Frequently Asked Questions and Answers

Applied Survey Data Analysis

February 5, 2014

Q: How do I calculate a design effect manually and should I use weighted or un-weighted numbers?

A: To correctly compute the design effect, the numerator will be "derived" using the estimated standard error from the design corrected means command (e.g., PROC SURVEYMEANS in SAS or svy: mean in Stata). Remember that you need the sampling variance which is se^2. The SRS variance of the proportion for the denominator is computed using the weighted estimate of the population proportion (not the unweighted estimates). Why? Because the weighted sample estimate is the unbiased estimate of the population proportion. Remember, we are trying to estimate the true SRS sampling variance P(1-P)/(n-1) by p(1-p)/(n-1). Our best estimate of this SRS variance uses the sample estimate, p, that is the best (i.e. unbiased) estimate of P. In samples with unequal probability sampling and weights, that unbiased estimate is the weighted sample estimate.

Q: How does one decide between multi-level modeling and design-based modeling?

A: This question is addressed in Chapter 12 of ASDA; see the Introduction (Section 12.1) for examples of applications where statistical models play a role. These areas of analysis include: survey data with multiple levels with important hierarchical relationships, longitudinal or repeated measures models, latent variable models (SEM), and small area estimation.

Additionally, the model-based approach allows for testing of hypotheses about the variation of covariate effects between clusters nested within the same stratum whereas the design-based approach does not. In situations where this is needed, a model-based approach is recommended. Also, see the recent references at the bottom of the ASDA web page on this topic.

Q: Where can I estimate percentiles along with correct replicated variances, and what is some example code?

A: Please see Section 5.3.5 of ASDA and Example 5.8 for an example of how to estimate population quantiles using Sudaan and WesVar. Sudaan uses the Taylor Series variance estimation method while WesVar uses the JRR or BRR repeated replication approach for variance estimation. There is also an example of using SAS v9.2+ posted in the analysis examples replication on the

ASDA website (Example 5.8) and this software also offers JRR and BRR methods as well (see the SAS documentation for more information).

Q: Are there design-based options for quantile regression?

A: There is an example of using R with a bootstrapping approach for complex sample adjusted variance for quantile regression on our website. Please see the Supplemental Code area for a link to the document.

Q: How does one construct an estimate of a population when using normalized weights?

A: Many survey weights (such as the NCS-R) are normalized to a mean 1.0 and sum of weights=n. If you would like to construct an estimate of a population, it requires expansion of the normalized weight back to a population estimate. See Example 5.3 in ASDA for details. The total size of the NCS-R survey population is 209,128,094 adults (US population as of 2002).

Q: When is a repeated replication method preferred to the Taylor Series Linearization method for variance estimation?

A: This topic is discussed in Chapter 3 of ASDA (Section 3.6.3). Statistically speaking, there are a few situations where a repeated replication approach is preferred. Note that TSL is useful if the estimate can be expressed as a function of sample totals and this method requires analytic manipulations and computation of derivatives. Therefore, the TSL is not directly suitable for percentiles such as the median or functions of percentiles and for these sorts of statistics, the repeated replication approach is a recommended choice.

Another common and more practical reason to use replication methods arises when data publishers release only replicate weights rather than design variables such as stratum and SECU (or PSU). In practice, many data publishers concerned about confidentiality protection will provide only replicate weights and this dictates the use of a JRR, BRR or other type of bootstrap approach for correct variance estimation. See Table 3.2 of ASDA for a comparison of results from TSL, JRR, and BRR methods.

Q: For logistic regression using the *svy: logistic* command in Stata, why don't you report the constant?

A: Unlike *svy: logit, svy: logistic* automatically computes odds ratios (exponentiated versions of the coefficients), and the exponentiated intercept has no meaning. This is why we do not report it in the textbook.