

Example and Tutorial for SynLBD Validation

Lars Vilhuber, Jorgen Harris, and Emin Dinlersoz

For comment only - not final
October 6, 2016

1 Introduction

The Synthetic Longitudinal Business Database (SynLBD) was created to allow researchers to carry out research and test hypotheses on the LBD without compromising its security or confidentiality. Because the SynLBD is similar in structure and statistical properties to the LBD, it can be used to generate results that approximate the LBD, and can be used to test code that can be securely run on the LBD by Census staff. Researchers can answer questions using the SynLBD either by requesting *validation* of their results on the LBD, or by simply requesting *export* of their SynLBD results only. This memo gives examples of extracts that meet validation requirements, that fail to meet validation requirements but meet export requirements, and that fail to meet validation and export requirements.

Validation ultimately leads to the release of results based on confidential data. The release of such results must satisfy the confidentiality requirements that an analysis executed in a Federal Statistical Research Data Center (FSRDC) would also need to satisfy (see [RDC Clearance Request Memo](#) and [RDC Researcher Handbook](#)). This limits what can be released, and needs to be taken into account by the researcher. In essence, validation should be limited to

- Simple descriptive statistics (e.g. 1-2 tables describing the sample), as would be directly printable in an article
- Tabular data is highly discouraged in general. This includes more complex or expansive summary statistics, and detailed moments of the data
- Model parameters for a limited number of coefficients and models.
- Parameters for dichotomous (dummy) variables are treated as tabular data would be, and need special treatment. In general, parameters derived from a small number of observations are discouraged.
- If summaries of many (100s) models or summary statistics are of interest (evolution of a parameter across hundreds of specifications, the distribution of the variance or IQR across many cells), then the validation request concerns the summary of those models and statistics, not the models and summary statistics themselves

In general, a good heuristic is: if you would print it as part of your paper, it is likely OK; if it looks like “raw data”, in particular if it will serve as input for some second-stage computation, it is not OK.

Export requirements concern only the results based on the synthetic data. Disclosure risk plays no role here, but the unauthorized distribution of the synthetic data itself is of issue. A similar heuristic as before applies: if it looks like “data”, it is not exportable: Straight subsets of the data, detailed summary tables of large parts of the data are not exportable. However, a limited number of results based on, for example, small subsamples, would pose no problem. Again, model parameters generally pose no problem and are

directly provided to the researcher. Large tables of summary statistics may involve some evaluation, but basic descriptive statistics of the type one would print in a published paper are not a problem.

2 Setup

The results in this exercise use a simple artificial dataset that loosely mimics the structure of the LBD. In this dataset, 1,000,000 establishments are generated with an industry (drawn from an exponential distribution and rounded up – what exactly is drawn here? No variable mentioned. Number of establishments in each industry is exponentially distributed, I believe?), and employment and payroll for each of three years (again drawn from an exponential distribution). As a result, there are lots of firms in industry 1 and many of them are lots of small firms, but few firms in industry 70, but with large numbers of employees.

This article uses the StatRep L^AT_EX package. The package is available for download at <http://support.sas.com/StatRepPackage>. To generate this document,
pdf_latex file.tex
sas file_SR.sas
pdf_latex file.tex
pdf_latex file.tex
or see Appendix A.

3 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=&minyear. to &maxyear.;
      emp= exp(ranuni(&seed2.)
               *(log(&maxemp.)-log(&minemp.))
               + log(&minemp.));
      payroll = emp*30*ranuni(&seed2.);
    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

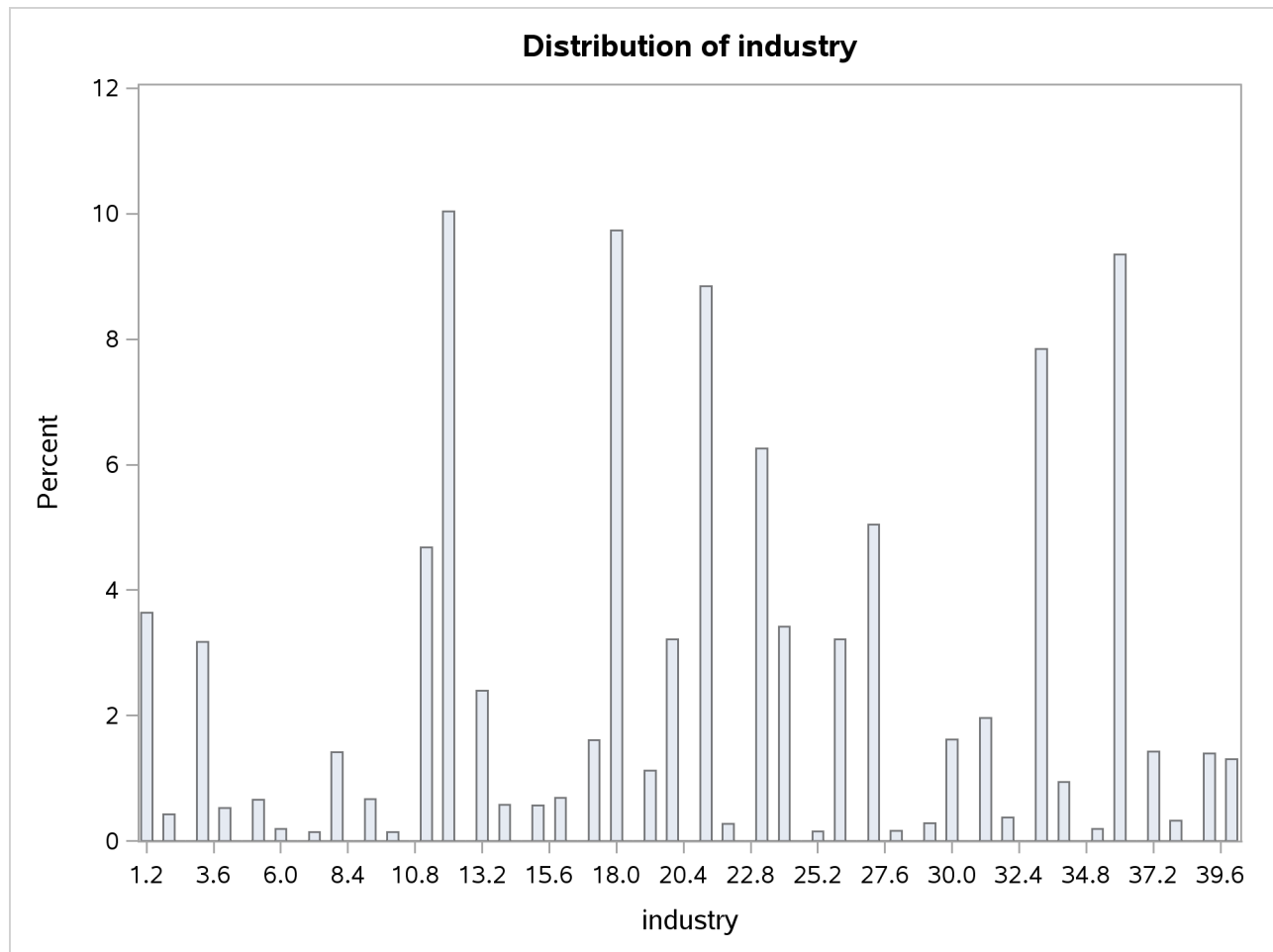
Let's have a look at the distribution of employment:

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

Figure 1 shows that 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



4 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 regressing average wages and number of employees on employment growth in industries 1 and 2. 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(payload);
lagw = lag(wage);
end;
empgrowth = emp/lage;
wagegrowth= wage/lagw;
run;

```

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

```

/*Regression of interest*/
/* program: 02_regression1.sas */
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;

```

Note that the researcher is capturing the output (using SAS ODS) in the file "param1". This will be useful for post-processing and outputting results. 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					
Variable	DF	Parameter Estimate	Standard 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. Note that is *unknown* to her at this point - she only knows those sample sizes in the synthetic data, and most *program* up the checks as a function of the *unknown* confidential numbers. However, the synthetic data are an ideal test ground for the program. Ideally, the results are 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:

```

/* Prepare disclosure avoidance analysis */
/* program: 03_prep_daa.sas */

/* create a count by industry */
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;
/* merge with parameter estimates */
/* flag possible problematic cases */
%let mincount=10;
data discreview1;
    merge discreview1(in=_a)
          param1(in=_b);
    by industry;
    flag_problem=(nEstabs<&mincount.);
    label flag_problem=" (nEstabs<&mincount.)";
run;

```

Finally, we can tabulate (for the synthetic data) the (potentially) problematic cases. Note that we set a particular value as the threshold - the Census Bureau might change that value, upon consultation with the Disclosure Review Board, but the program will still generate the right output.

```

proc freq data=discreview1;
table flag_problem;
run;

```

Figure 5: Potentially problematic parameter estimates

The FREQ Procedure				
(nEstabs<10)				
flag_problem	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	6	100.00	6	100.00

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;

```

Figure 6: Parameters to be released




		D		V		E		S		t		P	
i		e		a		s							
n		p		r		t							
d		e		i		i		t		V		P	
u		n		a		m		d		a		r	
s		o		d		a		E		l		o	
0		t		e		b		r		u		b	
b		r		e		l		D					
r		e		n		l		F					
s		y		l		t		e		r		e	
1	1	MODEL1	empgrowth	Intercept	1	504.27831	47.07337	10.71	<.0001				
2	1	MODEL1	empgrowth	lagE	1	-0.02017	0.00272	-7.41	<.0001				
3	1	MODEL1	empgrowth	lagw	1	-4.56316	2.64965	-1.72	0.0851				
4	2	MODEL1	empgrowth	Intercept	1	305.47283	116.69556	2.62	0.0090				
5	2	MODEL1	empgrowth	lagE	1	-0.01798	0.00655	-2.75	0.0062				
6	2	MODEL1	empgrowth	lagw	1	5.82236	6.66205	0.87	0.3824				

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

		n											
Obs		industry		Estabs		nObs		Model		Dependent		Variable	
1	1	1	3006	6011	MODEL1	empgrowth	Intercept	1					
2	1	1	3006	6011	MODEL1	empgrowth	lagE	1					
3	1	1	3006	6011	MODEL1	empgrowth	lagw	1					
4	2	2	350	700	MODEL1	empgrowth	Intercept	1					
5	2	2	350	700	MODEL1	empgrowth	lagE	1					
6	2	2	350	700	MODEL1	empgrowth	lagw	1					
Obs		Estimate		StdErr		tValue		Probt		flag_		problem	
1	504.27831	47.07337	10.71	<.0001	0								
2	-0.02017	0.00272	-7.41	<.0001	0								
3	-4.56316	2.64965	-1.72	0.0851	0								
4	305.47283	116.69556	2.62	0.0090	0								
5	-0.01798	0.00655	-2.75	0.0062	0								
6	5.82236	6.66205	0.87	0.3824	0								

This regression will meet validation requirements because each regression coefficient is identified from a large number of establishments, and because it is the type of regression that would be included as a table in a journal article.

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

- 02_synlbd_validation_SR.sas 
- validationtable1.csv 
- discreview1.csv 

Note that

- the result tables as shown here are based on the synthetic data, and both `discreview1.csv` and `validationtable1.csv` would be released to the researcher.
- the validation process would use the programs provided by the researcher to re-generate Figure 6, and the `discreview1.csv` file from the confidential data
- the actual results validated against the confidential data (both regression results and counts generated for the disclosure avoidance analysis) will differ from those reported here, and in particular, `discreview1.csv` would not be released to the researcher.
- the validation assumes in addition a reasonable number of results. Running a handful, or even a dozen regressions is OK, running several thousand to subsequently do a specification search on the confidential results is NOT OK, and such validation requests will be denied. The specification search would need to be programmed in a generic way, tested on the synthetic data, and re-run on the confidential data.

5 Project 2: Analysis that fails validation requirements, but meets export requirements

The following example is a project where the analysis meets export requirements, does not meet validation requirements. This project is interested in comparing wage growth by industry in firms with between 45 and 49 employees to wage growth in firms with between 51 and 55 employees. As before, the researcher prepares the data:

```
/*Prepare data*/
data analysis;
set fakedb;
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);

  empgrowth = emp/lage;
  wagegrowth= wage/lagw;
end;

/*Specialized establishment size category*/
if 45 le emp le 49 then empcat = 1;
else if 51 le emp le 55 then empcat =2;
else empcat = .;
run;

data analysis2;
set analysis;
by industry lbdnum year;
if not first.lbdnum;
run;
```

Then, the researcher produces his tabulations of interest:

```
/*Tabulation of interest*/
proc means data=analysis2 n mean variance;
class industry empcat year;
where year ge 2;
var empgrowth wagegrowth;
ods output summary=model2;
run;

proc print data=model2(obs=&printobs.);
where industry=5;
run;
```

The result of the analysis is the following tabulation, for all industries (Figure 11 lists the results for a single industry only):

Figure 7: Project 2: Tabulation

Obs	industry	empcat	year	NObs	VName_ empgrowth	empgrowth_ N
15	5	1	1980	3	empgrowth	3
16	5	1	1981	4	empgrowth	4
17	5	2	1980	3	empgrowth	3
18	5	2	1981	6	empgrowth	6

Obs	empgrowth_ Mean	empgrowth_ Var	VName_ wagegrowth	wagegrowth_ N	wagegrowth_ Mean	wagegrowth_ Var
15	4.2231758381	51.774648502	wagegrowth	3	0.6535617286	0.2265513224
16	0.6196349929	1.3040082891	wagegrowth	4	1.5559219994	1.2914206346
17	1.5038638959	6.0108663764	wagegrowth	3	4.8368784513	47.443898472
18	8.4462319483	167.12739751	wagegrowth	6	2.5389145824	9.122449814

The resulting table is not a simple descriptive statistic: It produces data for very fine cells for a large number of industries. Requesting this kind of summary statistics will lead to a denial of the validation request.¹ Considerations might also be that some of the variables in this table are similar to released Business Dynamics Statistics (BDS), thus preventing release of tabulations based on the underlying LBD.

However, the researcher could still request export of the synthetic data tables. Such a request should still create auxiliary information that will allow the Census Bureau staff the ability to discern whether this is a summary statistic or raw data. We suggest that cells contain statistics for at least 5 establishments, at any level of the data.

```
/* Export analysis --effective sample size of each cell*/
proc sql;
create table exreview2 as
select industry, empcat, year, nObs
from model2
;quit;
```

Figure 8: Number of establishments in cells

The UNIVARIATE Procedure			
Variable: NObs (N Obs)			
Moments			
N	148	Sum Weights	148
Mean	16.5540541	Sum Observations	2450
Std Deviation	18.4653689	Variance	340.969847
Skewness	1.42227177	Kurtosis	1.07390626
Uncorrected SS	90680	Corrected SS	50122.5676
Coeff Variation	111.545902	Std Error Mean	1.51784397

¹Note that the researcher could potentially create such a table as part of an approved FSRDC project, which includes preparing a memo for the Census Bureau's Disclosure Review Board. It is not the role of this document to speculate whether or not such a request would be denied or not.

Figure 8: *continued*

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	75.0
99%	69.0
95%	58.0
90%	49.0
75% Q3	22.5
50% Median	9.0
25% Q1	3.0
10%	1.0
5%	1.0
1%	1.0
0% Min	1.0

Note that the export request contains quite a few cells below the desired threshold. These need to be removed prior to export. For the export request for the synthetic data results, *both* tables are written out as CSV files:

```
/*Export table for export, export analysis table*/
data export2;
set model2;
where (NObs>5);
run;
proc export data=export2
  file="./validationtable2.csv" dbms=csv replace;
run;
```

Figure 9: Exportable data

Obs	industry	empcat2	year	NObs	empgrowth_ Mean
4958	5	1	1980	91	0.1961143588
4959	5	1	1981	77	0.1817454922
4960	5	2	1980	39	0.6280119222
4961	5	2	1981	41	1.0457472759
4962	5	3	1980	22	1.7755350662
4963	5	3	1981	28	2.0884064135
4964	5	4	1980	17	0.3938576963
4965	5	4	1981	13	3.4818185359
4966	5	5	1980	15	1.0447708794
4967	5	5	1981	9	1.5221417879

Obs	empgrowth_ Var	wagegrowth_ Mean	wagegrowth_ Var
4958	0.2247970469	2.6463802629	57.616790384
4959	0.2226578816	2.6556745273	33.143926696
4960	2.1915370921	1.1634892371	1.5183448131
4961	4.3987281536	5.1529027809	161.20627398
4962	10.274344658	2.6502103703	24.987044328
4963	13.927739462	2.0955781898	7.0558739373
4964	0.4514479462	2.231080752	10.881991035
4965	28.541890674	3.1715850584	68.977417516
4966	9.9126891294	1.5910220651	2.9229592333
4967	9.8136263525	10.732404276	805.37346024

```
proc export data=exreview2
  file="./exreview2.csv" dbms=csv replace;
run;
```

Figure 10: Export analysis of Project 2

Obs	industry	empcat	year	NObs
1	1	1	1980	26
2	1	1	1981	19
3	1	2	1980	21
4	1	2	1981	34
5	2	1	1980	4
6	2	2	1980	2
7	3	1	1980	24
8	3	1	1981	21
9	3	2	1980	21
10	3	2	1981	15

After inspection of the output, `discreview2.csv` and `validationtable2.csv` would be released to the researcher.

6 Project 2b: Tabular analysis that fails validation and export requirements:

This example is a project where the analysis does not meet export or validation requirements. Suppose the researcher in Project 2 wanted to look at wage and employment growth by establishment size (rounded to next multiple of five) for all establishment sizes, industries and years. Using the analysis dataset from Section 5, the researcher produces the tabulations of interest:

```
data analysis2b;
set analysis2;
empcat2=int(emp/5)+1;
run;

/*Tabulation of interest*/
proc means data=analysis2b n mean variance;
class industry empcat2 year;
where year ge 2;
var empgrowth wagegrowth;
ods output summary=model2b;
run;

proc print data=model2b(obs=10);
where industry=5;
run;
```

Figure 11 presents the result of the tabulation for a single industry only.

Figure 11: Project 2b: Tabulation

Obs	industry	empcat	year	NObs	VName_ empgrowth	empgrowth_ N
15	5	1	1980	3	empgrowth	3
16	5	1	1981	4	empgrowth	4
17	5	2	1980	3	empgrowth	3
18	5	2	1981	6	empgrowth	6

Obs	empgrowth_ Mean	empgrowth_ Var	VName_ wagegrowth	wagegrowth_ N	wagegrowth_ Mean	wagegrowth_ Var
15	4.2231758381	51.774648502	wagegrowth	3	0.6535617286	0.2265513224
16	0.6196349929	1.3040082891	wagegrowth	4	1.5559219994	1.2914206346
17	1.5038638959	6.0108663764	wagegrowth	3	4.8368784513	47.443898472
18	8.4462319483	167.12739751	wagegrowth	6	2.5389145824	9.122449814

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. This should be 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:

```
/*Disclosure avoidance review--effective sample size of each cell*/
proc sql;
create table discreview2b as
select industry
```

```

,empcat2
,year
,count(distinct lbdnum)as nEstabs
,count(*) as nObs
from analysis2b
where n(wagegrowth,empgrowth)=2
group by industry,empcat2,year
;
quit;

```

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

```

/*Export table for export, disclosure avoidance table*/
proc export data=model2b(drop=empgrowth_N wagegrowth_N)
  file="./validationtable2b.csv" dbms=csv replace;
run;

```

Obs	industry	empcat2	year	NObs	empgrowth_ Mean
4958	5	1	1980	91	0.1961143588
4959	5	1	1981	77	0.1817454922
4960	5	2	1980	39	0.6280119222
4961	5	2	1981	41	1.0457472759
4962	5	3	1980	22	1.7755350662
4963	5	3	1981	28	2.0884064135
4964	5	4	1980	17	0.3938576963
4965	5	4	1981	13	3.4818185359
4966	5	5	1980	15	1.0447708794
4967	5	5	1981	9	1.5221417879

Obs	empgrowth_ Var	wagegrowth_ Mean	wagegrowth_ Var
4958	0.2247970469	2.6463802629	57.616790384
4959	0.2226578816	2.6556745273	33.143926696
4960	2.1915370921	1.1634892371	1.5183448131
4961	4.3987281536	5.1529027809	161.20627398
4962	10.274344658	2.6502103703	24.987044328
4963	13.927739462	2.0955781898	7.0558739373
4964	0.4514479462	2.231080752	10.881991035
4965	28.541890674	3.1715850584	68.977417516
4966	9.9126891294	1.5910220651	2.9229592333
4967	9.8136263525	10.732404276	805.37346024

```

proc export data=discreview2b
  file="./discreview2b.csv" dbms=csv replace;
run;

```

Obs	industry	empcat2	year	n	
				Estabs	nObs
1	1	1	1980	475	475
2	1	1	1981	456	456
3	1	2	1980	223	223
4	1	2	1981	188	188
5	1	3	1980	131	131
6	1	3	1981	116	116
7	1	4	1980	68	68
8	1	4	1981	79	79
9	1	5	1980	68	68
10	1	5	1981	59	59

Such a table, however, is not a simple “descriptive table”. As Figure 12 shows, there are too many cells in this request - over 50,000.

Figure 12: Number of parameters

The UNIVARIATE Procedure			
Variable: nObs			
Moments			
N	55930	Sum Weights	55930
Mean	2.95063472	Sum Observations	165029
Std Deviation	21.1976363	Variance	449.339783
Skewness	36.9862517	Kurtosis	1747.93997
Uncorrected SS	25618065	Corrected SS	25131124.7
Coeff Variation	718.409368	Std Error Mean	0.0896324

Such a request would neither be exported nor validated.

7 Project 3: Regression analysis that fails validation and export requirements:

This example is a project where the analysis does not meet export or validation requirements. The researcher in Section 6 might conclude that their tabular output can be expressed as the output of a regression, using a series of dummy variables. However, this regression must meet the same requirements as the tabulation in Section 6, and thus cannot be exported or validated.

Using the analysis dataset from Section 6, the researcher produces the regression of interest. In order to reduce runtime, we use a random 1/100 sample of the data:

```
data analysis3;
set analysis2b;
testsamp=(ranuni(&seed2.) < 0.01);
run;

/*Regression of interest*/
ods exclude all;
proc glm data=analysis3(where=(testsamp));
class industry empcat2 year;
model empgrowth wagegrowth = empcat2*industry*year/solution;
ods output parameterestimates=model3b;
output out=obsds3 r=inc;
run;
ods exclude none;

/* we parse the Parameter variable to get back the characteristics we are after */
data model3b;
set model3b;
length industry empcat2 year sortorder 8 ;
sortorder=_n_;
industry=input(scan(Parameter,2," "),4.);
empcat2=input(scan(Parameter,3," "),20.);
year=input(scan(Parameter,4," "),4.);
run;
```

Figure 13 presents the regression coefficients for a single industry.

Figure 13: Project 3: Regression

	D	P	E	i	s
	e	a	s	n e	r
	p	r	t	d m	t
	e	a	i	P u p	o
	n	m	m	r s c	y r
	d	e	a	o t a	e d
0	e	t	t	b r t	a e
b	n	e	e	t y 2	r r
s	t	r			
1	empgrowth	Intercept	188.86452	0.0206	. . . 1
2	empgrowth	industr*empcat2*year 1 1 1980	-188.63143	0.0346	1 1 1980 2
3	empgrowth	industr*empcat2*year 1 1 1981	-188.85186	0.0320	1 1 1981 3
4	empgrowth	industr*empcat2*year 1 2 1980	-188.64757	0.0384	1 2 1980 4
5	empgrowth	industr*empcat2*year 1 2 1981	-188.84616	0.0447	1 2 1981 5
6	empgrowth	industr*empcat2*year 1 3 1980	-187.69891	0.0330	1 3 1980 6
7	empgrowth	industr*empcat2*year 1 4 1980	-186.04761	0.1061	1 4 1980 7
8	empgrowth	industr*empcat2*year 1 4 1981	-184.26911	0.1094	1 4 1981 8
9	empgrowth	industr*empcat2*year 1 6 1980	-188.77234	0.1011	1 6 1980 9
10	empgrowth	industr*empcat2*year 1 7 1981	-188.76544	0.1011	1 7 1981 10

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:

```
/*Effective sample size of each cell, meaning # included in each
dummy variable combo, by establishment and observations*/
proc sql;
create table discreview3b as
select industry
      ,empcat2
      ,year
      ,count(distinct lbdnum)as nEstabs
      , count(*) as nObs
from analysis3(where=(testsamp))
where n(wagegrowth,empgrowth)=2
group by industry,empcat2,year
;quit;

/* we sort for merge */
proc sort data=model3b;
by industry empcat2 year;
run;
/* the Discreview dataset contains the COMPLETE set of statistics */
data discreview3b;
merge model3b discreview3b;
by industry empcat2 year;
run;
```

This potential request has two problems: it has far too many observations, and each observations is for too

few entities. Both the export request and the validation request would be turned down.

Figure 14: Number of parameters

The UNIVARIATE Procedure			
Variable: nObs			
Moments			
N	2506	Sum Weights	2506
Mean	1.29369513	Sum Observations	3242
Std Deviation	1.29544579	Variance	1.67817979
Skewness	7.32234948	Kurtosis	64.9097786
Uncorrected SS	8398	Corrected SS	4203.84038
Coeff Variation	100.135322	Std Error Mean	0.02587788
Quantiles (Definition 5)			
Quantile	Estimate		
100% Max	18		
99%	7		
95%	3		
90%	2		
75% Q3	1		
50% Median	1		
25% Q1	1		
10%	1		
5%	1		
1%	1		
0% Min	1		

Figure 15: Count of observations associated with parameters

Obs	Dependent	industry	empcat2	year	Estimate	nObs
1	empgrowth	.	.	.	188.86452	.
2	empgrowth	1	1	1980	-188.63143	5
3	empgrowth	1	1	1981	-188.85186	6
4	empgrowth	1	2	1980	-188.64757	4
5	empgrowth	1	2	1981	-188.84616	3
6	empgrowth	1	3	1980	-187.69891	6
7	empgrowth	1	4	1980	-186.04761	1
8	empgrowth	1	4	1981	-184.26911	1
9	empgrowth	1	6	1980	-188.77234	1
10	empgrowth	1	7	1981	-188.76544	1

A 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 $\text{pdf}\text{\LaTeX}$ 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.