

The Asymptotic Variance of Semiparametric Estimators with Generated Regressors*

Jinyong Hahn
Department of Economics, UCLA

Geert Ridder
Department of Economics, USC

February 23, 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 nonparametric. Using Newey's (1994) approach 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 nonparametric regression, i.e., that of conditioning variable and that of argument. We consider three examples in more detail: the Olley and Pakes (1996) production function estimator, 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. 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 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 nonparametric 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 nonparametrically. A special instance of his method deals with the case of a linear regression model with a nonparametrically estimated generated regressor. The purpose of this note is to use Newey's insights to derive the asymptotic variance of three or even multi-step estimators in which one of the steps is a nonparametric regression with a generated regressor. This generated regressor can be estimated parametrically or nonparametrically. Therefore we generalize Newey's result to the case of a moment condition for a finite dimensional parameter vector that depends on a conditional expectation (to be estimated nonparametrically) that itself depends on a generated regressor. Since Newey (1994) a number of estimators have been suggested that have this structure. We consider three examples: (i) the partially linear production function estimator of Olley and Pakes (1996), (ii) the Average Treatment Effect (ATE) estimator for the case of unconfounded treatment assignment suggested by Heckman, Ichimura, and Todd (1998) that involves two nonparametric regressions on the estimated propensity score, (iii) a parametric control variate estimator that depends on a nonparametric regression on a first stage estimated residual. These examples illustrate the method that can be used to derive the asymptotic variance of other estimators with the same structure not covered here.

It turns out that the generalization of Newey's (1994) results is straightforward, although not trivial *ex ante*. The key issue is to account for the contribution of the first stage estimation error of the generated regressor on the 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 independent variable. However, there is a second contribution due to the effect of the first stage estimate on the conditional expectation if we condition on an estimated instead of a population value of the regressor. It is the latter contribution that is specific to our setup and its derivation is the modest contribution of this note. The approach that we take in the derivation is the same as in Newey (1994), i.e. we derive the influence function of the estimator of the finite dimensional parameter vector heuristically. We do not give regularity conditions on the smoothness of the conditional expectations and/or on the smoothing parameters of the nonparametric estimators that make the difference between the estimator and its asymptotically linear representation converge to 0 at a rate that is faster than the parametric rate. Therefore our results do not depend on the particular nonparametric estimator used.

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 note, where the conditioning variables are not estimated. Therefore our results are not a special case of, but rather complement 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 short 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 examples 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 $w_i = (y_i, x_i, z_i)$, $i = 1, \dots, n$ 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 $\hat{\alpha}$ such that

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

with $\mathbb{E}[\psi(x_i, z_i)] = 0$ 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 $v = \varphi(x, z, \alpha_*)$, i.e. the solution to $\min_{\gamma_1, \gamma_2, \gamma_3} \mathbb{E}[(y - \gamma_1 - \gamma_2 x - \gamma_3 v)^2]$. Because we do not know α_* , we use the estimated $\hat{v}_i = \varphi(x_i, z_i, \hat{\alpha})$, so that the estimator $\hat{\gamma}$ of γ_* is the OLS estimator of y on x, \hat{v} . The estimator of β_* is obtained in the third step $\hat{\beta} = \frac{1}{n} \sum_{i=1}^n (\hat{\gamma}_1 + \hat{\gamma}_2 x_i + \hat{\gamma}_3 \varphi(x_i, z_i, \hat{\alpha}))$, so that we have $\beta_* = \mathbb{E}[\gamma_{1*} + \gamma_{2*} x + \gamma_{3*} \varphi(x, z, \alpha_*)]$. 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

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

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 calculation¹ reveals that

$$\Delta = - [1 \quad \mathbb{E}[x] \quad \mathbb{E}[\varphi(x, z, \alpha_*)]] G_\gamma^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \begin{bmatrix} \varepsilon_i \\ x_i \varepsilon_i \\ \varphi(x_i, z_i, \alpha_*) \varepsilon_i \end{bmatrix} + o_p(1), \quad (1)$$

where

$$G_\gamma = -\mathbb{E} \begin{bmatrix} 1 & x & \varphi(x, z, \alpha_*) \\ x & x^2 & x\varphi(x, z, \alpha_*) \\ \varphi(x, z, \alpha_*) & x\varphi(x, z, \alpha_*) & \varphi(x, z, \alpha_*)^2 \end{bmatrix}.$$

¹See Appendix A.

The expansion (1) can be given an intuitive interpretation by considering an infeasible estimator. Assume that α_* is known to the econometrician, and $v_i = \varphi(x_i, z_i, \alpha_*)$ is used in the regression. Let $\tilde{\gamma}$ denote the resulting OLS estimator of γ_* . The first order asymptotic properties of $\tilde{\beta} = \frac{1}{n} \sum_{i=1}^n (\tilde{\gamma}_1 + \tilde{\gamma}_2 x_i + \tilde{\gamma}_3 \varphi(x_i, z_i, \alpha_*))$ can be analyzed using the expansion

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

It can be shown² that

$$\begin{aligned} &\begin{bmatrix} 1 & \mathbb{E}[x] & \mathbb{E}[\varphi(x, z, \alpha_*)] \end{bmatrix} \sqrt{n} (\tilde{\gamma} - \gamma_*) \\ &= - \begin{bmatrix} 1 & \mathbb{E}[x] & \mathbb{E}[\varphi(x, z, \alpha_*)] \end{bmatrix} G_\gamma^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \begin{bmatrix} \varepsilon_i \\ x_i \varepsilon_i \\ \varphi(x_i, z_i, \alpha_*) \varepsilon_i \end{bmatrix} + o_p(1) \end{aligned} \quad (2)$$

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

In order to understand this apparent puzzle, it is convenient to define $\hat{\gamma}(\alpha) = (\hat{\gamma}_1(\alpha), \hat{\gamma}_2(\alpha), \hat{\gamma}_3(\alpha))$ as the OLS estimator with y as the dependent and x and $v = \varphi(x, z, \alpha)$ as the independent variables. Note that $\hat{\gamma} = \hat{\gamma}(\hat{\alpha})$ and $\tilde{\gamma} = \hat{\gamma}(\alpha_*)$. 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 $\tilde{\beta}$ would use the following decomposition

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

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

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

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

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

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

This naïve decomposition is missing one additional term,³ i.e.,

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

²See Appendix A.

³See Appendix A.

where

$$G_\alpha = \mathbb{E} \begin{bmatrix} -\gamma_{3*} \frac{\partial \varphi(x, z, \alpha_*)}{\partial \alpha} \\ -\gamma_{3*} x_i \frac{\partial \varphi(x, z, \alpha_*)}{\partial \alpha} \\ -2\gamma_{3*} \varphi(x, z, \alpha_*) \frac{\partial \varphi(x, z, \alpha_*)}{\partial \alpha} \end{bmatrix}$$

As shown in Appendix A $-G_\gamma^{-1} G_\alpha \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(x_i, z_i)$ is the effect of the sampling variation in $\hat{\alpha}$ on the sampling distribution of $\hat{\gamma}$. Defining $\Psi(\alpha) = \mathbb{E}[\gamma_1(\alpha) + \gamma_2(\alpha)x + \gamma_3(\alpha)\varphi(x, z, \alpha_*)]$, we show in Appendix B that the missing term is asymptotically equivalent to $\sqrt{n}(\Psi(\hat{\alpha}) - \Psi(\alpha_*))$. The expression $\gamma_1(\alpha) + \gamma_2(\alpha)x + \gamma_3(\alpha)\varphi(x, z, \alpha_*)$ that appears in the definition of $\Psi(\alpha)$ can be given an interesting interpretation. It is the linear projection of y on $1, x, \varphi(x, z, \alpha)$ when after projection we substitute $\varphi(x, z, \alpha_*)$ for $\varphi(x, z, \alpha)$. Note that the linear projection of y on $1, x, \varphi(x, z, \alpha)$ has coefficients $\gamma(\alpha)$. This specifies a function of $x, \varphi(x, z, \alpha)$ 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 $x, \varphi(x, z, \alpha_*)$. If we substitute the estimator $\hat{\alpha}$ then the two correction terms that account for the estimation error in $\hat{\alpha}$ 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 general the first step estimation 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 $\gamma(\alpha)$. The challenge to the econometrician is that when the projection is nonparametric, as is the case when the generated regressor is used in a nonparametric 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 Semiparametric Three-Step Estimators

We now present our two main results on semiparametric three-step estimators. In the first step we estimate a regressor. In the second step we estimate a nonparametric 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 nonparametric 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 nonparametric method. As was emphasized in the introduction, our characterization is based on Newey's (1994) calculation.

3.1 Parametric First Step, Nonparametric Second Step

We assume that we observe i.i.d. observations $w_i = (y_i, x_i, z_i), i = 1, \dots, n$. The first step is identical to that in Section 2, i.e., we have an estimator $\hat{\alpha}$ such that $\sqrt{n}(\hat{\alpha} - \alpha_*) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(x_i, z_i) + o_p(1)$ 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 nonparametric 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 nonparametric 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 nonparametric 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 nonparametric 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) derivative based 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., $\frac{1}{\sqrt{n}} \sum_{i=1}^n (h(\gamma(x_i, \varphi(x_i, z_i, \widehat{\alpha}))) - h(\mu(x_i, \varphi(x_i, z_i, \alpha_*))))$. 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.⁴

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

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

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

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

We define

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

Note that the two roles that α plays are made explicit in $g(w, \alpha_1, \alpha_2)$ that is obtained by substituting $v^* = \varphi(x, z, \alpha_1)$ in $\gamma(x, v^*; \alpha_2)$. Note also that $\mu(x, v_*) = \gamma(x, v_*; \alpha_*)$.

With these definitions, we can now write

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

⁴The fact that Newey's approach can be used "term-by-term" is illustrated with a slightly different example in Appendix E.1. There, we consider the case where the moment function includes multiple nonparametric objects, all of which are obtained by nonparametric regressions with possibly different independent variables.

where $\hat{\alpha}_1 = \hat{\alpha}_2 = \hat{\alpha}$, but we keep them separate to emphasize the two roles of $\hat{\alpha}$. It is intuitive to deal with the two roles that $\hat{\alpha}$ plays in the expansion by linearization and this amounts to taking partial derivatives:

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

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

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

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

$$\mathbb{E} \left[\frac{\partial g(w, \alpha_*, \alpha_*)}{\partial \alpha_2} \right] = \mathbb{E} \left[\frac{\partial h(\mu(x, \varphi(x, z, \alpha_*)))}{\partial \mu} \frac{\partial \gamma(x, \varphi(x, z, \alpha_*); \alpha_*)}{\partial \alpha} \right] \quad (4)$$

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

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

$$\begin{aligned} & \left(\mathbb{E} \left[\frac{\partial g(w, \alpha_*, \alpha_*)}{\partial \alpha_1} \right] + \mathbb{E} \left[\frac{\partial g(w, \alpha_*, \alpha_*)}{\partial \alpha_2} \right] \right) \sqrt{n}(\hat{\alpha} - \alpha_*) \\ &= \mathbb{E} \left[\frac{\partial^2 h(\mu(x, \varphi(x, z, \alpha_*)))}{\partial \mu^2} (y - \mu(x, \varphi(x, z, \alpha_*))) \frac{\partial \mu(x, \varphi(x, z, \alpha_*))}{\partial v} \frac{\partial \varphi(x, z, \alpha_*)}{\partial \alpha} \right] \sqrt{n}(\hat{\alpha} - \alpha_*) + o_p(1). \end{aligned} \quad (5)$$

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

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

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

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

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

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

If we choose

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

we have for all α

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

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

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

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

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

we find the desired result \square

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 Nonparametric First Step, Nonparametric Second Step

We now assume that the first step is nonparametric. Again we have a random sample $w_i = (y_i, x_i, z_i)$, $i = 1, \dots, n$. 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 on z is denoted by $v_* = \varphi_*(x, z)$. The first step is to estimate this projection by nonparametric regression. In the second step we estimate $\gamma(x, v_*) = \mathbb{E}[y | x, v_*]$ by nonparametric regression of y on $x, \hat{v} = \hat{\varphi}(x, z)$. Our interest is to characterize the first order asymptotic properties of

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

We define

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

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

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

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

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

For now we assume that γ need not be estimated. 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,

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

we obtain that

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

for

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

which is linear in v_α . Second, for

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

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

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

The variables x in $\delta_1(x, z)$ (and the conditioning variables in the expectation) are the variables in the subvector of x that enters $\varphi(z, x)$ so that we average over the variables in x that do not enter in $\varphi(z, x)$.

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

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

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

We summarize the result in a theorem:

Theorem 2 (Contribution nonparametric 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

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

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

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

Armed with Newey (1994, pp. 1360 – 61), we can easily conclude that the adjustment to the influence function is equal to

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

where

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

4 The Olley and Pakes Estimator

We consider a simplified version of the production function estimator of Olley and Pakes (1996). The simplification is that we assume that firms cannot close and we also ignore firm ageing. The Cobb-Douglas production function for firm i in period t is

$$\begin{aligned} y_{it} &= \beta_0 + \beta_k k_{it} + \beta_l l_{it} + \omega_{it} + \eta_{it} \\ \mathbb{E}[\omega_{it} | k_{it}, \omega_{i,t-1}] &= g(\omega_{i,t-1}) \\ \omega_{it} &= h_t(i_{it}, k_{it}) \end{aligned}$$

with k_{it}, l_{it} the capital and labor inputs, respectively, ω_{it} a productivity index that follows a first-order Markov process, and i_{it} investment. The third equation is the inverse of the firm's optimal investment choice in period t . Substitution of the third equation in the first gives (from now on we omit i)

$$\begin{aligned} y_t &= \beta_l l_t + \phi_t(i_t, k_t) + \eta_t \\ \phi_t(i_t, k_t) &= \beta_0 + \beta_k k_t + \omega_t = \beta_0 + \beta_k k_t + h_t(i_t, k_t) \end{aligned}$$

Olley and Pakes suggest the following three step estimator for β_k . In the first step β_l and ϕ_t are estimated by standard methods for partially linear models, where β_l and ϕ_t are identified as the minimizers of

$$\mathbb{E}[(y_t - \beta_l l_t - \phi_t(i_t, k_t))^2]$$

This minimization proceeds in two steps: first, for given β_l the function is minimized at $\phi_t(i_t, k_t) = \alpha_1(i_t, k_t) - \beta_l \alpha_2(i_t, k_t)$ with $\alpha_1(i_t, k_t) = \mathbb{E}[y_t | i_t, k_t]$ and $\alpha_2(i_t, k_t) = \mathbb{E}[l_t | i_t, k_t]$. Substitution and minimization over β_l identifies that parameter. Because ϕ_t minimizes an objective function, the sampling variation of its estimator has no effect on the asymptotic distribution of the estimator of β_l . This is shown in general by Newey (1994), pp. 1357-58.

By the Markov assumption on the productivity index we have

$$\mathbb{E}[y_{t+1} - \beta_l l_{t+1} | k_{t+1}] = \beta_0 + \beta_k k_{t+1} + \mathbb{E}[\omega_{t+1} | \omega_t, k_{t+1}] = \beta_0 + \beta_k k_{t+1} + g(\omega_t)$$

Defining the forecasting error $\xi_{t+1} = \omega_{t+1} - \mathbb{E}[\omega_{t+1} | \omega_t, k_{t+1}] = \omega_{t+1} - g(\omega_t)$ we have

$$y_{t+1} - \beta_l l_{t+1} = \beta_k k_{t+1} + g(\phi_t(i_t, k_t) - \beta_k k_t) + \xi_{t+1} + \eta_{t+1}$$

where β_0 has been absorbed in g . The β_k and g are identified as the minimizers of

$$\mathbb{E}[(y_{t+1} - \beta_l l_{t+1} - \beta_k k_{t+1} - g(\phi_t(i_t, k_t) - \beta_k k_t))^2]$$

with respect to (β, g) where we substitute β_l and $\phi_t(i_t, k_t)$ that were identified in the first step. Hence the parameters in the first step are (β_l, ϕ_t) , the second step nonparametric regression function is g and the parameter that is estimated in the third step is β_k .

Although it seems that β_k and g are estimated jointly, we have the same structure as in a partially linear model, so that we first minimize over g for given β_k . The solution is $g(\phi_t(i_t, k_t) - \beta_k k_t) = \gamma_1(\phi_t(i_t, k_t) - \beta_k k_t) - \beta_k \gamma_2(\phi_t(i_t, k_t) - \beta_k k_t)$ with $\gamma_1(\phi_t(i_t, k_t) - \beta_k k_t) = \mathbb{E}[y_{t+1} - \beta_l l_{t+1} | \phi_t(i_t, k_t) - \beta_k k_t]$ and $\gamma_2(\phi_t(i_t, k_t) - \beta_k k_t) = \mathbb{E}[k_{t+1} | \phi_t(i_t, k_t) - \beta_k k_t]$. In estimation the conditional expectations are replaced by nonparametric regressions. Upon substitution the criterion function depends on β_k only, and the estimator of β_k is just the nonlinear least squares estimator. Just as for the partial linear model

by Newey (1994), pp. 1357-58 the estimation error in g has no effect on the asymptotic distribution of β_k .⁵

If we have a random sample $(y_{i,t+1}, y_{it}, k_{i,t+1}, k_{it}, l_{i,t+1}, l_{it}, i_{it}), i = 1, \dots, n$, then

$$\sqrt{n}(\hat{\beta}_k - \beta_{k*}) =$$

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \left(y_{i,t+1} - \hat{\beta}_l l_{i,t+1} - \beta_{k*} k_{i,t+1} - \hat{g} \left(\hat{\phi}_t(i_{it}, k_{it}) - \beta_{k*} k_{it} \right) \right) \left(k_{i,t+1} - \hat{g}' \left(\hat{\phi}_t(i_{it}, k_{it}) - \beta_{k*} k_{it} \right) k_{it} \right) + o_p(1)$$

with

$$\hat{\phi}_t(i_{it}, k_{it}) = \hat{\alpha}_1(i_t, k_t) - \hat{\beta}_l \hat{\alpha}_2(i_t, k_t)$$

$$\hat{g} \left(\hat{\phi}_t(i_{it}, k_{it}) - \beta_{k*} \right) = \hat{\gamma}_1 \left(y_{i,t+1} - \hat{\beta}_l l_{i,t+1} \mid \hat{\phi}_t(i_{it}, k_{it}) - \beta_{k*} k_{it} \right) - \beta_{k*} \hat{\gamma}_2 \left(k_{i,t+1} \mid \hat{\phi}_t(i_{it}, k_{it}) - \beta_{k*} k_{it} \right)$$

$$\hat{g}' \left(\hat{\phi}_t(i_{it}, k_{it}) - \beta_{k*} \right) =$$

$$\hat{\gamma}'_1 \left[y_{i,t+1} - \hat{\beta}_l l_{i,t+1} \mid \hat{\phi}_t(i_{it}, k_{it}) - \beta_{k*} k_{it} \right] - \beta_{k*} \hat{\gamma}'_2 \left[k_{i,t+1} \mid \hat{\phi}_t(i_{it}, k_{it}) - \beta_{k*} k_{it} \right] - \hat{\gamma}_2 \left[k_{i,t+1} \mid \hat{\phi}_t(i_{it}, k_{it}) - \beta_{k*} k_{it} \right]$$

where $\hat{\gamma}'_1, \hat{\gamma}'_2$ denote the derivatives of the nonparametric regression estimators. This estimator fits into the setup in Section 3 if we set $z_i = (i_{it}, k_{it}, l_{it})$, $x_i = (y_{it}, k_{i,t+1}, l_{i,t+1})$ and $y_i = y_{i,t+1}$.

Noting that first stage nonparametric estimator $\hat{\phi}_t(i_{it}, k_{it})$ appears as an argument in the second stage nonparametric regression estimators in \hat{g} , it plays the two roles discussed earlier, so that we can derive the influence function as in Section 3 with the only added complication that besides nonparametric regression estimators their derivatives appear in the second stage, which in this case does not matter, because the estimator g is an extremum estimator.

5 Regression on the Estimated Propensity Score

There has been an ongoing debate on the role of the propensity score in the efficient estimation of the Average Treatment Effect (ATE) of an intervention. Since Hahn (1998) derived the semi-parametric efficiency bound for the ATE, there is a clear target for any proposed (semiparametric) estimator. Let y_0, y_1 denote the potential outcomes, d the treatment indicator, $y = dy_1 + (1 - d)y_0$ the observed outcome and x a vector of covariates. As shown by Rosenbaum and Rubin (1983) unconfounded assignment, i.e., $y_1, y_0 \perp d \mid x$, implies that $y_1, y_0 \perp d \mid \varphi(x)$ with $\varphi(x) = \Pr(d = 1 \mid x)$. As a consequence the ATE given x can be identified by $\mathbb{E}[y \mid d = 1, x] - \mathbb{E}[y \mid d = 0, x]$ or by $\mathbb{E}[y \mid d = 1, \varphi(x)] - \mathbb{E}[y \mid d = 0, \varphi(x)]$. These observations have led to a large number of estimators that can be classified into three groups. The most popular 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

⁵More specifically, Newey (1994), p. 1357 considered the two step estimation where the second step is given by

$$m(z, \beta, h) = \frac{\partial q(z, \beta, h)}{\partial \beta}$$

$$h(F) = \underset{\tilde{h}}{\operatorname{argmax}} E_F \left[q \left(z, \beta, \tilde{h} \right) \right]$$

Here because the q depends on β , this parameter is also one of the arguments of h . He goes and concludes that, for this problem where β and h are simultaneous determined, the estimation error due to h is asymptotically irrelevant.

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 that 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 nonparametric 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 $\mathbb{E}[y|d=1, x]$, $\mathbb{E}[y|d=0, x]$ 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 nonparametric regression on a propensity score that is estimated in a preliminary step. This estimator that was suggested and analyzed by Heckman, Ichimura, and Todd (1998) fits into our framework and is analyzed here. Our conclusion is that it 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 semiparametric efficiency. That does not mean that there is no role for the propensity score in assessing the identification or in the small sample performance of ATE estimators.

The objective function for the estimator based on a regression on the propensity score has a structure to which the results in Section 3 do not apply directly, but the basic approach can be easily adapted.

5.1 Parametric First Step, Nonparametric Second Step

We have a random sample $w_i = (y_i, x_i, d_i)$, $i = 1, \dots, n$. In the first step, we estimate $\hat{\alpha}$ such that

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

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

$$\gamma(v_*) = (\mathbb{E}[y | v_*, d = 1], \mathbb{E}[y | v_*, d = 0])'$$

where $v_* = \varphi(x_i, \alpha_*)$ is the parametrization of the propensity score. Because we do not observe α_* , we use $\hat{v} = \varphi(x_i, \hat{\alpha})$ for the nonparametric regression.

Our interest is to characterize the first order asymptotic properties of

$$\hat{\beta} = \frac{1}{n} \sum_{i=1}^n (\hat{\gamma}_1(\varphi(x_i, \hat{\alpha})) - \hat{\gamma}_0(\varphi(x_i, \hat{\alpha})))$$

In the setup of Section 3 we have $z_i = x_i$, $x_i = d_i$, $y_i = y_i$ and the first step parameter is α , the second step parameters are γ_1, γ_2 and the third step parameter is the ATE β . To derive the influence function for the ATE, we define

$$\begin{aligned} \mu(v_*) &= (\mathbb{E}[y | d = 1, v_*], \mathbb{E}[y | d = 0, v_*])' \\ \gamma(v^*; \alpha) &= (\mathbb{E}[y | d = 1, \varphi(x, \alpha) = v^*], \mathbb{E}[y | d = 0, \varphi(x, \alpha) = v^*])' \\ m(\varphi(x, \alpha_1), \alpha_2, \gamma) &= \gamma_1(\varphi(x, \alpha_1); \alpha_2) - \gamma_2(\varphi(x, \alpha_1); \alpha_2) \end{aligned}$$

Note that we have $\mu(v_*) = \gamma(v_*; \alpha_*)$.

With these definitions, we can now write

$$\hat{\beta} = \frac{1}{n} \sum_{i=1}^n m(\varphi(x_i, \hat{\alpha}_1), \hat{\alpha}_2, \hat{\gamma})$$

where $\hat{\alpha}_1 = \hat{\alpha}_2 = \hat{\alpha}$. We keep them separate for accounting purposes, i.e., to indicate the two roles that $\hat{\alpha}$ plays.

We determine the contribution of $\hat{\alpha}_1$, $\hat{\alpha}_2$ and $\hat{\gamma}$ to the influence function separately. The contribution of $\hat{\gamma}$ that can be derived using Newey (1994) is given in Appendix D, so we focus on characterizing the adjustment of the influence function that reflects the contributions of $\hat{\alpha}_1$ and $\hat{\alpha}_2$. In Lemmas 1 and 2 we show that that contribution is equal to

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

Lemma 1

$$\begin{aligned} \mathbb{E} \left[\frac{\partial}{\partial \alpha} m(\varphi(x, \alpha_*), \alpha_*, \gamma) \right] &= \mathbb{E} \left[\frac{\partial}{\partial \alpha_1} m(\varphi(x, \alpha_*), \alpha_*, \gamma) \right] + \mathbb{E} \left[\frac{\partial}{\partial \alpha_2} m(\varphi(x, \alpha_*), \alpha_*, \gamma) \right] \\ &= \mathbb{E} \left[\left(\frac{\partial \mu_1(\varphi(x; \alpha_*))}{\partial v} - \frac{\partial \mu_2(\varphi(x; \alpha_*))}{\partial v} \right) \frac{\partial \varphi(x; \alpha_*)}{\partial \alpha_1} \right] \\ &\quad + \mathbb{E} \left[\frac{\partial \gamma_1(\varphi(x; \alpha_*); \alpha_*)}{\partial \alpha_2} - \frac{\partial \gamma_2(\varphi(x; \alpha_*); \alpha_*)}{\partial \alpha_2} \right] \end{aligned}$$

Lemma 2

$$\begin{aligned} &\mathbb{E} \left[\left(\frac{\partial \mu_1(\varphi(x; \alpha_*))}{\partial v} - \frac{\partial \mu_2(\varphi(x; \alpha_*))}{\partial v} \right) \frac{\partial \varphi(x; \alpha_*)}{\partial \alpha_1} \right] \\ &\quad + \mathbb{E} \left[\frac{\partial \gamma_1(\varphi(x; \alpha_*); \alpha_*)}{\partial \alpha_2} - \frac{\partial \gamma_2(\varphi(x; \alpha_*); \alpha_*)}{\partial \alpha_2} \right] \\ &= -\mathbb{E} \left[\left(\frac{\mathbb{E}[y|x, d=1] - \mu_1(\varphi(x; \alpha_*))}{\varphi(x; \alpha_*)} + \frac{\mathbb{E}[y|x, d=0] - \mu_2(\varphi(x; \alpha_*))}{1 - \varphi(x; \alpha_*)} \right) \frac{\partial \varphi(x; \alpha_*)}{\partial \alpha} \right] \end{aligned}$$

Proof See Appendix C.

5.2 Nonparametric First Step, Nonparametric Second Step

The setup is as in the previous section. The only difference is that the propensity score $\Pr(d=1|x) = \varphi_*(x) = v_*$ is nonparametric. We define for $v = \varphi(x)$

$$\begin{aligned} \gamma(v^*; v) &= (\mathbb{E}[y | d=1, \varphi(x) = v^*], \mathbb{E}[y | d=0, \varphi(x) = v^*])' \\ m(v_1, v_2, \gamma) &= \gamma_1(v_1; v_2) - \gamma_2(v_1; v_2) \end{aligned}$$

Note that $\gamma(v_1; v_2)$ solves the minimization problem

$$\min_{s_1, s_2} \mathbb{E} \left[d(y - s_1(\varphi(x)))^2 + (1-d)(y - s_2(\varphi(x)))^2 \right]$$

so that for all $(\tilde{s}_1(\varphi(x)), \tilde{s}_2(\varphi(x)))$

$$\mathbb{E} [d(y - \gamma_1(\varphi(x); \varphi)) \tilde{s}_1(\varphi(x)) + (1-d)(y - \gamma_2(\varphi(x); \varphi)) \tilde{s}_2(\varphi(x))] = 0$$

If we choose

$$\begin{aligned} \tilde{s}_1(\varphi(x)) &= \frac{1}{\varphi(x)} \\ \tilde{s}_2(\varphi(x)) &= -\frac{1}{1 - \varphi(x)} \end{aligned}$$

we find

$$\mathbb{E} \left[\frac{dy}{\varphi(x)} - \frac{(1-d)y}{1-\varphi(x)} \right] = \mathbb{E} \left[\frac{d\gamma_1(\varphi(x); v)}{\varphi(x)} - \frac{(1-d)\gamma_2(\varphi(x); v)}{1-\varphi(x)} \right]$$

We will now consider a path φ_α with $\varphi_{\alpha_*} = \varphi_*$. Using Lemma 2, we obtain

$$\begin{aligned} & \mathbb{E} \left[\frac{\partial}{\partial \alpha} m(\varphi(x, \alpha_*), \alpha_*, \gamma) \right] \\ &= -\frac{\partial}{\partial \alpha} \mathbb{E} \left[\left(\frac{\mathbb{E}[y|x, d=1] - \mu_1(\varphi_*(x))}{\varphi_*(x)} + \frac{\mathbb{E}[y|x, d=0] - \mu_0(\varphi_*(x))}{1-\varphi_*(x)} \right) \varphi(x) \varphi_\alpha \right] \end{aligned}$$

and the expectation on the right hand side is linear in φ_α . As in Section 3.2, Proposition 4 of Newey (1994) implies that the adjustment to the influence function for the estimation of φ is

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

which can be alternatively written as

$$\begin{aligned} & -\frac{\mathbb{E}[y|x, d=1] - \mathbb{E}[y|\varphi_*(x), d=1]}{\varphi_*(x)} d + (\mathbb{E}[y|x, d=1] - \mathbb{E}[y|\varphi_*(x), d=1]) \\ & + \frac{\mathbb{E}[y|x, d=0] - \mathbb{E}[y|\varphi_*(x), d=0]}{1-\varphi_*(x)} (1-d) - (\mathbb{E}[y|x, d=0] - \mathbb{E}[y|\varphi_*(x), d=0]) \end{aligned} \quad (6)$$

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 D and is equal to

$$\begin{aligned} & (\mathbb{E}[y|\varphi_*(x), d=1] - \mathbb{E}[y|\varphi_*(x), d=0] - \beta_*) + \\ & \frac{d}{\varphi_*(x)} (y - \mathbb{E}[y|\varphi_*(x), d=1]) - \frac{1-d}{1-\varphi_*(x)} (y - \mathbb{E}[y|\varphi_*(x), d=0]) \end{aligned} \quad (7)$$

Adding (6) and (7), we obtain the influence function of the estimator based on regressions on the estimated propensity score:

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

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

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

with $h_1(x) = \mathbb{E}[y|x, d=1]$, $h_2(x) = \mathbb{E}[y|x, d=0]$. 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 nonparametrically estimated propensity score does not result in an efficiency gain.

5.3 Approximating the Influence Function for the Nonparametric First Step with a Parametric First Step

We assume that

$$\varphi_*(x) = \mathbb{E}[d|x] = p(x)' \alpha_* = \varphi(x, \alpha_*)$$

where $p(x)$ is a large but still finite dimensional vector valued function of x . We can think of this expression as a series approximation with basis functions $p(x)$. The influence function for the least squares estimator of α is

$$\sqrt{n}(\hat{\alpha} - \alpha) = \sqrt{n} \left(\mathbb{E}[p(x)p(x)'] \right)^{-1} p(x) (d - \varphi_*(x)) \quad (8)$$

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

$$\begin{aligned} & - \mathbb{E} \left[\left(\frac{\mathbb{E}[y|x, d=1] - \mathbb{E}[y|\varphi(x, \alpha_*), d=1]}{\varphi(x, \alpha_*)} + \frac{\mathbb{E}[y|x, d=0] - \mathbb{E}[y|\varphi(x, \alpha_*), d=0]}{1 - \varphi(x, \alpha_*)} \right) \frac{\partial \varphi(x, \alpha_*)}{\partial \alpha'} \right] \sqrt{n}(\hat{\alpha} - \alpha) \\ & = - \mathbb{E}[\Psi(x)p(x)'] \sqrt{n}(\hat{\alpha} - \alpha) \end{aligned} \quad (9)$$

where

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

for simplicity. Combining (8) and (9), we conclude that the adjustment to the influence function can be written as

$$- \mathbb{E}[\Psi(x)p(x)'] \left(\mathbb{E}[p(x)p(x)'] \right)^{-1} p(x) (d - \varphi_*(x)) \quad (10)$$

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

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

where $\Pi(\cdot|p(x))$ denotes the projection on the linear space spanned by $p(x)$. If the dimension of $p(x)$ is sufficiently large, then approximately $\Pi(\Psi(x)|p(x)) \approx \mathbb{E}[\Psi(x)|x] = \Psi(x)$. It follows that the adjustment to the influence function in (10) is

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

which is the result in the previous section.

6 A Semiparametric 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 $w_i = (y_i, x_i, z_i)$, $i = 1, \dots, n$, and an estimator of a parameter β has the following three steps:

1. Estimate a finite dimensional parameter $\hat{\alpha}$ by nonlinear least squares of x on $\psi(z, \alpha)$ and obtain the residuals $\hat{v} = x - \psi(z, \hat{\alpha}) = \varphi(x, z, \hat{\alpha})$ that are our generated regressors.
2. Estimate $\mu(x, v_*) = \mathbb{E}[y | x, v_*]$ nonparametrically using the sample $(y_i, x_i, \hat{v}_i), i = 1, \dots, n$. Call the estimator $\hat{\mu}(x, \hat{v})$. Let $L(x) = \mathbb{E}_{v_*}[\mu(x, v_*)]$ and $\hat{L}(x) = \frac{1}{n} \sum_{j=1}^n \hat{\mu}(x, \hat{v}_j)$.
3. Assume that $L(x) = R(x, \beta_*)$ for a known function R and define $\hat{\beta}$ as the solution of the minimization problem

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

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

Let $\hat{\beta}$ denote the solution to the preceding minimization problem, which solves the moment condition

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

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

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

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

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

where an integral with respect to \hat{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 $\varphi(x, z)$. To distinguish this from the x over which we average separately, we use the notation $\tilde{z} = (x, z)$.

With these definitions, we can now write

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

where $\hat{\alpha}_1 = \hat{\alpha}_2 = \hat{\alpha}$ but written separately for accounting purposes. As in Section 4, we will evaluate the adjustment to the influence function term-by-term, i.e. one parameter at a time with the other parameters at their population values. The contribution of $\hat{\gamma}$ and \hat{F}_{xz} can be derived as in Newey (1994) and by the V-statistic projection theorem, respectively, and we concentrate on the contribution of $\hat{\alpha}$.

For the contribution of $\hat{\alpha}_1$ we compute

$$\begin{aligned} \mathbb{E}_x \left[\int \frac{\partial}{\partial \alpha} \gamma(x, \varphi(\tilde{z}, \alpha); \alpha_*) r(x) dF_{xz}(\tilde{z}) \right] \Big|_{\alpha=\alpha_*} &= \mathbb{E}_x \left[\int \frac{\partial}{\partial \alpha} \mu(x, \varphi(\tilde{z}, \alpha)) r(x) dF_{xz}(\tilde{z}) \right] \\ &= -\mathbb{E}_x \left[\int \frac{\partial \mu(x, \varphi(\tilde{z}, \alpha_*))}{\partial v} \left(-\frac{\partial \psi(\tilde{z}, \alpha_*)}{\partial \alpha} \right) r(x) dF_{xz}(\tilde{z}) \right] \equiv \Xi_1 \end{aligned}$$

To obtain the contribution of $\widehat{\alpha}_2$ we first observe that

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

Because $\gamma(x, \varphi(x, z, \alpha_1) ; \alpha_2)$ solves the minimization problem $\min_s \mathbb{E} \left[(y - s(x, \varphi(x, z, \alpha_1)))^2 \right]$ so that

$$0 = \mathbb{E}[(y - \gamma(x, \varphi(x, z, \alpha_1) ; \alpha_2)) s(x, \varphi(x, z, \alpha_1))]$$

for all square integrable function $s(x, \varphi(x, z, \alpha_1))$. In particular, we should have

$$0 = \mathbb{E} \left[(y - \gamma(x, \varphi(x, z, \alpha_1) ; \alpha_2)) r(x) \frac{f(x) f(\varphi(x, z, \alpha_1))}{f(x, \varphi(x, z, \alpha_1))} \right],$$

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

$$\begin{aligned}\mathbb{E} \left[\frac{\partial \gamma(x, v_* ; \alpha_*)}{\partial \alpha_2} r(x) \frac{f(x) f(v_*)}{f(x, v_*)} \right] &= \mathbb{E} \left[(y - \gamma(x, v_* ; \alpha_*)) r(x) \frac{\partial}{\partial v} \left(\frac{f(x) f(v_*)}{f(x, v_*)} \right) \frac{\partial \varphi(x, z, \alpha_*)}{\partial \alpha_1} \right] \\ &\quad - \mathbb{E} \left[\frac{\partial \gamma(x, v_* ; \alpha_*)}{\partial v} r(x) \frac{\partial \varphi(x, z, \alpha_*)}{\partial \alpha_1} \frac{f(x) f(v_*)}{f(x, v_*)} \right]\end{aligned}$$

We therefore obtain

$$\begin{aligned}\frac{\partial}{\partial \alpha} \mathbb{E} \left[\gamma(x, v_* ; \alpha_*) \frac{f(x) f(v_*)}{f(x, v_*)} \right] &= -\mathbb{E} \left[(y - \mu(x, v_*)) r(x) \frac{\partial}{\partial v} \left(\frac{f(x) f(v_*)}{f(x, v_*)} \right) \frac{\partial \psi(z, \alpha_*)}{\partial \alpha} \right] \\ &\quad + \mathbb{E} \left[\frac{\partial \mu(x, v_*)}{\partial v} r(x) \frac{\partial \psi(z, \alpha_*)}{\partial \alpha} \frac{f(x) f(v_*)}{f(x, v_*)} \right] = \Xi_2\end{aligned}$$

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 nonparametric regressions on a regressor that is estimated in the first step. The first step estimator is either parametric or nonparametric. Although we heavily use Newey's (1994) approach and no results beyond that paper are needed, the application of those results in the type of three step estimators that we consider is not trivial. Published results not always seem to account for the first step estimation in the proper way. The three examples that we study in detail are interesting in their own right, but it should be emphasized that our results can easily be adapted to other three step estimators.

Appendix

A Proof of (1)

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

$$\mathbb{E} \begin{bmatrix} y - \gamma_1 - \gamma_2 x - \gamma_3 \varphi(x, z, \alpha) \\ x(y - \gamma_1 - \gamma_2 x - \gamma_3 \varphi(x, z, \alpha)) \\ \varphi(x, z, \alpha)(y - \gamma_1 - \gamma_2 x - \gamma_3 \varphi(x, z, \alpha)) \\ \psi(x, z) - \alpha \end{bmatrix} = 0$$

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

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

where

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

$$G_\gamma = -\mathbb{E} \begin{bmatrix} 1 & x & \varphi(x, z, \alpha_*) \\ x & x^2 & x\varphi(x, z, \alpha_*) \\ \varphi(x, z, \alpha_*) & x\varphi(x, z, \alpha_*) & \varphi(x, z, \alpha_*)^2 \end{bmatrix}$$

and

$$G_\alpha = -\mathbb{E} \begin{bmatrix} \gamma_{3*} \frac{\partial \varphi(x, z, \alpha_*)}{\partial \alpha} \\ \gamma_{3*} x \frac{\partial \varphi(x, z, \alpha_*)}{\partial \alpha} \\ 2\gamma_{3*} \varphi(x, z, \alpha_*) \frac{\partial \varphi(x, z, \alpha_*)}{\partial \alpha} \end{bmatrix}$$

Likewise, we obtain from the population moment equation

$$\mathbb{E} \begin{bmatrix} y - \gamma_1 - \gamma_2 x - \gamma_3 \varphi(x, z, \alpha_*) \\ x(y - \gamma_1 - \gamma_2 x - \gamma_3 \varphi(x, z, \alpha_*)) \\ \varphi(x, z, \alpha_*) (y - \gamma_1 - \gamma_2 x - \gamma_3 \varphi(x, z, \alpha_*)) \end{bmatrix} = 0$$

that

$$\sqrt{n}(\tilde{\gamma} - \gamma_*) = -G_\gamma^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \begin{bmatrix} \varepsilon_i \\ x_i \varepsilon_i \\ \varphi(x_i, z_i, \alpha_*) \varepsilon_i \end{bmatrix} + o_p(1)$$

It follows that

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

Now note that

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

and therefore,

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

It follows that

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

B Interpretation of (3)

In order to understand the additional term

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

we examine

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

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

$$\mathbb{E} \begin{bmatrix} y - \gamma_1(\alpha) - \gamma_2(\alpha)x - \gamma_3(\alpha)\varphi(x, z, \alpha) \\ x(y - \gamma_1(\alpha) - \gamma_2(\alpha)x - \gamma_3(\alpha)\varphi(x, z, \alpha)) \\ \varphi(x, z, \alpha)(y - \gamma_1(\alpha) - \gamma_2(\alpha)x - \gamma_3(\alpha)\varphi(x, z, \alpha)) \end{bmatrix} = 0$$

which holds for all α , we can use the implicit function theorem to derive

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

It follows that

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

so that

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

C Proof of Lemmas in Section 5

Proof of Lemma 1

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

Proof of Lemma 2 Note that $\gamma(\varphi(x, \alpha); \alpha)$ solves the minimization problem

$$\min_{s_1, s_2} \mathbb{E} \left[d(y - s_1(\varphi(x, \alpha)))^2 + (1-d)(y - s_2(\varphi(x, \alpha)))^2 \right]$$

so that

$$\mathbb{E} [d(y - \gamma_1(\varphi(x, \alpha); \alpha)) \tilde{s}_1(\varphi(x, \alpha)) + (1-d)(y - \gamma_2(\varphi(x, \alpha); \alpha)) \tilde{s}_2(\varphi(x, \alpha))] = 0$$

for all functions $(\tilde{s}_1(\varphi(x, \alpha)), \tilde{s}_2(\varphi(x, \alpha)))'$. In particular, this should hold for

$$\begin{aligned}
\tilde{s}_1(\varphi(x, \alpha)) &= \frac{1}{\varphi(x, \alpha)} \\
\tilde{s}_2(\varphi(x, \alpha)) &= -\frac{1}{1 - \varphi(x, \alpha)}
\end{aligned}$$

or

$$\mathbb{E} \left[\frac{dy}{\varphi(x, \alpha)} - \frac{(1-d)y}{1 - \varphi(x, \alpha)} \right] = \mathbb{E} \left[\frac{d\gamma_1(\varphi(x, \alpha); \alpha)}{\varphi(x, \alpha)} - \frac{(1-d)\gamma_2(\varphi(x, \alpha); \alpha)}{1 - \varphi(x, \alpha)} \right] \quad (11)$$

Noting that

$$\begin{aligned}
\mathbb{E} \left[\frac{dy}{\varphi(x, \alpha)} \right] &= \mathbb{E} \left[\frac{\mathbb{E}[dy|x]}{\varphi(x, \alpha)} \right] = \mathbb{E} \left[\frac{\mathbb{E}[d|x] \mathbb{E}[y|x, d=1]}{\varphi(x, \alpha)} \right] \\
&= \mathbb{E} \left[\frac{\varphi(x, \alpha_*) \mathbb{E}[y|x, d=1]}{\varphi(x, \alpha)} \right],
\end{aligned}$$

we obtain

$$\frac{\partial}{\partial \alpha} \mathbb{E} \left[\frac{dy}{\varphi(x, \alpha)} \right] \Big|_{\alpha=\alpha_*} = -\mathbb{E} \left[\frac{\mathbb{E}[y|x, d=1]}{\varphi(x, \alpha_*)} \frac{\partial \varphi(x, \alpha_*)}{\partial \alpha} \right] \quad (12)$$

Analogously we obtain

$$\frac{\partial}{\partial \alpha} \mathbb{E} \left[\frac{(1-d)y}{1-\varphi(x, \alpha)} \right] \Big|_{\alpha=\alpha_*} = \mathbb{E} \left[\frac{\mathbb{E}[y|x, d=0]}{1-\varphi(x, \alpha_*)} \frac{\partial \varphi(x, \alpha_*)}{\partial \alpha} \right] \quad (13)$$

Combining (12) and (13), we have

$$\frac{\partial}{\partial \alpha} \mathbb{E} \left[\frac{dy}{\varphi(x, \alpha)} - \frac{(1-d)y}{1-\varphi(x, \alpha)} \right] \Big|_{\alpha=\alpha_*} = -\mathbb{E} \left[\left(\frac{\mathbb{E}[y|x, d=1]}{\varphi(x, \alpha_*)} + \frac{\mathbb{E}[y|x, d=0]}{1-\varphi(x, \alpha_*)} \right) \frac{\partial \varphi(x, \alpha_*)}{\partial \alpha} \right] \quad (14)$$

Now we note that

$$\begin{aligned} \frac{\partial}{\partial \alpha} \mathbb{E} \left[\frac{d\gamma_1(\varphi(x, \alpha); \alpha)}{\varphi(x, \alpha)} \right] \Big|_{\alpha=\alpha_*} &= \mathbb{E} \left[\frac{\partial}{\partial \alpha} \left(\frac{\varphi(x, \alpha_*) \gamma_1(\varphi(x, \alpha); \alpha)}{\varphi(x, \alpha)} \right) \right] \Big|_{\alpha=\alpha_*} \\ &= \mathbb{E} \left[\frac{\partial \gamma_1(\varphi(x, \alpha_*); \alpha_*)}{\partial v} \frac{\partial \varphi(x, \alpha_*)}{\partial \alpha_1} \right] \\ &\quad + \mathbb{E} \left[\frac{\partial \gamma_1(\varphi(x, \alpha_*); \alpha_*)}{\partial \alpha_2} \right] \\ &\quad - \mathbb{E} \left[\frac{\gamma_1(\varphi(x, \alpha_*); \alpha_*)}{\varphi(x, \alpha_*)} \frac{\partial \varphi(x, \alpha_*)}{\partial \alpha_1} \right] \end{aligned} \quad (15)$$

and likewise

$$\begin{aligned} \frac{\partial}{\partial \alpha} \mathbb{E} \left[\frac{(1-d)\gamma_2(\varphi(x, \alpha); \alpha)}{1-\varphi(x, \alpha)} \right] \Big|_{\alpha=\alpha_*} &= \mathbb{E} \left[\frac{\partial \gamma_2(\varphi(x, \alpha_*); \alpha_*)}{\partial v} \frac{\partial \varphi(x, \alpha_*)}{\partial \alpha_1} \right] \\ &\quad + \mathbb{E} \left[\frac{\partial \gamma_2(\varphi(x, \alpha_*); \alpha_*)}{\partial \alpha_2} \right] \\ &\quad + \mathbb{E} \left[\frac{\gamma_2(\varphi(x, \alpha_*); \alpha_*)}{1-\varphi(x, \alpha_*)} \frac{\partial \varphi(x, \alpha_*)}{\partial \alpha_1} \right] \end{aligned} \quad (16)$$

Combining (15) and (16), we obtain

$$\begin{aligned} &\frac{\partial}{\partial \alpha} \mathbb{E} \left[\frac{d\gamma_1(\varphi(x, \alpha); \alpha)}{\varphi(x, \alpha)} - \frac{(1-d)\gamma_2(\varphi(x, \alpha); \alpha)}{1-\varphi(x, \alpha)} \right] \Big|_{\alpha=\alpha_*} \\ &= \mathbb{E} \left[\left(\frac{\partial \mu_1(\varphi(x, \alpha_*))}{\partial v} - \frac{\partial \mu_2(\varphi(x, \alpha_*))}{\partial v} \right) \frac{\partial \varphi(x, \alpha_*)}{\partial \alpha_1} \right] \\ &\quad + \mathbb{E} \left[\frac{\partial \gamma_1(\varphi(x, \alpha_*); \alpha_*)}{\partial \alpha_2} - \frac{\partial \gamma_2(\varphi(x, \alpha_*); \alpha_*)}{\partial \alpha_2} \right] \\ &\quad - \mathbb{E} \left[\left(\frac{\mathbb{E}[y|\varphi(x, \alpha_*), d=1]}{\varphi(x, \alpha_*)} + \frac{\mathbb{E}[y|\varphi(x, \alpha_*), d=0]}{1-\varphi(x, \alpha_*)} \right) \frac{\partial \varphi(x, \alpha_*)}{\partial \alpha_1} \right] \end{aligned} \quad (17)$$

By (11) the left hand side of (17) is equal to the right hand side of (14), and upon substitution and rearranging the resulting expression we conclude that

$$\begin{aligned} &\mathbb{E} \left[\left(\frac{\partial \mu_1(\varphi(x, \alpha_*))}{\partial v} - \frac{\partial \mu_2(\varphi(x, \alpha_*))}{\partial v} \right) \frac{\partial \varphi(x, \alpha_*)}{\partial \alpha_1} \right] + \mathbb{E} \left[\frac{\partial \gamma_1(\varphi(x, \alpha_*); \alpha_*)}{\partial \alpha_2} - \frac{\partial \gamma_2(\varphi(x, \alpha_*); \alpha_*)}{\partial \alpha_2} \right] \\ &= -\mathbb{E} \left[\left(\frac{\mathbb{E}[y|x, d=1] - \mathbb{E}[y|\varphi(x, \alpha_*), d=1]}{\varphi(x, \alpha_*)} + \frac{\mathbb{E}[y|x, d=0] - \mathbb{E}[y|\varphi(x, \alpha_*), d=0]}{1-\varphi(x, \alpha_*)} \right) \frac{\partial \varphi(x, \alpha_*)}{\partial \alpha_1} \right] \end{aligned}$$

D The Influence Function of the Imputation Estimator

The ATE is

$$\beta_* = \mathbb{E}[h_1(x) - h_2(x)]$$

with

$$\begin{aligned} h_1(x) &= \mathbb{E}[y | d = 1, x] \\ h_2(x) &= \mathbb{E}[y | d = 0, x] \end{aligned}$$

The ATE satisfies the moment equation

$$0 = \mathbb{E}[m(x, \beta_*, h_1, h_2)]$$

where

$$m(x, \beta_*, h_1, h_2) = h_1(x) - h_2(x) - \beta_*$$

The imputation estimator for the ATE is

$$\hat{\beta} = \frac{1}{n} \sum_{i=1}^n (\hat{h}_1(x_i) - \hat{h}_2(x_i))$$

We are interested in

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (\hat{h}_1(x_i) - \hat{h}_2(x_i))$$

so that we need to consider

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

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

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}_θ denotes an expectation with respect to the distribution with density $f(x, \theta)$, then we define the corresponding paths for the projections $h_1(x, \theta) = \mathbb{E}_\theta[y | x, d = 1]$ and $h_2(x, \theta) = \mathbb{E}_\theta[y | x, d = 0]$. The path $h(x, \theta)$ is the minimizer of a *single* objective function

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

so that the following orthogonality condition holds

$$\mathbb{E}_\theta \left[d \left(y - h_1(x, \theta) \right) \tilde{h}_1(x) + (1 - d) \left(y - h_2(x, \theta) \right) \tilde{h}_2(x) \right] = 0$$

for all functions $(\tilde{h}_1(x), \tilde{h}_2(x))'$. Choose $(\tilde{h}_1(x), \tilde{h}_2(x)) = \left(\frac{1}{\varphi_*(x)}, -\frac{1}{1-\varphi_*(x)} \right)$, i.e.,

$$\mathbb{E}_\theta \left[\frac{d}{\varphi_*(x)} \left(y - h_1(x, \theta) \right) - \frac{1-d}{1-\varphi_*(x)} \left(y - h_2(x, \theta) \right) \right] = 0 \quad (18)$$

or

$$\begin{aligned} \mathbb{E}_\theta \left[\frac{d}{\varphi_*(x)} y - \frac{1-d}{1-\varphi_*(x)} y \right] &= \mathbb{E}_\theta \left[\frac{d}{\varphi_*(x)} h_1(x, \theta) - \frac{1-d}{1-\varphi_*(x)} h_2(x, \theta) \right] \\ &= \mathbb{E}_\theta [h_1(x, \theta) - h_2(x, \theta)] \end{aligned} \quad (19)$$

The final expression is useful to compute the derivative as in Newey (1994), equation (4.5). By the chain rule (evaluate the derivatives at $\theta = 0$)

$$\begin{aligned} \frac{\partial \mathbb{E}_\theta \left[\frac{d}{\varphi_*(x)} g_1(x, \theta) - \frac{1-d}{1-\varphi_*(x)} g_2(x, \theta) \right]}{\partial \theta} &= \frac{\partial \mathbb{E}_\theta \left[\frac{d}{\varphi_*(x)} h_1(x) - \frac{1-d}{1-\varphi_*(x)} h_2(x) \right]}{\partial \theta} \\ &+ \frac{\partial \mathbb{E} \left[\frac{d}{\varphi_*(x)} h_1(x, \theta) - \frac{1-d}{1-\varphi_*(x)} h_2(x, \theta) \right]}{\partial \theta} \end{aligned}$$

where we use the fact that the derivative of the projection paths at $\theta = 0$ are equal to h_1, h_2 . Therefore combining this with the result above

$$\begin{aligned} \frac{\partial \mathbb{E} [D(x)' h(x, \theta)]}{\partial \theta} &= \frac{\partial \mathbb{E} [h_1(x, \theta) - h_2(x, \theta)]}{\partial \theta} \\ &= \frac{\partial \mathbb{E} \left[\frac{d}{\varphi_*(x)} h_1(x, \theta) - \frac{1-d}{1-\varphi_*(x)} h_2(x, \theta) \right]}{\partial \theta} \\ &= \frac{\partial \mathbb{E}_\theta \left[\frac{d}{\varphi_*(x)} h_1(x, \theta) - \frac{1-d}{1-\varphi_*(x)} h_2(x, \theta) \right]}{\partial \theta} - \frac{\partial \mathbb{E}_\theta \left[\frac{d}{\varphi_*(x)} h_1(x) - \frac{1-d}{1-\varphi_*(x)} h_2(x) \right]}{\partial \theta} \\ &= \frac{\partial}{\partial \theta} \left(\mathbb{E}_\theta \left[\frac{d}{\varphi_*(x)} y - \frac{1-d}{1-\varphi_*(x)} y \right] - \mathbb{E}_\theta \left[\frac{d}{\varphi_*(x)} h_1(x) - \frac{1-d}{1-\varphi_*(x)} h_2(x) \right] \right) \end{aligned}$$

so that at $\theta = 0$

$$\begin{aligned} \frac{\partial \mathbb{E} [D(x)' h(x, \theta)]}{\partial \theta} &= \frac{\partial}{\partial \theta} \mathbb{E}_\theta \left[\frac{d}{\varphi_*(x)} (y - h_1(x)) - \frac{1-d}{1-\varphi_*(x)} (y - h_2(x)) \right] \\ &= \mathbb{E} \left[\left(\frac{d}{\varphi_*(x)} (y - h_1(x)) - \frac{1-d}{1-\varphi_*(x)} (y - h_2(x)) \right) S(y, d, x) \right], \end{aligned}$$

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

$$\frac{d}{\varphi_*(x)} (y - h_1(x)) - \frac{1-d}{1-\varphi_*(x)} (y - h_2(x))$$

and the influence function of the imputation estimator is

$$(h_1(x) - h_2(x) - \beta_*) + \frac{d}{\varphi_*(x)} (y - h_1(x)) - \frac{1-d}{1-\varphi_*(x)} (y - h_2(x)) \quad (20)$$

so this estimator is efficient.

The ATE is also equal to

$$\beta_* = \mathbb{E} [h_1(\varphi_*(x)) - h_2(\varphi_*(x))]$$

with

$$\begin{aligned} h_1(x) &= \mathbb{E} [y | d = 1, \varphi_*(x)] \\ h_2(x) &= \mathbb{E} [y | d = 0, \varphi_*(x)] \end{aligned}$$

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

$$(h_1(\varphi_*(x)) - h_2(\varphi_*(x)) - \beta_*) + \frac{d}{\varphi_*(x)} (y - h_1(\varphi_*(x))) - \frac{1-d}{1-\varphi_*(x)} (y - h_2(\varphi_*(x))) \quad (21)$$

The asymptotic variances implied by (20) and (21) are

$$\mathbb{E} \left[(\beta(x) - \beta_*)^2 + \frac{\text{Var}(y_1|x)}{\varphi_*(x)} + \frac{\text{Var}(y_0|x)}{1 - \varphi_*(x)} \right] \quad (22)$$

and

$$\mathbb{E} \left[(\beta(\varphi_*(x)) - \beta_*)^2 + \frac{(y_1 - h_1(\varphi_*(x)))^2}{\varphi_*(x)} + \frac{(y_0 - h_0(\varphi_*(x)))^2}{1 - \varphi_*(x)} \right] \quad (23)$$

where $\beta(x) = h_1(x) - h_2(x)$ and $\beta(\varphi_*(x)) = h_1(\varphi_*(x)) - h_2(\varphi_*(x)) = \mathbb{E}[\beta(x) | \varphi_*(x)]$. Using

$$\begin{aligned} \mathbb{E} \left[(y_1 - h_1(\varphi_*(x)))^2 \middle| x \right] &= E \left[((y_1 - h_1(x)) + (h_1(x) - h_1(\varphi_*(x))))^2 \middle| x \right] \\ &= \text{Var}(y_1|x) + (h_1(x) - h_1(\varphi_*(x)))^2 \end{aligned}$$

$$\mathbb{E} \left[(y_0 - h_0(\varphi_*(x)))^2 \middle| x \right] = \text{Var}(y_0|x) + (h_0(x) - h_0(\varphi_*(x)))^2$$

and

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

we note that

$$\begin{aligned} \mathbb{E} \left[\frac{(y_1 - h_1(\varphi_*(x)))^2}{\varphi_*(x)} \right] &= \mathbb{E} \left[\frac{\text{Var}(y_1|x)}{\varphi_*(x)} \right] + \mathbb{E} \left[\frac{(h_1(x) - h_1(\varphi_*(x)))^2}{\varphi_*(x)} \right] \\ \mathbb{E} \left[\frac{(y_0 - h_0(\varphi_*(x)))^2}{1 - \varphi_*(x)} \right] &= \mathbb{E} \left[\frac{\text{Var}(y_0|x)}{1 - \varphi_*(x)} \right] + \mathbb{E} \left[\frac{(h_0(x) - h_0(\varphi_*(x)))^2}{1 - \varphi_*(x)} \right] \\ \mathbb{E} \left[(\beta(x) - \beta_*)^2 \right] &= \mathbb{E} \left[(\beta(x) - \beta(\varphi_*(x)))^2 \right] + \mathbb{E} \left[(\beta(\varphi_*(x)) - \beta_*)^2 \right]. \end{aligned}$$

Therefore, we can see that the difference of (23) and (22) is equal to

$$\begin{aligned} &\mathbb{E} \left[\frac{(h_1(x) - h_1(\varphi_*(x)))^2}{\varphi_*(x)} + \frac{(h_0(x) - h_0(\varphi_*(x)))^2}{1 - \varphi_*(x)} \right] - \mathbb{E} \left[(\beta(x) - \beta(\varphi_*(x)))^2 \right] \\ &= \mathbb{E} \left[\frac{a(x)^2}{\varphi_*(x)} + \frac{b(x)^2}{1 - \varphi_*(x)} - (a(x) - b(x))^2 \right] \end{aligned}$$

for $a(x) = h_1(x) - h_1(\varphi_*(x))$ and $b(x) = h_0(x) - h_0(\varphi_*(x))$. Therefore, the difference of (23) and (22) is equal to

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

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

E Additional Results on Influence Functions with Multiple Non-parametric Components

E.1 Newey's Approach with Multiple Nonparametric Components

Although Newey's (1994) analysis is general enough to cover this topic, his analysis on pp. 1360 – 61 does not explicitly deal with multiple non-parametric objects with potentially different regressors. It is therefore useful to spell out exactly what is needed to deal with the situation.

Suppose that we are dealing with $m(z, h_1(x_1), \dots, h_J(x_J))$. Note that m depends on h 's only through their values. Using his equation (3.17), we have

$$\begin{aligned} \frac{\partial \mathbb{E}[m(z, h_1(\theta), \dots, h_J(\theta))]}{\partial \theta} &= \sum_{j=1}^J \mathbb{E} \left[D_j(z) \frac{\partial h_j(x_j, \theta)}{\partial \theta} \right] \\ &= \frac{\partial}{\partial \theta} \mathbb{E} \left[\sum_{j=1}^J D_j(z) h_j(x_j, \theta) \right] \\ &= \mathbb{E}[D(z, h_1(\theta), \dots, h_J(\theta))] \end{aligned}$$

for

$$D_j(z) = \left. \frac{\partial m(z, h_1, \dots, h_J)}{\partial h_j} \right|_{h_1=h_1(x_1), \dots, h_J=h_J(x_J)}$$

Note that

$$\mathbb{E}[D(z, \tilde{g}_1, \dots, \tilde{g}_J)] = \mathbb{E} \left[\sum_{j=1}^J D_j(z) \tilde{g}_j(x_j) \right] = \mathbb{E} \left[\sum_{j=1}^J \delta_j(x_j) \tilde{g}_j(x_j) \right]$$

for

$$\delta_j(x_j) = \mathbb{E}[D_j(z) | x_j]$$

Let $g_j(x_j, \theta) = \operatorname{argmin}_{\tilde{g}_j} \mathbb{E}_\theta [(y_j - \tilde{g}_j(x_j))^2]$ for a path. Note that $\delta_j(x_j)$ satisfies the orthogonality $\mathbb{E}_\theta [(y_j - g_j(x_j, \theta)) \delta_j(x_j)] = 0$, or

$$\mathbb{E}_\theta [\delta_j(x_j) g_j(x_j, \theta)] = \mathbb{E}_\theta [\delta_j(x_j) y_j]$$

Then by the chain rule,

$$\begin{aligned} \frac{\partial \mathbb{E}[D(z, h_1(\theta), \dots, h_J(\theta))]}{\partial \theta} &= \frac{\partial \mathbb{E} \left[\sum_{j=1}^J \delta_j(x_j) g_j(x_j, \theta) \right]}{\partial \theta} \\ &= \frac{\partial \mathbb{E}_\theta \left[\sum_{j=1}^J \delta_j(x_j) g_j(x_j, \theta) \right]}{\partial \theta} - \frac{\partial \mathbb{E}_\theta \left[\sum_{j=1}^J \delta_j(x_j) g_j(x_j) \right]}{\partial \theta} \\ &= \frac{\partial \mathbb{E}_\theta \left[\sum_{j=1}^J \delta_j(x_j) y_j \right]}{\partial \theta} - \frac{\partial \mathbb{E}_\theta \left[\sum_{j=1}^J \delta_j(x_j) g_j(x_j) \right]}{\partial \theta} \\ &= \frac{\partial \mathbb{E}_\theta \left[\sum_{j=1}^J \delta_j(x_j) (y_j - g_j(x_j)) \right]}{\partial \theta} \\ &= \mathbb{E} \left[\left(\sum_{j=1}^J \delta_j(x_j) (y_j - g_j(x_j)) \right) S(z) \right] \end{aligned}$$

It follows that the adjustment to the influence function is equal to

$$\sum_{j=1}^J \delta_j(x_j) (y_j - g_j(x_j))$$

This derivation simply verifies that the adjustment can be computed “term by term”.

E.2 Verification by Ai & Chen’s (2007) Asymptotic Variance Formula

Because Newey did not explicitly deal with multiple nonparametric components explicitly, it would be useful for readers’ peace of mind to make sure that the derivation is not flawed. For this purpose, it is useful to look at Ai & Chen (2007). They considered a very general model, but it can be easily mapped into ours:

$$\begin{aligned} \mathbb{E}[y_j - g_j(x_j) | x_j] &= 0, \quad j = 1, \dots, J \\ \mathbb{E}[m(z, g_1(x_1), \dots, g_J(x_J)) - \theta] &= 0 \end{aligned}$$

Their notation is a little messy, so we need to approach it with care. First of all, they are dealing with J moments, whereas we are dealing with $J + 1$ moments. See their equation (1).

Second, using their definition of m_j in the middle of page 9, we come up with

$$\begin{aligned} M_j &= \mathbb{E}[y_j | x_j] - g_j(x_j), \quad j = 1, \dots, J \\ M_{J+1} &= \mathbb{E}[m(z, g_1(x_1), \dots, g_J(x_J))] - \theta \end{aligned}$$

In order to avoid confusion, I used M instead of m .

Third, we need to get the pathwise derivative for the deviation $(\tilde{\theta}, \tilde{g}_1, \dots, \tilde{g}_J)$, which is written $\alpha - \alpha_*$ in Ai & Chen. Using the definition on page 15, we find

$$\begin{aligned} \frac{dM_{J+1}}{dh} [\alpha - \alpha_*] &= \sum_{j=1}^J \mathbb{E} \left[\frac{\partial m(z, g_1(x_1), \dots, g_J(x_J))}{\partial g_j} \tilde{g}_j(x_j) \right] \\ &= \mathbb{E} \left[\sum_{j=1}^J D_j(z) \tilde{g}_j(x_j) \right] \end{aligned}$$

$$\frac{dM_j}{dh} [\alpha - \alpha_*] = -\tilde{g}_j(x_j), \quad j = 1, \dots, J$$

Fourth, we need to find w^* as defined on page 18. (Because θ here is one dimensional, we do not need to work with multiple l s, and the dependence on l is suppressed.) According to their analysis on page 18, all we need to do is to find $w_1^*(x_1), \dots, w_J^*(x_J)$ that minimizes

$$\mathbb{E} \left[\sum_{j=1}^J (w_j(x_j))^2 + \left(\mathbb{E} \left[-1 - \sum_{j=1}^J D_j(z) w_j(x_j) \right] \right)^2 \right] = \left(\mathbb{E} \left[1 + \sum_{j=1}^J D_j(z) w_j(x_j) \right] \right)^2 + \sum_{j=1}^J \mathbb{E} \left[(w_j(x_j))^2 \right]$$

It can be shown that the solution is given by

$$w_j^*(x_j) = - \frac{\mathbb{E}[D_j(z) | x_j]}{1 + \mathbb{E} \left[\sum_{j=1}^J (\mathbb{E}[D_j(z) | x_j])^2 \right]}$$

Proof is provided later.

Fifth, we need to calculate as in their page 19

$$D_{w^*}(x) = \begin{bmatrix} w_1^*(x_1) \\ \vdots \\ w_J^*(x_J) \\ \mathbb{E} \left[-1 - \sum_{j=1}^J D_j(z) w_j^*(x_j) \right] \end{bmatrix}$$

We can ignore their V_{w^*} because there is no misspecification in our model.

Sixth, we calculate as in their Theorem 4.1 on page 21,

$$\mathbb{E} [D_{w^*}(x)' D_{w^*}(x)] = \left(\mathbb{E} \left[1 + \sum_{j=1}^J D_j(z) w_j^*(x_j) \right] \right)^2 + \sum_{j=1}^J \mathbb{E} \left[(w_j^*(x_j))^2 \right]$$

As for the first term on the right, we have

$$\begin{aligned} \mathbb{E} \left[1 + \sum_{j=1}^J D_j(z) w_j^*(x_j) \right] &= \mathbb{E} \left[1 - \frac{\sum_{j=1}^J D_j(z) \mathbb{E} [D_j(z) | x_j]}{1 + \mathbb{E} \left[\sum_{j=1}^J (\mathbb{E} [D_j(z) | x_j])^2 \right]} \right] \\ &= 1 - \frac{\mathbb{E} \left[\sum_{j=1}^J (\mathbb{E} [D_j(z) | x_j])^2 \right]}{1 + \mathbb{E} \left[\sum_{j=1}^J (\mathbb{E} [D_j(z) | x_j])^2 \right]} \\ &= \frac{1}{1 + \mathbb{E} \left[\sum_{j=1}^J (\mathbb{E} [D_j(z) | x_j])^2 \right]} \end{aligned}$$

As for the second term on the right, we have

$$\sum_{j=1}^J \mathbb{E} \left[(w_j^*(x_j))^2 \right] = \frac{\sum_{j=1}^J \mathbb{E} \left[(\mathbb{E} [D_j(z) | x_j])^2 \right]}{\left(1 + \mathbb{E} \left[\sum_{j=1}^J (\mathbb{E} [D_j(z) | x_j])^2 \right] \right)^2}$$

Therefore, we have

$$\begin{aligned} \mathbb{E} [D_{w^*}(x)' D_{w^*}(x)] &= \left(\frac{1}{1 + \mathbb{E} \left[\sum_{j=1}^J (\mathbb{E} [D_j(z) | x_j])^2 \right]} \right)^2 + \frac{\sum_{j=1}^J \mathbb{E} \left[(\mathbb{E} [D_j(z) | x_j])^2 \right]}{\left(1 + \mathbb{E} \left[\sum_{j=1}^J (\mathbb{E} [D_j(z) | x_j])^2 \right] \right)^2} \\ &= \frac{1}{1 + \mathbb{E} \left[\sum_{j=1}^J (\mathbb{E} [D_j(z) | x_j])^2 \right]} \end{aligned}$$

Seventh, we write

$$\rho = \begin{bmatrix} y_1 - g_1(x_1) \\ \vdots \\ y_J - g_J(x_J) \\ m(z, g_1(x_1), \dots, g_J(x_J)) - \theta \end{bmatrix}$$

and

$$\begin{aligned}
D_{w^*}(x)' \rho &= \sum_{j=1}^J w_j^*(x_j) (y_j - g_j(x_j)) + \mathbb{E} \left[-1 - \sum_{j=1}^J D_j(z) w_j^*(x_j) \right] (m - \theta) \\
&= \frac{\sum_{j=1}^J \mathbb{E}[D_j(z)|x_j] (y_j - g_j(x_j))}{1 + \mathbb{E} \left[\sum_{j=1}^J (\mathbb{E}[D_j(z)|x_j])^2 \right]} - \frac{1}{1 + \mathbb{E} \left[\sum_{j=1}^J (\mathbb{E}[D_j(z)|x_j])^2 \right]} (m - \theta) \\
&= \frac{(m - \theta) + \sum_{j=1}^J \mathbb{E}[D_j(z)|x_j] (y_j - g_j(x_j))}{1 + \mathbb{E} \left[\sum_{j=1}^J (\mathbb{E}[D_j(z)|x_j])^2 \right]}
\end{aligned}$$

calculate as in their Theorem 4.1 on page 21

$$\Omega_* = \text{Var} (D_{w^*}(x)' \rho) = \frac{\text{Var} \left((m - \theta) + \sum_{j=1}^J \mathbb{E}[D_j(z)|x_j] (y_j - g_j(x_j)) \right)}{\left(1 + \mathbb{E} \left[\sum_{j=1}^J (\mathbb{E}[D_j(z)|x_j])^2 \right] \right)^2}$$

Finally, using their Theorem 4.1, our asymptotic variance is equal to

$$\begin{aligned}
&\frac{\Omega_*}{\left(\mathbb{E} [D_{w^*}(x)' D_{w^*}(x)] \right)^2} \\
&= \frac{\text{Var} \left((m - \theta) + \sum_{j=1}^J \mathbb{E}[D_j(z)|x_j] (y_j - g_j(x_j)) \right) / \left(1 + \mathbb{E} \left[\sum_{j=1}^J (\mathbb{E}[D_j(z)|x_j])^2 \right] \right)^2}{1 / \left(1 + \mathbb{E} \left[\sum_{j=1}^J (\mathbb{E}[D_j(z)|x_j])^2 \right] \right)^2} \\
&= \text{Var} \left((m - \theta) + \sum_{j=1}^J \mathbb{E}[D_j(z)|x_j] (y_j - g_j(x_j)) \right)
\end{aligned}$$

Now, Inspection of this asymptotic variance formula indicates that it is equivalent to adjusting the influence function $m - \theta$ by $\sum_{j=1}^J \mathbb{E}[D_j(z)|x_j] (y_j - g_j(x_j))$. This is exactly what our earlier derivation would suggest.

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] 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.
- [10] Newey, W.K. (1984): “A Method of Moments Interpretation of Sequential Estimators,” *Economics Letters* 14, pp. 201 – 206.
- [11] Newey, W.K. (1994): “The Asymptotic Variance of Semiparametric Estimators,” *Econometrica* 62, pp. 1349 – 1382.
- [12] Olley, G.S. and A. Pakes (1996): “The Dynamics of Productivity in the Telecommunications Equipment Industry,” *Econometrica* 64, pp. 1263 – 1297.
- [13] Pagan, A. (1984): “Econometric Issues in the Analysis of Regressions with Generated Regressors,” *International Economic Review* 25, pp. 221 – 247.
- [14] Pakes, A. and G.S. Olley (1995): “A Limit Theorem for a Smooth Class of Semiparametric Estimators,” *Journal of Econometrics* 65, pp. 295 – 332.
- [15] Rosenbaum, P., and D. Rubin (1983): “The Central Role of the Propensity Score in Observational Studies for Causal effects, ” *Biometrika*, 70, pp. 41–55.
- [16] Shefrin, S. (1979): “Unanticipated Monetary Growth and Output Fluctuations,” *Economic Inquiry* 17, pp. 1–13.
- [17] Wooldridge, J.M. (2002): *Econometric Analysis of Cross Section and Panel Data*, Cambridge: MIT Press.