Using the New SURVEY Procedures from a Modeling Perspective
Jonas V. Bilenas, JP Morgan Chase, Wilmington, DE

ABSTRACT
This is 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):

```plaintext
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:

```plaintext
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
age        3   8  12  16  21  30  50        17.43         7.07
alcohol   40  40  40  43  54  61  65        46.82         7.59
rating    57  70  76  79  84  91  96        79.55         6.30
```

**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.

```plaintext
proc surveyselect data=scotch.scotch method=srs n=10 out=sample1;
run;
```

```
Statistics & Analysis
NESUG 2009

```
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.

```sas
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.

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

proc surveyselect data=scotch method=srs n=(1 2 2 2 ) out=sample2;
   strata region;
run;
title "Stratified Sample";
proc print data=sample2;
run;
```

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:

```sas
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:

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

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;

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

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:

![Scatter Plot Matrix](image)

**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:

```plaintext
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

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>1</td>
<td>19.81</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>alcohol</td>
<td>1</td>
<td>0.01</td>
<td>0.9421</td>
</tr>
<tr>
<td>age*alcohol</td>
<td>1</td>
<td>13.94</td>
<td>0.0002</td>
</tr>
<tr>
<td>region</td>
<td>3</td>
<td>0.96</td>
<td>0.8106</td>
</tr>
<tr>
<td>age*region</td>
<td>3</td>
<td>3.95</td>
<td>0.2667</td>
</tr>
<tr>
<td>alcohol*region</td>
<td>3</td>
<td>0.31</td>
<td>0.9587</td>
</tr>
<tr>
<td>woody</td>
<td>1</td>
<td>0.34</td>
<td>0.5577</td>
</tr>
<tr>
<td>age*woody</td>
<td>1</td>
<td>0.15</td>
<td>0.6979</td>
</tr>
<tr>
<td>alcohol*woody</td>
<td>1</td>
<td>0.06</td>
<td>0.8000</td>
</tr>
<tr>
<td>woody*region</td>
<td>2</td>
<td>2.30</td>
<td>0.3167</td>
</tr>
<tr>
<td>age*age</td>
<td>1</td>
<td>0.06</td>
<td>0.8126</td>
</tr>
<tr>
<td>alcohol*alcohol</td>
<td>1</td>
<td>0.29</td>
<td>0.5892</td>
</tr>
</tbody>
</table>

Code to generate the final model is:

```sas
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.

<table>
<thead>
<tr>
<th>Obs</th>
<th>Parameter</th>
<th>mean</th>
<th>ci2.5</th>
<th>ci5</th>
<th>ci95</th>
<th>ci97.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Intercept</td>
<td>57.2523</td>
<td>45.1756</td>
<td>47.1430</td>
<td>67.4013</td>
<td>69.3703</td>
</tr>
<tr>
<td>2</td>
<td>Scale</td>
<td>5.8663</td>
<td>5.3604</td>
<td>5.4321</td>
<td>6.3186</td>
<td>6.4175</td>
</tr>
<tr>
<td>3</td>
<td>age</td>
<td>1.6047</td>
<td>0.8899</td>
<td>1.0029</td>
<td>2.2076</td>
<td>2.2742</td>
</tr>
<tr>
<td>4</td>
<td>age*alcohol</td>
<td>-0.0291</td>
<td>-0.0441</td>
<td>-0.0421</td>
<td>-0.0161</td>
<td>-0.0137</td>
</tr>
<tr>
<td>5</td>
<td>alcohol</td>
<td>0.3946</td>
<td>0.1403</td>
<td>0.1747</td>
<td>0.6149</td>
<td>0.6740</td>
</tr>
</tbody>
</table>

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 SURVEYSELECT 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:

```plaintext
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:

```plaintext
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:

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

Now we run SURVEYREG:

```plaintext
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

| Parameter    | Estimate  | Standard 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

| Variable   | Label     | DF | Parameter Estimate | Standard 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

<table>
<thead>
<tr>
<th>Ordered Value</th>
<th>Total rating</th>
<th>Total Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>85</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>281</td>
</tr>
</tbody>
</table>

Probability modeled is rating='1'.

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>-7.6605</td>
<td>2.6649</td>
<td>8.2634</td>
<td>0.0040</td>
</tr>
<tr>
<td>age</td>
<td>1</td>
<td>0.5071</td>
<td>0.1449</td>
<td>12.2471</td>
<td>0.0005</td>
</tr>
<tr>
<td>alcohol</td>
<td>1</td>
<td>0.1105</td>
<td>0.0577</td>
<td>3.6659</td>
<td>0.0555</td>
</tr>
<tr>
<td>age*alcohol</td>
<td>1</td>
<td>-0.00922</td>
<td>0.00309</td>
<td>8.9010</td>
<td>0.0029</td>
</tr>
</tbody>
</table>

Association of Predicted Probabilities and Observed Responses

<table>
<thead>
<tr>
<th>Percent Concordant</th>
<th>Somers' D</th>
<th>0.363</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent Discordant</td>
<td>0.376</td>
<td></td>
</tr>
<tr>
<td>Percent Tied</td>
<td>0.130</td>
<td></td>
</tr>
<tr>
<td>Pairs</td>
<td>0.682</td>
<td></td>
</tr>
</tbody>
</table>
Bootstrap code and output is shown here. Note that the variable specified in the BY statement is “VARIABLE” for LOGISTIC REGRESSION.

```sas
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:

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

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

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Wald Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>-21.3959</td>
<td>8.9723</td>
<td>5.6866</td>
<td>0.0171</td>
</tr>
<tr>
<td>alcohol</td>
<td>1</td>
<td>0.3944</td>
<td>0.1907</td>
<td>4.2788</td>
<td>0.0386</td>
</tr>
<tr>
<td>age</td>
<td>1</td>
<td>1.1829</td>
<td>0.4665</td>
<td>6.4311</td>
<td>0.0112</td>
</tr>
<tr>
<td>alcohol*age</td>
<td>1</td>
<td>-0.0234</td>
<td>0.00997</td>
<td>5.5033</td>
<td>0.0190</td>
</tr>
</tbody>
</table>
LOGISTIC OUTPUT:

Analysis of Maximum Likelihood Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Wald Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>-21.3959</td>
<td>3.5300</td>
<td>36.7386</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>alcohol</td>
<td>1</td>
<td>0.3944</td>
<td>0.0732</td>
<td>29.0461</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>age</td>
<td>1</td>
<td>1.1829</td>
<td>0.1719</td>
<td>47.3357</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>alcohol*age</td>
<td>1</td>
<td>-0.0234</td>
<td>0.00361</td>
<td>42.0919</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

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:


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

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. © indicates USA registration. Other brand and product names are trademarks of their respective companies.

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