**Paper 1528-2014**

# Multivariate Ratio and Regression Estimators

Alan Ricardo da Silva, Universidade de Brasília, Dep. de Estatística, Brazil

## ABSTRACT

This paper considers showing **%mre** macro to estimate multivariate ratio estimates. Also, we can use PROC REG to estimate multivariate regression estimates and to show that regression estimates are superior to the ratio estimates.

## 1. INTRODUCTION

Ratio and regression estimators are used when an auxiliary variate X, correlated with the response variable Y, is used to obtain increased precision by taking advantage of the correlation between Y and X (Cochran, 1977). The main problem is that the population mean or total of the X must be known. However, the gain of precision is much superior to the simple random sampling estimator, when the correlation is moderate or high. When more than one covariate is available then we need to use the multivariate ratio or regression estimators. Multivariate ratio is not available in SAS until version 9.3 and multivariate regression can be easily obtained from PROC REG. So, the main objective of this paper is to show **%mre** macro to estimate multivariate ratio estimator and how to obtain multivariate regression estimates from REG procedure.

The paper is organized as follows. In Section 2, the theory about the Ratio and Regression estimators is given. In Section 3, a SAS® macro is presented and in Section 4 I introduce an illustration.

## 2 RATIO AND REGRESSION ESTIMATORS

The ratio estimate of the population mean is (Cochran, 1977):

$$\overline{Y}_R = \frac{\overline{y}}{\overline{x}}\overline{X} = \hat{R}\overline{X} \tag{1}$$

where $\overline{X}$ is the population mean of the covariate *X* and $\overline{y}$ , $\overline{x}$ are the sample means.

The estimated variance of (1) is given by (Cochran, 1977):

$$\hat{Var}(\overline{Y}_R) = \frac{(1-\frac{n}{N})}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{R}x_i)^2}{n-1} \tag{2}$$

where *N* is the population size and *n* is the sample size.

The ratio estimate of the population total is (Cochran, 1977):

$$\hat{Y}_R = \frac{\overline{y}}{\overline{x}}T_X = \hat{R}T_X \tag{3}$$

where $T_X$ is the population total of the covariate *X* and its estimated variance is given by (Cochran, 1977):

$$\hat{Var}(\hat{Y}_R) = \frac{N^2(1-\frac{n}{N})}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{R}x_i)^2}{n-1} \tag{4}$$

We can see that the difference between ratio estimate for the population mean and for population total is only the information about the population mean or population total of the covariate *X*.

In the case of the linear regression estimate, the population mean is calculated by (Cochran, 1977):

$$\overline{Y}_{\text{Reg}} = \overline{y} + B(\overline{X} - \overline{x}) \tag{5}$$

where *B* is the least square estimate and its estimated variance is given by (Cochran, 1977):

$$\hat{Var}(\overline{Y}_{\text{Reg}}) = \frac{(1-\frac{n}{N})}{n}S_y^2(1-\rho^2) \tag{6}$$

where $S_y^2$ is the variance of the *y* and $\rho = S_{XY} / S_X S_Y$ is the correlation between *y* and *x*.

The linear regression estimate of the population total is (Cochran, 1977):

$$\hat{Y}_{\text{Reg}} = N\overline{Y}_{\text{Reg}} \tag{7}$$

and its estimated variance is $N^2$ times the variance of the population mean.

## 2.1 MULTIVARIATE RATIO AND REGRESSION ESTIMATES

In the case of multivariate ratio estimate, Olkin (1958) provides the equations for the population mean as:

$$\overline{Y}_{MR} = w_1 \frac{\overline{y}}{\overline{x}_1} \overline{X}_1 + \ldots + w_p \frac{\overline{y}}{\overline{x}_p} \overline{X}_p \tag{8}$$

where $w = (w_1, \ldots, w_p)$, $\sum w_i = 1$ is a weighting function.

We can estimate *w* by (Olkin, 1958):

$$\hat{w} = \frac{\mathbf{eA}^{-1}}{\mathbf{eA}^{-1}\mathbf{e'}} \tag{9}$$

where $\mathbf{e} = (1, \ldots, 1)$ and $\mathbf{A} = \mathbf{TCT'}$, $\mathbf{C} = (c_{ij}) : (p+1) \times (p+1),\ c_{ij} = S_{ij} / \overline{X}_i \overline{X}_j$

$$\underset{p \times (p+1)}{T} = \begin{pmatrix} 1 & -1 & 0 & \ldots & 0 \\ 1 & 0 & -1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & 0 & 0 & \ldots & -1 \end{pmatrix} \tag{10}$$

The estimated variance of (8) is given by (Olkin, 1958):

$$\hat{Var}(\overline{Y}_{MR}) = \frac{1}{n(\mathbf{e}\hat{\mathbf{A}}^{-1}\mathbf{e'})} \tag{11}$$

The estimated population total is found changing the population mean of $\overline{X}_i$ by the population total $Tx_i$ in (8) and multiplying (11) by $N^2$.

In the case of multivariate regression estimate, Shukla (1965) provides the equations for the population mean as:

$$\overline{Y}_{\text{MReg}} = \overline{y} + \sum_{i=1}^{p} B_i (\overline{X}_i - \overline{x}_i) \tag{12}$$

and its variance is given by (Shukla, 1965):

$$\hat{Var}(\overline{Y}_{\text{MReg}}) = \frac{(1 - \frac{n}{N})}{n} S_y^2 (1 - R^2) \left\{ 1 + \frac{p}{n - p - 2} \right\} \tag{13}$$

where *p* is the number of auxiliary variates $X_i$ and $R^2$ is the square of multiple correlation coefficient of *y* on $(x_1, \ldots, x_p)$.

# 3 SAS® MACRO

The SAS® Macro for multivariate ratio estimator basically uses IML procedure and for multivariate regression estimator we use SQL and REG procedures and the parameters of the macro are:

DATA = specifies the dataset to be analyzed;
Y = specifies the response or dependent variable;
X = specifies the independent or explicative variables;
MX = specifies the population mean of the explicative variables;
TX = specifies the population total of the explicative variables;
POP = specifies the population size for finite population correction (fpc).

You should use the MX parameter when you want to estimate the population mean or the TX parameter when you want to estimate the population total. You can not use (and it is not necessary) both in the same call. When you use the TX parameter you should use the POP parameter also.

```
%mre(data =,y =,x =,Mx =,Tx =,pop =)
```

## 4 ILLUSTRATION

To illustrate the use of **%mre** macro, we consider the data presented by Olkin(1958) about the number of inhabitants in the 200 largest U.S. cities in 1930, excluding the five largest with Y=1950, X1=1940 and X2=1930. A simple random sample of size 50 was taken. The results of the **%mre** macro are as follows.

We can see that the correlation between the variables *p1950* and *p1930*, according Figure 1, is around .97 and between *p1950* and *p1930* is around .95. So, we expect the multivariate ratio estimate be better than univariate ratio estimate and that the covariate *p1940* be better than covariate *p1930*.

Pearson Correlation Coefficients, N = 50
Prob > |r| under H0: Rho=0

| | p1950 | p1940 | p1930 |
|---|---|---|---|
| p1950 | 1.00000 | 0.97524<br><.0001 | 0.95677<br><.0001 |
| p1940 | 0.97524<br><.0001 | 1.00000 | 0.99613<br><.0001 |
| p1930 | 0.95677<br><.0001 | 0.99613<br><.0001 | 1.00000 |

**Figure 1. Correlation between covariates p1950, p1940 and p1930.**

The macro call is:

```
%mre(data=cities,y=p1950,x=p1940 p1930,Mx=148.2 142.0,pop=200);
```

```
Covariates: p1940 p1930

Population Mean
P1940      P1930

148.2        142

Weights

P1940      P1930 Sum of Weights

2.8530993 -1.853099              1

Ratio Estimates

166.35734

SRS Estimate

185.6

Variance of Ratio Estimate

13.78202

Variance of SRS Estimate

603.90612

Deff

2.28%
```

**Figure 1. Multivariate Ratio Estimate using the covariates X1 and X2.**

Using only X1, the macro call is:

```
%mre(data=cities,y=p1950,x=p1940,Mx=148.2,pop=200);
```

```
Covariates: p1940

Population Mean
          P1940

          148.2

Weights

P1940 Sum of Weights

    1               1

Ratio Estimates

      162.46852

SRS Estimate

      185.6

Variance of Ratio Estimate

             34.545934

Variance of SRS Estimate

             603.90612

Deff

5.72%
```

**Figure 2. Univariate Ratio Estimate using the covariate X1.**

Using only X2, the macro call is:

```
%mre(data=cities,y=p1950,x=p1930,Mx=142.0,pop=200);
```

```
Covariates: p1930

Population Mean
          P1930

            142

Weights

P1930 Sum of Weights

    1               1

Ratio Estimates

      160.36996

SRS Estimate

      185.6

Variance of Ratio Estimate

          63.002473

Variance of SRS Estimate

         603.90612

Deff

      10.43%
```

**Figure 3. Univariate Ratio Estimate using the covariate X2.**

In this example 169.9 is the true value of $\overline{Y}$ , the population mean. We can see that the multivariate ratio is closer than real value and has the smaller variance. All ratio estimates has variance smaller than simple random sample (SRS), but the Design Effect (Deff) is increasing when the correlation between variables Y and X is decreasing.

The Multivariate regression estimate can be found using the REG procedure and Equations (12) and (13) as follows.

```
proc reg data=cities outest=reg rsquare;
  model p1950=p1940 p1930;
run;
quit;

data reg;set reg;
 call symput('r2',_RSQ_);
run;%put &r2;

proc sql noprint;
 select mean(p1950) into:yb from cities;
 select mean(p1940) into:Mp1940 from cities;
 select mean(p1930) into:Mp1930 from cities;
quit;
%put &yb &Mp1940 &Mp1930;
```

```
data reg;set reg;
 VARSRS=603.90612;
 Vyreg=VARSRS*(1-&r2);
 Vyreg2=VARSRS*(1-&r2)*(1+2/(50-2-2));
 Yreg=&yb+p1940*(148.2-&Mp1940)+p1930*(142-&Mp1930);
run;

proc print data=reg label noobs;
 var Vyreg Vyreg2 Yreg;
 label   Yreg="Regression   Estimate"   Vyreg="Variance   of
Regression Estimate";
run;
```

The final output is as follows:

| Variance of Regression Estimate | Variance of Regression Estimate with Correction | Regression Estimate |
|---|---|---|
| 12.6444 | 13.1941 | 167.083 |

**Figure 4. Multivariate Regression Estimate using the covariates X1 and X2.**

We can see that multivariate regression estimate is closer than real value and with smaller variance. The output variable "Variance of Regression estimate" ignores the term $\left\{1+\frac{p}{n-p-2}\right\}$ of Equation (13).

A **`%mrege`** macro from program above also was developed in order to facilitate the use of multivariate regression estimators and it is in appendix. The parameters are the same of **`%mre`** macro.

## CONCLUSIONS

This paper showed a simple way to estimate the multivariate regression estimator using the REG procedure and a macro named **%mre** to estimate the multivariate ratio estimator It is not difficult to show that regression estimator has always variance smaller than ratio estimator (Shukla, 1965; Cochran, 1977), as it was seen in the example. However, the ratio estimator is usually used.

## REFERENCES


Cochran, W.G. (1977). *Sampling Techniques*. John Wiley & Sons, 3rd edition.

*Olkin, I .(1958). Multivariate Ratio Estimates for Finite Population. Biometrika, 45, pp. 154-165.*

*Shukla, G. K. (1965). Multivariate Regression Estimate. Annual Conference on Indian Society of Agricultural Statistics.*

## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:


Name: Alan Ricardo da Silva
Enterprise: Universidade de Brasília
Address: Campus Universitário Darcy Ribeiro, Departamento de Estatística, Prédio CIC/EST sala A1 35/28
City, State ZIP: Brasília, DF, Brazil, 70910-900
Work Phone: +5561 3107 3672
Fax: +5561 3107 3672
E-mail: alansilva@unb.br
Web: www.est.unb.br

## APPENDIX I – SAS® MACRO

```
/******** Example of Olkin(1958) ************/

data cities;
input p1930 p1940 p1950;
cards;
670 672 677
104 101 116
50 64 95
292 385 593
130 173 204
55 54 58
102 97 130
54 58 70
52 62 87
71 69 68
55 50 51
900 878 915
47 48 53
79 82 84
50 49 54
115 115 112
55 57 60
113 110 109
65 70 82
64 62 63
65 67 74
46 49 56
148 203 334
115 110 113
62 71 70
260 268 326
68 69 79
59 56 56
451 456 504
116 117 131
58 59 66
328 325 332
781 771 601
```

```
100 101 97
57 54 55
106 112 125
156 152 163
578 587 637
75 78 91
63 65 74
105 108 117
51 46 58
46 51 80
195 194 203
364 387 427
102 101 102
114 111 121
63 63 64
308 319 369
54 59 74
;

*Multivariate Ratio Estimator;
%macro mre(data=,y=,x=,Mx=,Tx=,pop=);
%if &MX= and &TX= %then %do;
 %put ERROR: specifies Population Total or Mean;
%end;
%else %do;
proc iml;
print "Covariates: &x";
use &data;
read all var {&y} into y;
read all var {&x} into x;
p=ncol(x);
n=nrow(x);
%if &MX= %then %do;
 Mx={&Tx};
 print Mx[label="Population Total" colname={&x}];
%end;
%else %do;
 Mx={&Mx};
 print Mx[label="Population Mean" colname={&x}];
%end;

A=j(p,p,0);
do i=1 to p;
```

```
 do j=1 to p;
  ri=y[:]/x[:,i];
  rj=y[:]/x[:,j];
  A[i,j]=(y-ri*x[,i])`*(y-rj*x[,j])/(n-1);
 end;
end;

r=j(1,p,0);
do i=1 to p;
 r[i]=y[:]/x[:,i];
end;

detA=det(A);
e=j(1,p,1);
w=e*inv(A)*(inv(e*inv(A)*e`));
sw=w[+];
print "Weights",,w[label=" " colname={&x}]
sw[label="Sum of Weights"];
*print r[colname={&x}];
b=(w#r);
*print b[colname={&x}];
Yr=(w#r)*Mx`;
%if &MX= %then %do;
 yb=&pop*y[:];
%end;
%else %do;
 yb=y[:];
%end;
print Yr[label="Ratio Estimates"],, yb[label="SRS
Estimate"];
%if &pop= %then %do;
 %if &MX= %then %do;
  VarYr=&pop**2/(n*(e*inv(A)*e`));
  VarAAS=&pop**2*(y-y[:])`*(y-y[:])/((n-1)*n);
 %end;
 %else %do;
  VarYr=1/(n*(e*inv(A)*e`));
  VarAAS=(y-y[:])`*(y-y[:])/((n-1)*n);
 %end;
%end;
%else %do;
%if &MX= %then %do;
VarYr=&pop**2*(1-n/&pop)/(n*(e*inv(A)*e`));
```

```
VarAAS=&pop**2*(1-n/&pop)*(y-y[:])`*(y-y[:])/((n-1)*n);
%end;
 %else %do;
  VarYr=(1-n/&pop)/(n*(e*inv(A)*e`));
  VarAAS=(1-n/&pop)*(y-y[:])`*(y-y[:])/((n-1)*n);
 %end;
%end;
Deff=VarYr/VarAAS;
print VarYr[label="Variance of Ratio Estimate"],,
VarAAS[label="Variance of SRS Estimate"],,
Deff[label="Deff" format=percent10.2];
%end;
quit;
%mend mre;


*Multivariate Regression Estimator;
%macro mrege(data=,y=,x=,Mx=,Tx=,pop=);
%let nvar=%eval(%sysfunc(length(&x))-
%sysfunc(length(%sysfunc(compress(&x))))+1);
%put &nvar;
%if &MX= and &TX= %then %do;
%put ERROR: specifies Population Total or Mean;
%end;
%else %do;
proc reg data=&data outest=reg rsquare aic;
model &y=&x;
run;
quit;
data reg;
 set reg;
 call symput('r2',_RSQ_);
run;
%put &r2;
proc sql noprint;
 select mean(&y) into:yb from cities;
 select count(*) into:nobs from cities;
 %do i=1 %to &nvar;
  select mean(%scan(&x,&i)) into:Mx&i from &data;
 %end;
quit;
ods output statistics=varSRS;
```

```
proc surveymeans data=&data %if &pop ne %then
%do;total=&pop%end; var;
var &y;
run;
data varSRS;
 set varSRS;
 call symput('varSRS',var);
run;
data reg;set reg;
VARSRS=&varSRS;
%if &MX= %then %do;
VARSRS=&pop**2*&varSRS;
%end;
Vyreg=VARSRS*(1-&r2);
Vyreg2=VARSRS*(1-&r2)*(1+&nvar/(sum(_EDF_,_P_)-&nvar-
2));
Yreg=&yb;
%if &MX= %then %do;
Yreg=&yb*&pop;
%end;
%do i=1 %to &nvar;
%if &MX= %then %do;
Yreg=Yreg+%scan(&x,&i)*(%scan(&Tx,&i,' ')-&&Mx&i*&pop);
%end;
%else %do;
Yreg=Yreg+%scan(&x,&i)*(%scan(&Mx,&i,' ')-&&Mx&i);
%end;
%end;
run;
proc print data=reg label noobs;
var Vyreg Vyreg2 Yreg;
label Yreg="Regression Estimate" Vyreg="Variance of
Regression Estimate"
Vyreg2="Variance of Regression Estimate with
Correction";
run;
%end;
%mend mrege;
```