Title

stata.com

teffects intro advanced — Advanced introduction to treatment effects for observational data

Description

This entry provides a technical overview of treatment-effects estimators and their implementation in Stata. Those who are new to treatment-effects estimation may want to instead see [TE] **teffects intro**.

The `teffects` command estimates potential-outcome means (POMs), average treatment effects (ATEs), and average treatment effects among treated subjects (ATETs) using observational data.

Treatment effects can be estimated using regression adjustment (RA), inverse-probability weights (IPW), and "doubly robust" methods, including inverse-probability-weighted regression adjustment (IPWRA) and augmented inverse-probability weights (AIPW), and via matching on the propensity score or nearest neighbors.

The outcome models can be continuous, binary, count, or nonnegative. Continuous outcomes can be modeled using linear regression; binary outcomes can be modeled using logit, probit, or heteroskedastic probit regression; and count and nonnegative outcomes can be modeled using Poisson regression. The treatment model can be binary or multinomial. Binary treatments can be modeled using logit, probit, or heteroskedastic probit regression, while multinomial outcomes are modeled using multinomial logit regression.

Remarks and examples

stata.com

This entry presents a technical overview of treatment-effects estimators and their implementation in Stata. Users who are new to treatment-effects estimators for observational data should instead read [TE] **teffects intro**.

Remarks are presented under the following headings:

Introduction

The `teffects` commands estimate treatment effects from observed data. A treatment effect is the change in an outcome caused by a subject, often an individual, getting one treatment instead of another. We cannot estimate individual-level treatment effects, because we only observe each individual getting one or another treatment.

Potential-outcome models provide a solution to this missing-data problem and allow us to estimate the distribution of individual-level treatment effects. A potential-outcome model specifies the potential outcomes that each individual would obtain under each treatment level, the treatment assignment process, and the dependence of the potential outcomes on the treatment assignment process.

When the potential outcomes do not depend on the treatment levels, after conditioning on covariates, regression estimators, inverse-probability-weighted estimators, and matching estimators are commonly used.

What we call the potential-outcome model is also known as the Rubin causal model and the counterfactual model. See Rubin (1974); Holland (1986); Robins (1986); Heckman (1997); Heckman and Navarro-Lozano (2004); Imbens (2004); Cameron and Trivedi (2005, chap. 2.7); Imbens and Wooldridge (2009); and Wooldridge (2010, chap. 21) for more detailed discussions.

Defining treatment effects

Three parameters are often used to measure treatment effects: the average treatment effect (ATE), the average treatment effect on the treated (ATET), and the potential-outcome means (POMs). In this section, we define each of these terms and introduce the notation and parameters used in the rest of our discussion.

In the binary-treatment case, the two potential outcomes for each individual are y_{0i} and y_{1i}; y_{0i} is the outcome that would be obtained if i does not get the treatment, and y_{1i} is the outcome that would be obtained if i gets the treatment. y_{0i} and y_{1i} are realizations of the random variables y_0 and y_1. Throughout this entry, i subscripts denote realizations of the corresponding unsubscripted random variables. We do not discuss multivalued treatments here, because doing so only increases the number of parameters and notation required and detracts from the essential points; see [TE] **teffects multivalued** for information about multivalued treatments.

The parameters of interest summarize the distribution of the unobservable individual-level treatment effect $y_1 - y_0$. In defining the parameters, t denotes a random treatment, t_i denotes the treatment received by individual i, $t = 1$ is the treatment level, and $t = 0$ is the control level. Given this notation, we can now define our parameters of interest.

ATE The ATE is the average effect of the treatment in the population:

$$\text{ATE} = E(y_1 - y_0)$$

POM The POM for treatment level t is the average potential outcome for that treatment level:

$$\text{POM}_t = E(y_t)$$

ATET The ATET is the average treatment effect among those that receive the treatment:

$$\text{ATET} = E(y_1 - y_0 | t = 1)$$

For an illustration of these concepts, see *Defining treatment effects* in [TE] **teffects intro**.

The potential-outcome model

Next we specify a potential-outcome model that serves as a touchstone for the rest of our discussion. The model described here generates data in which y_i is the observed outcome variable, t_i is the treatment variable, $\mathbf{x}_i$ is a vector of covariates that affect the outcome, and $\mathbf{w}_i$ is a vector of covariates that affect the treatment assignment. $\mathbf{x}_i$ and $\mathbf{w}_i$ may have elements in common.

This potential-outcome model specifies that the observed outcome variable y is y_0 when $t = 0$ and that y is y_1 when $t = 1$. Algebraically, we say that

$$y = (1 - t)y_0 + ty_1$$

The functional forms for y_0 and y_1 are

$$y_0 = \mathbf{x}'\boldsymbol{\beta}_0 + \epsilon_0 \tag{1}$$
$$y_1 = \mathbf{x}'\boldsymbol{\beta}_1 + \epsilon_1 \tag{2}$$

where $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ are coefficients to be estimated, and ϵ_0 and ϵ_1 are error terms that are not related to $\mathbf{x}$ or $\mathbf{w}$. This potential-outcome model separates each potential outcome into a predictable component, $\mathbf{x}\boldsymbol{\beta}_t$, and an unobservable error term, ϵ_t.

The treatment assignment process is

$$t = \begin{cases} 1 \text{ if } \mathbf{w}'\boldsymbol{\gamma} + \eta > 0 \\ 0 \text{ otherwise} \end{cases} \tag{3}$$

where $\boldsymbol{\gamma}$ is a coefficient vector, and η is an unobservable error term that is not related to either $\mathbf{x}$ or $\mathbf{w}$. The treatment assignment process is also separated into a predictable component, $\mathbf{w}'\boldsymbol{\gamma}$, and an unobservable error term, η.

We emphasize six points about this model:

1. The observed data from this model contain y_i, t_i, $\mathbf{w}_i$, and $\mathbf{x}_i$. The data do not reveal both y_{0i} and y_{1i} for any given i.
2. The model for t determines how the data on y_0 and y_1 are missing.
3. The model separates the potential outcomes and treatment assignment into observable and unobservable components.
4. Whether η is independent of the vector (ϵ_0, ϵ_1) is essential in specifying the set of available estimators.
5. The coefficient vectors $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$, and $\boldsymbol{\gamma}$ are auxiliary parameters. We use estimates of these coefficient vectors to estimate the ATE, the POMs, and the ATET.
6. For notational simplicity, we represented y_0 and y_1 as linear functions. In practice, we can use other functional forms.

In specifying this potential-outcome model, we explicitly showed the functional forms for the potential outcomes and the treatment assignment process. To ease subsequent discussions, we refer to the set of functional forms for the potential outcomes as the "outcome model", and we refer to the treatment assignment process as the "treatment model".

Assumptions needed for estimation

As with any type of estimator, we must make some assumptions to use treatment-effects estimators. The particular assumptions we need for each estimator implemented by `teffects` and for each effect parameter vary, but some version of each of the following is required.

CI The conditional-independence CI assumption restricts the dependence between the treatment model and the potential outcomes.

Overlap The overlap assumption ensures that each individual could receive any treatment level.

i.i.d. The independent and identically distributed (i.i.d.) sampling assumption ensures that the potential outcomes and the treatment status of each individual are unrelated to the potential outcomes and treatment statuses of all other individuals in the population.

We now discuss each assumption in detail.

The CI assumption

After conditioning on covariates, when no unobservable variable affects both treatment assignment and the potential outcomes, the potential outcomes are conditionally independent of the treatment. In epidemiological jargon, there are no unmeasured confounders. In econometric jargon, we have selection on observables. If we observe enough covariates, the potential outcomes may indeed be conditionally independent of the treatment.

Intuitively, the CI assumption says that only the covariates $\mathbf{x}$ affect both the treatment and the potential outcomes. Any other factors that affect the treatment must be independent of the potential outcomes, and any other factors that affect the potential outcomes must be independent of the treatment. Formally, the CI assumption states that, conditional on covariates $\mathbf{x}$, the treatment t is independent of the vector of potential outcomes $(y_0, y_1)'$.

The CI assumption allows us to estimate the effects by regression adjustment (RA) methods, inverse-probability-weighting (IPW) methods, methods that combine RA and IPW concepts, and matching methods. The data only reveal information about $E(y_0|\mathbf{x}, \mathbf{w}, t = 0)$ and $E(y_1|\mathbf{x}, \mathbf{w}, t = 1)$, but we are interested in an average of $E(y_0|\mathbf{x}, \mathbf{w})$ and $E(y_1|\mathbf{x}, \mathbf{w})$, where $\mathbf{x}$ represents the outcome covariates and $\mathbf{w}$ the treatment-assignment covariates. The CI assumption allows us to estimate $E(y_0|\mathbf{x}, \mathbf{w})$ and $E(y_1|\mathbf{x}, \mathbf{w})$ directly from the observations for which $E(y_0|\mathbf{x}, \mathbf{w}, t = 0)$ and $E(y_1|\mathbf{x}, \mathbf{w}, t = 1)$, respectively.

For our potential-outcome model presented in (1) through (3), the CI assumption can be viewed as a set of restrictions on the covariance matrix of the error terms. Suppose that the vector of unobservables $(\epsilon_0, \epsilon_1, \eta)$ is normally distributed

$$\begin{pmatrix} \epsilon_0 \\ \epsilon_1 \\ \eta \end{pmatrix} \sim N\left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \rho_{01}\sigma_0\sigma_1 & \rho_{\eta 0}\sigma_0 \\ \rho_{01}\sigma_0\sigma_1 & \sigma_1^2 & \rho_{\eta 1}\sigma_1 \\ \rho_{\eta 0}\sigma_0 & \rho_{\eta 1}\sigma_1 & 1 \end{pmatrix} \right\} \tag{4}$$

where σ_0 is the standard deviation of ϵ_0, ρ_{01} is the correlation between ϵ_0 and ϵ_1, σ_1 is the standard deviation of ϵ_1, $\rho_{\eta 0}$ is the correlation between ϵ_η and ϵ_0, and $\rho_{\eta 1}$ is the correlation between ϵ_η and ϵ_1. As is standard in the normally distributed latent-variable specification of a binary-dependent variable, we normalize the variance of ϵ_η to 1.

CI specifies that $\rho_{\eta 0} = \rho_{\eta 1} = 0$ so that we can write (4) as

$$\begin{pmatrix} \epsilon_0 \\ \epsilon_1 \\ \eta \end{pmatrix} \sim N\left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \rho_{01}\sigma_0\sigma_1 & 0 \\ \rho_{01}\sigma_0\sigma_1 & \sigma_1^2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right\}$$

Writing the covariance matrix this way makes clear what we mean by conditional independence: unobserved shocks that affect whether a subject is treated do not have any effect on the potential outcomes, and unobserved shocks that affect a potential outcome do not affect treatment.

The command `teffects` implements estimators that require the CI assumption. See [TE] **etregress** and [TE] **etpoisson** for commands that handle two cases in which the CI assumption is replaced by precise specifications of the joint dependence among the unobservables. Brown and Mergoupis (2011) discuss the `itreatreg` command that extends [TE] **etregress**.

The CI assumption is also known as unconfoundedness and selection-on-observables in the literature. See Rosenbaum and Rubin (1983); Heckman (1997); Heckman and Navarro-Lozano (2004); Cameron and Trivedi (2005, sec. 25.2.1); Tsiatis (2006, sec. 13.3); Angrist and Pischke (2009, chap. 3); Imbens and Wooldridge (2009); and Wooldridge (2010, sec. 21.3). Some discussions with Stata commands can be found in Becker and Caliendo (2007), Nichols (2007), and Daniel, De Stavola, and Cousens (2011).

❑ Technical note

In fact, full CI is stronger than what we need to estimate the ATE, the ATET, or the POMs. For the estimators implemented in `teffects`, we only need a conditional mean independence (CMI) assumption. Intuitively, the CMI assumption says that after accounting for the covariates $\mathbf{x}_i$, the treatment does not affect the conditional mean of each potential outcome. Formally, the CMI requires that $E(y_0|\mathbf{x}, t) = E(y_0|\mathbf{x})$ and that $E(y_1|\mathbf{x}, t) = E(y_1|\mathbf{x})$. The CMI assumption allows the conditional variance to depend on the treatment, while the CI assumption does not.

The CI assumption implies the CMI assumption, but not vice versa.

See Wooldridge (2010, sec. 21.2 and 21.3) for an excellent introduction to this topic, and see Cattaneo, Drukker, and Holland (2013) for some discussion of the multiple treatment case.

❑

The overlap assumption

The overlap assumption requires that each individual have a positive probability of receiving each treatment level. Formally, the overlap assumption requires that for each possible $\mathbf{x}$ in the population and each treatment level $\widetilde{t}$, $0 < \Pr(t = \widetilde{t}|\mathbf{x}) < 1$. Rosenbaum and Rubin (1983) call the combination of the CI and overlap assumptions strong ignorability; see also Abadie and Imbens (2006, 237–238) and Imbens and Wooldridge (2009).

The i.i.d. assumption

The third of the three assumptions listed above is the i.i.d. assumption; it is the standard assumption of having an i.i.d. sample from the population. In potential-outcome models, i.i.d. sampling implies that the potential outcomes and treatment status of each individual are unrelated to the potential outcomes and treatment statuses of all the other individuals in the population. I.i.d. sampling rules out interactions among the individuals. For instance, models of vaccinations in epidemiology and general

equilibrium effects in economics violate the independence assumption. This third assumption is a part of what is known as the stable unit treatment value assumption (SUTVA); see Wooldridge (2010, 905) and Imbens and Wooldridge (2009).

Comparing the ATE and ATET

When comparing the ATE and the ATET, two points should be mentioned.

First, the assumptions required to estimate the ATET are less restrictive than the assumptions required to estimate the ATE. Estimating the ATET requires a weaker form of the CI assumption and a weaker version of the overlap assumption.

To estimate the ATE under CI, we require that the unobservables in the treatment model be conditionally independent of the unobservables in both potential outcomes. In contrast, we can estimate the ATET under CI when the unobservables in the treatment model are conditionally independent of just the control-level potential outcome; see Wooldridge (2010, 906–912).

Although the ATE version of overlap requires that all covariate patterns have a positive probability of being allocated to each treatment state, we can estimate the ATET when only the covariate patterns for which someone is treated have a positive probability of being allocated to each treatment state. This weaker overlap assumption can be important in some studies. For example, Heckman (1997) discusses how the ATET makes sense in job-training programs for which many types of individuals have zero chance of signing up. See also Wooldridge (2010, 911–913).

Second, the ATET reduces to the ATE when the mean of the covariates among the treated is the same as the mean of the covariates in the population and when the average contribution from the unobservables for the participants is zero.

Overview of treatment-effect estimators

We can classify the estimators implemented by `teffects` into five categories: 1) estimators based on a model for the outcome variable; 2) estimators based on a model for treatment assignment; 3) estimators based on models for both the outcome variable and the treatment assignment; 4) estimators that match on covariates; and 5) estimators that match on predicted probabilities of treatment. Within each category of estimator, there is a variety of choices about the functional forms for the models.

Because there are several categories of estimators, the user must decide whether to model the outcomes, the probability of treatment, or both. Ironically, some of the estimators that use both models only require that one of the two be correctly specified to consistently estimate the effects of interest, a property known as the double-robust property.

With the exception of using a matching estimator with a single continuous covariate, some choice of functional forms is required. There are two aspects one must consider when choosing the functional form for the outcome or treatment assignment. First, one must select the functional form for the conditional mean or conditional probability; depending on the variable being modeled, a linear, a binary choice, or an exponential model may be appropriate. Second, one must determine the appropriate polynomials of the covariates to include in the model. `teffects` offers a wide variety of options to specify different functional form choices for the conditional mean and conditional probability models. The factor variable notation in Stata allows us to easily specify the desired polynomial in the covariates.

We now provide some intuition behind each type of estimator.

RA estimators

RA estimators use means of predicted outcomes for each treatment level to estimate each POM. ATEs and ATETs are differences in estimated POMs.

The CI assumption implies that we can estimate $E(y_0|\mathbf{x})$ and $E(y_1|\mathbf{x})$ directly from the observations for which $t = 0$ and $t = 1$, respectively. Regression adjustment fits separate regressions for each treatment level and uses averages of the predicted outcomes over all the data to estimate the POMs. The estimated ATEs are differences in the estimated POMs. The estimated ATETs are averages of the predicted outcomes over the treated observations.

RA is a venerable estimator. See Lane and Nelder (1982); Cameron and Trivedi (2005, chap. 25); Wooldridge (2010, chap. 21); and Vittinghoff, Glidden, Shiboski, and McCulloch (2012, chap. 9). The usefulness of RA has been periodically questioned in the literature because it relies on specifying functional forms for the conditional means and because it requires having sufficient observations of each covariate pattern in each treatment level; see Rubin (1973) for an early salvo. Our experience is that RA is an exceptionally useful base-case estimator. We describe its relative advantages and disadvantages in the course of covering other estimators.

IPW estimators

IPW estimators use weighted averages of the observed outcome variable to estimate means of the potential outcomes. The weights account for the missing data inherent in the potential-outcome framework. Each weight is the inverse of the estimated probability that an individual receives a treatment level. Outcomes of individuals who receive a likely treatment get a weight close to one. Outcomes of individuals who receive an unlikely treatment get a weight larger than one, potentially much larger.

IPW estimators model the probability of treatment without any assumptions about the functional form for the outcome model. In contrast, RA estimators model the outcome without any assumptions about the functional form for the probability of treatment model.

IPW estimators become extremely unstable as the overlap assumption gets close to being violated. When the overlap assumption is nearly violated, some of the inverse-probability weights become very large, IPW estimators produce erratic estimates, and the large-sample distribution provides a poor approximation to the finite-sample distribution of IPW estimators. This instability occurs even though the functional form for the treatment model is correctly specified.

In contrast, when the overlap assumption is nearly violated, there are very few observations in a treatment level for some covariate patterns, so RA estimators use the model to predict in regions in which there are very little data. If the model is well specified and there are "enough" observations, an RA estimator will not become unstable as quickly as an IPW estimator, and the large-sample distribution of the RA estimator still provides a good approximation to the finite-sample distribution. However, in real situations in which "all models are approximate", relying on a correctly specified outcome model with little data is extremely risky.

IPW estimators are a general approach to missing-data problems that obey some CI assumptions. While IPW is an old idea in statistics that dates back to Horvitz and Thompson (1952), biostatisticians and econometricians have been actively working on extending it to handle modern problems and estimation methods. See Robins and Rotnitzky (1995); Robins, Rotnitzky, and Zhao (1994, 1995); and Wooldridge (2002, 2007). IPW has been used extensively in the modern treatment-effect estimation literature. See Imbens (2000); Hirano, Imbens, and Ridder (2003); Tan (2010); Wooldridge (2010, chap. 19); van der Laan and Robins (2003); and Tsiatis (2006, chap. 6).

To see the intuition behind IPW, consider a study with observed outcome variable y, treatment variable $t \in \{0, 1\}$, and potential outcomes y_0 and y_1. As part of this process, we need to estimate

the POM for treatment $t = 1$, $E(y_1)$. Using the observed data, $y_i t_i$ is y_{1i} when $t = 1$, but y_{1i} is unobserved when $t = 0$. An IPW estimator for $E(y_1)$ is $1/N \sum_{i=1}^{N} y_i t_i / p(\mathbf{x}_i)$, where $p(\mathbf{x}_i)$ is the probability that $t_i = 1$ and is a function of the covariates $\mathbf{x}_i$. If y_{1i} were always observed, the weights would all equal 1. This IPW estimator places a larger weight on those observations for which y_{1i} is observed even though its observation was not likely.

AIPW estimators

Instead of modeling either the outcome, like RA, or the treatment probability, like IPW, augmented inverse-probability-weighted (AIPW) estimators model both the outcome and the treatment probability. A surprising fact is that only one of the two models must be correctly specified to consistently estimate the treatment effects, a property of the AIPW estimators known as being "doubly robust". Given that two models instead of one are used, it is less surprising that the AIPW estimators can be more efficient than either the RA or the IPW estimators, though deriving this result is rather technical and relies on the theory of semiparametric estimators.

Intuitively, the AIPW estimator is an IPW that includes an augmentation term that corrects the estimator when the treatment model is misspecified. When the treatment is correctly specified, the augmentation term vanishes as the sample size becomes large. Like the IPW, the AIPW does not perform well when the predicted treatment probabilities are too close to zero or one.

AIPW estimators emerge naturally from a technique of producing more efficient estimators from estimators that have a few main parameters of interest and some auxiliary, or nuisance, parameters used in estimating the few main parameters. This method constructs efficient estimating equations for the main parameters that are orthogonal to the auxiliary parameters. The estimators produced by this method are known as efficient-influence function (EIF) estimators.

To gain some intuition, consider finding an EIF estimator from an IPW estimator for two POMs. Note that we only care about the two POM parameters and not about the many auxiliary parameters used to estimate the treatment probabilities. EIF estimators project the equations that yield the POM parameters onto the equations that yield the auxiliary treatment-model parameters and then use the residuals from this projection to estimate the POM parameters.

We refer to these estimators as "AIPW estimators" instead of "EIF estimators" because the former is commonly used in the biostatistical literature for some of the binary-treatment estimators and because the term "augmented inverse-probability-weighted" tells more about how these estimators relate to the other implemented estimators; see Tsiatis (2006) and Tan (2010). The estimators implemented in `teffects aipw` with the `wnls` option are based on those of Rubin and van der Laan (2008), which did well in simulations reported by Tan (2010), and denoted as $\widetilde{\alpha}_{RV}(\widehat{\pi})$ in Tan (2010, 663).

When either the outcome model or the treatment model is well specified, the AIPW estimators implemented in `teffects aipw` are more robust than either the RA or the IPW estimators because the AIPW estimators are doubly robust but the RA and IPW estimators are not. When both the outcome and the treatment model are misspecified, which estimator is more robust is a matter of debate in the literature; see Kang and Schafer (2007) and Robins et al. (2007) for some debate, and see Tan (2010) for a more recent discussion.

To the best of our knowledge, there is no general solution to the question of which estimator performs best when both the outcome and the treatment models are misspecified. We suspect that the answer depends on the true models, the implemented specifications, and the polynomials in the covariates used. To help users through this process, the estimators implemented in `teffects` offer many functional forms to approximate either the outcome process or the treatment process. In addition, Stata's factor-variable notation makes it easy to include polynomials in the covariates. Both of these approximation methods rely on having enough data. `teffects` also makes it easy to compare the results produced by different estimators.

The literature on these methods is vast and growing. For double-robust results and explanations, see Robins and Rotnitzky (1995); Robins, Rotnitzky, and Zhao (1995); van der Laan and Robins (2003, chap. 6); Bang and Robins (2005); Tsiatis (2006, chap. 13); Wooldridge (2007; 2010, sec. 21.3.4); and Tan (2010).

IPWRA estimators

Like AIPW estimators, inverse-probability-weighted regression-adjustment (IPWRA) estimators combine models for the outcome and treatment status; also like AIPW estimators, IPWRA estimators are doubly robust. IPWRA estimators emerge naturally from a robust approach to missing-data methods. IPWRA estimators use the inverse of the estimated treatment-probability weights to estimate missing-data-corrected regression coefficients that are subsequently used to compute the POMs.

As far as we know, there is no literature that compares the relative efficiency of AIPW estimators, which emerge from a general approach to creating efficient estimators, and the IPWRA estimators, which emerge from a robust-correction approach to missing-data analysis.

The IPWRA estimators are also know as "Wooldridge's double-robust" estimators because they were derived in Wooldridge (2007) and discussed at length in Wooldridge (2010, section 21.3.4).

Nearest-neighbor matching estimators

Matching estimators use an average of the outcomes of the nearest individuals to impute the missing potential outcome for each sampled individual. The difference between the observed outcome and the imputed potential outcome is an estimate of the individual-level treatment effect. These estimated individual-level treatment effects are averaged to estimate the ATE or the ATET.

`teffects nnmatch` determines the "nearest" by using a weighted function of the covariates for each observation. This type of matching is known as nearest-neighbor matching (NNM). `teffects psmatch` determines the "nearest" by using the estimated treatment probabilities, which are known as the propensity scores. This second type of matching is known as propensity-score matching (PSM).

NNM is nonparametric in that no explicit functional form for either the outcome model or the treatment model is specified. This flexibility comes at a price; the estimator needs more data to get to the true value than an estimator that imposes a functional form. More formally, the NNM estimator converges to the true value at a rate slower than the parametric rate, which is the square root of the sample size, when matching on more than one continuous covariate. `teffects nnmatch` uses bias correction to fix this problem. PSM provides an alternative to bias correction because it matches on a single continuous covariate, the estimated treatment probabilities.

Abadie and Imbens (2006, 2011) derived the rate of convergence of the NNM estimator and the bias-corrected NNM estimator and the large-sample distributions of the NNM and the bias-corrected NNM estimators. These articles provided the formal results that built on methods suggested in Rubin (1973, 1977).

`teffects nnmatch` is based on the results in Abadie and Imbens (2006, 2011) and a previous implementation in Abadie et al. (2004).

Propensity-score matching estimators

Instead of performing bias correction to handle the case of more than one continuous covariate, a common solution is to combine all the covariate information into estimated treatment probabilities, known as propensity scores, and use this single continuous covariate as the matching variable.

The term "propensity score" is widely used, but we continue to refer to it as the "treatment probability" to be consistent with the other estimators. We call the estimator that matches on the estimated treatment probabilities the "propensity-score matching (PSM) estimator" because the latter term is ubiquitous.

In effect, the PSM estimator parameterizes the bias-correction term in the treatment probability model. One advantage of matching on the estimated treatment probabilities over the bias-correction method is that one can explore the fit of different treatment probability models using standard methods before performing the nonparametric matching. For example, one can select the treatment model by minimizing an information criterion under i.i.d. sampling. We know of no counterpart for selecting the proper order of the bias-correction term for the NNM estimator.

Matching on estimated treatment probability models has been very popular since Rosenbaum and Rubin (1983) showed that if adjusting for covariates $\mathbf{x}_i$ is sufficient to estimate the effects, then one can use the probability of treatment to perform the adjustment. Abadie and Imbens (2012) derived a method to estimate the standard errors of the estimator that matches on estimated treatment probabilities, and this method is implemented in `teffects psmatch`.

Choosing among estimators

There is no definitive way to select one of the estimators implemented in `teffects` over the others. We offer three observations about the tradeoffs among the estimators.

First, if the outcome model is correctly specified, the RA estimator will break down more slowly than the IPW, AIPW, IPWRA, or PSM estimators as the overlap assumption begins to fail. This result depends critically on the ability of the RA estimator to predict into regions in which there are little data.

Second, if the overlap assumption holds, the AIPW and IPWRA estimators have the double-robust property for some functional form combinations. The double-robust property says that if either the outcome model or the treatment model is correctly specified, we can consistently estimate the effects. The properties of double-robust estimators when both models are misspecified are not known, although there is some discussion in the literature about the properties of the AIPW estimators; see Kang and Schafer (2007), Robins et al. (2007), and Tan (2010).

Third, all the estimators require the same assumptions, so if each is correctly specified, they should all produce similar results. Of course, just because they produce similar results does not mean that they are correctly specified; it is possible that they are just behaving similarly in response to some underlying problem.

Model choice

`teffects` offers a broad selection of functional form combinations so that you can choose a combination that fits your data. Picking a functional form that respects the values of the observed outcomes is usually best. Select `linear` for continuous outcomes over the real line; `logit`, `probit`, or `hetprobit` for binary outcomes; and `poisson` for counts or nonnegative outcomes.

For binary treatments, you can select among `logit`, `probit`, or `hetprobit` models. For multivalued treatments, only the multinomial logit is available to model the treatment probabilities.

Selecting a functional form of a given set of covariates is a famously difficult problem in statistics. In the treatment-effects context, Cattaneo, Drukker, and Holland (2013) found that model selection by minimizing an information criterion worked well. Cattaneo, Drukker, and Holland (2013) discuss a method and a user-written command to facilitate the process.

Acknowledgments

We thank Matias Cattaneo, Guido Imbens, Barbara Sianesi, and Jeffrey Wooldridge for helpful conversations and advice about the literature and methods in this area.

We thank Matias Cattaneo for his collaboration on Cattaneo, Drukker, and Holland (2013) and for many conversations about Cattaneo (2010), a part of which is implemented in `teffects ipw` and `teffects aipw`.

We thank Ashley Holland for her collaboration on Cattaneo, Drukker, and Holland (2013).

We thank Alberto Abadie, Jane Herr, and Guido Imbens for collaboration on Abadie et al. (2004), which is implemented in `teffects nnmatch`.

We thank Matias Cattaneo for sharing an extract of the dataset that he used in Cattaneo (2010) and Barbara Sianesi for sharing an extract of the dataset that she used in Blundell, Dearden, and Sianesi (2005).

We thank Alberto Abadie, Guido Imbens, and László Sándor for sharing their implementation of the standard errors derived in Abadie and Imbens (2012), which are implemented in `teffects psmatch`.

We thank Edwin Leuven and Barbara Sianesi for their inspirational command `psmatch2`, which computes several of the estimators in `teffects` and a few more that we have not implemented.

References

Abadie, A., D. M. Drukker, J. L. Herr, and G. W. Imbens. 2004. Implementing matching estimators for average treatment effects in Stata. *Stata Journal* 4: 290–311.

Abadie, A., and G. W. Imbens. 2006. Large sample properties of matching estimators for average treatment effects. *Econometrica* 74: 235–267.

——. 2011. Bias-corrected matching estimators for average treatment effects. *Journal of Business and Economic Statistics* 29: 1–11.

——. 2012. Matching on the estimated propensity score. Harvard University and National Bureau of Economic Research. http://www.hks.harvard.edu/fs/aabadie/pscore.pdf.

Angrist, J. D., and J.-S. Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Bang, H., and J. M. Robins. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics* 61: 962–972.

Becker, S. O., and M. Caliendo. 2007. Sensitivity analysis for average treatment effects. *Stata Journal* 7: 71–83.

Blundell, R., L. Dearden, and B. Sianesi. 2005. Evaluating the effect of education on earnings: Models, methods and results from the National Child Development Survey. *Journal of the Royal Statistical Society, Series A* 168: 473–512.

Brown, G. K., and T. Mergoupis. 2011. Treatment interactions with nonexperimental data in Stata. *Stata Journal* 11: 545–555.

Cameron, A. C., and P. K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

Cattaneo, M. D. 2010. Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics* 155: 138–154.

Cattaneo, M. D., D. M. Drukker, and A. D. Holland. 2013. Estimation of multivalued treatment effects under conditional independence. *Stata Journal* 13: 407–450.

Daniel, R. M., B. L. De Stavola, and S. N. Cousens. 2011. gformula: Estimating causal effects in the presence of time-varying confounding or mediation using the g-computation formula. *Stata Journal* 11: 479–517.

Heckman, J. 1997. Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *Journal of Human Resources* 32: 441–462.

Heckman, J., and S. Navarro-Lozano. 2004. Using matching, instrumental variables, and control functions to estimate economic choice models. *Review of Economics and Statistics* 86: 30–57.

Hirano, K., G. W. Imbens, and G. Ridder. 2003. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71: 1161–1189.

Holland, P. W. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 81: 945–960.

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: 663–685.

Imbens, G. W. 2000. The role of the propensity score in estimating dose-response functions. *Biometrika* 87: 706–710.

——. 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics* 86: 4–29.

Imbens, G. W., and J. M. Wooldridge. 2009. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* 47: 5–86.

Kang, J. D. Y., and J. L. Schafer. 2007. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* 22: 523–539.

Lane, P. W., and J. A. Nelder. 1982. Analysis of covariance and standardization as instances of prediction. *Biometrics* 38: 613–621.

Nichols, A. 2007. Causal inference with observational data. *Stata Journal* 7: 507–541.

Robins, J. M. 1986. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modeling* 7: 1393–1512.

Robins, J. M., and A. Rotnitzky. 1995. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* 90: 122–129.

Robins, J. M., A. Rotnitzky, and L. P. Zhao. 1994. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 89: 846–866.

——. 1995. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* 90: 106–121.

Robins, J. M., M. Sued, Q. Lei-Gomez, and A. Rotnitzky. 2007. Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable. *Statistical Science* 22: 544–559.

Rosenbaum, P. R., and D. B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70: 41–55.

Rubin, D. B. 1973. Matching to remove bias in observational studies. *Biometrics* 29: 159–183.

——. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66: 688–701.

——. 1977. Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics* 2: 1–26.

Rubin, D. B., and M. J. van der Laan. 2008. Empirical efficiency maximization: Improved locally efficient covariate adjustment in randomized experiments and survival analysis. *International Journal of Biostatistics* 4: 1–40.

Tan, Z. 2010. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika* 97: 661–682.

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

van der Laan, M. J., and J. M. Robins. 2003. *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer.

Vittinghoff, E., D. V. Glidden, S. C. Shiboski, and C. E. McCulloch. 2012. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. 2nd ed. New York: Springer.

Wooldridge, J. M. 2002. Inverse probability weighted M-estimators for sample selection, attrition, and stratification. *Portuguese Economic Journal* 1: 117–139.

——. 2007. Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics* 141: 1281–1301.

——. 2010. *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. Cambridge, MA: MIT Press.

Also see

[TE] **teffects** — Treatment-effects estimation for observational data

[TE] **teffects intro** — Introduction to treatment effects for observational data