Michigan SAS Users Group (MSUG):
2018 SAS Conference

The Latest and Greatest Capabilities of the SURVEY Procedures in SAS

Brady T. West, Ph.D.

Survey Research Center
Institute for Social Research
University of Michigan-Ann Arbor
Email: bwest@umich.edu
Phone: 734-647-4615

Presentation Overview

- A practical overview of the newest survey sampling and analysis procedures in SAS/STAT 14.3
- All current SURVEY procedures will be discussed, including their capabilities
- Examples of working code
- **Primary Focus:** New enhancements in SAS/STAT 14.3
- **Secondary Focus:** My personal wish list!
- A great new book on the topic by Taylor Lewis

An Important PSA To Get Started!

- Friends do not let friends analyze complex sample survey data in SAS without accounting for complex sample design features!
- Unfortunately this happens FAR TOO OFTEN; two recent publications on analytic error in peer-reviewed journal articles:

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0158120

https://surveyinsights.org/wp-content/uploads/2018/01/The-Need-to-Account-for-Complex-Sampling-Features-1.pdf

PROC SURVEYSELECT

- Used to select a probability sample from a given data set
- A wide variety of sampling procedures: simple random sampling, stratified sampling, stratified cluster sampling, PPS sampling, etc.
- The procedure automatically computes sampling weights (including replicate weights if applicable), and outputs the selected cases according to the design

SURVEYSELECT: What's New?

- The OUTORDER = RANDOM option in the PROC SURVEYSELECT statement randomly orders the selected observations in the output data set (so that they don't match the input data set)
- The REPNAME = option available when using the REPS = option in the PROC SURVEYSELECT statement provides a unique name for the replicate weight variables (rather than REPLICATE)

SURVEYSELECT: Example

- Here is example SAS code for drawing a stratified cluster sample of middle schools using systematic PP(e)S sampling:

```
data allocations;
   input stratum _nsize_ _seed_;
   datalines;
   1 14 120
   2 8 11
   3 6 100
   4 16 9
   5 14 80
   6 12 7
   7 14 60
   8 18 5
   9 10 40
   10 14 3
   11 6 20
   12 28 1
run;
```

```
proc surveyselect data = frame2
method = sys_pps sampsize =
allocations seed = allocations
out = my.sample;
   size g07;
   strata stratum;
run;
```

PROC SURVEYMEANS

- As the name suggests, used for estimation of population means, but also provides estimation of population totals and quantiles
- Enables design-based estimation and variance estimation, including all replication methods
- Also enables subgroup comparisons (**New!**)
- Generates very nice plots of weighted distributions for continuous variables

SURVEYMEANS: What's New?

- The bootstrap method of variance estimation is now possible in SAS/STAT 14.3, via the VARMETHOD = BOOTSTRAP option in the PROC SURVEYMEANS statement
- This is particularly important for quantiles
- Relatively new in SAS/STAT 14.2 (and important): subgroup comparisons, via the DIFF option in the DOMAIN statement

SURVEYMEANS: Example

- Request estimation of mean systolic blood pressure for males and females, including mean comparisons and graphs of weighted distributions (from NHANES):

```
ods graphics on;
proc surveymeans data = nhanes2012 plots = all
        varmethod = bootstrap;
   weight wtmec2yr;
   cluster sdmvpsu;
   strata sdmvstra;
   domain riagendr / diff;
   var systbp;
run;
ods graphics off;
```

PROC SURVEYFREQ

- Computes weighted estimates of frequency distributions on categorical variables
- Cross-tabulation analyses with a variety of design-adjusted chi-square tests (e.g., Rao-Scott first- and second-order corrections)
- Capable of producing plots of weighted frequency distributions
- Can also generate multi-way tables, with row and column percentages

SURVEYFREQ: What's New?

- VARMETHOD = BOOTSTRAP
- Ability to select which two-way sub-tables to display when requesting a multi-way table
- DOMAIN = ROW option in the TABLES statement: estimate frequency distributions and perform design-adjusted one-way chi-square tests for all rows of a given table
- Measures of agreement: the AC1 agreement coefficient, and Kappa statistics (PABAK)

SURVEYFREQ: Example

- Here is example code for plotting a weighted frequency distribution of marital status, and performing design-adjusted chi-square tests:

```
ods graphics on;
proc surveyfreq data =
   nhanes2012;
   weight wtint2yr;
   cluster sdmvpsu;
   strata sdmvstra;
   tables dmdmartl /
      plots = wtfreqplot;
   format dmdmartl matst.;
run;
ods graphics off;
```

```
proc surveyfreq data =
   nhanes2012;
   weight wtint2yr;
   cluster sdmvpsu;
   strata sdmvstra;
   tables female*dmdborn4
      / row chisq;
   format female fm.
      dmdborn4 cb.;
run;
```

PROC SURVEYREG

- Fits linear regression models to continuous dependent variables (normal residuals?) using weighted least squares (WLS) estimation
- Can handle categorical variables correctly via the CLASS statement
- Enables multi-parameter design-adjusted Wald tests for arbitrary blocks of parameters
- Subpopulation estimation via DOMAIN

SURVEYREG: What's New?

- VARMETHOD = BOOTSTRAP
- One can request that output for only specific levels of a domain variable be displayed when using DOMAIN for analysis

SURVEYREG: Example

- Fit a linear regression model with categorical predictors, and include a two-way interaction (along with lower-order coefficients):

```
proc surveyreg data = nhanes2012;
   weight wtint2yr;
   cluster sdmvpsu;
   strata sdmvstra;
   class female (ref = first) hsq571 (ref = '0');
   model pad630 = female|hsq571 ridageyr / solution;
run;
```

PROC SURVEYLOGISTIC

- Fits weighted binary logistic, ordinal logistic, multinomial logistic, and probit models using pseudo-maximum likelihood estimation
- Built-in computation of odds ratios and design-adjusted Wald tests
- Same DOMAIN functionality as SURVEYREG
- Same CLASS functionality as SURVEYREG
- Selected measures of goodness of fit

SURVEYLOGISTIC: What's New?

- VARMETHOD = BOOTSTRAP
- One can request that output for only specific levels of a domain variable be displayed when using DOMAIN for analysis

SURVEYLOGISTIC: Example

- Fit a binary logistic regression model with the same two-way interaction (note the use of the desc option to model the probability of a 1):

```
proc surveylogistic data = nhanes2012b;
   weight wtint2yr;
   cluster sdmvpsu;
   strata sdmvstra;
   class hsd010 (reference = '3') female
       (reference = 'male') / param = ref;
   model paq665 (desc) = hsd010|female ridageyr;
   format female fm.;
run;
```

PROC SURVEYPHREG

- Fits weighted Cox Proportional Hazards models to time-to-event data (possibly censored) using pseudo-maximum likelihood
- Can accommodate the same class of models enabled by PHREG, including models with time-varying covariates
- Generation of selected residuals and weighted plots displaying model estimates

SURVEYPHREG: What's New?

- Specification of two time variables in the MODEL statement, indicating specific time points when a subject is at risk
- The HAZARDRATIO statement, for requesting estimates of hazard ratios
- VARMETHOD = BOOTSTRAP
- The ATRISK option in the SURVEYPHREG statement: sums of weights for the number of units and the number of events among those at risk at a specific time point
- Subpopulation analyses using the DOMAIN statement

SURVEYPHREG: Example

- Fit a proportional hazards model predicting the hazard of developing a major depressive order, with right censoring (MDE = 0):

```
proc surveyphreg data=c10_ncsr ;
   strata sestrat ;
   cluster seclustr ;
   weight ncsrwtsh ;
   class mar3cat (ref=first) sex (ref=last) ed4cat
     (ref=first) racecat (ref=first) / param=ref ;
   model ageonsetmde*mde(0) = intwage sex mar3cat
      ed4cat racecat ;
run ;
```

PROC SURVEYIMPUTE

- Imputes missing survey data using Fully Efficient Fractional Imputation (FEFI)
- **NOTE:** This is not the same thing as multiple imputation, and in many cases is a more efficient method of imputing missing values
- Creates replicate weights enabling variance estimation based on the imputed values
- See Heeringa et al. (2017), *Applied Survey Data Analysis*, Chapter 12, for more details

SURVEYIMPUTE: What's New?

- Now possible to compute bootstrap replicate weights, via VARMETHOD = BOOTSTRAP in the PROC SURVEYIMPUTE statement

SURVEYIMPUTE: Example

- Impute missing values using SURVEYIMPUTE, and then analyze the imputed data using jackknife replicate weights in SURVEYFREQ:

```
proc surveyimpute data=c12_fefi
method=FEFI varmethod=Jackknife;
   id seqn;
   class age4cat povcat riagendr
    marcat ridreth1 bmicat dbpcat;
   var age4cat povcat riagendr
    marcat ridreth1 bmicat dbpcat;
   impjoint povcat bmicat ;
   strata sdmvstra ;
   cluster sdmvpsu ;
   weight wtmec2yr ;
   output out=nhanesFEFI
      outjkcoefs=nhanesJKCOEFS;
run;
```

```
proc surveyfreq data=nhanesFEFI
 varmethod=jackknife ;
   weight impwt ;
   repweights imprepwt_: /
      jkcoefs=nhanesjkcoefs ;
   tables bmicat dbpcat povcat
    high_dbp age4cat ridreth1
    marcat / cl ;
run ;
```

PROC GLIMMIX

- Enables weighted estimation of multilevel models (a very specialized technique)
- Important SAS white paper: Zhu (2014)
- Weight scaling needs to be performed PRIOR to fitting the models using GLIMMIX
- For more details: West et al. (2015, AJPH)
- We consider example syntax from this paper for fitting a weighted multilevel logistic model

GLIMMIX: Example

```
proc glimmix data = tempwts3;
   class facility_id _gend2_ps _prtype_ps _newrace_ps;
   model _sexrr_ps (event = "1") = _gend2_ps _prtype_ps
  _newrace_ps primcare_ps _specialist_ps integteam_ps
  num200 _rw_fund_ps _facilemr_ps
      / solution link=logit dist=bin obsweight=level1wts;
   random int / subject=facility_id weight=level2wt;
   covtest glm;
run;
```

- **Note: this is the basic syntax (many of the options in GLIMMIX are also available when using weights)**

My Wish List

- PROC SURVEYREG: Design-adjusted linear regression diagnostics (Cook statistics, residual plots, measures of influence, etc.)
- *PROC SURVEYGLM*: No idea why this hasn't been programmed yet! One cannot fit Poisson models, negative binomial models, etc.
- *PROC SURVEYMIXED*, rather than running everything through PROC GLIMMIX (minor)

My Wish List

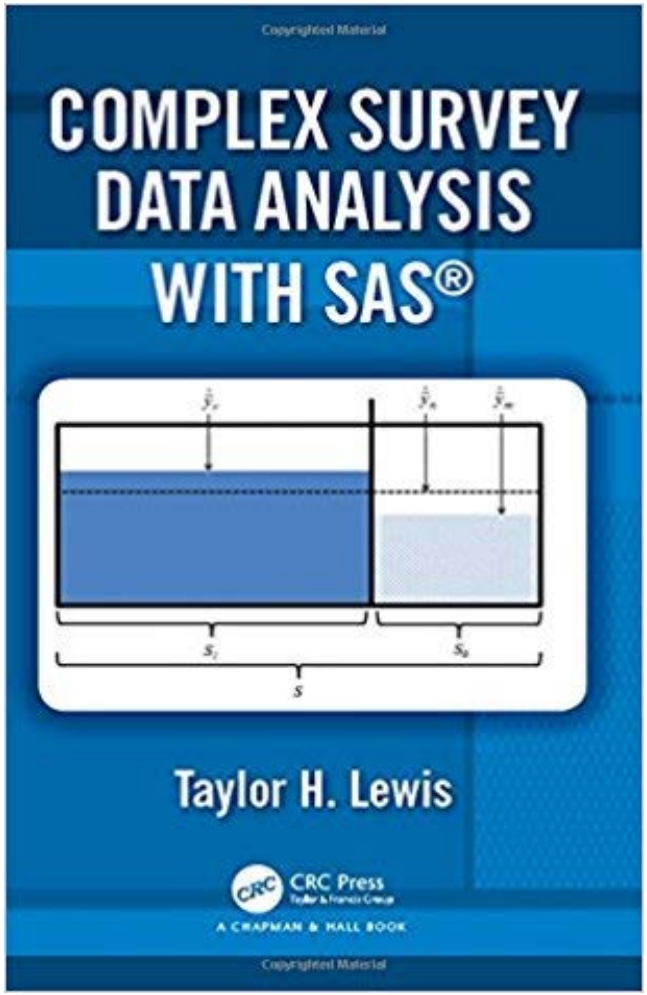
- Design-adjusted information criteria (AIC, BIC) for comparing alternative models, per recent work by Lumley and Scott (JSSAM, 2015)
- Goodness of fit tests for generalized linear models that account for complex sampling
- The ability to plot marginal predicted values based on weighted models, along with confidence intervals for the predictions

My Wish List

- An easy procedure for generating weighted Kaplan-Meier estimates of survivorship functions, in addition to correct design-adjusted standard errors for the estimates
- A procedure for fitting quantile regression models to complex sample survey data
- **Any others from people in the audience?**

A New Book!

- Taylor Lewis has written an excellent new book on this topic
- A fantastic desk reference if you perform these analyses a lot in SAS!



References

- Heeringa, S.G., West, B.T., and Berglund, P.A. (2017). *Applied Survey Data Analysis, Second Edition*. Chapman & Hall / CRC Press.
- SAS Online Documentation for SAS/STAT 14.3
- West, B.T., Beer, L., Gremel, W., Weiser, J., Johnson, C., Garg, S., and Skarbinski, J. (2015). Weighted Multilevel Models: A Case Study. *American Journal of Public Health*, 105(11), 2214-2215.
- Zhu, M. (2014). Analyzing Multilevel Models With the GLIMMIX Procedure. Cary, NC: SAS Institute Inc. Paper SAS026-2014.

Thanks! Questions?

- Thank you for attending today!
- Thanks to Brandy Sinco for the invitation!
- Selected examples were borrowed from this very helpful UCLA site:

https://stats.idre.ucla.edu/sas/seminars/sas-survey/

- For additional syntax examples, please visit:

http://isr.umich.edu/src/smp/asda

- For any follow-up inquiries or additional references, please email **bwest@umich.edu**.