The Asymptotic Variance of Semi-parametric Estimators with Generated Regressors*

Jinyong Hahn
Department of Economics, UCLA

Geert Ridder
Department of Economics, USC

October 1, 2010

Abstract

We study the asymptotic distribution of three-step estimators of a finite dimensional parameter vector where the second step consists of one or more nonparametric regressions on a regressor that is estimated in the first step. The first step estimator is either parametric or non-parametric. Using Newey's (1994) path-derivative method we derive the contribution of the first step estimator to the influence function. In this derivation it is important to account for the dual role that the first step estimator plays in the second step non-parametric regression, i.e., that of conditioning variable and that of argument. We consider three examples in more detail: the partial linear regression model estimator with a generated regressor, the Heckman, Ichimura and Todd (1998) estimator of the Average Treatment Effect and a semi-parametric control variable estimator.

JEL Classification: C01, C14.

Keywords: Semi-parametric estimation, generated regressors, asymptotic variance.

*Financial support for this research was generously provided through NSF SES 0819612 and 0819638. We thank Guido Imbens and seminar participants at UC Riverside, the Tinbergen Institute, Yale,FGV-Rio de Janeiro, FGV-São Paulo, Harvard/MIT, UCLA and Mannheim for comments. Geert Ridder thanks the Department of Economics, PUC, Rio de Janeiro for their hospitality. Addresses: Jinyong Hahn, Department of Economics, Bunche Hall, UCLA, Los Angeles, CA 90095, hahn@econ.ucla.edu; Geert Ridder, Department of Economics, Kaprilian Hall, USC, Los Angeles, CA 90089, ridder@usc.edu.

1 Introduction

In a seminal contribution Pagan (1984) derived the asymptotic variance of regression coefficient estimators in linear regression models, if (some of) the regressors are themselves estimated in a preliminary step. Pagan called such regressors generated regressors and he characterized the contribution of the estimation error in the generated regressors to the total asymptotic variance of the regression coefficient estimators. Examples of generated regressors are linear predictors or residuals from an estimated equation as in Barro (1977) or Shefrin (1979). The estimators considered by Pagan are special cases of standard two-step estimators, and such estimators can be conveniently analyzed as single-step GMM estimators, as in Newey (1984) or Murphy and Topel (1985). These methods of adjusting the asymptotic variance for the first-stage estimation error are now so well-understood that they can be found in textbooks such as Wooldridge (2002, Chapter 12.4).

Pagan (1984) considered parametric linear regression models with parametrically estimated generated regressors. However, econometrics has evolved since then, and the first step estimators these days can be non-parametric estimators obtained by kernel or sieve methods. Newey (1994) discusses a general method of characterizing the asymptotic variance of two-step GMM estimators of a finite dimensional parameter vector, if the moment condition depends on a conditional expectation or a density that is estimated non-parametrically. A special instance of his method deals with the case of a linear regression model with a non-parametrically estimated generated regressor. Newey uses path derivatives to obtain the influence function for semi-parametric GMM estimators. The asymptotically linear representation of the estimator gives the asymptotic variance of the estimator. After this derivation it still has to be shown that that the difference between the semi-parametric GMM estimator and its asymptotically linear representation converges to 0 at a rate that is faster than the parametric rate. Sufficient conditions for this in general depend on the non-parametric estimator and smoothness of the conditional expectation or density that is estimated. Given the complexity of the multi-step estimators it is useful to have the influence function before one considers the asymptotic properties of remainder terms.

The asymptotic properties of non-parametric two-step estimators where both the generated regressor and the second-stage regression are estimated non-parametrically have been studied by Sperlich (2009) and Song (2008). Non-parametric multi-step estimators are not considered in this paper. As in Newey (1994) we will only consider semi-parametric estimators for finite dimensional parameters. The difference with Newey is that we consider three-step estimators where the second step is a non-parametric regression on a generated regressor. As we discuss in this paper the effect of the first-stage estimation error on the asymptotic variance of estimator of the finite dimensional parameter is qualitatively different for the two- and three-step semi-parametric estimators. Also the results for two-step non-parametric estimators cannot be used directly to obtain the influence function for semi-parametric three-step estimators.

The purpose of this note is to use Newey's path-derivative method to derive the asymptotic variance of three- or even multi-step estimators of a finite dimensional parameter in which one of the steps is a non-parametric regression with a generated regressor. The generated regressor that is estimated in the first step can be estimated parametrically or non-parametrically. Since Newey (1994), a number of estimators have been suggested that have this structure with one

of the steps a non-parametric regression on a generated regressor. We consider three examples: (i) the partially linear regression model with a generated regressor in Wooldridge and Lee (2002) and Newey (2009), (ii) the Average Treatment Effect (ATE) estimator for the case of unconfounded treatment assignment suggested by Heckman, Ichimura, and Todd (1998) that involves two non-parametric regressions on the estimated propensity score, (iii) a parametric control variate estimator that depends on a non-parametric regression on a residual estimated in a first stage. These examples illustrate the method that can also be used to derive the asymptotic variance of other estimators with the same structure not covered here, for instance the production function estimators of Pakes and Olley (1995) and Olley and Pakes (1996).

The key issue in the application of Newey's path-derivative method is to account for the contribution of the first-stage estimation error of the generated regressor on the sampling variation of the second-stage nonparametric regression. This contribution consists of two parts. First, there is the effect of the first-step estimation error on the estimate of the generated regressor. However, there is a second contribution to the sampling variation of the conditional expectation, because we condition on an estimated instead of a population value of the regressor. It is the latter contribution that is easily forgotten.

One can wonder whether the reformulation of the two-step estimator of Pagan (1984) as a one-step GMM estimator as in Newey (1984) or Murphy and Topel (1985) can be generalized to the three or more step estimator considered here. In particular, Ai and Chen (2007) recently considered a variety of conditional moment restriction estimators, some with a more complicated structure than in this paper, where the conditioning variables are not estimated. Therefore our results are not a special case of, but rather complementary to the results in Ai and Chen. Whether our asymptotic variance can be derived from a one step GMM problem as in Ai and Chen (2007) is the subject of ongoing research.

This paper has the following structure. In Section 2, we present a parametric example that provides the basic intuition underlying our results. Our main result is in Section 3. In Sections 4, 5 and 6, we discuss the three applications mentioned above.

2 A Parametric Example

To gain intuition for the results later on we consider a fully parametric, be it somewhat artificial example. Consider the following scenario. We have a random sample $w_i = (y_i, x_i, z_i), i = 1, \ldots, n$ from a joint distribution. The scalar parameter β is estimated by a three-step estimator. In the first step, we estimate the scalar parameter α by $\widehat{\alpha}$ such that

$$\sqrt{n}\left(\widehat{\alpha} - \alpha_*\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi\left(x_i, z_i\right) + o_p\left(1\right)$$

with $\mathbb{E}\left[\psi\left(x_i, z_i\right)\right] = 0$ and α_* the population value of the parameter. In the second step, we estimate the coefficients $\gamma_* = (\gamma_{1*}, \gamma_{2*}, \gamma_{3*})$ of the linear projection of y on $1, x, v$ with $v = \varphi\left(x, z, \alpha_*\right)$, i.e., the solution to $\min_{\gamma_1,\gamma_2,\gamma_3} \mathbb{E}\left[\left(y - \gamma_1 - \gamma_2 x - \gamma_3 v\right)^2\right]$. Because we do not know α_*, we use the estimated $\widehat{v}_i = \varphi\left(x_i, z_i, \widehat{\alpha}\right)$, so that the estimator $\widehat{\gamma}$ of γ_* is the OLS estimator of y on $x, \widehat{v}$. The estimator of β_* is obtained in the third step $\widehat{\beta} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\gamma}_1 + \widehat{\gamma}_2 x_i + \widehat{\gamma}_3 \varphi\left(x_i, z_i, \widehat{\alpha}\right)\right)$, so that $\beta_* = \mathbb{E}[\gamma_{1*} + \gamma_{2*}x + \gamma_{3*}\varphi(x, z, \alpha_*)]$. Our interest is to characterize the first order asymptotic properties of this estimator.

A standard argument suggests that it suffices to consider the expansion of the form

$$\begin{aligned}\sqrt{n}\left(\widehat{\beta}-\beta_*\right) =& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\gamma_{1*}+\gamma_{2*}x_i+\gamma_{3*}\varphi\left(x_i,z_i,\alpha_*\right)-\beta_*\right)\\ &+\left[\begin{array}{ccc}1 & \mathbb{E}\left[x\right] & \mathbb{E}\left[\varphi\left(x,z,\alpha_*\right)\right]\end{array}\right]\sqrt{n}\left(\widehat{\gamma}-\gamma_*\right)\\ &+\mathbb{E}\left[\gamma_{3*}\frac{\partial\varphi\left(x,z,\alpha_*\right)}{\partial\alpha}\right]\sqrt{n}\left(\widehat{\alpha}-\alpha_*\right)+o_p\left(1\right).\end{aligned}$$

Let us now focus on the adjustments to the influence function that account for the estimation error in the first and second step, i.e., the sum of the second and third terms on the right, which we will call Δ. A routine calculation (presented in Appendix A) reveals that

$$\Delta=-\left[\begin{array}{ccc}1 & \mathbb{E}\left[x\right] & \mathbb{E}\left[\varphi\left(x,z,\alpha_*\right)\right]\end{array}\right]G_\gamma^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\begin{array}{c}\varepsilon_i\\ x_i\varepsilon_i\\ \varphi\left(x_i,z_i,\alpha_*\right)\varepsilon_i\end{array}\right]+o_p\left(1\right), \tag{1}$$

where

$$G_\gamma=-\mathbb{E}\left[\begin{array}{ccc}1 & x & \varphi\left(x,z,\alpha_*\right)\\ x & x^2 & x\varphi\left(x,z,\alpha_*\right)\\ \varphi\left(x,z,\alpha_*\right) & x\varphi\left(x,z,\alpha_*\right) & \varphi\left(x,z,\alpha_*\right)^2\end{array}\right].$$

The expansion (1) can be given an intuitive interpretation by considering an infeasible estimator. Assume that α_* is known to the econometrician, and $v_i=\varphi\left(x_i,z_i,\alpha_*\right)$ is used in the regression. Let $\widetilde{\gamma}$ denote the resulting OLS estimator of γ_*. The first order asymptotic properties of $\widetilde{\beta}=\frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{\gamma}_1+\widetilde{\gamma}_2x_i+\widetilde{\gamma}_3\varphi\left(x_i,z_i,\alpha_*\right)\right)$ can be analyzed using the expansion

$$\begin{aligned}\sqrt{n}\left(\widetilde{\beta}-\beta_*\right) =& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\gamma_{1*}+\gamma_{2*}x_i+\gamma_{3*}\varphi\left(x_i,z_i,\alpha_*\right)-\beta_*\right)\\ &+\left[\begin{array}{ccc}1 & \mathbb{E}\left[x\right] & \mathbb{E}\left[\varphi\left(x,z,\alpha_*\right)\right]\end{array}\right]\sqrt{n}\left(\widetilde{\gamma}-\gamma_*\right)+o_p\left(1\right)\end{aligned}$$

A routine calculation (presented in Appendix A) also establishes that

$$\begin{aligned}&\left[\begin{array}{ccc}1 & \mathbb{E}\left[x\right] & \mathbb{E}\left[\varphi\left(x,z,\alpha_*\right)\right]\end{array}\right]\sqrt{n}\left(\widetilde{\gamma}-\gamma_*\right)\\ &=-\left[\begin{array}{ccc}1 & \mathbb{E}\left[x\right] & \mathbb{E}\left[\varphi\left(x,z,\alpha_*\right)\right]\end{array}\right]G_\gamma^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\begin{array}{c}\varepsilon_i\\ x_i\varepsilon_i\\ \varphi\left(x_i,z_i,\alpha_*\right)\varepsilon_i\end{array}\right]+o_p\left(1\right)\end{aligned} \tag{2}$$

Comparing the correction terms (1) and (2) leads us to an interesting conclusion: The influence function for $\widehat{\beta}$ is equal to that of the unfeasible estimator $\widetilde{\beta}$ that ignores the estimation error in the first step, i.e., that in $\widehat{\alpha}$!

In order to understand this apparent puzzle, it is convenient to define $\widehat{\gamma}\left(\alpha\right)=\left(\widehat{\gamma}_1\left(\alpha\right),\widehat{\gamma}_2\left(\alpha\right),\widehat{\gamma}_3\left(\alpha\right)\right)$ as the OLS estimator with y as the dependent and x and $v=\varphi(x,z,\alpha)$ as the independent variables. Note that $\widehat{\gamma}=\widehat{\gamma}\left(\widehat{\alpha}\right)$ and $\widetilde{\gamma}=\widehat{\gamma}\left(\alpha_*\right)$. Also $\gamma(\alpha)$ is the vector of coefficients of the linear projection of y on $1,x,\varphi(x,z,\alpha)$. A naïve derivation of the influence function of $\widehat{\beta}$ would use the following decomposition

1. Main term that reflects the uncertainty left if we know γ_* and α_*:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\gamma_{1*}+\gamma_{2*}x_i+\gamma_{3*}\varphi\left(x_i,z_i,\alpha_*\right)-\beta_*\right)$$

2. A term that accounts for the sampling variation in $\widehat{\gamma}\left(\alpha_*\right)$ if we know α_*:

$$-\left[\begin{array}{ccc}1 & \mathbb{E}\left[x\right] & \mathbb{E}\left[\varphi\left(x,z,\alpha_*\right)\right]\end{array}\right]G_\gamma^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\begin{array}{c}\varepsilon_i\\ x_i\varepsilon_i\\ \varphi\left(x_i,z_i,\alpha_*\right)\varepsilon_i\end{array}\right]$$

3. A term that accounts for the sampling variation in $\widehat{\alpha}$:

$$\mathbb{E}\left[\gamma_{3*}\frac{\partial\varphi\left(x,z,\alpha_*\right)}{\partial\alpha}\right]\sqrt{n}\left(\widehat{\alpha}-\alpha_*\right)$$

This naïve decomposition is missing one additional term,[1] i.e.,

$$-\left[\begin{array}{ccc}1 & \mathbb{E}\left[x\right] & \mathbb{E}\left[\varphi\left(x,z,\alpha_*\right)\right]\end{array}\right]G_\gamma^{-1}G_\alpha\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi\left(x_i,z_i\right) \tag{3}$$

where

$$G_\alpha=\mathbb{E}\left[\begin{array}{c}-\gamma_{3*}\frac{\partial\varphi(x,z,\alpha_*)}{\partial\alpha}\\ -\gamma_{3*}x_i\frac{\partial\varphi(x,z,\alpha_*)}{\partial\alpha}\\ -2\gamma_{3*}\varphi\left(x,z,\alpha_*\right)\frac{\partial\varphi(x,z,\alpha_*)}{\partial\alpha}\end{array}\right]$$

As shown in Appendix A, $-G_\gamma^{-1}G_\alpha\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi\left(x_i,z_i\right)$ is the effect of the sampling variation in $\widehat{\alpha}$ on the sampling distribution of $\widehat{\gamma}$. Defining $\Psi\left(\alpha\right)=\mathbb{E}\left[\gamma_1\left(\alpha\right)+\gamma_2\left(\alpha\right)x+\gamma_3\left(\alpha\right)\varphi\left(x,z,\alpha_*\right)\right]$, we show in Appendix B that the missing term is asymptotically equivalent to $\sqrt{n}\left(\Psi\left(\widehat{\alpha}\right)-\Psi\left(\alpha_*\right)\right)$. The expression $\gamma_1\left(\alpha\right)+\gamma_2\left(\alpha\right)x+\gamma_3\left(\alpha\right)\varphi\left(x,z,\alpha_*\right)$ that appears in the definition of $\Psi\left(\alpha\right)$ can be given an interesting interpretation. It is the linear projection of y on $1,x,\varphi\left(x,z,\alpha\right)$ when after projection we substitute $\varphi\left(x,z,\alpha_*\right)$ for $\varphi\left(x,z,\alpha\right)$. Note that the linear projection of y on $1,x,\varphi\left(x,z,\alpha\right)$ has coefficients $\gamma(\alpha)$. This specifies a function of $x,\varphi\left(x,z,\alpha\right)$ that can be evaluated at any value of these arguments and here we choose the values $x,\varphi(x,z,\alpha_*)$. Hence, α plays two roles. First, it determines the functional form of the projection, here only the coefficients $\gamma(\alpha)$, because the projection is restricted to be linear. Second, α enters in the variables at which the (linear) projection is evaluated, here $x,\varphi\left(x,z,\alpha_*\right)$. If we substitute the estimator $\widehat{\alpha}$ then the two correction terms that account for the estimation error in $\widehat{\alpha}$ correspond to these two roles of α and in this example these two correction terms are opposites so that their sum is 0. The naïve derivation of the influence function ignores the effect of α on the coefficients of the linear projection.

In this paper we propose a method that accounts for the full contribution of $\widehat{\alpha}$ to the influence function, i.e., we improve on step 3 above. The full (accounting for the two distinct

[1] See Appendix A.

roles of α) contribution of the sampling variation of $\widehat{\alpha}$, i.e., with the projection coefficients equal to $\gamma_1(\widehat{\alpha}), \gamma_2(\widehat{\alpha}), \gamma_3(\widehat{\alpha})$, is

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\gamma_1(\widehat{\alpha})+\gamma_2(\widehat{\alpha})x_i+\gamma_3(\widehat{\alpha})\varphi\left(x_i,z_i,\widehat{\alpha}\right)\right. \\
&\left.-\gamma_{1*}-\gamma_{2*}x_i-\gamma_{3*}\varphi\left(x_i,z_i,\alpha_*\right)\right) \\
&=\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\alpha}(\gamma_1(\ \alpha)+\gamma_2(\alpha)x_i+\gamma_3(\alpha)\varphi\left(x_i,z_i,\alpha\right))\Big|_{\alpha=\alpha_*}\sqrt{n}(\widehat{\alpha}-\alpha_*)+o_p(1) \\
&=\frac{\partial}{\partial\alpha}\mathbb{E}[\gamma_1(\alpha)+\gamma_2(\alpha)x_i+\gamma_3(\alpha)\varphi\left(x_i,z_i,\alpha\right)]\Big|_{\alpha=\alpha_*}\sqrt{n}(\widehat{\alpha}-\alpha_*)+o_p(1)
\end{aligned}$$

Now the projection of y on $1, x, \varphi(x,z,\alpha)$ implies that for all constants s_1, s_2, s_3 and for all α

$$0=\mathbb{E}\left[\left(s_1\cdot 1+s_2\cdot x+s_3\cdot\varphi\left(x,z,\alpha\right)\right)\left(y-\gamma_1\left(\alpha\right)-\gamma_2\left(\alpha\right)x-\gamma_3\left(\alpha\right)\varphi\left(x,z,\alpha\right)\right)\right]$$

Taking $s_1=1$, $s_2=0$, and $s_3=0$, and differentiating the first equation with respect to α and evaluating the derivative at $\alpha=\alpha_*$, we obtain

$$\frac{\partial}{\partial\alpha}\mathbb{E}\left[\gamma_1\left(\alpha_*\right)+\gamma_2\left(\alpha_*\right)x+\gamma_3\left(\alpha_*\right)\varphi\left(x,z,\alpha_*\right)\right]=0$$

Therefore we conclude that the contribution of the sampling variation in $\widehat{\alpha}$ to the sampling variation of $\widehat{\beta}$ is 0. This derivation is simpler than that in Appendix A and can be generalized to the case of general projections that are not restricted to be linear.

In general the first step estimate plays these two distinct roles. The example in this section was relatively simple because the linear functional relation can be summarized by a finite dimensional vector $\gamma\left(\alpha\right)$. The challenge to the econometrician is that when the projection is non-parametric, as is the case when the generated regressor is used in a non-parametric regression, such simplicity disappears. By separately considering the two roles that sampling variation in the first step plays when we evaluate its effect on the second stage projection, we can properly adjust the influence function. In general the two corresponding correction terms are not opposite as in the simple example considered here.

3 The Influence Function of Semi-parametric Three-Step Estimators

We now present our two main results on semi-parametric three-step estimators. In the first step we estimate a regressor. In the second step we estimate a non-parametric regression with the generated regressor as one of the independent variables. In the third step we estimate a finite dimensional parameter (without loss of generality we consider the scalar case) that satisfies a moment condition that also depends on the non-parametric regression estimated in the second step. We distinguish between two cases. The first result concerns the case where in the first step the regressor is estimated by a parametric method. The second result concerns the case where in the first step the regressor is estimated by a non-parametric method. As was emphasized in the introduction, our characterization is based on Newey's (1994) path-derivative method.

3.1 Parametric First Step, Non-parametric Second Step

We assume that we observe i.i.d. observations $w_i = (y_i, x_i, z_i), i = 1, \ldots, n$. The first step is identical to that in Section 2, i.e., we have an estimator $\widehat{\alpha}$ such that $\sqrt{n}(\widehat{\alpha} - \alpha_*) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(x_i, z_i) + o_p(1)$ with $\mathbb{E}[\psi(x_i, z_i)] = 0$. The parameter vector α indexes a relation between a dependent variable that is a component of x (and that we later denote by u) and independent variables that are some or all of the other variables in x and those in z. Either the predicted value (Sections 4 and 5) or the residual (Section 6) of this relationship is an independent variable in the second step non-parametric regression. The notation $\varphi(x, z, \alpha)$ covers both cases. If φ is a residual then both x and φ can enter in the second step non-parametric regression. The second step is different from the parametric example, because our goal is to estimate

$$\mu(x, v_*) = \mathbb{E}[y \mid x, v_*]$$

where $v_* = \varphi(x, z, \alpha_*)$, i.e., we no longer restrict the projection to be linear. Because we do not observe α_*, we use $\widehat{v}_i = \varphi(x_i, z_i, , \widehat{\alpha})$ in the non-parametric regression. Our goal is to characterize the first order asymptotic properties of

$$\widehat{\beta} = \frac{1}{n} \sum_{i=1}^{n} h(\widehat{\gamma}(x_i, \varphi(x_i, z_i, \widehat{\alpha})))$$

with $\widehat{\gamma}$ the non-parametric regression of y on x and $\widehat{v}$. We can consider $\widehat{\beta}$ as the solution of a sample moment equation that is derived from a population moment equation that depends on β and $\mu(x, \varphi(x, z, \alpha_*))$. As will be seen below it matters whether h is linear (as in Section 2) or not.

Using Newey's (1994) path-derivative approach, we express the influence function of $\widehat{\beta}$ as a sum of three terms: (i) the main term

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (h(\mu(x_i, \varphi(x_i, z_i, \alpha_*))) - \beta_*)$$

(ii) a term that adjusts for the estimation of $\widehat{\gamma}$, i.e.,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (h(\widehat{\gamma}(x_i, \varphi(x_i, z_i, \alpha_*))) - h(\mu(x_i, \varphi(x_i, z_i, \alpha_*))))$$

and (iii) an adjustment related to the estimation of $\widehat{\alpha}$, i.e.,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (h(\gamma(x_i, \varphi(x_i, z_i, \widehat{\alpha}))) - h(\mu(x_i, \varphi(x_i, z_i, \alpha_*)))) \quad .$$

The decomposition here is based on the fact that Newey's approach can be used "term-by-term". Therefore, we may without loss of generality assume that α is a scalar.[2]

[2]The fact that Newey's approach can be used "term-by-term" is illustrated in an earlier version of the paper, which is available upon request. There, we consider the case where the moment function includes multiple non-parametric objects, all of which are obtained by non-parametric regressions with possibly different independent variables.

The second component in the decomposition can be easily analyzed as in Newey (1994, pp. 1360 – 61). It is equal to

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbb{E}\left[\left.\frac{\partial h\left(\mu\left(x_i,v_{*i}\right)\right)}{\partial\mu}\right|x_i,v_{*i}\right]\left(y_i-\mu\left(x_i,v_{*i}\right)\right)+o_p\left(1\right)$$

$$=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial h\left(\mu\left(x_i,v_{*i}\right)\right)}{\partial\mu}\left(y_i-\mu\left(x_i,v_{*i}\right)\right)+o_p\left(1\right)$$

As in Section 2 we therefore focus on the analysis of the third component

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(h\left(\gamma\left(x_i,\varphi\left(x_i,z_i,\widehat{\alpha}\right)\right)\right)-h\left(\mu\left(x_i,\varphi\left(x_i,z_i,\alpha_*\right)\right)\right)\right)$$

We define

$$\begin{aligned}\gamma\left(x,v^*;\alpha\right)&=\mathbb{E}\left[y\mid\ x,\varphi\left(x,z,\alpha\right)=v^*\right]\\ g\left(w,\alpha_1,\alpha_2,\gamma\right)&=h\left(\gamma\left(x,\varphi\left(x,z,\alpha_1\right);\alpha_2\right)\right)\end{aligned}$$

Note that the two roles that α plays are made explicit in $g\left(w,\alpha_1,\alpha_2,\gamma\right)$ that is obtained by substituting $v^*=\varphi(x,z,\alpha_1)$ in $\gamma(x,v^*;\alpha_2)$. Note also that $\mu\left(x,v_*\right)=\gamma\left(x,v_*;\alpha_*\right)$. The notation α_1,α_2 is just an expositional device, since $\alpha_1=\alpha_2=\alpha$.

With these definitions, we can now write

$$\frac{1}{n}\sum_{i=1}^{n}h\left(\gamma\left(x_i,\varphi\left(x_i,z_i,\widehat{\alpha}\right);\widehat{\alpha}\right)\right)=\frac{1}{n}\sum_{i=1}^{n}g\left(x_i,z_i,\widehat{\alpha}_1,\widehat{\alpha}_2,\hat{\gamma}\right)$$

where $\widehat{\alpha}_1=\widehat{\alpha}_2=\widehat{\alpha}$, but we keep them separate to emphasize the two roles of $\widehat{\alpha}$. This is helpful in order to deal with the two roles that $\widehat{\alpha}$ plays in the expansion by linearization, an expansion that amounts to taking partial derivatives:

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(h\left(\gamma\left(x_i,\varphi\left(x_i,z_i,\widehat{\alpha}\right);\widehat{\alpha}\right)\right)-h\left(\gamma\left(x_i,\varphi\left(x_i,z_i,\alpha_*\right);\alpha_*\right)\right)\right)\\&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(g\left(w_i,\widehat{\alpha}_1,\widehat{\alpha}_2,\hat{\gamma}\right)-g\left(w_i,\alpha_*,\alpha_*,\gamma_*\right)\right)\\&=\left(\mathbb{E}\left[\frac{\partial g\left(w,\alpha_*,\alpha_*,\gamma_*\right)}{\partial\alpha_1}\right]+\mathbb{E}\left[\frac{\partial g\left(w,\alpha_*,\alpha_*,\gamma_*\right)}{\partial\alpha_2}\right]\right)\sqrt{n}(\widehat{\alpha}-\alpha_*)+o_p\left(1\right)\end{aligned}$$

Therefore we must compute $\mathbb{E}\left[\frac{\partial g(w,\alpha_*,\alpha_*,\gamma_*)}{\partial\alpha_1}\right]$ and $\mathbb{E}\left[\frac{\partial g(w,\alpha_*,\alpha_*,\gamma_*)}{\partial\alpha_2}\right]$. The computation of the first expectation is easy. Because $\gamma\left(x,\varphi\left(x,z,\alpha\right);\alpha_*\right)=\mu\left(x,\varphi\left(x,z,\alpha\right)\right)$, we have

$$\mathbb{E}\left[\frac{\partial g\left(w,\alpha_*,\alpha_*,\gamma_*\right)}{\partial\alpha_1}\right]=\mathbb{E}\left[\frac{\partial h\left(\mu\left(x,\varphi\left(x,z,\alpha_*\right)\right)\right)}{\partial\mu}\frac{\partial\mu\left(x,\varphi\left(x,z,\alpha_*\right)\right)}{\partial v}\frac{\partial\varphi\left(x,z,\alpha_*\right)}{\partial\alpha}\right]$$

The headache is to compute the second expectation. By the chain rule

$$\mathbb{E}\left[\frac{\partial g\left(w,\alpha_{*},\alpha_{*},\gamma_{*}\right)}{\partial\alpha_{2}}\right]=\mathbb{E}\left[\frac{\partial h\left(\mu\left(x,\varphi\left(x,z,\alpha_{*}\right)\right)\right)}{\partial\mu}\frac{\partial\gamma\left(x,\varphi\left(x,z,\alpha_{*}\right);\alpha_{*}\right)}{\partial\alpha}\right] \tag{4}$$

Unfortunately, it is not obvious how to differentiate $\gamma\left(x,\varphi\left(x,z,\alpha_{*}\right);\alpha\right)$ with respect to α. After all, $\gamma\left(x,\varphi\left(w,\alpha_{*}\right);\alpha\right)$ has the functional form of $\mathbb{E}\left[y\mid\ x,\varphi\left(x,z,\alpha\right)=v^{*}\right]$ that depends on α. The next theorem gives the solution.

Theorem 1 (Contribution parametric first-stage estimator) *The adjustment to the influence function that accounts for the first-stage estimation error is*

$$\left(\mathbb{E}\left[\frac{\partial g\left(w,\alpha_{*},\alpha_{*},\gamma_{*}\right)}{\partial\alpha_{1}}\right]+\mathbb{E}\left[\frac{\partial g\left(w,\alpha_{*},\alpha_{*},\gamma_{*}\right)}{\partial\alpha_{2}}\right]\right)\sqrt{n}(\widehat{\alpha}-\alpha_{*}) \tag{5}$$

$$=\mathbb{E}\left[\frac{\partial^{2}h\left(\mu\left(x,\varphi\left(x,z,\alpha_{*}\right)\right)\right)}{\partial\mu^{2}}\left(y-\mu\left(x,\varphi\left(x,z,\alpha_{*}\right)\right)\right)\frac{\partial\mu\left(x,\varphi\left(x,z,\alpha_{*}\right)\right)}{\partial v}\frac{\partial\varphi\left(x,z,\alpha_{*}\right)}{\partial\alpha}\right]\sqrt{n}\left(\widehat{\alpha}-\alpha_{*}\right).$$

Proof See Appendix C.

Note that the form of the adjustment term implies that if h is linear, then the first-stage estimation error has no effect on the variance of the estimator of β. This was illustrated for the fully parametric case in Section 2.

3.2 Multivariate Generalization

Suppose now that the μ is multidimensional, i.e., y is a J-dimensional random vector. More specifically, suppose now that we have

$$\gamma_{j}\left(x,v^{*};\alpha\right)=\mathbb{E}\left[y_{j}\mid\ x,\varphi\left(x,z,\alpha\right)=v^{*}\right]$$

and

$$\widehat{\beta}=\frac{1}{n}\sum_{i=1}^{n}h\left(\gamma_{1}\left(x_{i},\varphi\left(x_{i},z_{i},\widehat{\alpha}\right)\right),\ldots,\gamma_{J}\left(x_{i},\varphi\left(x_{i},z_{i},\widehat{\alpha}\right)\right)\right)$$

The product rule of calculus suggests that we can tackle this problem by adding the derivatives. This is formalized in the next theorem.

Theorem 2 (Contribution parametric first-stage estimators) *The adjustment to the influence function that accounts for the first-stage estimation error is*

$$\sum_{j}\mathbb{E}\left[\frac{\partial^{2}h\left(\mu\left(x,\varphi\left(x,z,\alpha_{*}\right)\right)\right)}{\partial\mu_{j}^{2}}\left(y_{j}-\mu_{j}\left(x,\varphi\left(x,z,\alpha_{*}\right)\right)\right)\frac{\partial\mu_{j}\left(x,\varphi\left(x,z,\alpha_{*}\right)\right)}{\partial v'}\frac{\partial\varphi\left(x,z,\alpha_{*}\right)}{\partial\alpha}\right]\sqrt{n}\left(\widehat{\alpha}-\alpha_{*}\right).$$

Proof See Appendix C.

3.3 Non-parametric First Step, Non-parametric Second Step

We now assume that the first step is non-parametric. Again we have a random sample $w_i = (y_i, x_i, z_i), i = 1, \ldots, n$. The first-step projection of one of the components of x, that we denote by u, on some or all of the other components of x and z is denoted by $v_* = \varphi_*(x, z) = \mathbb{E}[u \mid x, z]$. The first step is to estimate this projection by non-parametric regression. In the second step we estimate $\gamma(x, v_*) = \mathbb{E}[y \mid x, v_*]$ by non-parametric regression of y on $x, \widehat{v} = \widehat{\varphi}(x, z)$. Our interest is to characterize the first order asymptotic properties of

$$\frac{1}{n}\sum_{i=1}^{n} h\left(\widehat{\gamma}\left(x_i, \widehat{\varphi}\left(x_i, z_i\right)\right)\right)$$

We define

$$\begin{aligned} \mu(x, v_*) &= \mathbb{E}[y \mid x, \varphi_*(x, z) = v_*] \\ \gamma(x, v^*; v) &= \mathbb{E}[y \mid x, \varphi(x, z) = v^*] \\ g(w, v_1, v_2, \gamma) &= h(\gamma(x, v_1; v_2)) \end{aligned}$$

with $v = \varphi(x, z)$ and with v_1 and v_2 playing the roles of α_1 and α_2.

With these definitions, we can now write

$$\frac{1}{n}\sum_{i=1}^{n} h\left(\widehat{\gamma}\left(x_i, \widehat{v}_1; \widehat{v}_2\right)\right) = \frac{1}{n}\sum_{i=1}^{n} g\left(w_i, \widehat{v}_1, \widehat{v}_2, \widehat{\gamma}\right)$$

where $\widehat{v}_1 = \widehat{v}_2 = \widehat{v}$. We keep them separate to emphasize their different roles. Our objective is to approximate

$$\frac{1}{n}\sum_{i=1}^{n} g\left(w_i, \widehat{v}_1, \widehat{v}_2, \widehat{\gamma}\right) - \frac{1}{n}\sum_{i=1}^{n} g\left(w_i, v_1, v_2, \gamma\right)$$

To find the contribution of the sampling variation in $\hat{v}$ we can take γ as known. As in Newey (1994) we consider a path v_α indexed by $\alpha \in \mathbb{R}$ such that $v_{\alpha_*} = v_*$. First, using the calculation in the previous section,

$$\begin{aligned} &\mathbb{E}\left[\frac{\partial}{\partial \alpha_1} g\left(w, \alpha_*, \alpha_*, \gamma_*\right)\right] + \mathbb{E}\left[\frac{\partial}{\partial \alpha_2} g\left(w, \alpha_*, \alpha_*, \gamma_*\right)\right] \\ &= \frac{\partial}{\partial \alpha}\mathbb{E}\left[\frac{\partial^2 h\left(\mu\left(x, v_*\right)\right)}{\partial \mu^2}\left(y - \mu\left(x, v_*\right)\right)\frac{\partial \mu\left(x, v_*\right)}{\partial v} v_\alpha\right] \end{aligned}$$

we obtain that

$$\left.\frac{\partial \mathbb{E}\left[h\left(\gamma\left(x, v_\alpha; v_\alpha\right)\right)\right]}{\partial \alpha}\right|_{\alpha=\alpha_*} = \frac{\partial \mathbb{E}\left[D\left(w, v_\alpha\right)\right]}{\partial \alpha}$$

for

$$D\left(w, v_\alpha\right) = \frac{\partial^2 h\left(\mu\left(x, v_*\right)\right)}{\partial \mu^2}\left(y - \mu\left(x, v_*\right)\right)\frac{\partial \mu\left(x, v_*\right)}{\partial v} v_\alpha.$$

which is linear in v_α. Second, for[3]

$$\delta_1(x,z) = \mathbb{E}\left[\frac{\partial^2 h(\mu(x,\varphi_*(x,z)))}{\partial \mu^2}(y-\mu(x,\varphi_*(x,z)))\frac{\partial \mu(x,\varphi_*(x,z))}{\partial v}\middle| x,z\right].$$

we have that for any $v=\varphi(x,z)$

$$\mathbb{E}[D(w,v)] = \mathbb{E}[\delta_1(x,z)\varphi(x,z)]$$

By Newey (1994) Proposition 4 these two facts imply that the adjustment to the influence function is equal to

$$\delta_1(x_i,z_i)(u_i - \mathbb{E}[u \mid x_i,z_i]) = \delta_1(x_i,z_i)(u_i - \varphi_*(x_i,z_i))$$

with u the component of of x that is projected on x,z.

We summarize the result in a theorem:

Theorem 3 (Contribution non-parametric first-stage estimator) *The adjustment to the influence function that accounts for the first-stage estimation error is*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\delta_1(x_i,z_i)(u_i-\varphi_*(x_i,z_i))$$

with $\varphi_*(x,z)=\mathbb{E}[u|x,z]$ *and*

$$\delta_1(x,z) = \mathbb{E}\left[\frac{\partial^2 h(\mu(x,\varphi_*(x,z)))}{\partial \mu^2}(y-\mu(x,\varphi_*(x,z)))\frac{\partial \mu(x,\varphi_*(x,z))}{\partial v}\middle| x,z\right]$$

Finally we consider the adjustment for the estimation of γ. This is essentially the adjustment to the influence function for

$$\frac{1}{n}\sum_{i=1}^{n} h(\widehat{\gamma}(x_i,v_{*i}))$$

By Newey (1994, pp. 1360 – 61), we conclude that the corresponding adjustment to the influence function is equal to

$$\delta_2(x_i,v_{*i})(y_i - \mathbb{E}[y \mid x_i,v_{*i}])$$

where

$$\delta_2(x,v_*) = \mathbb{E}\left[\frac{\partial h(\mu(x,v_*))}{\partial \mu}\middle| x,v_*\right] = \frac{\partial h(\mu(x,v_*))}{\partial \mu}$$

[3]If $\varphi(x_1,z)$ depends on a subvector of the variables x that enter in μ, then we average over the remaining variables in x.

3.4 Extension

So far, we have assumed that the parameter of interest is

$$\beta_* = \mathbb{E}[h(\mu(x, v_*))]$$

where h depends only on μ. We now consider the extension to

$$\beta_* = \mathbb{E}[h(w, \mu(x, v_*))]$$

where w is a vector of other variables that may have x, z as subvectors. We consider both the case that φ is parametric and the case that this function is non-parametric. Because as before the main term and the contribution of the estimation of $\mathbb{E}(y|x, v_*)$ do not raise new issues, the next two theorems only give the contribution of the first-stage estimator. In these theorems we use the function

$$\kappa(x, v) = \mathbb{E}\left[\frac{\partial h(w, \mu(x, v_*))}{\partial \mu}\middle| x, \varphi(x, z, \alpha_*) = v\right] \tag{6}$$

with an obvious adjustment for the non-parametric case.

Theorem 4 (Contribution parametric first-stage estimator) *The adjustment to the influence function that accounts for the first-stage estimation error is*

$$\left(\mathbb{E}\left[\left(\frac{\partial h(w, \mu(x, \varphi(x, z, \alpha_*)))}{\partial \mu} - \kappa(x, \varphi(x, z, \alpha_*))\right)\frac{\partial \mu(x, \varphi(x, z, \alpha_*))}{\partial v}\frac{\partial \varphi(x, z, \alpha_*)}{\partial \alpha}\right] + \mathbb{E}\left[\frac{\partial \kappa(x, \varphi(x, z, \alpha_*))}{\partial v}(y - \gamma(x, \varphi(x, z, \alpha_*); \alpha_*))\frac{\partial \varphi(x, z, \alpha_*)}{\partial \alpha}\right]\right)\sqrt{n}(\widehat{\alpha} - \alpha_*)$$

Now, we consider the case where the first step is non-parametric. The discussion preceding Theorem 3, which summarizes Newey's argument, implies that

Theorem 5 (Contribution non-parametric first-stage estimator) *The adjustment to the influence function that accounts for the first-stage estimation error is*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\delta_3(x_i, z_i)(u_i - \varphi_*(x_i, z_i))$$

with $\varphi_*(x, z) = \mathbb{E}[u|x, z]$ *and*

$$\delta_3(x, z) = \mathbb{E}\left[\left(\frac{\partial h(w, \mu(x, \varphi_*(x, z)))}{\partial \mu} - \kappa(x, \varphi_*(x, z))\right)\frac{\partial \mu(x, \varphi_*(x, z))}{\partial v}\middle| x, z\right] + \mathbb{E}\left[\frac{\partial \kappa(x, \varphi_*(x, z))}{\partial v}(y - \gamma(x, \varphi_*(x, z)))\middle| x, z\right]$$

Suppose that $\kappa(x, \varphi_*(x, z)) = 0$ in Theorem 4. The adjustment is then equal to the derivative with respect to α_1, i.e., the naive derivative (see equation (21) in the proof of Theorem 4). Therefore, it may be useful to check whether $\kappa(x, \varphi_*(x, z)) = 0$ in specific models. If it is the case, we need not worry about the effect of first-step estimation on the second-stage

non-parametric regression. Note also that the effect of the first-stage estimation now consists of two terms, the first of which is 0 in Theorem 1 and 3.

It is also useful to point out the theorems can be applied to general semi-parametric GMM estimators. If we consider the moment condition

$$\mathbb{E}[m(w, \mu(x, v_*), \beta_*)] = 0$$

and we linearize the corresponding sample moment condition we obtain

$$\sqrt{n}(\hat{\beta} - \beta_*) = \left(\mathbb{E}\left[\frac{\partial m(w, \mu(x, v_*), \beta_*)}{\partial \beta'} \right] \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} m\left(w_i, \hat{\gamma}(x_i, \hat{\varphi}(x_i, z_i)), \beta_*\right) + o_p(1)$$

Therefore, the contribution of the first-stage estimate to the asymptotic distribution of $\hat{\beta}$ can be found by applying Theorem 5 to $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} m\left(w_i, \hat{\gamma}(x_i, \hat{\varphi}(x_i, z_i)), \beta_*\right)$.

3.5 Discussion

The effect of the first-stage estimation error is qualitatively different for three-stage and two-stage semi-parametric estimators. To show this we contrast our results with two results available in the literature. First, consider the standard two-stage estimator (with a non-parametric first stage) of the form

$$\widehat{\beta} = \frac{1}{n} \sum_{i=1}^{n} h\left(x_i, \widehat{\varphi}\left(x_i, z_i\right)\right)$$

where $\widehat{\varphi}$ is an estimator of $\varphi\left(x, z\right) = \mathbb{E}\left[u \middle| x, z\right]$. As discussed in Newey (1994), among others, the contribution of the estimation of φ to the influence function is $\frac{\partial h(x, \varphi(x,z))}{\partial v}\left(u - \varphi\left(x, z\right)\right)$. This involves the first derivative of h, so that this contribution is nonzero if h is linear. This in contrast to the three-stage estimator, in which case the contribution is zero with h linear.

Second, we can compare our results with those on the asymptotic distribution of the non-parametric regression estimator $\widehat{\gamma}\left(x, \varphi\left(x, z, \widehat{\alpha}\right)\right)$ following a first-step parametric estimation. Because the $\widehat{\alpha}$ typically converges at the parametric rate, the asymptotic distribution of $\widehat{\gamma}\left(x, \varphi\left(x, z, \widehat{\alpha}\right)\right)$ for all x, z is unaffected by the first-step estimation error. If we would take this result to the third-step estimation of β_* by

$$\widehat{\beta} = \frac{1}{n} \sum_{i=1}^{n} h\left(\widehat{\gamma}\left(x_i, \varphi\left(x_i, z_i, \widehat{\alpha}\right)\right)\right)$$

we would incorrectly conclude that the first-step estimation of $\widehat{\alpha}$ does not affect the third-step estimator whether h is linear or not. This example makes it clear that our results cannot be derived from the results in, e.g., Song (2008) or Sperlich (2009) for the non-parametric regression on generated regressors estimated in the first step.

4 The Partial Linear Model with a Generated Regressor

In this section, we apply the results in the previous section to a semi-parametric model, the partial linear regression model,

$$y_i = x_i\beta_* + m\left(\upsilon\left(w_i, \alpha_*\right)\right) + \varepsilon_i,$$

where x_i is a component of w_i, and m is non-parametric. The error term ε_i satisfies $\mathbb{E}\left[\varepsilon_i \middle| x_i, \upsilon\left(w_i, \alpha_*\right)\right] = 0$. The parameter of interest is β_*. We initially consider the case that the generated regressor is estimated parametrically, but we also give the contribution to the influence function for the case that it is estimated non-parametrically.

The model can be estimated by regressing $y_i - \mathbb{E}\left[y_i \middle| \upsilon\left(w_i, \widehat{\alpha}\right)\right]$ on $x_i - \mathbb{E}\left[x_i \middle| \upsilon\left(w_i, \widehat{\alpha}\right)\right]$. By Newey (1994), Proposition 2 the estimation of the conditional expectation $\mathbb{E}\left[x_i \middle| \upsilon\left(w_i, \alpha_*\right)\right]$ has no contribution to the influence function of $\hat{\beta}$. By substitution we find that $\sqrt{n}\left(\widehat{\beta} - \beta_*\right)$ can be written as $\mathbb{E}\left[\left(x - \mathbb{E}\left[x \middle| \upsilon\left(w, \alpha_*\right)\right]\right)^2\right]^{-1}$ times

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(x_i - \mathbb{E}\left[x \middle| \upsilon\left(w_i, \widehat{\alpha}\right)\right]\right)\varepsilon_i \\
&+ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(x_i - \mathbb{E}\left[x \middle| \upsilon\left(w_i, \widehat{\alpha}\right)\right]\right)\left(m\left(\upsilon\left(w_i, \alpha_*\right)\right) - \mathbb{E}\left[m\left(\upsilon\left(w, \alpha_*\right)\right) \middle| \upsilon\left(w_i, \widehat{\alpha}\right)\right]\right) \\
&- \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(x_i - \mathbb{E}\left[x \middle| \upsilon\left(w_i, \widehat{\alpha}\right)\right]\right)\mathbb{E}\left[\varepsilon \middle| \upsilon\left(w_i, \widehat{\alpha}\right)\right] + o_p\left(1\right). \qquad (7)
\end{aligned}$$

To assess the contribution of the estimation error of $\widehat{\alpha}$ we linearize with respect to α. The coefficient in the linearization, i.e., that of $\sqrt{n}(\widehat{\alpha} - \alpha_*)$ is, using the notation $v_* = v(w, \alpha_*)$,

$$\begin{aligned}
&\frac{\partial}{\partial\alpha}\mathbb{E}\left[\left(x - \mathbb{E}\left[x \middle| \upsilon\left(w, \alpha\right)\right]\right)\varepsilon\right]\Big|_{\alpha=\alpha_*} \qquad (8)\\
&+ \frac{\partial}{\partial\alpha}\mathbb{E}\left[\left(x - \mathbb{E}\left[x \middle| \upsilon\left(w, \alpha\right)\right]\right)\left(m\left(\upsilon_*\right) - \mathbb{E}\left[m\left(\upsilon_*\right) \middle| \upsilon_*\right]\right)\right]\Big|_{\alpha=\alpha_*} \qquad (9)\\
&+ \frac{\partial}{\partial\alpha}\mathbb{E}\left[\left(x - \mathbb{E}\left[x \middle| \upsilon_*\right]\right)\left(m\left(\upsilon_*\right) - \mathbb{E}\left[m\left(\upsilon_*\right) \middle| \upsilon\left(w, \alpha\right)\right]\right)\right]\Big|_{\alpha=\alpha_*} \qquad (10)\\
&- \frac{\partial}{\partial\alpha}\mathbb{E}\left[\left(x - \mathbb{E}\left[x \middle| \upsilon\left(w, \alpha\right)\right]\right)\mathbb{E}\left[\varepsilon \middle| \upsilon_*\right]\right]\Big|_{\alpha=\alpha_*} \qquad (11)\\
&- \frac{\partial}{\partial\alpha}\mathbb{E}\left[\left(x - \mathbb{E}\left[x \middle| \upsilon_*\right]\right)\mathbb{E}\left[\varepsilon \middle| \upsilon\left(w, \alpha\right)\right]\right]\Big|_{\alpha=\alpha_*} \qquad (12)
\end{aligned}$$

Because $\mathbb{E}(\varepsilon|v_*) = 0$ and $\mathbb{E}(m(v_*|v_*)) = m(v_*)$, (9) and (11) are equal to 0. The other terms are analyzed using Theorem 4. For (8) we have for $\mu_1(v_*) = \mathbb{E}[x|v_*]$

$$h_1(x, \varepsilon, \mu_1) = (x - \mu_1(v_*))\varepsilon$$

so that

$$\frac{\partial h_1(x,\varepsilon,\mu_1)}{\partial \mu_1} = -\varepsilon$$

and $\kappa_1(v) = -\mathbb{E}(\varepsilon|v_* = v) = 0$ for all v with κ_1 as defined in (6). Therefore the coefficient of $\sqrt{n}(\widehat{\alpha} - \alpha_*)$ in the influence function is

$$-\mathbb{E}\left[\varepsilon \frac{\partial \mu_1(\upsilon_*)}{\partial \upsilon} \frac{\partial \upsilon(w,\alpha_*)}{\partial \alpha}\right]$$

For (10) we define

$$h_2(x,\mu_2) = (x - \mathbb{E}[x|\upsilon_*])(m(\upsilon_*) - \mu_2)$$

with $\mu_2(v) = \mathbb{E}[m(v_*)|v(w,\alpha) = v]$ so that

$$\frac{\partial h_2(x,\mu_2)}{\partial \mu_2} = -(x - \mathbb{E}[x|\upsilon_*])$$

and $\kappa_2(v) = 0$ for all v with κ_2 as defined in (6). Therefore by Theorem 4 the coefficient of $\sqrt{n}(\widehat{\alpha} - \alpha_*)$ in the influence function is

$$-\mathbb{E}\left[\eta \frac{\partial m(\upsilon_*)}{\partial \upsilon} \frac{\partial \upsilon(w,\alpha_*)}{\partial \alpha}\right]$$

for $\eta = x - \mathbb{E}[x|\upsilon_*]$, because $\mu_2(v)|_{\alpha=\alpha_*} = m(v)$.

Finally for (12) we define

$$h_3(x,\mu_3) = (x - \mathbb{E}[x|\upsilon_*])\mu_3$$

with $\mu_3(v) = \mathbb{E}[\varepsilon|\upsilon(w,\alpha)]$, so that

$$\frac{\partial h_3(x,\mu_3)}{\partial \mu_3} = x - \mathbb{E}[x|\upsilon_*]$$

and $\kappa_3(v) = 0$ for all v with κ_3 as defined in (6). Therefore by Theorem 4 the coefficient of $\sqrt{n}(\widehat{\alpha} - \alpha_*)$ in the influence function is

$$\mathbb{E}\left[(x - \mathbb{E}[x|\upsilon_*]) \frac{\partial \mu_3(\upsilon_*)}{\partial \upsilon}\bigg|_{\alpha=\alpha_*} \frac{\partial \upsilon(w,\alpha_*)}{\partial \alpha}\right] = 0$$

because $\mu_3(v)|_{\alpha=\alpha_*} = \mathbb{E}[\varepsilon|v^* = v] = 0$ for all v.

To conclude, the adjustment in the influence function of $\widehat{\beta}$ corresponding to the estimation error in $\widehat{\alpha}$ is

$$-\left(\mathbb{E}\left[\varepsilon \frac{\partial \mu_1(v(w,\alpha_*))}{\partial \upsilon} \frac{\partial \upsilon(w,\alpha_*)}{\partial \alpha}\right] + \mathbb{E}\left[\eta \frac{\partial m(v(w,\alpha_*))}{\partial \upsilon} \frac{\partial \upsilon(w,\alpha_*)}{\partial \alpha}\right]\right)\sqrt{n}(\widehat{\alpha} - \alpha_*)$$

Note that the $\mathbb{E}\left[\varepsilon \frac{\partial \mu_1(v(w,\alpha_*))}{\partial \upsilon} \frac{\partial \upsilon(w;\alpha_*)}{\partial \alpha}\right] = 0$ if we assume that $\mathbb{E}[\varepsilon|w] = 0$, and in that case our result is the same as in Newey (2009) or Li and Wooldridge (2002). Because $\kappa(v)$ as defined in Theorem 4 is 0 for all v, the effect of the estimation of $\hat{\alpha}$ on the conditional expectation is 0. In other words, the 'naive' linearization is valid.

Combining this result with Newey (1994) we find that the contribution in the case that $v(w)$ is estimated by non-parametric regression of u on w is equal to

$$-\left(\mathbb{E}[\varepsilon|w] \frac{\partial \mu_1(v(w,\alpha_*))}{\partial \upsilon} + \mathbb{E}[\eta|w] \frac{\partial m(v(w,\alpha_*))}{\partial \upsilon}\right)(u - \upsilon_*(w))$$

5 Regression on the Estimated Propensity Score

We consider an intervention with potential outcomes y_0, y_1 that are the control and treated outcome, respectively. The treatment indicator is d and $y = dy_1 + (1-d)y_0$ is the observed outcome. The vector x contains covariates that are not affected by the intervention. As shown by Rosenbaum and Rubin (1983) unconfounded assignment, i.e., the assumption that $y_1, y_0 \perp d|x$, implies $y_1, y_0 \perp d|\varphi(x)$ with $\varphi(x) = \Pr(d = 1|x)$ probability of selection or propensity score. As a consequence the ATE given x can be identified by $\mathbb{E}[y|d = 1, x] - \mathbb{E}[y|d = 0, x]$ or by $\mathbb{E}[y|d = 1, \varphi(x)] - \mathbb{E}[y|d = 0, \varphi(x)]$. These observations have led to a large number of estimators that can be classified into three groups. Most of these estimators rely on the propensity score, but some do not. The asymptotic variance of the estimators can be compared to the semi-parametric efficiency bound for the ATE derived by Hahn (1998).

The most popular estimators are the matching estimators that estimate the ATE given x or given $\varphi(x)$ by averaging outcomes over units with a 'similar' value of x or $\varphi(x)$ (and subsequently average over the distribution of x or $\varphi(x)$ to estimate the ATE). Abadie and Imbens (2009a), (2009b) are recent contributions. They show that matching estimators that have an asymptotic distribution that is notoriously difficult to analyze, are not asymptotically efficient. The second class of estimators do not estimate the ATE given x or $\varphi(x)$ but use the propensity scores as weights Hahn's (1998) estimator and the estimator of Hirano, Imbens and Ridder (2003) are examples of such estimators. These estimators are asymptotically efficient, which suggests that the propensity score is needed to achieve efficiency. The third class of estimators use non-parametric regression to estimate $\mathbb{E}[y|d = 1, x]$, $\mathbb{E}[y|d = 0, x]$ or $\mathbb{E}[y|d = 1, \varphi(x)]$, $\mathbb{E}[y|d = 0, \varphi(x)]$. Of these estimators the estimator based on $\mathbb{E}[y|d = 1, x]$, $\mathbb{E}[y|d = 0, x]$, the imputation estimator, is known to be asymptotically efficient, which suggests that there is no role for the propensity score. The missing result is that for the estimator that uses the non-parametric regression on a propensity score that is estimated in a preliminary step. This estimator that was suggested and analyzed by Heckman, Ichimura, and Todd (HIT) (1998) fits into our framework and is analyzed here.[4]

Our conclusion is that the HIT estimator has the same asymptotic variance as the imputation estimator, so that there is no efficiency gain in using the propensity score. This should settle the issue whether there is a role for the propensity score in achieving semi-parametric efficiency. That does not mean that there is no role for the propensity score in assessing the identification or in improved small sample performance of ATE estimators. Although the estimator based on regressions on the propensity score has the same structure as the general estimator discussed in Section 3, the results of that section have to be adapted, because the non-parametric regressions are for the treated and controls separately, i.e., for subpopulations.

5.1 Parametric First Step, Nonparametric Second Step

We have a random sample $w_i = (y_i, x_i, d_i), i = 1, \ldots, n$. The propensity score $\Pr(d = 1|x) = \varphi(x, \alpha)$ is parametric and its parameters α are estimated in the first step, by e.g. Maximum

[4]Heckman, Ichimura, and Todd actually consider an estimator of the Average Treatment Effect on the Treated (ATT) that we also analyze.

Likelihood are OLS (Linear Probability model) or any other method, such that

$$\sqrt{n}\left(\widehat{\alpha}-\alpha_*\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi\left(d_i,x_i\right)+o_p\left(1\right)$$

with $\mathbb{E}\left[\psi\left(d_i,x_i\right)\right]=0$. In the second step, we estimate

$$\mu\left(\varphi(x,\alpha_*)\right)=\left(\mathbb{E}\left[y\mid\varphi(x,\alpha_*),d=1\right],\mathbb{E}\left[y\mid\varphi(x,\alpha_*),d=0\right]\right)',$$

Because we do not observe α_*, we use $\varphi\left(x_i,\widehat{\alpha}\right)$ in the non-parametric regression.

Our interest is to characterize the first order asymptotic properties of

$$\widehat{\beta}=\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\gamma}_1\left(\varphi\left(x_i,\widehat{\alpha}\right)\right)-\widehat{\gamma}_2\left(\varphi\left(x_i,\widehat{\alpha}\right)\right)\right)$$

This estimator has the structure of that Section 3.2 with $h(\gamma)=\gamma_1-\gamma_2$, except for the fact that we do not regress y non-parametrically on $\varphi(x,\alpha_*)$ in the full population, but in the subpopulations of the treated and controls. This will require a modification of the proof of Theorem 3.

We define

$$\begin{aligned}\gamma\left(v;\alpha\right)&=\left(\mathbb{E}\left[y\mid\ d=1,\varphi\left(x,\alpha\right)=v\right],\mathbb{E}\left[y\mid\ d=0,\varphi\left(x,\alpha\right)=v\right]\right)'\\ g(w,\alpha_1,\alpha_2)&=h(\gamma(\varphi(x,\alpha_1);\alpha_2)\end{aligned}$$

The functions $\gamma\left(\varphi\left(x,\alpha\right);\alpha\right)$ solve the minimization problem

$$\min_{p_1,p_2}\mathbb{E}\left[d\left(y-p_1\left(\varphi\left(x,\alpha\right)\right)\right)^2+\left(1-d\right)\left(y-p_2\left(\varphi\left(x,\alpha\right)\right)\right)^2\right]$$

Note that this is equivalent to minimizing the first term with respect to p_1 and the second with respect to p_2. Therefore for all functions $\left(s_1\left(\varphi\left(x,\alpha\right)\right),s_2\left(\varphi\left(x,\alpha\right)\right)\right)'$

$$\mathbb{E}\left[d\left(y-\gamma_1\left(\varphi\left(x,\alpha\right);\alpha\right)\right)s_1\left(\varphi\left(x,\alpha\right)\right)\right]=0$$

$$\mathbb{E}\left[\left(1-d\right)\left(y-\gamma_2\left(\varphi\left(x,\alpha\right);\alpha\right)\right)s_2\left(\varphi\left(x,\alpha\right)\right)\right]=0$$

In particular, this should hold for

$$\begin{aligned}s_1\left(\varphi\left(x,\alpha\right)\right)&=\frac{1}{\varphi\left(x,\alpha\right)}\\ s_2\left(\varphi\left(x,\alpha\right)\right)&=\frac{1}{1-\varphi\left(x,\alpha\right)}\end{aligned}$$

These function s_1 is chosen in view of the fact that

$$\mathbb{E}\left[\frac{dy}{\varphi(x,\alpha)}\middle|\,\varphi(x,\alpha)\right]=\frac{\mathbb{E}[d|\varphi(x,\alpha)]\gamma_1(\varphi(x,\alpha);\alpha)}{\varphi(x,\alpha)}$$

i.e. the projection in the subpopulation is obtained by projecting the outcome in the subpopulation weighted by the probability of observation on $\varphi(x,\alpha)$. This gives γ_1 up to a correction factor that is equal to 1 if $\alpha = \alpha_*$. A similar observation can be made for s_2.

The orthogonality conditions yield the following two equations that hold for all α

$$\mathbb{E}\left[\frac{dy}{\varphi\left(x,\alpha\right)}\right] = \mathbb{E}\left[\frac{d\gamma_1\left(\varphi\left(x,\alpha\right);\alpha\right)}{\varphi\left(x,\alpha\right)}\right]$$

$$\mathbb{E}\left[\frac{\left(1-d\right)y}{1-\varphi\left(x,\alpha\right)}\right] = \mathbb{E}\left[\frac{\left(1-d\right)\gamma_2\left(\varphi\left(x,\alpha\right);\alpha\right)}{1-\varphi\left(x,\alpha\right)}\right]$$

with the left-hand sides equal to $\mathbb{E}\left[\frac{\varphi(x,\alpha_*)\mathbb{E}[y|x,d=1])}{\varphi(x,\alpha)}\right]$ and $\mathbb{E}\left[\frac{(1-\varphi(x,\alpha_*))\mathbb{E}[y|x,d=0]}{1-\varphi(x,\alpha)}\right]$, respectively. Differentiation with respect to α gives the derivatives of γ_1 and γ_2 with respect to the α in the conditioning variable, i.e., the parametric propensity score. Substitution gives the contribution of the estimation of $\hat{\alpha}$ to the influence function that is equal to (more details in Appendix D)

$$\mathbb{E}\left[\frac{\partial g\left(w,\alpha_*,\alpha_*,\gamma_*\right)}{\partial\alpha_1}\right] + \mathbb{E}\left[\frac{\partial g\left(w,\alpha_*,\alpha_*,\gamma_*\right)}{\partial\alpha_2}\right]\sqrt{n}(\hat{\alpha}-\alpha_*) =$$

$$-\mathbb{E}\left[\left(\frac{\mathbb{E}\left[y\right|x,d=1\right]-\mu_1\left(\varphi\left(x,\alpha_*\right)\right)}{\varphi\left(x,\alpha_*\right)} + \frac{\mathbb{E}\left[y\right|x,d=0\right]-\mu_2\left(\varphi\left(x,\alpha_*\right)\right)}{1-\varphi\left(x,\alpha_*\right)}\right)\frac{\partial\varphi\left(x,\alpha_*\right)}{\partial\alpha}\right]\sqrt{n}(\hat{\alpha}-\alpha_*)$$

The contribution of $\widehat{\gamma}$ can be derived using Newey (1994), and is given in the next section.

We also consider the HIT estimator of the Average Treatment Effect on the Treated (ATT)

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n}\frac{d_i}{p}\left(\widehat{\gamma}_1\left(\widehat{\varphi}(x_i)\right) - \widehat{\gamma}_2\left(\widehat{\varphi}(x_i)\right)\right)$$

with $p = \Pr(d=1)$. This estimator is a special case of that considered in Theorem 4 with $h(w,\gamma_1,\gamma_2) = \frac{d}{p}(\gamma_1-\gamma_2)$ except for the fact that the non-parametric regressions γ_1 and γ_2 are for subpopulations and the average is over the subpopulation of the treated. This requires some changes in the proof. The functions s_1, s_2 are now

$$s_1(\varphi(x,\alpha)) = \frac{1}{p}$$

$$s_2(\varphi(x,\alpha)) = \frac{\varphi(x,\alpha)}{p(1-\varphi(x,\alpha))}$$

These are obtained by multiplying the functions that we used above by $\frac{\varphi(x,\alpha)}{p}$, a factor that re-weights the orthogonality conditions from the full population to the subpopulation of the treated. This gives two equations that hold for all α

$$\mathbb{E}\left[\frac{\varphi(x,\alpha_*)\mathbb{E}[y|x,d=1]}{p}\right] = \mathbb{E}\left[\frac{\varphi(x,\alpha_*)\gamma_1\left(\varphi\left(x,\alpha\right);\alpha\right)}{p}\right]$$

$$\mathbb{E}\left[\frac{\left(1-\varphi(x,\alpha_*)\right)\varphi\left(x,\alpha\right)\mathbb{E}[y|x,d=0]}{p\left(1-\varphi\left(x,\alpha\right)\right)}\right] = \mathbb{E}\left[\frac{\left(1-\varphi(x,\alpha_*)\right)\varphi\left(x,\alpha\right)\gamma_2\left(\varphi\left(x,\alpha\right);\alpha\right)}{p\left(1-\varphi\left(x,\alpha\right)\right)}\right]$$

Differentiation with respect to α gives the derivatives of γ_1 and γ_2 with respect to the α in the conditioning variable, i.e., the parametric propensity score. Note that these derivatives are different from those for the estimation of the ATE which shows that these derivatives depend on the third stage of the estimator that is different for the ATE (averaging over the full population) and the ATT (averaging over the subpopulation of the treated). Substitution gives the contribution of the estimation of $\hat{\alpha}$ to the influence function that is equal to (more details in Appendix D). With

$$g(w,\alpha,\alpha,\gamma) = \frac{\varphi\left(x,\alpha_*\right)}{p}(\gamma_1(\varphi(x,\alpha);\alpha) - \gamma_2(\varphi(x,\alpha);\alpha))$$

we find that the contribution is

$$\left(\mathbb{E}\left[\frac{\partial g\left(w,\alpha_*,\alpha_*,\gamma_*\right)}{\partial\alpha_1}\right] + \mathbb{E}\left[\frac{\partial g\left(w,\alpha_*,\alpha_*,\gamma_*\right)}{\partial\alpha_2}\right]\right)\sqrt{n}(\hat{\alpha}-\alpha_*) =$$

$$-\mathbb{E}\left[\frac{\mathbb{E}[y|x,d=0] - \mu_2\left(\varphi\left(x,\alpha_*\right)\right)}{p\left(1-\varphi\left(x,\alpha_*\right)\right)}\frac{\partial\varphi(x,\alpha_*)}{\partial\alpha}\right]\sqrt{n}(\hat{\alpha}-\alpha_*)$$

5.2 Non-parametric First Step, Non-parametric Second Step

The analysis in the previous section combined with the results in Newey (1994) show that in the case that the first stage is non-parametric the contribution of the first-stage estimation to the influence function of the ATE estimator is

$$-\left(\frac{\mathbb{E}\left[y\middle|\,x,d=1\right]-\mu_1\left(\varphi_*\left(x\right)\right)}{\varphi_*\left(x\right)} + \frac{\mathbb{E}\left[y\middle|\,x,d=0\right]-\mu_2\left(\varphi_*\left(x\right)\right)}{1-\varphi_*\left(x\right)}\right)(d-\varphi_*(x))$$

which can be alternatively written as

$$\begin{aligned}&-\frac{\mathbb{E}\left[y\middle|\,x,d=1\right]-\mu_1\left(\varphi_*\left(x\right)\right)}{\varphi_*\left(x\right)}d + \left(\mathbb{E}\left[y\middle|\,x,d=1\right]-\mu_1\left(\varphi_*\left(x\right)\right)\right)\\&+\frac{\mathbb{E}\left[y\middle|\,x,d=0\right]-\mu_2\left(\varphi_*\left(x\right)\right)}{1-\varphi_*\left(x\right)}\left(1-d\right) - \left(\mathbb{E}\left[y\middle|\,x,d=0\right]-\mu_2\left(\varphi_*\left(x\right)\right)\right)\end{aligned} \tag{13}$$

To obtain the complete influence function of $\widehat{\beta}$ we need the contribution of the estimation error in $\widehat{\gamma}$. This contribution is derived in Appendix E and is equal to

$$\begin{aligned}&\left(\mu_1\left(\varphi_*\left(x\right)\right)-\mu_2\left(\varphi_*\left(x\right)\right)-\beta_*\right)+\\&\frac{d}{\varphi_*\left(x\right)}\left(y-\mu_1\left(\varphi_*\left(x\right)\right)\right) - \frac{1-d}{1-\varphi_*\left(x\right)}\left(y-\mu_2\left(\varphi_*\left(x\right)\right)\right)\end{aligned} \tag{14}$$

Adding (13) and (14), we obtain the influence function of the estimator based on regressions on the estimated propensity score:

$$\left(\mathbb{E}\left[y\middle|\,x,d=1\right]-\mathbb{E}\left[y\middle|\,x,d=0\right]-\beta_*\right)+\frac{d}{\varphi_*\left(x\right)}\left(y-\mathbb{E}\left[y\middle|\,x,d=1\right]\right)-\frac{1-d}{1-\varphi_*\left(x\right)}\left(y-\mathbb{E}\left[y\middle|\,x,d=0\right]\right)$$

which is the influence function of the efficient estimator and also that of the imputation estimator

$$\widehat{\beta}_I = \frac{1}{n}\sum_{i=1}^{n}(\widehat{\lambda}_1(x_i) - \widehat{\lambda}_2(x_i))$$

with $\lambda_1(x) = \mathbb{E}[y|x, d=1], \lambda_2(x) = \mathbb{E}[y|x, d=0]$. The imputation estimator involves nonparametric regressions on x and not on the estimated propensity score. However these two estimators have the same influence function which shows that regressing on the non-parametrically estimated propensity score does not result in an efficiency gain. The infeasible estimator that depends on non-parametric regressions on the population propensity score is less efficient than the estimator that uses the estimated propensity score.

For the estimator of the ATT the contribution of the first stage is

$$-\frac{\mathbb{E}[y|x, d=0] - \mu_2(\varphi_*(x))}{p(1-\varphi_*(x))}(d - \varphi_*(x))$$

The main term and the contribution of the estimation of the (infeasible) non-parametric regressions is

$$\frac{d}{p}(y - \mu_1(\varphi_*(x))) - \frac{(1-d)\varphi_*(x)}{p(1-\varphi_*(x))}(y - \mu_2(\varphi_*(x))) + \frac{d}{p}(\mu_1(\varphi_*(x)) - \mu_2(\varphi_*(x)) - \beta_*)$$

which can be derived using an argument virtually identical to Appendix E. Adding these expressions we obtain the full influence function

$$\frac{d}{p}(y - \mathbb{E}[y|x, d=1]) - \frac{(1-d)\varphi_*(x)}{p(1-\varphi_*(x))}(y - \mathbb{E}[y|x, d=0]) + \frac{d}{p}(\mathbb{E}[y|x, d=1] - \mathbb{E}[y|x, d=0] - \beta_*)$$

As in the case of the ATE the influence function is the same as that for the estimator that involves non-parametric regressions on x and not on the estimated propensity score, so that again there is no first-order asymptotic efficiency gain if we use the estimated propensity score in the non-parametric regressions.

It should be noted that the influence functions derived in this section are different from those found in the literature. Recently, Mammen, Rothe, and Schienle (2010) derived the influence function for the ATE estimator considered in this section. They concluded that it is identical to that of the infeasible estimator that regresses on the population propensity score. This is because they imposed the index assumption $E[y|d, x] = E[y|d, \varphi(x)]$, which is implicitly ruled out in standard program evaluation literature. HIT derived the influence function for the ATT estimator that is also different from ours. In this case, the derivation fails to account for the effect of the first-stage estimation on the conditional expectation in the second stage. Only the variability of the first-stage estimator as an argument is considered.

5.3 Approximating the Influence Function for the Non-parametric First Step with a Parametric First Step

We assume that for the population propensity score

$$\varphi_*(x) = \varphi(x, \alpha_*) = p(x)'\alpha_*$$

where $p(x)$ is a finite-, possibly high-dimensional vector of functions of x. We can think of this expression as a series approximation of the propensity score with basis functions in the vector $p(x)$. The influence function for the least squares estimator of α_* is

$$\left(\mathbb{E}\left[p(x)\,p(x)'\right]\right)^{-1} p(x)\,(d-\varphi_*(x)) \tag{15}$$

Using the result in subsection 5.1, the adjustment to the influence function for the first step estimation is

$$-\mathbb{E}\left[\left(\frac{\mathbb{E}\left[y|\,x,d=1\right]-\mu_1\left(\varphi(x,\alpha_*)\right)}{\varphi(x,\alpha_*)}+\frac{\mathbb{E}\left[y|\,x,d=0\right]-\mu_2\left(\varphi(x,\alpha_*)\right)}{1-\varphi(x,\alpha_*)}\right)\frac{\partial\varphi(x,\alpha_*)}{\partial\alpha'}\right]\sqrt{n}\left(\widehat{\alpha}-\alpha\right)$$

$$=-\mathbb{E}\left[\Psi(x)\,p(x)'\right]\sqrt{n}\left(\widehat{\alpha}-\alpha\right) \tag{16}$$

where

$$\Psi(x)=\frac{\mathbb{E}\left[y|\,x,d=1\right]-\mu_1\left(\varphi(x,\alpha_*)\right)}{\varphi(x,\alpha_*)}+\frac{\mathbb{E}\left[y|\,x,d=0\right]-\mu_2\left(\varphi(x,\alpha_*)\right)}{1-\varphi(x,\alpha_*)}$$

for simplicity. Combining (15) and (16), we conclude that the adjustment to the influence function can be written as

$$-\mathbb{E}\left[\Psi(x)\,p(x)'\right]\left(\mathbb{E}\left[p(x)\,p(x)'\right]\right)^{-1}p(x)\,(d-\varphi_*(x)) \tag{17}$$

Now $\left(\mathbb{E}\left[p(x)\,p(x)'\right]\right)^{-1}\mathbb{E}\left[p(x)\,\Psi(x)\right]$ are the coefficients of the linear projection of $\Psi(x)$ on $p(x)$. In other words, we can write

$$p(x)'\left(\mathbb{E}\left[p(x)\,p(x)'\right]\right)^{-1}\mathbb{E}\left[p(x)\,\Psi(x)\right]=\Pi\left(\Psi(x)|\,p(x)\right)$$

where $\Pi\left(\cdot|\,p(x)\right)$ denotes the projection on the linear space spanned by $p(x)$. If the dimension of $p(x)$ is sufficiently large, then approximately $\Pi\left(\Psi(x)|\,p(x)\right)\approx\mathbb{E}\left[\Psi(x)|\,x\right]=\Psi(x)$. It follows that the adjustment to the influence function in (17) is

$$\begin{aligned}
&-\mathbb{E}\left[\Psi(x)\,p(x)'\right]\left(\mathbb{E}\left[p(x)\,p(x)'\right]\right)^{-1}p(x)\,(d-\varphi_*(x))\\
&\approx-\Psi(x)\,(d-\varphi_*(x))\\
&=-\left(\frac{\mathbb{E}\left[y|\,x,d=1\right]-\mu_1\left(\varphi(x,\alpha_*)\right)}{\varphi_*(x)}+\frac{\mathbb{E}\left[y|\,x,d=0\right]-\mu_2\left(\varphi(x,\alpha_*)\right)}{1-\varphi_*(x)}\right)(d-\varphi_*(x))
\end{aligned}$$

which is the result in the previous section, i.e., if the parametric approximation to the population propensity score is good, then the influence function is close to efficient influence function.

6 A Semi-parametric Control Variable Estimator

Hahn, Hu and Ridder (2008) consider a model that is nonlinear in a mismeasured independent variable. The details of their model are not important here. For our purpose it suffices to note that their estimator uses a control variable and the asymptotic analysis requires dealing with a generated regressor in a V-statistic. Because of the V-statistic structure, the results in Section 3 do not apply directly, but the basic approach can be easily modified. Suppose that an econometrician observes a random sample $w_i=(y_i,x_i,z_i)\,,i=1,\ldots,n$. The estimator of a parameter β has the following three steps:

1. Estimate a finite dimensional parameter $\widehat{\alpha}$ by nonlinear least squares of x on $\psi(z,\alpha)$ and obtain the residual $\widehat{v} = x - \psi\left(z,\widehat{\alpha}\right) = \varphi(x,z,\widehat{\alpha})$ that is our generated regressor.

2. Estimate $\mu(x,v_*) = \mathbb{E}\left[y \mid x, v_*\right]$ nonparametrically using the sample $\left(y_i, x_i, \widehat{v}_i\right), i = 1,\ldots,n$. Call the estimator $\widehat{\gamma}(x,\widehat{v})$. Let $L\left(x\right) = \mathbb{E}_{v_*}\left[\mu(x,v_*)\right]$ and $\widehat{L}\left(x\right) = \frac{1}{n}\sum_{j=1}^{n}\widehat{\gamma}(x,\widehat{v}_j)$.

3. Assume that $L\left(x\right) = R\left(x,\beta_*\right)$ for a known function R and define $\widehat{\beta}$ as the solution of the minimization problem

$$\min_{\beta}\frac{1}{n}\sum_{i=1}^{n}1_C\left(x_i\right)\left(\widehat{L}\left(x_i\right) - R\left(x_i,\beta\right)\right)^2$$

for some set C. In the sequel we will ignore the indicator function 1_C for simplicity.

Let $\widehat{\beta}$ denote the solution to the preceding minimization problem that satisfies the moment condition

$$0 = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{L}\left(x_i\right) - R\left(x_i,\widehat{\beta}\right)\right)\frac{\partial R\left(x_i,\widehat{\beta}\right)}{\partial\beta}.$$

Characterization of asymptotic distribution of $\widehat{\beta}$ requires characterization of the influence function of

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\widehat{L}\left(x_i\right) - L\left(x_i\right)\right)r\left(x_i\right),$$

where $r\left(x_i\right) = \partial R\left(x_i,\beta_*\right)/\partial\beta$. We define

$$\begin{aligned}\varphi\left(x,z,\alpha\right) &= x - \psi\left(z,\alpha\right)\\ \gamma\left(x,v^*;\alpha\right) &= \mathbb{E}\left[y \mid\ x,\varphi\left(x,z,\alpha\right) = v^*\right]\\ g\left(x,\alpha_1,\alpha_2,\gamma,F_{xz}\right) &= \int\gamma\left(x,\varphi\left(\tilde{x},\tilde{z},\alpha_1\right);\alpha_2\right)r\left(x\right)\mathrm{d}F_{xz}(\tilde{x},\tilde{z})\end{aligned}$$

where an integral with respect to $\widehat{F}_{xz}$ is just an average over x,z. Note that because of the V statistic structure we integrate with respect to the distribution of x,z that appear in $\varphi(x,z,\alpha)$.

With these definitions, we can now write

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{L}\left(x_i\right)r\left(x_i\right) = \frac{1}{n}\sum_{i=1}^{n}g\left(w_i,\widehat{\alpha}_1,\widehat{\alpha}_2,\widehat{\gamma},\widehat{F}_{xz}\right),$$

where $\widehat{\alpha}_1 = \widehat{\alpha}_2 = \widehat{\alpha}$ but written separately to emphasize the dual role of α. The contribution of $\widehat{\gamma}$ and $\widehat{F}_{xz}$ can be derived as in Newey (1994) and by the V-statistic projection theorem, respectively, and we concentrate on the contribution of $\widehat{\alpha}$.

The contribution of the estimation error of $\widehat{\alpha}$ is that error multiplied by the sum of the derivatives with respect to α_1, i.e., the α that appears in the argument, and α_2, i.e., the α in the conditioning variable. We have

$$\frac{\partial}{\partial\alpha_1}\mathbb{E}_x\left[\int\gamma\left(x,\varphi\left(\tilde{x},\tilde{z},\alpha_1\right);\alpha_*\right)r\left(x\right)\mathrm{d}F_{xz}(\tilde{x},\tilde{z})\right]\Bigg|_{\alpha=\alpha_*} = \mathbb{E}_x\left[\int\frac{\partial}{\partial\alpha_1}\mu\left(x,\varphi\left(\tilde{x},\tilde{z},\alpha_1\right)\right)r\left(x\right)\mathrm{d}F_{xz}(\tilde{x},\tilde{z})\right]\Bigg|_{\alpha=\alpha_*}$$

$$= -\mathbb{E}_x \left[\int \frac{\partial \mu \left(x, \varphi \left(\tilde{x}, \tilde{z}, \alpha_* \right) \right)}{\partial v} \frac{\partial \psi \left(\tilde{x}, \tilde{z}, \alpha_* \right)}{\partial \alpha} r \left(x \right) \mathrm{d}F_{xz}(\tilde{x}, \tilde{z}) \right] \equiv \Xi_1$$

For the derivative with respect to α_2 we first observe that

$$\begin{aligned} \mathbb{E}_x \left[\int \gamma \left(x, \varphi \left(\tilde{x}, \tilde{z}, \alpha_* \right); \alpha \right) r \left(x \right) \mathrm{d}F_{xz}(\tilde{x}, \tilde{z}) \right] &= \int \int \gamma \left(x, v_*; \alpha \right) r \left(x \right) \frac{f \left(x \right) f \left(v_* \right)}{f \left(x, v_* \right)} f \left(x, v_* \right) \mathrm{d}x \mathrm{d}v_* \\ &= \mathbb{E} \left[\gamma \left(x, v_*; \alpha \right) r \left(x \right) \frac{f \left(x \right) f \left(v_* \right)}{f \left(x, v_* \right)} \right] \end{aligned}$$

We compute the derivative of the final expression. For that we note that $\gamma \left(x, \varphi \left(x, z, \alpha \right); \alpha \right)$ solves the minimization problem $\min_p \mathbb{E} \left[\left(y - p \left(x, \varphi \left(x, z, \alpha \right) \right) \right)^2 \right]$ so that

$$0 = \mathbb{E} \left[\left(y - \gamma \left(x, \varphi \left(x, z, \alpha \right); \alpha \right) \right) s \left(x, \varphi \left(x, z, \alpha \right) \right) \right]$$

for all square integrable function $s \left(x, \varphi \left(x, z, \alpha \right) \right)$ and all α. In particular, we have for all α

$$\mathbb{E} \left[\left(y - \gamma \left(x, \varphi \left(x, z, \alpha \right); \alpha \right) \right) r \left(x \right) \frac{f \left(x \right) f \left(\varphi \left(x, z, \alpha \right) \right)}{f \left(x, \varphi \left(x, z, \alpha \right) \right)} \right] = 0$$

If we differentiate with respect to α and evaluate at $\alpha = \alpha_*$ we obtain

$$\begin{aligned} \mathbb{E} \left[\frac{\partial \gamma \left(x, v_*; \alpha_* \right)}{\partial \alpha_2} r \left(x \right) \frac{f \left(x \right) f \left(v_* \right)}{f \left(x, v_* \right)} \right] &= \mathbb{E} \left[\left(y - \gamma \left(x, v_*; \alpha_* \right) \right) r \left(x \right) f \left(x \right) \frac{\partial \frac{f(v_*)}{f(x, v_*)}}{\partial v} \frac{\partial \varphi \left(x, z, \alpha_* \right)}{\partial \alpha} \right] \\ &\quad - \mathbb{E} \left[\frac{\partial \gamma \left(x, v_*; \alpha_* \right)}{\partial v} r \left(x \right) \frac{\partial \varphi \left(x, z, \alpha_* \right)}{\partial \alpha} \frac{f \left(x \right) f \left(v_* \right)}{f \left(x, v_* \right)} \right] \end{aligned}$$

We therefore obtain

$$\frac{\partial}{\partial \alpha_2} \mathbb{E} \left[\gamma \left(x, v_*; \alpha_* \right) r(x) \frac{f \left(x \right) f \left(v_* \right)}{f \left(x, v_* \right)} \right]$$

$$= \mathbb{E} \left[\left(\frac{\partial \mu \left(x, v_* \right)}{\partial v} - \left(y - \mu \left(x, v_* \right) \right) \left(\frac{\partial \ln f(v_*)}{\partial v} - \frac{\partial \ln f(x, v_*)}{\partial v} \right) \right) r \left(x \right) \frac{\partial \psi \left(z, \alpha_* \right)}{\partial \alpha} \frac{f \left(x \right) f \left(v_* \right)}{f \left(x, v_* \right)} \right] \equiv \Xi_2$$

The contribution of the first step estimation to the influence function is then

$$(\Xi_1 + \Xi_2) \sqrt{n} (\widehat{\alpha} - \alpha_*)$$

7 Conclusion

We studied the asymptotic distribution of three-step estimators of a finite dimensional parameter vector where the second step consists of one or more non-parametric regressions on a regressor that is estimated in the first step. The first step estimator is either parametric or non-parametric. We showed that Newey's (1994) path-derivative method can be used to determine the contribution of the first-step estimation error on the influence function. In doing so it is essential to recognize that the first-stage estimate has two effects on the sampling distribution of the finite-dimensional parameter vector. First, the first-stage estimate enters the argument

at which the conditional expectation is evaluated, second, the first-stage estimate changes the conditional expectation itself. In the literature the second contribution of the first-stage estimate to the influence function is sometimes forgotten. Our contribution is that we show how to derive this contribution so that we obtain the correct influence function for three- or more stage estimators.

Appendix

A Proof of (1)

We first examine the adjustment to the influence function of $\widehat{\gamma}$ to account for the estimation error of $\widehat{\alpha}$. Noting that $\widehat{\gamma}$ is an M-estimator corresponding to the population moment equation

$$\mathbb{E}\left[\begin{array}{c} y-\gamma_1-\gamma_2 x-\gamma_3\varphi\left(x,z,\alpha\right) \\ x\left(y-\gamma_1-\gamma_2 x-\gamma_3\varphi\left(x,z,\alpha\right)\right) \\ \varphi\left(x,z,\alpha\right)\left(y-\gamma_1-\gamma_2 x-\gamma_3\varphi\left(x,z,\alpha\right)\right) \\ \psi(x,z)-\alpha \end{array}\right]=0$$

we obtain upon linearizing the corresponding sample moment equation and upon solving for $\sqrt{n}(\widehat{\gamma}-\gamma_*)$

$$\sqrt{n}\left(\widehat{\gamma}-\gamma_*\right)=-G_\gamma^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\left[\begin{array}{c} \varepsilon_i \\ x_i\varepsilon_i \\ \varphi\left(x_i,z_i,\alpha_*\right)\varepsilon_i \end{array}\right]+G_\alpha\psi\left(x_i,z_i\right)\right)+o_p\left(1\right)$$

where

$$\varepsilon_i=y_i-\gamma_{1*}-\gamma_{2*}x_i-\gamma_{3*}\varphi\left(x_i,z_i,\alpha_*\right)$$

$$G_\gamma=-\mathbb{E}\left[\begin{array}{ccc} 1 & x & \varphi\left(x,z,\alpha_*\right) \\ x & x^2 & x\varphi\left(x,z,\alpha_*\right) \\ \varphi\left(x,z,\alpha_*\right) & x\varphi\left(x,z,\alpha_*\right) & \varphi\left(x,z,\alpha_*\right)^2 \end{array}\right]$$

and

$$G_\alpha=-\mathbb{E}\left[\begin{array}{c} \gamma_{3*}\frac{\partial\varphi(x,z,\alpha_*)}{\partial\alpha} \\ \gamma_{3*}x\frac{\partial\varphi(x,z,\alpha_*)}{\partial\alpha} \\ 2\gamma_{3*}\varphi\left(x,z,\alpha_*\right)\frac{\partial\varphi(x,z,\alpha_*)}{\partial\alpha} \end{array}\right]$$

Likewise, we obtain from the population moment equation

$$\mathbb{E}\left[\begin{array}{c} y-\gamma_1-\gamma_2 x-\gamma_3\varphi\left(x,z,\alpha_*\right) \\ x\left(y-\gamma_1-\gamma_2 x-\gamma_3\varphi\left(x,z,\alpha_*\right)\right) \\ \varphi\left(x,z,\alpha_*\right)\left(y-\gamma_1-\gamma_2 x-\gamma_3\varphi\left(x,z,\alpha_*\right)\right) \end{array}\right]=0$$

that

$$\sqrt{n}\left(\widetilde{\gamma}-\gamma_*\right)=-G_\gamma^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\begin{array}{c} \varepsilon_i \\ x_i\varepsilon_i \\ \varphi\left(w_i,\alpha_*\right)\varepsilon_i \end{array}\right]+o_p\left(1\right)$$

It follows that

$$\begin{aligned}\Delta = &- \left[\begin{array}{ccc} 1 & \mathbb{E}[x] & \mathbb{E}[\varphi(x,z,\alpha_*)] \end{array}\right] G_\gamma^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[\begin{array}{c} \varepsilon_i \\ x_i\varepsilon_i \\ \varphi(x_i,z_i,\alpha_*)\,\varepsilon_i \end{array}\right] \\ &- \left[\begin{array}{ccc} 1 & \mathbb{E}[x] & \mathbb{E}[\varphi(x,z,\alpha_*)] \end{array}\right] G_\gamma^{-1} G_\alpha \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(x_i,z_i) \\ &+ \mathbb{E}\left[\gamma_{3*} \frac{\partial \varphi(x,z,\alpha_*)}{\partial \alpha}\right] \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(x_i,z_i)\end{aligned}$$

Now note that

$$\begin{aligned}&- \left[\begin{array}{ccc} 1 & \mathbb{E}[x] & \mathbb{E}[\varphi(x,z,\alpha_*)] \end{array}\right] G_\gamma^{-1} \\ &= \left[\begin{array}{ccc} 1 & \mathbb{E}[x] & \mathbb{E}[\varphi(x,z,\alpha_*)] \end{array}\right] \left(\mathbb{E}\left[\begin{array}{ccc} 1 & x & \varphi(x,z,\alpha_*) \\ x & x^2 & x\varphi(x,z,\alpha_*) \\ \varphi(x,z,\alpha_*) & x\varphi(x,z,,\alpha_*) & \varphi(x,z,\alpha_*)^2 \end{array}\right]\right)^{-1} \\ &= \left[\begin{array}{ccc} 1 & 0 & 0 \end{array}\right]\end{aligned}$$

and therefore,

$$\begin{aligned}&\mathbb{E}\left[\gamma_{3*} \frac{\partial \varphi(x,z,\alpha_*)}{\partial \alpha}\right] - \left[\begin{array}{ccc} 1 & \mathbb{E}[x] & \mathbb{E}[\varphi(x,z,\alpha_*)] \end{array}\right] G_\gamma^{-1} G_\alpha \\ &= \mathbb{E}\left[\gamma_{3*} \frac{\partial \varphi(x,z,\alpha_*)}{\partial \alpha}\right] + \left[\begin{array}{ccc} 1 & 0 & 0 \end{array}\right] \mathbb{E}\left[\begin{array}{c} -\gamma_{3*} \frac{\partial \varphi(x,z,\alpha_*)}{\partial \alpha} \\ -\gamma_{3*} x \frac{\partial \varphi(x,z,\alpha_*)}{\partial \alpha} \\ -2\gamma_{3*}\varphi(x,z,\alpha_*) \frac{\partial \varphi(x,z,\alpha_*)}{\partial \alpha} \end{array}\right] \\ &= 0\end{aligned}$$

It follows that

$$\Delta = - \left[\begin{array}{ccc} 1 & \mathbb{E}[x] & \mathbb{E}[\varphi(x,z,\alpha_*)] \end{array}\right] G_\gamma^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[\begin{array}{c} \varepsilon_i \\ x_i\varepsilon_i \\ \varphi(x_i,z_i,\alpha_*)\,\varepsilon_i \end{array}\right].$$

B Interpretation of (3)

In order to understand the additional term

$$- \left[\begin{array}{ccc} 1 & \mathbb{E}[x] & \mathbb{E}[\varphi(x,z,\alpha_*)] \end{array}\right] G_\gamma^{-1} G_\alpha \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(x_i,z_i),$$

we examine

$$\begin{aligned}(\gamma_1(\widehat{\alpha}) + \gamma_2(\widehat{\alpha})\,\mathbb{E}[x] + \gamma_3(\widehat{\alpha})\,\mathbb{E}[\varphi(x,z,\alpha_*)]) - (\gamma_1(\alpha_*) + \gamma_2(\alpha_*)\,\mathbb{E}[x] + \gamma_3(\alpha_*)\,\mathbb{E}[\varphi(x,z,\alpha_*)]) \\ = \left[\begin{array}{ccc} 1 & \mathbb{E}[x] & \mathbb{E}[\varphi(x,z,\alpha_*)] \end{array}\right] (\gamma(\widehat{\alpha}) - \gamma(\alpha_*))\end{aligned}$$

Because $\gamma(\alpha)$ is defined by the moment equation

$$\mathbb{E}\left[\begin{array}{c} y-\gamma_1(\alpha)-\gamma_2(\alpha)x-\gamma_3(\alpha)\varphi(x,z,\alpha) \\ x\left(y-\gamma_1(\alpha)-\gamma_2(\alpha)x-\gamma_3(\alpha)\varphi(x,z,\alpha)\right) \\ \varphi(x,z,\alpha)\left(y-\gamma_1(\alpha)-\gamma_2(\alpha)x-\gamma_3(\alpha)\varphi(x,z,\alpha)\right) \end{array}\right]=0$$

which holds for all α, we can take the derivative with respect to α to derive

$$\frac{\partial\gamma(\alpha)}{\partial\alpha}=-G_\gamma^{-1}G_\alpha$$

It follows that

$$\begin{aligned}\frac{\partial}{\partial\alpha}\left(\gamma_1(\alpha)+\gamma_2(\alpha)\mathbb{E}[x]+\gamma_3(\alpha)\mathbb{E}[\varphi(x,z,\alpha_*)]\right)&=\left[\begin{array}{cccc}1 & \mathbb{E}[x] & \mathbb{E}[\varphi(x,z,\alpha_*)]\end{array}\right]\frac{\partial\gamma(\alpha)}{\partial\alpha}\\&=-\left[\begin{array}{cccc}1 & \mathbb{E}[x] & \mathbb{E}[\varphi(x,z,\alpha_*)]\end{array}\right]G_\gamma^{-1}G_\alpha\end{aligned}$$

so that

$$\begin{aligned}\sqrt{n}\left(\Psi(\widehat{\alpha})-\Psi(\alpha_*)\right)&=\sqrt{n}\left[\begin{array}{cccc}1 & \mathbb{E}[x] & \mathbb{E}[\varphi(x,z,\alpha_*)]\end{array}\right]\left(\gamma(\widehat{\alpha})-\gamma(\alpha_*)\right)\\&=\left[\begin{array}{cccc}1 & \mathbb{E}[x] & \mathbb{E}[\varphi(x,z,\alpha_*)]\end{array}\right]\frac{\partial\gamma(\alpha_*)}{\partial\alpha}\sqrt{n}\left(\widehat{\alpha}-\alpha_*\right)\\&=-\left[\begin{array}{cccc}1 & \mathbb{E}[x] & \mathbb{E}[\varphi(x,z,\alpha_*)]\end{array}\right]G_\gamma^{-1}G_\alpha\sqrt{n}\left(\widehat{\alpha}-\alpha_*\right)\\&=-\left[\begin{array}{cccc}1 & \mathbb{E}[x] & \mathbb{E}[\varphi(x,z,\alpha_*)]\end{array}\right]G_\gamma^{-1}G_\alpha\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(x_i,z_i)\end{aligned}$$

C Proof of Theorems in Section 3

Proof of Theorem 1 We compute the right hand side of (4)

$$\mathbb{E}\left[\frac{\partial h\left(\mu\left(x,\varphi(x,z,\alpha_*)\right)\right)}{\partial\mu}\frac{\partial\gamma\left(x,\varphi(x,z,\alpha_*);\alpha_*\right)}{\partial\alpha_2}\right]$$

We note that $\gamma(x,\varphi(w;\alpha),\alpha)$ solves the minimization problem

$$\min_p\mathbb{E}\left[\left(y-p\left(x,\varphi(x,z,\alpha)\right)\right)^2\right]$$

so that for all square integrable functions s of $x,\varphi(x,z,\alpha)$

$$\mathbb{E}\left[\left(y-\gamma\left(x,\varphi(x,z,\alpha);\alpha\right)\right)s\left(x,\varphi(x,z,\alpha)\right)\right]=0$$

If we choose

$$s\left(x,\varphi(x,z,\alpha)\right)=\frac{\partial h\left(\mu\left(x,\varphi(x,z,\alpha)\right)\right)}{\partial\mu}$$

we have for all α

$$\mathbb{E}\left[\left(y-\gamma\left(x,\varphi(x,z,\alpha);\alpha\right)\right)\frac{\partial h\left(\mu\left(x,\varphi(x,z,\alpha)\right)\right)}{\partial\mu}\right]=0$$

We now take the derivative and evaluate it at $\alpha = \alpha_*$. We find

$$\mathbb{E}\left[\frac{\partial g\left(w,\alpha_*,\alpha_*,\gamma_*\right)}{\partial \alpha_2}\right] = \mathbb{E}\left[\frac{\partial \gamma(x,\varphi(x,z,\alpha_*);\alpha_*)}{\partial \alpha_2}\frac{\partial h\left(\mu\left(x,\varphi\left(x,z,\alpha_*\right)\right)\right)}{\partial \mu}\right] =$$

$$-\mathbb{E}\left[\frac{\partial \gamma(x,\varphi(x,z,\alpha_*);\alpha_*)}{\partial v}\frac{\partial \varphi(x,z,\alpha_*)}{\partial \alpha}\frac{\partial h\left(\mu\left(x,\varphi\left(x,z,\alpha_*\right)\right)\right)}{\partial \mu}\right] +$$

$$\mathbb{E}\left[\left(y-\gamma\left(x,\varphi\left(x,z,\alpha_*\right);\alpha_*\right)\right)\frac{\partial^2 h\left(\mu\left(x,\varphi\left(x,z,\alpha_*\right)\right)\right)}{\partial \mu^2}\frac{\partial \mu\left(x,\varphi\left(x,z,\alpha_*\right)\right)}{\partial v}\frac{\partial \varphi(x,z,\alpha_*)}{\partial \alpha}\right]$$

Adding $\mathbb{E}\left[\frac{\partial g(w,\alpha_*,\alpha_*,\gamma_*)}{\partial \alpha_1}\right]$ and noting that

$$\mathbb{E}\left[\frac{\partial \gamma(x,\varphi(x,z,\alpha_*);\alpha_*)}{\partial v}\frac{\partial \varphi(x,z,\alpha_*)}{\partial \alpha}\frac{\partial h\left(\mu\left(x,\varphi\left(x,z,\alpha_*\right)\right)\right)}{\partial \mu}\right] =$$

$$\mathbb{E}\left[\frac{\partial \mu(x,\varphi(x,z,\alpha_*))}{\partial v}\frac{\partial \varphi(x,z,\alpha_*)}{\partial \alpha}\frac{\partial h\left(\mu\left(x,\varphi\left(x,z,\alpha_*\right)\right)\right)}{\partial \mu}\right]$$

we find the desired result □

Proof of Theorem 2 As before, we write

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(h\left(\gamma\left(x_i,\varphi\left(x_i,z_i,\widehat{\alpha}\right);\widehat{\alpha}\right)\right)-h\left(\gamma\left(x_i,\varphi\left(x_i,z_i,\alpha_*\right);\alpha_*\right)\right)\right)\\
&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(g\left(w_i,\widehat{\alpha}_1,\widehat{\alpha}_2\right)-g\left(w_i,\alpha_*,\alpha_*\right)\right)\\
&=\left(\mathbb{E}\left[\frac{\partial g\left(w,\alpha_*,\alpha_*\right)}{\partial \alpha_1}\right]+\mathbb{E}\left[\frac{\partial g\left(w,\alpha_*,\alpha_*\right)}{\partial \alpha_2}\right]\right)\sqrt{n}(\widehat{\alpha}-\alpha_*)+o_p\left(1\right)
\end{aligned}$$

Therefore we must compute $\mathbb{E}\left[\frac{\partial g(w,\alpha_*,\alpha_*,\gamma_*)}{\partial \alpha_1}\right]$ and $\mathbb{E}\left[\frac{\partial g(w,\alpha_*,\alpha_*,\gamma_*)}{\partial \alpha_2}\right]$. The computation of the first expectation is easy. Because $\gamma\left(x,\varphi\left(x,z,\alpha\right);\alpha_*\right)=\mu\left(x,\varphi\left(x,z,\alpha\right)\right)$, we have

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial g\left(w,\alpha_*,\alpha_*,\gamma_*\right)}{\partial \alpha_1}\right] &= \mathbb{E}\left[\frac{\partial h\left(\mu\left(x,\varphi\left(x,z,\alpha_*\right)\right)\right)}{\partial \mu'}\frac{\partial \mu\left(x,\varphi\left(x,z,\alpha_*\right)\right)}{\partial v'}\frac{\partial \varphi\left(x,z,\alpha_*\right)}{\partial \alpha}\right]\\
&= \sum_j \mathbb{E}\left[\frac{\partial h\left(\mu\left(x,\varphi\left(x,z,\alpha_*\right)\right)\right)}{\partial \mu_j}\frac{\partial \mu_j\left(x,\varphi\left(x,z,\alpha_*\right)\right)}{\partial v'}\frac{\partial \varphi\left(x,z,\alpha_*\right)}{\partial \alpha}\right]
\end{aligned}$$

where μ_j denotes the j-th component of μ, etc. We now tackle the second expectation. By the chain rule

$$\mathbb{E}\left[\frac{\partial g\left(w,\alpha_*,\alpha_*,\gamma_*\right)}{\partial \alpha_2}\right] = \mathbb{E}\left[\frac{\partial h\left(\mu\left(x,\varphi\left(x,z,\alpha_*\right)\right)\right)}{\partial \mu'}\frac{\partial \gamma\left(x,\varphi\left(x,z,\alpha_*\right);\alpha_*\right)}{\partial \alpha_2}\right] \tag{18}$$

We compute the right hand side of (18). We note that each component $\gamma_j(x, \varphi(w;\alpha), \alpha)$ of $\gamma(x, \varphi(w;\alpha), \alpha)$ solves the minimization problem

$$\min_p \mathbb{E}\left[\left(y_j - p(x, \varphi(x, z, \alpha))\right)^2\right]$$

for each component y_j of y, so that for all square integrable functions s of $x, \varphi(x, z, \alpha)$

$$\mathbb{E}\left[\left(y_j - \gamma_j(x, \varphi(x, z, \alpha); \alpha)\right) s(x, \varphi(x, z, \alpha))\right] = 0$$

If we choose

$$s(x, \varphi(x, z, \alpha)) = \frac{\partial h(\mu(x, \varphi(x, z, \alpha)))}{\partial \mu_j}$$

we have for all α

$$\mathbb{E}\left[\left(y_j - \gamma_j(x, \varphi(x, z, \alpha); \alpha)\right) \frac{\partial h(\mu(x, \varphi(x, z, \alpha)))}{\partial \mu_j}\right] = 0$$

We now take the derivative and evaluate it at $\alpha = \alpha_*$. We find

$$\mathbb{E}\left[\frac{\partial g(w, \alpha_*, \alpha_*, \gamma_*)}{\partial \alpha_2}\right] = \sum_j \mathbb{E}\left[\frac{\partial h(\mu(x, \varphi(x, z, \alpha_*)))}{\partial \mu_j} \frac{\partial \gamma_j(x, \varphi(x, z, \alpha_*); \alpha_*)}{\partial \alpha_2}\right] =$$

$$-\sum_j \mathbb{E}\left[\frac{\partial h(\mu(x, \varphi(x, z, \alpha_*)))}{\partial \mu_j} \frac{\partial \gamma_j(x, \varphi(x, z, \alpha_*); \alpha_*)}{\partial v'} \frac{\partial \varphi(x, z, \alpha_*)}{\partial \alpha}\right] +$$

$$\sum_j \mathbb{E}\left[\left(y_j - \gamma_j(x, \varphi(x, z, \alpha_*); \alpha_*)\right) \frac{\partial^2 h(\mu(x, \varphi(x, z, \alpha_*)))}{\partial \mu_j^2} \frac{\partial \mu_j(x, \varphi(x, z, \alpha_*))}{\partial v'} \frac{\partial \varphi(x, z, \alpha_*)}{\partial \alpha}\right]$$

Adding $\mathbb{E}\left[\frac{\partial g(w, \alpha_*, \alpha_*, \gamma_*)}{\partial \alpha_1}\right]$ and noting that

$$\mathbb{E}\left[\frac{\partial h(\mu(x, \varphi(x, z, \alpha_*)))}{\partial \mu_j} \frac{\partial \gamma_j(x, \varphi(x, z, \alpha_*); \alpha_*)}{\partial v'} \frac{\partial \varphi(x, z, \alpha_*)}{\partial \alpha}\right] =$$

$$\mathbb{E}\left[\frac{\partial h(\mu(x, \varphi(x, z, \alpha_*)))}{\partial \mu_j} \frac{\partial \mu_j(x, \varphi(x, z, \alpha_*))}{\partial v'} \frac{\partial \varphi(x, z, \alpha_*)}{\partial \alpha}\right]$$

we find the desired result.□

Proof of Theorem 4 The contribution of $\widehat{\alpha}$ is the sum of

$$\mathbb{E}\left[\frac{\partial h(w, \gamma(x, \varphi(x, z, \alpha_*); \alpha_*))}{\partial \alpha_1}\right] \sqrt{n}(\widehat{\alpha} - \alpha_*) \tag{19}$$

and

$$\mathbb{E}\left[\frac{\partial h(w, \gamma(x, \varphi(x, z, \alpha_*); \alpha_*))}{\partial \alpha_2}\right] \sqrt{n}(\widehat{\alpha} - \alpha_*) \tag{20}$$

Note that

$$\mathbb{E}\left[\frac{\partial h\left(w,\gamma\left(x,\varphi\left(x,z,\alpha_*\right);\alpha_*\right)\right)}{\partial\alpha_1}\right]=\mathbb{E}\left[\frac{\partial h\left(w,\mu\left(x,\varphi\left(x,z,\alpha_*\right)\right)\right)}{\partial\mu(x,\varphi(x,z,\alpha_*))}\frac{\partial\mu}{\partial v}\frac{\partial\varphi\left(x,z,\alpha_*\right)}{\partial\alpha}\right] \tag{21}$$

and

$$\mathbb{E}\left[\frac{\partial h\left(w,\gamma\left(x,\varphi\left(x,z,\alpha_*\right);\alpha_*\right)\right)}{\partial\alpha_2}\right]=\mathbb{E}\left[\frac{\partial h\left(w,\mu\left(x,\varphi\left(x,z,\alpha_*\right)\right)\right)}{\partial\mu}\frac{\partial\gamma\left(x,\varphi\left(x,z,\alpha_*\right);\alpha_*\right)}{\partial\alpha_2}\right] \tag{22}$$

Because $\frac{\partial\gamma(x,\varphi(x,z,\alpha_*);\alpha_*)}{\partial\alpha_2}$ is a function of $(x,\varphi\left(x,z,\alpha_*\right))$, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{\partial h\left(w,\gamma\left(x,\varphi\left(x,z,\alpha_*\right);\alpha_*\right)\right)}{\partial\alpha_2}\right]\\
&=\mathbb{E}\left[\frac{\partial h\left(w,\mu\left(x,\varphi\left(x,z,\alpha_*\right)\right)\right)}{\partial\mu}\frac{\partial\gamma\left(x,\varphi\left(x,z,\alpha_*\right);\alpha_*\right)}{\partial\alpha_2}\right]\\
&=\mathbb{E}\left[\mathbb{E}\left[\frac{\partial h\left(w,\mu\left(x,\varphi\left(x,z,\alpha_*\right)\right)\right)}{\partial\mu}\middle|x,\varphi\left(x,z,\alpha_*\right)\right]\frac{\partial\gamma\left(x,\varphi\left(x,z,\alpha_*\right);\alpha_*\right)}{\partial\alpha_2}\right]\\
&=\mathbb{E}\left[\kappa\left(x,\varphi\left(x,z,\alpha_*\right)\right)\frac{\partial\gamma\left(x,\varphi\left(x,z,\alpha_*\right);\alpha_*\right)}{\partial\alpha_2}\right]
\end{aligned} \tag{23}$$

We now note that $\gamma\left(x,\varphi\left(x,z,\alpha\right),\alpha\right)$ solves the minimization problem

$$\min_s\mathbb{E}\left[\left(y-s\left(x,\varphi\left(x,z,\alpha\right)\right)\right)^2\right]$$

we have that for all α

$$\mathbb{E}\left[\left(y-\gamma\left(x,\varphi\left(x,z,\alpha\right);\alpha\right)\right)\kappa\left(x,\varphi\left(x,z,\alpha\right)\right)\right]=0$$

We now take the derivative with respect to α and evaluate it at $\alpha=\alpha_*$:

$$\begin{aligned}
&\mathbb{E}\left[\frac{\partial\gamma\left(x,\varphi\left(x,z,\alpha_*\right);\alpha_*\right)}{\partial\alpha_2}\kappa\left(x,\varphi\left(x,z,\alpha_*\right)\right)\right]\\
&=\mathbb{E}\left[\left(y-\gamma\left(x,\varphi\left(x,z,\alpha_*\right);\alpha_*\right)\right)\frac{\partial\kappa\left(x,\varphi\left(x,z,\alpha_*\right)\right)}{\partial v}\frac{\partial\varphi\left(x,z,\alpha_*\right)}{\partial\alpha}\right]\\
&-\mathbb{E}\left[\kappa\left(x,\varphi\left(x,z,\alpha_*\right)\right)\frac{\partial\mu\left(x,\varphi\left(x,z,\alpha_*\right)\right)}{\partial v}\frac{\partial\varphi\left(x,z,\alpha_*\right)}{\partial\alpha}\right]
\end{aligned} \tag{24}$$

Combining (21) - (24), we conclude that

$$\begin{aligned}
&\mathbb{E}\left[\frac{\partial h\left(w,\gamma\left(x,\varphi\left(x,z,\alpha_*\right);\alpha_*\right)\right)}{\partial\alpha_1}\right]+\mathbb{E}\left[\frac{\partial h\left(w,\gamma\left(x,\varphi\left(x,z,\alpha_*\right);\alpha_*\right)\right)}{\partial\alpha_2}\right]\\
&=\mathbb{E}\left[\left(\frac{\partial h\left(w,\mu\left(x,\varphi\left(x,z,\alpha_*\right)\right)\right)}{\partial\mu}-\kappa\left(x,\varphi\left(x,z,\alpha_*\right)\right)\right)\frac{\partial\mu\left(x,\varphi\left(x,z,\alpha_*\right)\right)}{\partial v}\frac{\partial\varphi\left(x,z,\alpha_*\right)}{\partial\alpha}\right]\\
&+\mathbb{E}\left[\frac{\partial\kappa\left(x,\varphi\left(x,z,\alpha_*\right)\right)}{\partial v}\left(y-\gamma\left(x,\varphi\left(x,z,\alpha_*\right);\alpha_*\right)\right)\frac{\partial\varphi\left(x,z,\alpha_*\right)}{\partial\alpha}\right]
\end{aligned}$$

which gives us the desired result □

D Details of Derivations in Section 5

Derivation of the contribution of $\hat{\alpha}$ for ATE The first step is the same as in Theorem 3

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial g\left(w,\alpha_*,\alpha_*,\gamma_*\right)}{\partial\alpha}\right] &= \mathbb{E}\left[\frac{\partial g\left(w,\alpha_1,\alpha_*,\gamma_*\right)}{\partial\alpha_1}\right]+\mathbb{E}\left[\frac{\partial g\left(w,\alpha_*,\alpha_2,\gamma_*\right)}{\partial\alpha_2}\right] \qquad (25)\\
&= \mathbb{E}\left[\frac{\partial}{\partial\alpha_1}\left(\gamma_1\left(\varphi\left(x,\alpha_1\right);\alpha_*\right)-\gamma_2\left(\varphi\left(x,\alpha_1\right);\alpha_*\right)\right)\right]\Big|_{\alpha_1=\alpha_*}\\
&+ \mathbb{E}\left[\frac{\partial}{\partial\alpha_2}\left(\gamma_1\left(\varphi\left(x,\alpha_*\right);\alpha_2\right)-\gamma_2\left(\varphi\left(x,\alpha_*\right);\alpha_2\right)\right)\right]\Big|_{\alpha_2=\alpha_*}\\
&= \mathbb{E}\left[\left(\frac{\partial\mu_1\left(\varphi\left(x,\alpha_*\right)\right)}{\partial v}-\frac{\partial\mu_2\left(\varphi\left(x,\alpha_*\right)\right)}{\partial v}\right)\frac{\partial\varphi\left(x,\alpha_*\right)}{\partial\alpha}\right]\\
&+ \mathbb{E}\left[\frac{\partial\gamma_1\left(\varphi\left(x,\alpha_*\right);\alpha_2\right)}{\partial\alpha_2}\Big|_{\alpha_2=\alpha_*}-\frac{\partial\gamma_2\left(\varphi\left(x,\alpha_*\right);\alpha_2\right)}{\partial\alpha_2}\Big|_{\alpha_2=\alpha_*}\right]
\end{aligned}$$

Using the orthogonality conditions in Section 5 we find that for all α

$$\mathbb{E}\left[\frac{\varphi\left(x,\alpha_*\right)\mathbb{E}[y|x,d=1])}{\varphi\left(x,\alpha\right)}\right]=\mathbb{E}\left[\frac{d\gamma_1\left(\varphi\left(x,\alpha\right);\alpha\right)}{\varphi\left(x,\alpha\right)}\right]$$

$$\mathbb{E}\left[\frac{\left(1-\varphi\left(x,\alpha_*\right)\right)\mathbb{E}[y|x,d=0]}{1-\varphi\left(x,\alpha\right)}\right]=\mathbb{E}\left[\frac{\left(1-d\right)\gamma_2\left(\varphi\left(x,\alpha\right);\alpha\right)}{1-\varphi\left(x,\alpha\right)}\right]$$

Differentiating these two equations with respect to α we obtain

$$-\mathbb{E}\left[\frac{\mathbb{E}[y|x,d=1]}{\varphi\left(x,\alpha_*\right)}\frac{\partial\varphi\left(x,\alpha_*\right)}{\partial\alpha}\right]= \qquad (26)$$

$$\mathbb{E}\left[\left(\frac{\partial\gamma_1\left(\varphi\left(x,\alpha_*\right);\alpha_*\right)}{\partial v}-\frac{\gamma_1\left(\varphi\left(x,\alpha_*\right);\alpha_*\right)}{\varphi\left(x,\alpha_*\right)}\right)\frac{\partial\varphi\left(x,\alpha_*\right)}{\partial\alpha}\right]+\mathbb{E}\left[\frac{\partial\gamma_1\left(\varphi\left(x,\alpha_*\right);\alpha_*\right)}{\partial\alpha_2}\right]$$

and

$$\mathbb{E}\left[\frac{\mathbb{E}[y|x,d=1]}{1-\varphi\left(x,\alpha_*\right)}\frac{\partial\varphi\left(x,\alpha_*\right)}{\partial\alpha}\right]= \qquad (27)$$

$$\mathbb{E}\left[\left(\frac{\partial\gamma_2\left(\varphi\left(x,\alpha_*\right);\alpha_*\right)}{\partial v}+\frac{\gamma_2\left(\varphi\left(x,\alpha_*\right);\alpha_*\right)}{1-\varphi\left(x,\alpha_*\right)}\right)\frac{\partial\varphi\left(x,\alpha_*\right)}{\partial\alpha}\right]+\mathbb{E}\left[\frac{\partial\gamma_2\left(\varphi\left(x,\alpha_*\right);\alpha_*\right)}{\partial\alpha_2}\right]$$

Substituting (26) and (27) in (25) we obtain

$$\mathbb{E}\left[\frac{\partial g\left(w,\alpha_*,\alpha_*,\gamma_*\right)}{\partial\alpha}\right]=$$

$$-\mathbb{E}\left[\left(\frac{\mathbb{E}\left[y\right|x,d=1]-\mu_1\left(\varphi\left(x,\alpha_*\right)\right)}{\varphi\left(x,\alpha_*\right)}+\frac{\mathbb{E}\left[y\right|x,d=0]-\mu_2\left(\varphi\left(x,\alpha_*\right)\right)}{1-\varphi\left(x,\alpha_*\right)}\right)\frac{\partial\varphi\left(x,\alpha_*\right)}{\partial\alpha_1}\right]$$

Derivation of the contribution of $\hat{\alpha}$ for ATT The first step is the same as in Theorem 3

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial g\left(w,\alpha_*,\alpha_*,\gamma_*\right)}{\partial\alpha}\right] &= \mathbb{E}\left[\frac{\partial g\left(w,\alpha_1,\alpha_*,\gamma_*\right)}{\partial\alpha_1}\right]+\mathbb{E}\left[\frac{\partial g\left(w,\alpha_*,\alpha_2,\gamma_*\right)}{\partial\alpha_2}\right] \\
&= \mathbb{E}\left[\frac{\varphi(x,\alpha_*)}{p}\frac{\partial}{\partial\alpha_1}\left(\gamma_1\left(\varphi\left(x,\alpha_1\right);\alpha_*\right)-\gamma_2\left(\varphi\left(x,\alpha_1\right);\alpha_*\right)\right)\right]\Bigg|_{\alpha_1=\alpha_*} \\
&+ \mathbb{E}\left[\frac{\varphi(x,\alpha_*)}{p}\frac{\partial}{\partial\alpha_2}\left(\gamma_1\left(\varphi\left(x,\alpha_*\right);\alpha_2\right)-\gamma_2\left(\varphi\left(x,\alpha_*\right);\alpha_2\right)\right)\right]\Bigg|_{\alpha_2=\alpha_*} \\
&= \mathbb{E}\left[\frac{\varphi(x,\alpha_*)}{p}\left(\frac{\partial\mu_1\left(\varphi\left(x,\alpha_*\right)\right)}{\partial v}-\frac{\partial\mu_2\left(\varphi\left(x,\alpha_*\right)\right)}{\partial v}\right)\frac{\partial\varphi\left(x,\alpha_*\right)}{\partial\alpha}\right] \\
&+ \mathbb{E}\left[\frac{\varphi(x,\alpha_*)}{p}\left(\frac{\partial\gamma_1\left(\varphi\left(x,\alpha_*\right);\alpha_2\right)}{\partial\alpha_2}\Bigg|_{\alpha_2=\alpha_*}-\frac{\partial\gamma_2\left(\varphi\left(x,\alpha_*\right);\alpha_2\right)}{\partial\alpha_2}\Bigg|_{\alpha_2=\alpha_*}\right)\right]
\end{aligned}$$

As shown in Section 5 the following equations hold for all α

$$\mathbb{E}\left[\frac{\varphi(x,\alpha_*)\mathbb{E}[y|x,d=1]}{p}\right]=\mathbb{E}\left[\frac{\varphi(x,\alpha_*)\gamma_1\left(\varphi\left(x,\alpha\right);\alpha\right)}{p}\right]$$

$$\mathbb{E}\left[\frac{\left(1-\varphi(x,\alpha_*)\right)\varphi\left(x,\alpha\right)\mathbb{E}[y|x,d=0]}{p\left(1-\varphi\left(x,\alpha\right)\right)}\right]=\mathbb{E}\left[\frac{\left(1-\varphi(x,\alpha_*)\right)\varphi\left(x,\alpha\right)\gamma_2\left(\varphi\left(x,\alpha\right);\alpha\right)}{p\left(1-\varphi\left(x,\alpha\right)\right)}\right]$$

Differentiation with respect to α gives

$$\mathbb{E}\left[\frac{\varphi(x,\alpha_*)}{p}\frac{\partial\gamma_1\left(\varphi\left(x,\alpha_*\right);\alpha_2\right)}{\partial\alpha_2}\Bigg|_{\alpha=\alpha_*}\right]=-\mathbb{E}\left[\frac{\varphi(x,\alpha_*)}{p}\frac{\partial\mu_1(\varphi(x,\alpha_*))}{\partial v}\frac{\partial\varphi(x,\alpha_*)}{\partial\alpha}\right]$$

$$\mathbb{E}\left[\frac{\varphi(x,\alpha_*)}{p}\frac{\partial\gamma_2\left(\varphi\left(x,\alpha_*\right);\alpha_2\right)}{\partial\alpha_2}\Bigg|_{\alpha=\alpha_*}\right]=$$

$$\mathbb{E}\left[\left(\frac{\mathbb{E}[y|x,d=0]-\mu_2\left(\varphi\left(x,\alpha_*\right)\right)}{p\left(1-\varphi\left(x,\alpha_*\right)\right)}-\frac{\varphi(x,\alpha_*)}{p}\frac{\partial\mu_2(\varphi(x,\alpha_*))}{\partial v}\right)\frac{\partial\varphi(x,\alpha_*)}{\partial\alpha}\right]$$

Upon substitution we obtain

$$\mathbb{E}\left[\frac{\partial g\left(w,\alpha_*,\alpha_*,\gamma_*\right)}{\partial\alpha}\right]=$$

$$-\mathbb{E}\left[\frac{\mathbb{E}[y|x,d=0]-\mu_2\left(\varphi\left(x,\alpha_*\right)\right)}{p\left(1-\varphi\left(x,\alpha_*\right)\right)}\frac{\partial\varphi(x,\alpha_*)}{\partial\alpha}\right]$$

E The Influence Function of the Imputation Estimator

The ATE is

$$\beta_*=\mathbb{E}\left[\lambda_1\left(x\right)-\lambda_2\left(x\right)\right]$$

with

$$\lambda_1(x) = \mathbb{E}[y|\, d=1, x]$$
$$\lambda_2(x) = \mathbb{E}[y|\, d=0, x]$$

The ATE satisfies the moment equation

$$0 = \mathbb{E}[m(x, \beta_*, \lambda_1, \lambda_2)]$$

where

$$m(x, \beta_*, \lambda_1, \lambda_2) = \lambda_1(x) - \lambda_2(x) - \beta_*$$

The imputation estimator for the ATE is

$$\widehat{\beta} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\lambda}_1(x_i) - \widehat{\lambda}_2(x_i)\right)$$

so that we need to consider the linear functional

$$\mathbb{E}[D(x)'\lambda(x)]$$

with $D(x) = (1,-1)'$ and $D(x)'\lambda(x)$ is linear in λ.

Following Newey (1994) define a path indexed by the scalar parameter θ for the distribution of (y,d,x) with density $f(\cdot,\theta)$ where $f(\cdot,0) = f(\cdot)$ the population density of (y,d,x). If $\mathbb{E}_\theta$ denotes an expectation with respect to the distribution with density $f(x,\theta)$, then we define the corresponding paths for the projections $\lambda_1(x,\theta) = \mathbb{E}_\theta[y|x,d=1]$ and $\lambda_2(x,\theta) = \mathbb{E}_\theta[y|x,d=0]$. To determine the contribution of the estimation of λ_1, λ_2 to the influence function Newey (1994) suggests that we compute

$$\frac{\partial \mathbb{E}\left[D(x)'\lambda(x,\theta)\right]}{\partial\theta} = \frac{\partial \mathbb{E}\left[\lambda_1(x,\theta) - \lambda_2(x,\theta)\right]}{\partial\theta} \tag{28}$$

and evaluate the result at $\theta = 0$.

The path $\lambda(x,\theta)$ is the minimizer of a *single* objective function

$$\mathbb{E}_\theta\left[d\left(y - \widetilde{\lambda}_1(x)\right)^2 + (1-d)\left(y - \widetilde{\lambda}_2(x)\right)^2\right]$$

so that the following orthogonality condition holds

$$\mathbb{E}_\theta\left[d\left(y - \lambda_1(x,\theta)\right)s_1(x) + (1-d)\left(y - \lambda_2(x,\theta)\right)s_2(x)\right] = 0$$

for all functions $(s_1(x), s_2(x))'$. Choose $(s_1(x), s_2(x)) = \left(\frac{1}{\varphi_*(x)}, -\frac{1}{1-\varphi_*(x)}\right)$ with $\varphi_*(x) = \mathbb{E}[d|x]$, i.e., the propensity score is not that on the path, but the population propensity score. Therefore

$$\mathbb{E}_\theta\left[\frac{d}{\varphi_*(x)}\left(y - \lambda_1(x,\theta)\right) - \frac{1-d}{1-\varphi_*(x)}\left(y - \lambda_2(x,\theta)\right)\right] = 0 \tag{29}$$

or

$$\mathbb{E}_\theta\left[\frac{d}{\varphi_*(x)}y-\frac{1-d}{1-\varphi_*(x)}y\right]=\mathbb{E}_\theta\left[\frac{d}{\varphi_*(x)}\lambda_1(x,\theta)-\frac{1-d}{1-\varphi_*(x)}\lambda_2(x,\theta)\right] \tag{30}$$

which holds for all θ.

We differentiate the right-hand side of (30). By the chain rule (evaluate the derivatives at $\theta=0$)

$$\begin{aligned}\frac{\partial\mathbb{E}_\theta\left[\frac{d}{\varphi_*(x)}\lambda_1(x,\theta)-\frac{1-d}{1-\varphi_*(x)}\lambda_2(x,\theta)\right]}{\partial\theta}&=\frac{\partial\mathbb{E}_\theta\left[\frac{d}{\varphi_*(x)}\lambda_1(x)-\frac{1-d}{1-\varphi_*(x)}\lambda_2(x)\right]}{\partial\theta}\\&\quad+\frac{\partial\mathbb{E}\left[\frac{d}{\varphi_*(x)}\lambda_1(x,\theta)-\frac{1-d}{1-\varphi_*(x)}\lambda_2(x,\theta)\right]}{\partial\theta}\\&=\frac{\partial\mathbb{E}_\theta\left[\frac{d}{\varphi_*(x)}\lambda_1(x)-\frac{1-d}{1-\varphi_*(x)}\lambda_2(x)\right]}{\partial\theta}+\frac{\partial\mathbb{E}\left[\lambda_1(x,\theta)-\lambda_2(x,\theta)\right]}{\partial\theta}\end{aligned}$$

where we use the fact that the derivative of the projection paths at $\theta=0$ are equal to λ_1,λ_2.

Therefore combining this with the result above

$$\begin{aligned}\frac{\partial\mathbb{E}\left[D(x)'\lambda(x,\theta)\right]}{\partial\theta}&=\frac{\partial\mathbb{E}\left[\lambda_1(x,\theta)-\lambda_2(x,\theta)\right]}{\partial\theta}\\&=\frac{\partial\mathbb{E}_\theta\left[\frac{d}{\varphi_*(x)}\lambda_1(x,\theta)-\frac{1-d}{1-\varphi_*(x)}\lambda_2(x,\theta)\right]}{\partial\theta}-\frac{\partial\mathbb{E}_\theta\left[\frac{d}{\varphi_*(x)}\lambda_1(x)-\frac{1-d}{1-\varphi_*(x)}\lambda_2(x)\right]}{\partial\theta}\\&=\frac{\partial}{\partial\theta}\left(\mathbb{E}_\theta\left[\frac{d}{\varphi_*(x)}y-\frac{1-d}{1-\varphi_*(x)}y\right]-\mathbb{E}_\theta\left[\frac{d}{\varphi_*(x)}\lambda_1(x)-\frac{1-d}{1-\varphi_*(x)}\lambda_2(x)\right]\right)\end{aligned}$$

so that at $\theta=0$

$$\begin{aligned}\frac{\partial\mathbb{E}\left[D(x)'\lambda(x,\theta)\right]}{\partial\theta}&=\frac{\partial}{\partial\theta}\mathbb{E}_\theta\left[\frac{d}{\varphi_*(x)}\left(y-\lambda_1(x)\right)-\frac{1-d}{1-\varphi_*(x)}\left(y-\lambda_2(x)\right)\right]\\&=\mathbb{E}\left[\left(\frac{d}{\varphi_*(x)}\left(y-\lambda_1(x)\right)-\frac{1-d}{1-\varphi_*(x)}\left(y-\lambda_2(x)\right)\right)S(y,d,x)\right],\end{aligned}$$

with $S(\cdot)=\left.\frac{\partial\ln f(\cdot,\theta)}{\partial\theta}\right|_{\theta=0}$. Therefore the adjustment to the influence function is

$$\frac{d}{\varphi_*(x)}\left(y-\lambda_1(x)\right)-\frac{1-d}{1-\varphi_*(x)}\left(y-\lambda_2(x)\right)$$

and the influence function of the imputation estimator is

$$\left(\lambda_1(x)-\lambda_2(x)-\beta_*\right)+\frac{d}{\varphi_*(x)}\left(y-\lambda_1(x)\right)-\frac{1-d}{1-\varphi_*(x)}\left(y-\lambda_2(x)\right) \tag{31}$$

so this estimator is efficient, because this the efficient influence function of Hahn (1998).

The ATE is also equal to

$$\beta_* = \mathbb{E}\left[\mu_1\left(\varphi_*\left(x\right)\right) - \mu_2\left(\varphi_*\left(x\right)\right)\right]$$

with

$$\begin{aligned}\mu_1\left(x\right) &= \mathbb{E}\left[y\middle|\, d=1, \varphi_*\left(x\right)\right]\\ \mu_2\left(x\right) &= \mathbb{E}\left[y\middle|\, d=0, \varphi_*\left(x\right)\right]\end{aligned}$$

so that the same argument as above shows that the influence of the imputation estimator that uses regressions on the population propensity score is

$$\left(\mu_1\left(\varphi_*\left(x\right)\right) - \mu_2\left(\varphi_*\left(x\right)\right) - \beta_*\right) + \frac{d}{\varphi_*\left(x\right)}\left(y - \mu_1\left(\varphi_*\left(x\right)\right)\right) - \frac{1-d}{1-\varphi_*\left(x\right)}\left(y - \mu_2\left(\varphi_*\left(x\right)\right)\right) \quad (32)$$

The asymptotic variances implied by (31) and (32) are

$$\mathbb{E}\left[\left(\beta\left(x\right) - \beta_*\right)^2 + \frac{\mathrm{Var}\left(y_1\middle|\, x\right)}{\varphi_*\left(x\right)} + \frac{\mathrm{Var}\left(y_0\middle|\, x\right)}{1-\varphi_*\left(x\right)}\right] \quad (33)$$

and

$$\mathbb{E}\left[\left(\beta\left(\varphi_*\left(x\right)\right) - \beta_*\right)^2 + \frac{\left(y_1 - \mu_1\left(\varphi_*\left(x\right)\right)\right)^2}{\varphi_*\left(x\right)} + \frac{\left(y_0 - \mu_2\left(\varphi_*\left(x\right)\right)\right)^2}{1-\varphi_*\left(x\right)}\right] \quad (34)$$

where $\beta\left(x\right) = \lambda_1\left(x\right) - \lambda_2\left(x\right)$ and $\beta\left(\varphi_*\left(x\right)\right) = \mu_1\left(\varphi_*\left(x\right)\right) - \mu_2\left(\varphi_*\left(x\right)\right)$. Using

$$\begin{aligned}\mathbb{E}\left[\left(y_1 - \mu_1\left(\varphi_*\left(x\right)\right)\right)^2\middle|\, x\right] &= E\left[\left(\left(y_1 - \lambda_1\left(x\right)\right) + \left(\lambda_1\left(x\right) - \mu_1\left(\varphi_*\left(x\right)\right)\right)\right)^2\middle|\, x\right]\\ &= \mathrm{Var}\left(y_1\middle|\, x\right) + \left(\lambda_1\left(x\right) - \mu_1\left(\varphi_*\left(x\right)\right)\right)^2\end{aligned}$$

$$\mathbb{E}\left[\left(y_0 - \mu_2\left(\varphi_*\left(x\right)\right)\right)^2\middle|\, x\right] = \mathrm{Var}\left(y_0\middle|\, x\right) + \left(\lambda_2\left(x\right) - \mu_2\left(\varphi_*\left(x\right)\right)\right)^2$$

and

$$\begin{aligned}\mathbb{E}\left[\left(\beta\left(x\right) - \beta_*\right)^2\middle|\, \varphi_*\left(x\right)\right] &= \mathbb{E}\left[\left(\left(\beta\left(x\right) - \beta\left(\varphi_*\left(x\right)\right)\right) + \left(\beta\left(\varphi_*\left(x\right)\right) - \beta_*\right)\right)^2\middle|\, \varphi_*\left(x\right)\right]\\ &= \mathbb{E}\left[\left(\beta\left(x\right) - \beta\left(\varphi_*\left(x\right)\right)\right)^2\middle|\, \varphi_*\left(x\right)\right] + \left(\beta\left(\varphi_*\left(x\right)\right) - \beta_*\right)^2\end{aligned}$$

we note that

$$\begin{aligned}\mathbb{E}\left[\frac{\left(y_1 - \mu_1\left(\varphi_*\left(x\right)\right)\right)^2}{\varphi_*\left(x\right)}\right] &= \mathbb{E}\left[\frac{\mathrm{Var}\left(y_1\middle|\, x\right)}{\varphi_*\left(x\right)}\right] + \mathbb{E}\left[\frac{\left(\lambda_1\left(x\right) - \mu_1\left(\varphi_*\left(x\right)\right)\right)^2}{\varphi_*\left(x\right)}\right]\\ \mathbb{E}\left[\frac{\left(y_0 - \mu_2\left(\varphi_*\left(x\right)\right)\right)^2}{1-\varphi_*\left(x\right)}\right] &= \mathbb{E}\left[\frac{\mathrm{Var}\left(y_0\middle|\, x\right)}{1-\varphi_*\left(x\right)}\right] + \mathbb{E}\left[\frac{\left(\lambda_2\left(x\right) - \mu_2\left(\varphi_*\left(x\right)\right)\right)^2}{1-\varphi_*\left(x\right)}\right]\\ \mathbb{E}\left[\left(\beta\left(x\right) - \beta_*\right)^2\right] &= \mathbb{E}\left[\left(\beta\left(x\right) - \beta\left(\varphi_*\left(x\right)\right)\right)^2\right] + \mathbb{E}\left[\left(\beta\left(\varphi_*\left(x\right)\right) - \beta_*\right)^2\right].\end{aligned}$$

Therefore, we can see that the difference of (34) and (33) is equal to

$$\mathbb{E}\left[\frac{\left(\lambda_1(x)-\mu_1(\varphi_*(x))\right)^2}{\varphi_*(x)}+\frac{\left(\lambda_2(x)-\mu_2(\varphi_*(x))\right)^2}{1-\varphi_*(x)}\right]-\mathbb{E}\left[\left(\beta(x)-\beta(\varphi_*(x))\right)^2\right]$$
$$=\mathbb{E}\left[\frac{a(x)^2}{\varphi_*(x)}+\frac{b(x)^2}{1-\varphi_*(x)}-\left(a(x)-b(x)\right)^2\right]$$

for $a(x)=\lambda_1(x)-\mu_1(\varphi_*(x))$ and $b(x)=\lambda_2(x)-\mu_2(\varphi_*(x))$. Therefore, the difference of (34) and (33) is equal to

$$\mathbb{E}\left[\frac{1-\varphi_*(x)}{\varphi_*(x)}a(x)^2+\frac{\varphi_*(x)}{1-\varphi_*(x)}b(x)^2-2a(x)b(x)\right]$$
$$=\mathbb{E}\left[\left(\sqrt{\frac{1-\varphi_*(x)}{\varphi_*(x)}}a(x)-\sqrt{\frac{\varphi_*(x)}{1-\varphi_*(x)}}b(x)\right)^2\right]\geq 0$$

which establishes relative efficiency of imputation using on x over imputation using $\varphi_*(x)$.

References

[1] Abadie, A. and G. Imbens (2009a),"A Martingale Representation for Matching Estimators," unpublished working paper.

[2] Abadie, A. and G. Imbens (2009b),"Matching on the Estimated Propensity Score, " unpublished working paper.

[3] Ai, C. and X. Chen (2007): "Estimation of Possibly Misspecified Semiparametric Conditional Moment Restriction Models with Different Conditioning Variables," *Journal of Econometrics* 141, pp. 5 – 43.

[4] Barro, R.J. (1977): "Unanticipated Money Growth and Unemployment in the United States," *American Economic Review* 67, pp. 101–115.

[5] Hahn, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects, " *Econometrica*, 66,pp. 315–331.

[6] Hahn, J., Y. Hu, and G. Ridder (2008): "Instrumental Variable Estimation of Nonlinear Models with Nonclassical Measurement Error Using Control Variates," unpublished working paper.

[7] Heckman, J.J., H. Ichimura, and P. Todd (1998): "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies* 65, pp. 261 – 294.

[8] Hirano, K., G. Imbens, and G. Ridder (2003): "Efficient Estimation of the Average Treatment Effect Using the Estimated Propensity Score, " *Econometrica*, 71, pp. 1161–1189.

[9] Li, Q. and J.M. Wooldridge (2002): "Semiparametric Estimation of Partially Linear Models for Dependent Data with Generated Regressors," *Econometric Theory* 18, pp. 625–645.

[10] Mammen, E., C. Rothe, and M. Schienle (2010): "Nonparametric Regression with Nonparametrically Generated Covariates, " unpublished working paper.

[11] Murphy, K. M. and R. H. Topel (1985): "Estimation and Inference in Two-Step Econometric Models," *Journal of Business and Economic Statistics* 3, pp. 370 – 379.

[12] Newey, W.K. (1984): "A Method of Moments Interpretation of Sequential Estimators," *Economics Letters* 14, pp. 201 – 206.

[13] Newey, W.K. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica* 62, pp. 1349 – 1382.

[14] Newey, W.K. (2009): "Two-Step Series Estimation of Sample Selection Models," *Econometrics Journal* 12, pp. S217–S229.

[15] Olley, G.S. and A. Pakes (1996): "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica* 64, pp. 1263 – 1297.

[16] Pagan, A. (1984): "Econometric Issues in the Analysis of Regressions with Generated Regressors," *International Economic Review* 25, pp. 221 – 247.

[17] Pakes, A. and G.S. Olley (1995): "A Limit Theorem for a Smooth Class of Semiparametric Estimators," *Journal of Econometrics* 65, pp. 295 – 332.

[18] Rosenbaum, P., and D. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal effects, " *Biometrika* 70, pp. 41–55.

[19] Shefrin, S. (1979): "Unanticipated Monetary Growth and Output Fluctuations," *Economtic Iniquiry* 17, pp. 1–13.

[20] Song, K. (2008): "Uniform Convergence of Series Estimators over Function Spaces," *Econometric Theory* 24, pp. 1463–1499.

[21] Sperlich, S. (2009): "A Note on Nonparametric Estimation with Predicted Variables," *Econometrics Journal* 12, pp. 382–395.

[22] Wooldridge, J.M. (2002): *Econometric Analysis of Cross Section and Panel Data,* Cambridge: MIT Press.