Maternal Labor Force Participation on Child Development," *Journal of Labor Economics,* 23, 177-11.

Li, Q., (2000), "Efficient Estimation of Additive Partially Linear Models," *International Economic Review,* 41, 1073-1092.

Li, Q., and Racine, J. (2007), *Nonparametric Econometrics: Theory and Practice,* Princeton: Princeton University Press.

Mammen, E., Rothe, C., and Schienle, M. (2012), "Nonparametric Regression with Nonparametrically Generated Covariates," *Annals of Statistics,* 40, 1132-1170.

Martins-Filho, C., and Yao, F. (2012), "Kernel-Based Estimation of Semiparametric Regression in Triangular Systems," *Economics Letters,* 115, 24-27.

Masry, E. (1996), "Multivariate Local Polynomial Regression for Time Series: Uniform Strong Consistency Rates," *Journal of Time Series Analysis,* 17, 571-599.

Newey, W. K. (1997), "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics,* 79, 147-168.

Newey, W. K., and Powell, J. L., (2003), "Instrumental Variable Estimation of Nonparametric Models," *Econometrica,* 71, 1565-1578.

Newey, W. K., Powell, J. L., and Vella, F. (1999), "Nonparametric Estimation of Triangular Simultaneous Equation Models," *Econometrica,* 67, 565-603.

Pagan, A., and Ullah, A. (1999), *Nonparametric Econometrics,* Cambridge: Cambridge University Press.

Peisner-Feinberg, E., M. Burchinal, R. Clifford, M. Culkin, C. Howes, S. Kagan and N. Yazejian. (2001), "The Relation of Preschool Child-Care Quality to Childrens Cognitive and Social Developmental Trajectories through Second Grade," *Child Development,* 72(5), 1534-1553.

Pinkse, J. (2000), "Nonparametric Two-Step Regression Estimation When Regressors and Error Are Dependent," *Canadian Journal of Statistics,* 28, 289-300.

Powell, M. J. D. (1981), *Approximation Theory and Methods,* Cambridge: Cambridge University Press.

Roehrig, C.S. (1988), "Conditions for Identification in Nonparametric and Parametric models," *Econometrica,* 56, 433-47.

Schumaker, L. L. (2007), *Spline Functions: Basic Theory*, 3rd ed, Cambridge: Cambridge University Press.

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics.* John Wiley & Sons, New York.

Su, L., and Jin, S. (2012), "Sieve Estimation of Panel Data Models with Cross Section Dependence," *Journal of Econometrics,* 169, 34-47.

Su, L., Murtazashvili, I., and Ullah, A. (2013), "Local Linear GMM Estimation of Functional Coefficient IV Models with Application to the Estimation of Rate of Return to Schooling," *Journal of Business & Economic Statistics,* 31, 184-207.

Su, L., and Ullah, A. (2008), "Local Polynomial Estimation of Nonparametric Simultaneous Equations Models," *Journal of Econometrics,* 144, 193-218.

Ullah, A. (1985), "Specification Analysis of Econometric Models," *Journal of Quantitative Economics* 1, 187-209.

Vella, F. (1991), "A Simple Two-Step Estimator for Approximating Unknown Functional Forms in Models with Endogenous Explanatory Variables," *Working Paper*, Department of Economics, Australian National University.

Table 1: Monte Carlo Simulations for Conditional Mean and Gradients

	$\widehat{g}(\cdot,\cdot)$	$\frac{\partial \widehat{g}(\cdot,\cdot)}{\partial x}$	$\frac{\partial \widehat{g}(\cdot,\cdot)}{\partial z_1}$	$\widehat{g}(\cdot,\cdot)$	$\frac{\partial \widehat{g}(\cdot,\cdot)}{\partial x}$	$\frac{\partial \widehat{g}(\cdot,\cdot)}{\partial z_1}$	$\widehat{g}(\cdot,\cdot)$	$\frac{\partial \widehat{g}(\cdot,\cdot)}{\partial x}$	$\frac{\partial \widehat{g}(\cdot,\cdot)}{\partial z_1}$	$\widehat{g}(\cdot,\cdot)$	$\frac{\partial \widehat{g}(\cdot,\cdot)}{\partial x}$	$\frac{\partial \widehat{g}(\cdot,\cdot)}{\partial z_1}$
		200			400			800			1600	
RMSE												
DGP1	0.3955	0.3050	0.7647	0.3006	0.2450	0.5455	0.2356	0.1953	0.4359	0.1886	0.1606	0.3439
DGP2	0.4502	0.3421	0.8769	0.3390	0.2641	0.6429	0.2642	0.2138	0.5037	0.2092	0.1749	0.3887
DGP3	0.3741	0.7765	0.8033	0.2876	0.6401	0.6312	0.2256	0.5202	0.4915	0.1842	0.4493	0.3923
Variance												
DGP1	0.1457	0.0852	0.5261	0.0823	0.0564	0.2755	0.0496	0.0357	0.1812	0.0307	0.0247	0.1126
DGP2	0.1898	0.1100	0.6974	0.1064	0.0651	0.3820	0.0633	0.0426	0.2376	0.0386	0.0289	0.1462
DGP3	0.1333	0.4805	0.5943	0.0777	0.2616	0.3714	0.0468	0.1242	0.2333	0.0308	0.0512	0.1511
Bias												
DGP1	0.0943	0.0326	-0.0055	0.0849	0.0271	0.0189	0.0745	0.0160	0.0072	0.0686	0.0166	0.0067
DGP2	0.0949	0.0282	-0.0186	0.0838	0.0243	0.0007	0.0763	0.0177	0.0068	0.0698	0.0157	0.0083
DGP3	0.0524	0.3899	-0.0161	0.0582	0.3449	-0.0091	0.0573	-0.3426	0.0094	0.0533	0.3381	0.0046

These are the results for DGPs 1-3, where each DGP is five dimensional. Reported results are the mean values of the related measures.

Table 2: Monte Carlo Simulations for Conditional Mean and Gradients: Higher Dimensions

	$\widehat{g}(\cdot,\cdot)$	$\frac{\partial \widehat{g}(\cdot,\cdot)}{\partial x}$	$\frac{\partial \widehat{g}(\cdot,\cdot)}{\partial z_1}$	$\widehat{g}(\cdot,\cdot)$	$\frac{\partial \widehat{g}(\cdot,\cdot)}{\partial x}$	$\frac{\partial \widehat{g}(\cdot,\cdot)}{\partial z_1}$	$\widehat{g}(\cdot,\cdot)$	$\frac{\partial \widehat{g}(\cdot,\cdot)}{\partial x}$	$\frac{\partial \widehat{g}(\cdot,\cdot)}{\partial z_1}$	$\widehat{g}(\cdot,\cdot)$	$\frac{\partial \widehat{g}(\cdot,\cdot)}{\partial x}$	$\frac{\partial \widehat{g}(\cdot,\cdot)}{\partial z_1}$
		200			400			800			1600	
RMSE	0.5130	0.3082	0.9286	0.3706	0.23790	0.6496	0.2767	0.1913	0.4814	0.2092	0.1555	0.3747
Variance	0.2568	0.4578	0.0775	0.1344	0.2527	0.0385	0.0752	0.1380	0.0171	0.0431	0.0759	0.0093
Bias	-0.0314	0.0058	-0.0237	-0.0224	0.0090	0.0198	-0.01736	0.0065	0.0105	0.0119	0.0040	-0.0069

These are the results for DGP 4, and it is 10 dimensional. Reported results are the mean values of the related measures.

Table 3: Monte Carlo Simulations for the Conditional Mean: Nonparametric (NP) vs Additive Nonparametric Specifications

	200	400	800	1600
RMSE				
Additive	0.3955	0.3006	0.2356	0.1886
NP	0.6009	0.5079	0.4670	0.4062
Variance				
Additive	0.1457	0.0823	0.0496	0.0307
NP	0.3583	0.2562	0.2207	0.1645
Bias				
Additive	-0.0943	-0.0849	0.0745	-0.0686
NP	- 0.0055	- 0.0063	0.0015	–0.0073

These are the results for DGP 1 and nonparametric alternative of it. Both DGPs are 5 dimensional. Reported results are the mean values of the related measures.

Table 4: Monte Carlo Simulations for the Conditional Mean: Nonparametric (NP) vs Additive Nonparametric Specifications

	200	400	800	1600
RMSE				
Additive	0.5130	0.3706	0.2767	0.2092
NP	0.7555	0.6400	0.5670	0.5234
Variance				
Additive	0.2568	0.1344	0.0752	0.0431
NP	0.5681	0.4081	0.3207	0.2735
Bias				
Additive	-0.0314	0.0224	-0.0174	0.0119
NP	- 0.0017	- 0.0029	0.0015	-0.0002

These are the results for DGP 1 and nonparametric alternative of it. Both DGPs are 10 dimensional. Reported results are the mean values of the related measures.

Table 5: Monte Carlo Simulations: 2nd and 3rd Step Conditional Mean Estimates of DGP-1

	200	400	800	1600
RMSE				
3rd	0.3955	0.3006	0.2356	0.1886
2nd	0.5639	0.4219	0.3670	0.2393
Variance				
3rd	0.1457	0.0823	0.0496	0.0307
2nd	0.3132	0.1745	0.0207	0.0547
Bias				
3rd	-0.0943	-0.0849	0.0745	-0.0686
2nd	-0.0469	- 0.0492	0.0135	- 0.0779

Table 6: Monte Carlo Simulations: 1st and 2nd Step of the Conditional Mean Estimates of Y from DGP-4

	200	400	800	1600
RMSE				
3rd	0.5130	0.3706	0.2767	0.2092
2nd	0.6789	0.5039	0.3723	0.2761
Variance				
3rd	0.2568	0.1344	0.0752	0.0431
2nd	0.4578	0.2527	0.1380	0.0759
Bias				
3rd	-0.0314	0.0224	-0.0174	0.0119
2nd	-0.0011	0.0011	0.0003	0.0006

These are the results for DGP 1 and nonparametric alternative of it. Both DGPs are 10 dimensional. Reported results are the mean values of the related measures.

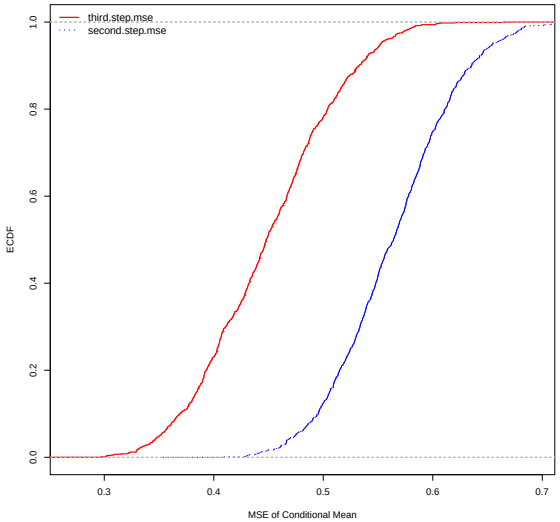
Table 7: Variables Used in the Empirical Application

Instruments (E):	
Policy Variables	Work requirements, Earning Disregards, Benefits and Expenditures of Childcare programs, Time Limits
Local Demand Conditions	State Unemployment rate, Hourly Wage rate
Controls (X_m, X_c):	
work:	= 1 is mother was employed prior to birth
age:	Age of the mother
educ:	Educational attainment of the mother
afqt:	AFQT score of the mother
exp:	Experience of the mother prior to birth
marital:	Marital status of the mother at birth
urban:	= 1 is resides in Urban area
race:	= 1 if Hispanic or Black
gender:	= 1 if male
weight:	Birthweight of the child
agec:	Age of child at test date
dPPVT:	= 1 if the corresponding test is Peabody Picture Vocabulary Test
dMATH:	= 1 if the corresponding test is Peabody Individual Achievement Test, Math Subtest
Endogenous Variables (C,I,N):	
income:	Household income
child:	Number of children under age 5
formal childcare	
informal childcare by relatives	
informal childcare by non-relatives	
Maternal Employment Indicator:	= 1 if mother always worked after birth, = 0 if mother never worked after birth.

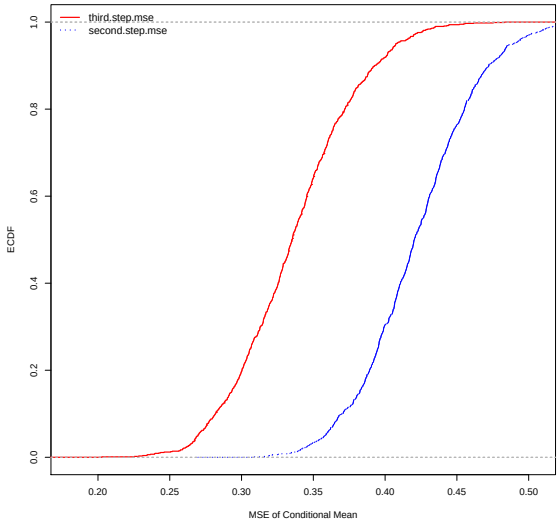
Table 8: Estimated Partial Effects on Test Scores of Children: Medians with Bootstrapped Standard Errors in Paranthesis

	Working Mothers	Non-Working Mothers
income	-0.0011 (0.0035)	-0.0134 (0.0122)
experience	0.0247 (0.0056)	-0.0178 (0.0438)
AFQT score	0.0069 (0.0034)	0.0008 (0.0034)
education	0.0621 (0.0146)	0.0707 (0.0758)
age	-0.0178 (0.0066)	-0.0122 (0.0206)
formal childcare	0.1825 (0.0054)	-0.4457 (0.0192)
informal childcare by relatives	-0.0712 (0.0054)	0.1423 (0.0192)
informal childcare by non-relatives	0.0665 (0.0043)	-0.0695 (0.0173)

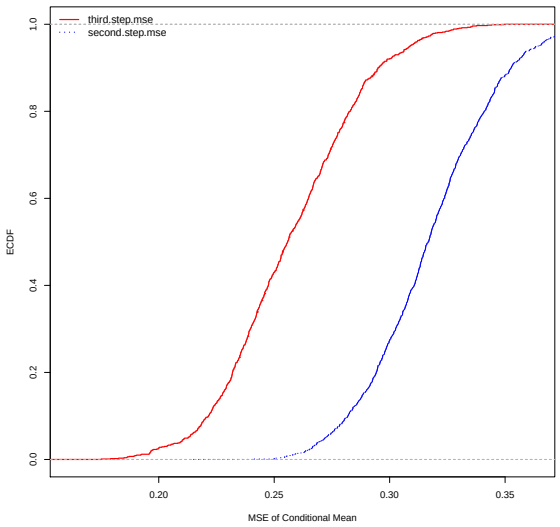
Figure 1: Empirical Distribution Functions from the MSE of 2nd and 3rd Step Conditional Mean Estimates: DGP-1



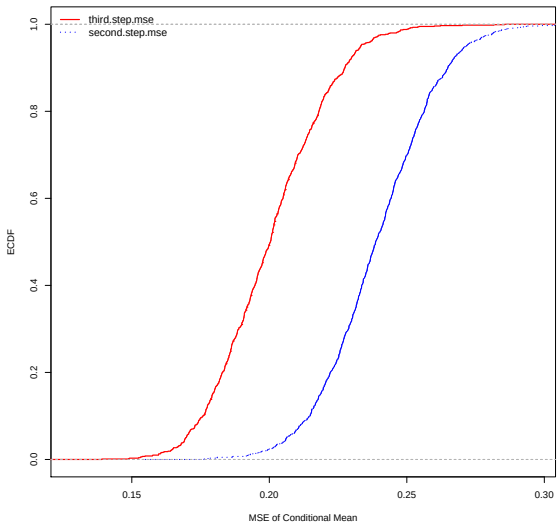
(a) n=200



(b) n=400

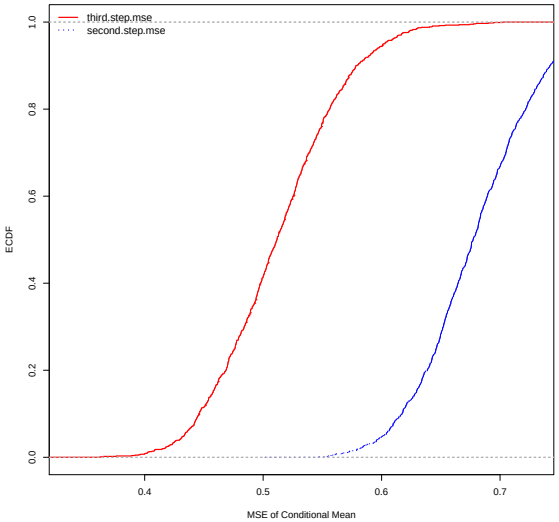


(c) n=800

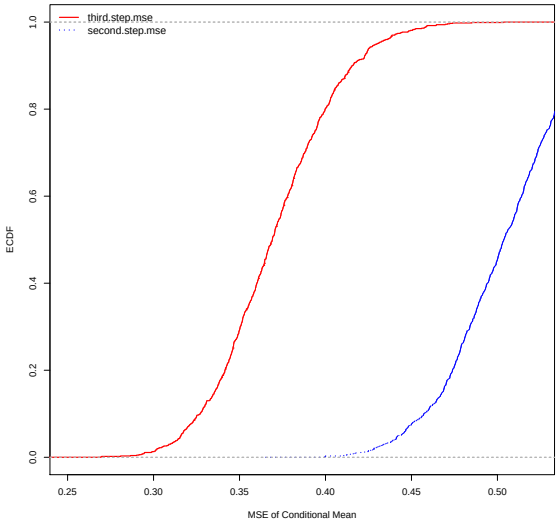


(d) n=1600

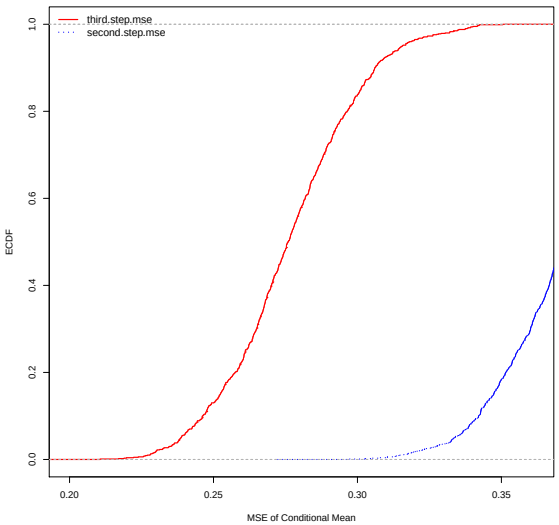
Figure 2: Empirical Distribution Functions from the MSE of 2nd and 3rd Step Conditional Mean Estimates: DGP-4



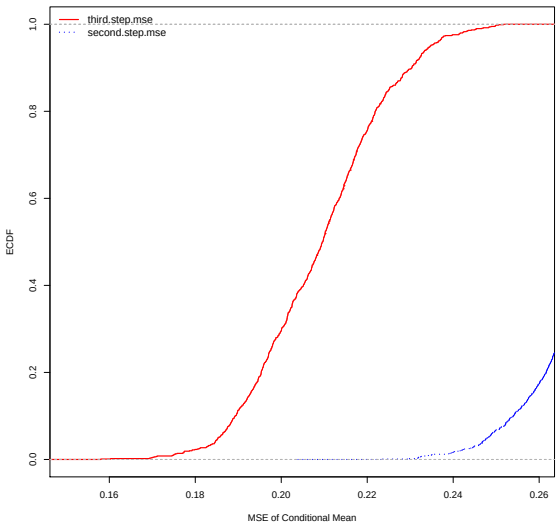
(a) n=200



(b) n=400

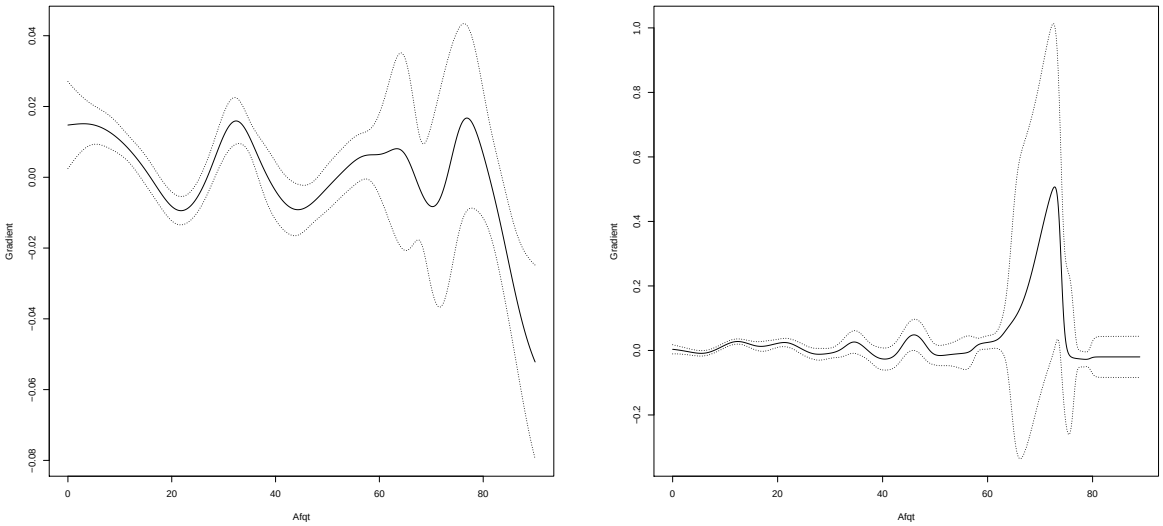


(c) n=800



(d) n=1600

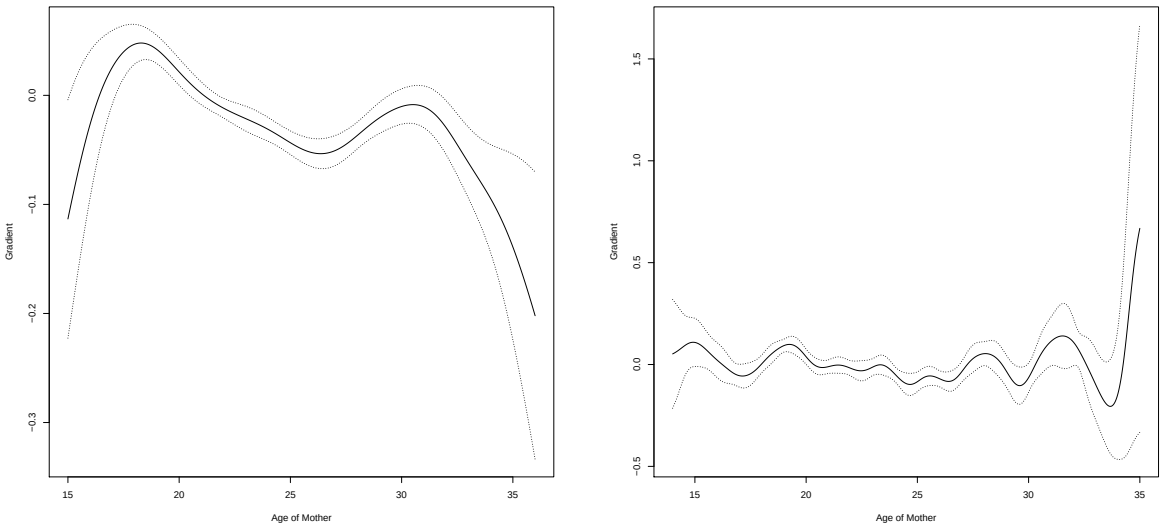
Figure 3: Gradients for AFQT score from Children with Working and Non-Working Mothers (with confidence bounds)



(a) Working Mothers

(b) Non-Working Mothers

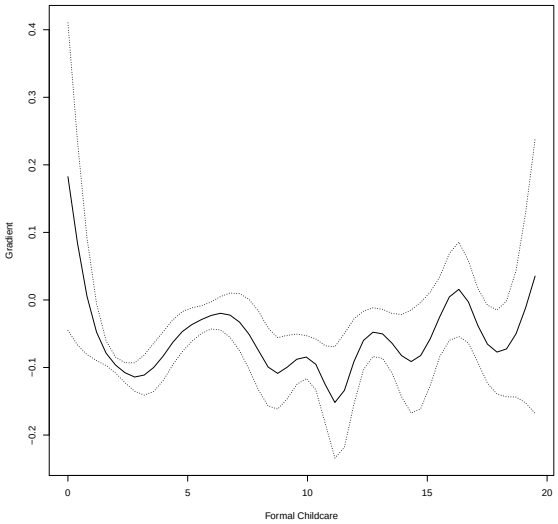
Figure 4: Gradients for Age of the Mother from Children with Working and Non-Working Mothers (with confidence bounds)



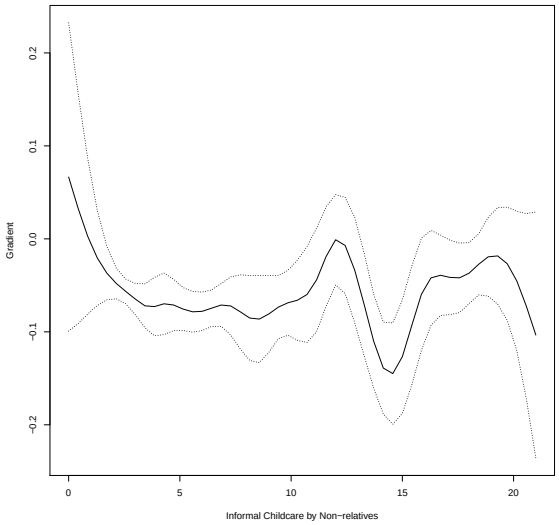
(a) Working Mothers

(b) Non-Working Mothers

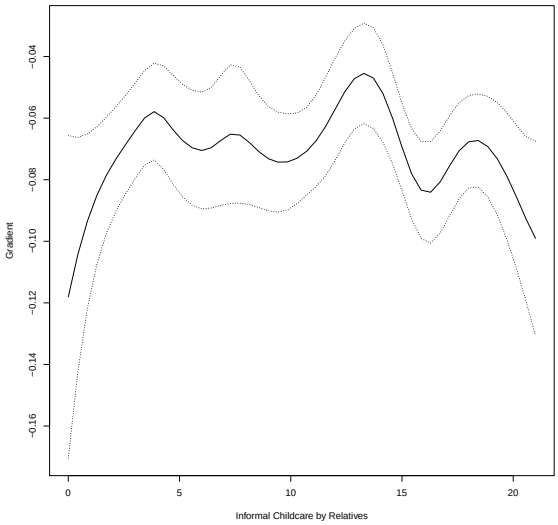
Figure 5: Gradients for Children with Working Mothers - Types of Childcare Use (with confidence bounds)



(a) Formal Childcare

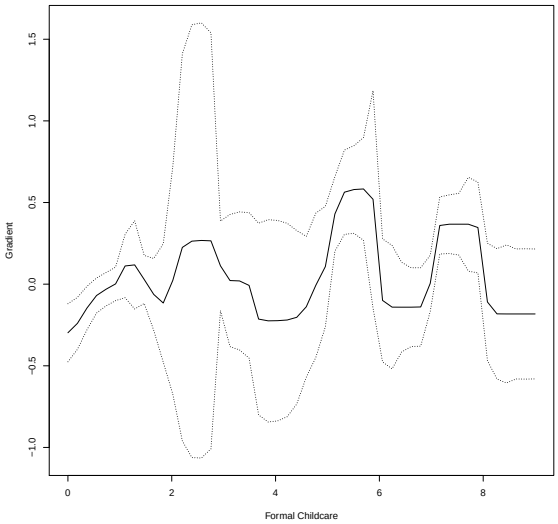


(b) Informal Childcare Provided by Non-Relatives

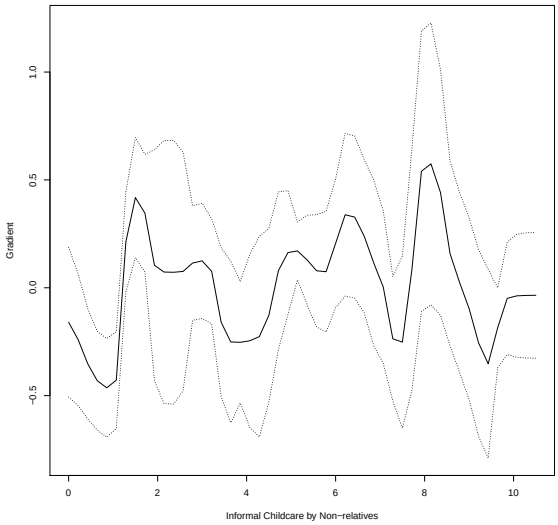


(c) Informal Childcare Provided by Relatives

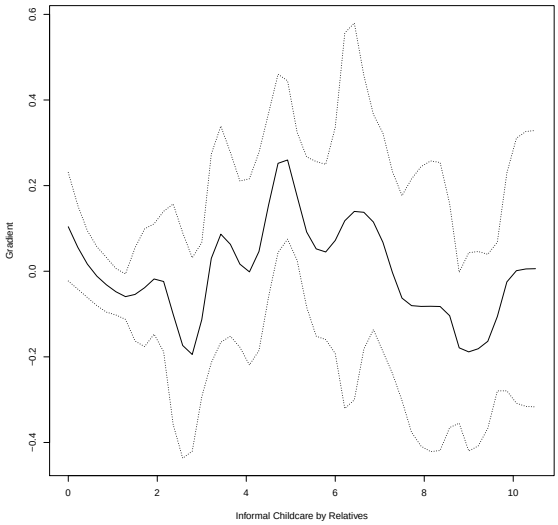
Figure 6: Gradients for Children with Non-Working Mothers - Types of Childcare Use (with confidence bounds)



(a) Formal Childcare

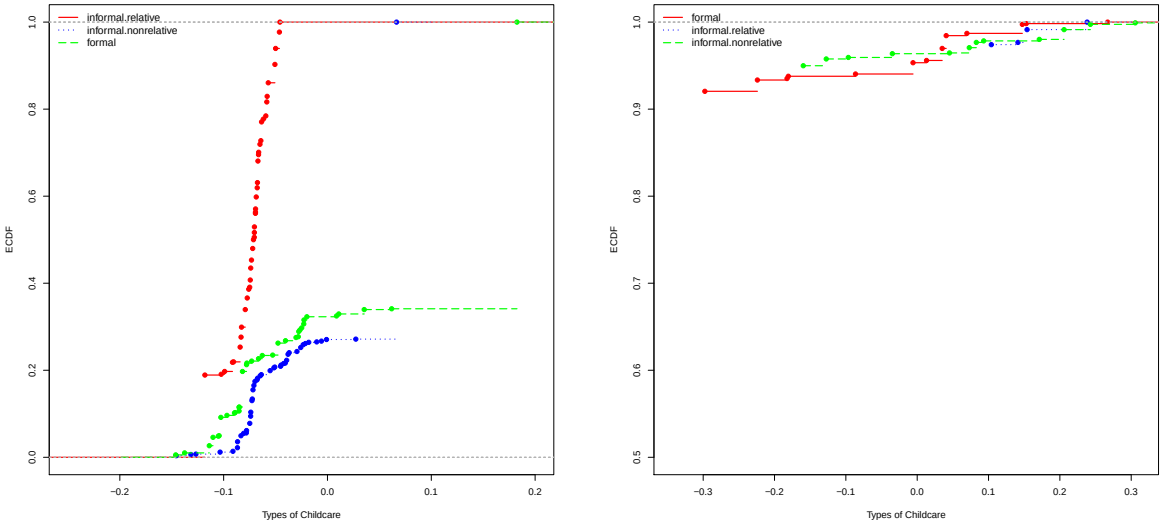


(b) Informal Childcare Provided by Non-Relatives



(c) Informal Childcare Provided by Relatives

Figure 7: Empirical Distribution Functions of Gradients - Types of Childcare Use



(a) Working Mothers

(b) Non-Working Mothers

8 Appendix

8.1 List of Terms used in the paper and the appendix:

$Z_{1i}^{*}(z_1) = \left[1, (Z_{1i} - z_1)'\right]$

$f_{X_1}(\cdot)$ denotes the probability density function (PDF) of X_{1i},

$|q(\cdot)|_r \equiv \max_{s \leq r} \sup_{v \in \mathcal{S}} |\partial^s q(v)|$. $\mathcal{S}$ is support, of function $q(\cdot)$ with r number of derivatives.

$\sigma^2(x) \equiv E(e_i^2 | X_i = x)$,

$\sigma^2(z_1) \equiv E(e_i^2 | Z_{1i} = z_1)$,

$\upsilon_{sj} \equiv \int v^s K(v)^j dv$ for s, $j = 0, 1, 2$.

$\mathbf{W}_i \equiv (\mathbf{X}_i', \mathbf{Z}_{1i}, \mathbf{U}_i', \mathbf{P}_i')'$,

$\mathbf{W}_{2i} \equiv (\mathbf{X}_i', \mathbf{Z}_{1i}, \bar{\pi}_i', \mathbf{P}_i')'$,

$\mathbf{Z}_i \equiv (\mathbf{Z}_{1i}, \mathbf{Z}_{2i})'$,

$e_i \equiv Y_i - \bar{g}(\mathbf{X}_i, \mathbf{Z}_{1i}, \mathbf{U}_i, \mathbf{P}_i)$

$\upsilon_i \equiv D_i - \bar{m}(\mathbf{Z}_{1i}, \mathbf{Z}_{2i})$

$\sigma_{1,i}^2 \equiv \sigma_1^2(\mathbf{Z}_i) \equiv E\left(\upsilon_i^2 | \mathbf{Z}_i\right)$

$L_{p,i} \equiv P^{\kappa_1}(\mathbf{Z}_i)$,

$\Phi_i \equiv \Phi^{\kappa}(\mathbf{W}_i)$,

$\widetilde{\Phi}_i \equiv \tau_i \circ \Phi^{\kappa}(\widetilde{\mathbf{W}}_i)$, where $\circ$ is the Hadamart product, and τ_i is the trimming function.

$Q_{n,PP} \equiv n^{-1} \sum_{i=1}^{n} P_i P_i'$,

$Q_{n,\Phi\Phi} \equiv n^{-1} \sum_{i=1}^{n} \Phi_i \Phi_i'$,

$\widetilde{Q}_{n,\Phi\Phi} \equiv n^{-1} \sum_{i=1}^{n} \widetilde{\Phi}_i \widetilde{\Phi}_i'$.

$Q_{\Phi\Phi} \equiv E[\Phi_i \Phi_i']$.

$Q_{LL_p} \equiv E[P^{\kappa_1}(\mathbf{Z}_1, \mathbf{Z}_2) P^{\kappa_1}(\mathbf{Z}_1, \mathbf{Z}_2)']$

$Q_{LL_x} \equiv E[P^{\kappa_1}(\mathbf{Z}_1, \mathbf{Z}_2) P^{\kappa_1}(\mathbf{Z}_1, \mathbf{Z}_2)']$

$Q_{LL_x,U} = E[P^{\kappa_1}(\mathbf{Z}_1, \mathbf{Z}_2) P^{\kappa_1}(\mathbf{Z}_1, \mathbf{Z}_2)' U^2]$.

$Q_{LL_x,\sigma} = E[P^{\kappa_1}(\mathbf{Z}_1, \mathbf{Z}_2) P^{\kappa_1}(\mathbf{Z}_1, \mathbf{Z}_2)' \sigma_1^2]$.

Let $\mathbf{Z}_i \equiv (\mathbf{Z}_{1i}', \mathbf{Z}_{2i}')'$.

$\|A\|$ ($\equiv [\mathrm{tr}(AA')]^{1/2}$), is the Frobenius norm of a real matrix A.

$\|A\|_{\mathrm{sp}}$ ($\equiv \sqrt{\lambda_{\max}(A'A)}$) is the its spectral norm of A.

$\equiv$ means "is defined as"

$\lambda_{\max}(\cdot)$ is the largest eigenvalue of a real symmetric matrix, and $\lambda_{\min}(\cdot)$ is the smallest eigenvalue of a real symmetric matrix.

$\dot{f}(\cdot)$ and $\ddot{f}(\cdot)$ are first and second derivatives of function $f(\cdot)$, respectively.

$\xrightarrow{D}$ denotes convergence in distribution.

$\xrightarrow{P}$ to denotes convergence in probability.

$\varsigma_{r\kappa} = O\left(\kappa^{r+1/2}\right)$

8.2 Assumptions A

Assumption A1.

(i) The triple $\{Y_i, \mathbf{X}_i, \mathbf{Z}_i\}$ are IID and random, for $i = 1, ..., n$.

(ii) $\mathbf{W}_i$ and $\mathbf{Z}_i$ have compact support.

(iii) $\mathbf{W}_i$ and $\mathbf{Z}_i$ are continuously distributed with respect to the Lebesgue measure.

Assumption A2.

(i) For every κ_1 that is sufficiently large, there exist $\underline{c}_{11}$ and $\bar{c}_{11}$ such that $0 < \underline{c}_{11} \leq \lambda_{\min}\left(Q_{LL_p}\right) \leq \lambda_{\max}\left(Q_{LL_p}\right) \leq \bar{c}_{11} < \infty$ and $\lambda_{\max}\left(Q_{LL_p,\sigma}\right) \leq \bar{c}_{11} < \infty$.

(ii) For every κ_1 that is sufficiently large, there exist $\underline{c}_{12}$ and $\bar{c}_{12}$ such that $0 < \underline{c}_{12} \leq \lambda_{\min}(Q_{LL_x}) \leq \lambda_{\max}(Q_{LL_x}) \leq \bar{c}_{12} < \infty$ and $\lambda_{\max}(Q_{LL_x,U}) \leq \bar{c}_{12} < \infty$.

(iii) For every κ that is sufficiently large, there exist $\underline{c}_2$ and $\bar{c}_2$ such that $0 < \underline{c}_2 \leq \lambda_{\min}(Q_{\Phi\Phi}) \leq \lambda_{\max}(Q_{\Phi\Phi}) \leq \bar{c}_2 < \infty$.

(iv) The functions $\{\pi_1(\cdot), \pi_2(\cdot), g_1(\cdot), g_2(\cdot), \lambda_1(\cdot), \lambda_2(\cdot)\}$, belong to the class of γ-smooth functions with $\gamma \geq 2$.

(v) There exist α's such that $\sup_{z\in\mathcal{Z}_1}|m_l(z)-\rho^{\kappa_1}(z)'\alpha_l| = O(\kappa_1^{-\gamma})$ for $l=1,2$, $\sup_{z\in\mathcal{Z}_2}|m_l(z)-\rho^{\kappa_1}(z)'\alpha_l| = O(\kappa_1^{-\gamma})$ for $l=1,2$.

(vi) There exist ψ's such that $\sup_{z\in\mathcal{Z}_1}|\pi_l(z)-\rho^{\kappa_1}(z)'\psi_l| = O(\kappa_1^{-\gamma})$ for $l=1,2$, $\sup_{z\in\mathcal{Z}_2}|\pi_l(z)-\rho^{\kappa_1}(z)'\psi_l| = O(\kappa_1^{-\gamma})$ for $l=1,2$.

(vii) There exist θ's such that $\sup_{x\in\mathcal{X}}|g_1(x)-\rho^{\kappa}(x)'\theta_l| = O(\kappa^{-\gamma})$,
$\sup_{z\in\mathcal{Z}_1}|g_2(z_1)-\rho^{\kappa}(z)'\theta_l| = O(\kappa^{-\gamma})$, $|\lambda_1(\cdot)-\rho^{\kappa}(\cdot)'\theta_l|_1 = O(\kappa^{-\gamma})$, and
$|\lambda_2(\cdot)-\rho^{\kappa}(\cdot)'\theta_l|_1 = O(\kappa^{-\gamma})$, for $l=1,,d$.

(viii) The set of basis functions, $\rho(\cdot)$, are twice continuously differentiable on the support of U_i. $\max_{0\le s\le r}\sup_{u\in\mathcal{U}}\|\partial^s\rho^{\kappa}(u)\| \le \varsigma_{r\kappa}$ for $r=0,1,2$.

(ix) The set of basis functions, $\rho(\cdot)$, are twice continuously differentiable on the support of P_i. $\max_{0\le s\le r}\sup_{p\in\mathcal{P}}\|\partial^s\rho^{\kappa}(p)\| \le \varsigma_{r\kappa}$ for $r=0,1,2$.

Assumption A3. (i) The marginal or joint probability densities of any single or two elements in $\mathbf{W}_i$ are bounded, bounded away from zero, and twice continuously differentiable.

(ii) Let $\sigma_i^2 \equiv \sigma^2(\mathbf{X}_i,\mathbf{Z}_i,\mathbf{U}_i,\mathbf{P}_i) \equiv E\left(e_i^2|\mathbf{X}_i,\mathbf{Z}_i,\mathbf{U}_i,\mathbf{P}_i\right)$ and $Q_{s,pp} \equiv E[p^{\kappa_1}(Z_{s,i})\,p^{\kappa_1}(Z_{s,i})'\,\sigma_i^2]$ for $s=1,2$. $\lambda(Q_{s,pp})$ is bounded uniformly in κ_1.

Assumption A4. $K(\cdot)$ is a bounded kernel such that:

(i) $\int k(v)\,dv = 1$ (ii) $k(v)\,dv = k(-v)\,dv$ (iii) $\int v^2k(v)\,dv > 0$ (iv) $k(v) = 0 for |v| > 1$ (v) For $T_j(v) = |v|^j k(v)$, $|T_j(v_1)-T_j(v_2)| \le C|v_2-v_1|$

Assumption A5. (i) $\kappa_1 \le \kappa$. As $n\to\infty$, $\kappa_1\to\infty$, $\kappa^3/n\to 0$ and $\tau_n \to c_1 \in [0,\infty)$, where $\tau_n \equiv \left(\kappa^{1/2}\varsigma_{0\kappa}+\varsigma_{1\kappa}\right)\nu_{1n}+\varsigma_{0\kappa}\varsigma_{2\kappa}\nu_{1n}^2$, $\nu_{1n} \equiv \kappa_1^{1/2}/n^{1/2}+\kappa_1^{-\gamma}$ and $\nu_n \equiv \kappa^{1/2}/n^{1/2}+\kappa^{-\gamma}$. (ii) As $n\to\infty$, $h\to 0$, $nh^3\log n\to\infty$, $nh\kappa^{-2\gamma}\to 0$, $\tau_n\nu_{1n} = o(n^{-1/2}h^{-1/2})$ and $[h^{1/2}\varsigma_{1\kappa}(1+n^{1/2}\kappa_1^{-\gamma}) + \varsigma_{2\kappa}n^{1/2}h^{1/2}\nu_{1n}^2](\nu_n+\nu_{1n}+\nu_{1n}) \to 0$.

Assumptions listed are either same or similar to those in Ozabaci et. al (2014). For completeness we list them all here.

A1 ensures that the random sample is independently and identically distributed, over compact support. By the help of compact support argument, we can employ univariate splines. However this assumption can be dropped, by the help of multivariate splines. Here we ensure that all variables, hence P and U to be continuously valued. However we will adopt a less restrictive assumption, on Assumption B1.

Assumption A2 is to ensure that the asymptotic covariance matrix exists and is nonsingular, functions satisfy certain smoothness conditions and approximation errors are as listed, for the first two stages. Please note that although the Q terms are similar to those in Ozabaci et. al (2014), here the assumption differs due to the existence of two first step regressions, and the trimming function included in ϕ.

Assumptions A3 and A4 are standard in the literature.

Assumptions A5 is directly adopted from Ozabaci et. al. (2014), except the final term. We assume that approximation terms for the two first stage regressions are equal, which gives us the similar results to those in Ozabaci et. al. (2014). However this assumption can be relaxed, which may effect the rates.

8.3 Proofs of first two theorems:

For our proofs, we employ the six Lemmas provided by Ozabaci et. al. (2014). Again, for completeness we write those lemmas here. Please note that they are modified in order to account for the sample selection, in addition to endogeneity. The details of the proofs provided here, as well as the proofs of the lemmas provided below, can be found in the online Supplemental Material.

Under assumptions A1-A5:

Lemma 8.1 *(i)* $\|Q_{n,LL_x}-Q_{LL_x}\|^2 = O_P\left(\kappa_1^2/n\right)$;
(ii) $\lambda_{\min}(Q_{n,LL_x}) = \lambda_{\min}(Q_{LL_x})+o_P(1)$ *and* $\lambda_{\max}(Q_{n,LL_x}) = \lambda_{\max}(Q_{LL_x})+o_P(1)$;
(iii) $\left\|Q_{n,LL_x}^{-1}-Q_{LL_x}^{-1}\right\|_{sp} = O_P\left(\kappa_1/n^{1/2}\right)$;
(iv) $\|Q_{n,\Phi\Phi}-Q_{\Phi\Phi}\|^2 = O_P\left(\kappa^2/n\right)$;
(v) $\lambda_{\min}(Q_{n,\Phi\Phi}) = \lambda_{\min}(Q_{\Phi\Phi})+o_P(1)$ *and* $\lambda_{\max}(Q_{n,\Phi\Phi}) = \lambda_{\max}(Q_{\Phi\Phi})+o_P(1)$;
(i) $\left\|Q_{n,LL_p}-Q_{LL_p}\right\|^2 = O_P\left(\kappa_1^2/n\right)$;
(ii) $\lambda_{\min}\left(Q_{n,LL_p}\right) = \lambda_{\min}\left(Q_{LL_p}\right)+o_P(1)$ *and* $\lambda_{\max}\left(Q_{n,LL_p}\right) = \lambda_{\max}\left(Q_{LL_p}\right)+o_P(1)$;
(iii) $\left\|Q_{n,LL_p}^{-1}-Q_{LL_p}^{-1}\right\|_{sp} = O_P\left(\kappa_1/n^{1/2}\right)$;

Lemma 8.2 *Where* $\xi_{1n} \equiv n^{-1}\sum_{i=1}^{n} L_{xi}U_i$,
$\xi_{pn} \equiv n^{-1}\sum_{i=1}^{n} L_{pi}v_i$,
$\zeta_{1n} \equiv n^{-1}\sum_{i=1}^{n} L_{xi}\left[\pi\left(\mathbf{Z}_i\right) - L'_{xi}\psi\right]$
$\zeta_{pn} \equiv n^{-1}\sum_{i=1}^{n} L_{xi}\left[m\left(\mathbf{Z}_i\right) - L'_{pi}\alpha\right]$,
(i) $\|\xi_{1n}\|^2 = O_P(\kappa_1/n)$;
(ii) $\|\zeta_{1n}\|^2 = O_P(\kappa_1^{-2\gamma})$;
(iii) $\|\xi_{pn}\|^2 = O_P(\kappa_1/n)$;
(iv) $\|\zeta_{pn}\|^2 = O_P(\kappa_1^{-2\gamma})$;
(v) $\widetilde{\psi}-\psi = Q_{\kappa_1}^{-1}n^{-1}\sum_{i=1}^{n} LL_{xi}U_i + Q_{\kappa_1}^{-1}n^{-1}\sum_{i=1}^{n} LL_{xi}\left[\pi\left(\mathbf{Z}_i\right) - P'_i\psi\right] + r_{1n}$;
(vi) $\widetilde{\alpha}-\alpha = Q_{\kappa_1}^{-1}n^{-1}\sum_{i=1}^{n} LL_{pi}V_i + Q_{\kappa_1}^{-1}n^{-1}\sum_{i=1}^{n} LL_{pi}\left[m\left(\mathbf{Z}_i\right) - LL'_{pi}\alpha\right] + r_{pn}$;
where $\|r_{nl}\|$ *and* $\|r_{nl}\|$ *are both* $O_P(\kappa_1/n + \kappa_1^{-\gamma+1/2}/n^{1/2})$.

Lemma 8.3 *(i)* $n^{-1}\sum_{i=1}^{n}\left(\widetilde{U}_i - U_i\right)^2\left[\sigma_i^2\right]^r = O_P\left(\nu_{1n}^2\right)$ *for* $r = 0, 1$;
(ii) $n^{-1}\sum_{i=1}^{n}\left(\widetilde{U}_i - U_i\right)^2\|\Phi_i\|^r = O_P\left(\varsigma_{0\kappa}^r\nu_{1n}^2\right)$ *for* $r = 1, 2$;
(iii) $n^{-1}\sum_{i=1}^{n}\left\|p^\kappa\left(\widetilde{U}_i\right) - p^\kappa\left(U_i\right)\right\|^2 = O_P\left(\varsigma_{1\kappa}^2\nu_{1n}^2\right)$;
(iv) $\left\|n^{-1}\sum_{i=1}^{n}\left[p^\kappa\left(\widetilde{U}_i\right) - p^\kappa\left(U_i\right)\right]\Phi'_i\right\| = O_P\left(\kappa^{1/2}\varsigma_{0\kappa}\nu_{1n} + \varsigma_{0\kappa}\varsigma_{2\kappa}\nu_{1n}^2\right)$;
(v) $\left\|n^{-1}\sum_{i=1}^{n}\left[p^\kappa\left(\widetilde{U}_i\right) - p^\kappa\left(U_i\right)\right]e_i\right\| = O_P\left(n^{-1/2}\varsigma_{1\kappa}\nu_{1n}\right)$;
(vi) $n^{-1}\sum_{i=1}^{n}\left(\widetilde{P}_i - P_i\right)^2\left[\sigma_i^2\right]^r = O_P\left(\nu_{1n}^2\right)$ *for* $r = 0, 1$;
(vii) $n^{-1}\sum_{i=1}^{n}\left(\widetilde{P}_i - P_i\right)^2\|\Phi_i\|^r = O_P\left(\varsigma_{0\kappa}^r\nu_{1n}^2\right)$ *for* $r = 1, 2$;
(viii) $n^{-1}\sum_{i=1}^{n}\left\|p^\kappa\left(\widetilde{P}_i\right) - p^\kappa\left(P_i\right)\right\|^2 = O_P\left(\varsigma_{1\kappa}^2\nu_{1n}^2\right)$;
(ix) $\left\|n^{-1}\sum_{i=1}^{n}\left[p^\kappa\left(\widetilde{P}_i\right) - p^\kappa\left(P_i\right)\right]\Phi'_i\right\| = O_P\left(\kappa^{1/2}\varsigma_{0\kappa}\nu_{1n} + \varsigma_{0\kappa}\varsigma_{2\kappa}\nu_{1n}^2\right)$;
(x) $\left\|n^{-1}\sum_{i=1}^{n}\left[p^\kappa\left(\widetilde{P}_i\right) - p^\kappa\left(P_i\right)\right]e_i\right\| = O_P\left(n^{-1/2}\varsigma_{1\kappa}\nu_{1n}\right)$.

Lemma 8.4 *(i)* $n^{-1}\sum_{i=1}^{n}\left\|\widetilde{\Phi}_i - \Phi_i\right\|^2 = O_P\left(\varsigma_{1\kappa}^2\nu_{1n}^2\right)$;
(ii) $\left\|n^{-1}\sum_{i=1}^{n}\left(\widetilde{\Phi}_i - \Phi_i\right)\Phi'_i\right\|_{sp} = O_P\left(\kappa^{1/2}\varsigma_{1\kappa}\nu_{1n}\right)$;
(iii) $\left\|\widetilde{Q}_{n,\Phi\Phi} - Q_{n,\Phi\Phi}\right\|_{sp} = O_P\left(\kappa^{1/2}\varsigma_{0\kappa}\nu_{1n} + \varsigma_{0\kappa}\varsigma_{2\kappa}\nu_{1n}^2\right)$;
(iv) $\left\|\widetilde{Q}_{n,\Phi\Phi}^{-1} - Q_{\Phi\Phi}^{-1}\right\|_{sp} = O_P\left(\kappa^{1/2}\varsigma_{0\kappa}\nu_{1n} + \varsigma_{0\kappa}\varsigma_{2\kappa}\nu_{1n}^2\right)$;
(v) $\left\|n^{-1}\sum_{i=1}^{n}\left(\widetilde{\Phi}_i - \Phi_i\right)e_i\right\| = O_P\left(n^{-1/2}\varsigma_{1\kappa}\nu_{1n}\right)$;
(vi) $n^{-1}\sum_{i=1}^{n}\left(\widetilde{\Phi}_i - \Phi_i\right)\left[\bar{g}\left(X_i, Z_{1i}, U_i, P_i\right) - \Phi'_i\theta\right] = O_P\left(\kappa_1^{-\gamma}\varsigma_{1\kappa}\nu_{1n}\right)$.

Lemma 8.5 $\xi_n \equiv n^{-1}\sum_{i=1}^{n}\Phi_i e_i$ *and* $\zeta_n \equiv n^{-1}\sum_{i=1}^{n}\Phi_i\left[\bar{g}\left(X_i, Z_{1i}, U_i, P_i\right) - \Phi'_i\theta\right]$.
(i) $\|\xi_n\| = O_P(\kappa^{1/2}/n^{1/2})$;
(iii) $\|\zeta_n\| = O_P(\kappa^{-\gamma})$;
(iii) $\left\|Q_{\Phi\Phi}^{-1}n^{-1}\sum_{i=1}^{n}\Phi_i\dot{p}^\kappa\left(U_i\right)'\theta\left(\widetilde{U}_i - U_i\right)\right\| = O_P\left(\nu_{1n}\right)$
(iv) $\left\|Q_{\Phi\Phi}^{-1}n^{-1}\sum_{i=1}^{n}\Phi_i\dot{p}^\kappa\left(P_i\right)'\theta\left(\widetilde{P}_i - P_i\right)\right\| = O_P\left(\nu_{1n}\right)$.

Lemma 8.6 *(i)* $S_{1n}\left(z_1\right) \equiv n^{-1/2}h^{1/2}\sum_{i=1}^{n} K_{iz_1}c'H^{-1}Z_{1i}^*\left(z_1\right)\left(\widetilde{U}_i - U_i\right)\dot{p}^\kappa\left(U_i\right) +$
$n^{-1/2}h^{1/2}\sum_{i=1}^{n} K_{iz_1}c'H^{-1}Z_{1i}^*\left(z_1\right)\left(\widetilde{P}_i - P_i\right)\dot{p}^\kappa\left(P_i\right) = h^{1/2}\varsigma_{1\kappa}O_P(1 + n^{1/2}\kappa_1^{-\gamma})$ *uniformly in* z_1;
(ii) $S_{2n}\left(z_1\right) \equiv n^{-1/2}h^{1/2}\sum_{i=1}^{n} K_{iz_1}\left|c'H^{-1}Z_{1i}^*\left(z_1\right)\right|\left(\widetilde{U}_i - U_i\right)^2 +$
$n^{-1/2}h^{1/2}\sum_{i=1}^{n} K_{iz_1}\left|c'H^{-1}Z_{1i}^*\left(z_1\right)\right|\left(\widetilde{P}_i - P_i\right)^2 = n^{1/2}h^{1/2}O_P(\nu_{1n}^2)$ *uniformly in* z_1.

Sketch proof of Theorem 4.1. $Y_i = \bar{g}(X_i, Z_{1i}, U_i, P_i) + e_i = \widetilde{\Phi}_i'\theta + e_i + [\bar{g}(X_i, Z_{1i}, U_i, P_i) - \widetilde{\Phi}_i'\theta]$, hence we can write that

$$\begin{aligned}
\widetilde{\theta} - \theta &= \widetilde{Q}_{n,\Phi\Phi}^{-1} n^{-1} \sum_{i=1}^{n} \widetilde{\Phi}_i Y_i - \theta = \widetilde{Q}_{n,\Phi\Phi}^{-1} n^{-1} \sum_{i=1}^{n} \widetilde{\Phi}_i e_i + \widetilde{Q}_{n,\Phi\Phi}^{-1} n^{-1} \sum_{i=1}^{n} \widetilde{\Phi}_i \left[\bar{g}(X_i, Z_{1i}, U_i, P_i) - \widetilde{\Phi}_i'\theta\right] \\
&= \widetilde{Q}_{n,\Phi\Phi}^{-1} \xi_n + \widetilde{Q}_{n,\Phi\Phi}^{-1} \zeta_n + \widetilde{Q}_{n,\Phi\Phi}^{-1} n^{-1} \sum_{i=1}^{n} \Phi_i \left(\Phi_i - \widetilde{\Phi}_i\right)' \theta + \widetilde{Q}_{n,\Phi\Phi}^{-1} n^{-1} \sum_{i=1}^{n} \left(\widetilde{\Phi}_i - \Phi_i\right) e_i \\
&\quad + \widetilde{Q}_{n,\Phi\Phi}^{-1} n^{-1} \sum_{i=1}^{n} \left(\widetilde{\Phi}_i - \Phi_i\right) [\bar{g}(X_i, Z_{1i}, U_i, P_i) - \Phi_i'\theta] - \widetilde{Q}_{n,\Phi\Phi}^{-1} n^{-1} \sum_{i=1}^{n} \left(\widetilde{\Phi}_i - \Phi_i\right) \left(\widetilde{\Phi}_i - \Phi_i\right)' \theta
\end{aligned}$$

The proof follows from Lemmas 3, 4 and 5.

Sketch proof of Theorem 4.2.
Let $Y_{1i} \equiv Y_i - \mu_y - g_2(Z_{1i}) - \lambda_1(P_i) - \lambda_2(U_i)$ and $\mathbf{Y}_1 \equiv (Y_{11}, ..., Y_{1n})'$. Again using the notation of Ozabaci et. al. (2014), we can write the local-linear estimator as:

$$\begin{aligned}
& \left[H^{-1}\mathbb{Z}_1(z_1)' \mathbb{K}_{z_1} \mathbb{Z}_1(z_1) H^{-1}\right]^{-1} H^{-1}\mathbb{Z}_1(z_1)' \mathbb{K}_{z_1} \mathbb{Z}_1(z_1) \mathbf{Y}_1 \\
& + \left[H^{-1}\mathbb{Z}_1(z_1)' \mathbb{K}_{z_1} \mathbb{Z}_1(z_1) H^{-1}\right]^{-1} H^{-1}\mathbb{Z}_1(z_1) \mathbb{K}_{z_1} (\widetilde{\mathbf{Y}}_1 - \mathbf{Y}_1) \\
\equiv \; & J_{1n}(z_1) + J_{2n}(z_1)
\end{aligned}$$

Lemma 8.7 *(i.)* $n^{-1}H^{-1} \times \mathbb{Z}_1(z_1)' \mathbb{K}_{z_1} \mathbb{Z}_1(z_1) H^{-1} = f_{Z_1}(z_1) \begin{pmatrix} 1 & 0 \\ 0 & \int u^2 K(u)\, du \end{pmatrix} + o_P(1)$ *uniformly in* z_1,

(ii.) $n^{1/2}h^{1/2} \times [J_{1n}(z_1) - b_1(z_1)] \xrightarrow{D} N(0, \Omega_1(z_1))$ *and* $\sup_{z_1} \in \mathcal{Z}_1 \|J_{1n}(z_1)\| = O_P\left((nh/\log n)^{-1/2} + h^2\right)$,

(iii.) $n^{-1/2}h^{1/2}H^{-1}\mathbb{Z}_1(z_1)' \mathbb{K}_{z_1}(\widetilde{\mathbf{Y}}_1 - \mathbf{Y}_1) = o_P(1)$ *uniformly in* z_1.

Proof of Lemma 8.7 follows from the standard arguments in Masry (1996) and Hansen (2008). We can decompose Lemma 8.7, (*iii.*) as following:

$$\begin{aligned}
n^{-1/2}h^{1/2}H^{-1}\mathbb{Z}_1(Z_1)' \mathbb{K}_{Z_1}(\widetilde{\mathbf{Y}}_1 - \mathbf{Y}_1) &= \sqrt{n}(\widetilde{\mu} - \mu) n^{-1}h^{1/2} \sum_{i=1}^{n} K_{iZ_1} H^{-1} Z_{1i}^*(Z_1) \\
&+ n^{-1/2}h^{1/2} \sum_{i=1}^{n} K_{iZ_1} H^{-1} Z_{1i}^*(Z_1) [\widetilde{g}_1(X_i) - g_1(X_i)] \\
&+ n^{-1/2}h^{1/2} \sum_{i=1}^{n} K_{iZ_1} H^{-1} Z_{1i}^*(Z_1) [\widetilde{g}_2(Z_{1i}) - g_2(Z_{1i})] \\
&+ n^{-1/2}h^{1/2} \sum_{i=1}^{n} K_{iZ_1} H^{-1} Z_{1i}^*(Z_1) \left[\widetilde{\lambda}_2(\widetilde{U}_i) - \lambda_2(U_i)\right] \\
&+ n^{-1/2}h^{1/2} \sum_{i=1}^{n} K_{iZ_1} H^{-1} Z_{1i}^*(Z_1) \left[\widetilde{\lambda}_1(\widetilde{P}_i) - \lambda_1(P_i)\right]
\end{aligned}$$

From Lemma 8.6 , we can show that Lemma 8.7, (*iii.*) is $= o_P(1)$. This completes the proof of Theorem 4.2. ■

8.4 Assumptions B

Assumption B1.
(i) The triple $\{Y_i, \mathbf{X}_i, \mathbf{Z}_i\}$ are IID and random, for $i = 1, ..., n$.

(ii) $\mathbf{Z}_i$ have compact support, and is continuously distributed with respect to the Lebesgue measure.
(iii) $\mathbf{X}_i$ has finite and known support.

Assumption B2. (i) For every κ_1 that is sufficiently large, there exist $\underline{c}_{11}$ and $\bar{c}_{11}$ such that $0 < \underline{c}_{11} \leq \lambda_{\min}\left(Q_{LL_p}\right) \leq \lambda_{\max}\left(Q_{LL_p}\right) \leq \bar{c}_{11} < \infty$ and $\lambda_{\max}\left(Q_{LL_p,\sigma}\right) \leq \bar{c}_1 < \infty$.

(ii) For every κ_1 that is sufficiently large, there exist $\underline{c}_{12}$ and $\bar{c}_{12}$ such that $0 < \underline{c}_{12} \leq \lambda_{\min}\left(Q_{LL_x}\right) \leq \lambda_{\max}\left(Q_{LL_x}\right) \leq \bar{c}_{12} < \infty$ and $\lambda_{\max}\left(Q_{LL_x,U}\right) \leq \bar{c}_1 < \infty$.

(iii) For every κ_2 that is sufficiently large, there exist $\underline{c}_{22}$ and $\bar{c}_{22}$ such that $0 < \underline{c}_{22} \leq \lambda_{\min}\left(Q_{\Phi\Phi_2}\right) \leq \lambda_{\max}\left(Q_{\Phi\Phi_2}\right) \leq \bar{c}_{22} < \infty$.

(iv) The functions $\{\pi(\cdot), g_2(\cdot), \lambda(\cdot)\}$, belong to the class of γ-smooth functions with $\gamma \geq 2$.

(v) There exist $\alpha_{\mathbf{l}}$'s such that $\sup_{z\in\mathcal{Z}_l} |m_l(z) - \rho^{\kappa_1}(z)'\alpha_l| = O(\kappa_1^{-\gamma})$ for $l = 1, 2$.

(vi) There exist $\psi_{\mathbf{l}}$'s such that $\sup_{z\in\mathcal{Z}_l} |\pi_l(z) - \rho^{\kappa_1}(z)'\psi_l| = O(\kappa_1^{-\gamma})$ for $l = 1, 2$.

(vii) There exist $\theta_{\mathbf{2}}$'s such that;
$\sup_{z\in\mathcal{Z}_1} |g_2(z_1) - \rho^{\kappa_2}(z_1)'\theta_{\mathbf{2}l}| = O(\kappa^{-\gamma})$,
$|\lambda(P_i) - \rho^{\kappa_2}(\cdot)'\theta_{\mathbf{2}k}| = O(\kappa_2^{-\gamma})$, for $l = 1$ and $k = 2$.

(viii) The set of basis functions, $\rho(\cdot)$, are twice continuously differentiable on the support of P_i. $\max_{0\leq s\leq r} \sup_{p\in\mathcal{P}} \|\partial^s \rho^\kappa(p)\| \leq \varsigma_{r\kappa}$ for $r = 0, 1, 2$.

Assumption B3. (i) Any two elements of $(\mathbf{Z_i}, \mathbf{P_i}, \pi_{\mathbf{i}})$ have bounded densities, which are bounded away from zero, and twice continuously differentiable.

(ii) Let $\sigma_i^2 \equiv \sigma^2(\mathbf{X}_i, \mathbf{Z}_i, \pi_{\mathbf{i}}, \mathbf{P}_i) \equiv E\left(e_i^2|\mathbf{X}_i, \mathbf{Z}_i, \pi_{\mathbf{i}}, \mathbf{P}_i\right)$ and $Q_{s,pp} \equiv E[p^{\kappa_1}(Z_{s,i})\, p^{\kappa_1}(Z_{s,i})'\, \sigma_i^2]$ for $s = 1, 2$. $\lambda_{max}(Q_{s,pp})$ is bounded uniformly in κ_1.

Assumption B4. $K(\cdot)$ is a bounded kernel such that:
(i) $\int k(v)\,dv = 1$ (ii) $k(v)\,dv = k(-v)\,dv$ (iii) $\int v^2 k(v)\,dv > 0$ (iv) $k(v) = 0 for |v| > 1$ (v) For $T_j(v) = |v|^j k(v)$, $|T_j(v_1) - T_j(v_2)| \leq C|v_2 - v_1|$.

Assumption B5. (i) $\kappa_1 \leq \kappa_2$. As $n \to \infty$, $\kappa_1 \to \infty$, $\kappa_2^3/n \to 0$ and $\tau_{2n} \to c_1 \in [0, \infty)$, where $\tau_{2n} \equiv \left(\kappa_2^{1/2}\varsigma_{0\kappa} + \varsigma_{1\kappa_2}\right)\nu_{1n} + \varsigma_{0\kappa}\varsigma_{2\kappa_2}\nu_{1n}^2$, $\nu_{1n} \equiv \kappa_1^{1/2}/n^{1/2} + \kappa_1^{-\gamma}$ and $\nu_{2n} \equiv \kappa_2^{1/2}/n^{1/2} + \kappa_2^{-\gamma}$. (ii) As $n \to \infty$, $h \to 0$, $nh^3 \log n \to \infty$, $nh\kappa_2^{-2\gamma} \to 0$, $\tau_n\nu_{1n} = o(n^{-1/2}h^{-1/2})$ and $[h^{1/2}\varsigma_{1\kappa_2}(1 + n^{1/2}\kappa_1^{-\gamma}) + \varsigma_{2\kappa_2} n^{1/2}h^{1/2}\nu_{1n}^2](\nu_{2n} + \nu_{1n}) \to 0$. (iii) $\nu_{1n} = o(n^{-1/2})$ and $\nu_{2n} = o(n^{-1/2})$.

8.5 Sketch Proof of Theorems 4.3 and 4.4

Since $\widetilde{\beta_0}$ can be estimated at the parametric $\sqrt{n}$ rate, the remainder of the results hold for the second model, as well.

Let's define $\widetilde{\theta}_3 = \widetilde{Q}_{n,\Phi_2\Phi_2}^{-1} n^{-1} \sum_{i=1}^n \widetilde{\Phi}_{2i}\left(Y_i - \widetilde{\beta}_0 X_i\right)$; where $\theta_2 = [\beta_0, \theta_3']'$.

Since $\widetilde{\beta}_0 - \beta_0$ is $op\left(n^{-1/2}\right)$, it is going to be converging faster than the nonparametric counterparts, hence $\widetilde{\theta}_3$ and θ_3 will be asymptotically equivalent. After reaching this conclusion, proof of Theorem 4.4 follows from proof of Theorem 4.2. ■