NBER WORKING PAPER SERIES

INVERSE PROBABILITY TILTING FOR MOMENT CONDITION MODELS WITH MISSING DATA

Bryan S. Graham
Cristine Campos de Xavier Pinto
Daniel Egel

Working Paper 13981
http://www.nber.org/papers/w13981

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

We would like to thank David Card, Stephen Cosslett, Jinyong Hahn, Patrick Kline, Justin McCrary, Richard Smith, Tom Rothenberg, members of the Berkeley Econometrics Reading Group and especially Michael Jansson for helpful discussions. We are particularly grateful to Gary Chamberlain, Guido Imbens, Geert Ridder, Enrique Sentana and three anonymous referees for detailed comments on earlier drafts. We also acknowledge feedback and suggestions from participants in seminars at the University of Pittsburgh, Ohio State University, University of Southern California, University of California - Riverside, University of California - Davis, University of Maryland, Georgetown University, Duke University, the University of California - Berkeley, CEMFI (Madrid), Harvard University, Pontifícia Universidade Católica do Rio de Janeiro and the 2009 Latin American Meetings of the Econometric Society. This is a heavily revised and extended version of an NBER Working Paper titled 'Inverse probability tilting and missing data problems'. Previous versions of this paper also circulated under the title "A new method of estimating moment condition models with missing data when selection is on observables." Material in Section 4 of the NBER paper is not included in this version of the paper, but may be found in the companion paper "Efficient estimation of data combination problems by the method of auxiliary-to-study tilting." All the usual disclaimers apply. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

Inverse Probability Tilting for Moment Condition Models with Missing Data
Bryan S. Graham, Cristine Campos de Xavier Pinto, and Daniel Egel
NBER Working Paper No. 13981
May 2008, Revised October 2010
JEL No. C14,C21,C23

ABSTRACT

Missing data are ubiquitous in applied econometric research. A useful, albeit controversial, approach to this problem assumes that conditional on always observed variables missingness occurs at random or that selection is on observables (e.g., Rosenbaum and Rubin, 1984; Heckman and Robb, 1985). This assumption justifies a simple procedure whereby complete cases, those units without any missing data, are reweighted by the inverse of the probability of selection or the propensity score (e.g., Wooldridge, 2007). This paper proposes a flexible parametric variant of inverse probability weighting, inverse probability tilting (IPT). Inverse probability tilting (IPT) replaces the conditional maximum likelihood estimate (CMLE) of the propensity score with a method of moments one. The methods of moments estimator chooses the propensity score parameter so that selected moments in the reweighted complete case subsample exactly coincide with their unweighted full sample counterparts. This alternative propensity score estimate increases the efficiency and robustness of inverse probability weighting. That replacing an efficient estimate of the propensity score with an inefficient one improves the properties of inverse probability weighting is ex ante surprising.

Bryan S. Graham
New York University
19 W 4th Street, 6FL
New York, NY 10012
and NBER
bsg1@nyu.edu

Daniel Egel
Department of Economics
University of California, Berkeley
Berkeley , CA 94720
egel@berkeley.edu

Cristine Campos de Xavier Pinto
Department of Economics
University of California Berkely
508-1 Evans Hall #3880
Berkeley, CA 94720

Abstract

Missing data are ubiquitous in applied econometric research. A useful, albeit controversial, approach to this problem assumes that conditional on always observed variables missingness occurs at random or that selection is on observables (e.g., Rosenbaum and Rubin, 1984; Heckman and Robb, 1985). This assumption justifies a simple procedure whereby complete cases, those units without any missing data, are reweighted by the inverse of the probability of selection or the propensity score (e.g., Wooldridge, 2007). This paper proposes a flexible parametric variant of inverse probability weighting, *inverse probability tilting* (IPT). Inverse probability tilting (IPT) replaces the conditional maximum likelihood estimate (CMLE) of the propensity score with a method of moments one. The methods of moments estimator chooses the propensity score parameter so that selected moments in the reweighted complete case subsample *exactly* coincide with their unweighted full sample counterparts. This alternative propensity score estimate increases the efficiency and robustness of inverse probability weighting. That replacing an efficient estimate of the propensity score with an inefficient one improves the properties of inverse probability weighting is ex ante surprising.

We also compare IPT with the class of augmented inverse probability weighting (AIPW) estimators introduced by Robins, Rotnitzky and Zhao (1994). Relative to AIPW, IPT exhibits lower higher order bias and has an implicit distribution function estimate which is guaranteed to be non-decreasing.

In an empirical application we revisit Johnson and Neal's (1998) analysis of the Black-White wage gap for young men in the United States. They find that 60 percent of the Black-White gap can be explained by group differences in cognitive skills acquired prior to labor market entry at age 18. We study the effect of group differences in skills acquired *prior* to adolescence (i.e., by age 12). Our analysis is complicated by the absence of a pre-adolescence test score for almost 90 percent of respondents. We use IPT to address this missing data problem. We find that two thirds of the pre-market effect found by Johnson and Neal (1998), or 40 percent of the overall gap, can be accounted for by group differences in cognitive skills already present by age 12. The IPT wage gap estimate is precisely determined relative to the standard IPW estimate with a standard error one half as large.

Operationally IPT coincides with standard IPW with the sole exception that the CMLE of the propensity score is replaced with our method of moments estimate. This estimate is the unique solution to a globally concave programming problem (Appendix C). In theory, and in practice, computing our IPT propensity score estimate is straightforward. A MATLAB routine for this purpose is available at `https://files.nyu.edu/bsg1/public/`.

JEL Classification: C14, C21, C23, J15, J70

Key Words: Missing Data, Semiparametric Efficiency, Double Robustness, (Augmented) Inverse Probability Weighting (IPW), Black-White Gap, Causal Inference, Average Treatment Effect (ATE), Two Sample Instrumental Variables (TSIV)

Inverse probability weighting (IPW) methods are widely used in economics and other disciplines to address missing data problems. If the data are missing at random (MAR) conditional on always-observed variables, or selection is on observables, then reweighting complete cases by the inverse of the probability of selection creates a pseudo-dataset that mimics a random sample from the population of interest (e.g., Rosenbaum, 1987). Standard estimation procedures applied to the reweighted complete cases can be used to consistently estimate the parameter of interest (e.g., Wooldridge, 2007).

Settings where IPW has proved useful include M-estimation under nonrandom sampling (e.g., Horvitz and Thompson, 1952; Manski and McFadden, 1981; Wooldridge, 1999; 2001), correcting for attrition in panel data (e.g., Robins and Rotnitzky, 1995; Robins, Rotnitzky and Zhao, 1995; Abowd, Crépon and Kramarz, 2001; Wooldridge, 2002), the construction of counterfactual distributions (e.g., Rosenbaum, 1987; Dinardo, Fortin and Lemieux, 1996), program evaluation under exogenous treatment assignment (e.g., Hirano, Imbens and Ridder, 2003), and dealing with mismeasured or missing regressors (e.g., Robins, Rotnitzky and Zhao, 1994; Chen, Hong and Tarozzi, 2004, 2008). Each of these problems belongs to a class of semiparametric missing data problems studied by Robins, Rotnitzky and Zhao (1994, Section 8). Wooldridge (2007) shows how these problems may be solved by appropriately constructed inverse probability weighted (IPW) M-estimators.

Three difficulties can arise when using IPW[2] methods. First, IPW is inefficient *unless* the selection probability or propensity score is modelled nonparametrically (Hirano, Imbens and Ridder, 2003). Unfortunately, due to the curse of dimensionality, nonparametric modelling of the propensity score can be impractical when more than a handful of the conditioning variables have continuous components.[3] Second, IPW "can attach large weights to small quantities of data, often in a fairly erratic manner" (Rosenbaum, 1987; p. 387; cf., Fortin, Lemieux, and Firpo, 2010). In settings where the probability of missingness is very high for some subpopulations, IPW estimates can be very sensitive to small changes in the propensity score model. Third, and

[2]Throughout we will use IPW as an abbreviation for both 'inverse probability weighted' and 'inverse probability weighting' with the intended meaning clear from the context.

[3]Unless, of course, a very large sample is available. Heckman, Ichimura and Todd (1998, p. 271) make related point noting that "...the dimensionality of X is a major drawback to practical application of the matching method or to the use of conventional nonparametric regression....For high dimensional X variables, neither method is feasible in samples of the size typically available to social scientists."

related to the second concern, the consistency of IPW methods requires that the propensity score is correctly modelled.[4]

In this paper we propose a modified version of inverse probability weighting, *inverse probability tilting* (IPT), which addresses each of these concerns. Our point of departure is a equivalence result due to Graham (2009): the variance bound for the semiparametric missing data model of Robins, Rotnitzky and Zhao (1994) coincides with that of a particular conditional moment problem. This problem suggests natural analog estimators. Operationally two issues arise. First, on which of the infinite number of unconditional moments implied by the conditional restriction should estimation be based? Second, how should any overidentification be dealt with? Our answer to these two questions results in an easily implementable variant of inverse probability weighting.

Our procedure coincides with the IPW estimator of, for example, Wooldridge (2007), *except* that we replace the conditional maximum likelihood estimate (CMLE) of the propensity score with an alternative method of moments estimator. We show that if the unconditional moments used to estimate the propensity score parameter are appropriately chosen our procedure (i) is locally efficient and (ii) remains consistent even if the propensity score is misspecified. These properties, *local efficiency* and *double robustness*, which we carefully define below, are not shared by the standard IPW estimator.[5] We show that the correct choice of moments is an implication of the maintained propensity score model *and* researcher beliefs about the conditional distribution of the missing variables given those always observed. As the first object is the foundation of reweighting and matching approaches to missing data and the second of imputation approaches (cf., Rubin, 1977; Gourieroux and Monfort, 1981; Little and Rubin, 2002; Browning and Søren Leth-Petersen, 2003), the process of optimal moment selection involves considerations familiar to applied researchers.

A key appeal of inverse probability weighting is its conceptual and operational

[4] In principal, if the richness of the propensity score model is allowed to grow with the sample size, as in Hirano, Imbens and Ridder (2003) and Chen, Hong and Tarozzi (2004, 2008), IPW will be consistent (as well as efficient) as long as the true propensity score is sufficiently smooth. The focus of this paper is on weighting methods that involve (flexible) parametric modelling of the propensity score. This is the case considered by Wooldridge (2007). In this case fragility vis-a-vis misspecification of the propensity score is relevant.

[5] To be more specific, IPW is locally efficient at a rather peculiar data generating process (DGP). Unfortunately this DGP is difficult to interpret and a priori implausible. We discuss this point below.

simplicity. Inverse probability tilting preserves this advantage, while offering improvements in terms of estimator efficiency and robustness. However other modifications of IPW exist. A leading one, which shares IPT's local efficiency and double robustness properties, is the augmented inverse probability weighting (AIPW) estimator introduced by Robins, Rotnitzky and Zhao (1994). While perhaps less familiar to econometricians, although Hirano and Imbens (2001), Imbens (2004), and Wooldridge (2007) are notable exceptions, AIPW methods are widely-studied (and used) by statisticians. Tsiatis (2006) provides a book length treatment. These estimators choose $\widehat{\gamma}$ to set a *parametric* estimate of the efficient score for γ_0 equal to zero. As shown by Newey (1990, Section 4), this is a general approach to constructing locally efficient estimators for semiparametric models.

Several variants of AIPW are now available, including versions due to Newey (1994a) and Cao, Tsiatis and Davidian (2009). We also show that the estimators proposed by Hirano and Imbens (2001) and Wooldridge (2007, Section 6.2) are AIPW ones. This is a priori non obvious. We demonstrate that each of these AIPW estimators belongs to a particular class of iterated GMM estimators (Hansen, Heaton and Yaron, 1996). The differences among them correspond to differences in the weight matrix being iterated over.

Using the iterated GMM representation we characterize the asymptotic bias of AIPW estimators. We also compute the asymptotic bias of our IPT estimator. These bias comparisons are interesting because IPT and AIPW are first order equivalent. Our derivations are based on stochastic expansions as in Rilstone, Srivastava and Ullah (1996) and Newey and Smith (2004). We find that IPT has smaller bias than AIPW. We also compare AIPW and IPW in terms of their implicit distribution function estimates (Back and Brown, 1993). While both estimates are efficient under a common prior restriction, the AIPW one can assign negative weights to some data points. Consequently it may be decreasing over some intervals. This phenomenon is likely to occur when the distribution of the always observed covariates is very different across the complete case and missing case subsamples (i.e., when overlap is limited).

In an empirical application we revisit Johnson and Neal's (1998) analysis of the Black-White wage gap for young men in the United States. They find that approximately 60 percent of the Black-White gap can be explained by group differences in cognitive skills acquired prior to labor market entry at age 18. We study the effect of group differences in skills acquired prior to adolescence (i.e., by age 12). We find

that pre-adolescent skill differences can explain about 40 percent of the overall wage gap and two thirds of pre-market effect found by Johnson and Neal (1998).

Our analysis is complicated by the fact that a pre-adolescence test score is available for just 11 percent of respondents. In addition to being few in number, these complete cases are unrepresentative of the sample as a whole. An analysis which ignores these facts is likely to be both inconsistent and imprecise. The IPT estimate of the wage gap conditional on the preadolescence test score corrects for the unrepresentativeness of the complete cases. The IPT point estimate is also precisely determined. Its standard error is one half the length of the one for the standard IPW estimate. Our application provides a concrete example of the type of efficiency gains IPT can provide. These gains arise despite the fact that we implement IPW with a heavily overparameterized propensity score model, which theory suggests should lead to a precisely determined point estimate (Hirano, Imbens and Ridder, 2003; Wooldridge, 2007).

All proofs are collected in the appendix. We also detail a computational algorithm that we have found to be reliable. Software implementing this procedure is available online at `https://files.nyu.edu/bsg1/public/`. A supplemental Web Appendix provides details of some of the more tedious underlying calculations.

Relationship to prior research Our paper is related to several distinct research programs. In econometrics various individuals have proposed globally efficient estimators for missing data models, particularly treatment effect models under exogenous treatment assignment (e.g., Cheng, 1994; Newey, 1994a; Hahn, 1998; Hirano, Imbens and Ridder, 2003; Ichimura and Linton, 2005; Imbens, Newey and Ridder, 2005; Chen, Hong and Tarrozi, 2004, 2007). All of these estimators require nonparametric estimation of (potentially) high dimensional objects. This limits their practical usefulness. While their first order asymptotic properties are not sensitive to the particulars of the nonparametric estimator used (including its dimension), their finite samples properties often will be (cf., Wang, Linton and Härdle, 2004). This motivates our focus on finding flexible parametric procedures with good efficiency and robustness properties. Practicality often demands that empirical work is of the flexible parametric variety, our theory is consequently concordant with (much of) actual practice.[6]

[6] The credibility of the MAR assumption often requires X to be high dimensional: the assumption of 'no selection on unobservables' is typically most persuasive when the researcher is able to condition on many observed unit characteristics (cf., Heckman, Ichimura and Todd, 1997). Of course, when deciding on which variables to include in X, a variety of considerations must be taken into account.

Our focus on flexible parametric modelling is shared by researchers in biostatistics (e.g., Robins, Rotnitzky and Zhao, 1994). Our proposed estimator, both operationally and in its derivation, differs from AIPW, the preferred approach in that literature. We are also the first, to our knowledge, to attempt higher order comparisons in the missing data context.[7] Our theoretical analysis suggests that IPT is attractive relative to both IPW and AIPW.

Our bias calculations are based on stochastic expansions of the first order conditions defining the IPT and AIPW estimators. In undertaking these calculations we use Lemma A.4 of Newey and Smith (2004, pp. 241 - 242). Some of the properties of IPT may be understood by making analogies with empirical likelihood (Imbens, 2002; Kitamura, 2007).[8] More generally the desirable properties of IPT are related to the theory of efficient estimation of expectations (e.g., Brown and Newey, 1998). Finally, our method of deriving a locally efficient estimator shares similarities with the literature on optimal instrumental variables in GMM (e.g., Chamberlain, 1987; Newey, 1993). We make some of these connections explicit in what follows.

1 A general moment condition model with missing data

In this section we describe a general moment condition model with data missing at random (MAR) or where selection is on observables (Rosenbaum and Rubin, 1983; Heckman and Robb, 1985). Our set-up is (essentially) the same as in Wooldridge (2007) except that our parameter is the solution to a moment condition, as opposed to a population optimization, problem (cf., Chen, Hong and Tarozzi, 2008). Let $Z = (Y_1', X')'$ be a random vector, γ_0 an unknown parameter, and assume that:

Assumption 1.1 (IDENTIFICATION) *For some known $K \times 1$ vector of functions $\psi(z, \gamma)$*

$$\mathbb{E}[\psi(Z, \gamma_0)] = 0,$$

with (i) $\mathbb{E}[\psi(Z, \gamma)] \neq 0$ for all $\gamma \neq \gamma_0$, $\gamma \in \mathcal{G} \subset \mathbb{R}^K$ and $\mathcal{G}$ compact, (ii) $|\psi(z, \gamma)| \leq$

[7] Ichimura and Linton (2005) and Imbens, Newey and Ridder (2005) do undertake mean squared error calculations, but not to compare alternative estimators.

[8] These connections were heavily emphasized in our NBER Working Paper (Egel, Graham and Pinto, 2008).

$b(z)$ *for all* $z \in \mathcal{Z}$ *with* $b(z)$ *a non-negative function on* $\mathcal{Z}$ *and* $\mathbb{E}[b(Z)] < \infty$, *(iii)* $\psi(z,\gamma)$ *is continuous on* $\mathcal{G}$ *for each* $z \in \mathcal{Z}$ *and continuously differentiable in a neighborhood of* γ_0, *(iv)* $\mathbb{E}\left[\|\psi(Z,\gamma_0)\|^2\right] < \infty$, *and (v)* $\mathbb{E}\left[\sup_{\gamma\in\mathcal{G}} \|\nabla_\gamma \psi(Z,\gamma)\|\right] < \infty$.

Assumption 1.1 provides a standard set of conditions under which the full sample method-of-moments estimate of γ_0, the solution to $\sum_{i=1}^N \psi(Z_i,\widehat{\gamma})/N = 0$, will be consistent and asymptotically normally (cf., Newey and McFadden 1994, Theorems 2.6 and 3.4).[9]

Here our interest is in identification and estimation when Y_1 is not observed for all units. Let D be a binary indicator variable. When $D = 1$ we observe Y_1 and X, while when $D = 0$ we observe only X. Our benchmark model is defined by Assumption 1.1 as well as:

Assumption 1.2 (Random Sampling) $\{D_i, X_i, Y_{1i}\}_{i=1}^N$ *is an independently and identically distributed random sequence. We observe* D, X *and* $Y = DY_1$ *for each sampled unit.*

Assumption 1.3 (Missing at Random) $\Pr(D = 1|X, Y_1) = \Pr(D = 1|X)$

Assumption 1.4 (Strong Overlap) *Let* $p_0(x) = \Pr(D = 1|X = x)$, *then* $0 < \kappa \leq p_0(x) \leq 1$ *for some* $0 < \kappa < 1$ *and all* $x \in \mathcal{X} \subset \mathbb{R}^{\dim(X)}$.

Assumption 1.5 (Propensity Score Model) *There is a unique* $\delta_0^* \in \mathcal{D}^* \subset \mathbb{R}^{\dim(\delta^*)}$, *known vector* $r(X)$ *of linearly independent functions of* X,*and known function* $G(\cdot)$ *such that (i)* $G(\cdot)$ *is strictly increasing, differentiable and maps into the unit interval with* $\lim_{v\to-\infty} G(v) = 0$ *and* $\lim_{v\to\infty} G(v) = 1$, *(ii)* $p_0(x) = G(r(x)'\delta_0^*)$ *for all* $x \in \mathcal{X}$, *and (iii)* $0 < \kappa \leq G(r(x)'\delta^*) \leq 1$ *for all* $\delta^* \in \mathcal{D}^*$ *and* $x \in \mathcal{X}$.

In what follows we refer to the model defined by Assumptions 1.1 to 1.5 as the semiparametric missing data model. Hahn (1998), Hirano, Imbens and Ridder (2003), and Chen, Hong and Tarozzi (2004, 2008) study this model without maintaining Assumption 1.5, that is, with the propensity score left nonparametric. As is well-known, removing Assumption 1.5 from the prior restriction does not affect the asymptotic precision with which γ_0 may be estimated (e.g., Hahn, 1998; Chen, Hong and Tarozzi,

[9]Note we assume that γ_0 is just-identified. Extending what follows to the overidentified case is straightforward. Extending what follows to the case where $\psi(Z,\gamma)$ is non-differentiable is also possible, albeit technically demanding (cf., Firpo, 2007).

2008). We nevertheless maintain it when deriving our local efficiency result (Theorem 2.1). Doing so is important for establishing regularity of our estimator. We also assess the properties of IPT when Assumption 1.5 fails (Theorem 2.2).

To get a sense of the range of problems to which our methods may be applied it is helpful to consider a few specific examples.

Example 1.1 (Mean of a variable missing at random) *Let Y_1 be a binary indicator for an individual's HIV status, let $D = 1$ if an individual is tested and zero otherwise; Y_1 is logically observable only when $D = 1$. We would like to estimate the population prevalence of HIV: $\gamma_0 = \mathbb{E}[Y_1]$. This corresponds to setting $\psi(Z, \gamma) = Y_1 - \gamma$. Assumption 1.3 says that the testing decision is independent of HIV status in subpopulations homogenous in X. This may be plausible if X includes measures of risk-taking behavior and other background characteristics so that it closely approximates an individual's own information set regarding their status. Assumption 1.4 requires that at least some individuals in every subpopulation defined in terms of $X = x$ get tested. Assumption 1.5 presumes the availability of a parametric model for the testing decision. This example is closely related to that of average treatment effect (ATE) estimation under exogenous assignment (e.g., Imbens, 2004; see Section 5 below).*

Example 1.2 (Regression function estimation with missing regressors) *Let X_1 be a vector of demographic characteristics, X_2 log earnings, Y_1 armed forces qualification test (AFQT) score, and X_3 a vector of always observed surrogates or proxies for Y_1 (e.g., scores on subcomponents of the test, on earlier tests, etc.). Let $D = 1$ if a unit's test score is available and zero otherwise. Let $X = (X_1', X_2', X_3')'$, $\gamma = (\gamma_1', \gamma_2')'$ and $\psi(Z, \gamma) = (X_1', Y')'(X_2 - X_1'\gamma_1 - Y_1'\gamma_2)$. Here γ corresponds to the coefficient vector indexing the linear predictor of log earnings given demographics and AFQT score as in Neal and Johnson (1996)*[10] *This corresponds to a linear regression model where the covariate of interest is subject to item non-response. Assumption 1.3 requires that across individuals with identical earnings (X_2), demographics (X_1), and test proxies (X_3) the probability of observing the AFQT score is independent of its value.*

[10] Using X_2 to denote a left-hand-side, and Y_1 a right-hand-side, variable is confusing but is done here to maintain consistency with the notation of Assumptions 1.1 to 1.5.

Chen, Hong and Tarozzi (2004), Wooldridge (2002, 2007) and Egel, Graham and Pinto (2008) survey additional examples of the semiparametric missing data model defined by Assumptions 1.1 to 1.5. See also Section 5 below.

2 Inverse probability tilting

Our approach to estimation involves inverse probability weighting (IPW) as in Rosenbaum (1987), Wooldridge (2002, 2007) and others. In what follows, unless noted explicitly otherwise, we use IPW to denote inverse probability weighting where (i) the propensity score is modelled parametrically and (ii) fitted by maximum likelihood. The main difference between our procedure, inverse probability tilting (IPT), and this standard IPW estimator is that we do not use a conditional maximum likelihood estimate (CMLE) of the propensity score.[11]

As the deliberate use of an inefficient estimate of a parameter (albeit a nuisance one) is non standard we spend some time explaining the reasoning behind our approach. Central to this approach is an equivalence result on semiparametric efficiency bounds for missing data problems due to Graham (2009).

2.1 Inverse probability weighting

Our first result shows that IPW, with the propensity score estimated by CMLE, is typically inefficient under Assumptions 1.1 to 1.5. The maximal asymptotic precision with which γ_0 can be estimated under these assumptions was characterized by Robins, Rotnitzky and Zhao (1994), Hahn (1998) and Chen, Hong and Tarozzi (2004, 2008) and is given by the inverse of

$$\mathcal{I}(\gamma_0) = \Gamma_0' \Lambda_0^{-1} \Gamma_0, \tag{1}$$

with

$$\Gamma_0 = \mathbb{E}\left[\frac{\partial \psi(Z, \gamma_0)}{\partial \gamma'}\right], \quad \Lambda_0 = \mathbb{E}\left[\frac{\Sigma(X; \gamma_0)}{p_0(X)} + q(X; \gamma_0)\, q(X; \gamma_0)'\right], \tag{2}$$

where $\Sigma(x; \gamma) = \mathbb{V}(\psi(Z, \gamma) | X = x)$ and $q(X; \gamma) = \mathbb{E}[\psi(Z, \gamma) | X = x]$.

[11] To the best of our knowledge all versions of IPW based on *parametric* models of the propensity score use CMLE. Hirano, Imbens and Ridder (2003) and Chen, Hong and Tarozzi (2004, 2008) use a sieve maximum likelihood (SML) estimate of the propensity score. Operationally this is nondistinguishable from a flexible parametric maximum likelihood estimate.

Let $r_i = r(X_i)$, $\psi_i(\gamma) = \psi(Z_i, \gamma)$ and $\psi_i = \psi(Z_i, \gamma_0)$. Similarly let $G_i(\delta^*) = G(r_i'\delta^*)$ and $G_i = G(r_i'\delta_0^*)$. Denote a random unit's contribution to the score of the propensity score log-likelihood evaluated at $\delta^* = \delta_0^*$ by

$$S_{\delta^*} = \frac{D - G}{G(1 - G)} G_1 r,$$

with $G_s(v) = \partial^s G(v) / \partial v^s$ for $s = 1, 2$. Finally let $q(X_i; \gamma) = \mathbb{E}[\psi(Z_i, \gamma) | X_i]$ and $q_i = q(X_i; \gamma_0)$. The inverse probability weighted estimate of γ_0 is given by the solution to

$$\frac{1}{N} \sum_{i=1}^{N} \frac{D_i \psi(Z, \widehat{\gamma}_{IPW})}{G(r(X_i)' \widehat{\delta}_{ML}^*)} = 0, \tag{3}$$

with $\widehat{\delta}_{ML}^*$ the CMLE estimate of δ_0^*. Proposition 2.1 summarizes the first order asymptotic properties of $\widehat{\gamma}_{IPW}$ (cf., Wooldridge, 2002, 2007).

Proposition 2.1 (Asymptotic Sampling Distribution of $\widehat{\gamma}_{IPW}$) *Suppose Assumptions 1.1 to 1.5 and additional regularity conditions hold, then (i)* $\sqrt{N}(\widehat{\gamma}_{IPW} - \gamma_0) \xrightarrow{D} \mathcal{N}(0, \text{AVar}(\widehat{\gamma}_{IPW}))$ *with*

$$\begin{aligned} \text{AVar}(\widehat{\gamma}_{IPW}) &= \mathcal{I}(\gamma_0)^{-1} \\ &\quad + \Gamma^{-1} \mathbb{E}\left[\left(\left(\frac{D}{G} - 1 \right) q - \Pi_S S_{\delta^*} \right) \left(\left(\frac{D}{G} - 1 \right) q - \Pi_S S_{\delta^*} \right)' \right] \Gamma^{-1\prime}, \end{aligned} \tag{4}$$

for $\Pi_S = \mathbb{E}\left[\frac{D}{G} \psi S_{\delta^*}'\right] \mathbb{E}\left[S_{\delta^*} S_{\delta^*}'\right]^{-1}$ *and (ii)* $k' \left[\text{AVar}(\widehat{\gamma}_{IPW}) - \mathcal{I}(\gamma_0)^{-1}\right] k \geq 0$ *for any vector of constants* k.

Proof. See Appendix A ■

While the inefficiency of IPW, part (ii) of Proposition 2.1, is well known (e.g., Hirano, Imbens and Ridder, 2003), the asymptotic variance expression (4) provides some insight into its large sample properties. Observe that $\Pi_S S_{\delta^*}$ equals the best (i.e., mean squared error minimizing) linear predictor of $\left(\frac{D}{G} - 1\right) q$ given S_{δ^*}.[12] If S_{δ^*} happens to be a good predictor of $\left(\frac{D}{G} - 1\right) q$, then IPW will be nearly efficient.

[12] Note that by the conditional mean zero property of the score function and Assumption 1.3

$$\mathbb{E}\left[\left(\frac{D}{G} - 1 \right) q S_{\delta^*}' \right] = \mathbb{E}\left[\frac{D}{G} q S_{\delta^*}' \right] = \mathbb{E}\left[\frac{D}{G} \psi S_{\delta^*}' \right].$$

Consider the case where the propensity score takes a logit form so that $G(v) = \exp(v)/[1+\exp(v)]$. Some basic calculations give $S_{\delta^*} = \left(\frac{D}{G} - 1\right) G \cdot r$; therefore if it so happens that q can be written as a linear function of $G \cdot r$, then the asymptotic variance of IPW will coincide with that of an efficient estimator. An interpretation of Hirano, Imbens and Ridder (2003) is that if the dimension of r is allowed to grow with the sample size, then q will eventually be arbitrarily well-approximated by a linear function of $G \cdot r$, so that this coincidence holds generally (i.e., for *all* data generating process (DGPs)). Wooldridge (2002, 2007) makes a related point: (4) cannot decrease if the dimension of r increases.

2.2 Optimal moments for propensity score estimation

Equation (4) suggests that if we make a random unit's contribution to the estimating equation for the propensity score proportional to $\left(\frac{D}{G} - 1\right) q$, then we can improve the asymptotic precision of IPW estimates of γ_0. One concern with this conjecture is that changing the estimating equation for δ_0^* necessarily involves replacing $\widehat{\delta}_{ML}^*$ in (3) with something weakly less efficient. However, since the variance bound for γ_0 is unaffected by prior restrictions on the propensity score, the increased sampling error in $\widehat{\delta}^*$ associated with using something other than the CMLE, should not affect $\widehat{\gamma}$'s sampling properties (at least to first order). Both of these conjectures turn out to be correct, as we now show.

Our starting point is the observation that the variance bound for γ_0 in the conditional moment problem

$$\mathbb{E}\left[\frac{D}{p_0(X)}\psi(Z,\gamma_0)\right] = 0 \tag{5}$$

$$\mathbb{E}\left[\frac{D}{p_0(X)} - 1 \middle| X\right] = 0, \tag{6}$$

coincides with the inverse of the information bound for γ_0 in the corresponding semiparametric missing data problem (Graham, 2009; Theorem 2.1). To clarify our argument we initially assume that the propensity score is known. In that case we could base estimation solely on (5). It is well known that the resulting known weights estimator is inefficient (e.g., Rosenbaum, 1987). The method of moments representation suggests a simple intuition for why this is so: the known weights estimator ignores the auxiliary conditional moment (6). This restriction implies that $D/p_0(X) - 1$ will

be uncorrelated with any function of X. Using unconditional moments of this form to augment (5) will typically improve the precision with which γ_0 is estimated. From Graham (2009) we know that the maximal asymptotic precision with which γ_0 can be estimated under (5) and (6) coincides with $\mathcal{I}(\gamma_0)^{-1}$. Achieving this bound in practice requires choosing the appropriate set of unconditional moments implied by (6) (out of the infinite number of valid ones).

To describe the optimal moment set we need some additional notation. Let C be a $K \times K$ matrix of constants, $t^*(X)$ a column vector with a 1 in the first row and known functions of X in the remaining rows, and

$$A(X) = \begin{pmatrix} C & 0 \\ 0 & t^*(X) \end{pmatrix}. \tag{7}$$

Conditions (5) and (6) imply, by iterated expectations, that

$$\mathbb{E}\left[A(X)\,\rho(Z, \gamma_0, p_0(X))\right] = 0, \tag{8}$$

for

$$\rho(Z, \gamma_0, p_0(X)) = \begin{pmatrix} \frac{D}{p_0(X)}\psi(Z, \gamma_0) \\ \frac{D}{p_0(X)} - 1 \end{pmatrix}.$$

By Chamberlain (1987) the asymptotic variance of any efficient estimator[13] based on (8) alone will be $V_A = \left(H_A' \Omega_A^{-1} H_A\right)^{-1}$ with

$$\begin{aligned} H_A &= \mathbb{E}\left[A(X)\,\frac{\partial \rho(Z, \gamma_0, p_0(X))}{\partial \gamma'}\right] \\ \Omega_A &= \mathbb{E}\left[A(X)\,\rho(Z, \gamma_0, p_0(X))\,\rho(Z, \gamma_0, p_0(X))'\,A(X)\right]. \end{aligned} \tag{9}$$

The following proposition characterizes the variance-minimizing choice of $A(X)$.

Proposition 2.2 (Optimal Moments For Missing Data Models) *Under conditions (5) and (6) with $p_0(X)$ known and satisfying Assumption 1.4 (i) the asymp-*

[13] Examples of efficient estimators in this context include two-step GMM and generalized empirical likelihood (GEL).

totic variance minimizing choice of $A(X)$ in (8) is

$$A_{\text{opt}}(X;\gamma_0) = \begin{pmatrix} I_K & 0 \\ 0 & q(X;\gamma_0) \end{pmatrix}, \tag{10}$$

and (ii) this variance bound equals $\mathcal{I}(\gamma_0)^{-1}$.

Proof. Let

$$\begin{aligned} m_A &= H_A' \Omega_A^{-1} A(X) \rho(Z, \gamma_0, p_0(X)) \\ m_q &= H_{A_{\text{opt}}}' \Omega_{A_{\text{opt}}}^{-1} A_{\text{opt}}(X;\gamma_0) \rho(Z, \gamma_0, p_0(X)). \end{aligned}$$

Note that $\mathbb{E}[m_a m_a'] = V_A^{-1}$, and, by iterated expectations, $\mathbb{E}[m_A m_q'] = V_A^{-1}$. For part (i) note that

$$\begin{aligned} V_A - \mathbb{E}[m_q m_q']^{-1} &= \mathbb{E}[m_A m_A']^{-1} - \mathbb{E}[m_q m_q']^{-1} \\ &= \mathbb{E}[m_A m_q']^{-1} \\ &\times \left(\mathbb{E}[m_A m_A'] - \mathbb{E}[m_A m_q'] \mathbb{E}[m_q m_q']^{-1} \mathbb{E}[m_q m_A'] \right) \mathbb{E}[m_q m_A']^{-1} \\ &= \mathbb{E}[RR'] \end{aligned}$$

with $R = \mathbb{E}[m_A m_q']^{-1} \left(m_A - \mathbb{E}[m_A m_q'] \mathbb{E}[m_q m_q']^{-1} m_q \right)$. Since $\mathbb{E}[RR']$ is positive definite $\mathbb{E}[m_q m_q']^{-1}$ is a lower bound for the asymptotic variance of all estimators based on the unconditional restriction $\mathbb{E}[A(X)\rho(Z, \gamma_0, p_0(X))] = 0$ (cf., Newey, 1993). Part (ii) follows from the fact that if $A(X) = A_{\text{opt}}(X;\gamma_0)$, then $\mathbb{E}[m_q m_q'] = \mathcal{I}(\gamma_0)$. ■

Observe that m_q is the influence function of an efficient estimator based on (8) with $A(X) = A_{\text{opt}}(X;\gamma_0)$. Since the gradient of m_q with respect to the propensity score is mean zero, the above result also holds when the known propensity score is replaced with a consistent estimate (cf. Newey, 1994b; Proposition 3). This is related to the fact that knowledge of the propensity score does not alter the efficiency bound for γ_0.

Proposition 2.2 suggests several possible approaches to estimation. One approach, inspired by work on optimal instrumental variables estimation (e.g., Newey, 1993), would be the following multi-step procedure:

Procedure 2.1 (OPTIMAL GMM) *(i) compute the conditional maximum likelihood estimate of the propensity score, $G(r(X)'\widehat{\delta}^*_{ML})$; (ii) compute the (inefficient) IPW estimate of γ_0 using the propensity score estimated in step (i); (iii) compute a flexible parametric estimate of $q(X;\gamma_0)$ based on a (weighted) least squares fit of $\psi(Z,\widehat{\gamma}_{IPW})$ onto functions of X in the complete-case ($D=1$) subsample; (iv) using the step (iii) fitted values construct the optimal instruments $\widehat{A}_{\text{opt}}(X;\widehat{\gamma}_{IPW})$; and (v) compute $\widehat{\gamma}$ by two-step GMM using $\widehat{A}_{\text{opt}}(X;\widehat{\gamma}_{IPW})\,\rho\left(Z,\gamma,G(r(X)'\widehat{\delta}^*_{ML})\right)$ as the moment vector.*

In Section 3 we show that the above procedure is closely related to the AIPW estimator first proposed by Robins, Rotnitzky and Zhao (1994).

2.3 A modified reweighting estimator: inverse probability tilting

One concern with an estimate based on Procedure 2.1 is that while estimation error in $\widehat{A}_{\text{opt}}(X;\widehat{\gamma}_{IPW})$ and $G(r(X)'\widehat{\delta}^*_{ML})$ will not affect its asymptotic sampling distribution, it may be important in finite samples. Our IPT estimator is formulated with this issue in mind.

Proposition 2.2 shows that the optimal instrument is a function of $q(X;\gamma_0) = \mathbb{E}[\psi(Z,\gamma_0)|X]$. Our IPT procedure requires formulating a working model for this object.

Assumption 2.1 (MOMENT CEF MODEL) *For some unique matrix Π^*_0 and vector of linear independent functions $t^*(X)$ with a constant 1 in the first row, we have*

$$\mathbb{E}[\psi(Z,\gamma_0)|X] = \Pi^*_0 t^*(X).$$

The precise content of Assumption 2.1 depends on the form of $\psi(Z,\gamma)$. If $\psi(Z,\gamma) = Y_1 - \gamma$, as in Example 1.1, then it is equivalent to assuming that the conditional mean of Y_1 is a linear function of $t^*(X)$. Example 1.2 provides a more complicated illustration. In that case

$$\mathbb{E}[\psi(Z,\gamma_0)|X] = \begin{pmatrix} X_1X_2 - X_1X_1'\gamma_1 - X_1\mathbb{E}[Y_1|X]'\gamma_2 \\ \mathbb{E}[Y_1|X]X_2 - \mathbb{E}[Y_1|X]X_1'\gamma_1 - \mathbb{E}[Y_1Y_1'|X]\gamma_2 \end{pmatrix},$$

so that selecting $t^*(X)$ requires formulating models for the first and second conditional moments of Y_1.[14]

When $\psi(Z,\gamma)$ is nonlinear in γ choosing $t(X)$ such that Assumption 2.1 holds is more difficult. In this case one can think of $t^*(X)$ as a vector of approximating functions as in the literature on nonparametric sieve estimation (e.g., Chen, 2007). We emphasize that any approach to missing data which involves imputation also requires formulating a model for $\mathbb{E}[\psi(Z,\gamma_0)|X]$ (cf., Little and Rubin, 2002; Browning and Leth-Petersen, 2003).

Under Assumption 2.1 we have the following Corollary.

Corollary 2.1 *Under Assumption 2.1, and the conditions of Proposition 2.2, an efficient estimator based on (8) with*

$$A(X) = \begin{pmatrix} I_K & 0 \\ 0 & t(X) \end{pmatrix}$$

has an asymptotic variance equal to $\mathcal{I}(\gamma_0)^{-1}$.

Proof. Recall that $V_A = \left(H_A'\Omega_A^{-1}H_A\right)^{-1}$. Since $H_A = (\Gamma_0', 0)'$ we have $V_A^{-1} = \Gamma_0'\left\{\Omega_A^{-1}\right\}_{1:K,1:K}\Gamma_0$, where $\left\{\Omega_A^{-1}\right\}_{1:K,1:K}$ is the upper-left-hand $K \times K$ block of Ω_A. Evaluating Ω_A and inverting gives

$$\begin{aligned}
\left\{\Omega_A^{-1}\right\}_{1:K,1:K} &= \mathbb{E}\left[\frac{\mathbb{E}[\psi\psi|X]}{p_0(X)} - \frac{1-p_0(X)}{p_0(X)}\Pi_0^* t^*(X)\, t^*(X)'\,\Pi_0^{*\prime}\right]^{-1} \\
&= \mathbb{E}\left[\frac{\mathbb{E}[\psi\psi|X]}{p_0(X)} - \frac{1-p_0(X)}{p_0(X)} q_0(X)\, q_0(X)'\right]^{-1} \\
&= \mathbb{E}\left[\frac{\Sigma_0(X)}{p_0(X)} + q_0(X)\, q_0(X)'\right]^{-1} \\
&= \Lambda_0^{-1},
\end{aligned}$$

with the second equality following from Assumption 2.1, the third from the definition of conditional variance, and the fourth from the definition of Λ_0 given in (2). ■

[14] To be explicit assume that $\mathbb{E}[Y_1|X] = h_1(X)'\pi_1$ and $vech(\mathbb{E}[Y_1Y_1'|X]) = h_2(X)'\pi_2$. Let $h_3(X)$ consist of $h_1(X)$ and all non-redundant interactions between its elements and those of X_1 and X_2, then setting $t^*(X) = (h_2(X)', h_3(X)')'$ with any redundant entries removed is sufficient for Assumption 2.1 to hold.

Corollary 2.1 justifies our approach to estimation. Let $t(X)$ denote the union of all distinct elements in $t^*(X)$ and $r(X)$ (recall that $r(X)$ are the functions of X entering the propensity score model in Assumption 1.5). Let $1+M$ equal the dimension of $t(X)$; this vector will include a constant and M known functions of X. Note that $t(X) = \left(r(X)', r^*(X)'\right)'$ where $r^*(X)$ is the relative complement of $r(X)$ in $t^*(X)$. Letting $\delta_0 = (\delta_0^{*\prime}, \eta_0')'$ we have under Assumptions 1.1 to 1.5 the following *just-identified* unconditional moment problem

$$\mathbb{E}\left[\frac{D}{G(t(X)'\delta_0)}\psi(Z,\gamma_0)\right] = 0 \tag{11}$$

$$\mathbb{E}\left[\left(\frac{D}{G(t(X)'\delta_0)} - 1\right)t(X)\right] = 0. \tag{12}$$

Our proposed estimator chooses $\widehat{\beta}_{IPT} = \left(\widehat{\gamma}_{IPT}', \widehat{\delta}_{IPT}'\right)'$ to set the sample analog of (11) and (12) equal to zero:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{D_i}{G(t(X_i)'\widehat{\delta}_{IPT})}\psi(Z_i,\widehat{\gamma}_{IPT}) = 0 \tag{13}$$

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i}{G(t(X_i)'\widehat{\delta}_{IPT})} - 1\right)t(X_i) = 0. \tag{14}$$

Several features of this estimator merit comment. First, if $r(X)$ is not contained within $t^*(X)$, then we add moments to the propensity score estimating equation, replacing $t^*(X)$ with $t(X)$. These additional moments do not improve the precision of $\widehat{\gamma}_{IPT}$, but they do ensure that (12) contains a sufficient number of moment restrictions to pin down the propensity score parameter.

Second, in the opposite case where $t^*(X)$ is not contained within $r(X)$, we enrich the propensity score model, replacing $r(X)'\delta_0^*$ with $t(X)'\delta_0$ in $G(\cdot)$. The effect of this replacement is to eliminate any overidentifying restrictions. To see this note that

$$t(X)'\delta_0 = r(X)'\delta_0^* + r^*(X)'\eta_0,$$

where, by Assumption 1.5, $\eta_0 = \underline{0}$. Nevertheless including $r^*(X)$ in the propensity score model ensures that the combined dimension of (11) and (12) coincides with $\dim(\gamma_0) + \dim(\delta_0) = K + 1 + M$ so that $\beta_0 = (\gamma_0', \delta_0')'$ is just-identified. This

approach to overidentification appeals to be novel.[15] Theorem 3.1 below shows that it results in attractive higher order properties.

Third, our propensity score parameter estimate, $\widehat{\delta}_{IPT}$, is not the conditional maximum likelihood estimate. Therefore $\widehat{\delta}_{IPT}$ is an inefficient estimate of $\delta_0 = (\delta_0^{*\prime}, \underline{0}')'$. Although IPT uses an inefficient propensity score estimate, Corollary 2.1 suggests that $\widehat{\gamma}_{IPT}$ will be locally efficient (Theorem 2.1 shows this formally). We show that estimating the propensity score in this way also endows IPT with an interesting robustness property (see Theorem 2.2).

An example helps to fix ideas. Let $\psi(Z, \gamma) = Y_1 - \gamma$ as in Example 1.1 with X scalar. We assume that Assumption 1.5 holds with $r(X) = (1, X)'$ so that the propensity score is, for example, logit with an index linear in X. In choosing $t^*(X)$ such that Assumption 2.1 holds we are concerned about possible nonlinearities in $\mathbb{E}[Y_1 | X = x]$, therefore we set $t^*(X) = (1, X, X^2)'$. This gives $t(X) = t(X)$ and $r^*(X) = X^2$. In this case we fit a propensity score model with an index that is quadratic in X despite the fact that Assumption 1.5 says that a linear one will suffice. We fit this model not by CMLE but by choosing $\widehat{\delta}_{IPT}$ to solve (14). Once we have fitted our propensity score we compute $\widehat{\gamma}_{IPT}$ by choosing it to solve (13).

Now consider the case where the analyst believes that the propensity score might vary sharply with X so that Assumption 1.5 requires $r(X) = (1, X, X^2)'$, but that $\mathbb{E}[Y_1 | X = x]$ is linear in X so that Assumption 2.1 requires only $t^*(X) = (1, X)'$. In this case $t(X) = r(X)$ and $r^*(X)$ is empty. Here the added moment serves only to tie down the propensity score parameter; it does not increase the precision of $\widehat{\gamma}_{IPT}$. There is no need to overfit the propensity score in this case.

Table 1 summarizes the main operational differences between IPW and IPT. The main difference is that we overfit the propensity score if Assumption 2.1 requires us to do so and (ii) we do not use CMLE to fit the propensity score. In Appendix C we show that the first step of our procedure requires solving a globally concave programming problem with unrestricted domain. In theory this is no harder than computing the CMLE associated with a binary choice logit model and in practice we have found this step to be straightforward. The second step of our procedure, as with the standard IPW one, can be completed by any M-estimation program that is able

[15] It is similar in spirit to the introduction of 'tilting' parameters in the context of generalized empirical likelihood (GEL) estimation of overidentified moment condition models (e.g., Imbens, 1997; Kitamura and Stutzer, 1997; Newey and Smith, 2004). This observation is the source of inverse probability tilting's name.

Table 1: Operational comparison of standard IPW and IPT

	IPW	**IPT**
Panel A: Modelling assumptions		
Propensity score model, $p_0(x)$	$p_0(x) = G(r(x)'\delta_0^*)$ with $G(\cdot)$ and $r(x)$ as in Ass. 1.5	$p_0(x) = G(t(x)'\delta_0)$ with $t(x)$ the union of all distinct elements in $r(x)$ (Ass. 1.5) and $t^*(x)$ (Ass. 2.1)
Moment CEF, $q_0(x)$	Not modelled	$q_0(x) = \Pi_0^* t^*(x)$; determined by Ass. 2.1; choice of $t^*(x)$ influences propensity score model
Panel B: Estimation		
First stage (i.e., propensity score)	Conditional maximum likelihood	Method-of-Moments; see Equation (14).
Second stage	Weighted GMM on complete case ($D = 1$) subsample	Weighted GMM on complete case ($D = 1$) subsample
Panel C: Inference		
Covariance matrix estimation	Computed as an application of two-step GMM (see Wooldridge, 2007)	Computed as an application of two-step GMM (see Equation (44) in Appendix A)

NOTES: For an asymptotic comparison of IPW and IPT see Proposition 2.1 and Theorems 2.1 and 2.2.

to accept user-specified weights.

The next two theorems characterize the first order asymptotic properties of $\widehat{\gamma}_{IPT}$. The first result shows that when Assumptions 1.1 to 1.5, *and* Assumption 2.1 hold, the asymptotic variance of $\widehat{\gamma}_{IPT}$ equals $\mathcal{I}(\gamma_0)^{-1}$. More precisely $\widehat{\gamma}_{IPT}$ is locally efficient for γ_0 in the semiparametric model defined by Assumptions 1.1 to 1.5 *at* the DGP which also satisfies Assumption 2.1.

Equation (1) is the information bound for γ_0 without imposing the additional auxiliary Assumption 2.1. This assumption imposes restrictions on the joint distribution of the data not implied by the baseline model. If these restrictions are added to the prior used to calculate the efficiency bound, then it is generally possible to estimate γ_0 more precisely. We emphasize that our estimator is not efficient with respect to this augmented model. Rather it attains the bound defined by (1) if Assumption 2.1 *happens to be true* in the population being sampled from, but *is not part of the prior restriction* used to calculate the bound. Newey (1990, p. 114), Robins, Rotnitzky and Zhao (1994, p. 852 - 3) and Tsiatis (2006) discuss the concept of local efficiency in detail.[16] In what follows we will, for brevity, say $\widehat{\gamma}_{IPT}$ is locally efficient at Assumption 2.1.

Theorem 2.1 (Local Semiparametric Efficiency of $\widehat{\gamma}_{IPT}$) *Consider the semiparametric missing data model defined by Assumptions 1.1 to 1.5 and additional regularity conditions, then for $\widehat{\gamma}_{IPT}$ the solution to (13), (i) $\widehat{\gamma}_{IPT}$ is regular and (ii) locally efficient at Assumption 2.1 with $\sqrt{N}(\widehat{\gamma}_{IPT} - \gamma_0) \xrightarrow{D} \mathcal{N}\left(0, \mathcal{I}(\gamma_0)^{-1}\right)$.*

Proof. See Appendix A. ■

Theorem 2.1 indicates that $\widehat{\gamma}_{IPT}$ has good efficiency properties. By choosing the estimating equation for the propensity score with the properties of $\mathbb{E}[\psi(Z, \gamma_0)|X]$ in mind, efficiency improvements over the standard IPW estimator are possible.[17]

[16] Examples of well-known locally efficient estimators include two-stage least squares, which is locally efficient under homoscedastic errors and a linear first stage, Robinson's (1988) estimator for the partial linear regression model (under homoscedastic errors), and Powell's (1986) symmetrically trimmed least squares estimator for the truncated regression model with a conditionally symmetric disturbance (under homoscedastic normality).

[17] We comment that the standard IPW estimator is also locally efficient. However this occurs not at DGPs which satisfies Assumption 2.1, but rather at ones where $\mathbb{E}[\psi(Z, \gamma_0)|X]$ is linear in $r(X) \cdot G\left(r(X)' \delta_0^*\right)$. We find this condition a bit awkward from a modelling standpoint, however it does help to explain why IPW is often nearly efficient in Monte Carlo experiments where the outcome equation is a direct function of the propensity score (e.g., Busso, DiNardo, and McCrary, 2009). If the data are missing completely at random (MCAR) such that $p_0(x) = \Pr(D = 1) = Q_0$ for all $x \in \mathcal{X}$, then IPW and IPT will be locally efficient at the same DGPs as long as $r(X) = t(X)$.

Researchers will often have views about plausible forms for $\mathbb{E}\left[\psi\left(Z,\gamma_0\right)|\,X\right]$. As models for $\mathbb{E}\left[\psi\left(Z,\gamma_0\right)|\,X\right]$ are central to imputation methods, insight into their formulation may be garnered from that literature (Little and Rubin, 2002).

Our next Theorem shows that IPW has a double robustness property (cf., Bang and Robins, 2005; Tsiatis, 2006; Wooldridge, 2007). Restrictions (11) and (12) were derived under the baseline missing data model defined by Assumptions 1.1 to 1.5. Consequently *regardless* of whether Assumption 2.1 also holds $\widehat{\gamma}_{IPT}$ will be consistent for γ_0 and asymptotically normal.[18] This is the first part of double robustness.

Now consider a DGP where Assumptions 1.1 to 1.4 and 2.1, but not 1.5 hold. That is a situation where the propensity score is misspecified but the implicit moment CEF model is not. In this case $\widehat{\delta} \xrightarrow{p} \delta_*$ where δ_* is the pseudo-true value which solves (12). This pseudo true value has an interesting property. Rearranging we get

$$\mathbb{E}\left[\frac{D}{G(t\left(X\right)'\delta_*)}t\left(X\right)\right] = \mathbb{E}\left[t\left(X\right)\right]. \tag{15}$$

The inverse probability weighted mean of $t\left(X\right)$ in the $D=1$ *subpopulation* coincides with its full population mean, $\mathbb{E}\left[t\left(X\right)\right]$. This property holds *regardless* of whether the true propensity score is of the form $G(t\left(X\right)'\delta)$ for some $\delta=\delta_0$.

In the sample, rearranging (14), we get

$$\sum_{i=1}^{N}\widehat{\pi}_{IPT,i}t\left(X_i\right) = \frac{1}{N}\sum_{i=1}^{N}t\left(X_i\right), \quad \widehat{\pi}_{IPT,i} = \frac{1}{N}\frac{D_i}{G(t\left(X_i\right)'\widehat{\delta}_{IPT})}, \tag{16}$$

so that the inverse probability weighted mean of $t\left(X\right)$ in the $D=1$ complete case *subsample* coincides with its full sample mean, $\sum_{i=1}^{N}t\left(X_i\right)/N$. By choosing the propensity score parameter to solve (14) we ensure that the estimated inverse probability weights satisfy an *exact balancing* property. For example, if $t\left(X_i\right)=\left(1,X,X^2\right)'$ with X scalar, then, after reweighting the complete case sample with $\widehat{\pi}_{IPT,i}$, the mean and variance of X will coincide with their full sample counterparts. Since the first element of $t\left(X_i\right)$ is a constant, the $\widehat{\pi}_{IPT,i}$ weights will also sum to 1.

Let $F\left(x,y_1\right)$ be the joint distribution of X, Y_1, then

$$\widehat{F}_{IPT}\left(x,y_1\right) = \sum\nolimits_{i=1}^{N}\widehat{\pi}_{IPT,i}\mathbf{1}\left(Y_{1i}\leq y_1\right)\mathbf{1}\left(X_i\leq x\right), \tag{17}$$

[18] Its asymptotic variance, however, will lie above $\mathcal{I}\left(\gamma_0\right)^{-1}$, in the matrix sense, unless Assumption 2.1 also holds.

is the estimate for the joint distribution of X and Y_1 implied by the IPT estimator (cf., Back and Brown, 1993; Imbens, 1997). By (16) this distribution function satisfies the exact balancing condition

$$\int t(x)\,\mathrm{d}\widehat{F}_{IPT}(x,y_1) = \int t(x)\,\mathrm{d}F_N(x), \tag{18}$$

where $F_N(x)$ is the full sample empirical distribution function of X. Since $F_N(x)$ is an efficient estimate of the distribution of X, it is reassuring that $\widehat{F}_{IPT}(x,y_1)$ satisfies (18). We discuss the properties of $\widehat{F}_{IPT}(x,y_1)$ further in Section 3.

The exact balancing property of $\widehat{F}_{IPT}(x,y_1)$ implies that $\widehat{\gamma}_{IPT}$ may be consistent for γ_0, even if the maintained propensity score model is incorrect. Let $\Pi_0 = (\Pi_0^*, \underline{0})$, under Assumption 2.1 we have $\Pi_0\mathbb{E}[t(X)] = \Pi_0^* t^*(X) = \mathbb{E}[\psi(Z,\gamma_0)]$. Using this equality, Assumption 1.3, and exact balancing (15) we get

$$\begin{aligned}
\mathbb{E}\left[\frac{D\psi(Z,\gamma)}{G(t(X)'\delta_*)}\right] &= \mathbb{E}\left[\frac{p_0(X)\,\psi(Z,\gamma)}{G(t(X)'\delta_*)}\right] \\
&= \mathbb{E}\left[\frac{p_0(X)}{G(t(X)'\delta_*)}\left\{\psi(Z,\gamma) - \Pi_0 t(X)\right\}\right] \\
&= \mathbb{E}\left[\frac{p_0(X)}{G(t(X)'\delta_*)}\left\{\mathbb{E}[\psi(Z,\gamma)|\,X] - \mathbb{E}[\psi(Z,\gamma_0)|\,X]\right\}\right] = 0
\end{aligned} \tag{19}$$

Therefore $\gamma = \gamma_0$ is a solution to the inverse probability weighted population moment even if there is no δ_0 such that $G(t(x)'\delta_0) = p_0(x)$ for all $x \in \mathcal{X}$. This is the second part of double robustness.

If $\psi(Z,\gamma)$ is linear in γ, as in Examples 1.1 and 1.2 above, then $\gamma = \gamma_0$ uniquely solves (19). In the general nonlinear case ensuring uniqueness of the solution to (19) may require the imposition of additional conditions, depending on the form of $\psi(Z,\gamma)$. As such conditions are model-specific we do not formulate them here, but note that doing so if facilitated by the fact that Assumption 1.4 and part (iv) of Assumption 1.5 ensures that $p_0(x)/G(t(x)'\delta_*)$ is bounded below by some positive constant.[19] Proceeding under the assumption that $\gamma = \gamma_0$ uniquely solves (19) we get our second result.

[19] Wooldridge (2001, pp. 458 - 459) develops conditions for consistency of unweighted M-estimators when the underlying sample is a stratified random one. His argument could be adapted to the current setting for cases where $E[\psi(Z,\gamma_0)] = 0$ corresponds to the first order condition of a population optimization problem.

Theorem 2.2 *(*Double Robustness of $\widehat{\gamma}_{IPT}$*) Suppose Assumptions 1.1 to 1.4, either Assumption 1.5 or 2.1, $\gamma = \gamma_0$ uniquely solves (19), and additional regularity conditions hold, then $\sqrt{N}(\widehat{\gamma}_{IPT} - \gamma_0) \xrightarrow{D} \mathcal{N}(0, \Psi_0)$, where the form of Ψ_0 depends on whether Assumption 1.5 or 2.1 holds (see Appendix A).*

Proof. See Appendix A. ■

Our formulation of the IPT estimator was undertaken with efficiency considerations at the forefront. This led to an approach where the propensity score was parameterized with two concerns in mind. First, the parametric propensity score family needs to be rich enough to contain the true score. Second, it needs to be rich enough to balance those functions of X which enter the CEF of $\psi(Z, \gamma_0)$. Theorem 2.2 shows that the dividend to this approach extends beyond local efficiency. Even if the propensity score is misspecified, IPT will remain consistent if $\mathbb{E}[\psi(Z, \gamma_0)|\, X]$ is linear in $t(X)$. More heuristically Theorem 2.2 suggests that IPT will perform well for moderately rich forms of $t(X)$ when *either* the propensity score or the conditional expectation of $\psi(Z, \gamma_0)$ are smooth in X. Researchers should choose $t(X)$ to be rich enough so that it accurately approximates whichever function, either $p_0(x)$ or $q_0(x) = \mathbb{E}[\psi(Z, \gamma_0)|\, X = x]$, is believed to be the least smooth.

We emphasize that consistency of the standard IPW estimate *requires* Assumption 1.5 to hold, while that of the parametric imputation estimate *requires* Assumption 2.1 to hold, in contrast, the IPT estimate is consistent if *either* one (or both) hold. Put differently while IPW and imputation respectively perform best when the propensity score and $\mathbb{E}[\psi(Z, \gamma_0)|\, X]$ are smooth in X, IPT will work well if either of the two objects are smooth. This is the practical content of 'double robustness'.

3 Other alternatives to IPW

Theorems 2.1 and 2.2 provide one argument for routine use of IPT: it is (i) more robust than either standard IPW or parametric imputation and (ii) locally efficient at Assumption 2.1. Computationally it is no harder than standard IPW (see Appendix C). Finally the exact balancing property is likely to be attractive to applied researchers. It is consistent with the intuition that reweighting makes the complete case subsample more like the full sample. Tables which assess balance after IPW are commonly featured in applied work (e.g., Hirano and Imbens, 2001; see also Appendix F in the Supplemental Web Appendix).

AIPW Estimator	$\omega_i(\delta)$	$\nu_i(\delta)$	Locally Efficient?	Doubly Robust?
Robins, Rotnitzky, and Zhao (1994)	$G_i(\delta)$	$\frac{D_i}{G_i(\delta)}$	Yes	Yes
Newey (1994a)	1	1	Yes	No
Cao, Tsiatis and Davidian (2009)	$\frac{1-G_i(\delta)}{G_i(\delta)}$	$\frac{D_i}{G_i(\delta)}$	Yes	Yes
Hirano and Imbens (2001) / Wooldridge (2007)	1	$\frac{D_i}{G_i(\delta)}$	Yes	Yes

Table 2: Weight functions for different AIPW estimators

While the argument privileging IPT over IPW appears to be straightforward, other alternatives to IPW exist. One such alternative is the class of augmented inverse probability weighting (AIPW) estimators introduced by Robins, Rotnitzky, and Zhao (1994). Like IPT, AIPW is locally efficient at Assumption 2.1. It is also doubly robust; consistency requires either 1.5 or 2.1 to hold. In this section we present two theoretical arguments for privileging our IPT method over AIPW ones. First we show that the implicit estimate of the joint distribution of X and Y_1 associated with IPT is attractive relative to the ones associated with AIPW. Second we compare the higher order bias of the two types of estimators.

3.1 A class of iterated AIPW estimators

Several versions of AIPW are now available. Here we describe a general set-up which captures many of them. Let $\omega_i(\delta) = \omega(X_i, \delta)$ and $\nu_i(\delta) = \nu(D_i, X_i, \delta)$ be known, scalar-valued, nonnegative weight functions. We require that $\nu(D_i, X_i, \delta)$ is such that $\mathbb{E}[\nu(D_i, X_i, \delta)|X] = 1$. Our family of AIPW estimators will be indexed by these two weight functions. Let $\widehat{\gamma}_{(\nu,\omega)}$ be an AIPW estimate in the family, which is defined as the solution to

$$\frac{1}{N}\sum\nolimits_{i=1}^{N}\left\{\frac{D_i}{G_i(\widehat{\delta}_{ML})}\psi\left(Z_i,\widehat{\gamma}_{(\nu,\omega)}\right)-\frac{\widehat{q}_{(\upsilon,\omega)}\left(X_i;\widehat{\gamma}_{(\nu,\omega)}\right)}{G_i(\widehat{\delta}_{ML})}\left(D_i-G_i(\widehat{\delta}_{ML})\right)\right\}=0, \quad (20)$$

with $\widehat{\delta}_{ML}$ the CMLE of the propensity score parameter. That is the solution to

$$\frac{1}{N}\sum_{i=1}^{N}\frac{D_i-\widehat{G}_i}{\widehat{G}_i(1-\widehat{G}_i)}\widehat{G}_{1i}t_i=0,$$

and

$$\widehat{q}_{(\upsilon,\omega)}(x;\gamma) = \left[\frac{1}{N}\sum_{i=1}^{N}\frac{D_i}{\widehat{G}_i}\widehat{\omega}_i\psi_i(\gamma)\,t_i'\right] \times \left[\frac{1}{N}\sum_{i=1}^{N}\widehat{\nu}_i\widehat{\omega}_i t_i t_i'\right]^{-1} t(x),$$

with $\widehat{G}_i = G_i(\widehat{\delta}_{ML})$, $\widehat{\nu}_i = \nu_i(\widehat{\delta}_{ML})$ and $\widehat{\omega}_i = \omega_i(\widehat{\delta}_{ML})$. Note that $\widehat{q}_{(\upsilon,\omega)}(x;\gamma)$ is the fitted value associated with a weighted least squares fit of $\psi_i(\gamma)$ onto t_i.

Setting $\nu_i(\delta) = D_i/G_i(\delta)$ and $\omega_i(\delta) = G_i(\delta)$ we get the original AIPW estimator of Robins, Rotnitzky and Zhao (1994); $\nu_i(\delta) = 1$ and $\omega_i(\delta) = 1$ yields Newey's (1994, Section 5.3) estimator, while $\nu_i(\delta) = D_i/G_i(\delta)$ and $\omega_i(\delta) = (1 - G_i(\delta))/G_i(\delta)$ gives the estimator suggested by Cao, Tsiatis and Davidian (2009) (see Table 2).[20]

Hirano and Imbens (2001) and Wooldridge (2007) propose a doubly robust estimator for the average treatment effect under exogenous treatment assignment.[21] It turns out that setting $\nu_i(\delta) = D_i/G_i(\delta)$ and $\omega_i(\delta) = 1$ gives their estimator. In the general moment model case their approach chooses $\widehat{\gamma}_{HIW}$ to solve

$$\frac{1}{N}\sum\nolimits_{i=1}^{N}\widehat{q}_{HIW}(X_i;\widehat{\gamma}_{HIW}) = 0 \tag{21}$$

where $\widehat{q}_{HIW}(x;\gamma)$ is the weighted least squares fit

$$\widehat{q}_{HIW}(x;\gamma) = \left[\frac{1}{N}\sum_{i=1}^{N}\frac{D_i}{\widehat{G}_i}\psi_i(\gamma)\,t_i'\right] \times \left[\frac{1}{N}\sum_{i=1}^{N}\frac{D_i}{\widehat{G}_i}t_i t_i'\right]^{-1} t(x). \tag{22}$$

The following Proposition shows that (21) is also a member of our class of AIPW estimators.

Proposition 3.1 *The solution to (21) is numerically identical to $\widehat{\gamma}_{(\nu,\omega)}$ with $\nu_i(\delta) = D_i/G_i(\delta)$ and $\omega_i(\delta) = 1$.*

Proof. Since the first element of t_i is a constant we have, by the first order condition associated with (22),

$$\frac{1}{N}\sum\nolimits_{i=1}^{N}\frac{D_i}{\widehat{G}_i}\left\{\psi(Z_i,\widehat{\gamma}) - \widehat{q}_{HIW}(x;\gamma)\right\} = 0. \tag{23}$$

[20] Many of the estimators listed in Table 2 were originally proposed in the context of a specific form for $\psi(Z,\gamma)$. We adapt to the general case as necessary. Newey (1994a) derives the large sample properties of his estimator where the dimension of $t(X)$ grows with N. Here we consider his estimator with the dimension of $t(X)$ held fixed.

[21] Wooldridge's (2007) estimator is actually slightly more general than the one described here in that $\widehat{q}_{HIW}(x;\gamma)$ need not correspond to a least squares fit.

Adding the left-hand side of (23) to (21) and re-arranging gives the result. ■

Finally note that Procedure 2.1, described in the previous section, is closely related to (20). Indeed they coincide if steps (iii) to (v) of Procedure 2.1 are iterated over until convergence. The different AIPW estimators listed in Table 2 are then recovered by using the appropriate weights when estimating $q(x;\gamma_0)$ in step (iii).

3.2 Implicit distribution function estimates

A useful way to understand the properties of first order equivalent estimators is in terms of their implicit distribution function estimates (e.g., Back and Brown, 1993; Imbens, 1997; Newey and Smith, 2004). After some simple algebra we can show that the solution to (20) coincides with that to

$$\sum\nolimits_{i=1}^{N} \widehat{\pi}_{(\upsilon,\omega),i}\psi(Z_i,\widehat{\gamma}_{(\upsilon,\omega)}) = 0,$$

where

$$\widehat{\pi}_{(\upsilon,\omega),i} = \frac{1}{N}\frac{D_i}{\widehat{G}_i}\widehat{\zeta}_{(\upsilon,\omega),i}, \qquad i = 1,\ldots,N \tag{24}$$

with

$$\widehat{\zeta}_{(\upsilon,\omega),i} = \left\{1 - \left[\frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i}{\widehat{G}_i} - 1\right)t_i\right]' \times \left[\frac{1}{N}\sum_{i=1}^{N}\widehat{\nu}_i\widehat{\omega}_i t_i t_i'\right]^{-1} \times \widehat{\omega}_i t_i\right\}, \qquad i = 1,\ldots,N. \tag{25}$$

This implies that the estimate of the joint distribution associated with $\widehat{\gamma}_{(\upsilon,\omega)}$ is

$$\widehat{F}_{(\upsilon,\omega)}(x,y_1) = \sum\nolimits_{i=1}^{N} \widehat{\pi}_{(\upsilon,\omega),i}\mathbf{1}(Y_{1i} \leq y_1)\,\mathbf{1}(X_i \leq x), \tag{26}$$

(see Back and Brown, 1993, Proposition 1).

This distribution function has several interesting properties. First if $\nu_i = D_i/G_i(\delta)$, which is true for all the estimators listed in Table 2 except Newey's (1994a), then

$$\int t(x)\,\mathrm{d}\widehat{F}_{(\upsilon,\omega)}(x,y_1) = \int t(x)\,\mathrm{d}F_N(x).$$

The re-weighted mean $t(X_i)$ in the complete case ($D = 1$) subsample coincides with its unweighted full sample mean. Since the unweighted full sample mean of $t(X_i)$ is an efficient estimate of its population analog, then so is the re-weighted complete case

sample mean. In this sense the $\widehat{F}_{(\upsilon,\omega)}(x, y_1)$ inherits some of the efficiency properties of $F_N(x)$. Since the first element of $t(X_i)$ is 1 the AIPW distribution function estimate also integrates to 1 (i.e., $\int \mathrm{d}\widehat{F}_{(\upsilon,\omega)}(x, y_1) = 1$).

As noted in the previous section the IPT distribution function estimate (17) also exactly balances the mean of $t(X_i)$ and integrates to one. However, it differs from $\widehat{F}_{(\upsilon,\omega)}(x, y_1)$ in that it is guaranteed to be non-decreasing, whereas $\widehat{F}_{(\upsilon,\omega)}(x, y_1)$ may be decreasing in x and/or y_1 over some ranges. Put differently some of the $\widehat{\pi}_{(\upsilon,\omega),i}$ weights may be negative, while the $\widehat{\pi}_{IPT,i}$ weights are positive by construction.

To gain further insight into this problem consider the distribution function estimator associated with standard IPW (e.g., Imbens, 2004):

$$\widehat{F}_{IPW}(x, y_1) = \sum\nolimits_{i=1}^{N} \widehat{\pi}_{IPW,i} \mathbf{1}(Y_{1i} \leq y_1) \mathbf{1}(X_i \leq x), \qquad \widehat{\pi}_{IPW,i} = \frac{1}{N}\frac{D_i}{\widehat{G}_i}. \tag{27}$$

Now consider a random sample where

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i}{\widehat{G}_i} - 1\right) t(X_i) > 0 \Leftrightarrow \sum_{i=1}^{N} \widehat{\pi}_{IPW,i} t(X_i) > \frac{1}{N}\sum_{i=1}^{N} t(X_i). \tag{28}$$

In this case the IPW estimate of the mean of $t(X_i)$ exceeds its unweighted full sample counterpart. The fact that the latter mean is efficient, implies that former is not. The AIPW distribution function estimator corrects this inefficiency by adjusting the IPW weights as follows

$$\widehat{\pi}_{(\upsilon,\omega),i} = \widehat{\pi}_{IPW,i} \times \widehat{\zeta}_{(\upsilon,\omega),i},$$

with $\widehat{\zeta}_{(\upsilon,\omega),i}$ as defined in (25). Under (28) large realizations of $t(X_i)$ are 'too frequent' in the complete case subsample (even after reweighting by the inverse of the estimated propensity score). In such a situation $\widehat{\zeta}_{(\upsilon,\omega),i}$ will be less than one for $D_i = 1$ units with large values of $t(X_i)$ and greater than one for units with small values. In extreme cases the resulting $\widehat{\pi}_{(\upsilon,\omega),i}$ may be negative or exceed one. Condition (28) is especially likely to occur when the propensity score model is misspecified. In that case $\widehat{G}_i$ corresponds to a quasi-MLE propensity score estimate and hence $\frac{1}{N}\sum_{i=1}^{N}\left(D_i/\widehat{G}_i - 1\right) t(X_i)$ may differ from zero even in large samples.

In practice the IPW and AIPW distribution functions can generate nonsensical estimates. Let $\psi(Z, \gamma) = Y_1 - \gamma$. Neither $\widehat{\gamma}_{IPW}$ and $\widehat{\gamma}_{(\upsilon,\omega)}$ are guaranteed to lie within the convex hull of the data. If $Y_1 \in \{0, 1\}$, for example, this means it is possible

for $\widehat{\gamma}_{IPW}$ and $\widehat{\gamma}_{(\upsilon,\omega)}$ to exceed one. In contrast $\widehat{\gamma}_{IPT}$ will lie in the convex hull of the data by construction. In our view an estimator which sets a weighted mean of $\psi(Z,\gamma)$ equal to zero, where these weights need not lie on the unit interval is *a priori* unattractive (cf., Robins, Sued, Lei-Gomez and Rotnitzky 2007).

3.3 Higher order bias

Another way IPT and AIPW can be compared is in terms of their higher order bias. In this section we present higher order bias expressions for both IPT and AIPW when Assumptions 1.1 to 1.5 and Assumption 2.1 hold. Bias comparisons are interesting in this case because IPT and AIPW are first order equivalent. Theorem 3.1, which is based on an application of Lemma A.4 of Newey and Smith (2004), gives the result.

Theorem 3.1 (Higher Order Bias) *Suppose Assumptions 1.1 to 1.5, Assumption 2.1, and additional regularity conditions hold, then*

$$\text{Bias}(\widehat{\gamma}_{(\upsilon,\omega)}) = C_O + C_V(\upsilon,\omega) \tag{29}$$

$$\text{Bias}(\widehat{\gamma}_{IPT}) = C_O, \tag{30}$$

where

$$\begin{aligned}
C_O &= -\frac{1}{2N}\sum_{k=1}^{K}\Gamma^{-1}\mathbb{E}\left[\frac{\partial^2\psi}{\partial\gamma_k\partial\gamma'}\right]\mathcal{I}(\gamma_0)^{-1}e_k \\
&\quad + \frac{1}{N}\Gamma^{-1}\mathbb{E}\left[\frac{\partial\psi}{\partial\gamma'}\Gamma^{-1}\frac{1}{G}\{\psi-\Pi_0 t\}\right] + \frac{1}{N}\Gamma^{-1}\mathbb{E}\left[\frac{\partial\psi}{\partial\gamma'}\Gamma^{-1}\Pi_0 t\right] \\
C_V(\upsilon,\omega) &= -\frac{\Gamma^{-1}}{N}\mathbb{E}\left[\frac{D}{G^2}\Sigma(X)\Lambda^{-1}\Pi_S S_\delta\right] \\
&\quad + \frac{\Gamma^{-1}}{N}\mathbb{E}\left[\left\{\frac{D}{G}\left(2\omega-\frac{1}{G}\right)-\omega v\right\}\Pi_0 tt'\Pi_0'\Lambda^{-1}\Pi_S S_\delta\right] \\
&\quad + \frac{\Gamma^{-1}}{N}\Pi_0\mathbb{E}\left[\omega\left(\frac{D}{G}-\nu\right)\left(\frac{D}{G}-1\right)tt'F_0^{-1}t\right],
\end{aligned}$$

with e_k denoting a $K\times 1$ vector with a 1 in the k^{th} row and zeros elsewhere.

Proof. See Appendix B. ■

These bias expressions have intuitive interpretations. The first bias component of $\widehat{\gamma}_{(\upsilon,\omega)}$ corresponds to the bias of the (infeasible) solution to

$$\frac{1}{N}\sum\nolimits_{i=1}^{N} A^{\text{opt}}\left(X_i;\gamma_0\right)\rho\left(Z_i,\widehat{\gamma}_I,p_0\left(X\right)\right)=0,$$

with $A^{\text{opt}}\left(X;\gamma_0\right)$ and $\rho\left(Z_i,\gamma,p_0\left(X\right)\right)$ as defined in Section 2. This estimator is infeasible because it uses the true optimal combination of moments, $A^{\text{opt}}\left(X_i;\gamma_0\right)$, and the true propensity score, $p_0\left(X\right)$. Theorem 3.1 shows that the higher order bias of $\widehat{\gamma}_{IPT}$ and this infeasible oracle estimator coincide. The coincidence is a consequence of the just identified nature of $\widehat{\gamma}_{IPT}$.

The additional term in (29) is due to sampling error in $A^{\text{opt}}(X;\widehat{\gamma}_{(\upsilon,\omega)})$ and $G(t\left(X\right)'\widehat{\delta}_{ML})$. Intuitively, the additional higher order bias in $\widehat{\gamma}_{(\upsilon,\omega)}$ arises from AIPW's separation of the tasks of (i) propensity score estimation and (ii) imposition of the optimal set of moments implied by (6). The first task generates no gains in terms of asymptotic precision, while at the same time introducing sampling error into the vector of estimating equations for $\widehat{\gamma}_{(\upsilon,\omega)}$. The second task results in an overidentified system of moment equations. The finite sample properties of $\widehat{\gamma}_{(\upsilon,\omega)}$ may degrade as a result. It is straightforward to constructed stylized examples where the bias of $\widehat{\gamma}_{(\upsilon,\omega)}$ increases with M, the dimension of $t\left(X\right)$, while that of $\widehat{\gamma}_{IPT}$ does not. This will be especially true if the distribution of $t\left(X\right)$ is skewed and/or that of $\psi\left(Z,\gamma_0\right)$ is heteroscedastic.

The contrast between the higher order bias of $\widehat{\gamma}_{IPT}$ and $\widehat{\gamma}_{(\upsilon,\omega)}$ in some ways parallels that between empirical likelihood (EL) and two-step GMM for general moment condition models (Newey and Smith, 2004). The empirical likelihood estimator transforms an overidentified moment conditional problem into a just-identified one by introducing a vector of tilting parameters (cf., Imbens, 1997). Our approach to overidentification, in contrast, involves overparameterizing the propensity score. The idea of overfitting a nuisance function to eliminate overidentification appears to be novel.

IPT, which heavily exploits the special structure of the missing data problem, has important computational advantages relative to a standard application of EL to the vector of moment equations suggested by the Corollary to Proposition 2.2. Let $L_r=\dim\left(r\left(X\right)\right)$ and $L_{t^*}=\dim\left(t^*\left(X\right)\right)$. Such an approach would apply EL to the

$K + L_{t^*} + L_r$ system of moments

$$\mathbb{E}\left[\begin{array}{c} \frac{D}{G(r(X)'\delta_0^*)}\psi\left(Z,\gamma_0\right) \\ \left(\frac{D}{G(r(X)'\delta_0^*)}-1\right)t^*\left(X\right) \\ \left(\frac{D-G(r(X)'\delta_0^*)}{G(r(X)'\delta_0^*)\left[1-G(r(X)'\delta_0^*)\right]}\right)G_1(r\left(X\right)'\delta_0^*)r\left(X\right) \end{array}\right] = 0.$$

Computation of $\widehat{\gamma}_{EL}$ would involve solving a saddle point problem with $2\left(K + L_r\right) + L_{t^*}$ parameters (Newey and Smith, 2004; Section 2). In contrast computing $\widehat{\gamma}_{IPT}$ requires solving a $1 + M \leq L_{t^*} + L_r$ dimensional globally concave problem and a just-identified moment condition problem with K parameters. Relative to EL, our approach substantially reduces the parameter space and sidesteps the need to solve a saddle point problem.

4 Basic skills and the Black-White wage gap

In an important pair of papers Neal and Johnson (1996) and Johnson and Neal (1998) document that a substantial portion of the Black-White gap in hourly earnings can be accounted for by differences in basic skills across the two groups acquired *prior* to labor market entry (i.e., by age 18). In particular they find that three fifths of the raw 28 percent Black-White gap in average hourly earnings can be accounted for by differences in Armed Forces Qualification Test (AFQT) scores, a measure of basic skills used by the military. Their finding suggests that pre-market differences in basic skills across racial groups may be relatively more important drivers of racial inequality than overt labor market discrimination.

A corollary to their finding is that further reductions in Black-White inequality will likely require interventions which facilitate skill acquisition among young Black men. A related question is when do those skill differences important for labor market outcomes first arise. A large literature documents that, in terms of standardized test scores, Blacks enter school behind their White counterparts and that this gap grows over time (e.g., Fryer and Levitt, 2006, forthcoming). The connection between the timing and magnitude of these gaps and labor market outcomes is not well-understood.

Here we repeat Johnson and Neal 's (1998) analysis after replacing AFQT scores

with measures of cognitive skills acquired *prior* to adolescence. The idea is to measure how much of Black-White differences in hourly earnings can be accounted for by differences in skills across the two groups manifest prior to adolescence. If a substantial portion of the wage gap can be so accounted for, then educational interventions which aim to ameliorate racial inequality might be more appropriately targeted toward younger children (cf., Carneiro and Heckman, 2004).

We reconstruct the National Longitudinal Survey of Youth 1979 (NLSY79) extract analyzed by Johnson and Neal (1998). This sample is a stratified random sample of young men from the United States born between 1962 and 1964. Measurements of average hourly wages over the 1990 to 1993 period, race, as well as AFQT scores are available for each individual. The NLSY79 also collected data from respondents' school records. In some cases these records included (nationally normed) percentile scores on IQ tests taken at various ages. We use those scores corresponding to tests taken between the ages of 7 and 12 as measures of cognitive skills acquired prior to adolescence.[22] Unfortunately these scores are missing for almost 90 percent of individuals. An unweighted analysis based on those individuals with complete information would be problematic for two reasons: (i) there are few complete cases making precise inference difficult and (ii) the complete cases are not representative of the full sample in terms of always-observed characteristics. Our IPT estimator is designed to address both of these problems.

Dataset description and replication of Johnson and Neal (1998) Our initial goal was to reconstruct the NLSY79 extract used by Johnson and Neal (1998). A preliminary inspection of the data, however, revealed that a preadolescence test score was available for only a handful of Hispanic respondents. We therefore decided to exclude Hispanics from our analysis. We targeted all non-Hispanic male respondents in the cross-sectional sample, as well as in the supplemental Black sample, born during or after 1962 ($N = 1,612$). These individuals were aged 16 to 18 when they took the NLSY79's administration of the Armed Services Vocational Aptitude Battery (ASVAB) from which the AFQT score is constructed.

Of the $1,612$ targeted individuals 159 were missing 1990 to 1993 average hourly

[22] If a respondent's record includes multiple test scores from the age 7 to 12 period, we use the average percentile score across all available tests. A *STATA* dictionary file of the NLSY79 extract used in our analysis as well as a do file which replicates the data manipulations described here is available online at `https://files.nyu.edu/bsg1/public/`.

wage data, 58 a valid AFQT score, and 24 both these items.[23] Our base sample therefore consists of the 1,371 individuals with valid wage and AFQT data. As in Johnson and Neal (1998) we condition on this non-response in what follows (cf., Johnson, Kitamura and Neal, 2000).

AFQT scores are reported in terms of normed percentiles (i.e., relative to population of American youths aged 18 to 23).[24] We transformed these scores onto the real line using the inverse normal CDF. The residual associated with the least squares fit of the transformed scores onto a vector of birth year dummies is our AFQT measure.[25] The average hourly wage for Whites in the base sample is $12.32 in 1993 prices. Blacks earn on average $2.87 less per hour. The mean AFQT score for Whites is 0.160, while that for Blacks is 1.031 standard deviations lower.

Columns 1 and 2 of Table 3 replicate Columns 1 and 2 of Table 14-1 in Johnson and Neal (1998, p. 483) (with the exception that we exclude Hispanics from our analysis).[26] The first column reports the least squares fit of LogWage onto a constant, YearOfBirth, and Black. The estimated wage gap between Blacks and Whites of the same age is 28 percent.

Column 2 adds AQFT to the set of explanatory variables. The wage gap between Blacks and Whites of the same age with the same pre-market AFQT score is only 11 percent. Seventeen percentage points of the unconditional Black-White hourly wage gap can be accounted for by average differences in pre-market AFQT scores across the two groups. That a substantial portion of racial differences in hourly wages can be accounted for by differences in skills acquired prior to entry into the labor market is Neal and Johnson's (1996) central result.

Racial wage gaps and preadolescent skill differences For 11 percent of the respondents ($N = 144$) in our base sample an IQ test score from between the ages of

[23] Hourly wages are equal to their average over the 1990 to 1993 survey waves. In cases where an individual was not employed or interviewed in a given year, the average is over those years with wage information. Reported wage values less that $1 and greater than $75 per hour are discarded. Wages are measured in 1993 dollars. As in Johnson and Neal (1998) we exclude any AFQT score where the testing protocol was non-standard.

[24] We used the 1989 scoring of the AFQT.

[25] Neal and Johnson (1996) appear to have used the residual associated with the least squares fit of AFQT *percentiles* onto a set of birth year dummies (see their Figure 2, p. 886). Even after standardizing this variable to have mean zero and unit variance its distribution is non-normal. Our approach results in an AFQT score distribution that is indistinguishable from a normal one.

[26] See also Columns 1 and 3 of Table 1 in Neal and Johnson (1996, p. 875).

Table 3: Replication of Table 14-1 of Johnson and Neal (1998) and unweighted complete case analysis with pre-adolescent test score

	(1) OLS	(2) OLS	(3) $CC-OLS$	(4) $CC-OLS$
YearOfBirth	−0.0458	−0.0466	−0.0947	−0.0940
	$(0.0151)^{**}$	$(0.0147)^{**}$	$(0.0464)^{*}$	$(0.0470)^{*}$
Black	−0.2776	−0.1079	−0.2708	−0.1606
	$(0.0261)^{**}$	$(0.0284)^{**}$	$(0.0833)^{**}$	$(0.0900)^{+}$
AFQT	–	0.1645	–	–
		$(0.0146)^{**}$		
EarlyTest	–	–	–	0.1011
				$(0.0539)^{+}$
R^2	0.062	0.183	0.068	0.120
N	1,371	1,371	144	144

NOTES: Estimation samples are as described in the main text. The 1979 baseline sampling weights are used in place of the empirical measure when computing all estimates. A '**', '*' and '+' denotes that a point estimate is significantly different from zero at the 1, 5 and 10 percent levels. Standard errors (in parentheses) allow for arbitrary patterns of heteroscedasticity and dependence across units residing in the same household at baseline.

7 and 12 is available. This score is recorded as a (nationally normed) percentile. As with the AFQT scores we transformed these scores onto the real line using the inverse normal CDF. The residual associated with the least squares fit of these transformed scores onto a vector of birth year and age when tested (in years) dummies is our EarlyTest measure. We view this score as measure of acquired cognitive skills not of innate ability (cf., Fryer and Levitt, forthcoming).

Columns 3 and 4 of Table 3 replicate the basic Johnson and Neal (1998) analysis after replacing AFQT with the earlier test score. This is an unweighted analysis based on the 144 complete cases. Conditioning on age alone, racial wage gaps in the complete case subsample are very similar to those computed using the full sample (Column 3). The wage gap conditional on the pre-adolescent test score is substantially lower (Column 4). Unfortunately these wage gap estimates are very imprecise; their standard errors are almost four times those of their Columns 1 and 2 counterparts.

A second problem with this analysis is that those individuals with early test scores differ systematically from those without them. This is shown in Table 4. Columns 1, 2 and 3 respectively report average values across the full sample, the complete

Table 4: Comparison of the full sample with the complete case subsample

	(1) Full Sample	(2) $D = 1$	(3) $D = 0$	(4) Difference
LogWage	6.984	6.999	6.982	0.0167 (0.0433)
YearOfBirth	62.99	62.91	62.99	−0.0842 (0.0824)
Black	0.155	0.111	0.160	−0.0490 (0.0218)*
AFQT	0.000	0.196	−0.025	0.2205 (0.0953)*
N	1,371	144	1,227	

Notes: Samples are as described in the main text. The 1979 baseline sampling weights are used when computing all summary statistics. A '**', '*' and '+' denotes that the Column 4 difference is significantly different from zero at the 1, 5 and 10 percent levels. Standard errors (in parentheses) allow for arbitrary patterns of heteroscedasticity and dependence across units residing in the same household at baseline.

case subsample ($D = 1$), and the subsample with missing early test scores ($D = 0$). Column 4 gives the Column 2 versus Column 3 difference. While the full and complete case samples are similar in terms of wages and age, they are rather different in terms of racial composition and AFQT.

To address bias due to non-randomness in the missingness process as well as to improve precision we re-estimated the Table 3, Column 4 model using our IPT procedure. To appropriately use IPT we require that EarlyTest is missing at random (Assumption 1.3). That is, conditional on YearOfBirth, Black, LogWage and AFQT, we require that the probability of observing EarlyTest is independent of its value. Since LogWage and AFQT are good measures of long run productivity and skills, characteristics which might inform the decision to be tested at a young age, this requirement is not unreasonable.

The the joint support of YearOfBirth and Black contains $3 \times 2 = 6$ points. We included in $t(X)$ five non-redundant dummy variables for YearOfBirth-by-Black cell (whites born in 1962 are the excluded group). This resulted in full distributional balance for the discretely-valued components of X. We also balanced the means, variance and covariance of AFQT and LogWage conditional on race alone, and age

alone, but not their interaction.[27] That is $t(X)$ also included AFQT, LogWage, AFQT^2, LogWage^2 and AFQT$\times$LogWage as well as the interactions of these variables with Black and the two year of birth dummies (1962 being the excluded cohort). This led to a specification of $t(X)$ with 26 elements.

Our choice of $t(X)$ was informed by two considerations. First, we wanted $t(X)$ to be rich enough to allow for rich forms for the propensity score (see Assumption 1.5) as well as for the conditional mean and variance of EarlyTest (see Assumption 2.1 and Example 1.2). Second, we wanted to reweight the 144 complete cases such that an analyst with access to these data alone would *numerically exactly reproduce* the results of Johnson and Neal (1998) (i.e., the point estimates in Columns 1 and 2 of Table 3). Our choice of $t(X)$ ensures that all those moments used to compute the full sample least squares fit are exactly balanced. Heuristically our hope is that if the point estimates associated with the Johnson and Neal (1998) specification when computed using the reweighted complete cases coincide with those computed using the unweighted full sample, that the reweighted complete case analysis with EarlyTest replacing AFQT will approximate the ideal, but infeasible, unweighted analysis based on full sample.

Column 2 of Table 5 reports IPT estimates of the best linear predictor of LogWage given, YearOfBirth, Black, and EarlyTest. For comparison the unweighted complete case estimates are reproduced in Column 1 of the table, while the standard inverse probability weighted (IPW) estimates are given in Column 3. Table 7 in Appendix F reports the underlying IPT and CMLE estimates of the propensity score parameter.

Relative to the unweighted complete case one, the IPT estimate of the Black-White wage gap, conditional on skills acquired prior to adolescence (EarlyTest), is larger in absolute value with a standard error almost two thirds smaller. Recall that the wage gap conditional on age alone was 28 percent (Table 3, Column 1). Conditioning on skills acquired prior to adolescence this gap falls to 18 percent. This is larger than the 11 percent gap present after conditioning on the later AFQT score, but substantially smaller than the unconditional gap. Put differently roughly 40 percent of the raw Black-White wage gap can be explained by differences in average skill levels across the two groups manifest prior to adolescence. This represents about

[27] Given the near normal distribution of AFQT and LogWage in our sample focusing on the first two moments of these variables seemed appropriate.

two-thirds of the pre-market effect found by Neal and Johnson (1996).

Column 3 of Table 5 reports IPW estimates of the same model. The IPW estimate of the Black-White wage gap is imprecisely determined with a standard error over twice as large as the IPT one. This provides a concrete example of the efficiency gains IPT can provide relative to IPW (see Proposition 2.1 and Theorem 2.1). Columns 4 through 7 report estimates based on the four implementations of AIPW described in Section 3. The AIPW point estimates, with the exception of Newey's (1994a), are very similar to their IPT counterpart, albeit with slightly larger standard errors.

In this particular example the implicit AIPW distribution function estimates are reasonably similar to the IPT one; AIPW does not give inordinate weight to any particular respondent and negative weight is attached to only a handful of units. The exception is Newey's (1994a) variant of AIPW. Theorem 3.1 suggests this variant of AIPW is more biased than the others, consistent with the empirical results.

Table 6 provides a synopsis of our main results. Conditioning on age alone the Black-White gap in wages is 28 percent (Column 1). Conditioning on AFQT, a measure of cognitive skills acquired by age 18, this gap falls to 11 percent. Conditioning on our early test measure, a measure of cognitive skills acquired by age 12, the gap is 18 percent. We interpret this result as implying that approximately two thirds of the pre-market effect found by Neal and Johnson (1996) reflects skill differences already present by age 12. This interpretation is justified by two additional pieces of evidence. First, when we include both AFQT and EarlyTest simultaneously in our model we find that the coefficient on the latter is insignificantly different from zero, while that on the former is insignificantly different from the estimate which conditions on AFQT alone (Column 4 of Table 6). This suggests that AFQT and EarlyTest measure similar types of skills, with the later measure being relevant for labor market outcomes. A similar interpretation is suggested by Table 9 in Appendix F of the Supplemental Web Appendix which shows that roughly two thirds of the AFQT gap across Blacks and Whites can be accounted for by skill differences present by age 12 (i.e., our EarlyTest variable).

Table 5: IPT, IPW and AIPW estimates of the Black-White wage gap conditional on preadolescent skills

	(1) $CC-OLS$	(2) IPT	(3) IPW	(4) $AIPW_{RRZ}$	(5) $AIPW_{NEWEY}$	(6) $AIPW_{HIW}$	(7) $AIPW_{CTD}$
YearOfBirth	−0.0940	−0.0537	−0.0968	−0.0533	−0.0353	−0.0535	−0.0543
	$(0.0470)^{*}$	$(0.0162)^{**}$	$(0.0302)^{**}$	$(0.0165)^{**}$	(0.0528)	$(0.0166)^{**}$	$(0.0167)^{**}$
Black	−0.1606	−0.1837	−0.2126	−0.1836	−0.0797	−0.1871	−0.1837
	$(0.0900)^{+}$	$(0.0356)^{**}$	$(0.0791)^{*}$	$(0.0418)^{**}$	(0.1518)	$(0.0392)^{**}$	$(0.0390)^{**}$
EarlyTest	0.1011	0.1112	0.1220	0.1049	0.0956	0.1072	0.1144
	$(0.0539)^{+}$	$(0.0296)^{**}$	$(0.0362)^{**}$	$(0.0358)^{**}$	(0.0360)	$(0.0346)^{**}$	$(0.0344)^{**}$

Notes: Samples are as described in the main text. The 1979 baseline sampling weights are used when computing all estimates. A '**', '*' and '+' denotes that a coefficient is significantly different from zero at the 1, 5 and 10 percent levels. Standard errors (in parentheses) allow for arbitrary patterns of heteroscedasticity and dependence across units residing in the same household at baseline.

Table 6: Pre-adolescent skills and adult wages: a synopsis

	(1) OLS	(2) OLS	(3) IPT	(4) IPT
YearOfBirth	−0.0458	−0.0466	−0.0537	−0.0443
	$(0.0151)^{**}$	$(0.0147)^{**}$	$(0.0162)^{**}$	$(0.0147)^{**}$
Black	−0.2776	−0.1079	−0.1837	−0.1132
	$(0.0261)^{**}$	$(0.0284)^{**}$	$(0.0356)^{**}$	$(0.0293)^{**}$
AFQT	–	0.1645	–	0.1866
		$(0.0146)^{**}$		$(0.0284)^{**}$
EarlyTest	–	–	0.1112	−0.0332
			$(0.0296)^{**}$	(0.0374)

Notes: Samples are as described in the main text. The 1979 baseline sampling weights are used when computing all estimates. A '**', '*' and '+' denotes that a coefficient is significantly different from zero at the 1, 5 and 10 percent levels. Standard errors (in parentheses) allow for arbitrary patterns of heteroscedasticity and dependence across units residing in the same household at baseline.

5 Extensions

Thus far we have focused on problems where Z is completely observed if $D = 1$. Now consider the case where $Z = (X', Y_0', Y_1')'$ with D, X and $Y = (1 - D) Y_0 + DY_1$ observed. That is we observe Y_0 if $D = 0$ and Y_1 if $D = 1$. Let the moment function take the separable form

$$\psi(z, \gamma) = \psi_1(y_1, x, \gamma) - \psi_0(y_0, x, \gamma).$$

Many problems fall into this basic set-up.

Example 5.1 (Average treatment effects (ATEs)) *Let $D = 1$ and $D = 0$ respectively denote assignment to an active and control program or intervention and Y_1 and Y_0 the corresponding potential outcomes. The Average Treatment Effect (ATE) is*

$$\gamma_0 = \mathbb{E}[Y_1 - Y_0],$$

which corresponds to setting $\psi_1(Y_1, X, \gamma) = Y_1$ and $\psi_0(Y_0, X, \gamma) = Y_0 + \gamma$. Since each unit can only be exposed to one intervention, either Y_1 or Y_0 is missing for all units.

Example 5.2 (Two sample instrumental variables (TSIV) estimation with

COMPATIBLE SAMPLES) *Assume that* $\dim(X) \geq \dim(Y_0)$ *and consider the following instrumental variables model*

$$Y_1 = Y_0'\gamma_0 + U, \qquad \mathbb{E}[UX] = 0.$$

This suggests a moment function with $\psi_1(Y_1, X, \gamma) = XY_1$ *and* $\psi_0(Y_0, X, \gamma) = XY_0'\gamma$. *Two independent random samples of size* N_1 *and* N_0 *from the same population are available. In the first sample* N_1 *values of* Y_1 *and* X *are recorded, while in the second* N_0 *values of* Y_0 *and* X *are recorded. For asymptotic analysis we assume that* $\lim_{N_1,N_0\to\infty} N_1/(N_1+N_0) = Q_0 > 0$. *This is the two-sample instrumental variables (TSIV) model analyzed by Angrist and Krueger (1992). Ridder and Moffitt (2007) provide a technical and historical overview (cf., Arellano and Meghir, 1992). This model is equivalent to a special case of the semiparametric missing data model, an observation that is apparently new. Assume* N *units are randomly drawn from some target population. With probability* Q_0 *the* i^{th} *unit's values for* Y_1 *and* X *are recorded, while with probably* $1-Q_0$ *its values of* Y_0 *and* X *are recorded. The indicator variable* D *denotes which set of variables are measured. The only difference between this sampling scheme and that of Angrist and Krueger (1992) is that in the latter* N_1 *and* N_0 *are fixed by the researcher, whilst in the missing data formulation they are random variables. An adaptation of the argument given by Imbens and Lancaster (1996, Sections 2.1-2.2) shows that this difference does not affect inference.*

To apply IPT to these problems we find the $\widehat{\delta}^0_{IPT}$, $\widehat{\delta}^1_{IPT}$ and $\widehat{\gamma}_{IPT}$ which solves

$$\frac{1}{N}\sum_{i=1}^{N}\left\{\frac{D_i\psi_1(Y_{1i}, X_i, \widehat{\gamma}_{IPT})}{G(t(X_i)'\widehat{\delta}^1_{IPT})} - \frac{(1-D_i)\psi_0(Y_{0i}, X_i, \widehat{\gamma}_{IPT})}{1-G(t(X_i)'\widehat{\delta}^0_{IPT})}\right\} = 0$$

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1-D_i}{1-G(t(X_i)'\widehat{\delta}^0_{IPT})} - 1\right)t(X_i) = 0$$

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i}{G(t(X_i)'\widehat{\delta}^1_{IPT})} - 1\right)t(X_i) = 0.$$

Note that this involves computing two propensity score parameter estimates. One which balances the mean of $t(X)$ in the $D=1$ subsample ($\widehat{\delta}^1_{IPT}$) with its full sample mean and one which balances the mean of $t(X)$ in the $D=0$ subsample with its full sample mean ($\widehat{\delta}^0_{IPT}$). Each of these propensity score estimates may be computed

using the algorithm described in Appendix C. The second step of estimation involves solving a just-identified moment condition problem.

It is straightforward to extend the arguments given above to show that the above estimator is locally efficient and doubly robust.[28] As before $t(X)$ should be rich enough to adequately model the propensity score. Local efficiency requires that $\mathbb{E}[\psi_0(Y_0, X, \gamma)|X]$ and $\mathbb{E}[\psi_1(Y_1, X, \gamma)|X]$ be linear in $t(X)$ (this is also the condition for double robustness). As in the examples outlined above the form of $t(X)$ is often suggested by the structure of the problem. Consider efficient estimation of the ATE by IPT. This requires choosing $t(X)$ such that the true propensity score is contained in the parametric family $G\left(t(X)'\delta\right)$ and the true potential outcome CEFs are linear in $t(X)$. Consistency requires only one of these two restrictions to be true.

[28] These results are contained in our NBER Working Paper (Egel, Graham and Pinto, 2008).

Appendix

This appendix outlines the proofs of the results given in the main text. Throughout the Appendix we assume that $t(X) = t^*(X) = r(X)$ so that $\Pi_0 = \left(\Pi_0^*, 0\right)$ and $\delta_0 = \left(\delta_0^*, 0\right)$. This is done only to simplify the notation and is without loss of generality. Appendix A outlines the proofs of the first order asymptotic properties of IPW and IPT stated in Section 2. Appendix B outlines our higher order bias derivations. Computation of the IPT propensity score estimate is detailed in Section C. All notation is as stated in the main text unless stated otherwise. A supplemental web appendix, available at `https://files.nyu.edu/bsg1/public/`, contains details of some of the more tedious calculations as well as additional tables related to the empirical application. Throughout we drop '0' subscripts, used to denote (true) population values, when doing so causes no confusion

A Derivation of the first order asymptotic properties of IPW, IPT and AIPW

This appendix details the derivation of the (first order) asymptotic properties of the inverse probability weighting (IPW) estimator of Wooldridge (2007), the inverse probability tilting (IPT) estimator introduced here, as well as those of a class of three-step augmented inverse probability weighting (AIPW) estimators.

A.1 The asymptotic variance of IPW (Proposition 2.1)

Wooldridge (2002, 2007) proves consistency and asymptotic normality of IPW M-estimators. Here we detail the derivation of the alternative variance expression given in Proposition 2.1, which clarifies that IPW is generally inefficient. Let $\beta = (\gamma', \delta')'$ and define the moment vector and derivative matrix

$$m_i(\beta) = \begin{pmatrix} \frac{D_i}{G_i(\delta)}\psi_i(\gamma) \\ \frac{D_i - G_i(\delta)}{G_i(\delta)(1-G_i(\delta))} G_{1i}(\delta)\, t_i \end{pmatrix}, \quad M_i(\beta) = \begin{bmatrix} \frac{D_i}{G_i(\delta)}\frac{\partial \psi_i(\beta)}{\partial \gamma'} & -\frac{D_i}{G_i(\delta)}\frac{G_{1i}(\delta)}{G_i(\delta)}\psi_i(\gamma)\, t_i' \\ 0 & -J_i(\delta) \end{bmatrix}, \tag{31}$$

where $J_i(\delta)$ is the i^{th} unit's contribution to the Hessian of the log-likelihood for the propensity score parameter δ. The solution to $\sum_{i=1}^{N} m_i\left(\widehat{\beta}\right)/N = 0$ corresponds to the IPW estimate. The covariance of m_i is given by

$$\Omega = \begin{pmatrix} \mathbb{E}\left[\frac{\psi\psi'}{G}\right] & \mathbb{E}\left[\frac{G_1}{G}\psi t'\right] \\ \mathbb{E}\left[\frac{G_1}{G} t\psi'\right] & \mathcal{I}(\delta) \end{pmatrix}, \tag{32}$$

while the population mean of M_i equals

$$M = \begin{pmatrix} \Gamma & -\mathbb{E}\left[\frac{G_1}{G}\psi t'\right] \\ 0 & -\mathcal{I}(\delta_0) \end{pmatrix}, \tag{33}$$

with $\mathcal{I}(\delta_0)$ the Fisher information for δ_0.

Standard results on GMM imply that $\sqrt{N}\left(\widehat{\beta} - \beta_0\right)$ has a limiting sampling variance of

$$M^{-1}\Omega M^{-1\prime} = \begin{pmatrix} \Gamma^{-1}\mathbb{E}\left[\frac{\psi\psi'}{G}\right]\Gamma^{-1\prime} - \Gamma^{-1}\mathbb{E}\left[\frac{G_1}{G}\psi t'\right]\mathcal{I}(\delta_0)^{-1}\mathbb{E}\left[\frac{G_1}{G} t\psi'\right]\Gamma^{-1\prime} & 0 \\ 0 & \mathcal{I}(\delta_0)^{-1} \end{pmatrix}. \tag{34}$$

An insightful rearrangement of the upper-left-hand block of (34) is (see the Supplemental Web Appendix)

$$\mathcal{I}(\gamma_0)^{-1} + \Gamma^{-1}\mathbb{E}\left[\left(\left(\frac{D}{G} - 1\right)q - \Pi_S S_\delta\right)\left(\left(\frac{D}{G} - 1\right)q - \Pi_S S_\delta\right)'\right]\Gamma^{-1\prime}, \tag{35}$$

with Π_S as defined in the statement of Proposition 2.1. Part (ii) of the result follows by inspection.

A.2 Local efficiency and double robustness of $\widehat{\gamma}_{IPT}$ (Theorems 2.2 and 2.1)

Consistency and double robustness We first consider the case where Assumptions 1.1 to 1.5 hold. In this case the consistency argument parallels that for the standard IPW estimator given in Wooldridge (2002, 2007). Assumptions 1.1 and 1.4 give

$$\mathbb{E}\left[\frac{|\psi(Z,\gamma)|}{p_0(X)}\right] \leq \kappa^{-1}\mathbb{E}\left[|\psi(Z,\gamma)|\right] < \infty,$$

Since $\kappa^{-1}|\psi(Z,\gamma)|$ dominates $D\psi(Z,\gamma)/p_0(X)$ we may apply the law of iterated expectations and the missing at random restriction (Assumption 1.3) to get

$$\mathbb{E}\left[\frac{D\psi(Z,\gamma)}{p_0(X)}\right] = \mathbb{E}\left[\psi(Z,\gamma)\right].$$

By Assumption 1.1 $\mathbb{E}[\psi(Z,\gamma)]$ is uniquely mean zero at $\gamma = \gamma_0$. Consistency follows from Theorem 2.6 of Newey and McFadden (1994).

Replacing $p_0(X)$ with the IPT estimate $\widehat{\delta}$ causes no real difficulties. We have $\widehat{\delta} \xrightarrow{p} \delta_0$ which gives the sample average of $D\psi(Z,\gamma)/G(t(X)'\widehat{\delta})$ converging uniformly in $\gamma \in \mathcal{G}$ to $\mathbb{E}[D\psi(Z,\gamma)/p_0(X)]$.

Next we consider the case where Assumptions 1.1 to 1.4 and 2.1 hold, but not 1.5 (we *do* assume that the $G(\cdot)$ function satisfies the stated regularity conditions; in particular that $G(t(x)'\delta) > 0$ for all $x \in \mathcal{X}$ and $\delta \in \mathcal{D}$). In this case the propensity score is misspecified such that $\widehat{\delta} \xrightarrow{p} \delta_*$ where δ_* is some pseudo-true value which solves $\mathbb{E}\left[\left(D/G(t(X)'\delta_*) - 1\right)t(X)\right] = 0$. This gives $\mathbb{E}\left[p_0(X)\,t(X)/G(t(X)'\delta_*)\right] = \mathbb{E}\left[t(X)\right]$ so that under Assumption 1.3 and 2.1 we have equation (19) of the main text. Therefore $\gamma = \gamma_0$ is a solution to the IPW population moment. If $\psi(Z,\gamma)$ is linear in γ, then this solution is also unique. Otherwise uniqueness follows by hypothesis.

Asymptotic normality Asymptotic normality of $\widehat{\gamma}_{IPT}$ follows under standard regularity conditions by Theorem 6.1 of Newey and McFadden (1994). As above let $\beta = (\gamma', \delta')'$. The $K + 1 + M \times 1$ moment vector and derivative matrix are now given by

$$\underset{K+1+M\times 1}{m_i(\beta)} = \begin{pmatrix} \frac{D_i}{G_i(\delta)}\psi_i(\gamma) \\ \left(\frac{D_i}{G_i(\delta)} - 1\right)t_i \end{pmatrix}, \quad M_i(\beta) = \begin{bmatrix} \frac{D_i}{G_i(\delta)}\frac{\partial\psi_i(\gamma)}{\partial\gamma'} & -\frac{D_i}{G_i(\delta)}\frac{G_{1i}(\delta)}{G_i(\delta)}\psi_i(\gamma)\,t_i' \\ 0 & -\frac{D_i}{G_i(\delta)}\frac{G_{1i}(\delta)}{G_i(\delta)}t_i t_i' \end{bmatrix}. \tag{36}$$

We first consider the case where Assumptions 1.1 to 1.5 hold. Let $M = \mathbb{E}[M_i(\beta_0)]$ and $\Omega = \mathbb{E}\left[m_i(\beta_0)\,m_i(\beta_0)'\right]$, then $\sqrt{N}(\widehat{\gamma} - \gamma_0) \xrightarrow{D} \mathcal{N}(0, \Psi_0)$ for $\Psi_0 = \left\{\left(M'\Omega^{-1}M\right)^{-1}\right\}_{1:K,1:K}$ (where $A_{1:K,1:K}$ is the upper left hand $K \times K$ block of A). The covariance of $m_i = m_i(\beta_0)$ equals

$$\Omega = \begin{pmatrix} \mathbb{E}\left[\frac{\psi\psi'}{G}\right] & E_0 \\ E_0' & F_0 \end{pmatrix}, \tag{37}$$

with

$$E_0 = \mathbb{E}\left[\frac{1-G}{G}\psi t'\right], \qquad F_0 = \mathbb{E}\left[\frac{1-G}{G}tt'\right]. \tag{38}$$

The the population mean of $M_i = M_i(\beta_0)$ equals

$$M = \begin{pmatrix} \Gamma & -\mathbb{E}\left[\frac{G_1}{G}\psi t'\right] \\ 0 & -\mathbb{E}\left[\frac{G_1}{G}tt'\right] \end{pmatrix}. \tag{39}$$

Standard GMM results then give a limiting sampling variance for $\sqrt{N}\left(\widehat{\beta}-\beta_0\right)$ equal to

$$M^{-1}\Omega M^{-1\prime}=\left(\begin{array}{c}\Gamma^{-1}\left(\mathbb{E}\left[\frac{\psi\psi'}{G}\right]-E_0F_0^{-1}E_0'\right)\Gamma^{-1\prime}+\Gamma^{-1}\mathbb{E}\left[\frac{G_1}{G}tt'\right]^{-1}\Delta_0'F_0\Delta_0\mathbb{E}\left[\frac{G_1}{G}tt'\right]^{-1}\Gamma^{-1\prime}\\ -\mathbb{E}\left[\frac{G_1}{G}tt'\right]^{-1}F_0\left\{E_0F_0^{-1}-\mathbb{E}\left[\frac{G_1}{G}\psi t'\right]\mathbb{E}\left[\frac{G_1}{G}tt'\right]^{-1}\right\}\Gamma^{-1\prime}\\ -\Gamma^{-1}\left\{E_0F_0^{-1}-\mathbb{E}\left[\frac{G_1}{G}\psi t'\right]\mathbb{E}\left[\frac{G_1}{G}tt'\right]^{-1}\right\}F_0\mathbb{E}\left[\frac{G_1}{G}tt'\right]^{-1}\\ V_{MM}\left(\delta\right)\end{array}\right), \tag{40}$$

where

$$\Delta_0=\mathbb{E}\left[\frac{D}{G}\left(\psi-E_0F_0^{-1}t\right)S_\delta'\right],\qquad V_{MM}\left(\delta_0\right)=\mathbb{E}\left[\frac{G_1}{G}tt'\right]^{-1}F_0\mathbb{E}\left[\frac{G_1}{G}tt'\right]^{-1}. \tag{41}$$

Now consider the case where Assumptions 1.1 to 1.4 and 2.1 hold, but not 1.5. Let $\beta_*=\left(\gamma_0',\delta_*'\right)'$, with δ_* the pseudo-true propensity score parameter. Let $G_*=G(t\left(X\right)'\delta_*)$ etc. Under this set of assumptions we have

$$\Omega_*=\left(\begin{array}{cc}\mathbb{E}\left[\frac{p_0(X)}{G_*^2}\psi\psi'\right] & \mathbb{E}\left[\frac{p_0(X)}{G_*}\left(\frac{1-G_*}{G_*}\right)\psi t'\right]\\ \mathbb{E}\left[\frac{p_0(X)}{G_*}\left(\frac{1-G_*}{G_*}\right)t\psi'\right] & \mathbb{E}\left[\left(\frac{p_0(X)}{G_*^2}-2\frac{p_0(X)}{G_*}+1\right)tt'\right]\end{array}\right),$$

and

$$M_*=\left(\begin{array}{cc}\mathbb{E}\left[\frac{p_0(X)}{G_*}\frac{\partial\psi}{\partial\gamma'}\right] & -\mathbb{E}\left[\frac{p_0(X)}{G_*}\frac{G_{1*}}{G_*}\psi t'\right]\\ 0 & -\mathbb{E}\left[\frac{p_0(X)}{G_*}\frac{G_{1*}}{G_*}tt'\right]\end{array}\right),$$

so that $\Psi_0=\left\{\left(M_*'\Omega_*^{-1}M_*\right)^{-1}\right\}_{1:K,1:K}$.

Local efficiency If Assumption 2.1 also holds we have $\mathbb{E}\left[\psi|\,X\right]=\Pi_0t=q$ so that $E_0F_0^{-1}=\Pi_0$ and hence

$$E_0F_0^{-1}E_0'=\mathbb{E}\left[\frac{1-G}{G}\Pi_0tt'\Pi_0'\right]=\mathbb{E}\left[\frac{1-G}{G}qq'\right], \tag{42}$$

which gives the equality $\Gamma^{-1}\left(\mathbb{E}\left[\frac{\psi\psi'}{G}\right]-E_0F_0^{-1}E_0'\right)\Gamma^{-1\prime}=\mathcal{I}\left(\gamma_0\right)^{-1}$. In that case we also have $\Delta_0=0$ since $\mathbb{E}\left[\psi|\,D,X\right]=E_0F_0^{-1}t$. Under these conditions (40) simplifies to

$$M^{-1}\Omega M^{-1\prime}=diag\left(\mathcal{I}\left(\gamma_0\right)^{-1},V_{MM}\left(\delta_0\right)\right). \tag{43}$$

Local efficiency at Assumption 2.1 follows if we can show that IPT is regular under Assumptions 1.1 to 1.5. The score function for a parametric submodel of the semiparametric missing data model is (e.g., Chen, Hong and Tarrozzi, 2004, 2008)

$$s_\eta\left(y,x,d;\eta\right)=ds_\eta\left(\,y_1|\,x;\eta\right)+\frac{d-G(t\left(x\right)'\delta)}{G(t\left(x\right)'\delta)\left[1-G(t\left(x\right)'\delta)\right]}G_1(t\left(x\right)'\delta)t\left(x\right)\times\left(\frac{\partial\delta}{\partial\eta}\right)+r_\eta\left(x;\eta\right).$$

Under Assumption 1.1 we have, differentiating under the integral and using iterated expectations,

$$\begin{aligned}\frac{\partial\gamma\left(\eta_0\right)}{\partial\eta}&=-\Gamma^{-1}\mathbb{E}\left[\psi\left(Z,\gamma_0\right)\frac{\partial\log f\left(Y_1,X;\eta_0\right)}{\partial\eta}\right]\\&=-\Gamma^{-1}\left\{\mathbb{E}\left[\psi\left(Z,\gamma_0\right)s_\eta\left(\,Y_1|\,X;\eta_0\right)\right]+\mathbb{E}\left[q\left(X;\gamma\right)r_\eta\left(X;\eta_0\right)\right]\right\}.\end{aligned}$$

Under Assumptions 1.1 to 1.5 standard calculations yield an asymptotically linear representation of $\widehat{\gamma}$ equal to:

$$\widehat{\gamma}=\gamma_0-\frac{1}{N}\sum_{i=1}^N\Gamma^{-1}\left\{\frac{D_i\psi\left(Z_i,\gamma_0\right)}{G(t\left(X_i\right)'\delta_0)}-M_{12}M_{22}^{-1}\left(\frac{D_i}{G(t\left(X_i\right)'\delta_0)}-1\right)t\left(X_i\right)\right\}+o_p\left(N^{-1/2}\right),$$

where $-\Gamma^{-1}$ times the term in $\{\cdot\}$ is the influence function and M_{12} and M_{22} denote the upper righthand $K \times 1+M$ and lower righthand $1+M \times 1+M$ blocks of M as given in (39) above. Let ϕ denote this influence function, by Theorem 2.2 of Newey (1990), regularity of $\widehat{\gamma}$ follows if

$$\frac{\partial \gamma(\eta_0)}{\partial \eta} = \mathbb{E}[\phi s_\eta(Y, X|\eta_0)] = -\Gamma^{-1}\left\{\mathbb{E}[\psi(Z,\gamma_0) s_\eta(Y_1|X;\eta_0)] + \mathbb{E}[q(X;\gamma_0) r_\eta(X;\eta_0)]\right\}.$$

We have, using the conditional mean zero property of scores, the MAR assumption, and the fact that $p_0(X) = G(t(X)'\delta_0)$

$$\begin{aligned}
\mathbb{E}[\phi s_\eta(Y,X|\eta_0)] &= -\Gamma^{-1}\mathbb{E}\left[\left\{\frac{D_i\psi(Z,\gamma_0)}{G(t(X)'\delta_0)} - M_{12}M_{22}^{-1}\left(\frac{D_i}{G(t(X)'\delta_0)} - 1\right)t(X_i)\right\}\{s_\eta(Y_1|X;\eta_0) + r_\eta(X;\eta_0)\}\right] \\
&= -\Gamma^{-1}\mathbb{E}\left[\frac{D_i\psi(Z,\gamma_0)}{G(t(X)'\delta_0)}\{s_\eta(Y_1|X;\eta_0) + r_\eta(X;\eta_0)\}\right] \\
&= -\Gamma^{-1}\mathbb{E}[\psi(Z,\gamma_0)\{s_\eta(Y_1|X;\eta_0) + r_\eta(X;\eta_0)\}] \\
&= -\Gamma^{-1}\left\{\mathbb{E}[\psi(Z,\gamma_0) s_\eta(Y_1|X;\eta_0)] + \mathbb{E}[q(X;\gamma_0) r_\eta(X;\eta_0)]\right\},
\end{aligned}$$

as required.

Consistent variance-covariance matrix estimation If Assumptions 1.1 to 1.4 and either 2.1 or 1.5 or both hold, then the asymptotic variance of $\widehat{\gamma}$ may be consistently estimated, under regularity conditions, by

$$\widehat{\Psi} = \left\{\left(\widehat{M}'\widehat{\Omega}^{-1}\widehat{M}\right)^{-1}\right\}_{1:K,1:K}, \tag{44}$$

with $\widehat{M} = \sum_{i=1}^{N} M_i\left(\widehat{\beta}\right)/N$ and $\widehat{\Omega} = \sum_{i=1}^{N} m_i\left(\widehat{\beta}\right) m_i\left(\widehat{\beta}\right)'/N$.

A.3 The asymptotic variance of three-step AIPW estimators

In this appendix we summarize the first order asymptotic properties of a class of three-step AIPW estimators under Assumptions 1.1 to 1.5. This class includes the estimator proposed by Robins, Rotnitzky and Zhao (1994). As well as the variations proposed by Newey (1994a), Hirano and Imbens (2001), and Cao, Tsiatis and Davidian (2009). While the first order properties of AIPW are well-known, we include the results below as they will prove useful for the higher order bias calculations.

We begin by developing some notation. Let $\beta = (\gamma', \delta')'$ and define the $K + 2(1+M) \times 1$ moment vector and derivative matrix

$$\underset{K+2(1+M)\times 1}{m_i(\beta)} = \begin{pmatrix} \frac{D_i}{G_i(\delta)}\psi_i(\gamma) \\ \left(\frac{D_i}{G_i(\delta)} - 1\right)t_i \\ \frac{D_i - G_i(\delta)}{G_i(\delta)(1-G_i(\delta))}G_{1i}(\delta)\, t_i \end{pmatrix}, \quad M_i(\beta) = \begin{pmatrix} \frac{D_i}{G_i(\delta)}\frac{\partial\psi_i(\beta)}{\partial\gamma'} & -\frac{D_i}{G_i(\delta)}\frac{G_{1i}(\delta)}{G_i(\delta)}\psi_i(\gamma)\, t_i' \\ 0 & -\frac{D_i}{G_i(\delta)}\frac{G_{1i}(\delta)}{G_i(\delta)}t_i t_i' \\ 0 & J_i(\delta) \end{pmatrix}, \tag{45}$$

with $J_i(\delta)$ as defined above. Further define the weight matrix

$$V_i(\beta) = \begin{bmatrix} \frac{D_i}{G_i(\delta)^2}\psi_i(\gamma)\psi_i(\gamma)' & \frac{D_i}{G_i(\delta)}\omega_i(\delta)\psi_i(\gamma)\, t_i' & 0 \\ \frac{D_i}{G_i(\delta)}\omega_i(\delta)\, t_i\psi_i(\gamma)' & \nu_i(\delta)\omega_i(\delta)\, t_i t_i' & 0 \\ 0 & \frac{D_i}{G_i(\delta)}\frac{G_{1i}(\delta)}{G_i(\delta)}t_i t_i' & -J_i(\delta) \end{bmatrix}, \tag{46}$$

where $\omega_i(\delta) = \omega(X_i, \delta)$ and $\nu_i(\delta) = \nu(D_i, X_i, \delta)$ are known, scalar-valued, weight functions (the latter with the property that $\mathbb{E}[\nu_i(\delta_0)|X] = 1$).

Finally let

$$\overline{M}(\beta) = \frac{1}{N}\sum_{i=1}^{N} M_i(\beta), \quad \overline{V}(\beta) = \frac{1}{N}\sum_{i=1}^{N} V_i(\beta), \quad \overline{m}(\beta) = \frac{1}{N}\sum_{i=1}^{N} m_i(\beta). \tag{47}$$

Our asymptotic analysis of three-step AIPW estimators exploits their representation as particular iterated GMM

estimators.

Lemma A.1 (Iterated GMM Representation of AIPW) *The AIPW estimate $\widehat{\gamma}$ which solves (20) in the main text is numerically identical to the iterated GMM estimate $\widehat{\beta} = \left(\widehat{\gamma}', \widehat{\delta}'\right)'$ which solves*

$$\overline{M}\left(\widehat{\beta}\right)\overline{V}\left(\widehat{\beta}\right)^{-1}\overline{m}\left(\widehat{\beta}\right) = 0. \tag{48}$$

Proof. See the Supplemental Web Appendix. ■

Invoking Lemma A.1 we proceed to characterize the large sample properties of (48). The population mean of the derivative of $m_i(\beta_0)$, as defined in (45) above, equals

$$M = \begin{bmatrix} \Gamma & -\mathbb{E}\left[\frac{G_1}{G}\psi t'\right] \\ 0 & -\mathbb{E}\left[\frac{G_1}{G}tt'\right] \\ 0 & -\mathcal{I}(\delta_0) \end{bmatrix}. \tag{49}$$

The probability limit of the weight matrix (46) is given by

$$V = \begin{bmatrix} \mathbb{E}\left[\frac{\psi\psi'}{G}\right] & E_\omega & 0 \\ E_\omega' & F_\omega & 0 \\ 0 & \mathbb{E}\left[\frac{G_1}{G}tt'\right] & \mathcal{I}(\delta_0) \end{bmatrix}, \tag{50}$$

with $E_\omega = \mathbb{E}[\omega\psi t']$ and $F_\omega = \mathbb{E}[\omega tt']$.

The covariance of the moment vector $m_i = m_i(\beta_0)$ is

$$\Omega = \begin{pmatrix} \mathbb{E}\left[\frac{\psi\psi'}{G}\right] & E_0 & \mathbb{E}\left[\frac{G_1}{G}\psi t'\right] \\ E_0' & F_0 & \mathbb{E}\left[\frac{G_1}{G}tt'\right] \\ \mathbb{E}\left[\frac{G_1}{G}t\psi'\right] & \mathbb{E}\left[\frac{G_1}{G}tt'\right] & \mathcal{I}(\delta_0) \end{pmatrix}, \tag{51}$$

with E_0 and F_0 as defined above.

Standard results on GMM imply that $\sqrt{N}\left(\widehat{\beta} - \beta_0\right)$ has a limiting sampling variance of

$$\begin{aligned} &\left(M'V^{-1}M\right)^{-1}M'V^{-1}\Omega V^{-1}M\left(M'V^{-1}M\right)^{-1\prime} \\ &= \begin{pmatrix} \begin{pmatrix} \Gamma^{-1}\left(\mathbb{E}\left[\frac{\psi\psi'}{G}\right] - E_0F_0^{-1}E_0' - \Delta_\omega\mathcal{I}(\delta_0)^{-1}\Delta_\omega'\right)\Gamma^{-1\prime} \\ +\Gamma^{-1}\left(E_0F_0^{-1} - E_\omega F_\omega^{-1}\right)F_0\left(E_0F_0^{-1} - E_\omega F_\omega^{-1}\right)'\Gamma^{-1\prime} \end{pmatrix} & 0 \\ 0 & \mathcal{I}(\delta_0)^{-1} \end{pmatrix}, \end{aligned} \tag{52}$$

with

$$\Delta_\omega = \mathbb{E}\left[\frac{D}{G}\left\{\psi - E_\omega F_\omega^{-1}t\right\}S_\delta'\right].$$

If Assumption 2.1 also holds (52) simplifies to $diag\left\{\mathcal{I}(\gamma_0)^{-1}, \mathcal{I}(\delta_0)^{-1}\right\}$.

B Derivation of the higher order bias of IPT and AIPW (Theorem 3.1)

In this appendix we outline the derivation of the $O\left(N^{-1}\right)$ bias expressions for the class of AIPW estimates of γ_0 discussed in the main text as well as for our IPT estimate (i.e., equations (29) and (30) in the main text). These bias expressions follow from stochastic expansions (i.e., second order, as opposed to the usual first order, Taylor series approximations of the estimating equations). Such expansions have a long history in statistics (e.g., Cox and Snell, 1968, Section 2; Rilstone, Srivastava and Ullah, 1996). Newey and Smith (2004, Lemma A.4, pp. 241 - 242) provide a general formula for the $O\left(N^{-1}\right)$ bias of M-estimators which they then specialize to analyze the bias properties of two-step GMM and GEL estimators (see also Newey (2002)). As each of the estimators we consider have M-estimator

representations we use their general result in our calculations. In what follows parameter's are evaluated at their population values unless stated otherwise; for this reason we drop '0' subscripts when doing so causes no confusion. In contrast to Appendix A we maintain Assumption 2.1 in what follows (in addition to Assumptions 1.1 to 1.5).

Let $\widehat{\theta}$ be the solution to

$$\underset{T\times 1}{\overline{b}(\widehat{\theta})} = \sum\nolimits_{i=1}^{N} b_i(\widehat{\theta}) = 0, \tag{53}$$

then under regularity conditions stated in Newey and Smith (2004, Lemma A.4) the asymptotic (higher order) bias of $\widehat{\theta}$ is given by

$$\text{Bias}(\widehat{\theta}) = \frac{-B^{-1}}{N}\left(\mathbb{E}\left[A_i\phi_i\right] + \frac{1}{2}\mathbb{E}\left[\sum\nolimits_{q=1}^{T}\phi_{q,i}B_q\phi_i\right]\right), \tag{54}$$

where e_q a $T\times 1$ column vector with a one in the q^{th} row and zeros elsewhere and

$$\underset{T\times T}{B} = \mathbb{E}\left[\frac{\partial b_i(\theta)}{\partial\theta'}\right], \quad \underset{T\times 1}{\phi_i} = -B^{-1}b_i(\theta), \quad \underset{T\times T}{A_i} = \frac{\partial b_i(\theta)}{\partial\theta'} - B, \quad \underset{T\times T}{B_q} = \mathbb{E}\left[\frac{\partial^2 b_i(\theta)}{\partial\theta_q\partial\theta'}\right]. \tag{55}$$

We assume that these objects are well-defined.

B.1 Higher order bias of $\widehat{\gamma}_{IPT}$

The IPT estimator of $\theta = (\gamma', \delta')'$ is given by the solution to (53) with

$$b_i(\theta) = \begin{pmatrix} \frac{D_i}{G_i(\delta)}\psi_i(\gamma) \\ \left(\frac{D_i}{G_i(\delta)} - 1\right)t_i \end{pmatrix}.$$

Objects, $\frac{\partial b_i(\theta_0)}{\partial\theta'}$, B and A_i of (55) above specialize to

$$\frac{\partial b_i(\theta_0)}{\partial\theta'} = \begin{bmatrix} \frac{D_i}{G_i}\frac{\partial\psi_i}{\partial\gamma'} & -\frac{D_i}{G_i}\frac{G_{1i}}{G_i}\psi_i t_i' \\ 0 & -\frac{D_i}{G_i}\frac{G_{1i}}{G_i}t_i t_i' \end{bmatrix}, \quad B = \begin{bmatrix} \Gamma & -\mathbb{E}\left[\frac{G_1}{G}\psi t'\right] \\ 0 & -\mathbb{E}\left[\frac{G_1}{G}tt'\right] \end{bmatrix}, \quad A_i = \begin{bmatrix} \frac{D_i}{G_i}\frac{\partial\psi_i}{\partial\gamma'} - \Gamma & -\frac{D_i}{G_i}\frac{G_{1i}}{G_i}\psi_i t_i' + \mathbb{E}\left[\frac{G_1}{G}\psi t'\right] \\ 0 & -\frac{D_i}{G_i}\frac{G_{1i}}{G_i}t_i t_i' + \mathbb{E}\left[\frac{G_1}{G}tt'\right] \end{bmatrix}.$$

Using the partitioned inverse formula we have

$$B^{-1} = \begin{bmatrix} \Gamma^{-1} & -\Gamma^{-1}\mathbb{E}\left[\frac{G_1}{G}\psi t'\right]\mathbb{E}\left[\frac{G_1}{G}tt'\right]^{-1} \\ 0 & -\mathbb{E}\left[\frac{G_1}{G}tt'\right]^{-1} \end{bmatrix}. \tag{56}$$

Combining the above expressions then gives

$$\mathbb{E}\left[A_i\phi_i\right] = -\begin{bmatrix} \mathbb{E}\left[\frac{\partial\psi}{\partial\gamma'}\Gamma^{-1}\frac{1}{G}\psi\right] - \mathbb{E}\left[\frac{1-G}{G}\frac{\partial\psi}{\partial\gamma'}\Gamma^{-1}\mathbb{E}\left[\frac{G_1}{G}\psi t'\right]\mathbb{E}\left[\frac{G_1}{G}tt'\right]^{-1}t\right] + \mathbb{E}\left[\frac{1-G}{G}\frac{G_1}{G}\psi t'\mathbb{E}\left[\frac{G_1}{G}tt'\right]^{-1}t\right] \\ \mathbb{E}\left[\frac{1-G}{G}\frac{G_1}{G}tt'\mathbb{E}\left[\frac{G_1}{G}tt'\right]^{-1}t\right] \end{bmatrix}. \tag{57}$$

Let $\Pi_* \overset{def}{=} \mathbb{E}\left[\frac{G_1}{G}\psi t'\right]\mathbb{E}\left[\frac{G_1}{G}tt'\right]^{-1}$; using (57) we have the first K rows of $-B^{-1}\mathbb{E}\left[A_i\phi_i\right]$ equal to

$$\begin{aligned}
&\Gamma^{-1}\mathbb{E}\left[\frac{\partial\psi}{\partial\gamma'}\Gamma^{-1}\frac{1}{G}\psi\right] - \Gamma^{-1}\mathbb{E}\left[\frac{1-G}{G}\frac{\partial\psi}{\partial\gamma'}\Gamma^{-1}\mathbb{E}\left[\frac{G_1}{G}\psi t'\right]\mathbb{E}\left[\frac{G_1}{G}tt'\right]^{-1}t\right] \\
&+\Gamma^{-1}\mathbb{E}\left[\frac{1-G}{G}\frac{G_1}{G}\psi t'\mathbb{E}\left[\frac{G_1}{G}tt'\right]^{-1}t\right] \\
&-\Gamma^{-1}\mathbb{E}\left[\frac{G_1}{G}\psi t'\right]\mathbb{E}\left[\frac{G_1}{G}tt'\right]^{-1}\mathbb{E}\left[\frac{1-G}{G}\frac{G_1}{G}tt'\mathbb{E}\left[\frac{G_1}{G}tt'\right]^{-1}t\right] \\
&=\Gamma^{-1}\mathbb{E}\left[\frac{\partial\psi}{\partial\gamma'}\Gamma^{-1}\frac{1}{G}\left\{\psi-\Pi_* t\right\}\right]+\Gamma^{-1}\mathbb{E}\left[\frac{\partial\psi}{\partial\gamma'}\Gamma^{-1}\Pi_* t\right] \\
&+\Gamma^{-1}\mathbb{E}\left[\frac{1-G}{G}\frac{G_1}{G}\left\{\psi-\Pi_* t\right\}t'\mathbb{E}\left[\frac{G_1}{G}tt'\right]^{-1}t\right]
\end{aligned} \tag{58}$$

Assumption 2.1 gives $q=\Pi_0 t$ so that $\Pi_*=\Pi_0$; therefore, applying the law of iterated expectations, gives the last term in the expression above identically equal to zero.

Now consider the second component of the bias expression (54). Evaluating $\mathbb{E}\left[\phi_i\phi_i'\right]$ yields

$$\mathbb{E}\left[\phi_i\phi_i'\right]=\begin{bmatrix}\mathcal{I}\left(\gamma_0\right)^{-1} & 0\\ 0 & V_{MM}\left(\delta\right)\end{bmatrix}. \tag{59}$$

For $q=1,...,K$, using the expression for $\partial b_i\left(\theta_0\right)/\partial\theta'$, we have

$$B_q=\mathbb{E}\begin{bmatrix}\frac{\partial^2\psi}{\partial\gamma_q\partial\gamma'} & -\frac{G_1}{G}\frac{\partial\psi}{\partial\gamma_q}t'\\ 0 & 0\end{bmatrix}, \tag{60}$$

for B_q as defined in (55) above. For $q=K+1,...,K+1+M\ (=T)$ we have instead

$$B_q=\mathbb{E}\begin{bmatrix}-\frac{G_1}{G}t_{q-K}\frac{\partial\psi}{\partial\gamma'} & \left(\frac{2G_1^2}{G^2}-\frac{G_2}{G}\right)t_{q-K}\psi t'\\ 0 & \left(\frac{2G_1^2}{G^2}-\frac{G_2}{G}\right)t_{q-K}tt'\end{bmatrix}. \tag{61}$$

Using (59), (60) and (61) the first K rows of $\frac{-B^{-1}}{2N}\mathbb{E}\left[\sum_{q=1}^{T}\phi_{q,i}B_q\phi_i\right]$ can be shown to equal

$$\left\{\frac{-B^{-1}}{2N}\mathbb{E}\left[\sum\nolimits_{q=1}^{T}\phi_{q,i}B_q\phi_i\right]\right\}_{1:K,:}=\left\{-\frac{B^{-1}}{2N}\sum_{q=1}^{T}B_q\mathbb{E}\left[\phi_i\phi_i'\right]e_q\right\}_{1:K,:}=-\frac{1}{2N}\sum_{k=1}^{K}\Gamma^{-1}\mathbb{E}\left[\frac{\partial^2\psi}{\partial\gamma_k\partial\gamma'}\right]\mathcal{I}\left(\gamma_0\right)^{-1}e_k. \tag{62}$$

Combining (58) and (62) yields C_O as given in the statement of the Theorem.

B.2 Higher order bias of $\widehat{\gamma}_{AIPW}$

Let $\theta=(\gamma',\delta',\lambda')$ be the $T=2K+3\left(1+M\right)$ vector of parameters of interest. Let $\widehat{\theta}$ be the solution to (53) with

$$b_i\left(\theta\right)=-\begin{bmatrix}M_i\left(\beta\right)'\lambda\\ m_i\left(\beta\right)+V_i\left(\beta\right)'\lambda\end{bmatrix}, \tag{63}$$

and $M_i\left(\beta\right)$ and $V_i\left(\beta\right)$ as defined by (45) and (46) above. The AIPW estimate corresponds to the first $K\times 1$ components of $\widehat{\theta}$ since $\overline{V}\left(\widehat{\beta}\right)^{-1}\overline{m}\left(\widehat{\beta}\right)=-\widehat{\lambda}$ so that $\overline{M}\left(\widehat{\beta}\right)'\widehat{\lambda}=\overline{M}\left(\widehat{\beta}\right)'\overline{V}\left(\widehat{\beta}\right)^{-1}\overline{m}\left(\widehat{\beta}\right)=0$ which, by Lemma A.1, is the first order condition for the AIPW estimator.

Similar to the treatment of two-step and iterated GMM by Newey and Smith (2004, cf., Lemmas A.5 and A.6), we derive the higher order bias properties of the AIPW estimate of γ_0 by considering those of a simplified $O_p\left(N^{-3/2}\right)$ equivalent estimate.

We have, by Lemma A.4 of Newey and Smith (2004), $\widehat{\theta}-\theta_0=\widetilde{\phi}/\sqrt{N}+O_p\left(1/N\right)$ with $\widetilde{\phi}=\sum_{i=1}^{N}\phi_i/\sqrt{N}$ for ϕ_i as defined in (55) above. This means that for $q=1,\ldots,K+1+M$ we have $\widehat{\beta}-\beta_{q0}=e_q'\widetilde{\phi}/\sqrt{N}+O_p\left(1/N\right)$. Now consider the mean value expansion

$$\begin{aligned}\overline{V}\left(\widehat{\beta}\right)&=\overline{V}\left(\beta_0\right)+\sum\nolimits_{q=1}^{K+1+M}\frac{\partial\overline{V}\left(\overline{\beta}\right)}{\partial\beta_q}\left(\widehat{\beta}-\beta_{q0}\right)\\&=\overline{V}\left(\beta_0\right)+\sum\nolimits_{q=1}^{K+1+M}\mathbb{E}\left[\frac{\partial V_i\left(\beta_0\right)}{\partial\beta_q}\right]\left(\widehat{\beta}-\beta_{q0}\right)+O_p\left(1/N\right)\\&=\overline{V}\left(\beta_0\right)+\sum\nolimits_{q=1}^{K+1+M}\mathbb{E}\left[\frac{\partial V_i\left(\beta_0\right)}{\partial\beta_q}\right]e_q'\widetilde{\phi}/\sqrt{N}+O_p\left(1/N\right),\end{aligned}$$

with $\overline{V}\left(\beta\right)$ and $V_i\left(\beta\right)$ as defined in (47) and (46) above; $\overline{\beta}$ is mean value between β_0 and $\widehat{\beta}$. Let $\xi_i=V_i-V+\sum_{q=1}^{K+1+M}\mathbb{E}\left[\frac{\partial V_i(\beta_0)}{\partial\beta_q}\right]e_q'\phi_i$ with V as given in (50) above. This gives $\overline{V}\left(\widehat{\beta}\right)=V+\frac{1}{N}\sum_{i=1}^{N}\xi_i+O_p\left(1/N\right)$. Now consider the solution to $\overline{b}^*\left(\widehat{\theta}\right)=\frac{1}{N}\sum_{i=1}^{N}b_i^*\left(\widehat{\theta}\right)=0$ with

$$b_i^*\left(\theta\right)=-\left[\begin{array}{c}M_i\left(\beta\right)'\lambda\\ m_i\left(\beta\right)+\left[V+\xi_i\right]'\lambda\end{array}\right]. \tag{64}$$

Using the definitions above we have

$$\begin{aligned}0&=\overline{b}^*\left(\widehat{\theta}\right) \qquad (65)\\&=\overline{b}\left(\widehat{\theta}\right)+\left(\begin{array}{c}0\\ \left[\overline{V}\left(\widehat{\beta}\right)-V-\frac{1}{N}\sum_{i=1}^{N}\xi_i\right]'\widehat{\lambda}\end{array}\right)\\&=\overline{b}\left(\widehat{\theta}\right)+O_p\left(N^{-3/2}\right),\end{aligned}$$

since $\overline{V}\left(\widehat{\beta}\right)-V-\frac{1}{N}\sum_{i=1}^{N}\xi_i=O_p\left(1/N\right)$ and $\widehat{\lambda}=\lambda_0+O_p\left(1/\sqrt{N}\right)$. Appealing to the equivalence implicit in the last line of (65) we henceforth analyze the bias properties of the solution to $\overline{b}^*\left(\widehat{\theta}\right)=0$. In what follows we redefine $b_i\left(\theta\right)$ to be equal to $b_i^*\left(\theta\right)$ as given in (64).

The terms defined in (55) are given by, recalling that $\lambda_0=0$,

$$B=-\left(\begin{array}{cc}0&M'\\ M&V\end{array}\right),\quad A_i=-\left(\begin{array}{cc}0&\left(M_i-M\right)'\\ \left(M_i-M\right)&\xi_i\end{array}\right),$$

with M and V given by (49) and (50) above.

The partitioned inverse formula gives

$$B^{-1}=-\left(\begin{array}{cc}-\Upsilon&H\\ H'&L\end{array}\right), \tag{66}$$

with

$$\Upsilon=\left(M'V^{-1}M\right)^{-1},\quad H=\Upsilon M'V^{-1},\quad L=V^{-1}-V^{-1}MH. \tag{67}$$

The supplemental web appendix shows that

$$\Upsilon=\left(\begin{array}{cc}\Gamma^{-1}\Lambda\Gamma^{-1\prime}&0\\ \Pi_S'\Gamma^{-1\prime}&\mathcal{I}\left(\delta_0\right)^{-1}\end{array}\right) \tag{68}$$

$$H=\left(\begin{array}{ccc}\Gamma^{-1}&-\Gamma^{-1}\Pi_0&0\\ 0&0&-\mathcal{I}\left(\delta_0\right)^{-1}\end{array}\right) \tag{69}$$

$$L=\left(\begin{array}{ccc}0&0&-\Lambda^{-1}\Pi_S\\ 0&F_\omega^{-1}&\Pi_0'\Lambda^{-1}\Pi_S\\ \Pi_S'\Lambda^{-1}&-\mathcal{I}\left(\delta_0\right)^{-1}\mathbb{E}\left[\frac{G_1}{G}tt'\right]\left(F_\omega-E_\omega'\mathbb{E}\left[\frac{\psi\psi'}{G}\right]^{-1}E_\omega\right)^{-1}&0\end{array}\right). \tag{70}$$

From (55) and the expressions above we also have $\phi_i = -\begin{pmatrix} H' & L' \end{pmatrix}' m_i$, where we evaluate at the population value of θ.

The first part of the AIPW bias formula is given by the first K rows of $\frac{-B^{-1}}{N}\mathbb{E}\left[A_i\phi_i\right]$. Manipulating, using the expressions given above, we have

$$-B^{-1}\mathbb{E}\left[A_i\phi_i\right] = -\mathbb{E}\left[\begin{pmatrix} -HM_iH + \Upsilon M_i'L - H\xi_iL \\ -LM_iH - H'M_i'L - L\xi_iL \end{pmatrix} m_i\right]. \tag{71}$$

We require expressions for the first K rows of the matrix $\mathbb{E}\left[\left(HM_iH - \Upsilon M_i'L + H\xi_iL\right)m_i\right]$. Let $\{A\}_{1:K,:}$ denote rows 1 to K of a matrix. Very tedious calculations detailed in the supplemental web appendix give

$$\{H\mathbb{E}\left[M_iHm_i\right]\}_{1:K,:} = \Gamma^{-1}\mathbb{E}\left[\frac{1}{G}\frac{\partial\psi\left(\beta\right)}{\partial\gamma'}\Gamma^{-1}\left(\psi - \Pi_0 t\right)\right] + \Gamma^{-1}\mathbb{E}\left[\frac{\partial\psi\left(\beta\right)}{\partial\gamma'}\Gamma^{-1}\Pi_0 t\right] \tag{72}$$

$$-\left\{\Upsilon\mathbb{E}\left[M_i'Lm_i\right]\right\}_{1:K,:} = \mathcal{I}\left(\gamma_0\right)^{-1}\mathbb{E}\left[\frac{D}{G}\left(\frac{\partial\psi}{\partial\gamma'}\right)'\Lambda^{-1}\Pi_S S_\delta\right] \tag{73}$$

$$\begin{aligned}\{H\mathbb{E}\left[\xi_iLm_i\right]\}_{1:K,:} = {} & \Gamma^{-1}\Pi_0\mathbb{E}\left[\omega\left(\frac{D}{G} - \nu\right)\left(\frac{D}{G} - 1\right)tt'F_0^{-1}t\right] \\ & - \Gamma^{-1}\mathbb{E}\left[\frac{D}{G}\psi\left(\frac{D}{G}\psi - \omega\Pi_0 t_0\right)'\Lambda^{-1}\Pi_S S_\delta\right] \\ & + \Gamma^{-1}\mathbb{E}\left[\omega\left(\frac{D}{G} - \nu\right)\Pi_0 tt'\Pi_0\Lambda^{-1}\Pi_S S_\delta\right].\end{aligned} \tag{74}$$

Collecting these terms yields a bias contribution of

$$\begin{aligned}C_L = {} & \frac{\Gamma^{-1}}{N}\mathbb{E}\left[\frac{1}{G}\frac{\partial\psi\left(\beta\right)}{\partial\gamma'}\Gamma^{-1}\left(\psi - \Pi_0 t\right)\right] + \frac{\Gamma^{-1}}{N}\mathbb{E}\left[\frac{\partial\psi\left(\beta\right)}{\partial\gamma'}\Gamma^{-1}\Pi_0 t\right] \\ & + \frac{\mathcal{I}\left(\gamma_0\right)^{-1}}{N}\mathbb{E}\left[\frac{D}{G}\left(\frac{\partial\psi}{\partial\gamma'}\right)'\Lambda^{-1}\Pi_S S_\delta\right] \\ & + \frac{\Gamma^{-1}}{N}\Pi_0\mathbb{E}\left[\omega\left(\frac{D}{G} - \nu\right)\left(\frac{D}{G} - 1\right)tt'F_0^{-1}t\right] \\ & - \frac{\Gamma^{-1}}{N}\mathbb{E}\left[\frac{D}{G}\psi\left(\frac{D}{G}\psi - \omega\Pi_0 t_0\right)'\Lambda^{-1}\Pi_S S_\delta\right] \\ & + \frac{\Gamma^{-1}}{N}\mathbb{E}\left[\omega\left(\frac{D}{G} - \nu\right)\Pi_0 tt'\Pi_0\Lambda^{-1}\Pi_S S_\delta\right].\end{aligned} \tag{75}$$

To compute the second component of (54) for $\widehat{\gamma}_{AIPW}$ we require some additional notation and results. Recall that $B_q = \mathbb{E}\left[\partial^2 b_i\left(\theta\right)/\partial\theta_q\partial\theta'\right]$. For $q = 1, \ldots, K+1+M$ we have this term equal to

$$B_q = -\begin{pmatrix} 0 & \mathbb{E}\left[\frac{\partial M_i'}{\partial\beta_q}\right] \\ \mathbb{E}\left[\frac{\partial M_i}{\partial\beta_q}\right] & 0 \end{pmatrix}, \tag{76}$$

while for $q = K+1+M+1, \ldots, 2K+3\left(1+M\right)$ is it given by

$$B_q = -\begin{pmatrix} \mathbb{E}\left[\frac{\partial^2 m_{q-K-1-M}(Z_i,\beta)}{\partial\beta\partial\beta'}\right] & 0 \\ 0 & 0 \end{pmatrix}. \tag{77}$$

We also have

$$\mathbb{E}\left[\phi_i\phi_i'\right] = \begin{pmatrix} H\Omega H' & H\Omega L' \\ L\Omega H' & L + L\left(\Omega - V\right)L' \end{pmatrix}, \tag{78}$$

with Ω as given by (51) above. Note that $LVL' = L$ so that $L\Omega L' = LVL' + L\left(\Omega - V\right)L' = L + L\left(\Omega - V\right)L'$.

Using these expressions we evaluate the second component of (54) (i.e., $-\frac{B^{-1}}{2N}\mathbb{E}\left[\sum_{q=1}^{T}\phi_{q,i}B_q\phi_i\right]$) as follows:

$$-\frac{B^{-1}}{2N}\mathbb{E}\left[\sum_{q=1}^{T}\phi_{q,i}B_q\phi_i\right] = -\frac{1}{2N}\sum_{q=1}^{K+1+M}\begin{pmatrix} -\Upsilon\mathbb{E}\left[\frac{\partial M_i'}{\partial\beta_q}\right]L\Omega H' + H\mathbb{E}\left[\frac{\partial M_i'}{\partial\beta_q}\right]H\Omega H' \\ +H'\mathbb{E}\left[\frac{\partial M_i'}{\partial\beta_q}\right]L\Omega H' + L\mathbb{E}\left[\frac{\partial M_i'}{\partial\beta_q}\right]H\Omega H' \end{pmatrix} e_q \tag{79}$$
$$-\frac{1}{2N}\sum_{q=1}^{K+2(1+M)}\begin{pmatrix} -\Upsilon\mathbb{E}\left[\frac{\partial^2 m_q(Z_i,\beta)}{\partial\beta\partial\beta'}\right]H\Omega L' \\ +H'\mathbb{E}\left[\frac{\partial^2 m_q(Z_i,\beta)}{\partial\beta\partial\beta'}\right]H\Omega L' \end{pmatrix} e_q.$$

The first K rows of (79) contribute to the bias expression for $\widehat{\gamma}_{AIPW}$. To determine the form of the first K rows of (79) we only require expressions for the first K rows of the two matrices in parentheses to the right of the equality in (79).

After tedious calculation we have, for $q = 1, \ldots, K$,

$$-\left\{\Upsilon\mathbb{E}\left[\frac{\partial M_i'}{\partial\beta_q}\right]L\Omega H'\right\}_{1:K,:} = \begin{pmatrix} 0 & -\mathcal{I}(\gamma_0)^{-1}\mathbb{E}\left[\frac{\partial^2\psi}{\partial\gamma_q\partial\gamma}\right]'\Lambda^{-1}\Pi_S \end{pmatrix} \tag{80}$$

$$\left\{H\mathbb{E}\left[\frac{\partial M_i}{\partial\beta_q}\right]H\Omega H'\right\}_{1:K,:} = \begin{pmatrix} \Gamma^{-1}\mathbb{E}\left[\frac{\partial^2\psi}{\partial\gamma_q\partial\gamma}\right]\mathcal{I}(\gamma_0)^{-1} & -\Gamma^{-1}\mathbb{E}\left[\frac{D}{G}\frac{\partial\psi}{\partial\gamma_q}S_\delta'\right]\mathcal{I}(\delta_0)^{-1} \end{pmatrix} \tag{81}$$

Similarly, for $q = K+1, \ldots, K+1+M$, we have

$$-\left\{\Upsilon\mathbb{E}\left[\frac{\partial M_i'}{\partial\beta_q}\right]L\Omega H'\right\}_{1:K,:} = \begin{pmatrix} 0 & \mathcal{I}(\gamma_0)^{-1}\mathbb{E}\left[\frac{D}{G}\frac{\partial\psi}{\partial\gamma'}S_{\delta,q-K}'\right]'\Lambda^{-1}\Pi_S \end{pmatrix} \tag{82}$$

$$\left\{H\mathbb{E}\left[\frac{\partial M_i}{\partial\beta_q}\right]H\Omega H'\right\}_{1:K,:} = \begin{pmatrix} -\Gamma^{-1}\mathbb{E}\left[\frac{D}{G}\frac{\partial\psi}{\partial\gamma'}S_{\delta,q-K}'\right]\mathcal{I}(\gamma_0)^{-1} & 0 \end{pmatrix} \tag{83}$$

Using (79) and (80) to (83) we get, after some manipulation, a bias contribution of

$$C_{NL1} = -\frac{1}{2N}\sum_{q=1}^{K}\Gamma^{-1}\mathbb{E}\left[\frac{\partial^2\psi}{\partial\gamma_q\partial\gamma}\right]\mathcal{I}(\gamma_0)^{-1}e_q \tag{84}$$
$$-\frac{\mathcal{I}(\gamma_0)^{-1}}{2N}\mathbb{E}\left[\frac{D}{G}\left(\frac{\partial\psi}{\partial\gamma'}\right)'\Lambda^{-1}\Pi_S S_\delta\right].$$

Now consider the second part of (79). For $q = 1, \ldots, K$ we have

$$-\left\{\Upsilon\mathbb{E}\left[\frac{\partial^2 m_q(Z_i,\beta)}{\partial\beta\partial\beta'}\right]H\Omega L'\right\}_{1:K} = \begin{pmatrix} \mathcal{I}(\gamma_0)^{-1}\mathbb{E}\left[\frac{D}{G}\frac{\partial\psi_q}{\partial\gamma}S_\delta'\right]\Pi_S'\Lambda^{-1} & \text{(Not Needed)} & \text{(Not Needed)} \end{pmatrix}, \tag{85}$$

while for $q = K+1, \ldots, K+1+M$

$$-\left\{\Upsilon\mathbb{E}\left[\frac{\partial^2 m_q(Z_i,\beta)}{\partial\beta\partial\beta'}\right]H\Omega L'\right\}_{1:K} = \begin{pmatrix} 0 & 0 & 0 \end{pmatrix}, \tag{86}$$

and finally for $q = K+1+M+1, \ldots, K+2(1+M)$

$$-\left\{\Upsilon\mathbb{E}\left[\frac{\partial^2 m_q(Z_i,\beta)}{\partial\beta\partial\beta'}\right]H\Omega L'\right\}_{1:K} = \begin{pmatrix} 0 & 0 & 0 \end{pmatrix}. \tag{87}$$

Using (79) and (85) to (87) we get, after some manipulation, a bias contribution of

$$C_{NL2} = -\frac{\mathcal{I}(\gamma_0)^{-1}}{2N}\mathbb{E}\left[\frac{D}{G}\left(\frac{\partial\psi}{\partial\gamma'}\right)'\Lambda^{-1}\Pi_S S_\delta\right]. \tag{88}$$

Equation (29) is given by the sum of (75), (84), and (88).

C Computation

Computation of the IPT estimate of $\widehat{\gamma}$ consists of two steps. In the first step, which is nonstandard and detailed here, $\widehat{\delta}$ is computed as the solution to (14). In the second $\widehat{\gamma}$ is computed as the solution to (13). The second step is identical to that associated with standard inverse probability weighting (IPW) (e.g., Wooldridge, 2002, 2007). As second step is both application specific, and typically straightforward to compute using standard software (that accepts user-specified weights), we do not detail it here.

The IPT propensity score parameter δ estimate corresponds to the solution to (14). This solution could be computed in any number of ways (e.g., by a nonlinear GMM program such as STATA's *gmm* routine). Here we outline an approach which we have found to be computationally convenient and very reliable in practice. This involves defining $\widehat{\delta}$ to be the solution to a globally concave programming problem with unrestricted domain. Such problems are straightforward to solve using standard gradient-based numerical minimization procedures (e.g., MATLAB's *fminunc*() routine). A MATLAB routine, $IPT\_LOGIT()$, implementing the algorithm described here is available online at `https://files.nyu.edu/bsg1/public/`. The inverse probability weights computed using this procedure can be imported into standard software in order to implement the second step of IPT estimation. A nonparametric bootstrap procedure, which repeats each of the two stages of estimation for each bootstrap draw, can be used to construct asymptotically valid standard errors and confidence intervals (cf., Fortin, Lemieux, and Firpo, 2010). Alternatively, analytic standard errors may be computed as described in Appendix A.

Consider the following function

$$\varphi(v) = \frac{v}{G(v)} + \int_{1/G(v)}^{a} G^{-1}\left(\frac{1}{t}\right) \mathrm{d}t, \tag{89}$$

with $G(\cdot)$ as defined in Assumption 1.5 of the main text. When the propensity score takes the logit $G(v) = (1 + \exp(-v))^{-1}$ form (89) exists in closed form (see below). We implement the logit specification in the empirical application and expect that most users will do likewise. If a different propensity score model is assumed, then (89) is easily evaluated numerically if a closed form expression is unavailable.

The first and second derivatives of $\varphi(v)$ are

$$\varphi_1(v) = \frac{1}{G(v)}, \qquad \varphi_2(v) = -\frac{G_1(v)}{G(v)^2}, \tag{90}$$

so that (89) is strictly concave.

We compute $\widehat{\delta}$ by solving the following optimization problem

$$\max_{\delta} l_N(\delta), \quad l_N(\delta) = \frac{1}{N}\sum_{i=1}^{N} D_i \varphi\left(t(X_i)'\delta\right) - \frac{1}{N}\sum_{i=1}^{N} t(X_i)'\delta. \tag{91}$$

Differentiating $l_N(\delta)$ with respect to δ gives an $1 + M \times 1$ gradient vector of

$$\underset{1+M\times 1}{\nabla_\delta l_N(\delta)} = \frac{1}{N}\sum_{i=1}^{N} D_i \varphi_1\left(t(X_i)'\delta\right) t(X_i) - \frac{1}{N}\sum_{i=1}^{N} t(X_i) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i}{G\left(t(X_i)'\delta\right)} - 1\right) t(X_i), \tag{92}$$

which coincides with (14) as required. The $1 + M \times 1 + M$ Hessian matrix is

$$\underset{1+M\times 1+M}{\nabla_{\delta\delta} l_N(\delta)} = \frac{1}{N}\sum_{i=1}^{N} D_i \varphi_2\left(t(X_i)'\delta\right) t(X_i)\, t(X_i)'. \tag{93}$$

This is a negative definite function of δ; the problem (91) is consequently concave with a unique solution (if one exists; cf., Albert and Anderson, 1984).

In practice (92) will have an 'exploding denominator' when $t(X_i)'\delta$ is a large negative number. This can lead to numerical instabilities by causing the Hessian to be ill-conditioned. We address this problem by noting that at a valid solution $\sum_{i=1}^{N} D_i/(Gt(X_i)'\widehat{\delta})/N = 1$. Since Assumption 1.5 implies that $G(v)$ is bounded below by zero, this means that $D_i/(Gt(X_i)'\widehat{\delta}) < N$ for all $i = 1, \ldots, N$. Letting $v_i = t(X_i)'\delta$ this inequality corresponds to requiring that

$$G^{-1}(D_i/N) < v_i, \qquad i = 1, \ldots, N \tag{94}$$

at $\delta = \widehat{\delta}$. Let $v_N^* = G^{-1}(1/N)$; note that $v_N^* \to -\infty$ as $N \to \infty$ suggesting that (94) will be satisfied for most values of δ in large enough samples. In small samples (94) may be violated for some i at some iterations of the maximization procedure (although not at a valid solution). Our approach to estimation involves replacing $\varphi(v)$ with a quadratic function when $v \leq v_N^*$; this ensures that the denominator in (92) is bounded. This will improve the condition of the Hessian with respect to δ without changing the solution. Owen (2001, Chapter 12) proposes a similar procedure in the context of empirical likelihood estimation of moment condition models.

Specifically we replace $\varphi(v)$ in (91), (92) and (93) with

$$\varphi_N^\circ(v) = \begin{cases} \varphi(v) & v > v_N^* \\ a_N + b_N v_N^* + \frac{c_N}{2}\left(v_N^*\right)^2 & v \leq v_N^* \end{cases}, \tag{95}$$

where a_N, b_N and c_N are the solutions to

$$\begin{aligned} c_N &= \varphi_2(v_N^*) \\ b_N + c_N v_N^* &= \varphi_1(v_N^*) \\ a_N + b_N v_N^* + \frac{c_N}{2}(v_N^*)^2 &= \varphi_0(v_N^*). \end{aligned}$$

This choice of coefficients ensures that $\varphi_N^\circ(v)$ equals $\varphi(v)$, as well as equality of first and second derivatives, *at* $v = v_N^*$.

When $G(v)$ is logit our algorithm is particularly simple to implement. To derive the closed form expressions for $\varphi(v)$ in this case involves an application of integration by substitution:

$$\int_{t=a}^{t=b} f(t)\,\mathrm{d}t = \int_{u=h^{-1}(a)}^{u=h^{-1}(b)} f(h(u))\,h'(u)\,\mathrm{d}u, \tag{96}$$

where $h'(u) \neq 0$ for all $u \in [a,b]$.

Let $u = 1/(t-1)$ so that $t = \frac{1+u}{u} = h(u)$ with $h'(u) = \frac{1}{u} - \frac{1+u}{u^2} = -\frac{1}{u^2}$ applying (96) for $G(v) = \exp(v)/[1+\exp(v)]$ we get

$$\varphi(v) = \frac{v}{G(v)} + \int_{\frac{G(v)}{1-G(v)}}^{1/(a-1)} \ln(u)\left(-\frac{1}{u^2}\right)\mathrm{d}u \propto v - \exp(-v).$$

Differentiating with respect to v then gives $\varphi_1(v) = 1 + \exp(-v)$ and $\varphi_2(v) = -\exp(-v)$.

We also have $v_N^* = G^{-1}(1/N) = \ln\left(\frac{1/N}{1-1/N}\right) = \ln\left(\frac{1}{N-1}\right)$ so that solving for a_N, b_N and c_N yields

$$a_N = -(N-1)\left[1 + \ln\left(\frac{1}{N-1}\right) + \frac{1}{2}\left[\ln\left(\frac{1}{N-1}\right)\right]^2\right], \quad b_N = N + (N-1)\ln\left(\frac{1}{N-1}\right), \quad c_N = -(N-1).$$

References

[1] Abowd, John M., Bruno Crépon, and Francis Kramarz. (2001). "Moment estimation with attrition: an application to economic models," *Journal of the American Statistical Association* 96 (456): 1223 - 1231.

[2] Albert, A. and J. A. Anderson. (1984). "On the existence of maximum likelihood estimates in logistic regression models," *Biometrika* 71(1):1-10.

[3] Angrist, Joshua D. and Alan B. Krueger. (1992). "The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples," *Journal of the American Statistical Association* 87 (418): 328 - 336.

[4] Arellano, Manuel and Costas Meghir. (1992). "Female labour supply and on-the-job search: an empirical model estimated using complementary data sets," *Review of Economic Studies* 59 (3): 537 - 559.

[5] Back, Kerry and David P. Brown. (1993). "Implied probabilities in GMM estimators," *Econometrica* 61 (4): 971 - 975.

[6] Bang, Heejung and James M. Robins. (2005). "Doubly robust estimation in missing data and causal inference models," *Biometrics* 61 (4): 962 - 972.

[7] Brown, Bryan W. and Whitney K. Newey. (1998). "Efficient semiparametric estimation of expectations," *Econometrica* 66 (2): 453 - 464.

[8] Browning, Martin and Søren Leth-Petersen. (2003). "Imputing consumption from income and wealth Information," *Economic Journal* 113 (488): F282 - F301.

[9] Busso, Matias, John DiNardo and Justin McCrary. (2009). "Finite sample properties of semiparametric estimators of average treatment effects," *Mimeo.*

[10] Carneiro, Pedro and James J. Heckman (2004). "Human capital policy," *Inequality in America: What Role for Human Capital Policies?: 77 - 240.* Cambridge, MA: The MIT Press.

[11] Cao, Weihua, Anastasios A. Tsiatis and Marie Davidian. (2009). "Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data," *Biometrika* 96 (3): 723 - 734.

[12] Chamberlain, Gary. (1987). "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics* 34 (1): 305 - 334.

[13] Chen, Xiaohong. (2007). "Large sample sieve estimation of semi-nonparametric models," *Handbook of Econometrics* 6 (B): 5549 - 5632. (J.J. Heckman & E.E. Leamer, Eds.). Amsterdam: North-Holland.

[14] Chen, Xiaohong, Han Hong and Alessandro Tarozzi. (2004). "Semiparametric efficiency in GMM models of nonclassical measurement errors, missing data and treatment effects, *Mimeo.*

[15] Chen, Xiaohong, Han Hong and Alessandro Tarozzi. (2008). "Semiparametric efficiency in GMM models with auxiliary data," *Annals of Statistics* 36 (2): 808 - 843.

[16] Cheng, Philip E. (1994). "Nonparametric estimation of mean functionals with data missing at random," *Journal of the American Statistical Association* 89 (425): 81 - 87.

[17] Cox, D. R. and E. J. Snell. (1968). "A general definition of residuals," *Journal of the Royal Statistical Society B* 30(2): 248 - 275.

[18] Dinardo, John, Nicole M. Fortin, Thomas Lemieux. (1996). "Labor market institutions and the distribution of wages, 1973 - 1992: a semiparametric approach," *Econometrica* 64 (5): 1001 - 1044.

[19] Egel, Daniel, Bryan S. Graham and Cristine Campos de Xavier Pinto. (2008). "Inverse probability tilting and missing data problems," *NBER Working Paper No. 13981.*

[20] Firpo, Sergio. (2007). "Efficient semiparametric estimation of quantile treatment effects," *Econometrica* 75 (1): 259 - 276.

[21] Fortin, Nicole, and Thomas Lemieux, and Sergio Firpo. (2010). "Decomposition methods in economics," *NBER Working Paper No. 16045.*

[22] Fryer, Roland G. and Steven D. Levitt. (2006). "The Black-White test score gap through third grade," *American Law and Economics Review* 8 (2): 249 - 281.

[23] Fryer, Roland G. and Steven D. Levitt. (forthcoming). "Testing for racial differences in the mental ability of young children," *American Economic Review.*

[24] Gourieroux, Christian and Alain Monfort. (1981). "On the problem of missing data in linear models," *Review of Economic Studies* 48 (4): 579 - 586.

[25] Graham, Bryan S. (2009). "Efficiency bounds for missing data models with semiparametric restrictions," *Mimeo.*

[26] Hahn, Jinyong. (1998). "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica* 66 (2): 315 - 331.

[27] Hansen, Lars Peter, John Heaton and Amir Yaron. (1996). "Finite-sample properties of some alternative GMM estimators," *Journal of Business and Economic Statistics* 14 (3): 262 - 280.

[28] Heckman, James J. Richard Robb, Jr. (1985). "Alternative methods for evaluating the impact of interventions," *Longitudinal Analysis of Labor Market Data*: 156 - 245 (J.J. Heckman & B. Singer, Eds.). Cambridge: Cambridge University Press.

[29] Heckman, James J., Hidehiko Ichimura and. Petra E. Todd. (1997). "Matching as an econometric evaluation estimator: evidence from evaluating a job training programme," *Review of Economic Studies* 64 (4): 605 - 654.

[30] Heckman, James J., Hidehiko Ichimura and. Petra E. Todd. (1998). "Matching as an econometric evaluation estimator," *Review of Economic Studies* 65 (2): 261 - 294.

[31] Hirano, Keisuke and Guido W. Imbens. (2001). "Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization," *Health Services and Outcomes Research* 2 (3-4): 259 -278.

[32] Hirano, Keisuke, Guido W. Imbens and Geert Ridder. (2003). "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica* 71 (4): 1161 - 1189.

[33] Horvitz, D. G. and D. J. Thompson. (1952). "A generalization of sampling without replacement from a finite universe," *Journal of the American Statistical Association* 47 (260): 663 - 685.

[34] Ichimura, Hidehiko and Oliver Linton. (2005). "Asymptotic expansions for some semiparametric program evaluation estimators," *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*: 149 -170 (D.W.K Andrews & J.H. Stock, Eds). Cambridge: Cambridge University Press.

[35] Imbens, Guido. W. (1997). "One-step estimators of over-identified generalized method of moments models," *Review of Economic Studies* 64 (3): 359 - 383.

[36] Imbens, Guido. W. (2002). "Generalized method of moments and empirical likelihood," *Journal of Business and Economic Statistics* 20 (4): 493 - 506.

[37] Imbens, Guido W. (2004). "Nonparametric estimation of average treatment effects under exogeneity: a review," *Review of Economics and Statistics* 86 (1): 4 - 29.

[38] Imbens, Guido W. and Tony Lancaster. (1996). "Efficient estimation and stratified sampling," *Journal of Econometrics* 74 (2): 289 - 318.

[39] Imbens, Guido W., Whitney K. Newey and Geert Ridder (2005). "Mean-square-error calculations for average treatment effects," *IEPR Working Paper 05.34.*

[40] Johnson, William R. and Derek A. Neal (1998). "Basic skills and the black-white earnings gap," *The Black-White Test Score Gap*: 480 - 500. (C. Jencks & M. Phillips, Eds.). Washington, D.C.: The Brookings Institution.

[41] Johnson, William R., Yuichi Kitamura and Derek A. Neal. (2000). "Evaluating a simple method for estimating black-white gaps in median wages," *American Economic Review* 90 (2): 339 - 343.

[42] Kitamura, Yuichi. (2007). "Empirical likelihood methods in econometrics: theory and practice," *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress* 3: 174 - 237 (R. Blundell, W. Newey & T. Persson, Eds.). Cambridge: Cambridge University Press.

[43] Kitamura, Yuichi and Michael Stutzer. (1997). "An information-theoretic alternative to generalized method of moments estimation," *Econometrica* 65 (4): 861 - 874.

[44] Little, Roderick J. A. and Donald B. Rubin. (2002). *Statistical Analysis with Missing Data.* Hoboken, N.J.: John Wiley & Sons, Inc.

[45] Manski, Charles F. and Daniel McFadden. (1981). "Alternative estimators and sample designs for discrete choice analysis," *Structural Analysis of Discrete Data and Econometric Applications*: 2 - 50. (C.F. Manski & D. McFadden, Eds.). Cambridge, MA: The MIT Press.

[46] Neal, Derek A. and William R. Johnson. (1996). "The role of premarket factors in black-white wage differences," *Journal of Political Economy* 104 (5): 869 - 895.

[47] Newey, Whitney K. (1990). "Semiparametric efficiency bounds," *Journal of Applied Econometrics* 5 (2): 99 - 135.

[48] Newey, Whitney K. (1993). "Efficient estimation of models with conditional moment restrictions," *Handbook of Statistics* 11: 419 - 453. (G.S. Maddala, C.R. Rao and H.D. Vinod, Eds.). Amsterdam: North-Holland.

[49] Newey, Whitney K. (1994a). "Series estimation of regression functionals," *Econometric Theory* 10 (1): 1 - 28.

[50] Newey, Whitney K. (1994b). "The asymptotic variance of semiparametric estimators," *Econometrica* 62 (6): 1349 - 1382.

[51] Newey, Whitney K. (2002). "Stochastic expansion for M-estimator," *Mimeo.*

[52] Newey, Whitney K. and Daniel McFadden. (1994). "Large sample estimation and hypothesis testing," *Handbook of Econometrics 4:* 2111 - 2245 (R.F. Engle & D.F. McFadden, Eds.). Amsterdam: North-Holland.

[53] Newey, Whitney K. and Richard J. Smith. (2004). "Higher order properties of GMM and generalized empirical likelihood estimators," *Econometrica* 72 (1): 219 - 255.

[54] Owen, Art. B. (2001). *Empirical Likelihood.* New York: Chapman & Hall/CRC.

[55] Powell, James L. (1986). "Symmetrically trimmed least squares estimation for Tobit models," *Econometrica* 54 (6): 1435 - 1460.

[56] Ridder, Geert and Robert Moffitt. (2007). "The econometrics of data combination," *Handbook of Econometrics* 6 (2): 5469 - 5547 (J.J. Heckman & E Leamer, Eds.). New York: North-Holland.

[57] Rilstone, Paul, V. K. Srivastava and Aman Ullah. (1996). "The second-order bias and mean squared error on nonlinear estimators," *Journal of Econometrics* 75 (2): 369 - 395.

[58] Robins, James, Mariela Sued, Quanhong Lei-Gomez and Andrea Rotnitzky. (2007). "Comment: performance of double-robust estimators when "inverse probability" weights are highly variable," *Statistical Science* 22 (4): 544 - 559.

[59] Robins, James M. and Andrea Rotnitzky. (1995). "Semiparametric efficiency in multivariate regression models with missing data,".*Journal of the American Statistical Association* 90 (429): 122 - 129.

[60] Robins, James M., Andrea Rotnitzky and Lue Ping Zhao. (1994). "Estimation of regression coefficients when some regressors are not always observed,".*Journal of the American Statistical Association* 89 (427): 846 - 866.

[61] Robins, James M., Andrea Rotnitzky and Lue Ping Zhao. (1995). "Analysis of semiparametric regression models for repeated outcomes in the presence of missing data," *Journal of the American Statistical Association* 90 (429): 106 - 121.

[62] Robinson, Peter M. (1988). "Root-N-consistent semiparametric regression, *Econometrica* 56 (4): 931 - 954.

[63] Rosenbaum, Paul R. (1987). "Model-based direct adjustment," *Journal of the American Statistical Association* 82 (398): 387 - 394.

[64] Rosenbaum, Paul R. and Donald B. Rubin. (1983). "The central role of the propensity score in observational studies for causal effects," *Biometrika* 70 (1): 41 - 55.

[65] Rubin, Donald B. (1977). "Assignment to treatment group on the basis of a covariate," *Journal of Educational Statistics* 2 (1): 1 - 26.

[66] Tsiatis, Anastasios A. (2006). *Semiparametric Theory and Missing Data.* New York: Springer.

[67] Wang, Qihua, Oliver Linton and Wolfgang Härdle. (2004). "Semiparametric regression analysis with missing response at random," *Journal of the American Statistical Association* 99 (466): 334 - 345.

[68] Wooldridge, Jeffrey M. (1999). "Asymptotic properties of weighted M-estimators for variable probability samples," *Econometrica* 67 (6): 1385 - 1406.

[69] Wooldridge, Jeffrey M. (2001). "Asymptotic properties of weighted M-estimators for standard stratified samples," *Econometric Theory* 17 (2): 451 - 470.

[70] Wooldridge, Jeffrey M. (2002). "Inverse probability weighted M-estimators for sample selection, attrition and stratification," *Portuguese Economic Journal* 1 (2): 117 - 139.

[71] Wooldridge, Jeffrey M. (2007). "Inverse probability weighted estimation for general missing data problems," *Journal of Econometrics* 141 (2): 1281 - 1301.