Using bootstrap weights with Wes Var and SUDAAN

By Owen Phillips

Abstract

For the purpose of design-based variance estimation, a number of Statistics Canada surveys supply bootstrap weights with their microdata. While the use of bootstrap weights is not explicitly supported by commercially available software such as SUDAAN and WesVar, by taking advantage of similarities between a commonly used bootstrap technique and the method of Balanced Repeated Replication (BRR), these software can be used to produce bootstrap variance estimates. This article examines the reasoning behind this, and shows, by way of example, how this might be accomplished. The paper concludes with a brief discussion of other design-based approaches to variance estimation as well as software, programs and procedures where these methods have been employed.

Introduction

A bootstrap approach to design-based variance estimation is used increasingly in the survey sampling community. Several Statistics Canada surveys—Survey of Labour and Income Dynamics (SLID), National Population Health Survey (NPHS), and the General Social Survey (GSS), to name but a few—all provide bootstrap weights, or variants thereof, with their microdata for the purpose of variance estimation.

The (survey) bootstrap belongs to a family of variance estimation techniques known as replicate based variance estimation. A detailed discussion of replication methods can be found in Lohr (1999), Rust and Rao (1996) or Wolter (1985). Such methods use the existing sample to build ‘synthetic’ samples, called replicates. Balanced Repeated Replication (BRR) is another such method, and has been implemented in commercially available software such as SUDAAN and WesVar. While the bootstrap and BRR differ in the way in which the replicates are built, bootstrap weights can be used to produce bootstrap variance estimates in software that will accommodate BRR weights, a point that the software documentation fails to mention in great detail.

The following sections will elaborate on the differences and similarities between the bootstrap and BRR, and will, by way of example, show how to use bootstrap weights in SUDAAN and WesVar. A variant of the bootstrap employed by the GSS and the Workplace and Employee Survey (WES) known as the mean bootstrap will be contrasted against Fay’s variant of BRR. The paper will conclude with a brief discussion of other design-based variance estimation techniques and the software and programs that incorporate the many techniques discussed.

For simplicity, this paper presents a very general discussion of the process of producing survey weights and ignores many of its complexities like non-response adjustments and post-stratification. However, it is assumed that the reader is familiar with basic concepts of survey sampling. For those wishing more information on the sampling process, please refer to *Survey* *methods and practices* (Statistics Canada 2003). Some familiarity with SAS, SUDAAN and/or WesVar is also assumed.

II. Bootstrap methods

Many Statistics Canada surveys, including SLID, NPHS, the National Longitudinal Survey of Children and Youth (NLSCY), the Canadian Community Health Survey (CCHS), the Ethnic Diversity Survey (EDS) and the Youth in Transition Survey (YITS), are using a bootstrap method to estimate sampling error. Without going into too much detail, bootstrap replicates are generated by randomly choosing, with replacement, a sample of primary sampling units (PSUs) within each stratum and adjusting the original sampling weights of the units in the selected PSUs to reflect the probability of selection into the subsample. If a unit does not appear in the bootstrap replicate, its bootstrap weight variable is set to zero. This process of selecting samples and reweighting is repeated *B* times to arrive at *B* bootstrap samples, *B* bootstrap weight variables and consequently *B* bootstrap estimates.

The variance of the estimate $\hat{\theta}$ of the finite population parameter θ of interest—for example a regression coefficient, population mean, ratio of two totals, etc.—is estimated by

$$\hat{V}_{BOOT}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_b - \hat{\theta}\right)^2 \qquad (1)$$

where $\hat{\theta}$ is obtained using the full-sample weight variable and the estimates $\hat{\theta}_b$, b=1,…,B are obtained in exactly the same manner using the bootstrap weight variables[1,2].

A variant of the bootstrap, called the ***mean bootstrap,*** is used by GSS and WES. This method was originally proposed to address confidentiality issues arising from the release of bootstrap weights with public use microdata (See Yung 1997). Ultimately, it amounts to calculating the bootstrap weights as above and then averaging the bootstrap weights over *C* bootstrap samples. For example, in certain cycles of the GSS, 5000 bootstrap weight variables were produced. These weights were then averaged in groups of size *C*=25 to obtain the *B*=200 mean bootstrap weights that accompany the microdata. Similarly, WES provides 100 mean bootstrap weights, each of which is the mean of *C*=50 bootstrap weights.

The mean bootstrap variance estimator is given by:

[1] The sampling weight reflects the probability of selection of a unit to the full sample: it can be thought of as the number of units in the survey population represented by the sampled unit. The sampling weight is used to estimate the parameter of interest. The bootstrap weight is used for the purpose of estimating the sampling error associated with the parameter of interest. Like the sampling weight, a bootstrap weight might be thought of as the number of individuals in the survey population represented by a unit in the reduced (bootstrap) sample.

[2] Research is ongoing into approaches for obtaining estimates $\hat{\theta}_b$ from the bootstrap samples, other than by mirroring the approach used to estimate $\hat{\theta}$ from the full sample. These approaches may provide more stable variance estimates for some situations (see Roberts *et al*, 2003).

$$\hat{V}_{MBOOT}(\hat{\theta}) = \frac{C}{B}\sum_{b=1}^{B}\left(\hat{\theta}_b - \hat{\theta}\right)^2 \qquad (2)$$

where $\hat{\theta}_b$ is obtained using the bth mean bootstrap weight variable.

III. Balanced repeated replication

Balanced Repeated Replication (BRR) is applicable to survey designs where two and only two PSU are selected per stratum. As many Statistics Canada surveys sample more than two units at the first stage, BRR cannot be used as a variance estimation method for those surveys. However, certain features of the BRR method allow us to use software intended for BRR variance estimation to obtain bootstrap variance estimation.

The BRR method consists of generating half-samples by selecting one PSU in each stratum. In a prescribed way[3], a 'balanced' set of *G* half samples is selected; in each of these, the sampling weights of units within the selected PSUs are multiplied by 2, while units in non-selected PSUs are given a weight of 0. The half sample estimates $\hat{\theta}_g$ are then used to compute

$$\hat{V}_{BRR}(\hat{\theta}) = \frac{1}{G}\sum_{g=1}^{G}\left(\hat{\theta}_g - \hat{\theta}\right)^2 \qquad (3)$$

A possible variant of BRR is ***Fay's method***. The methodology used for selecting the half-samples is the same as for BRR; however, the weighting is done differently: the weights of units in the selected PSUs are multiplied by a factor (2-*K*); the weights of units in 'non-selected' PSUs are multiplied by *K*, where *K* is a fixed constant in the interval [0,1). In this way, all observed units contribute to each estimate $\hat{\theta}_g$, which is particularly useful when working with small domains. Variances are then estimated by

$$\hat{V}_{FAY}(\hat{\theta}) = \frac{1}{G(1-K)^2}\sum_{g=1}^{G}\left(\hat{\theta}_g - \hat{\theta}\right)^2 \qquad (4)$$

The Programme for International Student Assessment (PISA) uses Fay's method, with *K*=0.5.

IV. What the software documentation doesn't tell you

As we have seen in Sections II and III, the methodologies for creating bootstrap and BRR weights are different. That being said, the form of the bootstrap variance estimator is the same as that for BRR (i.e. (1) and (3) are equivalent provided that *G*=*B*). In other words, if the bootstrap weights are provided, but designated to be BRR weights, software such as SUDAAN and WesVar that allow BRR variance estimation will calculate bootstrap variance estimates

[3] The methodology of producing the balanced half samples is not necessary for this discussion. The interested reader may consult Section 9.3.1 of Lohr (1999).

appropriately. Similarly for the mean bootstrap and Fay's method, by setting $K = 1 - C^{-\frac{1}{2}}$ in (4), (2) and (4) are equivalent, and software that will accommodate replicate weights calculated using Fay's method can also be used to calculate variances using mean bootstrap weights.

Using bootstrap weights in SUDAAN

Specification of the variance estimation method to be used by SUDAAN is done in the call to a particular analytic procedure. The following process is the same for all SUDAAN procedures that allow for BRR variance estimation[4]:

- The bootstrap is implemented in SUDAAN by specifying DESIGN=BRR.
- The REPWGT statement is used to indicate the names of the variables containing the bootstrap weights.
- The WEIGHT statement is not mandatory, but should be used when the variable containing the final weight is available[5].
- For surveys that provide mean bootstrap weights, set option ADJFAY=*C* (note that, at the time of the writing of this article, for GSS, *C*=25 and for WES, *C*=50).

Using bootstrap weights in WesVar

In WesVar, the variance estimation method is specified when creating a new WesVar data file. The resulting file is then used to define workbooks where table and regression requests are carried out. To define a WesVar data file with bootstrap or mean bootstrap weights:

- Move the replicate weight variables to *Replicates* box.
- Move the final weight variable to the *Full sample* box
- For bootstrap, specify the *Method* as BRR
- For mean bootstrap, specify the *Method* as Fay and specify $Fay\_K = 1 - C^{-\frac{1}{2}}$
- Move analysis variables to the *Variables* box, a unique identifier to the *ID* box (optional) and save the file.

V. Examples of using WesVar and SUDAAN

The following examples are applications of the information given in Section IV. The particulars of defining new variables and manipulating the data into a format suitable for use with the software are, for the most part, ignored. The goal of these examples is simply to show how the bootstrap may be implemented in the two software packages, and thus ignores the interpretation of the resulting output. The details of reading and interpreting output from SUDAAN and WesVar are found in the respective software user guides (see RTI, 2001 and Westat, 2002).

[4] In SUDAAN Release 8.02 and earlier, DESIGN=BRR cannot be specified for PROC SURVIVAL.

[5] In the absence of the weight statement, SUDAAN uses $\hat{\hat{\theta}}_{(b)} = \frac{1}{B}\sum_{b}\hat{\theta}_b$ in place of $\hat{\theta}$.

SLID bootstrap example

The following example examines the transition from low earnings for longitudinal individuals from 1996 to 1998 using Panel 1 on SLID. Given the selected years, either Panel 1, Panel 2, or the combination of the two might have been used. The analysis variable *low96* takes on a value of 1, if in 1996 the longitudinal person's earnings at their main job fell below some prescribed level, and zero otherwise. The variable *low98* follows a similar definition. Data are stored in a SAS dataset named *mobility*. The bootstrap weights *bs1-bs1000* corresponding to the Panel 1 longitudinal weight *ilgwt26* have been merged to the analysis file.

(a) SUDAAN

PROC CROSSTAB is used in SAS-callable SUDAAN to produce estimates of those who have moved into or out of low earnings between the two survey occasions. Using the information above along with the instructions provided in Section IV, PROC CROSSTAB would be set up as follows to make use of the bootstrap weights:

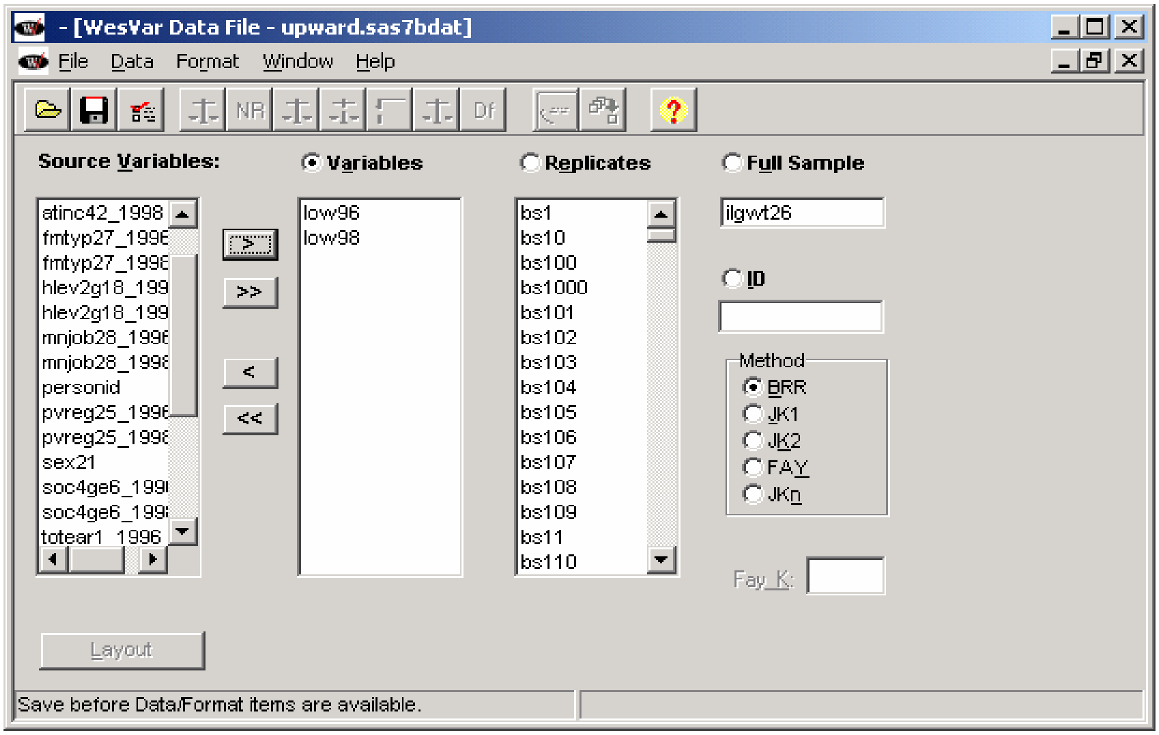
```
proc crosstab data=mobility design=BRR;
  weight ilgwt26;
  repwgt bs1-bs1000;
  recode low96=(0 1) low98=(0 1);
  subgroup low96 low98;
  levels 2 2;
  tables low96*low98;
run;
```

For more information on PROC CROSSTAB and SUDAAN, please refer to the user's guide (Research Triangle Institute, 2001).

(b) WesVar

Following the instructions in Section IV and given the same SAS datafile, a WesVar datafile would be defined as in Figure 1. The resulting file would then be saved and used to define the desired tables in a WesVar workbook. Instructions for creating a workbook and additional information on WesVar can be found in the user's guide (Westat 2002).

Figure 1: SLID bootstrap example using WesVar



Source Variables: atinc42_1998, fmtyp27_1996, fmtyp27_1998, hlev2g18_199, hlev2g18_199, mnjob28_1996, mnjob28_1998, personid, pvreg25_1996, pvreg25_1998, sex21, soc4ge6_199, soc4ge6_199, totear1_1996

● Variables: low96, low98

○ Replicates: bs1, bs10, bs100, bs1000, bs101, bs102, bs103, bs104, bs105, bs106, bs107, bs108, bs109, bs11, bs110

○ Full Sample: ilgwt26

○ ID

Method
● BRR
○ JK1
○ JK2
○ FAY
○ JKn

Save before Data/Format items are available.

GSS mean bootstrap example

In this example, GSS Cycle 14 data are used in a logistic regression to examine the association among various demographic and socio-economic factors and the probability of internet use. The dependent variable is *netuse* (equal to 1 if the responding individual uses the internet; and 0 if not). All independent variables used in the model are categorical. The final weight variable on the *internet* SAS datafile is *fwgt*, with corresponding mean bootstrap weights *bsw1-bsw200*. Recall that for GSS, *C*=25.

(a) SUDAAN

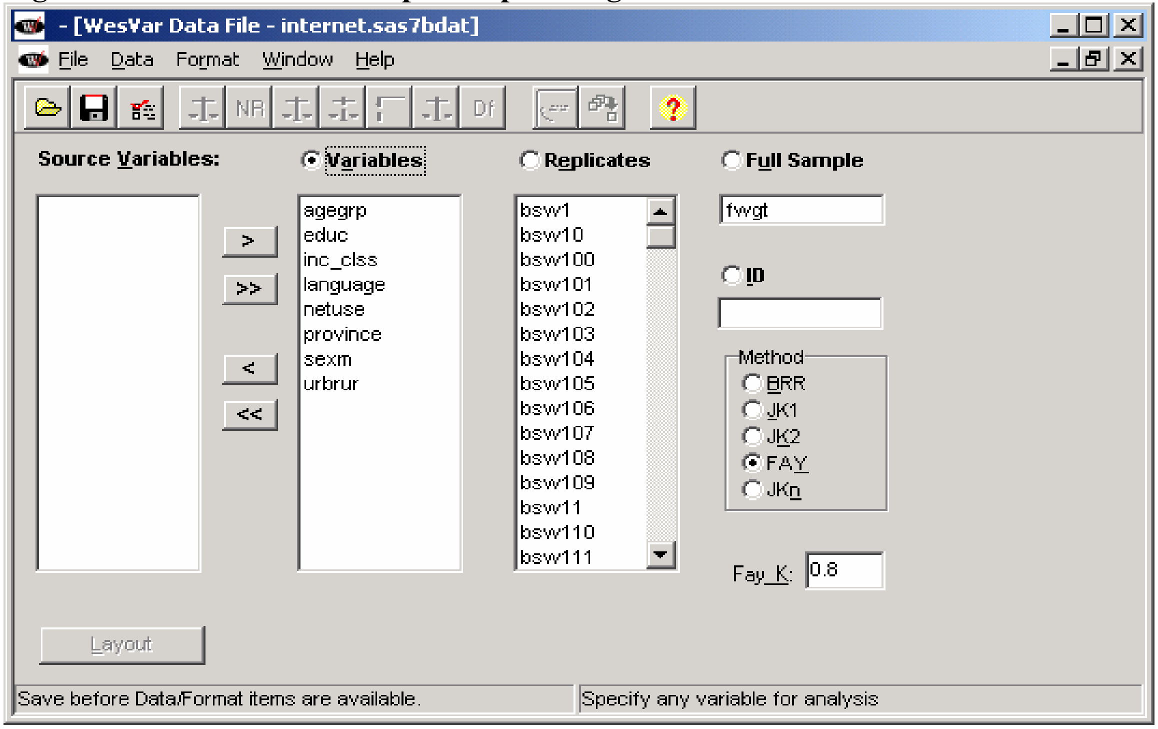
The following code shows how to set up PROC RLOGIST in SAS-callable SUDAAN given the above information and instructions in Section IV:

```
proc rlogist data=internet design=BRR;
  weight fwgt;
  repwgt bsw1-bsw200 / adjfay=25;
  subpopn province=59;
  subgroup sexm agegrp educ urbrur inc_clss language;
  levels 2 3 4 5 5 3;
  model netuse=sexm agegrp educ urbrur inc_clss language;
run;
```

(b) WesVar

Figure 2 shows how the same information is used to define a WesVar datafile. For GSS, $Fay\_K = 1 - 25^{-\frac{1}{2}} = 0.8$. As before, the resulting file would be saved and used to define the regression model in a WesVar workbook.

Figure 2: GSS mean bootstrap example using WesVar



● Variables
agegrp
educ
inc_clss
language
netuse
province
sexm
urbrur

○ Replicates
bsw1
bsw10
bsw100
bsw101
bsw102
bsw103
bsw104
bsw105
bsw106
bsw107
bsw108
bsw109
bsw11
bsw110
bsw111

○ Full Sample
fwgt

○ ID

Method
○ BRR
○ JK1
○ JK2
● FAY
○ JKn

Fay_K: 0.8

Save before Data/Format items are available.

Specify any variable for analysis

VI. Other approaches and software for design-based variance estimation

The bootstrap and BRR are not the only replication methods for obtaining design-based estimates of variance. The delete-1 jackknife, another replication method, involves deleting from the existing sample a single PSU and reweighting the remaining PSUs in the same stratum to account for the loss of sample and other weight adjustments. Each PSU is deleted once and only once, so that there are as many jackknife replicates, and consequently jackknife weights, as there are PSUs in the sample. There are other possible jackknife variants that are not described here.

Taylor series approximations, or linearization methods, are another approach to variance estimation. This is the approach implemented by SPSS and SAS in their specialized complex survey procedures. Non-linear parameters of interest (such as ratios and regression coefficients) are expressed as smooth, linear functions of simple statistics like means and totals for which an analytical formula for the form of the variance estimator is known. The desired parameter is then approximated through the first-order Taylor series expansion of this function about the true value for the parameter of interest, and the sampling error can thus be approximated. This approach is not suitable for statistics that cannot be well approximated by a linear function: chi-squares, for example.

Stratum and PSU identifiers must be supplied with the microdata in order to implement linearization methods or to implement a jackknife method where jackknife weights have not been provided. Additionally, software such as SUDAAN, WesVar and Stata are unable to account for the impact of all of the weight adjustments, such as non-response, post-stratification and other, when using Taylor linearization or when creating jackknife weights. That being said, for surveys that do not provide bootstrap weights, or in instances where the desired analysis cannot make use of bootstrap weights (e.g. PROC SURVIVAL in SUDAAN 8.02), it is recommended that one of these methods be used.

Chapter 9 of Lohr (1999) provides a more detailed overview of the replication and linearization methods discussed in this paper.

Commercially available products such as SUDAAN, WesVar, Stata and SAS, provide analytic procedures and commands capable of selected design-based analysis. Additionally, any software that offers an analytic procedure or command that can produce weighted estimates of the parameters of interest and also has the flexibility of a programming language, may be used recursively to obtain bootstrap variance estimates. Based upon this principle, SAS and SPSS macros have been constructed by Statistics Canada methodologists and are packaged together and provided with survey microdata (NPHS's and CCHS's *Bootvar*[6] and the *NLSCY Variance Estimation System* (*VES*), for example). A similar SAS-based program, *Bootmac* (written by an independent researcher), is available in the Research Data Centres (RDC). The user-defined Stata command *Bswreg*, can also be used to obtain bootstrap variance estimates for many of Stata's existing regression commands[7]. The benefits of this program were explained and exploited in the last issue of this bulletin (see Piérard *et al*, 2004).

The table in the Appendix compares the capabilities of many of the software and programs mentioned above for producing design-based variance estimates. It identifies the methods of variance estimation supported by the software, and the analytic procedures available to the user.

[6] A generic version of the Bootvar program is being produced to satisfy the need for a variance estimation tool for a number of surveys providing bootstrap weights with their microdata. This tool should be widely available in the Fall of 2004. Eventually, its capabilities will be expanded to include, for example, the estimation of sampling errors for quantiles and design-based tests of independence and homogeneity.

[7] Bswreg will not handle STATA regression commands that involve more than one line of code in order to implement the regression procedure. For example, it will not handle the Cox proportional hazards model.

References

Lohr, S. 1999. *Sampling: Design and Analysis*. Duxbury Press, USA.

Piérard, E., Buckley, N., Chowhan, J. Bootstrapping made easy: A Stata ADO file. *The Research Data Centres Information and Technical Bulletin* 1(1): 20-36.

Research Triangle Institute. 2001. *SUDAAN User's Manual, Release 8.0*. Research Triangle Institute, Research Triangle Park, NC.

Roberts, G., Binder, D., Kovacevic, M., Pantel, M., Phillips, O. 2003. Using an estimating function bootstrap approach for obtaining variance estimates when modelling complex health survey data. *Proceedings of the Survey Methods Section*, Statistical Society of Canada.

Rust, K.F., Rao, J.N.K. 1996. Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research* 5: 283-310.

Statistics Canada. 2003. *Survey Methods and Practices*, 12-587-XPE.

Westat. 2002. *WesVar 4.2 User's Guide*. Westat, USA.

Wolter, K.M. 1985. *Introduction to Variance Estimation*. Springer-Verlag, New York.

Yung, W. 1997. Variance estimation for public use microdata files. *Proceedings of Statistics Canada Symposium 97*, 91-95.

Appendix: Design-based analysis tools available in selected software

Software	**SUDAAN 8.02**	**WesVar 4.2**	**Stata 8.0**	**SAS 8.2**	**SAS 9.1**	**Bootvar and NLSCY VES**	**Bootmac**
Variance estimation approaches	**BRR (Bootstrap)** **Jackknife** **Taylor Series**	**BRR (Bootstrap)** **Jackknife**	**Taylor Series**	**Taylor Series**	**Taylor Series**	**Bootstrap**	**Bootstrap**
Modelling							
linear regression	*proc regress*	yes	*svyreg*	*proc surveyreg*	*proc surveyreg*	yes	yes
instrumental variable regression	no	no	*svyireg*	no	no	no	no
interval regression	no	no	*svyintrg*	no	no	no	yes
logistic regression	*proc logistic* (*rlogist*)	yes	*svylogit*	no	*proc surveylogistic*	yes	yes
probit regression	no	no	*svyprobt*	no	*proc surveylogistic*	no	yes
generalized logit models	*proc mulitlog*	yes	*svymlog*	no	*proc surveylogistic*	no	no
proportional odds models	*proc multilog*	no	*svyolog*	no	*proc surveylogistic*	no	yes
ordered probit regression	no	no	*svyoprob*	no	*proc surveylogistic*	no	yes
poisson and log-linear regression	*proc loglink*	no	*svypois*	no	no	no	yes
Heckman models	no	no	*svyheck*	no	no	no	no
proportional hazards models	*proc survival**	no	no	no	no	no	yes
Descriptive							
means	*proc descript*	yes	*svymean*	*proc surveymeans*	*proc surveymeans*	yes	yes
totals	*proc descript*	yes	*svytotal*	*proc surveymeans*	*proc surveymeans*	yes	yes
proportions	*proc descript*	yes	*svyprop*	no	no	yes	yes
ratios	*proc ratio*	yes	*svyratio*	no	no	yes	yes
tests of independence	*proc crosstab*	yes	*svytab*	no	*proc surveyfreq*	no	yes
quantiles	*proc descript*	yes	no	no	no	no	yes
Plausible values	no	yes**	no	no	no	no	no

* Taylor Series only

** For descriptive statistics and linear regression