

Developing A Parameter Evaluation Methodology using SAS®

Richard W. Graham, Bureau of the Census, Washington, D.C.

Abstract

In another paper I am presenting here at the conference, *Developing Ratio Edit Parameters with SAS® for the Economic Census*, I detail a sophisticated methodology for developing ratio edit parameters. As presented in that paper, developing parameters for the Economic Census is an arduous computer and labor intensive effort. Many months are devoted to developing and reviewing the edit parameters for nearly 2,900 industry classifications. Even with a full complement of 30 subject-matter survey statisticians to review the parameters, it would be impossible to review all of the parameters generated by the resistant fences methodology in a timely manner.

This paper describes both an interactive and batch-oriented graphical evaluation procedure for evaluating all of the industry and ratio combination parameters. Both of these applications were developed in an effort to significantly reduce the time to review the edit boundaries and the imputation factors. These procedures were developed in SAS using SAS/AF and the Screen Control Language (SCL), Base SAS, and SAS/GRAPH on the PC platform using version 6.12.

Introduction

The task of reviewing for consistency the boundary and imputation parameters for the Economic Census is an arduous one. With 2,866 industries and 555 explicit ratio relationships for which parameters are

developed, the complete set of explicit relationships requiring review reach well into the 100s of thousands.

When developing the ratio edit parameters for the census, previous-census originally-reported (i.e., unedited) data are used in the development of "cold-deck" boundary parameters. And for "cold-deck" imputation factors, previous-census edited reported data are used. This, of course, implies two completely different sets of data on which ratio edit parameters are generated.

Since two different sets of data are used to generate editing parameters, a consistency and evaluation procedure was needed to ensure that the imputation factors fall within the constraints of the ratio edit boundaries. This paper details such a methodology.

For the 1997 Economic Census, specifically the services division industries (i.e., retail and wholesale trade; finance, insurance and real estate industries; service industries; and transportation, communication and public utilities industries), we include two types of imputation methodologies for which parameters are developed. The first methodology is a simple ratio, or industry average, relationship between two variables (e.g., sales and annual payroll). The second methodology employs one of three linear regression models depending on the paired relationship. These three models are the log-linear, the square root-linear, and the untransformed linear regression models.

When editing records through the ratio edit module, each record's defined data items (in pairs) are subjected to the ratio boundary tests to ensure accuracy and consistency of the data. Items that are missing, or items that fail the boundary tests are marked for imputation and forwarded to the imputation algorithms. It is at this point that the abovementioned imputation factors are used to produce estimates for the missing or aberrant data.

To ensure that appropriate imputations occur and that the imputed data are consistent with other edit-passing reported data, the imputation factors must fall within the ratio edit boundaries. Thus, prior to data editing ratio edit boundaries are compared with the imputation factors to ensure that potential imputes would pass the edit boundary constraints.

The Application

To evaluate the parameters we developed a SAS/AF batch procedure which would evaluate all industry and ratio combination parameters for a particular trade area (e.g., retail, wholesale, etc.). To launch the application, the analyst simply clicked on an icon on their desktop which loaded the SAS system and started the application. Figure 1 displays the selection screen of four listboxes. The first listbox requests you to select a library where the parameter data sets reside, the second listbox requests you to select a parameter data set, the third displays the variables contained in the data set and requests a selection of the classification variable to use (i.e., industry), and the fourth lists all of the unique classification values. In the batch procedure, the fourth listbox has no selection capability, rather it is

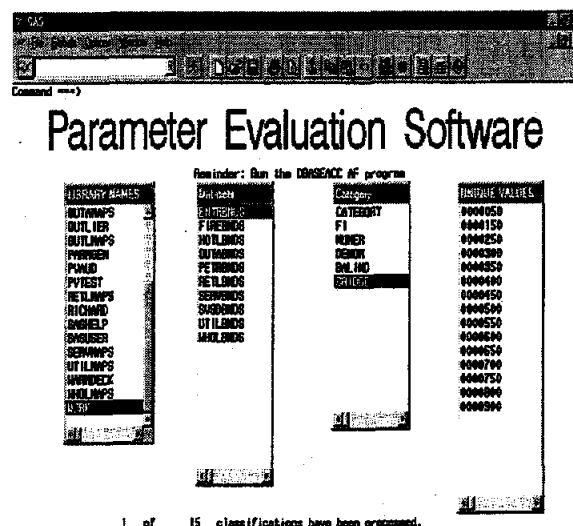


Figure 1

informational. In contrast, the interactive procedure allows individual selections from this listbox.

Also in the batch procedure, at the bottom of the screen, the procedure dynamically updates a message indicating progress of the process.

During the batch procedure the behind-the-scenes process calculates for each ratio an upper boundary value, a lower boundary value, an industry average value, and a regression impute value for independent values ranging from 10 to 10,000. That is, the upper ratio boundary is multiplied by 10, 20, 30, etc. up to 10,000 and stored in a temporary data set. Likewise, the lower boundary and industry average are multiplied by the same factors. These values also are plugged into the appropriate regression model (depending on which relationship) as the independent value. The following code, performed in a do loop, illustrates this process:

```

do i = 1 to numrec;
.
tvar1= 'u' || substr(cnum,2,3)
|| substr(cden,2,3);
tvar2= 'l' || substr(cnum,2,3)
|| substr(cden,2,3);
tvar3= 'a' || substr(cnum,2,3)
|| substr(cden,2,3);
tvar4= 'b' || substr(cnum,2,3)
|| substr(cden,2,3);
tvar5= 'in' || substr(cnum,2,3)
|| substr(cden,2,3);
.
up=GETVARN(dsid2,v3);
lw=GETVARN(dsid2,v4);
av=GETVARN(dsid2,v5);
beta_a=GETVARN(dsid1,v1);
beta_b=GETVARN(dsid1,v2);
infl=GETVARN(dsid1,v3);

submit continue;
data work.parmplot;
set work.parmplot;
    &tvar5=independ;
    &tvar1=(independ*&up)*&infl;
    &tvar2=(independ*&lw)*&infl;
    &tvar3=(independ*&av)*&infl;
    &tvar4=(&beta_a+&beta_b
        *independ)*&infl;
run;
endsubmit;
end;

```

Once all values are calculated and stored in the temporary data set, the evaluation takes place. In this part of the procedure, the calculated boundaries and imputes are tested to determine whether any of the imputes fall outside of the boundary values.

```

do i = 1 to numvar;
    rc=INSERTC(varn,VARNAME(dsid,i),-1);
end;
do i = 2 to numvar-4 by 5;
    rc=REWIND(dsid);
    itm1=getitemc(varn,i);
    itm2=getitemc(varn,i+1);
    itm3=getitemc(varn,i+2);
    itm4=getitemc(varn,i+3);
    itm5=getitemc(varn,i+4);
    itm6='C' || substr(itm1,3,3);
    itm7='C' || substr(itm1,6,3);
do n = 1 to 1000;

```

```

rc=FETCHOBS(dsid,n);
val1=getvarn(dsid,i+1);
val2=getvarn(dsid,i+2);
val3=getvarn(dsid,i+3);
val4=getvarn(dsid,i+4);
if val4>0 then do;
    if val3 > val1 or val3 < val2 or
        val4 > val1 or val4 < val2 then do;
        rc=FPOS(FID,1);
        rc=FPUT(FID,text4);
        rc=FPOS(FID,8);
        rc=FPUT(FID,itm6||' '||itm7);
        rc=FAPPEND(FID);
        n = numobs + 1;
        wrtind='Y';
        link plot;
    end;
end;
end;

```

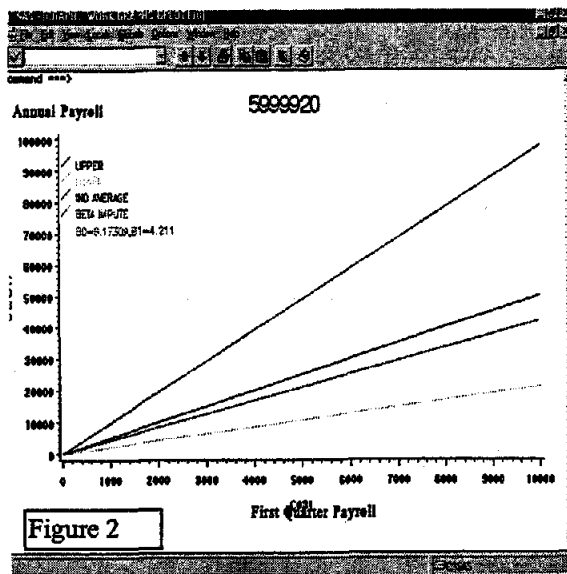
In the above, val1 is the upper bound, val2 is the lower bound, val3 is the industry average, and val4 is the beta impute. As can be seen from the code, if any impute value falls outside the boundaries, the routine passes the information to a plot procedure, *link plot*, where using PROC GPLOT we create a plot of all of the values for the particular failing industry and paired ratio combination. These graphs are produced only for failing relationships and are stored in a permanent catalog for subsequent review using PROC GREPLAY.

With the interactive procedure, the front-end SAS/AF screen is the same as in Figure 1 with two minor differences. The first is that the fourth listbox allows for selection of a particular industry. Second, the progress message is not displayed. Rather, once the processing for the selected industry is complete, you are asked if you want to process another industry. If yes is entered, the previous fourth listbox selection is cleared allowing for selection of a new industry. If no is selected, the procedure executes the PROC GREPLAY pointing to

the WORK library and temporary catalog containing the industry/ratio plots.

Upon selection of an industry in the interactive system, the basic process described above is executed. However, there are two differences. The first is that no evaluation occurs in this procedure. Only the values are calculated for each ratio pair. The second difference is that all ratio relationships are sent to the PROC GPLOT procedure, not just the failing relationships as in the batch system. This allows the reviewer the ability to evaluate all relationships graphically for a particular industry.

Figure 2 displays a plot for an industry and ratio pair combination. This plot graphs the points for the annual payroll to first quarter payroll ratio. Each of the 1,000 values,



ranging from 10 to 10,000, are the pseudo first quarter payroll values and plotted on the X, or horizontal axis. These values are multiplied by the various boundary parameters and imputation factors and plotted as the imputed annual payroll value

on the Y, or vertical, axis.

One of the nice features of the PROC GPLOT procedure is the ability to annotate user-specified legends. This is accomplished through the following code:

```
submit continue;
  goptions nodisplay device=win;
  title1 '          &text4';
  symbol1 c=red v=point i=join;
  symbol2 c=gold v=point i=join;
  symbol3 c=blue v=point i=join;
  symbol4 c=green v=point i=join;
  proc gplot data=work.parmplot
    gout=work.graphs;
  plot &itm2*&itm1=1 &itm3*&itm1=2
    &itm4*&itm1=3 &itm5*&itm1=4 / overlay
    annotate=parmgen.legend1;
    label &itm2='&itm6' &itm1='&itm7';
  run;
  quit;
  goptions display;
endsubmit;
```

Note the /OVERLAY option on the PLOT statement. The overlay allows you to annotate information from a data set. In this case, PARMGEN.LEGEND1 contains the necessary information about the graph. One particularly nice feature about this is the ability to dynamically update the legend given the ratio pair. Referring again to Figure 2, notice the beta factor values associated with the regression imputation model. Figure 3 displays the LEGEND1 data set.

Within the LEGEND1 data set the last five records provide the necessary information for annotating the beta factor values. These values are obtained from the beta factor imputation parameter file.

OBS	TEXT	FUNCTION	STYLE	X	Y	COLOR	POSITION
1		move	simplex	2	1	2.0	30.0
2	UPPER	draw	solid	a	a	1.5	0.0 red
3		move	simplex	a	a	1.0	-1.0
4		contxt	simplex				
5	UPPER	label	simplex				red
6		move	simplex	2	1	2.0	85.0
7	LOWER	draw	solid	a	a	1.5	0.0 gold
8		move	simplex	a	a	1.0	-1.0
9		contxt	simplex				
10	LOWER	label	simplex				gold
11		move	simplex	2	1	2.0	80.0
12	IND AVERAGE	draw	solid	a	a	1.5	0.0 blue
13		move	simplex	a	a	1.0	-1.0
14		contxt	simplex				
15	IND AVERAGE	label	simplex				blue
16		move	simplex	2	1	2.0	75.0
17	BETA INPUT	draw	solid	a	a	1.5	0.0 green
18		move	simplex	a	a	1.0	-1.0
19		contxt	simplex				
20	BETA INPUT	label	simplex				green
21		move	simplex	2	1	2.0	70.0
22	00-0.32629,01=1.14672	move	simplex	a	a	1.5	0.0 green
23		move	simplex	a	a	1.0	-1.0
24		contxt	simplex				
25	00-0.32629,01=1.14672	label	simplex				green

Figure 3

Conclusions

The frame-based applications described in this paper are fully-functional production SAS applications. The use of these applications has brought about efficiencies for review and enhanced the accuracy of the ratio edit parameter development.

Although these programs are tailored to a specific application, with some minor modifications, the procedures could be applied to other similar applications.

The full SCL code behind these applications will be available at the conference.

SAS, SAS/AF, and SAS/GRAPH are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries.

® indicates USA registration.

References

Amadeus Software Ltd. (1997), Building Frame Entry Applications in SAS® Software. Oxfordshire, UK

SAS Institute Inc.(1987), SAS/AF Guide for Personal Computers, Version 6 Edition. Cary, NC: SAS Institute.

SAS Institute Inc.(1990), SAS/GRAPH Software: Reference, Version 6, First Edition, Volumes 1 & 2. Cary, NC: SAS Institute.

SAS Institute Inc.(1990), Procedures Guide, Version 6, Third Edition. Cary, NC: SAS Institute.

Questions or comments to:

Richard W. Graham
Department of Commerce
Bureau of the Census
Room 2517 Bldg. 3
Washington, DC 20233

Phone: (301) 457-2606
Fax: (301) 457-1777
e-mail: richard.w.graham@ccmail.census.gov