Using the New SURVEY Procedures from a Modeling Perspective

Jonas V. Bilenas, JP Morgan Chase, Wilmington, DE

ABSTRACT

An introduction to the use of PROC SURVEYSELECT, SURVEYREG and SURVEYLOGISTIC to illustrate applications in modeling. We will show how efficient it is to use SURVEYSELECT to generate BOOTSTRAP data that is far superior than having hold-out validation samples. A number of examples will be shown detailing the SURVEYSELECT procedure. We will conclude with how to use SURVEYREG and SURVEYLOGISTIC when you have sampled data to calculate the correct p-values for modeling coefficients.

INTRODUCTION

SAS® introduced SURVEY procedures in SAS8 and SAS9. These procedures have many applications in modeling. Applications include generating samples, running bootstrap regression models, and obtaining the correct p-values when using sampled data in your models. We will look at examples of each of these applications in this paper.

Data Used in this Paper

I typically would use hypothetical data from the credit industry. However, I am reminded by some sampling a few of us used to partake in after work in the 1990's; single malt scotch whisky tasting at the now closed North Star Pub in lower Manhattan. The pub had a copy of Jackson's Guide to Single Malt Scotch and we used to read the descriptions in some blind taste tests. Taking the 4th edition of the book (Jackson, 1999) I created a SAS dataset with the following fields:

- Name of Whisky
- Region of Scotland
- Age of Whisky
- Alcohol by Volume
- Any special wood aging
- Rating

Any records with missing values of any of the above variables were removed from the analysis. As an exercise for this paper, the question is can we build a model to predict the rating of the whisky? First some summary statistics are generated. Looking at REGION we see that most of the whisky comes from the Highland Region of Scotland:

```
proc freq data=scotch.scotch;
  table region;
run;
```

Output:

region

region	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Cambeltown	6	1.64	6	1.64
Highlands	292	79.78	298	81.42
Islay	42	11.48	340	92.90
Lowlands	26	7.10	366	100.00

Looking the frequency of any type of special wood (yes or no):

```
proc format;
  value $wood  ' ' = 'NO Wood'
               other = 'WOOD'
  ;
run;
proc freq data=scotch.scotch;
  table  wood/missing;
  format wood $wood.;
run;
```

Output:

```
                            wood

                                    Cumulative    Cumulative
wood       Frequency     Percent     Frequency      Percent

NO Wood          336       91.80           336        91.80
WOOD              30        8.20           366       100.00
```

Looking at some additional stats on the numeric fields:

```
proc tabulate data=scotch.scotch noseps
              formchar='           ';
  var age alcohol rating;
  table age alcohol rating
        ,
            (min p5 p25 p50 p75 p95 max)*f=3. mean std
            /rts=20 row=float;
run;
```

Output:

```
                    Min P5  P25 P50 P75 P95 Max     Mean          Std

 age                  7   9  10  15  18  40  50        16.70         9.27
 alcohol             40  40  40  43  43  57  61        43.37         4.89
 rating              74  75  80  85  89  95  96        84.80         6.34
```

Generating Samples

Samples can be generated using PROC SURVEYSELECT. In the credit industry we often sample events that are rare (like response to offers) at 100% and non-events at a smaller rate. For the data we are working with here, the sampling maybe more geared to selecting a sample of whiskies to sample in a tasting party. Here is some code that selects 10 sample whiskies. The SRS specification is for a simple random sample.

```
proc surveyselect data=scotch.scotch method=srs n=10 out=sample1;
run;
title "PROC SURVEYSELECT";
proc print data=sample1;
run;
```

Output:

PROC SURVEYSELECT

Obs	whisky	region	age	alcohol	wood	rating
1	BRACKLA	Highlands	10	43.0		74
2	AN CNOC	Highlands	20	59.6		76
3	GLENFARCLAS	Highlands	25	43.0		88
4	GLENFIDDICH	Highlands	18	43.0		79
5	GLENMORANGIE	Highlands	21	43.0		82
6	GLEN MORAY	Highlands	17	43.0		77
7	JURA	Highlands	32	56.6		79
8	STRATHMILL	Highlands	12	43.0		71
9	TAMDHU	Highlands	27	49.5		76
10	TOMATIN	Highlands	12	43.0		76

The problem with the above sample is that we only selected Highland whisky. Let's modify the code to run a stratified sample. Note that the data must be sorted by the strata variable.

```
proc sort data=stuff.scotch out=scotch;
  by region;
run;

proc surveyselect data=scotch method=srs n=2 out=sample2;
  strata region;
run;
title "Stratified Sample";
proc print data=sample2;
run;
```

Output shows 2 samples from each region. Note that the B/S indication for WOOD is indication that the whisky aged in both bourbon and sherry barrels.

Stratified Sampling

Obs	region	whisky	age	alcohol	wood	rating	Prob	Weight
1	Cambeltown	GLENSCOTIA	14	40.0		87	0.33333	3
2	Cambeltown	SPRINGBANK	12	46.0		84	0.33333	3
3	Highlands	BALVENIE	12	40.0	B/S	87	0.00685	146
4	Highlands	AN CNOC	21	57.5		77	0.00685	146
5	Islay	BOWMORE	15	43.0		87	0.04762	21
6	Islay	CAOLILA	20	61.3		82	0.04762	21
7	Lowlands	BLADNOCH	10	43.0		85	0.07692	13
8	Lowlands	BLADNOCH	17	57.2		84	0.07692	13

We can modify the n= option. What if we only wanted to select 1 bottling from Cambeltown and 2 from the other regions. This can be done as follows:

```
proc surveyselect data=scotch method=srs n=(1 2 2 2) out=sample2;
  strata region;
run;

title "Stratified Sample";
proc print data=sample2;
run;
```

Output:

Obs	region	whisky	age	alcohol	wood	rating	Selection Prob	Sampling Weight
1	Cambeltown	SPRINGBANK	12	46.0		84	0.16667	6
2	Highlands	BLAIR ATHOL	8	40.0		75	0.00685	146
3	Highlands	DALMORE	12	40.0		79	0.00685	146
4	Islay	BRUICHLADDICH	10	40.0		77	0.04762	21
5	Islay	CAOLILA	8	60.4		78	0.04762	21
6	Lowlands	AUCHENTOSHAN	10	40.0		83	0.07692	13
7	Lowlands	LITTLEMILL	8	43.0		78	0.07692	13

In addition to the n= option we can provide sampling rates using rate=(*rates*) options. Rate values indicate what percentage of each strata you wish to include in your sample. Here is an example and output:

```
proc surveyselect data=scotch method=srs rate=(.16, .006, .04, .07)
out=sample2;
  strata region;
run;

title "Stratified Sample";
proc print data=sample2;
run;
```

Obs	region	whisky	age	alcohol	wood	rating	Selection Prob	Sampling Weight
1	Cambeltown	GLENSCOTIA	14	40.0		87	0.16667	6
2	Highlands	HIGHLAND PARK	12	40.0		90	0.00685	146
3	Highlands	HIGHLAND PARK	14	52.6		93	0.00685	146
4	Islay	BRUICHLADDICH	10	40.0		77	0.04762	21
5	Islay	CAOLILA	20	61.3		82	0.04762	21
6	Lowlands	AUCHENTOSHAN	10	40.0		83	0.07692	13
7	Lowlands	BLADNOCH	10	43.0		85	0.07692	13

Regression examples

Let's generate some scatter plots for the numeric data to see if we see any trends. Code:

```
goptions reset=all;

ods html;
ods graphics on;

proc corr data=scotch.scotch plots=matrix;
  var rating alcohol age;
run;

ods graphics off;
ods html close;
```

Output of Scatter Plot generated by ODS GRAPHICS is shown in figure 1:

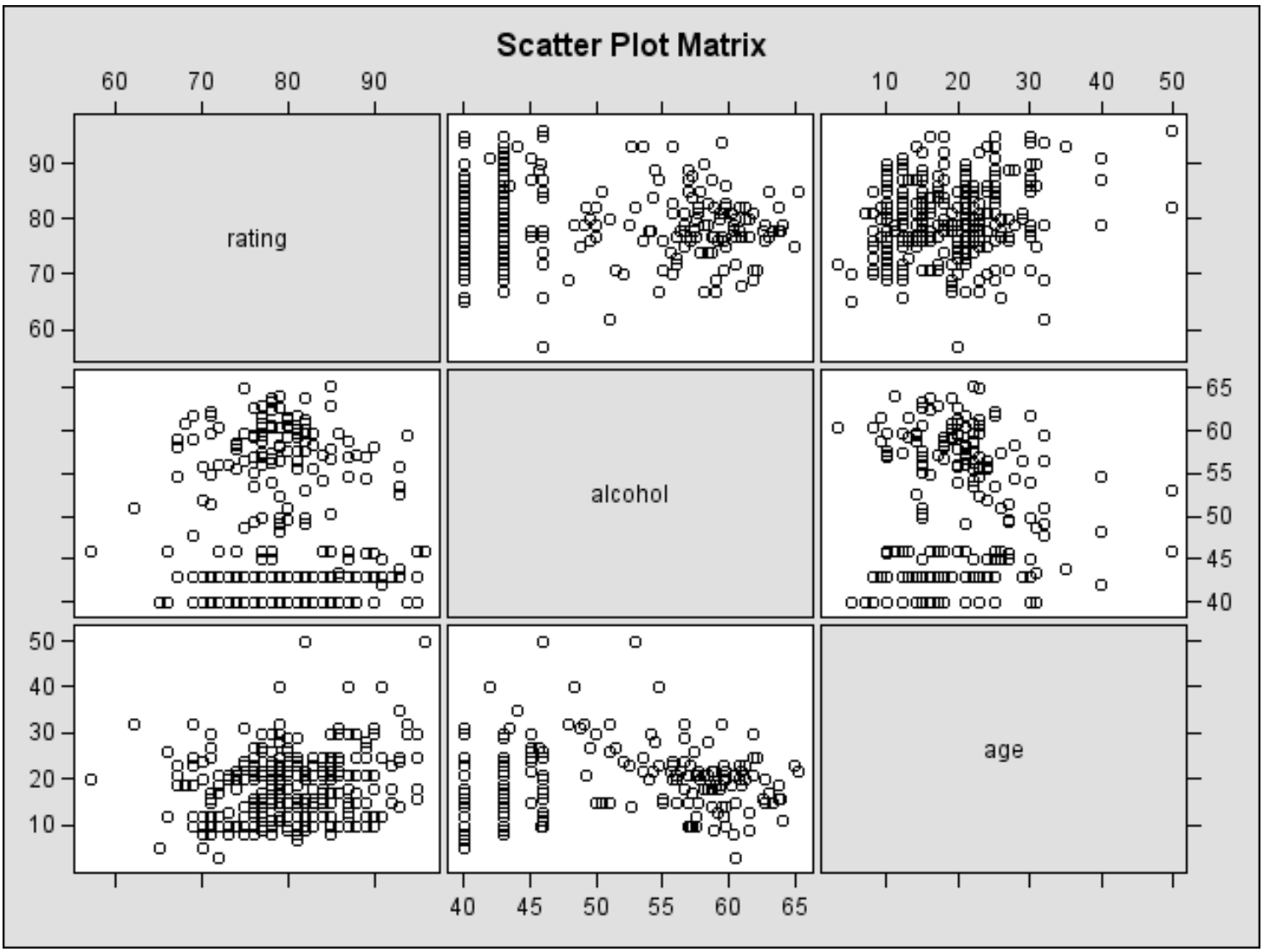

Figure 1. ODS GRAPHICS Output

We see a stronger relationship between AGE and RATING than ALCOHOL and RATING. Let us build a few models. Code:

```
data scotch;
  set scotch.scotch;
  woody=(wood ne ' ');
run;

proc genmod data=scotch;
  class region/param=glm;
  model rating = age|alcohol|region|woody@2
                 age*age alcohol*alcohol
                         /wald type3 dist=normal;
run;
```

Looking at the WALD TYPE 3 stats we can probably build a model using AGE, ALCOHOL and interaction of the 2 to predict rating:

Wald Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
age	1	19.81	<.0001
alcohol	1	0.01	0.9421
age*alcohol	1	13.94	0.0002
region	3	0.96	0.8106
age*region	3	3.95	0.2667
alcohol*region	3	0.31	0.9587
woody	1	0.34	0.5577
age*woody	1	0.15	0.6979
alcohol*woody	1	0.06	0.8000
woody*region	2	2.30	0.3167
age*age	1	0.06	0.8126
alcohol*alcohol	1	0.29	0.5892

Code to generate the final model is:

```
proc genmod data=scotch;
  class region/param=glm;
  model rating = age|alcohol
                 /wald type3 dist=normal;
run;
```

Output:

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	57.4903	5.9209	45.8855	69.0950	94.28	<.0001
age	1	1.5878	0.3278	0.9453	2.2304	23.46	<.0001
alcohol	1	0.3881	0.1287	0.1358	0.6404	9.09	0.0026
age*alcohol	1	-0.0286	0.0069	-0.0422	-0.0150	17.04	<.0001
Scale	1	5.9130	0.2186	5.4998	6.3572		

Bootstrap Regression examples

As an alternative to hold out validation samples, we can generate many samples using PROC SURVEYSELECT and evaluate in a bootstrap context. Code, suggested by Cassell (2007), is listed here. The URS option specifies that samples are drawn using Unrestricted Random Sampling. This generates a simple random sample with replacement.

```
proc surveyselect data=scotch out=outdata seed=20060510
    rep=1000 method=urs samprate=1 outhits;
  run;

ods output ParameterEstimates=bout;
proc genmod data=outdata;
  class region/param=glm;
  by replicate;
  model rating = age|alcohol
                 /wald type3 dist=normal;
run;
ods output close;
proc sort data=bout force noequals;
  by parameter;
run;
proc univariate data=bout;
  by parameter;
  var estimate;
  output out=final pctlpts=2.5, 5, 95, 97.5 pctlpre=ci mean=mean;
run;

proc print data=final;
run;
```

Results are shown here.

Obs	Parameter	mean	ci2_5	ci5	ci95	ci97_5
1	Intercept	57.2523	45.1756	47.1430	67.4013	69.3703
2	Scale	5.8663	5.3604	5.4321	6.3186	6.4175
3	age	1.6047	0.8899	1.0029	2.2076	2.2742
4	age*alcohol	-0.0291	-0.0441	-0.0421	-0.0161	-0.0137
5	alcohol	0.3946	0.1403	0.1747	0.6149	0.6740

One thing to look for in the output above is a switch in sign from parameter estimates going from ci2_5 (2.5 percentile) and ci97_5 (97.5 percentile). This would indicate that the variable is not significant at the 2 sided alpha of .05.

Which model works best. Looking at correlations between the 2 models, the bootstrap has a 0.34183 correlation with the actual rating and the single regression has 0.34188 correlation. I would still go with the bootstrap model since we ran 1000 samples through regressions.

As an aside, you may want to turn off ODS LISTING CLOSE during the regression runs. Also, to speed up processing, you may want to load the data into memory during the SURVEYSELCT step. This is done with a SASFILE SCOTCH LOAD statement before the runs and SASFILE SCOTCH CLOSE statement at the end:

```
ODS LISTING CLOSE:
SASFILE SCOTCH LOAD;
proc surveyselect data=scotch out=outdata seed=20060510
    rep=1000 method=urs samprate=1 outhits;
run;
SASFILE SCOTCH CLOSE;
ODS LISTING;
```

Dealing with Sampled Data

If your data is built from samples, the p-values will not be correct in regression models. You will need to run PROC SURVEYREG or PROC SURVEYLOGISTIC for binary data. Let us run an example. Take a stratified sample of the data:

```
proc surveyselect data=scotch method=srs rate=(.20, .1, .1, .1) out=sample2;
  strata region;
run;
```

To run SURVEYREG we need a dataset with the totals for each region: This we have from the original frequency:

```
data strat_totals;
  input region $ _TOTAL_;
  datalines;
Cambelown      6
Highlands    292
Islay         42
Lowlands      26
;;
run;
```

We need to specify an interaction term for the model:

```
data sample_int;
  set sample2;
  alc_age=alcohol*age;
run;
```

Now we run SURVEYREG:

```
proc surveyreg data=sample_int total=Strat_Totals;
  strata region / list;
  model rating = alcohol age alc_age / covb;
  weight SamplingWeight;
run;
```

Selected Output from SURVEYREG:

```
Fit Statistics

R-square            0.1725
Root MSE            6.4642
Denominator DF          36
```

```
The SURVEYREG Procedure

Regression Analysis for Dependent Variable rating

             Estimated Regression Coefficients

                            Standard
Parameter      Estimate         Error    t Value    Pr > |t|

Intercept    41.8319823    15.7632129       2.65      0.0118
alcohol       0.7148644     0.3492696       2.05      0.0480
age           2.8604972     0.8980007       3.19      0.0030
alc_age      -0.0537964     0.0192858      -2.79      0.0084
```

If we were to compare a PROC REG with a WEIGHT statement, the code:

```
proc reg data=sample_int;
 model rating = alcohol age alc_age / covb;
  weight SamplingWeight;
run;
```

And Selected REG OUTPUT shows that parameter estimates are identical but the p-values are different:

```
Root MSE             19.55353    R-Square     0.1725
Dependent Mean       80.26903    Adj R-Sq     0.1035
Coeff Var            24.35999


                              Parameter Estimates

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

Intercept    Intercept      1       41.83198       19.88241       2.10      0.0424
alcohol      alcohol        1        0.71486        0.41854       1.71      0.0962
age          age            1        2.86050        1.19588       2.39      0.0221
alc_age                     1       -0.05380        0.02461      -2.19      0.0354
```

Some LOGISTIC Examples

To complete the presentation let's look at how we would handle SURVEY options with logistic regression. Normally I would not convert a continuous (ordinal, interval, or ratio) variable into a binned or binary data but for this exercise what if we wanted to predict the probability of a rating greater than or equal to 85? Here is some code and selected output which includes the ROC curve generated with ODS GRAPHICS in figure 2:

```
proc format;
  value high 85-high = '1'
             other   = '0'
;
run;
```

```
goptions reset=all;
ods html;
ods graphics on;

proc logistic data=scotch descending;
  format rating high.;
  model rating (event='1') = age|alcohol
                             /outroc=roc1;
                             ;
run;
ods graphics off;
```

OUTPUT:

Response Profile

Ordered Value	rating	Total Frequency
1	1	85
2	0	281

Probability modeled is rating='1'.

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-7.6605	2.6649	8.2634	0.0040
age	1	0.5071	0.1449	12.2471	0.0005
alcohol	1	0.1105	0.0577	3.6659	0.0555
age*alcohol	1	-0.00922	0.00309	8.9010	0.0029

Association of Predicted Probabilities and Observed Responses

Percent Concordant	66.5	Somers' D	0.363
Percent Discordant	30.2	Gamma	0.376
Percent Tied	3.3	Tau-a	0.130
Pairs	23885	c	0.682

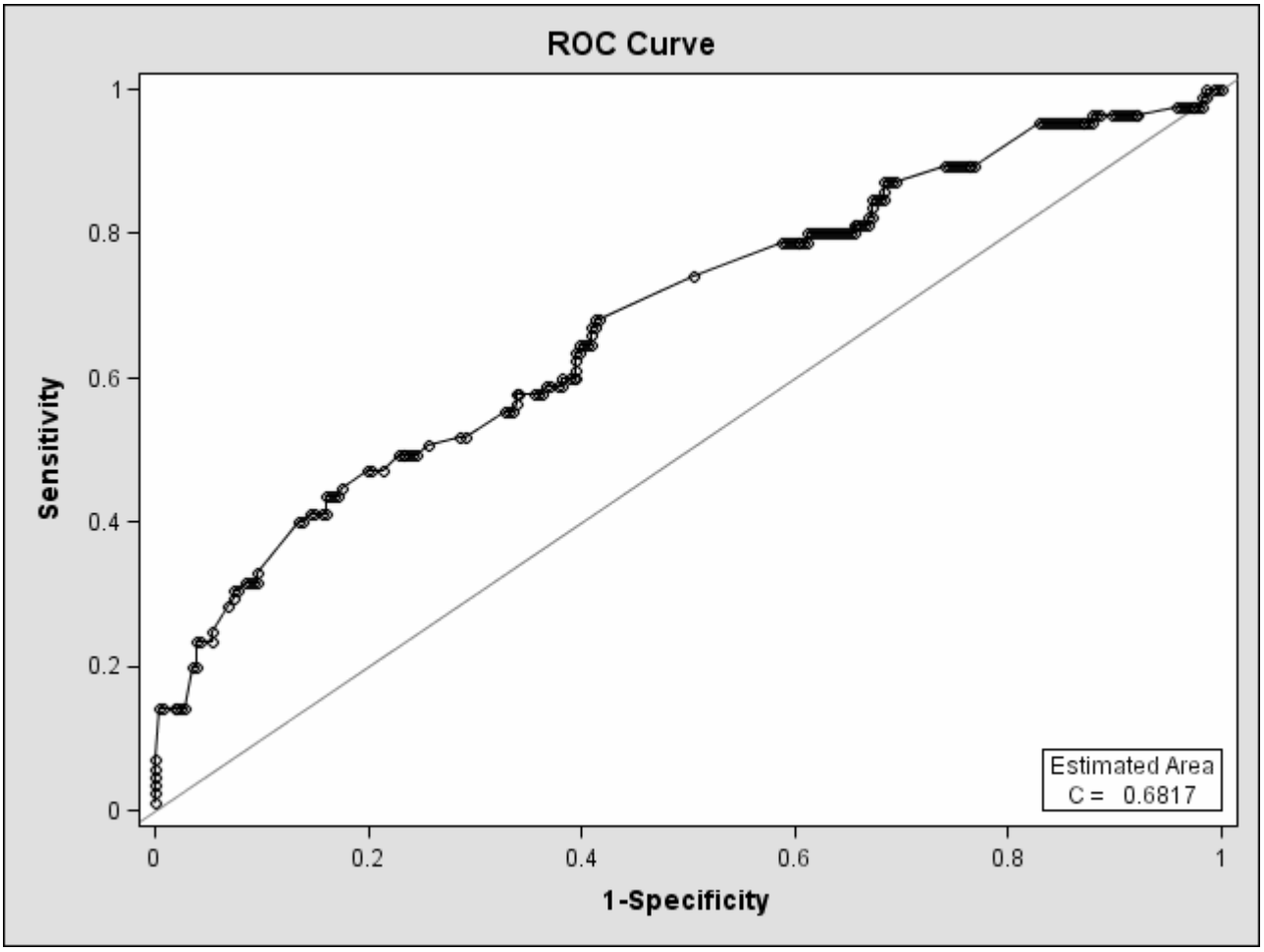

Figure 2. ROC Curve generated from ODS GRAPHICS

Bootstrap code and output is shown here. Note that the variable specified in the BY statement is "VARIABLE" for LOGISTIC REGRESSION.

```
sasfile scotch open;
proc surveyselect data=scotch out=outdata seed=20060510
    rep=1000 method=urs samprate=1 outhits;
  run;
sasfile scotch close;

ods listing close;
ods output ParameterEstimates=bout;
proc logistic data=outdata;
      by replicate;
  format rating high.;
  model rating (event='1') = age|alcohol;
      run;
ods output close;

proc sort data=bout;
  by variable;
run;
```

```
proc univariate data=bout;
  by variable;
  var estimate;
  output out=final pctlpts=2.5, 5, 95, 97.5 pctlpre=ci mean=mean;
run;

ods listing;
proc print data=final;
run;
```

OUTPUT:

Obs	Variable	mean	ci2_5	ci5	ci95	ci97_5
1	Intercept	-7.47205	-13.5865	-12.3924	-2.15211	-1.10659
2	age	0.50296	0.1713	0.2183	0.79534	0.84284
3	age*alcohol	-0.00911	-0.0169	-0.0155	-0.00297	-0.00163
4	alcohol	0.10545	-0.0345	-0.0154	0.21423	0.23470

The final example we provide is using PROC SURVEYLOGISTIC with the same samples we used in PROC SURVEYREG.

CODE:

```
data sample2;
  set sample2;
  high=put(rating,high.);
run;

proc surveylogistic data=sample2
                    total=strat_totals;
  class region;
      model high (event='1') = alcohol|age;
      weight SamplingWeight;
      stratum Region;
run;

proc logistic data=sample2;
      model high (event='1') = alcohol|age;
      weight SamplingWeight;
run;
```

SURVEYLOGISTIC OUTPUT:

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-21.3959	8.9723	5.6866	0.0171
alcohol	1	0.3944	0.1907	4.2788	0.0386
age	1	1.1829	0.4665	6.4311	0.0112
alcohol*age	1	-0.0234	0.00997	5.5033	0.0190

LOGISTIC OUTPUT:

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-21.3959	3.5300	36.7386	<.0001
alcohol	1	0.3944	0.0732	29.0461	<.0001
age	1	1.1829	0.1719	47.3357	<.0001
alcohol*age	1	-0.0234	0.00361	42.0919	<.0001

In the above outputs, notice that the coefficients are identical between LOGISTIC and SURVEYLOGISTIC. However we see that the standard errors and statistical tests for parameter estimates are different with the correct SURVEYLOGISTIC generating the correct results.

For cases where sampling is done on the event, SURVEYLOGISTIC will also provide the correct statistics. Sampling by event is common in credit models where the analyst will sample the rare event at 100% and sample the non-event at a much lower rate.

More detail on using the SURVEYREG and SURVEYLOGISTIC procedures can be found in Cassell (2006).

CONCLUSION

In this paper we have seen how simple and stratified samples can be generated with a few lines of code using PROC SURVEYSELECT. We have also seen how the procedure can generate samples to run bootstrap regressions. We also noticed that when we run regressions on sampled data we must use the appropriate SURVEY regression PROC to obtain the correct p-values.

REFERENCES AND ADDITIONAL INFORMATION:

- Cassell, D. (2007). Don't Be Loopy: Re-Sampling and Simulation the SAS® Way. SAS Global Forum 2007. Paper 183-2007.
- Cassell, D. (2006) Wait Wait, Don't Tell Me… You're Using the Wrong Proc! SUGI31. Paper 193-31.
- Effron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical Science, 1:54-77.
- Jackson, M. (1999), Michael Jackson's Complete Guide to Single Malt Scotch, 4th Edition. Running Press.

CONTACT INFORMATION

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

Jonas V. Bilenas
JP Morgan Chase Bank
Wilmington, DE 19801
Email: Jonas.Bilenas@chase.com
jonas@jonasbilenas.com

This work is an independent effort and does not necessarily represent the practices followed at JP Morgan Chase Bank.