NATIONAL SURVEY ON DRUG USE AND HEALTH

A COMPARISON OF VARIANCE ESTIMATION METHODS FOR REGRESSION ANALYSES WITH THE MENTAL HEALTH SURVEILLANCE STUDY CLINICAL SAMPLE

Substance Abuse and Mental Health Services Administration
Center for Behavioral Health Statistics and Quality
Rockville, Maryland

June 2017

This page intentionally left blank

NATIONAL SURVEY ON DRUG USE AND HEALTH

A COMPARISON OF VARIANCE ESTIMATION METHODS FOR REGRESSION ANALYSES WITH THE MENTAL HEALTH SURVEILLANCE STUDY CLINICAL SAMPLE

RTI Project No. 0213984.100.001.010
Contract No. HHSS283201300001C

RTI Authors:
Phillip S. Kott
Dan Liao

SAMHSA Authors:
Matthew Williams
Sarra Hedden

Project Director:
David Hunter

SAMHSA Project Officer:
Peter Tice

For questions about this report, please e-mail Peter.Tice@samhsa.hhs.gov.

Prepared for Substance Abuse and Mental Health Services Administration, Rockville, Maryland

Prepared by RTI International, Research Triangle Park, North Carolina

June 2017

Recommended Citation: Center for Behavioral Health Statistics and Quality. (2017). *National Survey on Drug Use and Health: A Comparison of Variance Estimation Methods for Regression Analyses with the Mental Health Surveillance Study Clinical Sample*. Substance Abuse and Mental Health Services Administration, Rockville, MD.

Acknowledgments

This methodological document was prepared for the Substance Abuse and Mental Health Services Administration (SAMHSA), Center for Behavioral Health Statistics and Quality (CBHSQ), by RTI International (a registered trademark and a trade name of Research Triangle Institute). In addition to the listed authors, significant contributors include Rachel Harter at RTI and Jonaki Bose and Arthur Hughes at CBHSQ. Debbie Bond word processed and formatted the report, while Richard Straw copyedited it and Teresa Bass coordinated its web production. One of the SAMHSA contributors, Arthur Hughes, has retired from SAMHSA.

Table of Contents

Chapter **Page**

Appendix

This page intentionally left blank

List of Tables

Table **Page**

This page intentionally left blank

1. Overview

The National Survey on Drug Use and Health (NSDUH), conducted by the Substance Abuse and Mental Health Services Administration (SAMHSA), is one of the primary sources of data for population-based prevalence estimates of substance use and mental health indicators in the United States. The NSDUH interview includes several self-administered indicators of mental health, such as assessments of lifetime and past year major depressive episode (MDE), past month and past year psychological distress and functional impairment, as well as past year suicidality. From 2008 to 2012, a subsample of NSDUH adult respondents was selected to participate in the Mental Health Surveillance Study (MHSS), which was a telephone interview that included clinical assessments of the presence of selected mental disorders. MHSS clinicians administered semistructured diagnostic interviews to this subsample to assess the presence of selected mental disorders (Aldworth et al., 2010).

The main purpose of the MHSS clinical component was to generate prevalence estimates of serious mental illness (SMI) and any mental illness (AMI) among adults aged 18 years or older at the national and state levels by developing a statistical model that could be applied annually to the full NSDUH adult sample. In addition to that purpose, the 2008 to 2012 MHSS clinical data have been used to generate nationally representative prevalence estimates of past year mental disorders among the adult civilian, noninstitutionalized population in 2008 to 2012 across a wide spectrum of diagnostic categories that do not depend on a statistical model. For more details, see the Center for Behavioral Health Statistics and Quality (CBHSQ) data review by Karg et al. (2014).

Kott, Bose, Hedden, Liao, and Hughes (2014) and CBHSQ (2016) described how direct (nonmodeled) mental disorder prevalence estimates (and other estimated means) covering the 2008 to 2012 time period were computed using clinical sample weights that had undergone a number of calibration adjustments. The 2016 report focused on the last adjustment—the annual calibration of the clinical sample to the control totals computed from the full NSDUH adult sample (i.e., the "poststratification" adjustment). Several Taylor series linearization-based methods for computing the standard error (SE) measures of the prevalence were discussed. Based on the analyses in that report, SAMHSA decided to compute SEs for prevalence estimates using a linearization methodology that accounts for the annual calibration of the clinical sample to adult NSDUH control totals. The recommended methodology typically leads to a reduction in SEs and can be implemented using the WTADJX procedure in SUDAAN®, Release 11.0 (RTI International, 2012).[1]

A standard "naïve" linearization method, which treats final analysis weights as if they were nonrandom, is used to measure SEs for all estimates derived from the full NSDUH sample (CBHSQ, 2014a). For prevalences and totals derived from the MHSS clinical sample, however, analysts can use a more sophisticated version of linearization that captures the potential variance reduction due to the calibration of the clinical sample to controls estimated from the larger

[1] Some technical details of this calibration-weighting software routine are described in Section A.1 of Appendix A in this report.

NSDUH sample. The final MHSS clinical sample weights were calibrated to the following annual adult totals estimated from the full NSDUH:

- six categories of gender (male and female) by age (18 to 25, 26 to 34, 35 or older) categories,
- four race/ethnicity categories (Hispanic, non-Hispanic white, non-Hispanic black, other),
- past year suicidal thoughts,
- past year and lifetime MDE,
- interaction between a psychological distress measure based on the Kessler-6 (K6) score[2] and the three age categories, and
- interaction between a functional impairment measure based on either the World Health Organization Disability Assessment Schedule (WHODAS) score or the Sheehan Disability Scale (SDS) score for the 2008B sample[3] and the three age categories.

In addition to estimating totals computed with final MHSS clinical sample weights meeting these targets (note that a mean is a ratio of totals), the calibration-weighting software routine (WTADJX) limited the size of the final individual weights depending on the year (2008 through 2012).

The purpose of this report is to compare alternative methods for producing measures of SEs for regression models for the MHSS clinical sample with the goal of producing more accurate and potentially smaller SEs. As with SE estimates for means and totals, smaller SEs for regression coefficients lead to narrower confidence intervals (CIs) and increased power for tests of statistical significance. The software routine described above, while producing calibrated weights and measuring SEs of estimated means and totals computed with those weights, cannot produce measures of SEs for regression coefficients estimated with those weights. In light of this limitation, analysts have been using a naïve linearization routine that ignores the potential impact of weighting the MHSS clinical sample to the full adult NSDUH sample on variance estimates of the coefficients from a logistic or linear regression. Therefore, this study was conducted to provide alternative methods for producing SEs for parameter estimates in regression models that fully account for the weights created for the MHSS sample.

Analyses that are described in this report assess whether the current use of the naïve linearization method is justified for regression analyses. To that end, the SEs of certain logistic regression coefficients are measured with naïve linearization and a particular jackknife replication technique, the latter of which captures the impact of the annual calibration of the MHSS clinical sample to the full adult NSDUH sample (the particular jackknife and alternative replication techniques are described in Appendix A).

The analyses find that there is a marked tendency for SE measures to decrease when the calibration is accounted for using this version of jackknife replication. Moreover, this jackknife,

[2] For details on the K6 score, see the 2012 MHSS design and estimation report (CBHSQ, 2014b).

[3] For details on the WHODAS and SDS scores and the 2008B sample, see the report mentioned in footnote 2.

although asymptotically unbiased, may have a tendency to be too large in finite samples. As a result, using this jackknife method, which requires adding 100 sets of jackknife replication weights to the clinical sample data, would likely remove some but not all of the tendency of SE measures for regression coefficients to be larger than the actual SEs, a tendency that leads to statistical CIs that are larger than they should be. Using this jackknife method will often be an improvement, producing narrower CIs and sharper inferences than the naïve linearization method currently used by analysts. That is, using the jackknife method rather than the current method is more likely to reject a null hypothesis that is incorrect; at the same time, the probability of rejecting a correct hypothesis at, for example, the .05 level will be only 5 percent.

This page intentionally left blank

2. Estimating MHSS Statistics with a (Traditional) Jackknife

The Mental Health Surveillance Study (MHSS) clinical sample has 100 variance strata, each with two variance primary sampling units (PSUs). To create one replicate of naïve jackknife weights, one would randomly select one of the variance PSUs in a single variance stratum and remove it from the replicate sample while doubling the weights of the respondents in the other PSU in the stratum to compensate. The weights in all of the other variance strata remain unchanged. The resulting weights make up one set of replicate jackknife weights. Repeating this process for each variance stratum creates 100 sets of naïve replicate weights.

As described more fully in Appendix A, a jackknife variance estimator for a statistic t computed with all of the weights has the following form:

$$v_J(t) = \sum \left(t - t^{(r)}\right)^2, \tag{1}$$

where $t^{(r)}$ is analogous to t but is computed using the r^{th} set of jackknife replicate weights. The summation is over 100 replicate estimates in this case, each computed with its own set of replicate weights. When applied using the final MHSS clinical sample weights, the variance estimator in equation (1) is asymptotically equivalent to the naïve linearization variance estimator computed by ignoring the annual calibration to the adult National Survey on Drug Use and Health (NSDUH) sample.

By contrast, it is not difficult to create calibrated jackknife replicate weights that account for the final calibration (which is called "the traditional jackknife" in Appendix A). First, create 100 sets of jackknife replicate weights from the input weights used in the final calibration step of the MHSS sample weighting, then apply the calibration routine described by Kott et al. (2014) and the Center for Behavioral Health Statistics and Quality (CBHSQ, 2016) to each set in turn (including the scaled full adult NSDUH sample weights[4]) to create 100 sets of calibration-adjusted jackknife weights. When equation (1) is computed with those jackknife replicate weights, the impact of the final calibration is captured.

One practical impediment to the jackknife for a calibrated estimator is the possibility that the calibration equations are not solvable with one or more sets of jackknife weights. That is, weighted totals computed with calibration-adjusted MHSS jackknife replicate weights cannot be forced to equal weighted totals computed with analogous NSDUH jackknife replicate weights. This happened 3 times out of 100 with the MHSS clinical data. In these cases, the restriction on the sizes of the final weights had to be relaxed.

[4] See Kott et al. (2014) and CBHSQ (2016) for a discussion of scaling weighting when combining years of MHSS clinical data.

Once jackknife weights have been created, they can be used with available software. Exhibit 1 shows the SUDAAN® and Stata® commands, which are only slight modifications from the default code that uses the naïve linearization method.

Exhibit 1. Jackknife Commands for Use within SUDAAN® and Stata®

In SUDAAN,
DESIGN = JACKWGTS;
WEIGHT MHFNLWGT;
JACKWGTS WFINAL1-WFINAL100/ADJJACK=1.
In Stata,
SVYSET [PW = MHFNLWGT], JKRWEIGHT(WFINAL1
WFINAL100, MULTIPLIER(1)) VCE(JACK) MSE

The following example of SAS-callable SUDAAN code could be used to produce estimates for any given outcome variable for a specified domain with the MHSS clinical data. This example code provides estimated means and totals and naïve measures of their standard errors (computed using Taylor series linearization) by respondent gender (IRSEX) and adult age category (CATAGMH2) for serious mental illness (SMI) using MHSS clinical interview (i.e., Structured Clinical Interview for DSM-IV [SCID]) data.

```
proc descript data= MHSS /*dataset name*/ filetype=sas design=wr;
nest MHVESTR MHVEREP;
weight MHFNLWGT;
var SCID_SMI;
/*outcome of interest */
class IRSEX CATAGMH2;
tables IRSEX CATAGMH2;
/*domain(s) of interest*/
OUTPUT mean semean total setotal;
run;
```

When using jackknife replicates, the bold portion of the standard code based on Taylor series linearization is replaced.

```
proc descript data= MHSS /*dataset name*/ filetype=sas design= jackwgts;
weight MHFNLWGT;
JACKWGTS JKNWT1-JKNWT100/ADJJACK=1;
var SCID_SMI;
/*outcome of interest */
class IRSEX CATAGMH2;
tables IRSEX CATAGMH2;
/*domain(s) of interest*/
OUTPUT mean semean total setotal;
run;
```

The same replacement of the **design**, **nest**, and **weight** commands applies when using other procedures, such as ***proc rlogist*** for logistic regression.

In Stata, the code for the naïve method is as follows:

svyset **mhverep** [pw = mhfnlwgt], **strata(mhvestr)**
svy: mean scid_smi, over(catagmh2 irsex)

The Stata code using jackknife replicates is as follows:

svyset [pw = mhfnlwgt], **jkrweight(jknwt1-jknwt100, multiplier(1))**
vce(jackknife) mse
svy: mean scid_smi, over(catagmh2 irsex)

For logistic regression, the ***svy: logit*** command is used instead of the ***svy: mean*** command.

This page intentionally left blank

3. Comparing the Jackknife and Linearization Measures of Standard Error

Kott et al. (2014) and the Center for Behavioral Health Statistics and Quality (CBHSQ, 2016) computed alternative standard error (SE) measures for 43 different estimated means of mental and substance use disorders. Table 1 displays summaries of how some SE measures for the 43 estimated means differ. Four SE measures, two old and two new, are compared. The naïve linearization (n-lin) and naïve jackknife (n-jk) measures do not account for the variance-reducing potential of the calibration weights. The calibration-adjusted linearization (lin) and calibration-adjusted jackknife (jk) do account for the potential (see Appendix A for more details). The summaries are distributions computed over the 43 estimates using this symmetric measure of average percentage change:[5]

$$Log\left(\frac{\text{SE of Estimate under Method 1}}{\text{SE of Estimate under Method 2}}\right) = Log\left(\text{SE under Method 1}\right) - Log\left(\text{SE under Method 2}\right). \quad (2)$$

Table 1 shows that there is virtually no difference between using the two naïve measures. Using the calibration-adjusted jackknife reduces the SE measure by roughly 7 percent on average relative to the naïve measures (the median decrease is roughly 3 percent). Using the calibration-adjusted linearization measure reduces the measured SE another 6 percent on average.

Table 1. Distribution of Differences among Standard Error Measures for MHSS Estimates

Difference (on Log Scale)	Mean	Median	Minimum	First Quartile	Third Quartile	Maximum
$Log\left(SE_{n\text{-}jk} / SE_{n\text{-}lin}\right)$	0.0014	0.0012	-0.0069	-0.0019	-0.0002	0.0010
$Log\left(SE_{jk} / SE_{n\text{-}jk}\right)$	-0.0714	-0.0293	-0.7144	-0.1910	0.0319	0.2992
$Log\left(SE_{jk} / SE_{n\text{-}lin}\right)$	-0.0729	-0.0311	-0.7213	-0.1920	0.0317	0.2991
$Log\left(SE_{lin} / SE_{jk}\right)$	-0.0623	-0.0462	-0.3837	-0.0712	-0.0186	0.1095

jk = jackknife; lin = linearization; MHSS = Mental Health Surveillance Study; n = naïve; SE = standard error.
NOTE: Distribution is across the measured standard errors for 43 estimates.

Table 2 looks at the estimated domain means for the 43 variables. These are three age groups (18 to 25, 26 to 49, 50 or older), gender, and four races/ethnicities (non-Hispanic white, non-Hispanic black, non-Hispanic other, and Hispanic). Again, the two naïve measures show almost no difference, but the SE using the jackknife measure is now, on average, only around 2.5 percent lower than the naïve measures for SE, while the linearization measure remains, on average, roughly an additional 6 percent lower than the jackknife measure.

[5] The $Log\left(a / b\right) \approx \left(a - b\right) / b$, when a is within 25 percent of b. In addition, $Log\left(a / b\right) = -Log\left(b / a\right)$, while $\left(a - b\right) / b \neq -\left(b - a\right) / a$.

Table 2. Distribution of Differences among Standard Error Measures for MHSS Estimated Domain Means (Race/Ethnicity, Age Group, Gender)

Difference (on Log Scale)	Mean	Median	Minimum	First Quartile	Third Quartile	Maximum
$Log\left(SE_{n\text{-}jk} / SE_{n\text{-}lin}\right)$	-0.0018	-0.0003	-0.0550	-0.0034	0.0013	0.0258
$Log\left(SE_{jk} / SE_{n\text{-}jk}\right)$	-0.0255	-0.0180	-0.7072	-0.0723	0.0279	0.6045
$Log\left(SE_{jk} / SE_{n\text{-}lin}\right)$	-0.0273	-0.0180	-0.7204	-0.0709	0.0272	0.6070
$Log\left(SE_{lin} / SE_{jk}\right)$	-0.0634	-0.0314	-0.7132	-0.0877	-0.0032	0.1204

jk = jackknife; lin = linearization; MHSS = Mental Health Surveillance Study; n = naïve; SE = standard error.
NOTE: Distribution is across the measured standard errors for 387 estimated domain means.

Several theoretical reasons can explain why the linearization estimate may have a downward bias while the jackknife measure has an upward bias. Both are asymptotically unbiased, but the samples are finite. The linearization estimator plugs in a sample error term in place of the full-population difference between the variable of interest and a prediction of its value based on the calibration variables. Because the sample value has been designed to fit the sample (rather than the population), the linearization variance estimator tends to underestimate.

The jackknife measure, by contrast, implicitly uses the correct errors but requires the jackknife replicate weights to adjust to meet the calibration targets. The added variability in the replicate weights tends to overestimate variances. It may be that either having the constraints on the sizes of jackknife weights or lifting those constraints in three cases increases the tendency for the jackknife measure to overestimate variances.

4. Logistic Regression Examples

This chapter describes the investigation of standard error (SE) measures for some logistic regression coefficients. In this chapter's tables, unlike the ones in Chapter 3, there are no calibration-adjusted linearization SE measures because the calibration-adjusted linearization method is not applicable for regression analysis.

Table 3 looks at the estimated coefficients and SE measures for models with *any disorder* (SCIDANY_R), *anxiety disorder* (ANXIETY_2), and *substance use disorder* (SUBDIS_2), respectively, as the dependent variables and the three age groups, gender, and four race/ethnicity categories as the explanatory variables. The coefficients for the first age group, female, and non-Hispanic white are missing from the table because they were treated as reference levels.

Table 3. Estimated Coefficients and Their Standard Error Measures for Three Simple Logistic Regression Models

Dependent Variable/ Independent Variable	Beta	SE: Naïve Linearization	SE: Naïve Jackknife	SE: Cal-adj Jackknife
Any Disorder				
Intercept	-0.78	0.1663	0.1598	0.1438
Aged 26 to 49	-0.28	0.1385	0.1384	0.1343
Aged 50 or Older	-0.85	0.2090	0.1974	0.1451
Male	0.09	0.1259	0.1231	0.1132
Non-Hispanic Black	-0.25	0.1931	0.1943	0.1916
Non-Hispanic Other	-0.43	0.2345	0.2325	0.2225
Hispanic	0.05	0.3160	0.2910	0.2322
Anxiety Disorder				
Intercept	-3.54	0.5618	0.5009	0.3942
Aged 26 to 49	0.02	0.3066	0.2879	0.2838
Aged 50 or Older	-0.29	0.6452	0.5320	0.4087
Male	1.14	0.2795	0.2554	0.1978
Non-Hispanic Black	-0.84	0.3011	0.2900	0.3344
Non-Hispanic Other	-1.11	0.3434	0.3402	0.3102
Hispanic	0.58	0.7476	0.5749	0.4247
Substance Use Disorder				
Intercept	-1.18	0.1556	0.1546	0.1706
Aged 26 to 49	-0.64	0.1936	0.1954	0.2137
Aged 50 or Older	-1.80	0.2583	0.2554	0.2428
Male	-1.00	0.1803	0.1794	0.1646
Non-Hispanic Black	0.26	0.2581	0.2584	0.2586
Non-Hispanic Other	-0.79	0.3350	0.3404	0.3608
Hispanic	-0.15	0.2840	0.2972	0.3920

Cal-adj = calibration-adjusted; SE = standard error.

Table 4 summarizes the differences across the 21 estimated coefficients using the log measure in Chapter 3, and Table 5 looks at *p* values for adjusted Wald F statistics on the impact of age group, gender, and race ethnicity on each of the three dependent variables. Unlike the *t* values of regression coefficients, the choice of reference level does not affect F statistics, which is why they are reviewed here.

Table 4. Distribution of Differences among Coefficient Standard Error Measures for Three Simple Logistic Regression Models

Difference (on Log Scale)	Mean	Median	Minimum	First Quartile	Third Quartile	Maximum
$Log\left(SE_{n\text{-}jk} / SE_{n\text{-}lin}\right)$	-0.0441	-0.0114	-0.2627	-0.0629	-0.0006	0.0454
$Log\left(SE_{jk} / SE_{n\text{-}jk}\right)$	-0.0691	-0.0507	-0.3080	-0.2257	0.0007	0.2768
$Log\left(SE_{jk} / SE_{n\text{-}lin}\right)$	-0.1132	-0.0771	-0.5656	-0.3082	0.0021	0.3223

jk = jackknife; lin = linearization; n = naïve; SE = standard error.

NOTE: Distribution is across the measured standard errors for 21 of the coefficients.

Table 5. Estimated *p* Values of the Adjusted Wald F Tests for Three Simple Logistic Regression Models Using Different Standard Error Measures

Dependent Variable/ Independent Variable	*p* Value: Naïve Linearization	*p* Value: Naïve Jackknife	*p* Value: Cal-adj Jackknife
Any Disorder			
Age	0.0005	0.0002	0.0000
Gender	0.4838	0.4743	0.4363
Race/Ethnicity	0.0763	0.0790	0.1344
Anxiety Disorder			
Age	0.8143	0.7436	0.5365
Gender	0.0001	0.0000	0.0000
Race/Ethnicity	0.0003	0.0002	0.0001
Substance Use Disorder			
Age	0.0000	0.0000	0.0000
Gender	0.0000	0.0000	0.0000
Race/Ethnicity	0.0708	0.0763	0.0850

Cal-adj = calibration-adjusted.

Tables 6 through 8 mimic the previous three tables with a single logistic model of any disorder against the existence of past exposure to one or more potentially traumatic episodes (PTEs) and a set of additional explanatory variables selected for their apparent significance after some model fitting. In Table 6, the reference level for a categorical variable with more than two levels is the larger of the first or last level.

Tables 4 and 7 show that, unlike with estimated means, SE measures computed with the naïve jackknife measure are, on average, smaller than those computed with the naïve linearization measure. The means in both tables are much larger than the medians (roughly 4.4 vs. 1.1 percent for the means and 3.0 vs. 1.3 percent for the medians). This may be because the largest naïve linearization estimates displayed in Tables 3 and 6 are outliers. Observe the .6452

and .7476 in Table 3 and the .4249 and .3172 in Table 5. Unfortunately, there is no way to verify that these are indeed outliers.

Tables 4 and 7 also show that the calibration-adjusted jackknife SE measure is 4 to 7 percent lower, on average, than that of the naïve jackknife measure. The differences tend to be less than those between the calibration-adjusted jackknife and the naïve linearization measures. This may be of some concern because of the possibility that while the currently used naïve linearization SE measures may be too large, the calibration-adjusted jackknife measure may be too small. Any worry about the calibrated jackknife underestimating SEs, however, is mitigated by the known and demonstrated tendency of the jackknife to overestimate SEs for means and domain means. These means can themselves be expressed as linear and logistic regression coefficients.

Table 6. Estimated Coefficients and Their Standard Error Measures for a Logistic Regression Model of Any Disorder on Potentially Traumatic Episodes and Other Variables

Independent Variable	Beta	SE: Naïve Linearization	SE: Naïve Jackknife	SE: Cal-adj Jackknife
Intercept	-1.12	0.2303	0.2305	0.2099
Past Exposure to a Potentially Traumatic Episode	0.58	0.1146	0.1104	0.1028
Aged 26 to 49	-0.13	0.1459	0.1462	0.1474
Aged 50 or Older	-0.47	0.2152	0.2007	0.1878
Non-Hispanic Black	-0.34	0.1905	0.1925	0.1885
Non-Hispanic Other	-0.49	0.2447	0.2433	0.2333
Hispanic	-0.19	0.2791	0.2506	0.2246
Small Metropolitan County	-0.12	0.1374	0.1331	0.1405
Nonmetropolitan County	-0.35	0.1636	0.1601	0.1479
Married	-0.13	0.1350	0.1355	0.1351
Widowed	-0.90	0.4249	0.3963	0.3924
Divorced/Separated	0.10	0.1996	0.2016	0.2275
Less Than High School	0.57	0.3172	0.2835	0.1987
High School Graduate	0.10	0.1484	0.1465	0.1475
Some College	-0.07	0.1403	0.1397	0.1423
Health Insurance	-0.21	0.1662	0.1650	0.1703
Moves	0.27	0.0921	0.0907	0.0866
Military Service	-0.57	0.2026	0.2006	0.1810
Attends Fewer Than Three Religious Services in a Year	0.30	0.1695	0.1548	0.1278

Cal-adj = calibration-adjusted; SE = standard error.

Table 7. Distribution of Differences among Coefficient Standard Error Measures for a Logistic Regression Model of Any Disorder on Potentially Traumatic Episodes and Other Variables

Difference (on Log Scale)	Mean	Median	Minimum	First Quartile	Third Quartile	Maximum
$Log\left(SE_{n\text{-}jk} / SE_{n\text{-}lin}\right)$	- 0.0299	-0.0128	-0.1125	-0.0695	0.0010	0.0107
$Log\left(SE_{jk} / SE_{n\text{-}jk}\right)$	-0.0502	-0.0420	-0.3554	-0.0936	0.0082	0.1205
$Log\left(SE_{jk} / SE_{n\text{-}lin}\right)$	-0.0801	-0.0610	-0.4679	-0.1130	0.0099	0.1308

jk = jackknife; lin = linearization; n = naïve; SE = standard error.

NOTE: Distribution is across the measured standard errors for one of nine coefficients.

Finally, note in Tables 5 and 8 how significance levels are changed by using the three different measures. Surprisingly, using the calibration-adjusted jackknife replicate weights in place of either naïve jackknife or naïve linearization replicate weights removes the significance at the .1 level of difference among races/ethnicities in any disorder (Table 5). In the expected direction, using the calibrated-jackknife measure increases the significance level of age group on any disorder in the logistic regression, including past exposure to a PTE (making it significant at the .05 level; see Table 8). The calibrated-jackknife measure also increases the significance level of the county's urbanicity, the respondent's marital status, education, and attending fewer than three religious services in a year.

Table 8. Estimated *p* Values of the Adjusted Wald F Tests for a Logistic Regression Model of Any Disorder on Potentially Traumatic Episodes and Other Variables Using Different Standard Error Measures

Independent Variable	Beta	*p* Value: Naïve Linearization	*p* Value: Naïve Jackknife	*p* Value: Cal-adj Jackknife
Past Exposure to a Potentially Traumatic Episode	0.58	0.0000	0.0000	0.0000
Age Group	-0.13	0.0937	0.0617	0.0229
Race/Ethnicity	-0.19	0.0238	0.0264	0.0303
County's Urbanicity	-0.12	0.1076	0.1009	0.0674
Marital Status	-0.13	0.0730	0.0552	0.0480
Education	0.57	0.1778	0.1175	0.0069
Health Insurance	-0.21	0.2122	0.2080	0.2232
Moves	0.27	0.0041	0.0036	0.0024
Military Service	-0.57	0.0056	0.0052	0.0020
Attends Fewer Than Three Religious Services in a Year	0.30	0.0840	0.0590	0.0227

Cal-adj = calibration-adjusted.

5. Conclusions

The jackknife variance estimator using the calibration-adjusted jackknife measure shows an average decrease in standard errors (SEs) compared with the naïve methods for variance estimation for both totals (3 to 7 percent) and regression coefficients (8 to 11 percent) based on selected estimates from the Mental Health Surveillance Study (MHSS) clinical sample. The Taylor series linearization method in the WTADJX procedure provides an additional decrease in SE (6 percent) for totals beyond the savings from the jackknife method, but it is not available for regression analyses.[6]

Overall, these savings (reductions in SEs) lead to more precise analyses with narrower confidence intervals and more power to detect statistical significance. Therefore, it is recommended that these calibration-adjusted methods be used for all analyses of the MHSS clinical dataset, with the calibration-adjusted linearization method still recommended for means and totals and the calibration-adjusted jackknife method recommended for regression.

The calibration-adjusted versions of the jackknife and Taylor series linearization methods for estimating SEs both tend to yield (but not always) smaller SEs than those computed via naïve Taylor series linearization, which ignores the annual calibration of the MHSS sample to the corresponding annual adult full National Survey on Drug Use and Health (NSDUH) sample. Linearization may produce SEs that are slightly smaller than they should be, but jackknife SEs are likely to be at least equally biased in the other direction. Moreover, the impact of the constraint on the weights of the jackknife replicates—and of lifting those constraints as was done in three cases—has never been theoretically or empirically fully evaluated. A sensitivity analysis is encouraged where multiple methods are run and results are compared.

[6] This procedure is available in SUDAAN®, Release 11.0 (RTI International, 2012).

This page intentionally left blank

References

Aldworth, J., Colpe, L. J., Gfroerer, J. C., Novak, S. P., Chromy, J. R., Barker, P. R., Barnett-Walker, K., Karg, R. S., Morton, K. B., & Spagnola, K. (2010). The National Survey on Drug Use and Health Mental Health Surveillance Study: Calibration analysis. *International Journal of Methods in Psychiatric Research, 19*(Suppl. 1), 61-87. https://doi.org/10.1002/mpr.312

Center for Behavioral Health Statistics and Quality. (2014a). *2012 National Survey on Drug Use and Health: Methodological resource book (Section 12, Person-level sampling weight calibration)*. Retrieved from https://www.samhsa.gov/data/

Center for Behavioral Health Statistics and Quality. (2014b). *2012 National Survey on Drug Use and Health: Methodological resource book (Section 16a, 2012 Mental Health Surveillance Study: Design and estimation report)*. Retrieved from https://www.samhsa.gov/data/

Center for Behavioral Health Statistics and Quality. (2016). *2014 National Survey on Drug Use and Health: Mental health estimates computed directly from the clinical sample of the Mental Health Surveillance Study and measures of their standard errors*. Retrieved from https://www.samhsa.gov/data/

Judkins, D. R. (1990). Fay's method for variance estimation. *Journal of Official Statistics, 6*, 223-239. Retrieved from http://www.jos.nu/Articles/abstract.asp?article=63223

Karg, R. S., Bose, J., Batts, K. R., Forman-Hoffman, V. L., Liao, D., Hirsch, E., Pemberton, M. R., Colpe, L. J., & Hedden, S. L. (2014, October). *Past year mental disorders among adults in the United States: Results from the 2008-2012 Mental Health Surveillance Study*. CBHSQ Data Review. Retrieved from https://www.samhsa.gov/data/

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology, 32*, 133-142. Retrieved from https://www.academia.edu/4229836/Using_Calibration_Weighting_to_Adjust_for_Nonresponse_and_Coverage_Errors

Kott, P. S. (2011). A nearly pseudo-optimal method for keeping calibration weights from falling below unity in the absence of nonresponse or frame errors. *Pakistan Journal of Statistics*, 27, 391-396. Retrieved from https://www.semanticscholar.org/paper/A-Nearly-Pseudo-optimal-Method-for-Keeping-Calibra-Kott/aa2d7e3dbcdb7f6b288160cab7ac364b1b148b82

Kott, P. S., Bose, J., Hedden, S., Liao, D., & Hughes, A. (2014). Mental health estimates computed directly from the clinical sample of the Mental Health Surveillance Study and measures of their standard errors. In *Proceedings of the 2014 Joint Statistical Meetings, American Statistical Association, Survey Research Methods Section, Boston, MA* (pp. 1458-1472). Retrieved from https://ww2.amstat.org/sections/srms/Proceedings/y2014/Files/311701_85604.pdf

Kott, P. S., & Liao, D. (2015). One step or two? Calibration weighting from a complete list frame with nonresponse. *Survey Methodology, 41*, 165-181. Retrieved from http://www.statcan.gc.ca/pub/12-001-x/2015001/article/14172-eng.pdf

Krewski, D., & Rao, J. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics, 9*, 1010-1019. https://doi.org/10.1214/aos/1176345580

RTI International. (2012). *SUDAAN language manual, Release 11.0.* Research Triangle Park, NC: Author.

Rust, K. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics, 1*, 381-397. Retrieved from http://www.jos.nu/Articles/abstract.asp?article=14381

SAS Institute Inc. (2008). *SAS/STAT® 9.2 user's guide*. Cary, NC: Author.

Appendix A: Description of Alternative Replicate Methods

The Mental Health Surveillance Study (MHSS) complex sample design features two variance primary sampling units (PSUs)[7] within each of 100 variance strata (see Chapter 3 in Center for Behavioral Health Statistics and Quality [CBHSQ], 2014b). For variance estimation purposes, the PSUs are treated as if they were selected with replacement. Moreover, the weights before the final calibration, hereinafter called the "calibration," are treated as design weights (i.e., as the inverses of overall selection probabilities). See Kott et al. (2014) and CBHSQ (2016) for more details.

Two general types of replicate weights can be reasonably computed given such a two-variance-PSU-per-variance-stratum design: Fay's balanced repeated replication (F-BRR) weights and jackknife weights (see Judkins, 1990, and the references therein). This appendix discusses the pros and cons of four potential methods of creating replicate weights, whether F-BRR or jackknife weights. Each method results in standard error (SE) measures with expectations nearly (i.e., asymptotically) equal to those produced by the WTADJX procedure for the prevalence estimate.[8] This appendix also explains why a stratified bootstrap is not recommended for this purpose.

The four potential methods of F-BRR and jackknife weighting parallel each other. First, one can create either 100 or 50 sets of replicate weights. Second, one can create those weights in either of two ways. *Traditional replication* first replicates the before-calibration weights, then calibrates each set of replicate weights in an exact mirror of how the original before-calibration weights were calibrated. *Recalibrated replication* replicates the calibrated weights, then recalibrates each set using a simplified version of calibration.

Section A.1 reviews how the linearized variance of an MHSS prevalence estimate is computed with WTADJX. Section A.2 describes constructing jackknife weights for the MHSS clinical sample, and Section A.3 describes constructing the F-BRR weights. Both of the latter two sections also discuss variants of less interest. For example, with 100 strata, one can construct either 100 or 200 sets of jackknife replicate weights, but there is virtually no gain from doing the latter. Similarly, when the weights are calibrated, there is a reason for preferring Fay's BRR over simple BRR (Section A.4 discusses the difference). A stratified bootstrap has the same weakness as simple BRR and is usually not as efficient (i.e., the variance of a BRR variance estimator is never greater than that of a stratified bootstrap). Section A.4 also explains why the jackknife method was chosen for the analysis in the main text of this report.

One restriction placed on the methods chosen for this appendix is that their expectations need to be nearly equal to the SE measures for the prevalence estimates produced by WTADJX. Another restriction is that the methods need to be computable in SUDAAN® (RTI International, 2012) and SAS/STAT® (SAS Institute Inc., 2008).

[7] PSUs have been called "replicates" elsewhere, but that term has a different meaning when replicate variance estimation is used.

[8] When a variance estimator is (nearly) unbiased, its square root, the analogous SE measure, is nearly unbiased. The WTADJX procedure is explained in RTI International (2012).

A.1 Linearized Variance of an MHSS Prevalence Estimate

The MHSS clinical samples were calibrated separately in each year. The following, adapted from Kott et al. (2014) and CBHSQ (2016), describes how calibration was done in a particular year. The 2008 National Survey on Drug Use and Health (NSDUH) main adult sample was randomly divided in half with a separate MHSS clinical subsample selected from each half.[9] Therefore, the 2008A and 2008B half samples are treated as if they were sampled from different years.

Let S denote an annual NSDUH main adult respondent sample, w_k the weight attached to main survey respondent k, and q_k the respondent's clinical sample weight after all adjustments for coverage and nonresponse but before the final calibration to the NSDUH main adult sample. By convention, $q_k = 0$ when adult k is a respondent to the NSDUH main interview but either was not subsampled or did not respond to the clinical interview.

Let $a_k = q_k / w_k$. This is given a vector of calibration variables $\mathbf{z}_k$ (to be defined shortly) and a scalar $T_k = .05\left(\sum_S w_k\right)$ when k is from the 2008A sample, and $.04\left(\sum_S w_k\right)$ otherwise, where the summations are over the sample in the same year as k.

The calibration adjustment factor for a clinical interview respondent had this form:

$$f_k = \frac{\exp\left(\frac{U_k}{U_k - 1} a_k \mathbf{z}_k{}^T \mathbf{g}\right)}{1 + \left[\exp\left(\frac{U_k}{U_k - 1} a_k \mathbf{z}_k{}^T \mathbf{g}\right) - 1\right] \Big/ U_k}, \tag{A1}$$

where $\mathbf{g}$ was chosen by successive linearizations (Newton's method) to satisfy the calibration equation:

$$\sum_S q_k f_k \mathbf{z}_k = \sum_S w_k \mathbf{z}_k, \tag{A2}$$

and $U_k = T_k / q_k$. The w_k in the NSDUH main adult respondent sample had been calibrated so that their sum equals the adult population size.

The vector $\mathbf{z}_k$ consists of the following components: (a) indicators for six categories of gender (male and female) by age (18 to 25, 26 to 34, 35 or older), (b) indicators for four race/ethnicity categories (Hispanic, non-Hispanic white, non-Hispanic black, other), (c) an indicator for past year suicidal thoughts, (d) indicators from the NSDUH main interview for a past year and lifetime major depressive episode (MDE), (e) interaction terms between an alternative Kessler-6 (K6) score and the three age categories, and (f) interaction terms between an alternative World Health Organization Disability Assessment Schedule (WHODAS) score

[9] Serious mental illness (SMI) was modeled using a different measure of functional impairment in each 2008 subsample.

(or an alternative Sheehan Disability Scale [SDS] score for the 2008B sample[10]) and the three age categories.

The a_k in equation (A1) renders the adjustment factors nearly pseudo-optimal (Kott, 2011). One can express a calibration-weighted total $t = \sum_S \omega_k y_k$, where ω_k is the calibration weight for adult k, as

$$t = \sum_S w_k \mathbf{z}_k{}^T \mathbf{b} + \sum_S \omega_k e_k = \sum_S w_k \left(\mathbf{z}_k{}^T \mathbf{b} + [\omega_k / w_k] e_k \right),$$

where, for technical reasons explained in Kott and Liao (2015), the quasi-randomization regression coefficient is

$$\mathbf{b} = \left(\sum_S q_k f_k \left[(U_k - f_k) / (U_k - 1) \right] a_k \mathbf{z}_k \mathbf{z}_k^T \right)^{-1} \sum_S q_k f_k \left[(U_k - f_k) / (U_k - 1) \right] a_k \mathbf{z}_k y_k, \quad \text{(A3)}$$

and $e_k = y_k - \mathbf{z}_k{}^T \mathbf{b}$.

This decomposition is effectively what WTADJX does internally. Each $\mathbf{x}_k = a_k \mathbf{z}_k$ in $\mathbf{b}$ can be viewed as a vector of model variables, while $\mathbf{z}_k{}^T$ in both $\mathbf{b}$ and e_k can be viewed as a (transposed) vector of calibration variables.

Kott et al. (2014) and CBHSQ (2016) showed how WTADJX was used to compute SEs for prevalence estimates. The samples for the years were combined and the w_k and q_k scaled for reasons explained there. The $\mathbf{z}$ vector contained a separate set of components for each year.

A.2 Jackknife Replication and Jackknife Weights

For this section, S is redefined as the NSDUH main adult sample from 2008 to 2012 and the $\mathbf{z}$ vector redefined accordingly. Consider first a prevalence estimate for a 0/1 variable y computed with the precalibration MHSS weights:

$$p = \frac{\sum_S q_k y_k}{\sum_S q_k}, \quad \text{(A4)}$$

(Recall that $q_k = 0$ when k is not in the MHSS clinical sample.)

With H variance strata each containing two PSUs, the usual jackknife variance estimator for p is

[10] See CBHSQ (2014b, Chapter 2) for details on the alternative K6, WHODAS, and SDS scores.

$$v_J'(p) = \tfrac{1}{2}\sum_{h=1}^{H}\sum_{j=1}^{2}\left(p_{(hj)} - p\right)^2 = \tfrac{1}{2}\sum_{h=1}^{H}\sum_{j=1}^{2}\left(\frac{\sum_S q_{k(hj)}y_k}{\sum_S q_{k(hj)}} - \frac{\sum_S q_k y_k}{\sum_S q_k}\right)^2 \qquad \text{(A5)}$$

$$= \tfrac{1}{2}\sum_{h=1}^{H}\left(\frac{\sum_S q_{k(h1)}y_k}{\sum_S q_{k(h1)}} - \frac{\sum_S q_k y_k}{\sum_S q_k}\right)^2 + \left(\frac{\sum_S q_{k(h2)}y_k}{\sum_S q_{k(h2)}} - \frac{\sum_S q_k y_k}{\sum_S q_k}\right)^2,$$

where, according to Rust (1980),

$$q_{k(hj)} = \left\{\begin{array}{ll} q_k & \text{when } k \text{ is not in variance stratum } h \\ 0 & \text{when } k \text{ is in variance stratum } h \text{ and variance PSU } j \\ 2q_k & \text{when } k \text{ is in variance stratum } h \text{ but not variance PSU } j \end{array}\right\}. \qquad \text{(A6)}$$

The $2H$ sets of jackknife replicate weights $\{q_{k(hj)}\}$, one for each variance stratum/variance PSU combination, can be used to estimate the variance of any prevalence estimator and any estimator nearly (i.e., asymptotically) equal to a smooth function of linear estimators of the form $t = \sum_S q_k y_k$.[11] Observe that p in equation (A4) is a smooth function of linear estimators as is an estimated regression coefficient, while a logistic regression coefficient can be asymptotically approximated by a smooth function of linear estimators.

To construct a jackknife variance estimator for an estimator θ that is (approximately) a smooth function of linear estimators, one simply replaces p in equation (A5) with θ, and the $p_{(hj)}$ with $\theta_{(hj)}$ computed like θ but with $q_{k(hj)}$ everywhere replacing q_k.

Krewski and Rao (1981) discussed versions of the jackknife where the p in equation (A5) is replaced by other values, but equation (A5) is most often computed in practice.[12] When θ is a linear function $t = \sum_S q_k y_k$, $v_J'(\theta)$ is identical to both the following jackknife variance estimator,

$$v_J(\theta) = \sum_{h=1}^{H}\left(\theta_{(h1)} - \theta\right)^2, \qquad \text{(A7)}$$

---

[11] If $\mathbf{y}$ is a vector of linear estimates with expected values $\mathbf{Y}$ with $\mathbf{y} - \mathbf{Y} = \mathbf{O}_p\left(1/\sqrt{\mathbf{H}}\right)$, its linear and jackknife variance estimates are identical. Moreover, the variance estimate for a smooth function of $\mathbf{y}$, $g(\mathbf{y}) = g(\mathbf{Y}) + g'(\mathbf{Y})(\mathbf{y} - \mathbf{Y}) + \mathbf{O}_p(1/\mathbf{H})$, where $g'(\mathbf{Y})$ is the row vector of partial derivatives evaluated at $\mathbf{Y}$, can also be computed with the same jackknife weights and estimation formula because it is asymptotically identical to a linear combination of the $\mathbf{y} - \mathbf{Y}$.

[12] Krewski and Rao (1981) also provided formal analyses of applying jackknife and BRR variance estimators to smooth function of linear estimators.

and the linearization variance estimator,

$$v_L(\theta) = \sum_{h=1}^{H} \left(\sum_{k \in S_{h1}} q_k y_k - \sum_{k \in S_{h2}} q_k y_k \right)^2, \tag{A8}$$

where $\theta = \sum_S q_k y_k = \sum_{h=1}^{H} \sum_{j=1}^{2} \sum_{k \in S_{hj}} q_k y_k$, and S_{hj} is the sample within PSU hj.

For a nonlinear estimator, $v_J(\theta)$ and $v_J'(\theta)$ remain relative simple to compute once the jackknife weights are determined, while computing $v_L(\theta)$ becomes more difficult. All three are nearly equal. Because one can compute $v_J(\theta)$ with H sets of replicate weights rather than the $2H$ required by $v_J'(\theta)$, yet the result is nearly the same, the former is preferred.

In the calibration estimator computed with the MHSS clinical sample, p in equation (A4) is replaced by

$$p_C = \frac{\sum_S \omega_k y_k}{\sum_S \omega_k} = \frac{\sum_S q_k f_k y_k}{\sum_S q_k f_k},$$

where f_k is defined by equation (A1), and $\mathbf{g}$ in that equation is chosen to satisfy equation (A2).[13] In traditional replication, jackknife replicate weights are computed using

$$\omega_{k(hj)}^{trad} = q_{k(hj)} f_{k(hj)}, \tag{A9}$$

where

$$f_{k(hj)} = \frac{\exp\left(\frac{U_k}{U_k - 1} a_k \mathbf{z}_k{}^T \mathbf{g}_{(hj)}\right)}{1 + \left[\exp\left(\frac{U_k}{U_k - 1} a_k \mathbf{z}_k{}^T \mathbf{g}_{(hj)}\right) - 1\right] \Big/ U_k},$$

$\mathbf{g}_{(hj)}$ is computed for each set of replicate weights hj such that

$$\sum_S w_{k(hj)} \mathbf{z}_k = \sum_S q_{k(hj)} f_{k(hj)} \mathbf{z}_k, \tag{A10}$$

and the $w_{k(hj)}$ are analogous to the $q_{k(hj)}$ in equation (A6) (recall that S is the entire NSDUH main adult sample). That is to say, the jackknife replicate calibration weights are computed

[13] The linearization estimator for the variance of p_C that accounts for the final calibration $w_k / \Sigma_S w_i$ replaces q_k in equation (A8), and $\mathbf{z}_k{}^T \mathbf{b} + [\omega_k / w_k] e_k$ replaces y_k, where $\mathbf{b}$ and e_k are defined by equation (A3).

exactly how the original calibration weights were computed treating the original NSDUH main adult sample as if variance PSU j in variance stratum h were missing and replaced by a repeat of the other variance PSU in stratum h.

A potential problem with traditional jackknife calibration is that equation (A10) may not be solvable for every hj or every $h1$ because of the restrictions of the replicate adjustment factors $f_{k(hj)}$ due to the U_k (which do not change). Although many software packages (such as SUDAAN and SAS) simply omit replicates when equation (9) cannot be solved, that ad hoc technique is not theoretically justified and cannot be recommended.

A recalibrated jackknife can be designed to avoid this problem. Let

$$\omega_{k(hj)}^{recal} = \omega_{k(hj)}\tilde{f}_{k(hj)}, \tag{A11}$$

where $\tilde{f}_{k(hj)} = \exp\left(a_k f_k[(U_k - f_k)/(U_k - 1)]\mathbf{z}_k{}^T\tilde{\mathbf{g}}_{(hj)}\right)$,[14] and $\tilde{\mathbf{g}}_{(hj)}$ is computed for each set of replicate weights hj such that

$$\sum_S w_{k(hj)}\mathbf{z}_k = \sum_S \omega_{k(hj)}\tilde{f}_{k(hj)}\mathbf{z}_k, \tag{A12}$$

and the $w_{k(hj)}$ and $\omega_{k(hj)}$ are analogous to the $q_{k(hj)}$ in equation (A6). It is unclear, however, whether recalibration is actually needed in this case. Moreover, the recalibrated jackknife does not have extensive empirical verification, which is why it was not used in this report's chapters.

It should also be noted that the naïve jackknife, where $\omega_{k(hj)}^{naïve} = q_{k(hj)}f_k$, also avoids this problem, but fails to capture the impact of calibration on the variance. The naïve linearization estimator is computed using equation (A8) with $\omega_k / \Sigma_S\omega_i$ replacing q_k.

The proof that an estimator that is (approximately) equal to a smooth function of estimators in the form of p_C can have its variance estimated in a nearly unbiased fashion using jackknife replicate weights closely parallels that for a smooth function of linear estimators (see footnote 12).

Even choosing $v_J(\theta)$ in equation (A7) rather than $v_J{}'(\theta)$ forces the computation and storing of 100 sets of jackknife replicate weights when there are 100 MHSS variance strata. It is tempting, and possible, to create 50 variance strata rather than 100 without biasing the resulting variance estimates.[15] Whether H is 50 or 100, the nominal degrees of freedom for a regression

[14] Kott (2006) proposed the recalibrated jackknife. In that version, the calibration adjustment factor $\tilde{f}_{k(hj)}$ was linear, $\tilde{f}_{k(hj)} = 1 + a_k f_k[(U_k - f_k)/(U_k - 1)]\mathbf{z}_k{}^T\tilde{\mathbf{g}}_{(hj)}$, which would allow jackknife replicate weights to be negative. This is avoided here, but a solution to equation (12) is no longer ensured. The proof that recalibration produces a nearly unbiased variance estimator remains the same, but it is too complex to repeat here.

[15] This was, in fact, done when it was thought the clinical samples could produce yearly estimates. See Kott et al. (2014).

estimator with p covariates other than the intercept is $H - p$, a number that is not likely to be uncomfortably small in practice when $H = 50$, but $H = 100$ is safer. Moreover, jackknife SE measures for prevalences will be closer to their already computed linearization analogues when $H = 100$, which is what was used in this report's chapters.

A.3 Fay's BRR and F-BRR Weights

Many of the properties of the variance estimators using Fay's BBR and F-BRR weights parallel the variance estimators using the jackknife weights. The H sets of the F-BRR replicate weights for a two-variance-PSU-per-two-variance stratum design are created using an $H \times H$ Hadamard matrix. For each set of replicate weights in a linear estimator $t = \sum_S q_k y_k$, the matrix dictates the variance PSU in a particular variance stratum whose elements have replicate weights that are $2 - \varepsilon$ times their original weights while the element weights of the other variance PSU are ε times their original values. In F-BRR, ε is a nonnegative value of less than 1. In traditional BRR, $\varepsilon = 0$.

For a prevalence estimator using precalibrated weights, the F-BRR variance estimator is

$$v_{F-BRR}(p) = \frac{1}{H(1-\varepsilon)^2} \sum_{r=1}^{H} \left(\frac{\sum_S q_{k(r)} y_k}{\sum_S q_k} - \frac{\sum_S q_k y_k}{\sum_S q_k} \right)^2, \qquad \text{(A13)}$$

where r denotes a particular F-BRR replicate. To create F-BRR replicated weights for the calibration estimator p_C, one simply replaces $q_{k(r)}$ with $\omega_{k(r)}^{trad}$ or $\omega_{k(r)}^{recal}$ while the hj in equations (A10) through (A12) are replaced by r as needed.

A stratified bootstrap variance estimator for p has the same form as equation (A13) with $\varepsilon = 0$ and H replaced by R, the number of bootstrap replicates. Moreover, the determination of which variance PSU in a replicate has its elements' weights doubled and which are set to 0 is made randomly within each variance stratum. The expectation of this variance estimator is nearly the same as that from using the jackknife or Fay's BRR. Its variance will usually be higher, however. The "balanced" in BRR means that the patterns of replicate weighting have been chosen to minimize the extra noise in variance estimation.[16]

A.4 Choosing a Replicate Method for the MHSS Clinical Sample

Variance estimators for the smooth function of calibration estimators are nearly—that is, asymptotically—the same whether they are computed with jackknife weights or F-BRR weights.

---

[16] For a linear estimator, a variance estimator based on a single Fay replicate is $(1-\varepsilon)^{-2}(\sum_S q_{k(r)} y_{k(r)} - \sum_S q_k y_k)^2 = (1-\varepsilon)^{-2}(\sum^H \sum^2 \sum_{k \in S_{hj}} [q_{k(r)} - q_k] y_{k(r)})^2$, $\sum^H X_h{}^2 + \sum_{h \neq h'}^H X_h X_{h'} = (\sum^H X_h)^2$, where $X_h = (1-\varepsilon)^{-1} \sum^2 \sum_{S_{k \in hj}} [q_{k(r)} - q_k] y_{k(r)}$. It has nearly the same expected value as Fay's BRR variance estimator. Balancing eliminates the added noise from the $X_h X_{h'}$ terms. These terms do not contribute to the asymptotic bias of the variance estimator, but they do contribute to its variance.

What differentiates them is that when estimating the variance of a nonsmooth function of a calibration estimator, such as a median prevalence rate, using jackknife replicate weights will often be biased, while the bias from using F-BRR weights shrinks nearly to zero when $\varepsilon = 0$ (Judkins, 1990). Competing with this is the increased possibility of not being able to solve the F-BRR equivalents of equation (A10) or even (A12) with F-BRR weights. The lower the ε value, the less likely one will find a solution for all sets of replicate weights. In practice, ε is most often set to ½ as a compromise.

An important empirical question with the MHSS clinical sample is whether and when a replicate fails to calibrate. That is, no solution to equation (A10) or (A12) exists for the jackknife or its F-BRR analogue, ruling out the use of that method.

Of lesser importance, but still germane, is the size of the difference in the SE measures of estimated percentiles between a potential replication method and those computed using the WTADJX procedure with 100 variance strata. Even when replication and linearization methods produce the same results asymptotically, these results will still be different given actual finite sample sizes. How different is largely an empirical question. Heuristically, all replication methods that successfully calibrate should overestimate variance slightly because the variability of replicate calibration weights will likely be greater than that of the original calibration weights. Under this reasoning, a jackknife variance estimator should have less bias than an analogous F-BRR, which in turn should be less biased the larger ε is.

This report's chapters were not concerned with quantiles. Consequently, SEs were measured with the traditional jackknife using equation (A7) and 100 variance strata for reasons discussed earlier. The constraint on U_k was removed for three sets of replicate weights. Except for these few easily fixed calibration failures, there was no need to compute the recalibrated jackknife. Moreover, the effectiveness of that jackknife technique has not been demonstrated empirically in the literature.