**DOCUMENTS DE TRAVAIL CEMOI / CEMOI WORKING PAPERS**

# A SAS macro to estimate Average Treatment Effects with Matching Estimators

**Nicolas Moreau**[1]

**http://cemoi.univ-reunion.fr/publications/**

**Centre d'Economie et de Management de l'Océan Indien**
**Université de La Réunion**

**October 2016**

Abstract

This paper presents a SAS macro to estimate the Average Treatment Effect (ATE) and the Average Treatment Effect for the Treated (ATET) with nearest-neighbor matching.

JEL : C210

[1] E-mail: nicolas.moreau@univ-reunion.fr

## Introduction

In this paper, we present the SAS macro *nn matching*. *nn matching* estimates nearest neighbor matching with replacement for the average treatment effect (ATE) and average treatment effect for the treated (ATET). In this respect, we draw heavily on Abadie et al. (2004) and Abadie and Imbens (2002, 2011). We refer the reader to these articles for a clear and comprehensive presentation of the matching estimator.

Following these authors, *nn_matching* provides the simple and bias-corrected matching estimators. Variance estimation is conducted by assuming homoscedasticity of the conditional variance or allowing for heteroscedasticity of the conditional variance. The source code is available at http://cemoi.univ-reunion.fr.

## Syntax of *nn_matching*

The syntax is %nn_matching(data=,y=,w=,x=,M=,scaling=,covarbias=);

where *data* specifies the data set, *y* the outcome variable, *w* the binary variable treatment indicator, *x* the list of covariates to be used in the matching, and *M* the number of matches to be made per observation. *M* could be any integer between 1 and the minimum of the number of treated units and controls in the sample.

If there are ties and if different matched pairs $(i,j)$ and $(i,l)$ lead to the same distance $d_{ij}= d_{il}$, then the number of matches per unit is greater than *M*.

*Scaling* specifies the metric for measuring the distance between two vectors of covariates. Following Abadie et al. (2004), the default is the diagonal matrix of the inverses of the sample variances of each covariate in *x* when scaling is not specified by the user. If *scaling*=1, the Mahalabonis metric is used; it is the inverse of the sample covariance matrix of the covariates. If *scaling*=2, the identity matrix is used instead. *Covarbias* specifies the list of covariates to be used to compute the bias-corrected matching estimator. The default list is *x* when *covarbias* is not specified by the user.

Note that all variables in *y*, *x*, *w* and *covarbias* must be numeric.

## Results presentation and output data files

ATE and ATET are automatically computed. For each estimated parameter, estimated standard errors that assume homoscedasticity of the conditional variance or allow for heteroscedasticity of the conditional variance are supplied.

The bias-corrected matching estimator developed in Abadie and Imbens (2002, 2011) is automatically computed if the number of "continuous" covariates in *x* is greater than 1. In *nn_matching*, all variables that are not binary are considered as continuous.

The first two output tables summarize the model specification and estimation options. The third table provides summary statistics, such as the number of treated units, the number of controls matched to treated units, and so on. The fourth and final table shows the main results. The "Estimate" column reports the estimated ATE and ATET. The next column shows the corresponding robust standard errors. "*z*" corresponds to the z-statistics to test whether ATE or ATET are 0; these are computed as the estimated parameters divided by their corresponding standard error. The "P-value" column reports the p-values for the z-statistics for a two-sided test.

The last two columns show the lower and upper bounds of the 95% confidence interval for the z-statistics.

To assess the validity of the balancing hypothesis, *nn matching* provides normalized differences (see Abadie and Imbens, 2011; Austin, 2009) so as box-plots and empirical distribution functions to compare the covariate distributions between treatment groups before and after matching. Theses plots are edited for continuous variables alone. For binary variables, contingency tables are provided.

Two temporary output data files are created, which need to be stored in a specific folder with a libname statement to become permanent.

*Outdata1* includes an internal identification number for observation *i* created by the program that is based on the original sort order and called _id_. *nbelements* specifies the number of matches for unit *i*. *count* is the number of times unit *i* is used as a match. *km_i* specifies the number of times unit *i* is used as a match for any observation *j* of the opposite treatment group weighted by the total number of matches for the given observation *j*. *Outdata1* also includes the outcome variable *y*, covariates *x*, and treatment group indicator *w*.

*Outdata2* includes the list of indices for the *M* closest matches for unit *i*. *id* is the internal identification number for observation *i* and *idM* the corresponding identification number of *i*'s closest matches in the opposite treatment group. For each *id* , there is one row per match. For instance, if unit 3 is matched with units 5, 6, and 10, there are three rows in *outdata2* that correspond to *id* =3, the first with *idM*=5, the second with *idM*=6, and the last with *idM*=10. For each matched pair (*i,j*), the distance (as an absolute value) between unit *i* and unit *j* of the opposite treatment group is stored as unit *j*'s outcome value.

**An example**

Following Abadie et al. (2004), we use the particular data set constructed by Dehejia and Wahba (1999) from Lalonde (1986) to examine the effect of participation in a job-training program on individuals' earnings in 1978.

re78: individual earnings in 1978
treat = 1 if the individual participates to the job-training program, 0 otherwise
educ: years of education
black=1 if Afro-American, 0 otherwise
hisp=1 if Hispanic, 0 otherwise
married=1 if married, 0 otherwise
re74 (re75): individual earnings in 1974 (1975)
u74 (u75)=1 if unemployed in 1974 (1975), 0 otherwise.

The macro statement:

%nn_matching(data=lib.lalonde,y=re78,w=treat,x=age educ black hisp married re74 re75 u74 u75,M=4,scaling=,covarbias=);

replicate Abadie et al. (2004)'s results apart from the standard error of ATET with bias-correction under homoscedasticity of the conditional variance which is larger in Abadie et al. (2004). The results are presented below.

ESTIMATING AVERAGE TREATEMENT EFFECTS

**Model specification**

**Outcome variable:** re78

**Binary treatment:** Treat

**Matching variables:** age educ black hisp married re74 re75 u74 u75

**Estimation options**

**Number of matches requested:** 4

**Scaling matrix used:** Inverse variances

**Covariates used for bias correction:** age educ black hisp married re74 re75 u74 u75

**Summary statistics**

| | |
|---|---|
| **Number of observations:** | 445 |
| **Number of control units:** | 260 |
| **Number of treated units:** | 185 |
| **Number of treated units matched to controls:** | 174 |
| **Number of control units matched to treated:** | 239 |
| **Number of times a treated unit is used as a match (MIN):** | 1 |
| **Number of times a treated unit is used as a match (MAX):** | 21 |
| **Number of times a control unit is used as a match (MIN):** | 1 |
| **Number of times a control unit is used as a match (MAX):** | 11 |

**Estimation results assuming homoscedasticity of the conditional variance**

| | Estimate | Std.Error | z | P-value | L. bound 95% CI | U. bound 95% CI |
|---|---|---|---|---|---|---|
| **Average Treatment Effect (ATE)** | 1903.326 | 720.215 | 2.643 | 0.008 | 491.731 | 3314.921 |
| **ATE with bias correction** | 1717.726 | 720.341 | 2.385 | 0.017 | 305.883 | 3129.569 |
| **Average Treatment Effect for the Treated (ATET)** | 1994.622 | 712.729 | 2.799 | 0.005 | 597.699 | 3391.544 |
| **ATET with bias correction** | 1838.424 | 712.824 | 2.579 | 0.01 | 441.314 | 3235.534 |

**Estimation results allowing for heteroscedasticity of the conditional variance**

| | Estimate | Std.Error | z | P-value | L. bound 95% CI | U. bound 95% CI |
|---|---|---|---|---|---|---|
| **Average Treatment Effect (ATE)** | 1903.326 | 745.421 | 2.553 | 0.011 | 442.328 | 3364.324 |
| **ATE with bias correction** | 1717.726 | 745.421 | 2.304 | 0.021 | 256.729 | 3178.724 |

**Estimation results allowing for heteroscedasticity of the conditional variance**

| | Estimate | Std.Error | z | P-value | L. bound 95% CI | U. bound 95% CI |
|---|---|---|---|---|---|---|
| **Average Treatment Effect for the Treated (ATET)** | 1994.622 | 752.634 | 2.65 | 0.008 | 519.486 | 3469.757 |
| **ATET with bias correction** | 1838.424 | 752.634 | 2.443 | 0.015 | 363.289 | 3313.56 |

Normalized covariate mean differences are presented in the following table:

**Normalized covariate mean differences between treated and controls**

| | Unmatched Samples | Matched samples (T) | Matched samples (C) |
|---|---|---|---|
| **AGE** | 0.107 | 0.115 | 0.06 |
| **EDUC** | 0.141 | 0.097 | 0.105 |
| **BLACK** | 0.044 | -0.029 | 0.081 |
| **HISP** | -0.175 | -0.079 | -0.16 |
| **MARRIED** | 0.094 | 0.091 | 0.05 |
| **RE74** | -0.002 | 0.057 | -0.012 |
| **RE75** | 0.084 | 0.095 | 0.096 |
| **U74** | -0.094 | -0.092 | -0.084 |
| **U75** | -0.177 | -0.154 | -0.182 |

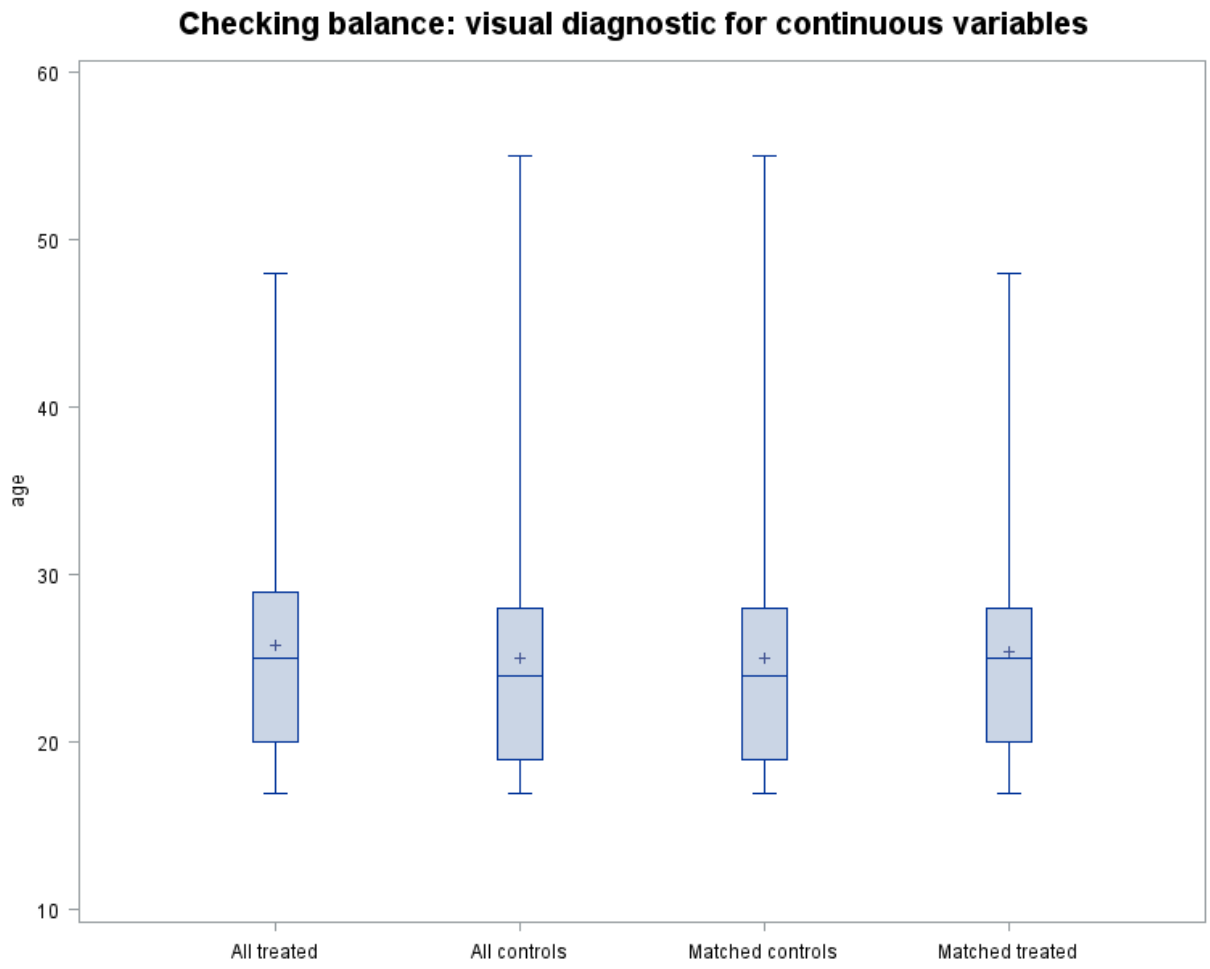
Note: 'Matched samples (T)' is for normalized mean differences between all sample treated and their matches, 'Matched samples (C)' for normalized mean differences between all sample controls and their matches.

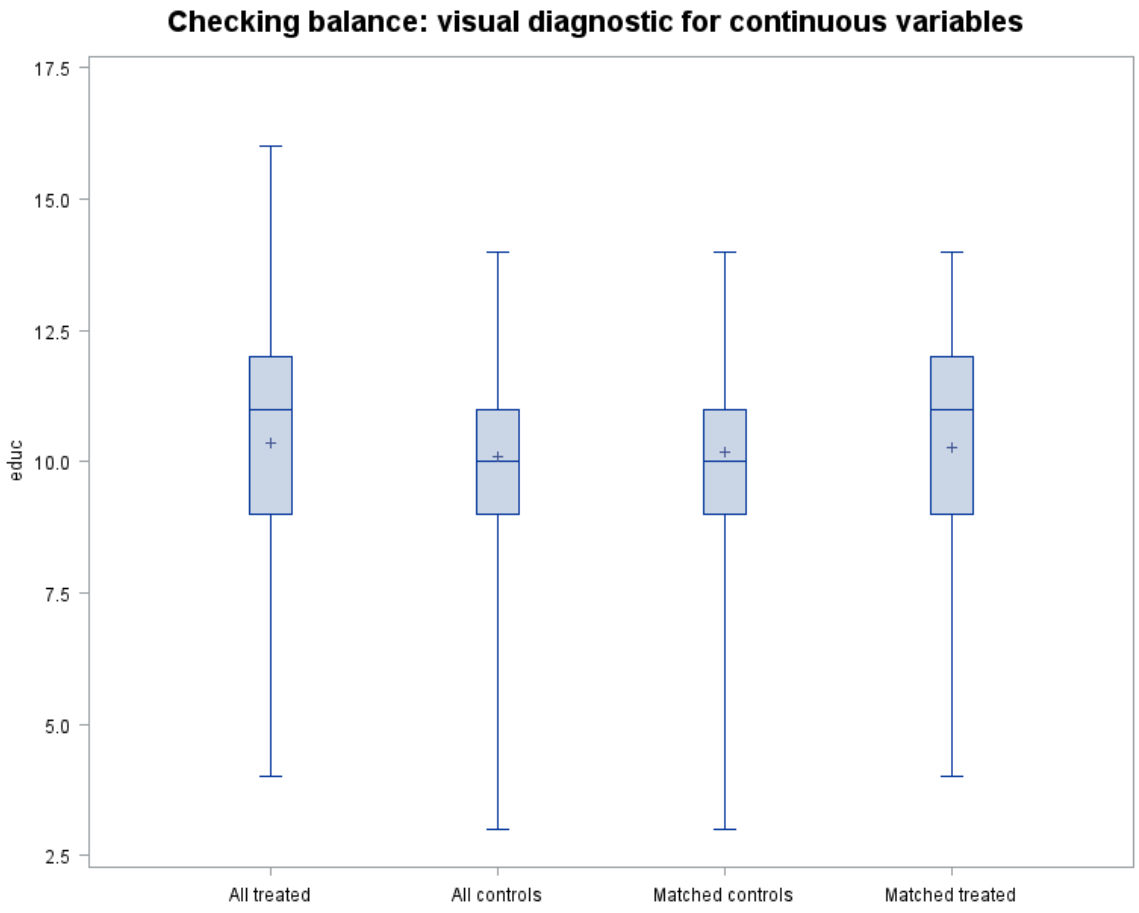
Comparing binary covariate distributions between treatment groups before and after matching

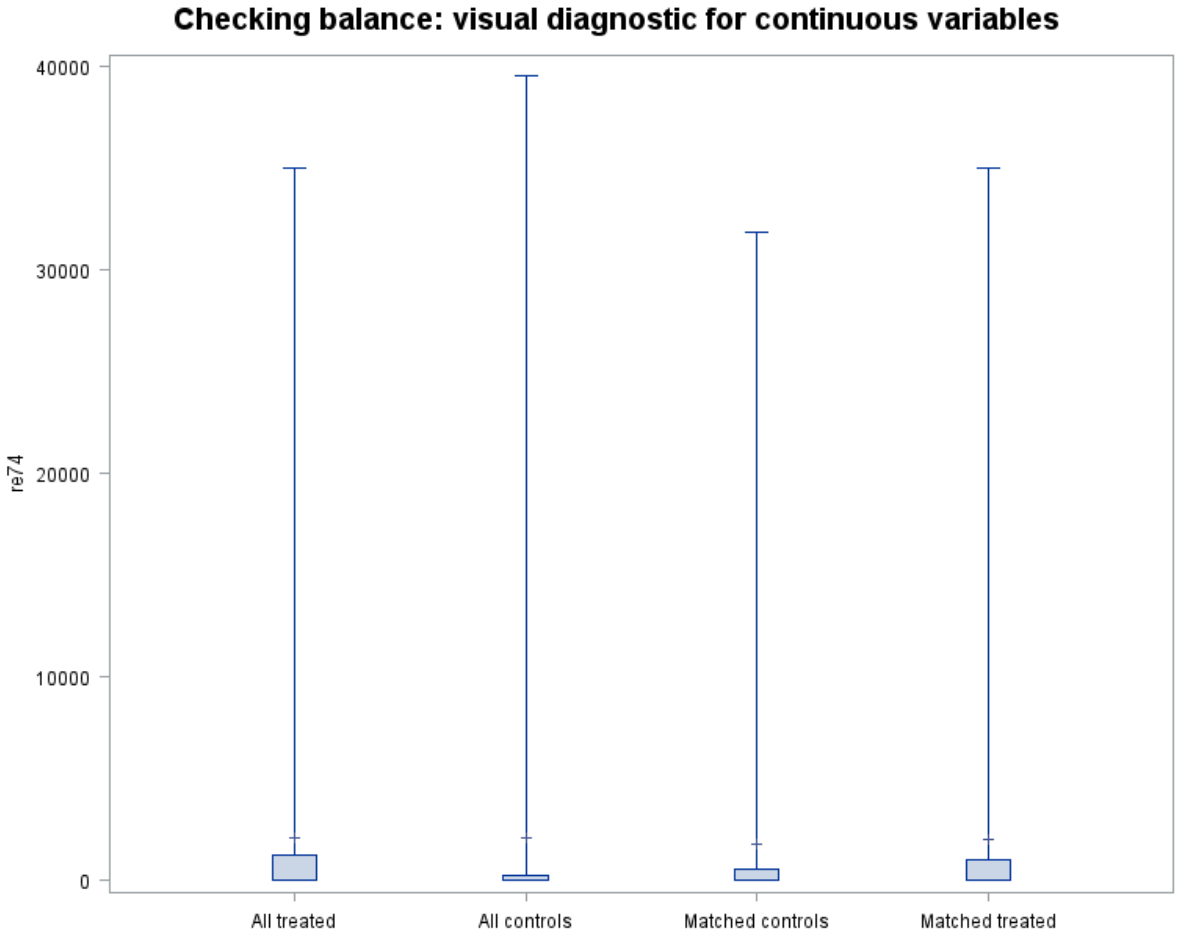
| | | SampleType | | | |
|---|---|---|---|---|---|
| | | All treated | All controls | Matched controls | Matched treated |
| black | | | | | |
| 0 | Distribution in % | 15.68 | 17.31 | 14.37 | 14.64 |
| 1 | Distribution in % | 84.32 | 82.69 | 85.63 | 85.36 |
| hisp | | | | | |
| 0 | Distribution in % | 94.05 | 89.23 | 93.68 | 92.05 |
| 1 | Distribution in % | 5.95 | 10.77 | 6.32 | 7.95 |
| married | | | | | |
| 0 | Distribution in % | 81.08 | 84.62 | 82.76 | 84.52 |
| 1 | Distribution in % | 18.92 | 15.38 | 17.24 | 15.48 |
| u74 | | | | | |

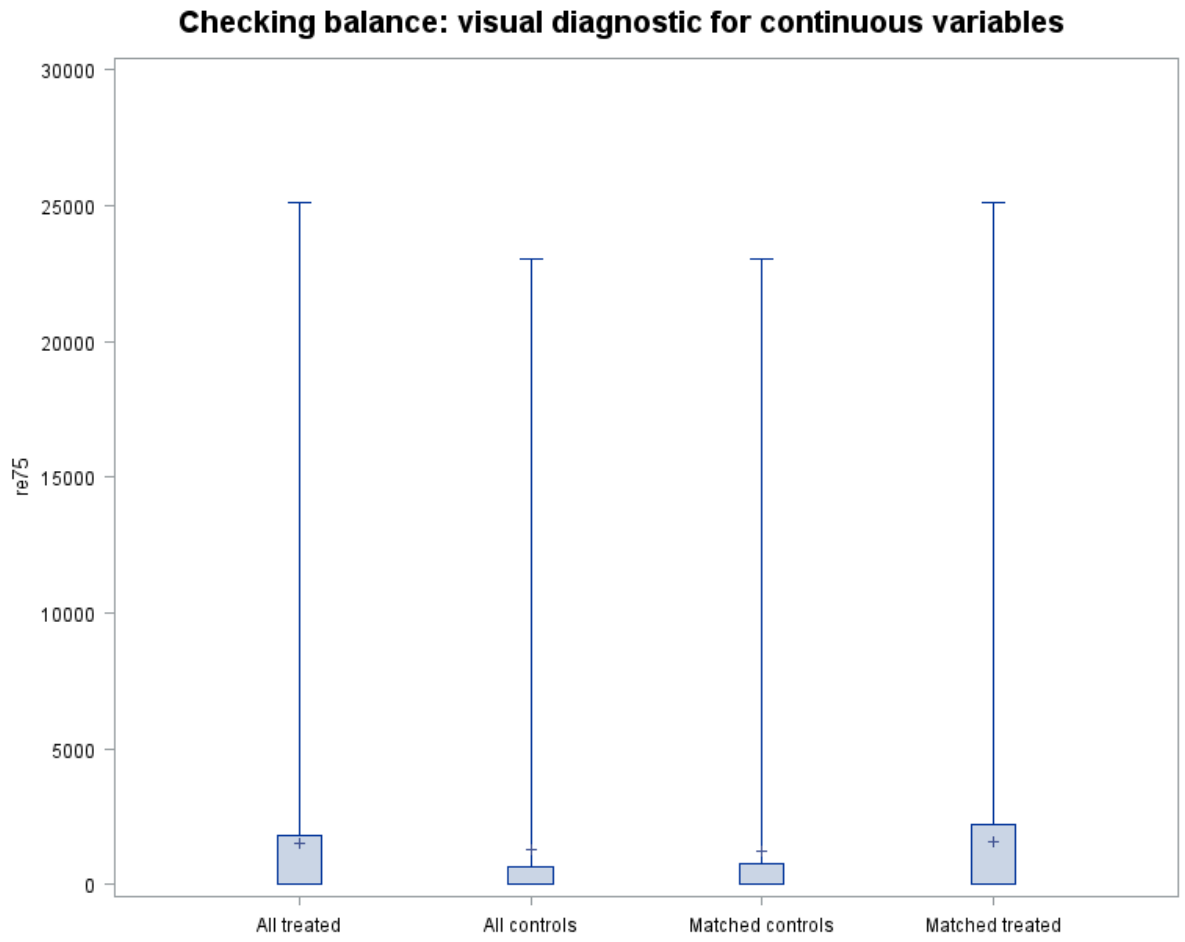
| | | | | | |
|---|---|---|---|---|---|
| 0 | Distribution in % | 29.19 | 25.00 | 28.74 | 25.10 |
| 1 | Distribution in % | 70.81 | 75.00 | 71.26 | 74.90 |
| u75 | | | | | |
| 0 | Distribution in % | 40.00 | 31.54 | 40.23 | 32.64 |
| 1 | Distribution in % | 60.00 | 68.46 | 59.77 | 67.36 |

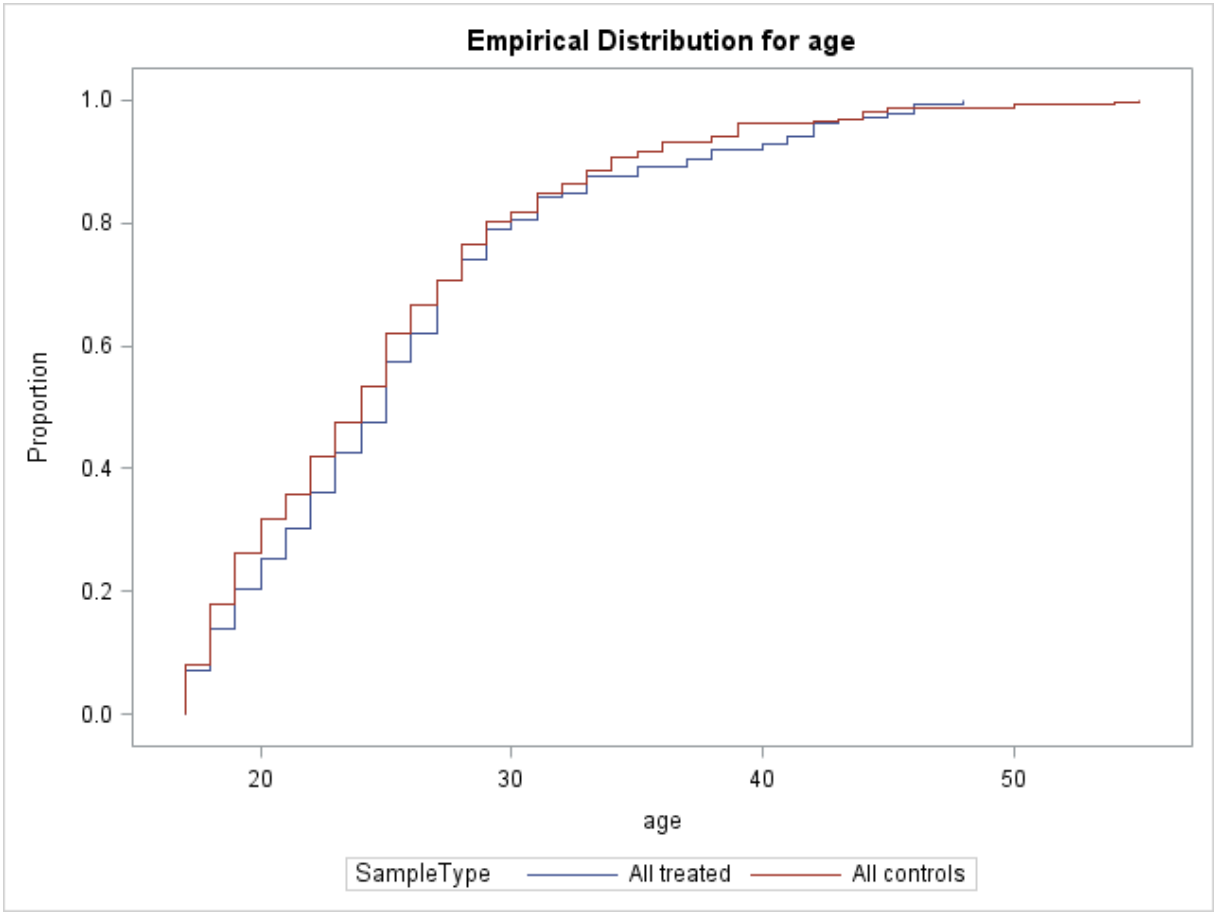
Visual diagnostic to assess the validity of the balancing hypothesis for continuous variables is then provided:

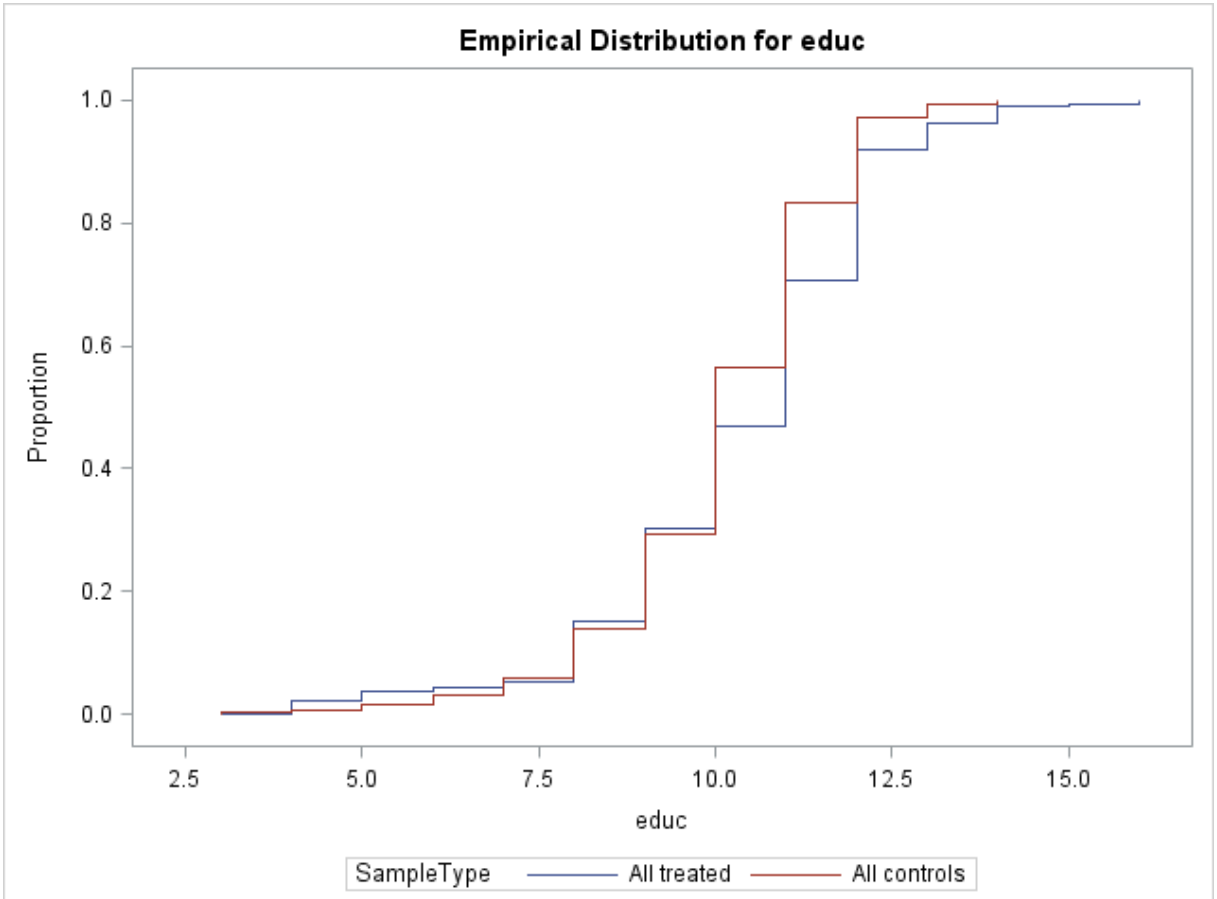


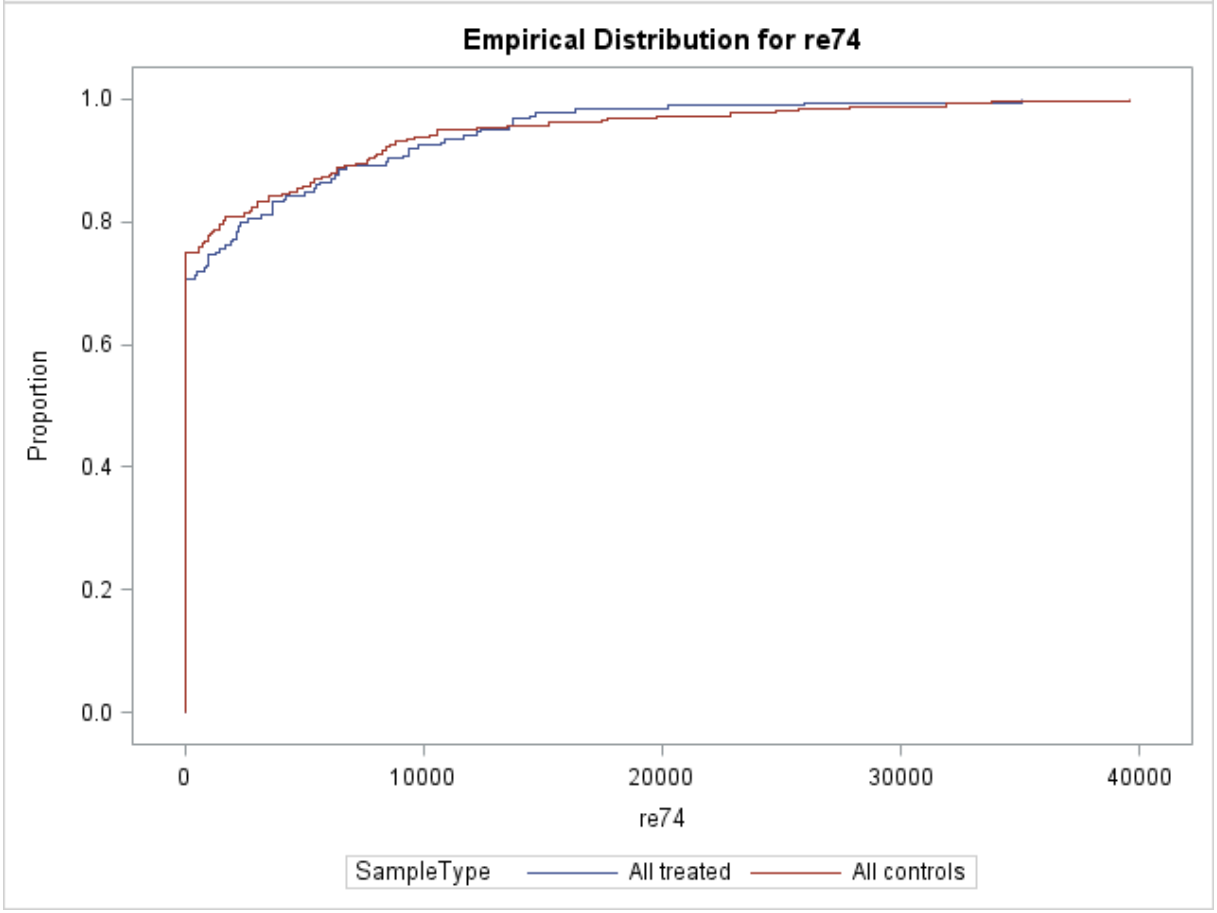

Checking balance: visual diagnostic for continuous variables
17.5
15.0
12.5
10.0
7.5
5.0
2.5
educ
All treated
All controls
Matched controls
Matched treated

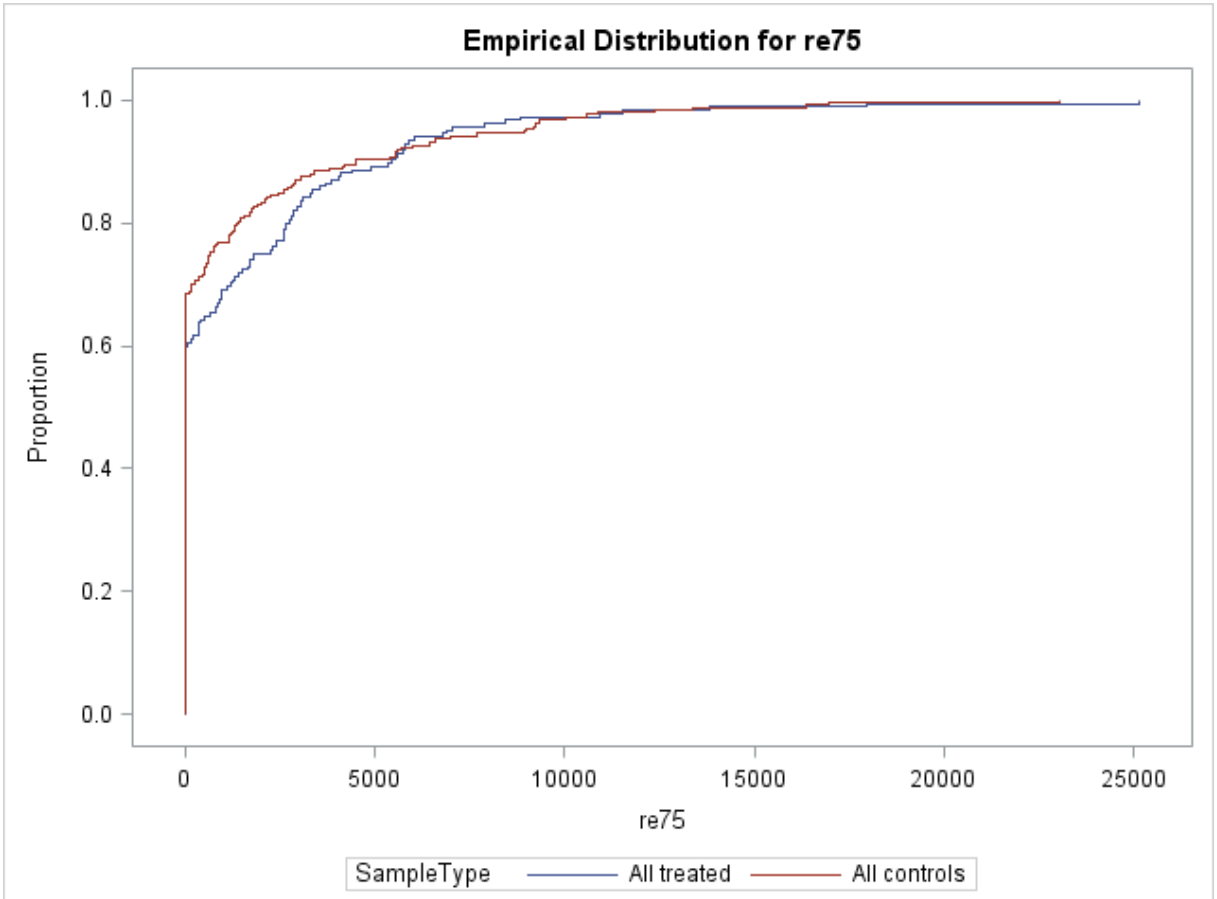

Checking balance: visual diagnostic for continuous variables
40000
30000
20000
10000
0
re74
All treated
All controls
Matched controls
Matched treated

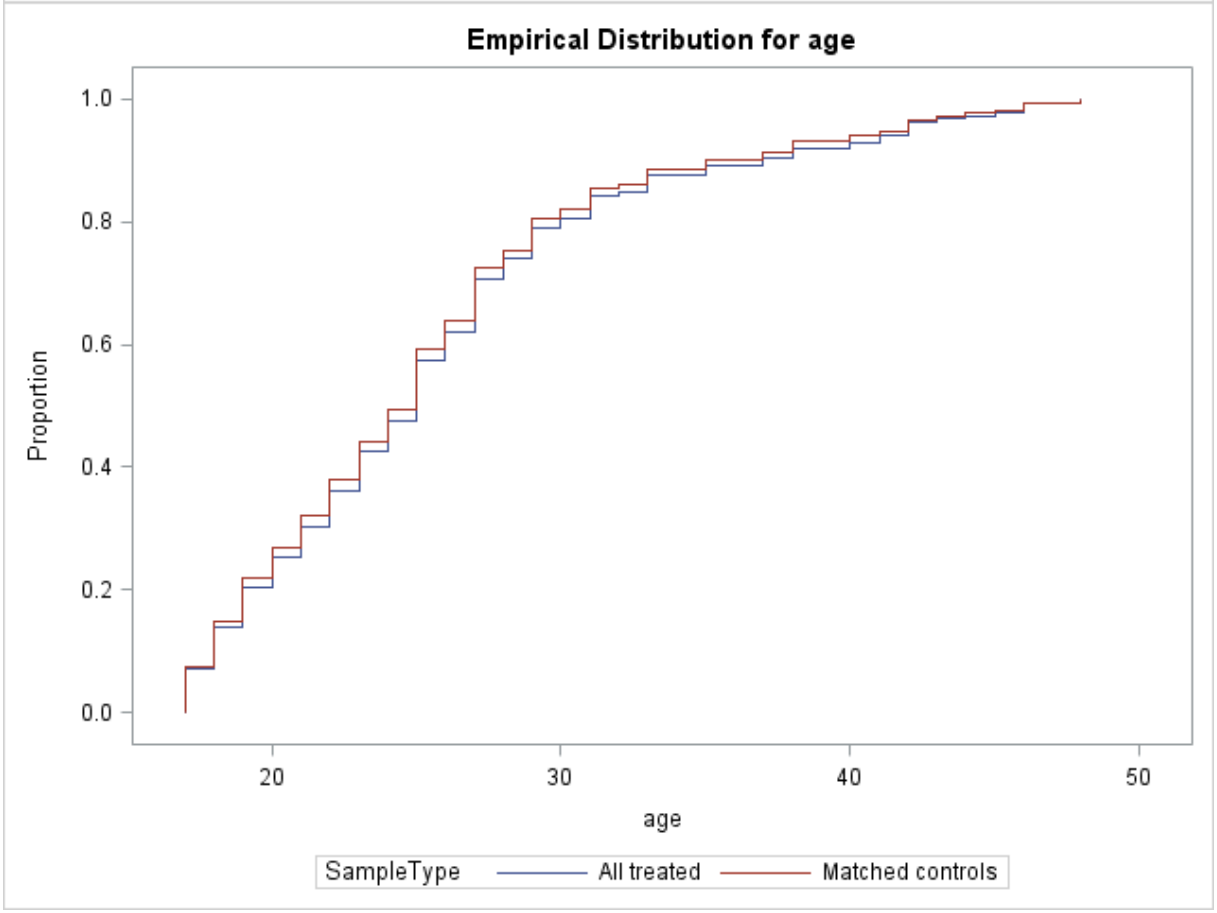

Checking balance: visual diagnostic for continuous variables
re75
30000
25000
20000
15000
10000
5000
0
All treated
All controls
Matched controls
Matched treated

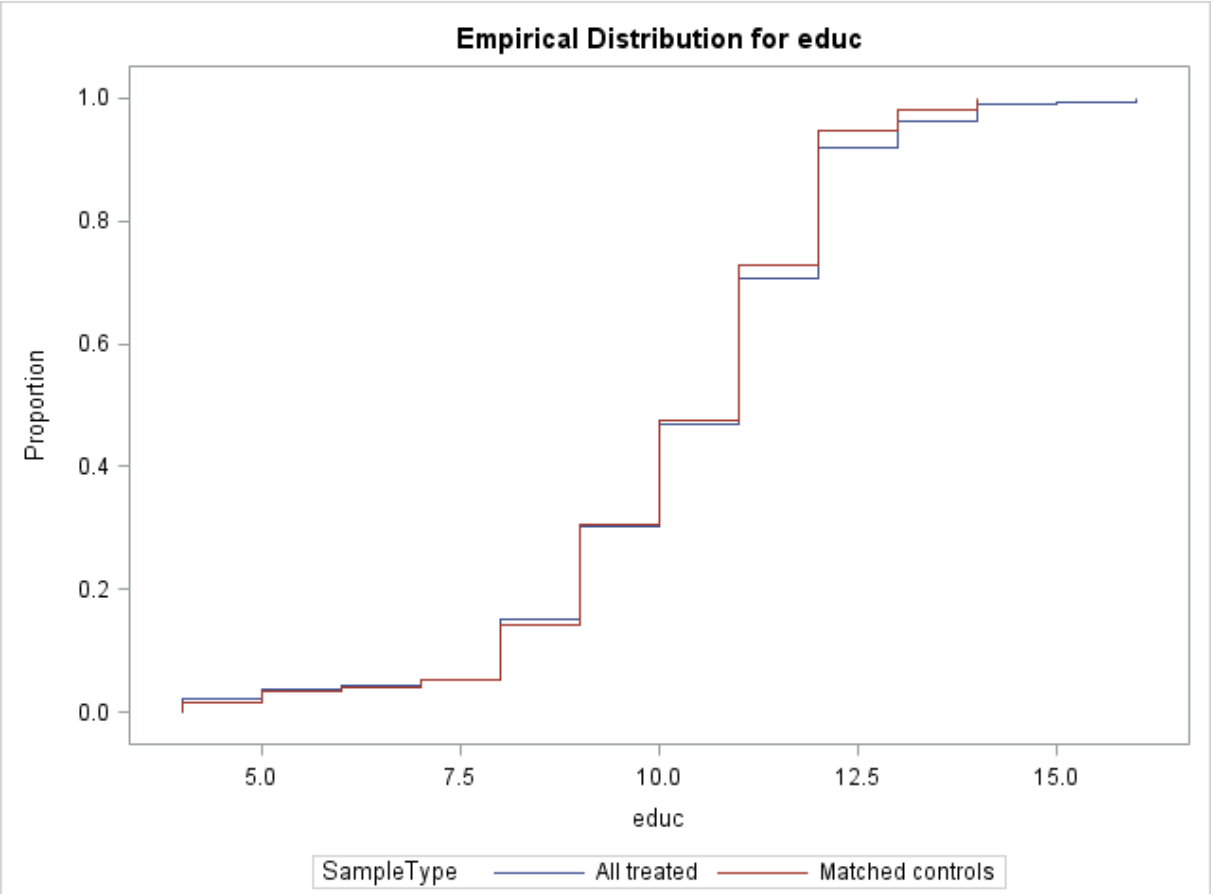

Empirical Distribution for age
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
20
30
40
50
age
SampleType
All treated
All controls

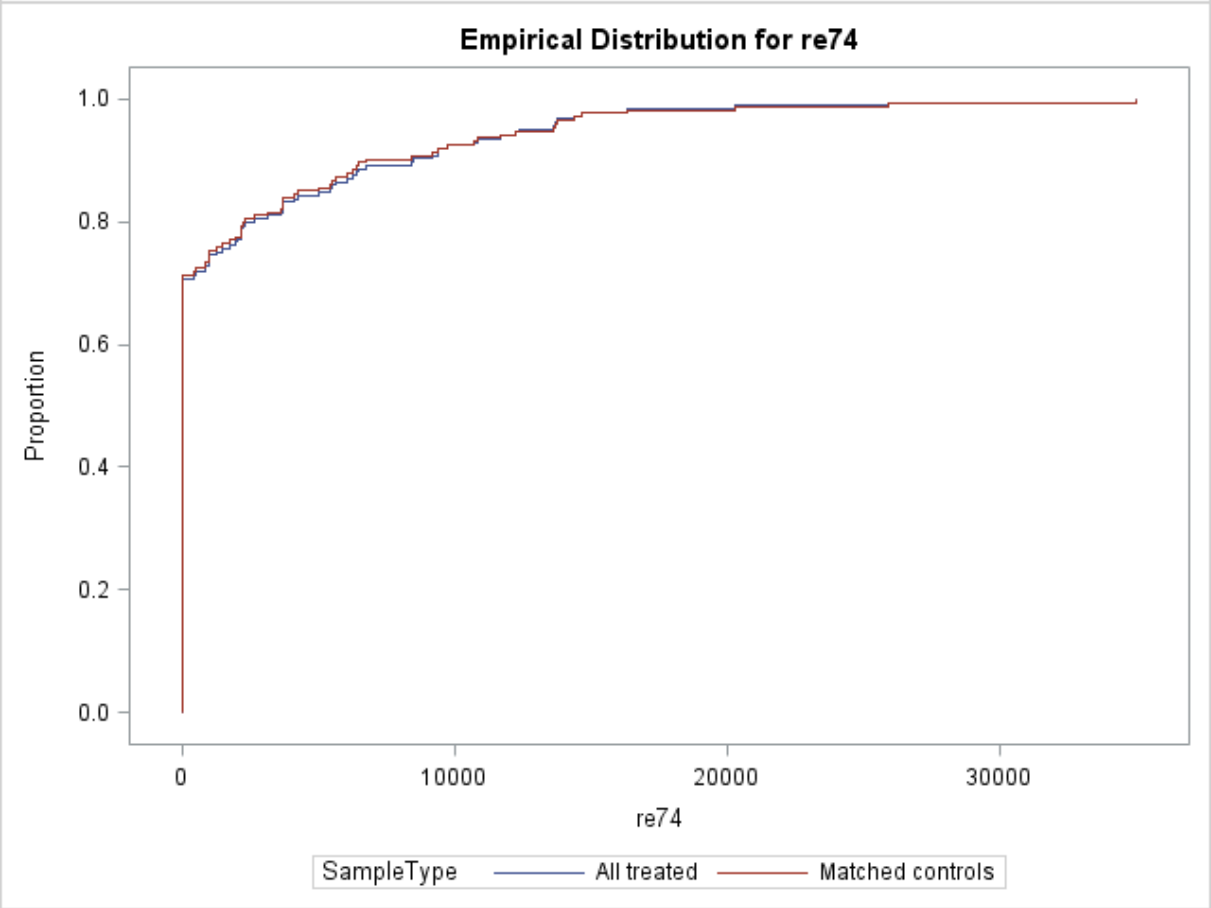

Empirical Distribution for educ
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
2.5
5.0
7.5
10.0
12.5
15.0
educ
SampleType
All treated
All controls

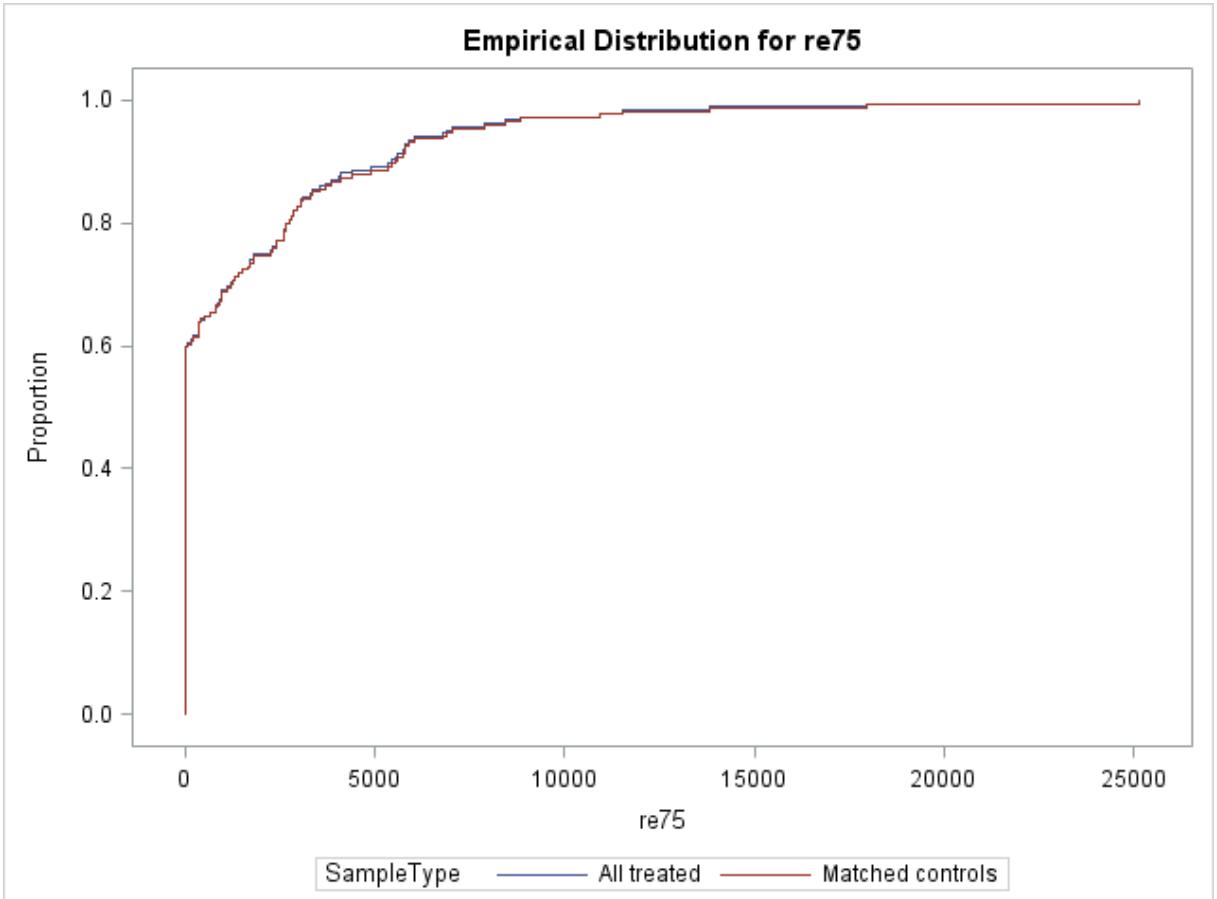

Empirical Distribution for re74
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
0
10000
20000
30000
40000
re74
SampleType
All treated
All controls

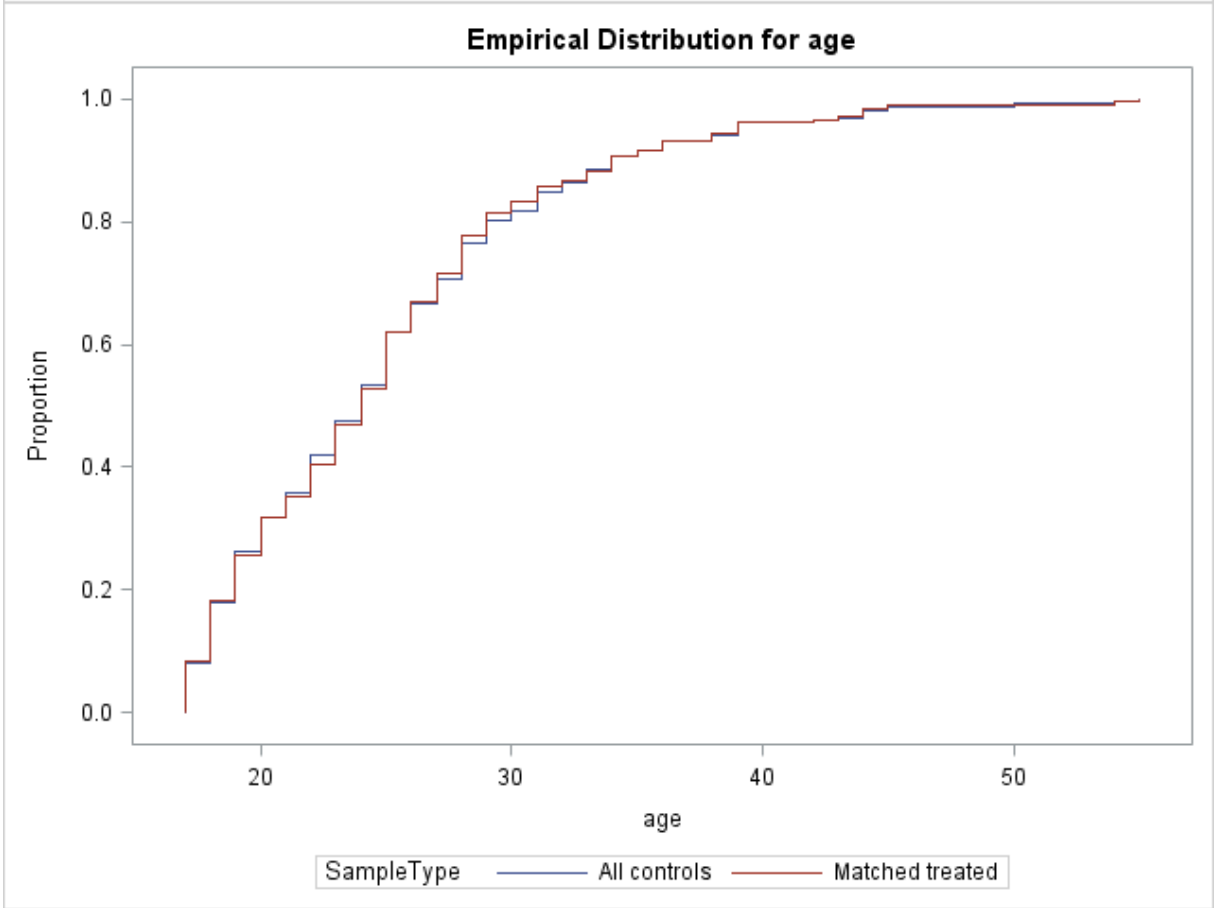

Empirical Distribution for re75
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
0
5000
10000
15000
20000
25000
re75
SampleType
All treated
All controls

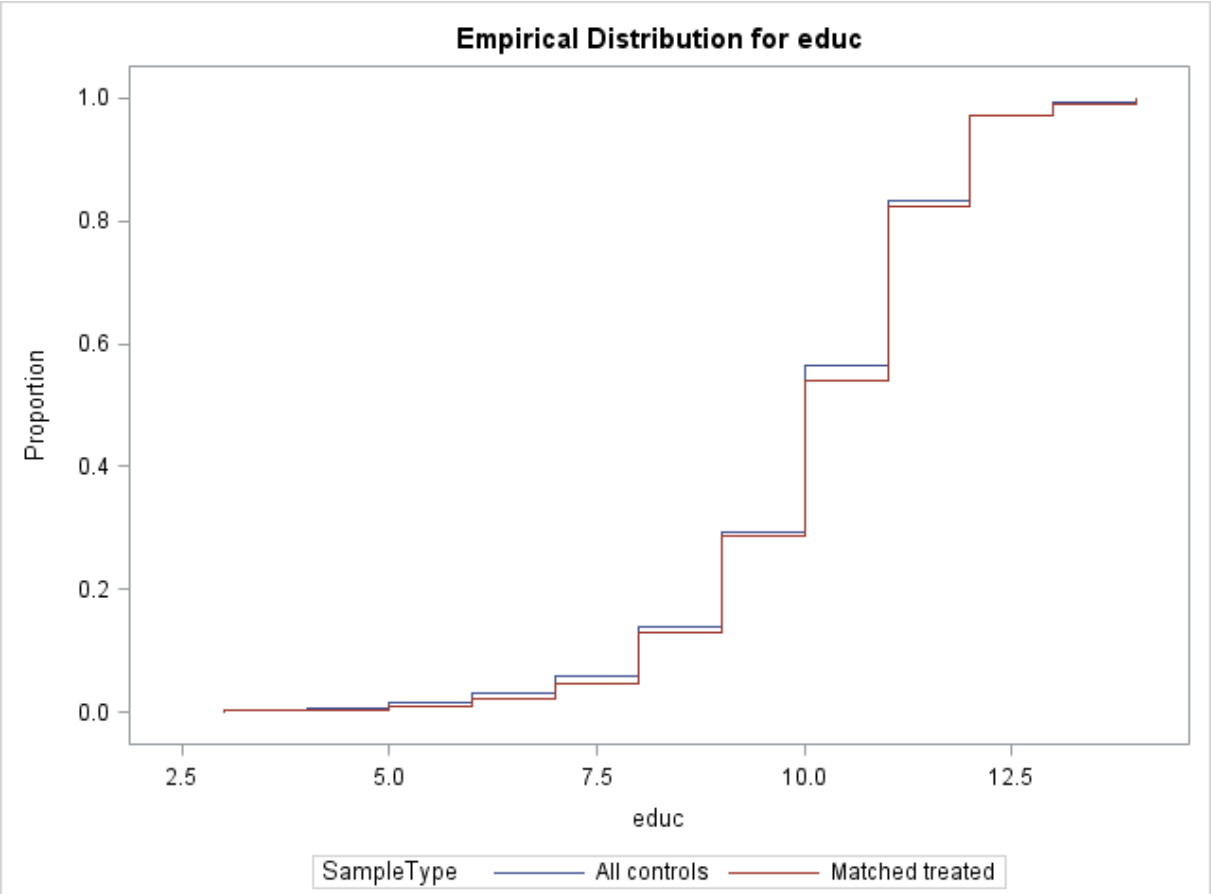

Empirical Distribution for age
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
20
30
40
50
age
SampleType
All treated
Matched controls

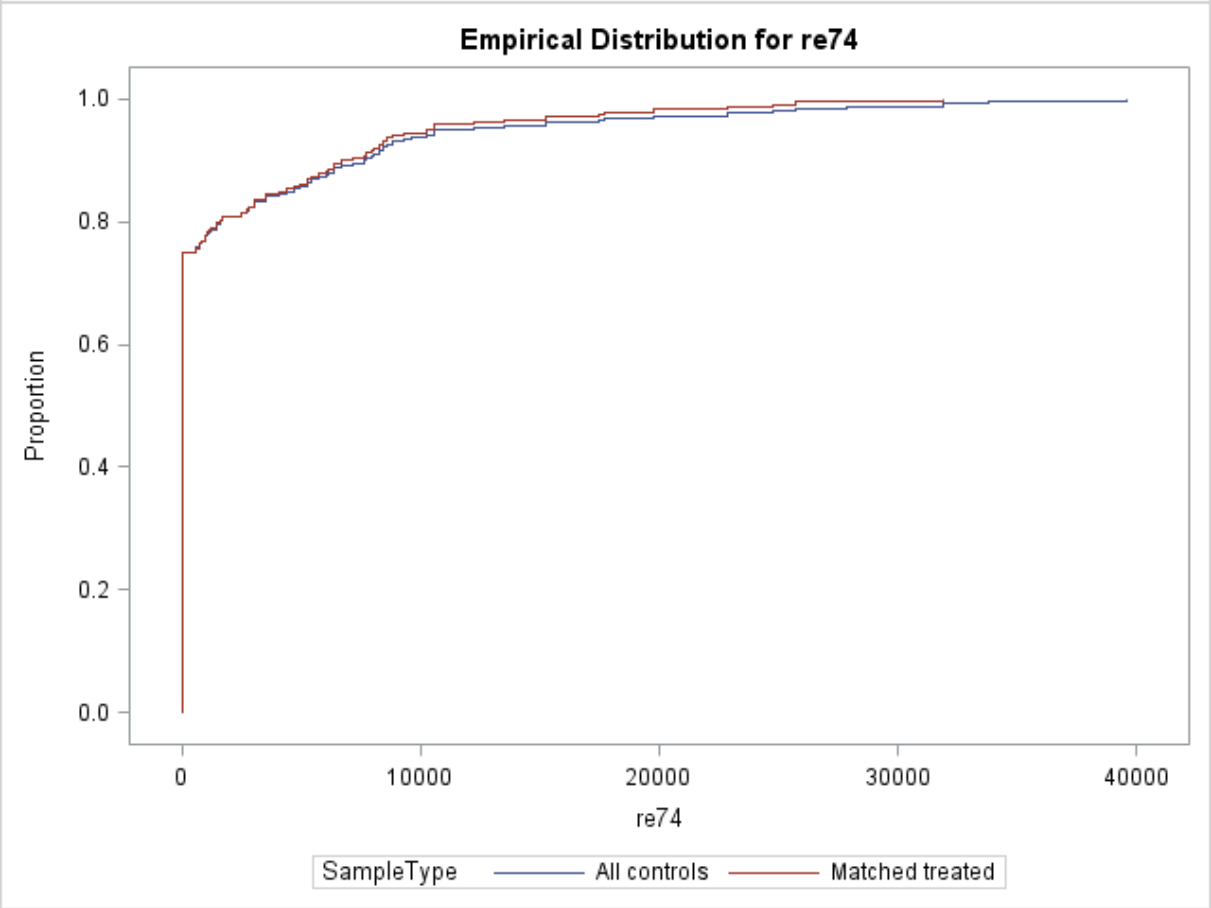

Empirical Distribution for educ
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
5.0
7.5
10.0
12.5
15.0
educ
SampleType
All treated
Matched controls

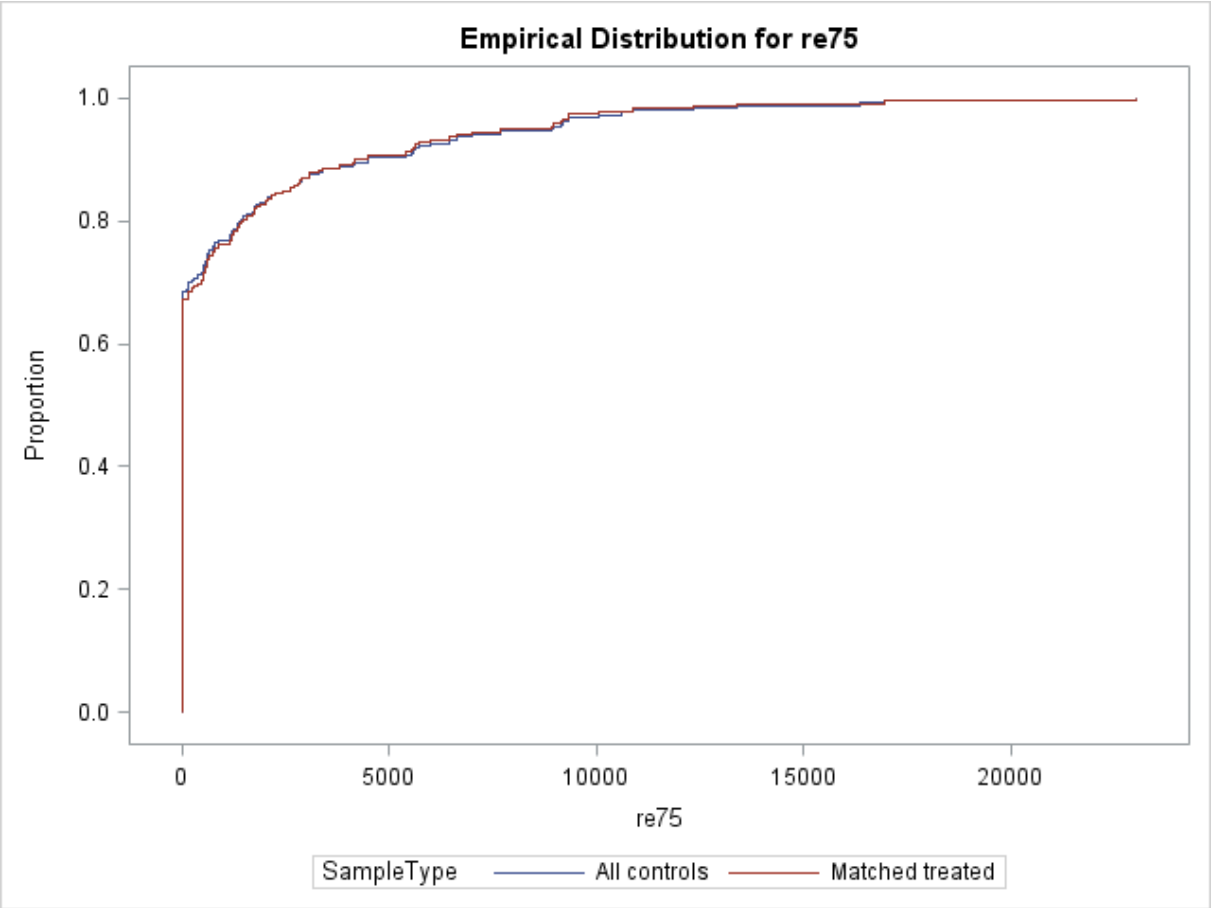

Empirical Distribution for re74
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
0
10000
20000
30000
re74
SampleType
All treated
Matched controls


Empirical Distribution for re75
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
0
5000
10000
15000
20000
25000
re75
SampleType
All treated
Matched controls


Empirical Distribution for age
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
20
30
40
50
age
SampleType
All controls
Matched treated


Empirical Distribution for educ
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
2.5
5.0
7.5
10.0
12.5
educ
SampleType
All controls
Matched treated


Empirical Distribution for re74
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
0
10000
20000
30000
40000
re74
SampleType
All controls
Matched treated


Empirical Distribution for re75
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
0
5000
10000
15000
20000
re75
SampleType
All controls
Matched treated

**References**

Abadie A., Drukker D., Leber Herr J. and Imbens G. W. (2004), "Implementing matching estimators for average treatment effects in Stata", The Stata Journal, vol. 4, n°3: 290-311.

Abadie A., and Imbens G. W. (2002), "Simple and bias-corrected matching estimators for average treatment effects", NBER Technical Working Paper 283.

Abadie A., and Imbens G. W. (2011), "Bias-Corrected Matching Estimators for Average Treatment Effects", Journal of Business and Economic Statistics, vol. 29, n°1: 1-11.

Austin P. C. (2009), "Balance diagnostics for comparing the distribution of baseline covariates between treatment group in propensity-score matched samples",
Statistics in Medicine, vol. 28: 3083–3107.

Dehehia R. H., and Wabba S. (1999), "Causal effects in non-experimental studies: Re-evaluation of the evaluation of training programs", Journal of the American Statistical Association, vol. 94: 1053-1062.

Lalonde R. J. (1986), "Evaluating the econometric evaluations of training programs", American Economic Review, vol. 74, n°4: 604-620.