Methods Used for Estimating Statistics in EdSurvey 1.0.5

Developed by Paul Bailey & Michael Cohen[*†]

April 26, 2017

This document describes estimation procedures for the `EdSurvey` package. It includes estimation of means (including regression analysis) and percentages; 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 out 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 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
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 ith unit's full sample weight
x_i	The ith unit's value for some variable
$\boldsymbol{X}$	A matrix of predictor variables in a regression
$\boldsymbol{y}$	A vector of predicted variables in a regression
$\boldsymbol{\beta}$	The regression coefficients in a regression
ϵ	The residual term in a regression
$\mathcal{A}$	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 next section describes estimation of means, and the second section describes 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.

*This publication was prepared for NCES under Contract No. ED-IES-12-D-0002 with 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.

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

Estimation of Weighted Means

This section concerns 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

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

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

In the case of regression of the form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

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

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

where $\boldsymbol{X}_i$ is the ith row of $\boldsymbol{X}$ and $\text{ArgMin}_{\boldsymbol{b}}$ means the value of $\boldsymbol{b}$ that minimizes the expession that follows it.

Estimation of weighted means when plausible values are present

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

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

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

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

$$\boldsymbol{\beta} = \frac{1}{m} \sum_{p=1}^{m} \boldsymbol{\beta}_p$$

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

Estimation of the coefficient of determination in a weighted linear regression

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

$$R^2 = \frac{1}{m} \sum_{p=1}^{m} R_p^2$$

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

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

$$R^2 = 1 - \frac{RSS}{SYY}$$

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

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 ($\boldsymbol{V}_J$) is estimated as

$$\boldsymbol{V}_J = \boldsymbol{V}_{jrr,0} = \sum_{j=1}^{J} (\boldsymbol{\beta}_j - \boldsymbol{\beta}_0)^2$$

where $\boldsymbol{\beta}_j$ is the coefficients estimated with the jth jackknife replicate weights and $\boldsymbol{\beta}_0$ is the coefficients estimated with the sample weights, and J is the total number of jackknife replicate weights.

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 ($\boldsymbol{V}_{JP}$) is estimated as the sum of a variance component from the plausible values (also called imputation values, so the variance term is called $\boldsymbol{V}_{imp}$) and the sampling variance using plausible values ($\boldsymbol{V}_{jrr,P}$) according to the following formula:

$$\boldsymbol{V}_{JP} = \boldsymbol{V}_{imp} + \boldsymbol{V}_{jrr,P}$$

The sampling variance is

$$\boldsymbol{V}_{jrr,P} = \frac{1}{m^*} \sum_{i=1}^{m^*} \boldsymbol{V}_{jrr,p} \ .$$

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 above equation, $\boldsymbol{V}_{jrr,P}$ is the average of $\boldsymbol{V}_{jrr,p}$ over the plausible values and the values of $\boldsymbol{V}_{jrr,p}$ are calculated in a way analogous to $\boldsymbol{V}_{jrr,0}$ in the previous section, except that the pth plausible values are used within each step:

$$\boldsymbol{V}_{jrr,p} = \sum_{j=1}^{J} (\boldsymbol{\beta}_{jp} - \boldsymbol{\beta}_{0p})^2$$

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

$$\boldsymbol{V}_{imp} = \frac{m+1}{m} \sum_{p=1}^{m} (\boldsymbol{\beta}_p - \boldsymbol{\beta})^2$$

where m is the number of plausible values, $\boldsymbol{\beta}_p$ is the vector of coefficients calculated with the pth set of plausible values, and $\boldsymbol{\beta}$ is the estimated coefficient vector averaged over all 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 ($\boldsymbol{V}_T$) is estimated as[3]

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

[2] This option is included because any value for m^* gives an estimate of $\boldsymbol{V}_{jrr}$ with the same properties as a 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.

[3] This 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 (2002; see Tools: Procedures: Other Available Procedures: Regression: Details).

where

$$\boldsymbol{D} = (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1}$$

the $\boldsymbol{X}$ is the matix of regressors, $\boldsymbol{W}$ matrix is a diagonal matrix with the ith weight in the ith diagonal element, and

$$\boldsymbol{Z} = \sum_{j=1}^{J} \frac{n_s}{n_s - 1} \sum_{u=1}^{n_s} \boldsymbol{z}_{uj} \boldsymbol{z}_{uj}^T$$

where the inner sum is over the sampled 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 $\boldsymbol{z}_{uj}$ is a scalar; for a regression, $\boldsymbol{z}_{uj}$ is a vector with an entry for each regressor. In what follows, when the estimand is a mean, $\boldsymbol{X}$ simply would be a column vector of ones.

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

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$$

and define the term

$$U_{ik} = e_i X_{ik} w_i$$

where i indexes the matrix row (observation) and k indexes the matrix column (regressor), then the kth entry of $\boldsymbol{z}_{uj}$ is given by

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

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

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 ($\boldsymbol{V}_{TP}$) is estimated as

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

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

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

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

$$\boldsymbol{e} = \boldsymbol{y}_p - \boldsymbol{X}\boldsymbol{\beta}_p$$

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

[4] This 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.

Estimation of weighted percentages

Percentages are used to estimate the proportion of individuals in a group that have some characteristic. For example, the percentage of Blacks who are female. This is often 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,

$$\Pi = 100 \times \frac{|\tilde{\mathcal{A}}|}{|\tilde{\mathcal{U}}|}$$

where $|\cdot|$ is the cardinality, or 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 is[6]

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

where π is the estimated percent.

Another statistic of interested 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

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

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 ($\boldsymbol{V}_{\pi,J}$) is estimated as

$$\boldsymbol{V}_{\pi,J} = 100^2 \times \boldsymbol{V}_{jrr,f,0} \ ,$$

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

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

where the jackknife variance of the fraction is given by

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

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

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 ($\boldsymbol{V}_{\pi,JP}$) is estimated as

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

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

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

where the subscript j is used to indicate that the weights for the jth 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 $V_{imp,f}$ is given by

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

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

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

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 Taylor series, the variance covariance matrix $\boldsymbol{V}_{\pi,JP}$ of the coefficients is estimated as

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

where the block matrix has elements $DZD \in \mathbb{R}^{c-1 \times c-1}$; the cth row and column are then products of DZD and the vector $\mathbf{1} \in \mathbb{R}^{c-1}$ which has a one in every element; the definition of $\boldsymbol{D}$ is the inverse of a matrix of derivatives of a score vector, taken with respect to $\boldsymbol{\pi}$ (shown below); and 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

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

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 above

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

For π_c, the definition is that

$$\pi_c = 1 - \sum_{k=1}^{c-1} \pi_k$$

and with some algebric rearangement this becomes

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

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 of the percentages add up to 100), this is done for the first $c-1$ items and the variance of π_c is seperately calculated. Taking the derivative of $S(\pi)$ and then inverting it shows that $\boldsymbol{D} \in \mathbb{R}^{c-1 \times c-1}$ is a diagonal matrix with entries $\frac{1}{\sum_{i=1}^{n} w_i}$.

Then the $\boldsymbol{Z}$ matrix accounts is given by

$$\boldsymbol{Z} = \sum_{s=1}^{N_s} \frac{n_s}{n_s - 1} \sum_{j=1}^{n_s} \boldsymbol{U}_{sk}^T \boldsymbol{U}_{sk}$$

where N_s is the number of strata, n_s is the number of primary sampling units (PSUs) in a stratum, and $\boldsymbol{U}_{sk}$ is the vector of mean score deviates given by

$$\boldsymbol{U}_{sk} = \sum_{l=1}^{n_{sk}} \boldsymbol{S}_{skl}(\boldsymbol{\pi}) - \frac{1}{n_s} \sum_{j=1}^{n_s} \sum_{l=1}^{n_{sj}} \boldsymbol{S}_{sjl}(\boldsymbol{\pi})$$

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

$$\boldsymbol{S}_{skl}(\boldsymbol{\pi}) = w_{skl} \boldsymbol{e}_{skl} - w_{skl} \boldsymbol{\pi}$$

where $\boldsymbol{e}_{skl}$ 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 $\boldsymbol{e}_{skl}$ would be $(0, 1)^T$.

This gives the covariance, matrix for the first $c-1$ elements of the $\boldsymbol{\pi}$ 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 setion.

However, 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 exchangable, so the same formula that applies to the first term applies to the final term under reodering. Thus any term in the covariance matrix can be found by simply permuting the covariance matrix so that the term is not in the cth row or column. Because of that, the method for calcuating the upper left portion of the block matrix clearly applies to the cth 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.

References

Binder, D. A. (1983). On the Variances of Asymptotically Normal Estimators From Complex Surveys. *International Statistical Review*, *51*(**3**): 279–92.

Cohen, J. (2002). *AM Manual*. Washington, DC: American Institutes for Research.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York, NY: Wiley.

Weisberg, S. (1985). *Applied Linear Regression*. New York, NY: Wiley.