Example and Tutorial for SynLBD Validation

Lars Vilhuber, Jorgen Harris, and Emin Dinlersoz

September 14, 2016

Introduction

We want to provide an example of proper validation criteria, using a fake dataset as our input.

This article uses the StatRep LaTeX package. The package is available for download at `http://support.sas.com/StatRepPackage`. To generate this document,

```
pdflatex file.tex
sas file_SR.sas
pdflatex file.tex
pdflatex file.tex
```

or see Appendix .

The Fake Dataset

We create a dataset that approximates, very roughly, the characteristics of the establishment and employment distribution of a real dataset such as the SynLBD and LBD. We use this so that the document is maximally portable, given distribution restrictions for both SynLBD and LBD.

```
/* draw estab count distribution across industries*/
/* use log normal */
data industries;
  do industry = 1 to &indcnt.;
  estabs= exp(ranuni(&seed1.)*
                (log(&maxestabcnt.)
                -log(&minestabcnt.))
                + log(&minestabcnt.));
  output;
  end;
  run;
```

```
/* now draw employment for each estab in each industry */

data fakelbd;
  set industries;
  by industry;
  drop i;
  do lbdnum=100000*industry+1 to 100000*industry+estabs;
    do year=1 to 3;
       emp= exp(ranuni(&seed2.)
         *(log(&maxemp.)-log(&minemp.))
         + log(&minemp.));
    payroll  = emp*30*ranuni(3153);
    output;
  end; end;
run;
```

We can assess the distributions, first of establishments:

Figure 1: Statistics on establishments

```
The UNIVARIATE Procedure
Variable:  estabs

             Basic Statistical Measures

    Location                    Variability

Mean     2063.382     Std Deviation              2446
Median   1113.014     Variance                5980753
Mode          .       Range                      8165
                      Interquartile Range        2408
```

Let's have a look at the distribution of employment:

Figure 2: Statistics on employment

```
The UNIVARIATE Procedure
Variable:  emp

             Basic Statistical Measures

    Location                    Variability

Mean     3942.874     Std Deviation              8203
Median    203.816     Variance                67288482
Mode          .       Range                     41998
                      Interquartile Range        2897
```

The number of establishments across industries varies, which will lead to difficulties if we want to obtain results for certain industries:

Figure 3: Number of obs per industry

The FREQ Procedure

industry	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	9018	3.64	9018	3.64
2	1050	0.42	10068	4.07
3	7872	3.18	17940	7.25
4	1302	0.53	19242	7.77
5	1632	0.66	20874	8.43
6	483	0.20	21357	8.63
7	354	0.14	21711	8.77
8	3495	1.41	25206	10.18
9	1656	0.67	26862	10.85
10	360	0.15	27222	11.00
11	11586	4.68	38808	15.68
12	24849	10.04	63657	25.72
13	5931	2.40	69588	28.11
14	1428	0.58	71016	28.69
15	1416	0.57	72432	29.26
16	1710	0.69	74142	29.95
17	3972	1.60	78114	31.56
18	24105	9.74	102219	41.29
19	2772	1.12	104991	42.41
20	7971	3.22	112962	45.63
21	21888	8.84	134850	54.47
22	672	0.27	135522	54.75
23	15495	6.26	151017	61.01
24	8454	3.42	159471	64.42
25	366	0.15	159837	64.57
26	7971	3.22	167808	67.79
27	12480	5.04	180288	72.83
28	396	0.16	180684	72.99
29	702	0.28	181386	73.27
30	3999	1.62	185385	74.89
31	4857	1.96	190242	76.85
32	927	0.37	191169	77.23
33	19410	7.84	210579	85.07
34	2325	0.94	212904	86.01
35	489	0.20	213393	86.20
36	23148	9.35	236541	95.55
37	3519	1.42	240060	96.98
38	810	0.33	240870	97.30
39	3456	1.40	244326	98.70
40	3219	1.30	247545	100.00

Project 1: Analysis that meets validation requirements

This example is a project where the analysis, and the validation request, meet the requirements. This project is interested in ... First, the researcher prepares the data:

```
/*Prepare data*/
/* program: 01_prepdata.sas */
data analysis1;
set fakelbd;
by industry lbdnum year;
wage = payroll/emp;
if first.lbdnum then do;
lagE = .;
lagp = .;
lagw = .;
end;
else                  do;
lagE = lag(emp);
lagp=lag(payroll);
lagw = lag(wage);
end;
empgrowth = emp/lage;
wagegrowth= wage/lagw;
run;
```

Then, the regression of interest to the researcher is run:

```
/*Regression of interest*/
proc reg data=analysis1;
by industry;
where industry le 2;
model empgrowth = lagE lagw;
output out=obsds1 r=inc;
ods output parameterestimates=param1;
run;
ods trace off;
```

The result of the regression is the following output (here for the first industry only):

Figure 4: Project 1: Parameter estimates

```
industry=1

The REG Procedure
Model: MODEL1
Dependent Variable: empgrowth

                        Parameter Estimates

                     Parameter       Standard
Variable     DF       Estimate          Error    t Value    Pr > |t|

Intercept     1      504.27831       47.07337      10.71      <.0001
lagE          1       -0.02017        0.00272      -7.41      <.0001
lagw          1       -4.56316        2.64965      -1.72      0.0851
```

In order to prepare for validation and disclosure avoidance review of the *confidential* analysis, the researcher must determine the effective sample size of each parameter in terms of establishments and total observations. Ideally, this is provided as an "augmented" results table that allows the Census Bureau disclosure officer to assess the whole picture. The following code will generate that information:

```
proc sql;
create table discreview1 as
select industry,count(distinct lbdnum) as nEstabs,count(*) as nObs
from obsds1
where inc ne .
group by industry
;quit;
data discreview1;
  merge discreview1(in=_a)
                param1(in=_b);
   by industry;
 run;
```

Finally, in order to prepare the validation request, as well as the release request for the synthetic data results, *both* tables are written out as CSV files:

```
/*Export validation table and sample size table*/
proc export data=param1 file="./validationtable1.csv" dbms=csv replace;
run;
```

Obs	industry	Dependent	Model	Effect	DF	Estimate	StdErr	tValue	Probt
1	1	empgrowth	MODEL1	Intercept	1	504.27831	47.07337	10.71	<.0001
2	1	empgrowth	MODEL1	lagE	1	-0.02017	0.00272	-7.41	<.0001
3	1	empgrowth	MODEL1	lagw	1	-4.56316	2.64965	-1.72	0.0851
4	2	empgrowth	MODEL1	Intercept	1	305.47283	116.69556	2.62	0.0090
5	2	empgrowth	MODEL1	lagE	1	-0.01798	0.00655	-2.75	0.0062
6	2	empgrowth	MODEL1	lagw	1	5.82236	6.66205	0.87	0.3824

```
proc export data=discreview1 file="./discreview1.csv" dbms=csv replace;
run;
```

Obs	industry	nEstabs	nObs	Model	Dependent	Variable
1	1	3006	6011	MODEL1	empgrowth	Intercept
2	1	3006	6011	MODEL1	empgrowth	lagE
3	1	3006	6011	MODEL1	empgrowth	lagw
4	2	350	700	MODEL1	empgrowth	Intercept
5	2	350	700	MODEL1	empgrowth	lagE
6	2	350	700	MODEL1	empgrowth	lagw

Obs	DF	Estimate	StdErr	tValue	Probt
1	1	504.27831	47.07337	10.71	<.0001
2	1	-0.02017	0.00272	-7.41	<.0001
3	1	-4.56316	2.64965	-1.72	0.0851
4	1	305.47283	116.69556	2.62	0.0090
5	1	-0.01798	0.00655	-2.75	0.0062
6	1	5.82236	6.66205	0.87	0.3824

In fact, if using LaTeX, the researcher could attach all programs and output from the synthetic data to the validation request, and submit it:

- 01_synlbd_validation_SR.sas 

- validationtable1.csv 
- discreview1.csv 

Note that the result tables as shown here would be based on the synthetic data, and both `discreview1.csv` and `validationtable1.csv` would be released to the researcher. However, the results validated against the confidential data would differ from those reported in Figure **??**, and the `discreview1.csv` file generated from the confidential data, using the code submitted by the researcher, would NOT be released.

Appendix: How to compile a StatRep document

When you use the StatRep LaTeX package, you use the following four-step process to create an executable document that enables you to ensure that your research results are reproducible:

1. Create your LaTeX document so that it contains your text, data, and SAS code.
2. Compile your document with pdfLaTeX to generate the SAS program.
3. Execute the SAS program to capture your output. For each code block in your document, SAS creates a SAS Output Delivery System (ODS) document that contains the resulting output.

   For each output request in your document, SAS replays the specified output objects to external files. All your requested output is generated and captured when you execute the generated SAS program.
4. Recompile your LaTeX document. In this step, the requested outputs are embedded in the resulting final PDF document.

   You might need to repeat this step so that LaTeX can measure the listing outputs to ensure that they are framed appropriately.