

Statistical Methods Used in EdSurvey

Developed by Paul Bailey and Michael Cohen^{*†}

November 2, 2018

This document describes estimation procedures for the **EdSurvey** package. It includes the estimation of means (including regression analysis) and percentages; the estimation of correlation coefficients is covered in a vignette in the **wCorr** package.¹

Which estimation procedure is used for any statistic appears in the Help file for the function that creates the statistic. For example, to find the estimation procedure used for the standard error of the regression coefficients, use `?lm.sdf` to see the manual entry.

This document uses many symbols; a table of the symbols is shown here as a reference. Terms used only once are defined immediately above or below equations, so they do not appear in this table.

Symbol	Meaning
A	A random variable
B	Another random variable
i	An index used for observations
j	An index used for jackknife replicates
J	The number of jackknife replicates
m	The number of plausible values
m^*	The number of plausible values used in a calculation
n	The number of units in the sample
p	An index used for plausible values
w_i	The i th unit's full sample weight
x_i	The i th unit's value for some variable
\mathbf{X}	A matrix of predictor variables in a regression
\mathbf{y}	A vector of predicted variables in a regression
β	The regression coefficients in a regression
ϵ	The residual term in a regression
γ	The sampling variance multiplier
\mathcal{A}	The set of sampled units who are in a population of interest (e.g., Black females)
$\tilde{\mathcal{A}}$	The set of population units who are in a population of interest (e.g., Black females)
\mathcal{U}	The set of sampled units who are in a population that contains \mathcal{A} (e.g., Black individuals)
$\tilde{\mathcal{U}}$	The set of population units who are in a population that contains \mathcal{A} (e.g., Black individuals)

The remainder of this document describes estimation procedures that are used in the **EdSurvey** package. The

^{*}This publication was prepared for NCES under Contract No. ED-IES-12-D-0002 with the American Institutes for Research. Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government.

[†]The authors would like to thank Dan Sherman, Qingshu Xie, and Ting Zhang for reviewing this document.

¹See `vignette("wCorrFormulas", package="wCorr")`

next section describes the estimation of means, and the second section describes the estimation of percentages. Each section starts by describing estimation of the statistic, followed by estimation procedures of the variances of the statistic. Separate sections address situations where plausible values are present and situations where plausible values are not present. For sections on variance estimation, separate sections address the jackknife or Taylor series variance estimators.

Estimation of Weighted Means

This section concerns the estimation of means, including regression coefficients and the standard errors of means and regression coefficients.

Estimation of Weighted Means When Plausible Values Are Not Present

Weighted means are estimated according to

$$\mu_x = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i} \quad (1)$$

where x_i and w_i are the outcome and weight of the i th unit (respectively) and n is the total number of units in the sample.

In the case of regression of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (2)$$

a weighted regression is used so that the estimated coefficients ($\boldsymbol{\beta}$) minimize the weighted square residuals

$$\boldsymbol{\beta} = \text{ArgMin}_{\mathbf{b}} \sum_{i=1}^n w_i (y_i - \mathbf{X}_i \mathbf{b})^2 \quad (3)$$

where \mathbf{X}_i is the i th row of \mathbf{X} and $\text{ArgMin}_{\mathbf{b}}$ means the value of \mathbf{b} that minimizes the expression that follows it.

Estimation of Weighted Means When Plausible Values Are Present

When the variable x has plausible values, then these are used to form the mean estimate (μ) according to

$$\mu = \frac{1}{m} \sum_{p=1}^m \frac{\sum_{i=1}^n w_i x_{ip}}{\sum_{i=1}^n w_i} \quad (4)$$

where x_{ip} is the p th plausible value for the i th unit's outcome, and there are m plausible values for each unit.

For regressions, the coefficient estimates are simply averaged over the plausible values

$$\boldsymbol{\beta} = \frac{1}{m} \sum_{p=1}^m \boldsymbol{\beta}_p \quad (5)$$

where $\boldsymbol{\beta}_p$ is the vector of estimated regression coefficients, calculated using the p th set of plausible values.

Estimation of the Coefficient of Determination in a Weighted Linear Regression

In regression analysis, statistics such as the coefficient of determination (or R -squared) are estimated across all observations. For these statistics, their values are averaged across the regression runs (one per set of plausible values). For example,

$$R^2 = \frac{1}{m} \sum_{p=1}^m R_p^2 \quad (6)$$

where R_p^2 is the R -squared value for the regression run with the p th set of plausible values.

For a particular regression, the R -squared is defined in Weisberg (1985, eq. 2.31) as

$$R^2 = 1 - \frac{RSS}{SYY} \quad (7)$$

where $RSS = \mathbf{e}^T \mathbf{W} \mathbf{e}$ (Weisberg, 1985, eq. 4.2), and $SYY = (\mathbf{y} - \bar{y})^T \mathbf{W} (\mathbf{y} - \bar{y})$, \bar{y} is the weighted mean of the outcome, and $\mathbf{e} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$.

Estimation of Standard Deviations

Weighted variance estimates dispersion in the variable in question and are calculated according to

$$\hat{s}^2 = \frac{\sum_{i=1}^n w_i (x_i - \mu_x)^2}{\sum_{i=1}^n w_i} \quad (8)$$

where μ_x is the weighted mean. When there are several plausible values, the variance is averaged across the plausible values (calculating μ_x per plausible value).

The estimate of the standard deviation is the square root of the estimated variance.

Estimation of Standardized Regression Coefficients

Using the definition of the standardized regression coefficients (b)

$$b_j = \frac{\tilde{\sigma}_y}{\tilde{\sigma}_{X_j}} \beta_j \quad (9)$$

where b_j is the standardized regression coefficient associated with the j th regressor, $\tilde{\sigma}_y$ is the weighted standard deviation of the outcome variable, and $\tilde{\sigma}_{X_j}$ is the standard deviation of the j th regressor.

Default Variance Estimation of Standardized Regression Coefficients

The default standard error of the standardized regression coefficients then treats the standard error estimates as constants and is

$$\sigma_{b_j} = \frac{\sigma_y}{\sigma_{X_j}} \sigma_{\beta_j} \quad (10)$$

Sampling Variance Estimation of Standardized Regression Coefficients

An alternative method estimates the standardized regression coefficients using the same process but estimates their standard error accounting for the design-based sampling variance.

In this method, the standardized regression coefficients are estimated per plausible value and jackknife replicate. When the standardized regression coefficient is estimated for a plausible value, the overall variance of the outcome ($\tilde{\sigma}_y$) and the regressors ($\tilde{\sigma}_{X_j}$) is used. For a jackknife replicate, the values of $\tilde{\sigma}_y$ and $\tilde{\sigma}_{X_j}$ are updated with the jackknife replicate weights.

Estimating the variance of the standardized regression coefficient proceeds identically to estimating the variance of the regressors, and the weighted standard deviations are also updated with the jackknife replicate weights.

Estimation of Standard Errors of Weighted Means When Plausible Values Are Not Present, Using the Jackknife Method

When the predicted value does not have plausible values and the requested variance method is jackknife, the variance of the coefficients (V_J) is estimated as

$$V_J = V_{jrr,0} = \gamma \sum_{j=1}^J (\beta_j - \beta_0)^2 \quad (11)$$

where γ is a constant equal to one for the jackknife variance estimation method, its inclusion allows us to extend the equations to other variance estimation methods such as balanced repeated replication (Wolter, 2007) or Fay's method (Judkins, 1990); β_j are the coefficients estimated with the j th jackknife replicate weights, and β_0 are the coefficients estimated with the sample weights; and J is the total number of jackknife replicate weights.

The covariance between β_l and β_m ($C_{J;lm}$) is esimated as

$$C_{J;lm} = C_{jrr,0;lm} = \gamma \sum_{j=1}^J (\beta_{j;l} - \beta_{0;l})(\beta_{j;m} - \beta_{0;m}) \quad (12)$$

where subscripts after the semicolon indicate the matrix element (two subscripts) of the covariance matrix (C) or the vector element (one subscript) for the estimate vector β . The other subscripts are as with the variance estimation.

Estimation of Standard Errors of Weighted Means When Plausible Values Are Present, Using the Jackknife Method

When the predicted value has plausible values and the requested variance method is jackknife, the variance (V_{JP}) is estimated as the sum of a variance component from the plausible values (also called imputation values, so that the variance term is called V_{imp}) and the sampling variance using plausible values ($V_{jrr,P}$) is estimated according to the following formula:

$$V_{JP} = V_{imp} + V_{jrr,P} \quad (13)$$

The sampling variance is

$$V_{jrr,P} = \frac{1}{m^*} \sum_{i=1}^{m^*} V_{jrr,p} \quad (14)$$

Note that in this equation, m^* is a number that can be as small as one or as large as the number of plausible values.² In the previous equation, $V_{jrr,P}$ is the average of $V_{jrr,p}$ over the plausible values, and the values of $V_{jrr,p}$ are calculated in a way analogous to $V_{jrr,0}$ in the previous section, except that the p th plausible values are used within each step:

$$V_{jrr,p} = \gamma \sum_{j=1}^J (\beta_{jp} - \beta_{0p})^2 \quad (15)$$

The imputation variance is estimated according to Rubin (1987):

$$V_{imp} = \frac{m+1}{m(m-1)} \sum_{p=1}^m (\beta_p - \beta)^2 \quad (16)$$

²This option is included because any value for m^* gives an estimate of V_{jrr} with the same properties as larger values of m^* (they are unbiased under the same conditions), but larger values of m^* can take substantially longer to compute. The value of m^* is set with the `jrrIMax` argument; note that `jrrIMax` affects the estimation of V_{jrr} only.

where m is the number of plausible values, β_p is the vector of coefficients calculated with the p th set of plausible values, and β is the estimated coefficient vector averaged over all plausible values.

Covariance terms between β_l and β_m are estimated according to

$$C_{JP;lm} = C_{imp;lm} + C_{jrr,P;lm} \quad (17)$$

where subscripts after a semicolon indicate the indexes of the covariance term being identified,

$$C_{jrr,p;lm} = \gamma \sum_{j=1}^J (\beta_{jp;l} - \beta_{0p;l})(\beta_{jp;m} - \beta_{0p;m}) \quad (18)$$

and

$$C_{imp;lm} = \frac{m+1}{m(m-1)} \sum_{p=1}^m (\beta_{p;l} - \beta_l)(\beta_{p;m} - \beta_m) \quad (19)$$

where β_l and β_m are the estimates averaged across all of the plausible values.

Estimation of Standard Errors of Weighted Means When Plausible Values Are Not Present, Using the Taylor Series Method

When the predicted value does not have plausible values and the requested variance method is the Taylor series, the variance of the coefficients (\mathbf{V}_T) is estimated as³

$$\mathbf{V}_T = \mathbf{V}_{Taylor,0}(\beta) = \mathbf{D}^T \mathbf{Z} \mathbf{D} \quad (20)$$

When the predicted value does not have plausible values and the requested variance method is the Taylor series, the variance of the coefficients will be denoted by \mathbf{V}_T . Notice that \mathbf{V}_T is a matrix, so the variance estimates are along the diagonal while covariances are the off-diagonal elements. Then \mathbf{V}_T is estimated by

$$\mathbf{D} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \quad (21)$$

\mathbf{X} is the matrix of regressors, the \mathbf{W} matrix is a diagonal matrix with the i th weight in the i th diagonal element, and

$$\mathbf{Z} = \sum_{j=1}^J \frac{n_s}{n_s - 1} \sum_{u=1}^{n_s} \mathbf{z}_{uj} \mathbf{z}_{uj}^T \quad (22)$$

where the inner sum is over the sampled Primary Sampling Units (PSUs; u), of which there are n_s , and the outer sum is over all units in the jackknife replicate strata (j), of which there are still J . Note that only strata with at least two PSUs that have students are included in the sum; others are simply excluded.⁴ For a mean, \mathbf{z}_{uj} is a scalar; for a regression, \mathbf{z}_{uj} is a vector with an entry for each regressor. In what follows, when the estimand is a mean, \mathbf{X} simply would be a column vector of ones.

Define the estimated residual vector (\mathbf{e}) as

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\beta \quad (23)$$

and define the term

$$U_{ik} = e_i X_{ik} w_i \quad (24)$$

³This is a slight generalization of Binder (1983, sec. 4.2) to the weighted case. This is derived in more detail and with notation more closely aligned to Binder in the AM manual (Cohen, 2002; see Tools: Procedures: Other Available Procedures: Regression: Details).

⁴This leads to a downward bias in the estimated variance. When the number of units excluded by this rule is proportionally small, the bias also should be proportionally small.

where i indexes the matrix row (observation) and k indexes the matrix column (regressor). Then the k th entry of \mathbf{z}_{uj} is given by

$$z_{ujk} = \sum_{i \in \mathcal{Q}_{uj}}^{n_s} \left[U_{ik} - \left(\frac{1}{n_s} \sum_{i' \in \mathcal{Q}_j}^{n_s} U_{i'k} \right) \right] \quad (25)$$

where \mathcal{Q}_{uj} is the indices for observations in the u th PSU of the j th stratum, and \mathcal{Q}_j is the indices for observations in the j th stratum (across all PSUs). Thus, when two PSUs are selected per stratum, the value of \mathbf{z} will be related by $\mathbf{z}_{1j} = -\mathbf{z}_{2j}$.

Estimation of Standard Errors of Weighted Means When Plausible Values Are Present, Using the Taylor Series Method

When the predicted value has plausible values and the requested variance method is the Taylor series, the variance of the coefficients (\mathbf{V}_{TP}) is estimated as

$$\mathbf{V}_{TP} = \mathbf{V}_{Taylor,P}(\boldsymbol{\beta}) + \mathbf{V}_{imp} \quad (26)$$

where the equation for \mathbf{V}_{imp} (and \mathbf{C}_{imp}) are given in the section on the jackknife variance estimator and where the $\mathbf{V}_{Taylor,P}$ are averaged over the plausible values according to

$$\mathbf{V}_{Taylor,P} = \frac{1}{m} \sum_{p=1}^m \mathbf{V}_{Taylor}(\boldsymbol{\beta}_p) \quad (27)$$

where $\mathbf{V}_{Taylor}(\boldsymbol{\beta}_p)$ is calculated as in the previous section, using the p th plausible values to form \mathbf{e} , so that

$$\mathbf{e} = \mathbf{y}_p - \mathbf{X}\boldsymbol{\beta}_p \quad (28)$$

and the remainder of the calculation of U_{ik} and z_{ujk} are otherwise identical.

Estimation of Standard Errors of Differences of Means

In some cases, two means (μ_1 and μ_2) are calculated for potentially overlapping groups. When this is done, the difference ($\Delta = \mu_1 - \mu_2$) may be of interest. When a covariance term is available, the variance of this difference is then estimated using the formula

$$\sigma_{\Delta}^2 = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12} \quad (29)$$

where σ_{12} is the covariance of μ_1 and μ_2 .

When there is not a covariance term available the approximate equation is used

$$\sigma_{\Delta}^2 = \sigma_1^2 + \sigma_2^2 - 2p\sigma_m^2 \quad (30)$$

where a subscript of m is used to indicate the subset that has more observations and p is the ratio of the number of observations used in calculating both μ_1 and μ_2 divided by the n -size of the larger of the two samples.

When one group is a subset of the other, this is a part-whole comparison. Part-whole comparisons in the context of the National Assessment of Educational Progress (NAEP) applies to comparisons between jurisdictions. Please use with caution when implementing this method to other types of gap comparisons.

Estimation of Weighted Percentages

Percentages are used to estimate the proportion of individuals in a group who have some characteristic; for example, the percentage of Blacks who are female. This often is called a “domain.” In the population, the

universe is the set $\tilde{\mathcal{U}}$; in the example, $\tilde{\mathcal{U}}$ is Blacks who are eligible for sampling. The tilde is used to indicate that this set is in the population.⁵ The sought-after percentage is then the percentage of individuals in the subset $\tilde{\mathcal{A}} \subseteq \tilde{\mathcal{U}}$. In the example, $\tilde{\mathcal{A}}$ is the set of Black females who are eligible for sampling. The percentage for which an estimate is desired is then 100 times the number of individuals in $\tilde{\mathcal{A}}$ divided by the number of individuals in $\tilde{\mathcal{U}}$. Mathematically,

$$\Pi = 100 \times \frac{|\tilde{\mathcal{A}}|}{|\tilde{\mathcal{U}}|} \quad (31)$$

where $|\cdot|$ is the cardinality, or the count of the number of members in a set. Note that in this example, $\tilde{\mathcal{U}}$ was itself a subset of the entire eligible population. In other cases, $\tilde{\mathcal{U}}$ simply could be the population of eligible individuals. Then the value Π would represent the percentage of eligible individuals who were Black females.

The remainder of this section describes statistics meant to estimate Π and the variance of those estimates.

Estimation of Weighted Percentages When Plausible Values Are Not Present

In the sample, units are identified as in \mathcal{A} and \mathcal{U} (where the tilde is dropped to indicate they are the sampled sets) and the estimator is⁶

$$\pi = 100 \times \frac{\sum_{i \in \mathcal{A}} w_i}{\sum_{i \in \mathcal{U}} w_i} \quad (32)$$

where π is the estimated percentage.

Another statistic of interest is the weighted sample size of \mathcal{A} , or an estimate of the number of individuals in the population who are members of $\tilde{\mathcal{A}}$. This is calculated with $\sum_{i \in \mathcal{A}} w_i$.

Estimation of Weighted Percentages When Plausible Values Are Present

If membership in \mathcal{A} or both \mathcal{A} and \mathcal{U} is dependent on a measured score being in a range, then the value of Π is estimated once for each set of plausible values (indexed by p) by

$$\pi = 100 \times \frac{1}{m} \sum_{p=1}^m \frac{\sum_{i \in \mathcal{A}_p} w_i}{\sum_{i \in \mathcal{U}_p} w_i} \quad (33)$$

In the case where membership in \mathcal{U} is not associated with the plausible value, \mathcal{U}_p will be the same for all sets of plausible values. The same applies for \mathcal{A}_p .

Estimation of the Standard Error of Weighted Percentages When Plausible Values Are Not Present, Using the Jackknife Method

When membership in \mathcal{A} and \mathcal{U} are not dependent on plausible values and the requested variance method is jackknife, the variance of the percentage ($V_{\pi,J}$) is estimated as

$$V_{\pi,J} = 100^2 \times V_{jrr,f,0} \quad (34)$$

where the jackknife variance of the fraction is given by

$$V_{jrr,f,0} = \gamma \sum_{j=1}^J \left(\frac{\sum_{i \in \mathcal{A}} w_{ij}}{\sum_{i \in \mathcal{U}} w_{ij}} - \frac{\sum_{i \in \mathcal{A}} w_i}{\sum_{i \in \mathcal{U}} w_i} \right)^2 \quad (35)$$

The subscript j is used to indicate that the weights for the j th jackknife replicates are being used, and weights that do not contain a second subscript are the student full sample weights.

⁵When the tilde is not present, the set is just of individuals in the sample.

⁶The notation $i \in \mathcal{A}$ is a bit of an abuse of notation. Strictly speaking, it is the unit that is in \mathcal{A} and \mathcal{U} , not the indices.

Estimation of the Standard Error of Weighted Percentages When Plausible Values Are Present, Using the Jackknife Method

When membership in \mathcal{A} and \mathcal{U} are dependent on plausible values and the requested variance method is jackknife, the variance of the percentage ($\mathbf{V}_{\pi,JP}$) is estimated as

$$\mathbf{V}_{\pi,TP} = 100^2 \times (\mathbf{V}_{jrr,f,P} + \mathbf{V}_{imp,f}) \quad (36)$$

Here, the only modification to $\mathbf{V}_{jrr,f}$ to make it $\mathbf{V}_{jrr,f,P}$ is that the sets \mathcal{A} and \mathcal{U} must be modified to regard one set of plausible values.

$$\mathbf{V}_{jrr,f,P} = \gamma \frac{1}{m^*} \sum_{p=1}^{m^*} \sum_{j=1}^J \left(\frac{\sum_{i \in \mathcal{A}_p} w_{ij}}{\sum_{i \in \mathcal{U}_p} w_{ij}} - \frac{\sum_{i \in \mathcal{A}_p} w_i}{\sum_{i \in \mathcal{U}_p} w_i} \right)^2 \quad (37)$$

where the subscript j is used to indicate that the weights for the j th jackknife replicates are being used, weights that do not contain a second subscript are the student full sample weights, and the subscript p indicates the plausible values being used. Note that in some situations, the \mathcal{A}_p will be identical to each other across all plausible values, and the \mathcal{U}_p will be identical to each other in a broader set of situations.

The value of $\mathbf{V}_{imp,f}$ is given by

$$\mathbf{V}_{imp,f} = \frac{m+1}{m(m-1)} \sum_{p=1}^m \left(\frac{\sum_{i \in \mathcal{A}_p} w_i}{\sum_{i \in \mathcal{U}_p} w_i} - \frac{1}{m} \sum_{p'=1}^m \frac{\sum_{i \in \mathcal{A}_{p'}} w_i}{\sum_{i \in \mathcal{U}_{p'}} w_i} \right)^2 \quad (38)$$

so that the second sum is simply the average over all plausible values and represents the estimate itself (π), and the expression could be rewritten slightly more compactly as

$$\mathbf{V}_{imp,f} = \frac{m+1}{m(m-1)} \sum_{p=1}^m \left(\frac{\sum_{i \in \mathcal{A}_p} w_i}{\sum_{i \in \mathcal{U}_p} w_i} - \frac{\pi}{100} \right)^2 \quad (39)$$

Estimation of the Standard Error of Weighted Percentages When Plausible Values Are Not Present, Using the Taylor Series Method

When membership in $\tilde{\mathcal{A}}$ and $\tilde{\mathcal{U}}$ are not dependent on plausible values and the requested variance method is the Taylor series, the variance-covariance matrix $\mathbf{V}_{\pi,JP}$ of the coefficients is estimated as

$$\mathbf{V}_{\pi,T} = 100^2 \times \begin{bmatrix} \mathbf{DZD} & -\mathbf{DZD}\mathbf{1} \\ -\mathbf{1}^T \mathbf{DZD} & \mathbf{1}^T \mathbf{DZD}\mathbf{1} \end{bmatrix} \quad (40)$$

where the block matrix has elements $\mathbf{DZD} \in \mathbb{R}^{(c-1) \times (c-1)}$; the c th row and column are then products of \mathbf{DZD} and the vector $\mathbf{1} \in \mathbb{R}^{c-1}$ has a one in every element; the definition of \mathbf{D} is the inverse of a matrix of derivatives of a score vector, taken with respect to $\boldsymbol{\pi}$; and \mathbf{Z} is a variance estimate of the proportions based on the sample survey. This is based on results derived here, following Binder (1983, sec. 3.2).

The score function in question is

$$S(\pi_h) = \left(\sum_{i=1}^n w_i \mathbb{I}(\text{unit } i \text{ is in class } h) \right) - \left(\sum_{i=1}^n \pi_h w_i \right) \quad (41)$$

Setting the score function to zero and solving yields the parameter estimator shown in the section “Estimation of Weighted Percentages When Plausible Values Are Present,” less the factor of 100 that converts a proportion to a percentage.

For the first $c-1$ elements of $\boldsymbol{\pi}$, when this function is solved for π_h , the solution is the estimate of π_h shown earlier:

$$\pi_h = \frac{\sum_{i=1}^n w_i \mathbb{I}(\text{unit } i \text{ is in class } h)}{\sum_{i=1}^n w_i} \quad (42)$$

For π_c , the definition is that

$$\pi_c = 1 - \sum_{k=1}^{c-1} \pi_k \quad (43)$$

and with some algebraic rearrangement, this becomes

$$= \frac{\sum_{i=1}^n w_i \mathbb{I}(\text{unit } i \text{ is in class } c)}{\sum_{i=1}^n w_i} \quad (44)$$

The value of D is then the derivative of $S(\pi)$ with respect to π . Because this derivative must be calculated in total equilibrium (so that all the percentages add up to 100), this is done for the first $c - 1$ items, and the variance of π_c is separately calculated. Taking the derivative of $S(\pi)$ and then inverting it shows that $D \in \mathbb{R}^{(c-1) \times (c-1)}$ is a diagonal matrix with entries $\frac{1}{\sum_{i=1}^n w_i}$.

Then the Z matrix accounts is given by

$$Z = \sum_{s=1}^{N_s} \frac{n_s}{n_s - 1} \sum_{j=1}^{n_s} U_{sk}^T U_{sk} \quad (45)$$

where N_s is the number of strata, n_s is the number of PSUs in a stratum, and U_{sk} is the vector of mean score deviates given by

$$U_{sk} = \sum_{l=1}^{n_{sk}} S_{skl}(\pi) - \frac{1}{n_s} \sum_{j=1}^{n_s} \sum_{l=1}^{n_{sj}} S_{sjl}(\pi) \quad (46)$$

where n_{sk} is the number of observations in PSU k and in stratum s , l is an index for individuals within the stratum and PSU, and the score vector is given by

$$S_{skl}(\pi) = w_{skl} e_{skl} - w_{skl} \pi \quad (47)$$

where e_{skl} is a vector that is 0 in all entries except for a single 1 for the class that the unit is in. For example, if a respondent is a male and the possible levels are ("Female", "Male") then their level of e_{skl} would be $(0, 1)^T$.

This gives the covariance matrix for the first $c - 1$ elements of the π vector. Using the usual formula for variance and covariance, it is easy to see that the variance for the final row and column are as shown at the beginning of this section.⁷

Estimation of Degrees of Freedom

This section and the next two subsections regards how degrees of freedom are estimated when a statistic is calculated entirely with one survey. After that there is a description of how degrees of freedom are pooled when a statistic is calculated across two cohorts or samples.

⁷However, the matrix need not be calculated in this fashion. Instead, the final row and column (the covariance terms associated with the value π_c) need not be dropped. They can be simply included in the formulation of D and S along with every other term.

Two heuristic arguments are offered for this. First, the variance terms are all exchangeable, so the same formula that applies to the first term applies to the final term under reordering. Thus, any term in the covariance matrix can be found by simply permuting the covariance matrix so that the term is not in the c th row or column. Because of that, the method for calculating the upper left portion of the block matrix clearly applies to the c th row and column, which can be calculated directly. Some experiments with NAEP data show that the two methods agree.

The second heuristic argument is that the values of π_h already meet the requirement of summing up to one when the score vector is set equal to zero and solved. This means that the constraint does not need to be imposed a second time.

One method of estimating the degrees of freedom for education survey results is to find the sum of the number of PSUs (schools) and subtract from it the number of strata. For NAEP surveys, this results in an estimate of 62.

However, many estimates do not use variation across all the PSUs, so the degrees of freedom should be expected to be smaller.

Estimation of Degrees of Freedom, Using the Jackknife

When the jackknife estimator is used, the Welch-Satterthwaite (Satterthwaite 1946; Welch, 1947) degrees of freedom estimate (dof_{WS}) is

$$dof_{WS} = \frac{1}{m^*} \sum_{p=1}^{m^*} \frac{\left[\sum_{j=1}^J [(A_{jp} - A_{0p}) - (B_{jp} - B_{0p})]^2 \right]^2}{\sum_{j=1}^J [(A_{jp} - A_{0p}) - (B_{jp} - B_{0p})]^4} \quad (48)$$

where A_{jp} is the estimate of A using the j th jackknife replicate value and the p th plausible value, and A_{0p} is the estimate of A using the full sample weights and the p th plausible value; and the same is true for the B subscripts. In the case of a regression coefficient $(A_{jp} - A_{0p}) - (B_{jp} - B_{0p})$ is replaced by the j th deviate from the full sample weight coefficient $\beta_{jp} - \beta_{0p}$.

The Johnson Rust corrected degrees of freedom (dof_{JR}) is

$$dof_{JR} = \left(3.16 - \frac{2.77}{\sqrt{J}} \right) dof_{WS} \quad (49)$$

Estimation of Degrees of Freedom, Using the Taylor Series

For the Taylor series estimator, the degrees of freedom estimator also uses the Welch-Satterthwaite (WS) degrees of freedom estimate (dof_{WS}). However, because the jackknife replicate estimates are not available, a different equation is used. The WS weights require an estimate of the degrees of freedom per group; when using the jackknife variance estimator, this was estimated using the jackknife replicate weights. For the Taylor series, this is estimated per stratum.

Following Binder (1983) and Cohen (2002), the contribution c_s to the degrees of freedom from stratum s is defined as z_{uj} from the section, “Estimation of Standard Errors of Weighted Means When Plausible Values Are Not Present, Using the Taylor Series Method,”

$$c_s = w_s \frac{n_s}{n_s - 1} \sum_{u=1}^{n_s} z_{uj} z_{uj}^T \quad (50)$$

where u indexes the PSUs in the stratum, of which there are n_s , and w_s is the stratum weight, or the sum, in that stratum, of all the unit’s full sample weights. The stratum weight is not used in the calculation of estimated degrees of freedom for the jackknife but is applied here to approximately account for the size of the stratum in the degrees of freedom calculation. Using the c_s values, the degrees of freedom is

$$dof_{WS} = \frac{(\sum c_s)^2}{\sum c_s^2} \quad (51)$$

With multiple plausible values, the average of dof_{WS} across the plausible values is used.

Estimation of Degrees of Freedom for Statistics Pooled Across Multiple Surveys

When a statistic is estimated from two surveys then a different formula is used. In this case there is a statistic from each survey that is combined into the final estimate; for example a difference in means would use a

mean estimate from each survey. The degrees of freedom for this pooled statistic is

$$dof = \frac{(s_1^2 + s_2^2)^2}{\frac{s_1^4}{dof_1} + \frac{s_2^4}{dof_2}} \quad (52)$$

where s_1^2 and dof_1 are the squared standard error and degrees of freedom, respectively, of the stastic calculated from the first survey, and s_2^2 and dof_2 are the squared standard error and degrees of freedom, respectively, of the statistic calculated from the second survey. These intermediate degrees of freedom would be calculated as described in one of the previous two sections.

Estimation of Weighted Mixed Models

Weighted mixed models are estimated, by plausible value, using the **WeMix** package and the results are combined as described in this section.

The fixed effects are estimated from the **WeMix** results using

$$\beta = \frac{1}{M} \sum_{p=1}^M \beta_p \quad (53)$$

where p indexes the plausible values, or which there are M . The exact same formula is used for the estimate of the variance of the random effects.

Weighted mixed model variance estimation

The standard error of the mixed effects are cluster robust, and so are already aggregated to the highest cluster level in the model. We recomend that the highest level align with the sampling design, for example school in a study that selects schools; then the cluster robust variance estimator includes the sampling design and the sampling variance can be estimated by simply averaging it across the plausible values. This is true for both the fixed effects and the variances of the random effects.

The imputation variance is then calculated using Rubin’s rule. Only the formula for the fixed effects is shown, but the same rule applies to the estimated variance of the random effects.

$$V_{IMP} = \frac{M+1}{M(M-1)} \sum_{p=1}^M (\beta_p - \beta)(\beta_p - \beta)' \quad (54)$$

where V_{IMP} is the imputation covariance matrix. The total covariance is the sum of the sampling covariance and the imputation covariance.

$$V = V_{IMP} + V_{SMP} \quad (55)$$

The variance is the diagonal of the covariance matrix.

References

- Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51(3), 279–292.
- Cohen, J. (2002). *AM manual*. Washington, DC: American Institutes for Research.
- Johnson, E. G., & Rust, K. F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Education Statistics*, 17(2), 175–190.
- Judkins, D. R. (1990). Fay’s method for variance estimation. *Journal of Official Statistics*, 6(5), 223–239.

- National Center for Education Statistics. (2009). *t test for partly overlapping groups*. Retrieved from https://nces.ed.gov/nationsreportcard/tdw/analysis/2004_2005/infer_compare2_overlap.aspx
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley.
- Satterthwaite, F. E. (1946) An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2, 110–114.
- Weisberg, S. (1985). *Applied linear regression*. New York, NY: Wiley.
- Welch, B. L. (1947) The generalization of ‘student’s’ problem when several different population variances are involved. *Biometrika*, 34, 28–35.
- Wolters, K. M. (2007). *Introduction to variance estimation*, 2nd ed. New York, NY: Springer.