A SAS macro to assess covariate balance between the treated and control samples

Nicolas Moreau

DOCUMENTS DE TRAVAIL CEMOI / CEMOI WORKING PAPERS

A SAS macro to assess covariate balance between the treated and control samples

Nicolas Moreau[1]

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

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

September 2020

Abstract

This paper presents a SAS macro to assess overlap in covariate distributions by treatment status.

JEL : C210

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

Introduction

In this paper, we present the SAS macro *overlap*. *overlap* computes four measures of the differences between covariate distributions and three measures of balance between multivariate distributions. We draw heavily on Imbens and Rubin (2015) and refer the reader to chapter 14 of their book for a clear and comprehensive presentation of the different measures.

For each covariate, *overlap* provides the normalized difference in averages between groups, the conventional *t*-statistic to test the null hypothesis of equal means, the logarithm of the ratio of standard deviations by treatment status to compare dispersion, and the proportions of both the treated and the control units with covariate values outside the central 95% part of the distribution for the other treatment group.
overlap also provides three summary measures of the overall difference in covariate distribution between the treated and control groups. It calculates the distance between the mean vectors of each group using the Mahalanobis distance, and two overlap statistics that measure the proportion of units in each group for which there is at least one unit from the other group with a similar linearized propensity score.

The source code is available at https://cemoi.univ-reunion.fr/econometrie-avec-r-et-sas.

Syntax of *overlap*

The syntax is %overlap(data=, treatment=, covariates=, varpscore=, similar=);

where *data* specifies the data set, *treatment* the binary variable treatment indicator, *covariates* the list of covariates to be used, and *varpscore* the list of variables to be used to estimate the propensity score. *similar* indicates whether the user requires the calculation of the two overlap statistics. If *similar* is empty, the Mahalanobis distance is automatically computed. If *similar=no*, *similar*=0, or *similar*=none (or whatever character/value except a blank), these statistics are not computed.[2]

Note that all variables in *treatment*, *covariates* and *varpscore* must be numeric.

Results presentation

Four histograms are displayed to graphically represent the estimated distributions of the propensity score and linearized propensity score, for treated and controls respectively.

The first table shows the first four measures to assess overlap in covariate distributions between treated and controls. **NbObs_c** and **NbObs_t** are the number of nonmissing values of the covariate for the control group and the treated group, **Mean_c** ($\bar{X}_c$) and **Mean_t** ($\bar{X}_t$) the sample average of the covariate values for the control group and the treated group, and **Std_c** (s_c) and **Std_t** (s_t) the sample standard deviation of the covariate values for the control and treated group, respectively.
Let N_c denote the number of control units, and N_t the number of treated units. Let W_i denote the treatment indicator. It is equal to 1 if unit *i* is treated and 0 otherwise. Then, the sample averages of the covariates values are $\bar{X}_c = \frac{1}{N_c}\sum_{i:\, W_i=0} X_i$ for the controls and $\bar{X}_t = \frac{1}{N_t}\sum_{i:\, W_i=1} X_i$ for the treated. The standard deviations are the square root of the within-group sample variances of the covariate which are $s_c^2 = \frac{1}{N_c-1}\sum_{i:\, W_i=0}(X_i - \bar{X}_c)^2$ for the controls and $s_t^2 = \frac{1}{N_t-1}\sum_{i:\, W_i=1}(X_i - \bar{X}_t)^2$ for the treated.
The conventional *t*-statistic is named **TStat**. It is equal to $\frac{\bar{X}_t - \bar{X}_c}{\sqrt{s_t^2/N_t + s_c^2/N_c}}$.
The normalized difference in average covariate values is called **NormDif**. It is equal to $\frac{\bar{X}_t - \bar{X}_c}{\sqrt{(s_t^2 + s_c^2)/2}}$.
LogRatioSTD denotes the logarithm of the ratio of standard deviations. It is equal to $\ln(s_t) - \ln(s_c)$.

[2] To calculate these statistics, a table containing the Cartesian product of the linearized propensity score vectors for the control and treated units is created. If there are N_c control units and N_t treated units in the original dataset, this new table will include $N_c \times N_t$ observations, which can be very large.

OUTSIDE95_c denotes the proportion of control units with covariate values outside the central 95% part of the distribution for the treated group. It is equal to:

$$\left(1-\hat{F}_c\left(\hat{F}_t^{-1}(0.975)\right)\right)+\hat{F}_c\left(\hat{F}_t^{-1}(0.025)\right),$$

where $\hat{F}_c(\cdot)$ and $\hat{F}_t(\cdot)$ are the empirical cumulative distribution functions of the covariate in the control and treated groups, respectively.

In the same way, **OUTSIDE95_t** denotes the proportion of treated units with covariate values outside the central 95% part of the distribution for the control group. It is equal to:

$$\left(1-\hat{F}_t\left(\hat{F}_c^{-1}(0.975)\right)\right)+\hat{F}_t\left(\hat{F}_c^{-1}(0.025)\right).$$

The second table exhibits the same measures for the estimated propensity score and the linearized estimated propensity score. Let $\hat{e}(x)$ and $\hat{\ell}(x)$ denote the estimated propensity score and linearized estimated propensity score, respectively. Then, $\hat{\ell}(x)=\ln\left(\frac{\hat{e}(x)}{1-\hat{e}(x)}\right)$.

The last table includes measures of the overall difference in covariate distribution between the treated and control groups. **Distance** is the Mahalanobis distance between the mean vectors. The mean vectors consist of the means of each covariate. Let $\boldsymbol{X}$ denote the *K*-vector of covariates. Then **Distance** is equal to $\sqrt{\left(\overline{\boldsymbol{X}}_t-\overline{\boldsymbol{X}}_c\right)^T\left(\frac{\hat{\Sigma}_c+\hat{\Sigma}_t}{2}\right)^{-1}\left(\overline{\boldsymbol{X}}_t-\overline{\boldsymbol{X}}_c\right)}$, where $\hat{\Sigma}_c$ and $\hat{\Sigma}_t$ are the $K\times K$ sample covariance matrices of the covariates in the control and treated groups, respectively. These matrices are calculated as:

$$\hat{\Sigma}_c=\frac{1}{N_c-1}\sum_{i:\,W_i=0}(\boldsymbol{X}_i-\overline{\boldsymbol{X}}_c)\cdot(\boldsymbol{X}_i-\overline{\boldsymbol{X}}_c)^T \text{ and } \hat{\Sigma}_t=\frac{1}{N_t-1}\sum_{i:\,W_i=1}(\boldsymbol{X}_i-\overline{\boldsymbol{X}}_t)\cdot(\boldsymbol{X}_i-\overline{\boldsymbol{X}}_t)^T.$$

Finally, the measures **Overlap_c** and **Overlap_t** represent the proportion of units in each group for which there is at least one unit from the other group with a similar linearized propensity score (that is, the difference should not exceed 0.1). Define (see Imbens and Rubin, 2015, page 318), for each unit *i*, the indicator ς_i that takes on the value one if there is at least one unit i' with $W_{i'}=1-W_i$ that has a similar value for the linearized propensity score and zero otherwise:

$$\varsigma_i=\begin{cases}1 & \text{if } \sum_{i':\,W_{i'}\neq W_i}\mathbf{1}_{|\hat{\ell}(\boldsymbol{x}_{i'})-\hat{\ell}(\boldsymbol{x}_i)|\le 0.1}\ \ge 1,\\ 0 & \text{otherwise.}\end{cases}$$

Then **Overlap_c** is equal to $\frac{1}{N_c}\sum_{i:\,W_i=0}\varsigma_i$ and **Overlap_t** to $\frac{1}{N_t}\sum_{i:\,W_i=1}\varsigma_i$.

An example

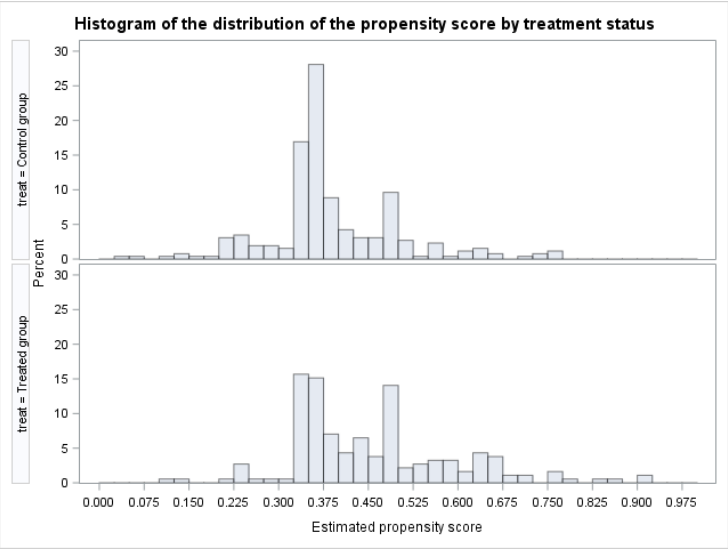
Following Imbens and Rubin (2015), we use the particular data set constructed by Dehejia and Wahba (1999) from Lalonde (1986) to examine the effect of participation in a job-training program on individuals' earnings in 1978.

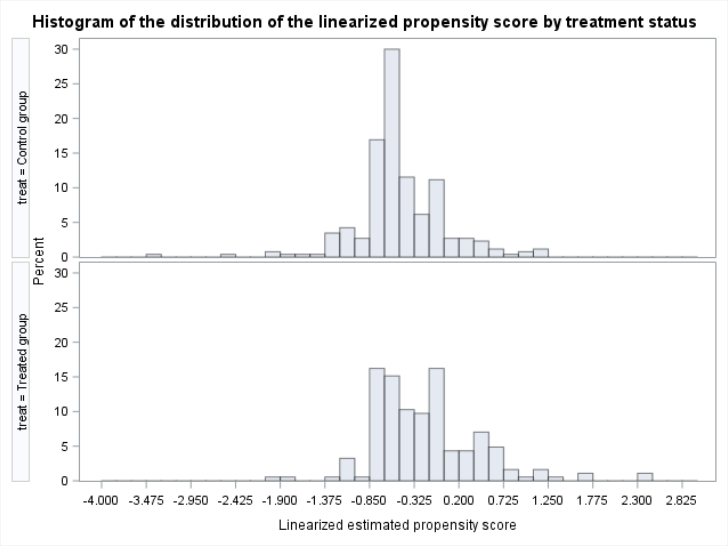
re78: individual earnings in 1978
treat = 1 if the individual participates to the job-training program, 0 otherwise
educ: years of education
black=1 if Afro-American, 0 otherwise
hisp=1 if Hispanic, 0 otherwise
married=1 if married, 0 otherwise
re74 (re75): individual earnings in 1974 (1975)
u74 (u75)=1 if unemployed in 1974 (1975), 0 otherwise.
nodegr=1 if not graduated, 0 oterwise.
nodegeduc is the product of nodegr with educ.
re74nodegr is the product of re74 with nodegr.
u75educ is the product of u75 with educ.

The macro statement:

%overlap(data=lib.lalonde,treatment=treat,covariates=black hisp age married nodegr educ re74 u74 re75 u75, varpscore=re74 u74 re75 u75 nodegr hisp educ nodegeduc re74nodegr u75educ,similar=);

replicates Imbens and Rubin (2015)'s results. Note that not all of the variables used to calculate the propensity score are included in the list of covariates. The outputs are shown below.





Assessing Balance in Covariate Distributions between Treated and Controls

Covariate	NbObs_c	Mean_c	Std_c	NbObs_t	Mean_t	Std_t	TStat	NormDif	LogRatioSTD	OUTSIDE95_c	OUTSIDE95_t
black	260	0.8269	0.37904	185	0.8432	0.36456	0.45778	0.04389	-0.03897	0.000000	0.000000
hisp	260	0.1077	0.31059	185	0.0595	0.23712	-1.85654	-0.17456	-0.26989	0.000000	0.000000
age	260	25.0538	7.05774	185	25.8162	7.15502	1.11404	0.10728	0.01369	0.011538	0.027027
married	260	0.1538	0.36150	185	0.1892	0.39272	0.96684	0.09364	0.08285	0.000000	0.000000
nodegr	260	0.8346	0.37224	185	0.7081	0.45587	-3.10850	-0.30399	0.20265	0.000000	0.000000
educ	260	10.0885	1.61432	185	10.3459	2.01065	1.44218	0.14122	0.21954	0.007692	0.075676
re741000	260	2.1070	5.68791	185	2.0956	4.88662	-0.02275	-0.00216	-0.15184	0.042308	0.010811
u74	260	0.7500	0.43385	185	0.7081	0.45587	-0.97469	-0.09414	0.04951	0.000000	0.000000
re751000	260	1.2669	3.10298	185	1.5321	3.21925	0.86921	0.08386	0.03678	0.019231	0.027027
u75	260	0.6846	0.46557	185	0.6000	0.49123	-1.82997	-0.17681	0.05366	0.000000	0.000000

Assessing Balance in Multivariate Distributions between Treated and Controls

Pscore	NbObs_c	Mean_c	Std_c	NbObs_t	Mean_t	Std_t	TStat	NormDif	LogRatioSTD	OUTSIDE95_c	OUTSIDE95_t
Propensity score	260	0.38734	0.11240	185	0.45564	0.13869	5.52968	0.54107	0.21013	0.061538	0.086486
Linearized propensity score	260	-0.48570	0.52913	185	-0.17733	0.62824	5.44247	0.53093	0.17168	0.061538	0.086486

Assessing Balance in Multivariate Distributions between Treated and Controls

Distance	Overlap_c	Overlap_t
0.43561	0.97692	0.96757

References

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.

Imbens G. W., and Rubin D. B., Causal Inference for Statistics, Social, and Biomedical Sciences, Cambridge University Press, 2015, chapter 14.

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