

---

*2013 UK Stata Users Group meeting*

Cass Business School

September 12–13, 2013

---

**TREATREW: a user-written Stata routine for estimating  
Average Treatment Effects by *reweighting* on propensity score**

Giovanni Cerulli

*Ceris-CNR*

*National Research Council of Italy*

*Institute for Economic Research on Firms and Growth*

E-mail: [g.cerulli@ceris.cnr.it](mailto:g.cerulli@ceris.cnr.it)

## BACKGROUND

**TREATREW** is a Stata routine for estimating Average Treatment Effects by *reweighting on propensity score*:

- provides consistent estimation of Average Treatment Effects under the hypothesis of **selection on observables**, conditional on a pre-specified set of observable exogenous variables  $\mathbf{x}$
- estimates the **Average Treatment Effect (ATE)**, the one on **Treated (ATET)** and the one on **Non-Treated (ATENT)**, as well as the same ones conditional on the observable factors  $\mathbf{x}$  (i.e.,  $ATE(\mathbf{x})$ ,  $ATET(\mathbf{x})$  and  $ATENT(\mathbf{x})$ )
- implements the reweighting on propensity score estimator as proposed by **Rosenbaum and Rubin (1983)** in their seminal article

## MAIN CONTRIBUTION

- **TREATREW** provides **analytical standard errors** for the estimation of ATE, ATET and ATENT using the approximation suggested by Wooldridge (2010, p. 920-930)
- **Bootstrapped** standard errors can also be easily computed

## REWEIGHTING ON PROPENSITY-SCORE INVERSE PROBABILITY

Assumptions about the **Data Generating Process** (DGP):

- i.  $y_1 = g_1(\mathbf{x}) + \varepsilon_1$  ,  $E(\varepsilon_1) = 0$
- ii.  $y_0 = g_0(\mathbf{x}) + \varepsilon_0$  ,  $E(\varepsilon_0) = 0$
- iii.  $y = w y_1 + y_0(1-w)$
- iv. Conditional Mean Independence (CMI) holds, so that:  
$$E(y_1 | w, \mathbf{x}) = E(y_1 | \mathbf{x}) \quad \text{and} \quad E(y_0 | w, \mathbf{x}) = E(y_0 | \mathbf{x})$$
- v.  $\mathbf{x}$  exogenous

### *LEGEND*

- $y_1$  and  $y_0$  is the unit's outcome when it is treated and untreated respectively;
- $g_1(\mathbf{x})$  and  $g_0(\mathbf{x})$  is the unit's reaction function to the confounder  $\mathbf{x}$  when the unit is treated and untreated respectively;
- $w$  is the treatment binary indicator taking value 1 for treated and 0 for untreated units;
- $\varepsilon_0$  and  $\varepsilon_1$  are two error terms with unconditional zero mean;
- $\mathbf{x}$  is a set of observable, exogenous confounding variables driving the non-random assignment into treatment.

Under previous DGP, we get that:

$$\text{ATE} = \text{E} \left\{ \frac{[w - p(\mathbf{x})]y}{p(\mathbf{x})[1 - p(\mathbf{x})]} \right\}$$

$$\text{ATET} = \text{E} \left\{ \frac{[w - p(\mathbf{x})]y}{p(w = 1)[1 - p(\mathbf{x})]} \right\}$$

$$\text{ATENT} = \text{E} \left\{ \frac{[w - p(\mathbf{x})]y}{p(w = 0)p(\mathbf{x})} \right\}$$

NOTE: This is equivalent to the **Horvitz–Thompson estimator**, due to Daniel G. Horvitz and Donovan J. Thompson in 1952. In sampling theory it is a method for estimating the *total* and *mean* of a super-population in a stratified sample. **Inverse Probability Weighting (IPW)** is generally applied to account for “*different proportions of observations within strata in a target population*”.

## Sample estimation and standard errors for ATE, ATET and ATENT

Assuming that the **propensity score is correctly specified**, we can estimate previous parameters simply by using the “*sample equivalent*” of the population parameters, that is:

$$\widehat{\text{ATE}} = \frac{1}{N} \sum_{i=1}^N \frac{[w_i - \hat{p}(\mathbf{x}_i)]y_i}{\hat{p}(\mathbf{x}_i)[1 - \hat{p}(\mathbf{x}_i)]}$$

$$\widehat{\text{ATET}} = \frac{1}{N} \sum_{i=1}^N \frac{[w_i - \hat{p}(\mathbf{x}_i)]y_i}{\hat{p}(w=1)[1 - \hat{p}_i(\mathbf{x})]}$$

$$\widehat{\text{ATENT}} = \frac{1}{N} \sum_{i=1}^N \frac{[w_i - \hat{p}_i(\mathbf{x}_i)]y_i}{\hat{p}(w=0)\hat{p}(\mathbf{x}_i)}$$

Estimation follows in **two steps**:

- (i) *estimate the propensity score  $p(\mathbf{x}_i)$ , thus getting  $\hat{p}(\mathbf{x}_i)$ ;*
- (ii) *substitute  $\hat{p}(\mathbf{x}_i)$  into previous formulas to get parameters.*

Observe that **consistency** is guaranteed by the fact that these estimators are **M-estimators**.

## How to get analytical standard errors for previous estimators?

We can exploit some results from the case in which the first step is a ML-estimation and the second step is a M-estimation:

- First step: ML on Logit (or Probit)
- Second step: standard M-estimator.

Wooldridge (2007; 2010, p. 922-924) has proposed a straightforward procedure to get *analytical standard errors*, provided that the propensity score is correctly specified.

## Standard error estimation for ATE

**First:** define the estimated ML-score of the first step (Probit or Logit). It is, by definition, equal to:

$$\hat{\mathbf{d}}_i = \hat{\mathbf{d}}(w_i, \mathbf{x}_i, \hat{\boldsymbol{\gamma}}) = \frac{[\nabla_{\boldsymbol{\gamma}} \hat{p}(\mathbf{x}_i, \hat{\boldsymbol{\gamma}})]' \cdot [w_i - \hat{p}(\mathbf{x}_i, \hat{\boldsymbol{\gamma}})]}{\hat{p}(\mathbf{x}_i, \hat{\boldsymbol{\gamma}})[1 - \hat{p}(\mathbf{x}_i, \hat{\boldsymbol{\gamma}})]}$$

Observe that  $\mathbf{d}$  is a row-vector of the  $R-1$  parameters  $\boldsymbol{\gamma}$  and represents the gradient of the function  $p(\mathbf{x}, \boldsymbol{\gamma})$ .

**Second:** define the generic estimated summand of ATE as:

$$\hat{k}_i = \frac{[w_i - \hat{p}(\mathbf{x}_i)]y_i}{\hat{p}(\mathbf{x}_i)[1 - \hat{p}(\mathbf{x}_i)]}$$

**Third:** calculate OLS *residuals* from this regression:

$$\hat{k}_i \quad \text{on} \quad (1, \hat{\mathbf{d}}_i') \quad \text{with} \quad i = 1, \dots, N$$

and call them  $\hat{e}_i$  ( $i = 1, \dots, N$ ).

The *asymptotic standard error* for ATE is equal to:

$$\frac{\left[ \frac{1}{N} \sum_{i=1}^N \hat{e}_i^2 \right]^{1/2}}{\sqrt{N}}$$

and we can use it to test the significance of ATE.

Of course, **d** will have a different expression according to the probability model considered.

Here, we consider two cases:

- (i) **Logit**
- (ii) **Probit**



## Case 1: Logit

Suppose that the correct probability follows a *logistic distribution*. It means that:

$$p(\mathbf{x}_i, \gamma) = \frac{\exp(\mathbf{x}_i \gamma)}{1 + \exp(\mathbf{x}_i \gamma)} = \Lambda(\mathbf{x}_i \gamma)$$

Thus, by simple algebra, we get that:

$$\underbrace{\hat{\mathbf{d}}'_i}_{1 \times R} = \mathbf{x}_i (w_i - \hat{p}_i)$$

## Case 2: Probit

Suppose that the right probability follows a *Normal distribution*. It means that:

$$p(\mathbf{x}_i, \gamma) = \Phi(\mathbf{x}_i \gamma)$$

Thus, by simple algebra, we get that:

$$\hat{\mathbf{d}}'_i = \frac{\phi(\mathbf{x}_i, \hat{\gamma}) \mathbf{x}_i \cdot [w_i - \Phi(\mathbf{x}_i \gamma)]}{\Phi(\mathbf{x}_i \gamma) [1 - \Phi(\mathbf{x}_i \gamma)]}$$

where  $\Phi(\cdot)$  and  $\phi(\cdot)$  are the Normal cumulative distribution and density function respectively. Observe that one can add also functions of  $\mathbf{x}$  to estimate previous formulas. This reduces standard errors if these functions are partially correlated with  $k$ .

## The Stata routine `treatrew`: syntax and use

- `treatrew` estimates previous parameters with either *analytical* or *bootstrapped* standard errors.
- The syntax is rather simple and follows the typical Stata command syntax.
- The user has to declare:
  - (a) *outcome variable*, i.e. target-variable of the treatment (*outcome*)
  - (b) *binary treatment* variable (*treatment*)
  - (c) *set of confounding variables* (*varlist*)
  - (d) *options*.

Two options are important:

- option `model(modeltype)` sets the type of model, Probit or Logit, that has to be used in estimating the propensity score;
- option `graphic` and the related option `range(a b)` produce a chart where the distribution of  $ATE(\mathbf{x})$ ,  $ATET(\mathbf{x})$  and  $ATENT(\mathbf{x})$  are jointly plotted within the interval  $[a; b]$ .

The syntax of `treatrew` is:

```
treatrew outcome treatment [varlist] [if] [in] [weight],  
      model(modeltype) [GRange range(a b) conf(number) vce(robust)]
```

As e-class command, `treatrew` provides an `ereturn list` of objects (such as scalars and matrices) to be used in next elaborations. In particular, the values of **ATE**, **ATET** and **ATENT** are returned in the scalars `e(ate)`, `e(atet)` and `e(atent)` and they can be used to get *bootstrapped standard errors*.

Observe that - by default - `treatrew` provides *analytical* standard errors.

## Example: the impact of *education* on *fertility*

For a practical application of `treatreg`, we employ an instructional dataset called **FERTIL2.DTA** accompanying the manual “*Introductory Econometrics: A Modern Approach*” by Wooldridge (2000) collecting cross-sectional data on 4,361 women of childbearing age in Botswana. This dataset is freely downloadable at <http://fmwww.bc.edu/ec-p/data/wooldridge/FERTIL2.dta>. It contains 28 variables on various woman and family characteristics.

Using **FERTIL2.DTA**, we are interested in evaluating the impact of the variable “**educ7**” (taking value 1 if a woman has more than or exactly seven years of education, and 0 otherwise) on the number of family children (“**children**”). Several conditioning (or confounding) observable factors are included in the dataset, such as: the age of the woman (*age*), whether or not the family owns a TV (*tv*), whether or not the woman lives in a city (*urban*), and so forth.

We estimate ATE, ATET and ATENT (as well as:  $ATE(\mathbf{x})$ ,  $ATET(\mathbf{x})$  and  $ATENT(\mathbf{x})$ ) by “reweighting” using `treatrew`.

We compare reweighting results with other popular program evaluation methods, that is:

- (i) **Difference-In-Mean** (DIM), taken as benchmark
- (ii) **OLS random-coefficient model** with “heterogeneous reaction to confounders”, estimated through the user-written Stata routine `ivtreatreg` provided by Cerulli (2012)
- (iii) **One-to-one nearest neighbor Matching**, computed by the `psmatch2` Stata module provided by Leuven and Sianesi (2003)

NOTE: Matching estimators can be seen as specific reweighting procedures (Busso, DiNardo and McCrary, 2009).

# The `treatrew` help file

## Title

`treatrew` - Estimation of Average Treatment Effects by reweighting on propensity score

## Syntax

```
treatrew outcome treatment [varlist] [if] [in] [weight], model(modeltype) [GRaphic  
range(a b) conf(number) vce(robust)]
```

`fweights`, `iweights`, and `pweights` are allowed; see `weight`.

## Description

`treatrew` estimates Average Treatment Effects by reweighting on propensity score. Depending on the model specified, `treatrew` provides consistent estimation of Average Treatment Effects under the hypothesis of "selection on observables". Conditional on a pre-specified set of observable exogenous variables  $x$  - thought of as those driving the non-random assignment to treatment - `treatrew` estimates the Average Treatment Effect (ATE), the Average Treatment Effect on Treated (ATET) and the Average Treatment Effect on Non-Treated (ATENT), as well as the estimates of these parameters conditional on the observable factors  $x$  (i.e.,  $ATE(x)$ ,  $ATET(x)$  and  $ATENT(x)$ ). Parameters standard errors are provided either analytically (following Wooldridge, 2010, p. 920-930) and via bootstrapping. `treatrew` assumes that the propensity score specification is correct.

According to the syntax:

`outcome`: is the target variable over which measuring the impact of the treatment

`treatment`: is the binary treatment variable taking 1 for treated, and 0 for untreated units

`varlist`: is the set of pre-treatment (or observable confounding) variables

### Options

`model(modeltype)` specifies the model for estimating the propensity score, where `modeltype` must be one out of these two: "probit" or "logit". It is always required to specify one model.

`graphic` allows for a graphical representation of the density distributions of  $ATE(x)$ ,  $ATET(x)$  and  $ATENT(x)$  within their whole support.

`range(a b)` allows for a graphical representation of the density distributions of  $ATE(x)$ ,  $ATET(x)$  and  $ATENT(x)$  within the support  $[a;b]$  specified by the user. It has to be specified along with the `graphic` option.

`vce(robust)` allows for robust regression standard errors in the probit or logit estimates.

`conf(number)` sets the confidence level of probit or logit estimates equal to the specified number. The default is `number=95`.

<i>modeltype_options</i>	<i>description</i>
<code>probit</code>	The propensity score is estimated by a probit regression
<code>logit</code>	The propensity score is estimated by a logit regression

`treatrew` creates a number of variables:



`ATE_x` is an estimate of the idiosyncratic Average Treatment Effect.

`ATET_x` is an estimate of the idiosyncratic Average Treatment Effect on treated.

`ATENT_x` is an estimate of the idiosyncratic Average Treatment Effect on Non-Treated.

**treatrew** returns the following scalars:

`e(N)` is the total number of (used) observations.

`e(N1)` is the number of (used) treated units.

`e(N0)` is the number of (used) untreated units.

`e(ate)` is the value of the Average Treatment Effect.

`e(atet)` is the value of the Average Treatment Effect on Treated.

`e(atent)` is the value of the Average Treatment Effect on Non-treated.

Results in column (1) refer to the Difference-In-Mean (DIM) and are obtained by typing:

```
. reg children educ7
```

Results on column (2) refer to CF-OLS and are obtained by typing:

```
. ivtreatreg children educ7 age agesq evermarr urban electric tv , ///  
hetero(age agesq evermarr urban electric tv) model(cf-ols)
```

In the case of CF-OLS, standard errors for ATET and ATENT are obtained via bootstrap and can be obtained in STATA by typing:

```
. bootstrap atet=r(atet) atent=r(atent), rep(200): ///  
ivtreatreg children educ7 age agesq evermarr urban electric tv , ///  
hetero(age agesq evermarr urban electric tv) model(cf-ols)
```

Results set out in columns (3)-(6) refer to the reweighting estimator (REW). In column (3) and (4) standard errors are computed analytically, whereas in column (5) and (6) via bootstrap for the Logit and Probit model respectively. These results can be retrieved by typing sequentially:

```
. treatrew children educ7 age agesq evermarr urban electric tv , /// model(probit)
```

```
. treatrew children educ7 age agesq evermarr urban electric tv , /// model(logit)
```

```
. bootstrap e(ate) e(atet) e(atent) , reps(200): ///  
treatrew children educ7 age agesq evermarr urban electric tv , model(probit)
```

```
. bootstrap e(ate) e(atet) e(atent) , reps(200): ///  
treatrew children educ7 age agesq evermarr urban electric tv , model(logit)
```

Finally, column (7) presents an estimation of ATEs obtained by implementing a **One-to-one nearest-neighbor Matching** on propensity-score (MATCH). Here, the standard error for ATE is obtained analytically, whereas those for ATE and ATENT are computed by bootstrapping. Matching results can be obtained by typing:

```
. psmatch2 educ7 age agesq evermarr urban electric tv, ate out(children) com

. bootstrap r(ate) r(atu) : psmatch2 educ7 $xvars , ate out(children) com
```

where the option “com” restrict the sample to units with common support. In order to test the balancing property for such a Matching estimation, we provide a DIM on the propensity score *before* and *after* matching treated and untreated units, using the `psmatch2`'s post estimation command `pstest`:

```
. pstest _pscore
```

Variable	Sample	Mean		%reduct		t-test	
		Treated	Control	%bias	bias	t	p> t
_pscore	Unmatched	.65692	.42546	111.7		37.05	0.000
	Matched	.65692	.65688	0.0	100.0	0.01	0.994

This test suggests that, with regard to the propensity score, the Matching procedure implemented by `psmatch2` is balanced, so that we can sufficiently trust Matching results (indeed, the propensity score was unbalanced before Matching and it becomes balanced after Matching).

## Comparison of ATE, ATET and ATENT estimation among DIM, CF-OLS, REW and MATCH.

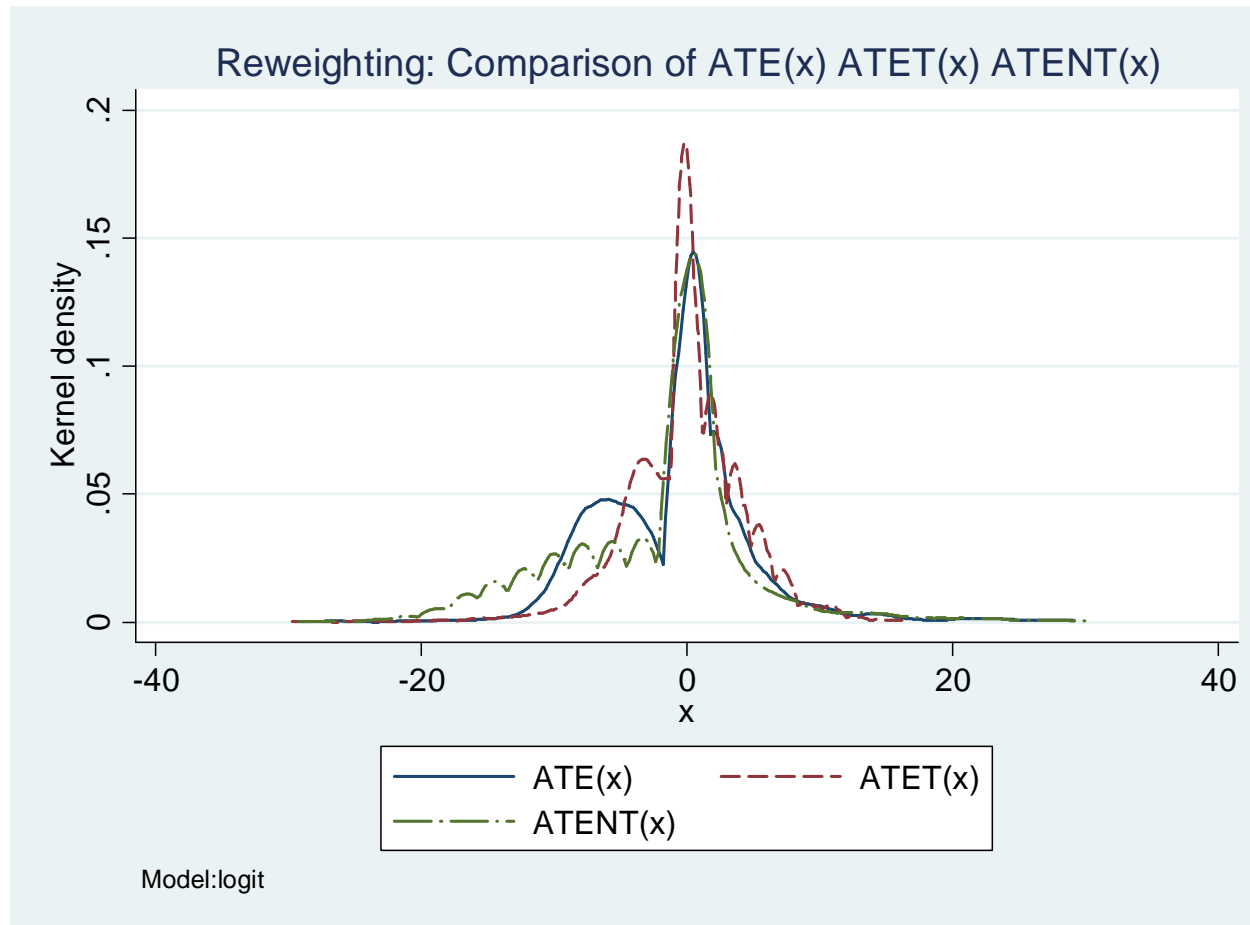
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	DIM	CF-OLS	REW (probit) Analytical Std. Err.	REW (logit) Analytical Std. Err.	REW (probit) Bootstrapped Std. Err.	REW (logit) Bootstrapped Std. Err.	MATCH <sup>(a)</sup>
ATE	-1.77 *** 0.062 -28.46	-0.374 *** 0.051 -7.35	-0.43 *** 0.068 -6.34	-0.415 *** 0.068 -6.09	-0.434 *** 0.070 -6.15	-0.415 *** 0.071 -5.87	-0.316 *** 0.080 -3.93
ATET		-0.255 *** 0.048 -5.37	-0.355 ** 0.15 -2.37	-0.345 *** 0.104 -3.33	-0.355 *** 0.065 -5.50	-0.345 *** 0.054 -6.45	-0.131 0.249 -0.52
ATENT		-0.523 *** 0.075 -7.00	-0.532 *** 0.19 -2.81	-0.503 ** 0.257 -1.96	-0.532 *** 0.115 -4.61	-0.503 *** 0.119 -4.21	-0.549 *** 0.135 -4.07

Note: b/se/t; DIM: Difference-in-Mean; CF-OLS: Control-function OLS; REW: Reweighting on propensity score; MATCH: One-to-one nearest-neighbor Matching on propensity-score. <sup>(a)</sup> Standard errors for ATE and ATENT are computed by bootstrapping. \*\*\* = 1%, \*\* = 5%, \* = 10% of significance.

## SOME COMMENTS ON RESULTS

- If confounding variables were not considered, as in **DIM**, the negative effect would appear dramatically higher.
- Results from **CF-OLS** and **REW** are fairly comparable: linearity is an acceptable approximation.
- By looking at the value of **ATET**, as obtained by **REW**: an educated woman in Botswana would have been – *ceteris paribus* – significantly more fertile if she had been less educated ==> “**education**” has a *negative* impact on “**fertility**”, leading to have around 0.5 children less.
- **REW** results using Wooldridge’s analytical standard errors in the case of Probit and Logit: very similar results.
- **REW** results when standard errors are obtained via bootstrap similar to analytical formulas. But bootstrap seems to *increase significance* both for ATET and ATENT, while ATE’s standard error is in line with the analytical one.

When `treatrew` is used with options `graphic` and `range(-30 30)`, we get kernel density for the distribution of  $ATE(\mathbf{x})$ ,  $ATET(\mathbf{x})$  and  $ATENT(\mathbf{x})$ .



## References

- Alberto, A.A. and Imbens, G.W. 2008. On the Failure of the Bootstrap for Matching Estimators. *Econometrica*, 76, 6, 1537–1557.
- Austin, N. 2007. Causal inference with observational data. *The Stata Journal*, 7, 507-541.
- Brunell, T.L. and DiNardo, J. 2004. A propensity score reweighting approach to estimating the partisan effects of full turnout in american presidential elections. *Political Analysis*, 12, 28-45.
- Busso, M., DiNardo, J., and McCrary, J. (2009). Finite sample properties of semiparametric estimators of average treatment effects. Forthcoming in the *Journal of Business and Economic Statistics*.
- Cerulli, G. 2012. IVTREATREG: a new STATA routine for estimating binary treatment models with heterogeneous response to treatment under observable and unobservable selection, *Working Paper CNR-Ceris*, 03/2012.
- Horvitz, D.G. and Thompson, D.J. 1952. A generalization of sampling without replacement from a finite universe source. *Journal of the American Statistical Association*, 47, 663- 685.
- Hirano, K., Imbens, G.W. and Ridder, G. 2003. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71, 1161-1189.

Li, Q., Racine, J.S. and Wooldridge, J.M. 2009. Efficient estimation of average treatment effects with mixed categorical and continuous data. *Journal of Business and Economic Statistics*, 27, 206-223.

Lunceford, J.K. and Davidian, M. 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 15, 2937-2960.

Leuven, E. and Sianesi, B. 2003. “PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing”.  
<http://ideas.repec.org/c/boc/bocode/s432001.html>.

Morgan, S.L. and Harding, D.J. 2006. Matching estimators of causal effects: prospects and pitfalls in theory and practice. *Sociological Methods and Research*, 35, 3-60.

Robins, J.M., Hernan, M.A. and Brumback, B.A. 2000. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11, 550-560.

Rosenbaum, P. and Rubin, D.B. 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70, 41-55.

Wooldridge, J.M. 2000. *Introductory Econometrics: A Modern Approach*. South-Western College Publishing.



Wooldridge, J.M. 2007. Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, 141, 1281–1301.

Wooldridge, J.M. 2010. *Econometric analysis of cross section and panel data*. 2nd Edition, Chapter 21. The MIT Press, Cambridge.