A SAS® Macro for measuring and testing global balance of categorical covariates

Camillo, Furio and D'Attoma,Ida
Dipartimento di Scienze Statistiche, Università di Bologna
via Belle Arti,41- 40126-
Bologna, Italy

November 22, 2010

Professor Furio Camillo is an Associate Professor of Business Statistics and Data Mining at Department of Statistical Sciences, University of Bologna (Italy). He studies the applications of data mining in private and public organizations in the areas of marketing, customer relationship management and policy evaluation.
E-mail: furio.camillo@unibo.it

Dr. Ida D'Attoma is a Research Fellow at Department of Statistical Sciences, University of Bologna (Italy). Her research interests include micro data mining, causal inference in observational studies, subgroup analysis, and new methods for public and private program evaluation.
E-mail: Ida.dattoma2@unibo.it

Abstract

We developed a SAS® Macro [1] to simultaneously measure and test global balance of categorical covariates. The purpose of the %BALANCE macro is to check global balance and test it by subgroups, no matter how groups are obtained. A subgroup could be a bin of a Propensity Score Analysis or a group of any classification method.
For each group we measure and test global balance according to the multivariate measure and its test introduced in Camillo and D'Attoma (2010) [2] and D'Attoma and Camillo (2010) [3].
The %BALANCE Macro use the SAS/Iml matrix language to obtain such measure. It generates as output the global balance measure and its test by subgroups. The user will choose a set of parameters that define the balance measure and its multivariate test.

The use of such Macro can significantly reduce the amount of time required to simultaneously test balance of any number of categorical covariates by subgroups.

Description

Given a set of pre-treatment categorical covariates involved in the selection processs, the %BALANCE macro simultaneously checks balance of all categorical covariates and tests its statistically significance in a multivariate ways accordinding to the GI formula and the multivariate test introduced in D'Attoma and Camillo (2010) [3].

The macro uses the Sas/iml matrix language that allows to compute the GI formula [3]:

$$\mathbf{GI} = \frac{\mathbf{1}}{\mathbf{Q}} \sum_{t=1}^{T} \sum_{j=1}^{JQ} \frac{\mathbf{b}_{tj}^{2}}{\mathbf{k}_{.t}\mathbf{k}_{.j}} - \mathbf{1}$$

where $\mathbf{Q}$ is the number of all categorical covariates introduced in the analysis, $\mathbf{b}_{tj}$ is the number of units with category $\mathbf{j} \in \mathbf{J}_{Q}$ in the treatment group $\mathbf{t} \in \mathbf{T}$, $\mathbf{k}_{.t}$ is the treatment group size and $\mathbf{k}_{.j}$ is the number of units with category $\mathbf{j} \in \mathbf{J}_{Q}$. Furthermore, the %BALANCE macro allows to test the GI significance according to the following confidence interval [3]:

$$\mathbf{GI} \in \left(\mathbf{0}, \frac{\chi^{2}_{(T-1)(J-1),\alpha}}{\mathbf{nQ}} \right)$$

where T is the number of treatment levels, J is the number of the categories of the Q categorical covariates and n is the number of units.

Usage

The %BALANCE Macro takes the following named parameters. The arguments may be listed within parentheses in any order, separated by commas. The %BALANCE macro uses the %DUMMY macro (Friendly,2001) [4] to create the disjunctive table given the Q pre-treatment categorical covariates. The %DUMMY macro must be downloaded from http://www.datavis.ca/sasmac/dummy.html and saved in the specified PATH=.

The %DUMMY macro must be parametrized as follows:

%DUMMY(data=_&cluster, out=disj_&cluster, var=&balance_var &treat, base=_last_, prefix=, format=, name=VAL, fullrank=0);

Parameters

LIBRARY=	The name of the SAS library
DATA=	The name of the input dataset. The input dataset contains the categorical covariates, the treatment indicator variable, the ID variable and a variable that indicates the group membership for each unit. A group could be the result of any classification analysis or a bin of a Propensity Score Analysis, conducted separately before running such Macro.
OUT=	The name of the output dataset. For each group it reports the number of units within the group (n), the number of units in the treatment group (n_t1) , the number of units within the control group (n_t2), the group membership indicator (id_clu), the value of the balance measure (GI), the upper limit of the confidence interval (CHI) and the balance result (BALANCE). BALANCE=yes if the GI measure is between 0 and the upper limit of the confidence interval; BALANCE=no otherwise.
FIRSTCLU=	The number of the first group to analyze in the GROUP_VAR. It is a numeric value.
LASTCLU=	The number of the last group to analyze in the GROUP_VAR. It is a numeric value.
GROUP_VAR=	The name of the input variable denoting the units group membership.
BALANCE_VAR=	The name (s) of the input variable(s) to be balance checked. The name(s) may be listed in any order. The variable(s) must be numeric. No missing values are allowed.
Q=	The number of categorical variables on which simultaneously check balance. The variable must be numeric.
TREAT=	The name of the treatment indicator variable. The variable must be numeric.
ALPHA=	The alpha level for the multivariate imbalance test. It is a numeric value.
PATH=	The path where the folder containing datasets and dummy.sas file is located

```
/*****************************************************************************/
/* MACRO NAME: Balance                                                       */
/* TITLE:      Macro to simultaneously test global balance of categorical    */
/*             covariates                                                    */
/*****************************************************************************/
/* AUTHORS:    Camillo, Furio   furio.camillo@unibo.it                       */
/*             D'Attoma, Ida    Ida.dattoma2@unibo.it                        */
/* CREATED:    07 June 2010                                                  */
/* REVISED:    22 November 2010                                              */
/*****************************************************************************/

%MACRO Balance(
                 library=    ,   /* name of the SAS lilbrary                 */
                 data=       ,   /* name of input dataset                    */
                 out=        ,   /* name of output dataset                   */
                 firstclu=   ,   /* the number of the first group to analyze */
                 lastclu=    ,   /* the number of the last group to analyze  */
                 group_var=  ,   /* the name of the classification variable  */
                 balance_var=,   /* the name(s) of variable(s) to be balance
                                    checked                                  */
                 Q=          ,   /* the number of categorical variables to be
                                    balance checked                          */
                 treat=      ,   /* the name of the treatment indicator variable*/
                 alpha=      ,   /* the alpha level for the multivariate test   */
                 path=       ;   /* the path where the dummy.sas macro is saved */
                 );
/*****************************************************************************/
/*                            The loop over groups                           */
/*****************************************************************************/

/*****************************************************************************/
/*                      creates a dataset for each group                     */
/*****************************************************************************/


%do I=&firstclu %to &lastclu ;
    %let cluster=&I;
     data _&cluster;
     set  &library..&data;
       if &group_var=&cluster then output;
     run;
/*****************************************************************************/
/*          counts treatment and control units within each group             */
/*****************************************************************************/


     proc freq data=_&cluster noprint;
     table &treat/out=freq_&cluster;
     run;

     proc transpose data=freq_&cluster out=trasp_&cluster
     name=count
     prefix=n_t;
     run;

     data trasp_&cluster;
```

```
     set  trasp_&cluster;
     if count ne "COUNT" then delete;
     run;

     data trasp_&cluster;
     set  trasp_&cluster;
     id_clu=&cluster;
     drop count _label_;
     run;
%end;

 %do n=&firstclu %to &lastclu;
   %let nidcluster=&n;
   data nt;
   set
   %do h=1 %to &nidcluster;
   trasp_&h
   %end;
   ;
   run;
 %end;
   ;
 run;

/*****************************************************************************/
/*      creates a disjunctive table for each group using the dummy.sas macro     */
/*          Produces as output for each group a disj_&cluster dataset            */
/*****************************************************************************/

 %do J=&firstclu %to &lastclu ;
      %let cluster=&J;
      %include &path;
 %end;
  run;


%do I=&firstclu %to &lastclu ;
      %let cluster=&I;
            data disj_&cluster;
            set  disj_&cluster;
            drop &balance_var &treat &group_var;
%end;
      run;

%do k=&firstclu %to &lastclu ;
      %let cluster=&k;
            data t_&cluster;
            set  disj_&cluster;
            keep &treat:;
%end;
      run;

/*****************************************************************************/
/*                      Compute the GI measure for each group                */
/*    Z includes the Q original categorical covariates in disjunctive form       */
/*                      L includes the t treatment levels                    */
/*                Q denotes the number of categorical  variables             */
/*                n denotes the number of rows in the disjunctive table(s)       */
```

```
/*                c denotes the number of columns of the disjunctive table(s)    */
/*             j denotes the number of levels of the Q categorical covariates    */
/*                              B is the Burt table                               */
/*                             band is the Burt band                             */
/*             kt denotes the number of treatment and comparison cases           */
/*                      Between denotes the final GI formula                     */
/*****************************************************************************/

%do I=&firstclu %to &lastclu ;
      %let cluster=&I;
proc iml;
use disj_&cluster;
read all into X;
use t_&cluster;
read all into L;
n=nrow(X);
C=ncol(X);
levelt=ncol(L);
j=C-levelt;
Z=X[,2:C-levelt];
A=T(Z);
B= A*Z;
T=X[,c-(levelt-1):c ];
band=A*T;
band_2=band#band;
kt=T[+,];
burt_diag=VECDIAG(B);
inversa=1/burt_diag;
frac_t=band_2#inversa;
sum_t=frac_t[+,];
invkt=1/kt;
bet1=invkt#sum_t;
invQ=1/&Q;
bet2=invQ*bet1;
bet3=bet2[,+];
gdl=(levelt-1)*(j-1);
Between=bet3-1;

/*****************************************************************************/
/*                                  chi-square quantile                      */
/*                     cinv(p,df,nc) returns the pth quantile, 0<p<1,        */
/*                  of the chi-square distribution with degrees of freedom df.    */
/*      If the optional non-centrality parameter nc is not specified, nc=0.      */
/*             nc is the sum of the squares of the means. Examples:           */
/*                              y=cinv(0.95,5);   z=cinv(0.95,5,3);           */
/*****************************************************************************/

chi=cinv(1-&alpha,gdl)/(n*&Q);
if (Between > chi) | (Between < 0) then Balance='Unbalanced'; else
Balance='Balanced';
inertia_tot=(j/&Q)-1;
within=inertia_tot-Between;
MIC=(1-(within/inertia_tot))*100;
chi=cinv(1-&alpha,gdl)/(n*&Q);
results_&cluster =    Between || chi;
  cname = {"GI" "CHI"  };
  create GI_&cluster from results_&cluster [ colname=cname ];
  append from results_&cluster;
```

```
quit;
data GI_&cluster;
set  GI_&cluster;
length Balance $ 17;
if GI le CHI then Balance="Yes";
else if (GI=0 and CHI='.') then Balance="no common support";
else Balance="No";
run;
data GI_&cluster;
set  GI_&cluster;
if  Balance="no common support" then GI='.';else GI=GI;
run;
%END;
run;

/*****************************************************************************/
/*                            Create output dataset                          */
/*****************************************************************************/

%do n=&firstclu %to &lastclu;
%let nidcluster=&n;
data GI;
set
%do h=1 %to &nidcluster;
GI_&h
%end;
;
run;
%end;
;
run;

/*****************************************************************************/
/*                      Delete  temporary output dataset(s)                  */
/*****************************************************************************/

%do n=&firstclu %to &lastclu;
%let nidcluster=&n;
proc datasets;
delete GI_&nidcluster freq_&nidcluster trasp_&nidcluster
t_&nidcluster _&nidcluster _cats_ disj_&nidcluster;
%end;
run;

data GI;
set  GI;
id_clu=_n_;
run;

proc sort data=GI;
  by id_clu;
run;

proc sort data=Nt;
  by id_clu;
run;

data &library..&out;
```

```
merge Nt(in=A) Gi(in=B);
 by id_clu;
if A and B;
run;

data &library..&out;
set  &library..&out;
n=sum(n_t1, n_t2);
run;

%MEND Balance;

%Balance;
```

Example: the output dataset

Libref.Balance_results

	n_t1	n_t2	id_clu	GI	CHI	Balance	n
1	4	18	1	0.0684900426	0.1694093371	Yes	22
2	4	19	2	0.0467372905	0.1523628248	Yes	23
3	1	14	3	0.2293792517	0.233622998	Yes	15
4	10	39	4	0.0815735385	0.0805605146	No	49
5	8	14	5	0.0281895214	0.1490540725	Yes	22
6	3	39	6	0.0606602389	0.083436785	Yes	42
7	4	13	7	0.0238404758	0.1658723295	Yes	17
8	5	29	8	0.0488967805	0.1096178063	Yes	34
9	4	16	9	0.0633293722	0.1525586504	Yes	20
10	4	35	10	0.0317939736	0.0955642414	Yes	39
11	5	20	11	0.1089966168	0.1490802166	Yes	25
12	26	.	12	.	.	no common support	26
13	5	17	13	0.0415849673	0.1694093371	Yes	22
14	14	43	14	0.0098278653	0.053529351	Yes	57
15	12	.	15	.	.	no common support	12
16	19	47	16	0.0268490851	0.0598100791	Yes	66
17	6	14	17	0.0217037557	0.14099148	Yes	20
18	4	18	18	0.024072527	0.1386896822	Yes	22

Suppose you have 18 subgroups on which test balance of covariates among treatment groups before estimate a treatment effect. The balance macro allows you to understand where an effect could be estimated in a easy way: the output dataset of %BALANCE Macro gives you information about balance for each subgroup.

In particular, the output result displays for each group the number of units within treatment group (n_t1), the number of units in the control group (n_t2), the number of units within each subgroup (n), the group membership indicator (id_clu), the GI measure (GI), the upper limit of the confidence interval (CHI) and the balance result (BALANCE). Balance equals ‘yes’ if the GI measure is lower than the upper limit of the confidence interval; otherwise, it equals ‘no’ if the GI measure is greater than the upper limit of the confidence interval. Finally, Balance equals ‘no common support’ when the GI measure and its test cannot computed because there are only treated units or only controls.

Conclusions

The %BALANCE Macro enables one to measure global balance of categorical covariates and to test it in a multivariate way, thereby overcoming limitation of standard variable-by-variable tests. The advantage of using the GI measure is that it considers baseline covariates simultaneously, and as such is also able to considers interactions among covariates. In sum, the %BALANCE Macro makes easy to measure treatment effects on balanced subgroups under non-experimental conditions, where a subgroup could be the result of any classification analysis or a bin of a Propensity Score Analysis.

References

[1] SAS/MACRO is a registered trademark or trademark of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

[2] Camillo, F. & I. D'Attoma. (2010). A new data mining approach to estimate causal effects of policy interventions. *Expert Systems with Applications*, 37: 171-181.

[3] D'Attoma, I. & Camillo, F. (2010). A multivariate strategy to measure and test global imbalance in observational studies. *Expert Systems with Applications*. Doi: 10.1016/j.eswa.2010.08.132.

[4] Friendly, M. (2001). dummy.sas: A Macro to create dummy variables. Available at http://www.datavis.ca/sasmac/dummy.html