The Stata Journal (2001) **1**, Number 1, pp. 115

Inverse Probability Tilting Estimation of Average Treatment Effects in Stata

Bryan S. Graham
New York University and NBER

Cristine Campos de Xavier Pinto
Escola de Economia de São Paulo, FGV

Daniel Egel
RAND Corporation

Abstract.

Graham, Pinto and Egel (forthcoming) introduce a new inverse probability weighting type estimator for general moment condition models with data missing at random. Estimation of the Average Treatment Effect (ATE) under exogenous treatment assignment is included in their framework. In this paper we describe a Stata implementation of their method for ATE estimation: the `iptATE` command.

Keywords: iptATE, Inverse Probability Tilting, Program Evaluation, Average Treatment Effect

When comparing subpopulations of 'treated' and 'control' units it is generally desirable to adjust for differences in observed baseline characteristics. Under strong conditions Rubin (1977) showed that such comparisons have a causal interpretation, however they may be of interest even in settings where this interpretation is untenable (cf., Imbens, 2004). In recent years a large literature has emerged proposing various methods of covariate adjustment. Imbens and Wooldridge (2011) provide a recent and comprehensive review.

When the sample size is small, and/or the number of observed baseline characteristics large, it is common for different methods of covariate adjustment to result in different average treatment effect (ATE) estimates. Furthermore, researcher chosen tuning parameters, such as the number of matches to use per unit in matching procedures, may nontrivially affect the location and estimated precision of point estimates (Abadie et al., 2004). Angrist and Pischke (2009), in reaction to the perceived sensitivity of ATE point estimates to researcher chosen implementation details, advocate for simple linear regression-based methods of covariate adjustment (pp. 80 - 91). A different response to this perception is embodied in the work of Kalyanaraman (2008), who develops automated tuning parameter choice procedures for a particular ATE estimator.

Graham, Pinto and Egel (forthcoming) propose a semiparametric method of covariate adjustment called Inverse Probability Tilting (IPT). Implementing their method requires the researcher to formulate two models. The first is for the probability of assignment to treatment conditional on characteristics or the propensity score. The second is actually a pair of models: for the conditional expectation functions (CEFs) of the potential outcomes under active and control treatment given characteristics. While the IPT point estimate of the ATE will be sensitive to them, these modelling decisions are arguably less abstract than that of, say, tuning parameter choice in the context of a

fully nonparametric method of covariate adjustment.

In addition to its conceptual simplicity the IPT estimator has a number of desirable theoretical properties. First, to use the language of Bang and Robins (2005), it is locally efficient and doubly robust. Second, relative to other estimators with similar first order asymptotic properties, such as the class of augmented inverse probability weighting (AIPW) estimators introduced by Robins, Rotnitzky and Zhao (1994), IPT has low higher order asymptotic bias.[1] Third, it provides a operational extension of methods of 'direct adjustment' or post-stratification from discretely- to continuously-valued covariates.

To understand this last feature consider a setting with a single baseline covariate, gender. If men and women differentially select into treatment, the gender composition of both the treated and control subsamples will depart from that of the full sample. In such a situation it is straightforward to reweight the two subsamples such that the distribution of gender in them mirrors that in the full sample (i.e., to restore 'balance'; cf., Rosenbaum, 1987).

When covariates are continuously-valued reweighting to restore balance is considerably more complex. IPT uses a reweighting of the data that balances a finite number of sample moments across treatment and control units. If, for example, units differentially select into treatment on the basis of their baseline earnings, then IPT can reweight the data such that the mean and variance of baseline earnings is *identical* across treatment and control units. IPT implements the idea that reweighting should make the treatment and control subsamples more comparable – in terms of the distribution of baseline characteristics – in a very concrete and aesthetically attractive way.

The underlying theory of IPT for general moment condition problems is developed in Graham, Pinto and Egel (forthcoming). The appendix to that paper also provides computational details. Other variants of inverse probability weighting (IPW) are discussed by Rosenbaum (1987), Wooldridge (2007) and Hirano, Imbens and Ridder (2003). This article presumes familiarity with the notation and language of the econometric program evaluation literature. Imbens and Wooldridge (2011) is a convenient reference for this material.

1 Review of inverse probability tilting

We seek to estimate the average effect of a binary treatment, D, on the scalar outcome Y. We let Y_1 denote a randomly sampled unit's potential outcome given assignment to the active treatment ($D = 1$) and Y_0 the corresponding potential outcome under control ($D = 0$). The average treatment effect (ATE) is

$$\gamma_0^{\text{ATE}} = \mathbb{E}\left[Y_1 - Y_0\right]. \tag{1}$$

1. For an implementation of AIPW in Stata see Emsley, Lunt, Pickles and Dunn (2008).

If both Y_1 and Y_0 were observed for all sampled units, then an analog estimate of (1) would be straightforward to construct. In practice we only observe

$$Y = (1 - D) Y_0 + DY_1,$$

or Y_1 for treated units and Y_0 for control units.

Let X denote a vector of baseline characteristics and $p_0(x) = \Pr(D = 1 | X = x)$ the propensity score or probability of assignment to the active treatment given characteristics. We assume the availability of the random sample $\{(D_i, X_i, Y_i)\}_{i=1}^N$ from the population of interest and use N_1 and N_0 to respectively denote the number of treated and control units.

If (i) D is independent of (Y_0, Y_1) conditional on $X = x$ for all $x \in \mathbb{X} \subset \mathbb{R}^{\dim(X)}$ and (ii) $\kappa < p_0(x) < 1 - \kappa$ for some $\kappa > 0$, then it is straightforward to show that γ_0 is identified (e.g., Rosenbaum and Rubin, 1983) and estimable at parametric rates (e.g., Robins, Rotnitzky and Zhao, 1994).

The following approach to identification is useful for our purposes. Rewriting (1) in integral form we have

$$\gamma_0^{\text{ATE}} = \int\int y_1 f_{Y_1,X}(y_1, x)\, \mathrm{d}y_1 \mathrm{d}x - \int\int y_0 f_{Y_0,X} f(y_0, x)\, \mathrm{d}y_0 \mathrm{d}x. \tag{2}$$

From condition (i) above we have the equalities

$$f_{Y_1|X}(y_1 | x) = f_{Y|X,D}(y | x, d = 1), \quad f_{Y_0|X}(y_0 | x) = f_{Y|X,D}(y | x, d = 0). \tag{3}$$

Bayes' Law gives the additional pair of equalities

$$f_X(x) = f_{X|D}(x | d = 1) \frac{Q_0}{p_0(x)} = f_{X|D}(x | d = 0) \frac{1 - Q_0}{1 - p_0(x)}, \tag{4}$$

where $Q_0 = \mathbb{E}[D]$ is the marginal frequency of treatment.

Substituting (3) and (4) into (2) and consolidating terms we get the following representation of the ATE

$$\gamma_0^{\text{ATE}} = \int\int y f_{Y,X|D}(y, x | d = 1) \frac{Q_0}{p_0(x)} \mathrm{d}y \mathrm{d}x - \int\int y f_{Y,X|D}(y, x | d = 0) \frac{1 - Q_0}{1 - p_0(x)} \mathrm{d}y \mathrm{d}x.$$

Each component on the right-hand-side of the above expression is identified by the joint distribution of the observed data (D, X, Y).

To construct an analog estimator based on the above representation we replace $f_{Y,X|D}(y, x | d = 1)$ and $f_{Y,X|D}(y, x | d = 0)$ with the empirical measures of the treated and control subsamples. These measures, which may be viewed as nonparametric maximum likelihood estimates (NPMLEs), place weight N_1^{-1} and N_0^{-1} on each treated and

control unit (Owen, 2001). Replacing Q_0 and $p_0(x)$ with the estimates $\widehat{Q}$ and $\widehat{p}(x)$ we get

$$\begin{aligned}\widehat{\gamma}^{\text{ATE}}_{(\widehat{Q},\widehat{p})} &= \sum_{i=N_0+1}^{N_1} \frac{\widehat{Q}}{N_1}\frac{1}{\widehat{p}(X_i)}Y_i - \sum_{i=1}^{N_0}\frac{1-\widehat{Q}}{N_0}\frac{1}{1-\widehat{p}(X_i)}Y_i \\ &= \sum_{i=N_0+1}^{N_1}\widehat{\pi}_1 Y_i - \sum_{i=1}^{N_0}\widehat{\pi}_0 Y_i,\end{aligned} \tag{5}$$

where

$$\begin{aligned}\widehat{\pi}_{0i} &= \frac{1-\widehat{Q}}{N_0}\frac{1}{1-\widehat{p}(X_i)}, \quad i=1,\ldots,N_0 \\ \widehat{\pi}_{1i} &= \frac{\widehat{Q}}{N_1}\frac{1}{\widehat{p}(X_i)}, \quad i=N_0+1,\ldots,N.\end{aligned} \tag{6}$$

Observe that, assuming $\widehat{Q}$ and $\widehat{p}(x)$ are consistent estimates,

$$\begin{aligned}\widehat{F}_{Y_0,X}(y_0,x) &= \sum_{i=1}^{N_0}\widehat{\pi}_0 \mathbf{1}(X_i \le x)\mathbf{1}(Y_i \le y_0) \\ \widehat{F}_{Y_1,X}(y_1,x) &= \sum_{i=N_0+1}^{N_1}\widehat{\pi}_1 \mathbf{1}(X_i \le x)\mathbf{1}(Y_i \le y_1)\end{aligned}$$

are consistent estimates of $F_{Y_0,X}(y_0,x)$ and $F_{Y_1,X}(y_1,x)$. These estimates, and consequently the ultimate ATE point estimate, $\widehat{\gamma}^{\text{ATE}}_{(\widehat{Q},\widehat{p})}$, vary with the choice of $\widehat{Q}$ and $\widehat{p}(x)$. In practice it is common to replace $\widehat{Q}$ and $\widehat{p}(X_i)$ with maximum likelihood estimates (MLEs). This approach requires the propensity score to be parametrically specified. Assume that $G\left(t(x)'\delta_0\right) = p_0(x)$ for all $x \in \mathbb{X}$ and some δ_0 where $t(X)$ is a $1+M$ column vector of known functions of X with a constant as its first element and $G(\cdot)$ a known, strictly increasing, and differentiable, function which maps the real line onto the unit interval (e.g., the logit function $G(v) = 1/(1+\exp(-v))$).

Let $\widehat{\delta}_{ML}$ and $\widehat{Q} = N_1/N$ denotes the MLEs of, respectively, δ_0 and Q_0.[2] Using these estimates we have, plugging into (5),

$$\widehat{\gamma}^{\text{ATE}}_{(\frac{N_1}{N},\widehat{p}_{ML})} = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{D_i}{G\left(t(X_i)'\widehat{\delta}_{ML}\right)} - \frac{1-D_i}{1-G\left(t(X_i)'\widehat{\delta}_{ML}\right)}\right\}Y_i, \tag{7}$$

which is the inverse probability weighting ATE estimator discussed by Imbens (2004) and Wooldridge (2007).

2. In the present setting the MLE of Q_0 is actually $\sum_{i=1}^{N} G\left(t(X_i)'\widehat{\delta}_{ML}\right)/N$. In the most common situation, where $G(\cdot)$ is assumed to take the logit form, we have, by the estimating equations of the logit MLE, the equality $\sum_{i=1}^{N} G\left(t(X_i)'\widehat{\delta}_{ML}\right)/N = N_1/N$.

While it is intuitive to replace $\widehat{Q}$ and $\widehat{p}(x)$ in (5) with their MLEs, it turns out that this choice has some undesirable consequences. First, in this case, $\widehat{F}_{Y_0,X}(y_0,x)$ and $\widehat{F}_{Y_1,X}(y_1,x)$ need not integrate to one (i.e., the IPW weights do not sum to one). Second, the resulting ATE point estimate is inefficient. Specifically its asymptotic sampling variance is generally greater than the semiparametric variance bound derived by Hahn (1998).

Wooldridge (2007) shows that the asymptotic sampling variance of (7) can be reduced by overfitting the propensity score (i.e., by including terms in $t(X)$ that don't enter the true selection probability). Hirano, Imbens and Ridder (2003) generalize this result, showing that if the dimension of $t(X)$ grows with the sample size in a specific way and $G(\cdot)$ takes the logit form, then IPW is semiparametrically efficient.

Graham, Pinto and Egel (forthcoming) propose an alternative variant of (5) which they term inverse probability tilting (IPT). The key difference between their approach and (7) is that they replace the maximum likelihood estimate of the propensity score with a particular method of moments one. In fact they utilize two propensity score estimates, which we now describe.

Let $\widehat{\delta}^1_{IPT}$ be the solution (if it exists, see below) to

$$\frac{1}{N}\sum_{i=1}^{N}\left\{\frac{D_i}{G\left(t(X_i)'\widehat{\delta}^1_{IPT}\right)}-1\right\}t(X_i)=0. \tag{8}$$

Plugging $G\left(t(X_i)'\widehat{\delta}^1_{IPT}\right)$ and $\widehat{Q}=N_1/N$ into (6) we get

$$\widehat{\pi}_{1i}=\frac{1}{N}\frac{1}{G\left(t(X_i)'\widehat{\delta}^1_{IPT}\right)}. \tag{9}$$

Rearranging (8) then yields

$$\sum_{i=N_0+1}^{N_1}\widehat{\pi}_{1i}t(X_i)=\frac{1}{N}\sum_{i=1}^{N}t(X_i). \tag{10}$$

Equation (10) indicates that IPT chooses the propensity score parameter so that, *after reweighting*, the mean of $t(X_i)$ across treated individuals ($\sum_{i=N_0+1}^{N_1}\widehat{\pi}_{1i}t(X_i)$) is *numerically identically* to the full sample mean ($\frac{1}{N}\sum_{i=1}^{N}t(X_i)$). This is intuitively attractive: the goal of reweighting is to make the treated units 'more like' the population as a whole. IPT ensures *exact* comparability in terms of a finite number of moments of X. For example if $t(X)$ includes, in addition to a constant, X and its square, then, after reweighting, the mean and variance of X across treated units will be identical to the corresponding objects calculated using the full sample.

We refer to $\{\widehat{\pi}_{1i}\}_{i=N_0+1}^{N}$ as the inverse probability tilt of the treated subsample. This tilt can be given an information theoretic interpretation. Specifically, it represents

a reweighting that satisfies (10), while being 'closest' to the empirical measure of the treated subsample (which places weight N_1^{-1} on all units). The discrepancy metric by which 'closest' is defined is related to the presumed form of the propensity score (see the Appendix of Graham, Pinto and Egel (forthcoming) and also Hirano, Imbens, Ridder and Rubin, 2001).

We also compute an inverse probability tilt of the control subsample ($\{\widehat{\pi}_{0i}\}_{i=1}^{N_0}$). Let $\widehat{\delta}_{IPT}^{0}$ be the solution to

$$\frac{1}{N}\sum_{i=1}^{N}\left\{\frac{1-D_i}{1-G\left(t\left(X_i\right)'\widehat{\delta}_{IPT}^{0}\right)}-1\right\}t\left(X_i\right)=0. \tag{11}$$

Plugging $G\left(t\left(X_i\right)'\widehat{\delta}_{IPT}^{0}\right)$ and $\widehat{Q}=N_1/N$ into (6) we get

$$\widehat{\pi}_{0i}=\frac{1}{N}\frac{1}{1-G\left(t\left(X_i\right)'\widehat{\delta}_{IPT}^{1}\right)} \tag{12}$$

Rearranging (11) we get a control subsample analog of (10):

$$\sum_{i=1}^{N_0}\widehat{\pi}_{0i}t\left(X_i\right)=\frac{1}{N}\sum_{i=1}^{N}t\left(X_i\right). \tag{13}$$

Equating (10) and (13) we have

$$\sum_{i=1}^{N_0}\widehat{\pi}_{0i}t\left(X_i\right)=\sum_{i=N_0+1}^{N_1}\widehat{\pi}_{1i}t\left(X_i\right)=\frac{1}{N}\sum_{i=1}^{N}t\left(X_i\right).$$

The IPT tilts, $\{\widehat{\pi}_{0i}\}_{i=1}^{N_0}$ and $\{\widehat{\pi}_{1i}\}_{i=N_0+1}^{N}$, are constructed such that the reweighted treatment and control subsample means of $t\left(X_i\right)$ are numerically identical to the unweighted full sample mean.

The IPT average treatment effect estimate is given by (5) with $\widehat{Q}=N_1/N$ and the first and second instances of $\widehat{p}\left(X_i\right)$ respectively replaced by $G\left(t\left(X_i\right)'\widehat{\delta}_{IPT}^{1}\right)$ and $G\left(t\left(X_i\right)'\widehat{\delta}_{IPT}^{0}\right)$:

$$\begin{aligned}\widehat{\gamma}_{IPT}^{\text{ATE}} &= \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{D_i}{G\left(t\left(X_i\right)'\widehat{\delta}_{IPT}^{1}\right)}-\frac{1-D_i}{1-G\left(t\left(X_i\right)'\widehat{\delta}_{IPT}^{0}\right)}\right\}Y_i \\ &= \sum_{i=N_0+1}^{N_1}\widehat{\pi}_{1i}Y_i-\sum_{i=1}^{N_0}\widehat{\pi}_{0i}Y_i,\end{aligned} \tag{14}$$

where $\widehat{\pi}_{1i}$ and $\widehat{\pi}_{0i}$ are as defined by (9) and (12).

The first line of (14) indicates that IPT is a inverse probability weighting type estimator, albeit based on distinctive propensity score estimates. The second line shows that IPT estimates the ATE by a difference between a weighted average outcome across the treated minus a weighted average outcome across controls. The precise weighting used in this computation is chosen so that, after reweighting, the two subsamples share identical means of $t(X)$.

A concrete example helps to illustrate the main idea. Say X equals years of completed schooling, setting

$$t(X) = (1, \mathbf{1}(X < 12), \mathbf{1}(X = 12), \mathbf{1}(12 < X < 16), \mathbf{1}(X = 16))'$$

will generate reweightings that ensure that the (weighted) sample fractions of high school dropouts, high school graduates, those with some college, those with a 4-year degree, and those with at least some graduate education are identical across treated and control units. Alternatively we might choose

$$t(X) = (1, X, X^2)',$$

which would ensure equality of the average and variance of years of completed schooling.

An interesting feature of $\widehat{\gamma}_{IPT}^{\text{ATE}}$ is that it is based on two separate propensity score estimates. This feature suggests a simple specification test. If the propensity score model is correctly specified, then the probability limits of $\widehat{\delta}_{IPT}^{1}$ and $\widehat{\delta}_{IPT}^{0}$ will coincide. If the model is misspecified, then this need not be true. Let δ_*^1 and δ_*^0 denote the possibility different plims of $\widehat{\delta}_{IPT}^{1}$ and $\widehat{\delta}_{IPT}^{0}$. A Wald test of the null hypothesis of equality of these two vectors is test of correct specification of the propensity score. `iptATE` reports this test statistics as part of its default output.

Choosing $t(X)$: To use the IPT estimator the researcher must choose which functions of X to include in $t(X)$. It is obvious that $t(X)$ should be rich enough to ensure that $G\left(t(x)'\delta_0\right) = p_0(x)$ for all $x \in \mathbb{X}$ and some δ_0. Less obviously, the choice of $t(X)$ should also be guided by researcher beliefs regarding the forms of the potential outcome regression functions. Specifically assume that $t(X)$ is rich enough such that

$$\mathbb{E}[Y_1|X] = \Pi_1 t(X), \qquad \mathbb{E}[Y_0|X] = \Pi_0 t(X), \tag{15}$$

for some Π_1 and Π_0. Graham, Pinto and Egel (forthcoming) show that, under the maintained assumption that the propensity score model is correctly specified, the asymptotic sampling variance of $\sqrt{N}\left(\widehat{\gamma}_{IPT}^{\text{ATE}} - \gamma_0\right)$ coincides with the bound derived by Hahn (1998) for all data generating processes which satisfy (15). This is a 'local' efficiency property. Their result suggests that the precision of $\widehat{\gamma}_{IPT}^{\text{ATE}}$ will be greatest when, in addition to being rich enough to provide a good approximation of the propensity score, $t(X)$ also includes enough elements such that a linear combination of them closely approximates both $\mathbb{E}[Y_1|X = x]$ and $\mathbb{E}[Y_0|X = x]$.

Graham, Pinto and Egel (forthcoming) also show that $\widehat{\gamma}_{IPT}^{\text{ATE}}$ remains consistent for γ_0^{ATE} even if the propensity score model is misspecified as long as (15) holds. This

is a double robustness property. While other locally efficient and doubly robust ATE estimators are available, Graham, Pinto and Egel (forthcoming) show that $\widehat{\gamma}_{IPT}^{\text{ATE}}$ has lower higher order bias than a large class of them. Bang and Robins (2005) and Tsiatis (2006) provide a recent and accessible introductions to double robust ATE estimation.

A more heuristic suggestion for choosing $t(X)$ is as follows. First, transform any continuously-valued components of X such that they are approximately Gaussian. For example we might take the log-transform of an earnings variable. Include the transformed X and its square as elements of $t(X)$. For discretely-valued variables include a dummy variable for each point of support (if the variable has many points of support some aggregation may be required). Finally include all pairwise interactions of the above variables (cf., Anderson, 1982).

Convex hull condition: For a solution to (8) to exist the full sample mean of $t(X_i)$ must lie inside the convex hull of the treated subsample data (cf., Owen, 2001). If the propensity score is strictly bounded between zero and one this condition will be satisfied in large enough samples. However, it may fail in small samples, particularly if overlap is weak. When $\sum_{i=1}^{N} t(X_i)/N$ lies near the boundary of the convex hull of the treated subsample data, the computation of $\widehat{\delta}_{IPT}^{1}$ may become difficult. While Graham, Pinto and Egel (forthcoming) develop a reliable computational algorithm, which is implemented in `iptATE`, users should be aware that the existence of $\widehat{\delta}_{IPT}^{1}$ is not automatic. The inability of `iptATE` is solve for $\widehat{\delta}_{IPT}^{1}$, is generally indicative of a weak research design with poor overlap. Users should be mindful that the convex hull condition will impose practical limitations on the richness of $t(X_i)$ in finite samples. For example, if baseline earnings are highly predictive of treatment status and only a few hundred units are available, it is unlikely that an inverse probability tilt of the data which balances the first eight moments of baseline earnings exists. If may be possible, however, to balance the first two moments of baseline earnings in such a situation. Analogous considerations apply to $\widehat{\delta}_{IPT}^{0}$.

Consistent variance estimation: Graham, Pinto and Egel (forthcoming) show that $\widehat{\gamma}_{IPT}^{\text{ATE}}$ is a sequential method of moments estimate. Consequently the sampling variance of $\widehat{\gamma}_{IPT}^{\text{ATE}}$ may be consistently estimated using standard results (e.g., Newey and McFadden, 1994). The relevant sample moments are

$$\sum_{i=1}^{N} \texttt{smplwgt}_i \times \begin{pmatrix} \left\{ \frac{D_i}{G\left(t(X_i)'\widehat{\delta}_{IPT}^{1}\right)} - 1 \right\} t(X_i) \\ \left\{ \frac{1-D_i}{1-G\left(t(X_i)'\widehat{\delta}_{IPT}^{0}\right)} - 1 \right\} t(X_i) \\ \frac{D_i}{G\left(t(X_i)'\widehat{\delta}_{IPT}^{1}\right)} - \frac{1-D_i}{1-G\left(t(X_i)'\widehat{\delta}_{IPT}^{0}\right)} \left(Y_i + \widehat{\gamma}_{IPT}^{\text{ATE}}\right) \end{pmatrix} = 0. \tag{16}$$

Where `smplwgt` is a user-specified sampling weight (i.e., a 'pweight' in Stata terminology). If no weights are specified `iptATE` replaces $\texttt{smplwgt}_i$ by $\frac{1}{N}$ in (16).

Stata's built in robust and 'clustered' robust variance estimators automatically implement a degrees of freedom correction (see the `regress` entry in the Stata 11 *Reference Q - Z*). Similar degrees of freedom corrections are *not* implemented by `iptATE`.The default standard errors reported by `iptATE` allow for heteroscedasticity (specifically the conditional variances of Y_1 and Y_0 can vary across subpopulations defined in terms of $X = x$). When the primary sampling unit is not a single unit, `iptATE` provides a 'clustered' variance-covariance estimator option.

2 A simple empirical example

In this section we use the dataset constructed by Graham and Powell (2008) to illustrate the use of IPT in practice. The data were collected in conjunction with an external evaluation of the Nicaraguan conditional cash transfer program Red de Protección Social (RPS) (see IFPRI, 2005). The RPS evaluation sample is a panel of 1,581 households from 42 rural communities in the departments of Madriz and Matagalpa, located in the northern part of the Central Region of Nicaragua. Each sampled household was first interviewed in August/September 2000 with follow-ups attempted in October of both 2001 and 2002. Here we analyze a balanced panel of 1,358 households from all three waves. The dataset includes a measure of total calories available per capita for each household, total real expenditure per capita, measures of household size and demographic structure, as well as a binary indicator for whether the household was located in a treatment or control village. Graham and Powell (2008) provide full details of the sample and variable construction. The data file used here, `RPSPolicyEvalData.dta`, is available online at `https://files.nyu.edu/bsg1/public/`. A Stata Do file which replicates the results reported below is also available.

We study the effect of RPS participation on calorie availability per capita in 2002. In the dataset the variable `log_calories_pc2002` equals the logarithm of calorie availability per capita in 2002 and `RPS` denotes whether the household resides in a treatment village. Because the `RPS` was a randomly assigned to communities the coefficient on the treatment indicator in a least squares fit of `log_calories_pc2002` onto a constant and `RPS` provides a consistent estimate of the ATE.

```
. use "$BASE\RPSPolicyEvalData", clear
(LdB 2000 : Annual food consumption -- components)

. reg log_calories_pc2002 RPS, cluster(village)

[Output removed]
                              (Std. Err. adjusted for 42 clusters in village)
------------------------------------------------------------------------------
             |               Robust
log_cal~2002 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         RPS |   .2080808   .0644476     3.23   0.002     .0779262    .3382354
```

```
       _cons |    7.62022     .05181   147.08   0.000     7.515588    7.724853
------------------------------------------------------------------------------
```

The simple treatment versus control difference suggests that the RPS program increased calorie availability per capita by about 20 percent. This is a very large program effect. We could increase the precision of this estimate by including various pretreatment household characteristics in our regression model, instead we illustrate how to estimate the ATE using the Stata command: `iptATE`. The syntax of the `iptATE` command is

`iptATE` *outcomevar treatmentvar controlvars* [`weight`] [`if` *exp*] [`in` *range*] [, `cluster`(*clustervar*) `optroutine`(`{e1`|`e2}`) `balance`]

where *outcomevar* is the variable name for the scalar outcome of interest, *treatmentvar* the variable name for the treatment indicator (*treatmentvar* = 1 denotes treatment, 0 control) and *controlvars* is a list of variable names which define $t(X)$; a constant is automatically included. Sampling weights or `pweights` are allowed; see the Stata 11 *User's Guide* for more information (`[U] 20.18`).

Options

`cluster`(*clustervar*) specifies the relevant sampling unit. If this option is omitted each observation is assumed to be an independent random draw from the population of interest. When the program of interest is 'assigned' at a group or aggregate level (e.g., at the village level in the case of the RPS program) then this option should be specified with *clustervar* the name of a group-identifying variable.

`optroutine`(`{e1`|`e2}`) specifies the optimization algorithm used to compute $\widehat{\delta}^0_{IPT}$ and $\widehat{\delta}^1_{IPT}$. The default is `e1`, which means that $\widehat{\delta}^0_{IPT}$ and $\widehat{\delta}^1_{IPT}$ are computed using a quasi-Newton procedure with analytic first derivatives. Specifically `iptATE` uses the implementation of the Broyden-Fletcher-Goldfarb-Shanno algorithm found in Mata's `moptimize()` command (see Stata 11 *Mata Matrix Programming*). If `e2` is specified a modified Newton-Raphson procedure is used. This procedures uses both analytic first and second derivatives. This second method is generally quicker and more accurate, particularly for problems where the convex hull condition is amply satisfied. The criterion function used by `iptATE` is described in the Appendix to Graham, Pinto and Egel (forthcoming).

`balance` when invoked `iptATE` will produce a table with the (unweighted) means of *controlvars* by treatment status as well as the treatment versus control difference. If `pweights` are also specified, then these means will be appropriately weighted.

As an initial illustration we include in $t(X)$ total household size as well as the logarithm per capita calorie availability and expenditure in 2000:

```
. iptATE log_calories_pc2002 RPS log_calories_pc2000 log_real_exp_pc2000 HHSize2000,
                                  cluster(village) optroutine(e2) balance
```

```
Mean covariate values by treatment status
------------------------------------------------------------------------------
-                Variable     T = 0          T = 1          Diff            p-value    -
------------------------------------------------------------------------------
      log_calories_pc2000         7.485          7.498          0.013
                             (    0.057)    (    0.053)    (    0.078)       0.868

      log_real_exp_pc2000         8.115          8.205          0.089
                             (    0.069)    (    0.059)    (    0.091)       0.330

                HHSize2000         6.135          5.916         -0.218
                             (    0.078)    (    0.132)    (    0.154)       0.162
------------------------------------------------------------------------------

Computing treated subsample tilt
initial:       f(p) = -754.35262
rescale:       f(p) = -702.39165
Iteration 0:   f(p) = -702.39165
Iteration 1:   f(p) = -691.96712
Iteration 2:   f(p) = -691.85999
Iteration 3:   f(p) = -691.85996

Computing control subsample tilt
initial:       f(p) =  -738.4736
rescale:       f(p) =       -653
Iteration 0:   f(p) =       -653
Iteration 1:   f(p) = -638.89437
Iteration 2:   f(p) =  -638.5741
Iteration 3:   f(p) = -638.57396
Iteration 4:   f(p) = -638.57396

Inverse probability tilting propensity score & ATE estimates
Outcome variable    : log_calories_pc2002
Treatment indicator : RPS
Control variables   : log_calories_pc2000 log_real_exp_pc2000 HHSize2000
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
delta1       |
log_cal~2000 |  -.3093419   .2788068    -1.11   0.267    -.8557933    .2371094
log_rea~2000 |   .4153759   .2982215     1.39   0.164    -.1691276    .9998794
  HHSize2000 |  -.0063623   .0223771    -0.28   0.776    -.0502207    .0374961
       _cons |  -.9568784   2.215386    -0.43   0.666    -5.298954    3.385197
-------------+----------------------------------------------------------------
```

```
delta0       |
log_cal~2000 |  -.3044234    .245151    -1.24   0.214    -.7849106    .1760637
log_rea~2000 |   .3993781   .2611813     1.53   0.126    -.1125278     .911284
  HHSize2000 |  -.0148753    .024551    -0.61   0.545    -.0629944    .0332437
       _cons |  -.8119363   2.139014    -0.38   0.704    -5.004327    3.380455
-------------+----------------------------------------------------------------
ate          |
       gamma |   .1854696   .0486575     3.81   0.000     .0901025    .2808366
------------------------------------------------------------------------------
Total number of primary sampling units: 42
Total number of observations: 1358

Test of equality of two tilting coefficient vectors

 ( 1)  [delta1]_cons - [delta0]_cons = 0
 ( 2)  [delta1]log_calories_pc2000 - [delta0]log_calories_pc2000 = 0
 ( 3)  [delta1]log_real_exp_pc2000 - [delta0]log_real_exp_pc2000 = 0
 ( 4)  [delta1]HHSize2000 - [delta0]HHSize2000 = 0

           chi2(  4) =    0.06
         Prob > chi2 =    0.9995
```

The initial part of the above output gives the means of `log_calories_pc2000`, `log_real_exp_pc2000` and `HHSize2000` by treatment status. The second-to-last column reports the treatment versus control difference (standard errors are in parentheses). As would be expected in the context of a social experiment, there is no evidence of significant average differences in these variables by treatment status (see the last column of the table for p-values).

The next portion of the output shows that Stata was able to compute $\widehat{\delta}^1_{IPT}$ and $\widehat{\delta}^0_{IPT}$ in, respectively, 3 and 4 Newton-Raphson iterations (`optroutine(e2)` was specified).[3] The next part of the output reports the IPT estimates $\widehat{\delta}^1_{IPT}$ and $\widehat{\delta}^0_{IPT}$, along with standard errors. These standard errors are robust to arbitrary patterns of dependence across households residing in the same village (`cluster(village)` was specified). Finally the IPT point estimate of the ATE is reported. The IPT point estimate is similar to the raw treatment versus control difference in means reported above, as one would expect given random assignment of `RPS` to communities, but its estimated standard error is about 10 percent smaller.

The final part of the output gives a χ^2 statistics for the null that δ^1_* and δ^0_* coincide. We accept, by a wide margin, the null that the propensity score model is correctly specified. This is unsurprising since we know that the true score is a constant.

Next we consider a richer specification for $t(X)$. Since the underlying data are from an experiment, the main potential advantage of including more elements in $t(X)$

3. Choosing `optroutine(e1)` instead requires, respectively, 9 and 8 iterations.

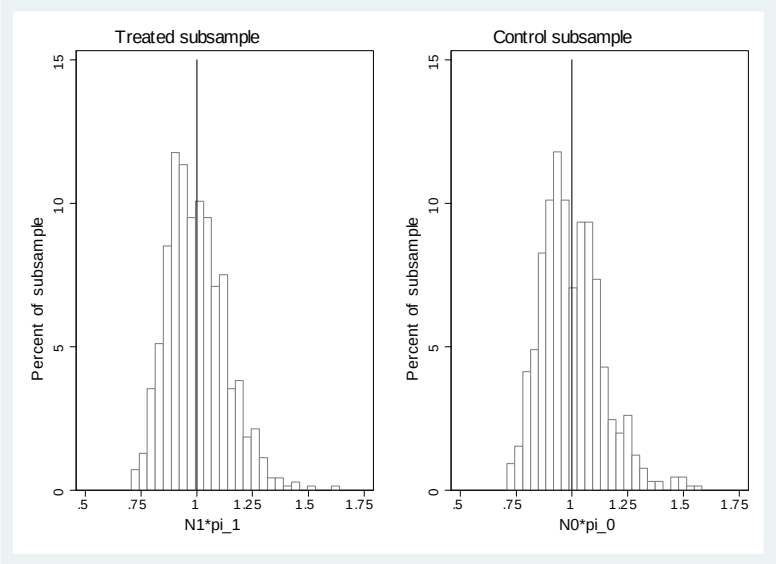

Figure 1. Histograms of $N_1\widehat{\pi}_{1i}$ and $N_0\widehat{\pi}_{0i}$

is improved precision. We created 5 dummy variables for `log_calories_pc2000` lying in the regions $(-\infty, 7.00]$, $(7.00, 7.25]$, $(7.25, 7.50]$, $(7.50, 7.75]$, $(7.75, 8.00]$. This grid of points was chosen to span the support of baseline calorie availability. We created a similar set of dummy variables based on `log_real_exp_pc2002` lying in the regions $(-\infty, 7.0]$, $(7.0, 7.5]$, $(7.5, 8.0]$, $(8.0, 8.5]$, $(8.5, 9.0]$ and `HHSize2000` lying in the regions $(0, 2]$, $(2, 6]$, $(6, 9]$. Finally we added the square of `log_calories_pc2000` and `log_real_exp_pc2002` to $t(X)$. This resulted in a specification of $t(X)$ with nineteen elements. This choice of $t(X)$ ensures, after reweighting, exact balance of the mean and variance of baseline calorie availability and expenditure across treatment and controls. It also ensures that the fraction of households in different regions of calorie availability and expenditure are the same. Finally it ensures average household size, as well as its general distribution, is the same across the two groups.

```
. iptATE log_calories_pc2002 RPS log_calories_pc2000 lc00_sq lc00_g1-lc00_g5
>                                log_real_exp_pc2000 lre00_sq lre00_g1-lre00_g5
>                                HHSize2000 hhs00_g1-hhs00_g3,
                                 cluster(village) optroutine(e2)

[Output removed]

-------------+----------------------------------------------------------------
ate          |
       gamma |   .1841529    .048632     3.79   0.000      .088836    .2794699
```

```
------------------------------------------------------------------------------

[Output removed]
```

While `iptATE` has no trouble computing the IPT tilts based on the enriched specification of $t(X)$, neither the ATE point estimate nor its estimated precision is appreciably affected.

Post estimation features

`iptATE` saves in `e()`:

Scalars:

`e(N)` - number of observations

Matrices:

e(b) - coefficient vector

e(V) - variance-covariance matrix

Functions:

`e(sample)` - marks estimation sample

`iptATE` stores the estimated coefficient vector and variance-covariance matrix in multiple equation form. The first equation is called `delta1`, corresponding to the coefficients indexing the treated subsample tilt, the second equation is called `delta0`, corresponding to the coefficients indexing the control subsample tilt, the final equation is called `ate`, corresponding to the average treatment effect.

A useful postestimation diagnostic is to inspect the two tilts ($\{\widehat{\pi}_{0i}\}_{i=1}^{N_0}$ and $\{\widehat{\pi}_{1i}\}_{i=N_0+1}^{N}$). One way to do this is to plot histograms of $N_0\widehat{\pi}_{0i}$ and $N_1\widehat{\pi}_{1i}$. If $N_1\widehat{\pi}_{1i} = 4$, then this indicates that the IPT upweights the i^{th} treated unit by a factor of four relative to the empirical measure of the treated subsample. A downweighting by a factor of four occurs if $N_1\widehat{\pi}_{1i} = 1/4$.

Since both of the inverse probability tilts sum to one, and attach positive weight to all units in the appropriate subsample, we know that $0 < N_0\widehat{\pi}_{0i} < N_0$ for $i = 1, \ldots, N_0$ and $0 < N_1\widehat{\pi}_{1i} < N_1$ for $i = N_0+1, \ldots, N$. Boundedness and positivity of the IPT weights is an important feature of the estimator (Graham, Pinto and Egel, forthcoming). Settings where large weight is attached to a handful of units, and very small weight to the remaining units, are indicative of poor overlap. To plot these histograms we use Stata's `predict` and `hist` commands. The following commands calculate $N_0\widehat{\pi}_{0i}$ and $N_1\widehat{\pi}_{1i}$ for, respectively, all control and treated units.

```
predict pi_1 if RPS==1, equation(delta1) xb
replace pi_1 = (1/e(N))*(1/invlogit(pi_1)) if RPS==1

predict pi_0 if RPS==0, equation(delta0) xb
replace pi_0 = (1/e(N))*(1/(1-invlogit(pi_0))) if RPS==0

quietly reg RPS if RPS==1
g N1Xpi_1 = e(N)*pi_1 if RPS==1

quietly reg RPS if RPS==0
g N0Xpi_0 = e(N)*pi_0 if RPS==0
```

Histograms of `N1Xpi_1` and `N0Xpi_0` are given in Figure 1.

3 References

Abadie, Alberto, David Drukker, Jane Leber Herr and Guido W. Imbens. (2004). "Implementing Matching Estimators for Average Treatment Effects in Stata," *The Stata Journal* 4 (3): 290 - 311.

Anderson, J.A. (1982). "Logistic discrimination," *Handbook of Statistics* 2: 169 - 191 (P.R. Krishnaiah & L.N. Kanal, Eds.). Amsterdam: North-Holland.

Angrist, Joshua D. and Jörn-Steffen Pischke. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton, NJ: Princeton University Press.

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

Emsley, Richard, Mark Lunt, Andrew Pickles, and Graham Dunn. (2008). "Implementing double-robust estimators of causal effects," *The Stata Journal* 8 (3): 334 - 353.

Graham, Bryan S., Cristine Pinto and Daniel Egel. (forthcoming). "Inverse probability tilting for moment condition models with missing data," *Review of Economic Studies.*

Graham, Bryan S. and James L. Powell (2008). "Identification and estimation of average partial effects in 'irregular' correlated random coefficient panel data models," *NBER Working Paper w14469.*

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

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.

Hirano, Keisuke, Guido W. Imbens, Geert Ridder, Donald B. Rubin. (2001). "Combining panel data sets with attrition and refreshment samples," *Econometrica* 69 (6):

1645 - 1659.

International Food Policy Research Institute (IFPRI). (2005). *Nicaraguan RPS evaluation data (2000-02): overview and description of data files (April 2005 Release).* Washington D.C.: International Food Policy Research Institute.

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

Imbens, Guido W., and Jeffrey M. Wooldridge. (2009). "Recent developments in the econometrics of program evaluation," *Journal of Economic Literature* 47 (1): 5 – 86.

Kalyanaraman, Karthik. (2008). "Bandwidth choice for regression functionals with application to average treatment effects," *Mimeo.*

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.

Owen, Art B. (2001). *Empirical Likelihood.* Boca Raton: Chapman & Hall, CRC.

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.

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

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.

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

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

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