# Estimating the R-indicator, Its Standard Error and Other Related Statistics with SAS and SUDAAN

Michael B. Witt

RTI International, P.O. Box 12194, Research Triangle Park, NC 27709

**Abstract**

The Representativity Indicators (or R-indicators) are statistics designed to supplement the traditional response rate when evaluating the potential for nonresponse bias from a survey. The response rate provides information on the amount of nonresponse in a survey and the R-indicators are designed to measure the similarity between respondents and the original sample or population. In this paper, the estimation of the R-indicator is discussed using a standard Logistic Model and using a Generalized Exponential Model (GEM) to estimate response propensity. The GEM is used to compute weight adjustments, such as nonresponse adjustment factors in SUDAAN Release 10. The linearized variance estimate of the R-indicator is presented as well as several related statistics including the population variance and standard deviation of the response propensity as well as correlation between response propensity and other study variables of interest. The variance estimates will account for both the complex design and the model used to estimate response propensity.

**Keywords:** Response Propensity, R-indicator, Survey Quality, Weight Adjustment

## 1. Introduction

Nonresponse bias associated with an estimate can be expressed as a function of two parameters, the response rate and the difference between the respondents and nonrespondents [see for example, (Bose, 2001) or (Biemer and Lyberg, 2003)]. Mathematically, this can be expressed as:

$$Bias(\bar{y}_r) = \bar{Y} - \bar{Y}_r = \left\{ \rho \bar{Y}_r - (1-\rho)\bar{Y}_{nr} \right\} - \bar{Y}_r = (1-\rho)\left\{ \bar{Y}_r - \bar{Y}_{nr} \right\} \qquad (1.1)$$

Where $\rho$ is the response rate, $\bar{Y}$ is the mean over the entire population, $\bar{Y}_r$ is the mean over respondents and $\bar{Y}_{nr}$ is the mean over nonrespondents. So Equation (1.1) suggests the bias will increase as the response rate decreases and as the "similarity" between the respondents and nonrespondents decreases with respect to some characteristic $y_i$ for each population unit $i$.

Given equation (1.1), to the extent possible survey researchers are interested in monitoring and measuring the two main components of nonresponse bias both during data collection and at its conclusion. Specifically, researchers will look at:

- **Response Rates.** Both weighted and unweighted response rates are typically examined. Survey researchers will often compared these rates between surveys that are methodologically similar and between groups within a survey to identify any notable and statistically significant differences.

- **Difference in an Estimate Between Respondents and Nonrespondents.** This can be more difficult to measure because, by definition, a researcher will not know the value of an estimate for the nonrespondents, i.e. the $\bar{Y}_{nr}$. In some cases, additional studies are done such as a nonresponse follow-up study in order to estimate the $\bar{Y}_{nr}$. In other cases,

researchers might identify variables that are correlated with $y_i$ and are known for both respondents and nonrespondents in order to evaluate the potential difference $\{\overline{Y}_r - \overline{Y}_{nr}\}$.

Equation (1.1) also illustrates another important point when evaluating the potential for nonresponse bias in an estimate. Namely, measuring response rates only is not sufficient to draw conclusions about the potential for nonresponse bias. For example, a larger nonresponse bias can exist in an estimate in the presence of a higher response rate – this will occur when the difference between the $\overline{Y}_r$ and the $\overline{Y}_{nr}$ is large.

In the last few years, some survey researchers have begun looking at a set of statistics to indirectly measure nonresponse bias in estimates produced from a respondent sample prior to making any correction for the nonresponse, i.e. prior to adjusting the respondent data using either imputation or some sort of nonresponse adjustment factor applied to the sample weight. These are called **Representativity Indicators**, or **R-indicators** for short. See for example, (Schouten and Cobben, 2007), (Schuten, Cobben and Bethlehem, 2009); (Skinner, Shlomo, Schouten, Zang, Bethlehem, 2009) and (Cobben, Schouten, 2007). The R-indicators provide a measure of how representative the respondents are compared to the sample or population from which they were drawn.

One of the more common R-indicators is one proposed by Schouten and Cobben (2007) and is defined as:

$$R = 1 - 2S_\rho = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\rho_i - \overline{\rho})^2} \tag{1.2}$$

This R-indicator is a function of the standard deviation of the response propensities ($S_\rho$) associated with each unit of the population *i*. It is fairly straightforward to show that *R* is bounded by 0 and 1, similar to a response rate. Furthermore, note that:

- A smaller value of *R* will result when the standard deviation of the response propensity is large. A higher standard deviation is an indication of a greater degree of variability in response propensity and an indication that a respondent sample is less likely to be "representative" of the population or sample from which it was drawn before any nonresponse adjustment is applied. When a respondent sample is less likely to be representative of the population or sample from which it was drawn one would expect the difference $\left|\overline{Y}_r - \overline{Y}_{nr}\right|$ will likely be greater. Clearly, however, this is not *always* true.

- It follows then that large values of *R* result when the standard deviation of the response propensity is small. A smaller standard deviation means the response propensities tend to be more similar and therefore a sample will likely be more "representative" of a population. And again, intuitively if the respondent sample is more representative of a population than one would expect the difference $\left|\overline{Y}_r - \overline{Y}_{nr}\right|$ to be smaller.

The R-indicator was proposed by Schouten and Cobben as a measure of deviation from what they refer to as **strong representativity**. They indicate an R-indicator value of 1 would indicate strong representativity and a value of 0 indicates the maximum deviation from strong representativity. Here, strong representativity is defined as:

$$\rho_i = \Pr[r_i = 1 \mid s_i = 1] = \rho \quad \text{for all } i \text{ in the population.} \tag{1.3}$$

Where $r_i$ is a 0/1 indicator for response and $s_i$ is a 0/1 indicator for being sampled.

The Schouten and Cobben (2007) paper further discusses how this R-indicator is related to the maximum absolute bias of an estimate. The paper notes that the absolute bias of any estimate is bounded by a function of the R-indicator as follows:

$$|B(\hat{\bar{y}}_{HT})| = |E_{\rho}(\bar{Y}_r) - \bar{Y}| = \left| \frac{\sum_{i=1}^{N} \rho_i y_i}{N\bar{p}} - \frac{N\bar{p}\bar{Y}}{N\bar{p}} \right| = \left| \frac{\sum_{i=1}^{N}(\rho_i - \bar{\rho})(y_i - \bar{Y})}{N\bar{p}} \right| \leq \frac{S_p S_y}{\bar{\rho}} = \frac{(1-R) \cdot S_y}{2\bar{\rho}} \qquad (1.4)$$

The inequality in equation (1.4) is due to an approximation by (Bethlehem, 1988). In Equation (1.4), the $\hat{\bar{y}}_{HT}$ is the Horvitz-Thompson estimator of the mean of some variable $y_i$ computed among respondents only and $S_y$ is the standard deviation of the $y_i$'s.

In practice, the response propensities for each unit $i$ in the population are not known. Therefore, the R-indicator defined in Equation (1.2) is estimated from the sample by the following:

$$\hat{R} = 1 - 2S_{\hat{\rho}} = 1 - 2 \sqrt{\frac{1}{\sum_{i=1}^{n} w_i - 1} \sum_{i=1}^{n} w_i \left(\hat{\rho}_i - \bar{\hat{\rho}}\right)^2} \qquad (1.5)$$

Where $w_i$ is the sample design weight associated with unit $i$. In this discussion we assumed the response propensities $\hat{\rho}_i$ are estimated using some model (e.g. a Logistic Model) as $\hat{\rho}_i = f(\tilde{x}_i, \hat{\beta})$ where $\tilde{x}_i$ is a row vector of explanatory variables and $\hat{\beta}$ is a column vector of the estimated model parameters.

In this paper, the discussion of the R-indicator and several related statistics are addressed. Specifically, this paper will present a discussion of:

- **The Linearized Variance Estimator for $\hat{R}$.** This is discussed in **Section 3**. The variance estimator includes both the contribution of the sample design to the variance as well as the variance associated with estimating the model parameters.

- **Alternate Statistics Similar to $\hat{R}$ and Their Linearized Variance Estimators.** Using the same underlying ideas of the R-indicator, a few additional statistics are proposed that would be more appropriate to examine when computing weight adjustments (e.g. nonresponse weight adjustments). These are discussed in **Section 4**. These include the simple standard deviation of the estimated response propensity and the relative standard deviation.

- **Correlations, Partial Correlations and Semi-Partial Correlations and Their Linearize Variance Estimators.** Another method of examining the relationship between response propensity and a particular item of interest is to look at correlations, partial correlations or semi-partial correlations. These are discussed in **Section 5.**

The discussion presented in this paper is centered on two aspects of the R-indicator and the other related statistics: (1) the computation of these statistics from a sample when one is estimating response propensity directly using a Logistic Model or indirectly using a weight adjustment model and (2) the linearized variance estimator for these statistics. An example comparison of these different statistics, along with their estimated standard errors, is presented in **Section 6.**

All of the statistics discussed in this paper are planned enhancements for the next release of the SUDAAN Software product (SUDAAN, 2008) scheduled for August, 2011. A user

will be able to request these statistics, and their variances, from SUDAAN's LOGISTIC and WTADJUST procedures. In preparation for this enhancement, all statistics discussed in the example presented in **Section 6** were created using SAS Version 9.2 so that this example could be used as one of the test jobs for the next release of SUDAAN. This paper concludes with a brief discussion that provides a few programming suggestions for testing the computation of linearized variates in SAS.

## 2. Estimating the Response Propensity

In practice, there are numerous methods of estimating response propensity to a survey, i.e. deriving the $\hat{\rho}_i = f(\tilde{x}_i, \hat{\beta})$. In this paper, two methodologies are considered:

1. **Standard Logistic Model**. In this case, the response propensity for $i$ is estimated by $\hat{\rho}_i = f(\tilde{x}_i, \hat{\beta}) = \left(1 + e^{-\tilde{x}_i \hat{\beta}}\right)^{-1}$. In a weighted logistic regression analysis, the maximum likelihood estimates of the model parameters, i.e. the $\hat{\beta}$, are determined by solving the solving the weighted score functions:

$$\sum_{i=1}^{n} w_i (r_i - \hat{\rho}_i)\tilde{x}_i' = 0, \quad \text{where } r_i \text{ is a 0/1 response indicator} \tag{2.1}$$

2. **Set Propensity Equal to Inverse of Nonresponse Adjustment.** A second method of estimating response propensity is to create a nonresponse weight adjustment directly using some technique such as a weighting class approach, a modeling approach or some other calibration approach. The estimated response propensity would then be defined as the reciprocal of the weight adjustment.

To illustrate the second method, in this paper a modeling/calibration approach to deriving a nonresponse weight adjustment $\left(i.e. \ \hat{\alpha}_i = \hat{\rho}_i^{-1}\right)$ is considered. Specifically, the Generalized Exponential Model that is used in SUDAAN Release 10 (SUDAAN, 2008) and discussed in (Folsom and Singh, 2001) is considered. This model is based on work originally proposed by (Deville and Sarndal, 1992). With this model, the weight adjustment $\hat{\alpha}_i$ is estimated by:

$$\hat{\alpha}_i = \frac{1}{\hat{\rho}_i} = \frac{l_i(u_i - c_i) + u_i(c_i - l_i)\exp(A_i \tilde{x}_i \hat{\beta})}{(u_i - c_i) + (c_i - l_i)\exp(A_i \tilde{x}_i \hat{\beta})} \tag{2.2}$$

In this model, $l_i, u_i$ and $c_i$ are constants set by the user. Note that that as $\tilde{x}_i \hat{\beta} \to -\infty$ then $\hat{\alpha}_i \to l_i$ and as $\tilde{x}_i \hat{\beta} \to \infty$ then $\hat{\alpha}_i \to u_i$. So $l_i$ and $u_i$ are absolute bounds set by the user on the estimated $\hat{\alpha}_i$. The parameter $c_i$ is a user-specified centering constant and $A_i = \frac{(u_i - l_i)}{(u_i - c_i)(c_i - l_i)}$ is a constant used to facilitate convergence.

With this weight adjustment model, the $\hat{\beta}$ are determined by solving the calibration equations $\sum_{i=1}^{n} w_i (\hat{\alpha}_i r_i - 1)\tilde{x}_i' = \sum_{i=1}^{n} w_i \left(\frac{r_i}{\hat{\rho}_i} - 1\right)\tilde{x}_i' = 0$.

## 3. R-indicator and Its Linearized Variance Estimate

We consider the following estimator of the Schouten and Cobben R-indicator for any particular domain of interest $\Omega$:

$$\hat{R} = 1 - 2S_{\hat{\rho}} = 1 - 2\sqrt{S_{\hat{\rho}}^2} = 1 - 2\sqrt{\frac{1}{\sum\limits_{i=1}^{n} d_i w_i - 1} \sum\limits_{i=1}^{n} d_i w_i \left(\hat{\rho}_i - \bar{\hat{\rho}}\right)^2} \tag{3.1}$$

Equation (3.1) is the same as equation (1.5) except for the addition of the domain indicator $d_i$. The variable $d_i = 1$ if unit $i$ belongs to $\Omega$ and $d_i = 0$ otherwise. This term is inserted into the discussion at this point as a way of clarifying the formulas when a particular domain of interest is being considered.

Since $\hat{R}$ is a sample estimate, it has a standard error which in this case is both a function of the sample design as well as the model used to estimate the response propensities $\rho_i$. The process of deriving an estimated standard error estimate of $\hat{R}$ is discussed in this section.

Computing an estimate of the $Var(\hat{R})$ is relatively straightforward with replication methods such as the Jackknife method, Bootstrapping or the Method of Random Groups. An example that estimated the $Var(\hat{R})$ using bootstrap samples is presented in (Schouten, Cobben and Bethlehem, 2009).

In addition to replication methods, another method for estimating variances of a nonlinear statistic is based on the first-order Taylor series approximation of the deviation of the estimate from its expected value. This technique involves first computing the Taylor series linearization for a statistic. We will refer to these linearized values as $\tau_k$ for each sample unit $k$. These linearized variates $\tau_k$'s are then inserted into a formula that correctly reflects the sample design in order to yield the desired variance estimate.

For example, consider one of the most common designs – a design that involves sampling with replacement at the first stage within some strata and with or without replacement sampling at subsequent stages. If $h$ represents strata, $i$ represents the first stage unit and $k$ the unit selected within each first stage unit then an estimate of the variance of a total is:

$$\text{var}(\hat{T}) = \text{var}\left(\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{k=1}^{n_{hi}} \tau_{hik}\right) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1}\sum_{i=1}^{n_h}\left(\tau_{hi} - \bar{\tau}_h\right)^2, \text{ where } \tau_{hi} = \sum_{k=1}^{n_{hi}} \tau_{hik} \tag{3.2}$$

For simplicity, the reference to stratum and PSU is dropped from the discussion and we consider the linearized variate $\tau_k$. Binder (1996) presents an easy-to-implement, general discussion of how to compute the linearized variates $\tau_k$ when the original statistic can be written as a function of weighted totals. In a similar manner, Shah (2002) discusses how the unweighted linearized variate for a statistic associated with unit $k$ can be found by differentiating the statistic with respect to its sample weight. The final weighted linearized variate $\tau_k$ that is used in the sample design formula such as Equation (3.2) is then the sample weight times this derivative. Shah's shortcut method is used to derive the linearized variates associated with $\hat{R}$ in this paper.

Suppose we want to compute the linearized variate $\tau_k$ for record $k$ associated with $\hat{R}$. To obtain the proper derivative of $\hat{R}$ with respect to $w_k$, and to properly account for the model used to estimate the response propensities, we'll need to compute $\dfrac{\partial \rho_i}{\partial w_k}$ for each unit $i$ in the sample. This is obtained by implicitly differentiating the score functions or calibration equations noted in **Section 2** depending on whether one is estimating response

propensity using a standard Logistic Model for using the Generalized Exponential Model. For illustrations purposes, consider the score functions of the Logistic Model [equation (2.1)]. Implicitly differentiating this with respect to $w_k$ yields:

$$\left(r_k - \hat{\rho}_k\right)\tilde{x}_k' - \sum_{i=1}^{n} w_i \hat{\rho}_i (1 - \hat{\rho}_i)\tilde{x}_i'\tilde{x}_i \frac{\partial \beta}{\delta w_k} = 0 \tag{3.3}$$

Which after a little re-arranging, yields:

$$\frac{\partial \beta}{\delta w_k} = J_{logistic}^{-1}\tilde{x}_k'(y_k - \hat{\rho}_k), \text{ Where } J_{logistic} = \sum_{i=1}^{n} w_i \tilde{x}_i'\tilde{x}_i \hat{\rho}_i (1 - \hat{\rho}_i) \tag{3.4}$$

Next, note that since we are considering a Logistic Model then $\hat{\rho}_i = \left(1 + e^{-\tilde{x}_i\hat{\beta}}\right)^{-1}$ so

$$\frac{\partial \hat{\rho}_i}{\partial w_k} = \frac{\partial}{\partial \beta}\left\{\left(1 + e^{-\tilde{x}_i\hat{\beta}}\right)^{-1}\right\} \cdot \frac{\partial \beta}{\partial w_k} = \hat{\rho}_i(1 - \hat{\rho}_i)\tilde{x}_i J_{logistic}^{-1}\tilde{x}_k'(r_k - \hat{\rho}_k) \tag{3.5}$$

Therefore, the weighted linearized variate associated with $\hat{R}$ for record $k$, assuming a Logistic Model is used to estimate the $\hat{\rho}_i$'s, is:

$$\tau_k = w_k \cdot \frac{-1}{S_{\hat{\rho}}} \cdot \left\{\frac{\partial(S_{\hat{\rho}}^2)}{\partial w_k}\right\} \quad \text{where } S_{\hat{\rho}}^2 = \frac{1}{\sum_{i=1}^{n} d_i w_i - 1}\sum_{i=1}^{n} d_i w_i\left(\hat{\rho}_i - \bar{\hat{\rho}}\right)^2 \tag{3.6}$$

or $\qquad \tau_k = w_k \cdot \frac{-1}{S_{\hat{\rho}}} \cdot \frac{1}{\sum_{i=1}^{n} d_i w_i - 1} \cdot \left\{d_k\left[\left(\hat{\rho}_k - \bar{\hat{\rho}}\right)^2 - S_{\hat{\rho}}^2\right] + 2\sum_{i=1}^{n} d_i w_i\left(\hat{\rho}_i - \bar{\hat{\rho}}\right)\frac{\partial \hat{\rho}_i}{\partial w_k}\right\} \tag{3.7}$

In comparison, if we assume the $\hat{\rho}_i$'s are known without error, then $\frac{\partial \hat{\rho}_i}{\partial w_k} = 0$ and the linearized variate associated with $\hat{R}$ becomes:

$$\tau_k = w_k \cdot \frac{-1}{S_{\hat{\rho}}} \cdot \frac{1}{\sum_{i=1}^{n} d_i w_i - 1} \cdot \left\{d_k\left[\left(\hat{\rho}_k - \bar{\hat{\rho}}\right)^2 - S_{\hat{\rho}}^2\right]\right\} \tag{3.8}$$

So the second term in equation (3.7) provides the estimated contribution to the variance associated with estimating the $\hat{\rho}_i$'s from a Logistic Model.

If the Generalized Exponential Model is used to estimate $\hat{\rho}_i$'s the mathematics follows in a similar fashion. The linearized variate associated with $\hat{R}$ for record $k$, assuming a Generalized Exponential Model is used to estimate $\alpha_i = \hat{\rho}_i^{-1}$, is:

$$\tau_k = w_k \cdot \frac{-1}{S} \cdot \frac{1}{\sum_{i=1}^{n} d_i w_i - 1}\left\{d_k\left[\left(\hat{\rho}_k - \bar{\hat{\rho}}\right)^2 - S^2\right] + 2\sum_{i=1}^{n} d_i w_i\left(\hat{\rho}_i - \bar{\hat{\rho}}\right)\frac{\partial \hat{\rho}_i}{\partial w_k}\right\} \tag{3.9}$$

Where $\frac{\partial \hat{\rho}_i}{\partial w_k} = \phi_i \hat{\rho}_i^2 \tilde{x}_i J_{GEM}^{-1}\tilde{x}_k'(\alpha_k r_k - 1)$, $J_{GEM} = \sum_{i=1}^{n} w_i \phi_i r_i \tilde{x}_i'\tilde{x}_i$ and $\phi_i = \frac{(\alpha_i - l_i)(u_i - \alpha_i)}{(c_i - l_i)(u_i - c_i)}$

## 4. Standard Deviation and Relative Standard Deviation of the Response Propensity, and Their Linearized Variance Estimate

In practice, response propensities are not always estimated directly. In many instances, survey researchers will account for nonresponse to a study by computing a nonresponse adjustment factor to the sample weights using some technique such as a Weighting Class

method, a Model-based method or some other Calibration method. If a nonresponse adjustment procedure is used in a study that, by definition, yields adjustment factors that are greater than or equal to one, defining an estimated response propensity as the inverse of the nonresponse adjustment factor and using the ideas presented in **Section 3** are a reasonable way to estimate the R-indicator. However, if the adjustment methodology yields factors that are less than one then defining response propensity as the inverse of the adjustment factor will yield estimated response propensities that are greater than one – a situation that is not particularly appealing when estimating the R-indicator. This would be similar to estimating a response rate using a response indicator $r_i$ that is in fact greater than one for some members of the sample.

**Exhibit 1. Alternate Statistics to the R-indicator and Their Linearized Variate**

| Statistic | Estimate | Weighted Linearized Variate |
|---|---|---|
| 1. Response Rate | $\bar{r} = \sum_{k \in S} \dfrac{w_k d_k r_k}{\hat{N}}$ | $\tau_k^{(1)} = \dfrac{w_k}{\hat{N}} \left\{ d_k (r_k - \bar{r}) \right\}$, where $\hat{N} = \sum_{k \in S} d_k w_k$ |
| **Statistics Based on Response Propensities** | | |
| 2. Mean | $\bar{\bar{\rho}} = \sum_{k \in S} \dfrac{w_k d_k \hat{\rho}_k}{\hat{N}}$ | $\tau_k^{(2)} = \dfrac{w_k}{\hat{N}} \left\{ d_k (\hat{\rho}_k - \bar{\bar{\rho}}) + \sum_{i \in S} w_i d_i \dfrac{\partial \hat{\rho}_i}{\partial w_k} \right\}$ |
| 3. Population Variance | $\hat{S}_\rho^2 = \dfrac{1}{\hat{N}-1} \sum_{k \in S} w_k d_k \left( \hat{\rho}_k - \bar{\bar{\rho}} \right)^2$ | $\tau_k^{(3)} = \dfrac{w_k}{\hat{N}-1} \left\{ d_k \left[ \left( \hat{\rho}_k - \bar{\bar{\rho}} \right)^2 - \hat{S}_\rho^2 \right] + 2 \sum_{i \in S} w_i d_i \left( \hat{\rho}_i - \bar{\bar{\rho}} \right) \dfrac{\partial \hat{\rho}_i}{\partial w_k} \right\}$ |
| 4. Population Standard Deviation | $\hat{S}_\rho = \sqrt{\hat{S}_\rho^2}$ | $\tau_k^{(4)} = \dfrac{1}{2\hat{S}_\rho} \tau_k^{(3)}$ |
| 5. Relative Standard Deviation | $\hat{\theta}_\rho = \dfrac{\hat{S}_\rho}{\bar{\bar{\rho}}}$ | $\tau_k^{(5)} = \dfrac{1}{\bar{\bar{\rho}}} \left[ \tau_k^{(4)} - \hat{\theta}_\rho \cdot \tau_k^{(2)} \right]$ |
| 6. R-indicator | $\hat{R} = 1 - 2\hat{S}_\rho$ | $\tau_k^{(6)} = \dfrac{-1}{\hat{S}_\rho} \tau_k^{(3)}$ |
| **Statistics Based on Weight Adjustments (Inverse of Response Propensities)** | | |
| 7. Mean | $\bar{\bar{\alpha}} = \sum_{k \in S} \dfrac{w_k d_k \hat{\alpha}_k}{\hat{N}}$ | $\tau_k^{(7)} = \dfrac{w_k}{\hat{N}} \left\{ d_k (\hat{\alpha}_k - \bar{\bar{\alpha}}) - \sum_{i \in S} w_i d_i \alpha_i^2 \dfrac{\partial \hat{\rho}_i}{\partial w_k} \right\}$ |
| 8. Population Variance | $\hat{S}_\alpha^2 = \dfrac{1}{\hat{N}-1} \sum_{k \in S} w_k d_k \left( \hat{\alpha}_k - \bar{\bar{\alpha}} \right)^2$ | $\tau_k^{(8)} = \dfrac{w_k}{\hat{N}-1} \left\{ d_k \left[ \left( \hat{\alpha}_k - \bar{\bar{\alpha}} \right)^2 - \hat{S}_\alpha^2 \right] - 2 \sum_{i \in S} w_i d_i \left( \alpha_i - \bar{\bar{\alpha}} \right) \alpha_i^2 \dfrac{\partial \hat{\rho}_i}{\partial w_k} \right\}$ |
| 9. Population Standard Deviation | $\hat{S}_\alpha = \sqrt{\hat{S}_\alpha^2}$ | $\tau_k^{(9)} = \dfrac{1}{2\hat{S}_\alpha} \tau_k^{(8)}$ |
| 10. Relative Standard Deviation | $\hat{\theta}_\alpha = \dfrac{\hat{S}_\alpha}{\bar{\bar{\alpha}}}$ | $\tau_k^{(10)} = \dfrac{1}{\bar{\bar{\alpha}}} \left[ \tau_k^{(9)} - \hat{\theta}_\alpha \cdot \tau_k^{(7)} \right]$ |

The obvious alternative to the R-indicator that will still provide a measure of representativeness of a respondent sample is to examine the standard deviation of weight adjustments and the inverse of the weight adjustments (see **Exhibit 1.**) Unlike the R-indicator, the standard deviation of a weight adjustment (or the inverse of the weight adjustment) is not bounded above by 1, so a comparison of a standard deviation between surveys can be difficult to interpret. To account for this we also propose looking at the relative standard deviation of the weight adjustment (or inverse weight adjustment) as a way of normalizing the standard deviation. This will allow one to compare representativity between studies with a little more confidence, although even the relative standard deviation is not bounded above by any particular number.

**Exhibit 1** displays the statistics and their linearized variates for several statistics related to response propensity and a nonresponse weight adjustment. These linearized variates were computed in a manner that was analogous to that described in **Section 3**.

## 5. Pearson Correlation, Partial Correlation and Semi-Partial Correlation and Their Linearized Variance Estimate

As noted in **Section 1**, the R-indicator indirectly measures the potential bias of an estimate due to nonresponse by providing a measure of representativity of the respondent sample. The R-indicator clearly has many advantages. For instance, similar to a response rate:

- The R-indicator is bounded by 0 and 1.

- One can interpret larger values of the R-indicator to be "better" in some sense than smaller values. Larger values indicate a smaller standard deviation in the response propensities and a greater likelihood that the respondent sample "represents" the population or sample from which it originated from.

- And the computation of the R-indicator (and its estimate, $\hat{R}$) is independent of any particular data item collected in the study. Therefore, similar to the response rate, this statistic can be interpreted as a quality measure for the entire survey, i.e. it does not apply only to a subset of the items collected from respondents.

However, one of the primary disadvantages of the R-indicator is that it provides information on the potential for nonresponse bias in an estimate computed from the respondent sample, *prior to making any adjustments to the respondent sample for the nonresponse*. In practice, survey statisticians typically apply some correction to the respondents – such as a nonresponse weight adjustment or imputation – to account for at least some of the potential nonignorable nonresponse mechanism underlying the response process. With this in mind, it would also be advantageous to examine the potential for nonresponse bias after the respondent sample has been "corrected" to account for the nonresponse.

Much like the R-indicator, another set of measures that can provide some information on the potential relationship between response propensity and another variable of interest is the **Pearson correlation** and other statistics related to this including the **partial correlation** and **semi-partial correlation**. In general, the partial correlation is designed to measure the strength of a relationship between two variables after correcting each variable for a set of predictors. The set of predictors used to correct each variable is often the same, but this is not necessary. The partial correlation is computed using the residuals from two models that measure the difference between the variables and their model-based expected values. The semi-partial correlation is somewhat similar to the partial correlation. The semi-partial correlation is designed to measure the strength of a relationship between a variable after correcting it for a set of predictors, and some second variable. The semi-partial correlation is defined as the correlation between the residuals from a one variable and a second variable.

Since both the partial correlation and the semi-partial correlation are measuring the strength of an association <u>after</u> correcting at least one of them for a set of predictors, they can be used to provide information on the potential for nonresponse bias after weights have been corrected for the nonresponse. This can be done by computing the correlation between residuals formed using the response propensity and some second variable (semi-partial correlation) or by computing the correlation between the residuals using response

propensity and the residuals obtained after modeling the second variable (partial correlation). For example, if a nonresponse adjustment factor is defined as the inverse response propensity estimated from a Logistic Model with age, race/Ethnicity and gender as explanatory variables then a partial correlation or a semi-partial correlation will provide information on the strength of the relationship between the propensity residuals and some other variable, after the response propensity has been adjusted for the demographic variables.

Some notable features of the correlation statistics include:

- Like the R-indicator, the correlation is bounded above and below (in this case by 1 and -1.)  This allows one to easily compare correlations between data items and between surveys.

- Larger absolute values of the correlation indicate a greater strength of association between the variables.

- As noted above, a partial correlation and the semi-partial correlation offers the advantage of providing information on the relationship between response propensity and some second variable of interest after the response propensity is corrected for some set of variables. This enables one to measure the potential for nonresponse bias after weights are corrected for nonresponse since the partial correlation is designed to measure the strength of association between items <u>after</u> correcting the propensity for a set of variables.

On the other hand, two disadvantages of the correlation measures are:

- By definition, it will vary for different survey items so these would not be a very good measure of "total" survey quality.

- Any correlation involving response propensity is typically computed using both the survey respondents and nonrespondents. This means the second variable should be known for nonrespondents. So typically the correlations are restricted to data items available on the sample frame or obtained in earlier stages of a multistage sample design.

The second limitation can be addressed by fitting the response propensity model using both respondents and nonrespondents, and subsequently treating the respondents as a domain of interest when computing the correlations. Treating the respondents as a domain of interest means the second data item only needs to be known for respondents. In practice, computing a partial correlation for a subdomain defined by respondents does not seem to work well, however computing a standard correlation by restricting the domain of interest to the respondents does. An example that illustrates this is discussed in **Section 6**.

Mathematically, a sample estimate of the correlation between some function of the model-predicted response propensity $F_i(\hat{\rho}_i)$ and a function of a second variable $G_i = G(\hat{\psi}_i)$ {here $\hat{\psi}_i = g(\tilde{y}_i, \hat{\eta})$, $\tilde{y}_i$ is a vector of explanatory variables and $\hat{\eta}$ are estimated model parameters} is defined by $CORR_{\rho\psi} = \dfrac{S_{FG}^2}{\sqrt{S_{FF}^2 S_{GG}^2}}$, where:

$$S_{FG}^2 = \sum_i d_i w_i (F_i - \overline{F})(G_i - \overline{G}) = \left(\sum_i d_i w_i F_i G_i\right) - \left(\hat{N} \cdot \overline{F} \cdot \overline{G}\right) \qquad (5.1)$$

$$S_{FF}^2 = \sum_i d_i w_i (F_i - \overline{F})^2 = \left(\sum_i d_i w_i F_i^2\right) - \left(\hat{N} \cdot \overline{F}^2\right),$$

$$S_{GG}^2 = \sum_i d_i w_i (G_i - \overline{G})^2 = \left(\sum_i d_i w_i G_i^2\right) - \left(\hat{N} \cdot \overline{G}^2\right)$$

Since $CORR_{\rho\psi}$ is a sample estimate, it has a standard error that depends on both the sample design and the model used to estimate $\rho_i$ (and if applicable, the model used to estimate $\psi_i$.) Using an approach that is similar to what was outlined in **Section 3**, the linearized variate corresponding to the correlation for record $k$ is:

$$w_k \cdot CORR_{\rho\psi} \left\{ \frac{\partial S_{FGk}^2}{S_{FG}^2} - \frac{1}{2}\left( \frac{\partial S_{FFk}^2}{S_{FF}^2} + \frac{\partial S_{GGk}^2}{S_{GG}^2} \right) \right\} \tag{5.2}$$

Where $\partial S_{FGk}^2 = \dfrac{\partial S_{FG}^2}{\partial w_k} = d_k\left(F_k - \overline{F}\right)\left(G_k - \overline{G}\right) + \sum_{i \in S} d_i w_i \left(F_i - \overline{F}\right)\dfrac{\partial G_i}{\partial w_k} + d_i w_i \left(G_i - \overline{G}\right)\dfrac{\partial F_i}{\partial w_k}$ (5.3)

$$\partial S_{FFk}^2 = \frac{\partial S_{FF}^2}{\partial w_k} = d_k\left(F_k - \overline{F}\right)^2 + 2\sum_{i \in S} d_i w_i \left(F_i - \overline{F}\right)\frac{\partial F_i}{\partial w_k}$$

$$\partial S_{GGk}^2 = \frac{\partial S_{GG}^2}{\partial w_k} = d_k\left(G_k - \overline{G}\right)^2 + 2\sum_{i \in S} d_i w_i \left(G_i - \overline{G}\right)\frac{\partial G_i}{\partial w_k}$$

For example, $F_i, G_i, \dfrac{\partial F_i}{\partial w_k}$ and $\dfrac{\partial G_i}{\partial w_k}$ are defined as follows if one were interested in estimating the partial correlation between response propensity and another variable $\psi_i$:

$$F_i = r_i - \hat{\rho}_i, \quad \frac{\partial F_i}{\partial w_k} = -1 \cdot \frac{\partial \hat{\rho}_i}{\partial w_k}, \quad G_i = \psi_i - \hat{\psi}_i \text{ and } \frac{\partial G_i}{\partial w_k} = -1 \cdot \frac{\partial \hat{\psi}_i}{\partial w_k}$$

As a second example, if one were interested in estimating the correlation between a weight adjustment $\hat{\alpha}_i$ and another variable $\psi_i$ then

$$F_i = \frac{1}{\hat{\rho}_i} = \hat{\alpha}_i, \quad \frac{\partial F_i}{\partial w_k} = -1 \cdot \hat{\alpha}_i^2 \frac{\partial \hat{\rho}_i}{\partial w_k}, \quad G_i = \psi_i \text{ and } \frac{\partial G_i}{\partial w_k} = 0$$

Other similar correlations, semi-partial correlations and partial correlations follow in a similar fashion.

## 6. Example

As noted in **Section 1**, the purpose of this research was primarily to derive the R-indicator and several related statistics when response propensity is estimated using a Logistic or a Generalized Exponential Model. In addition, we computed the linearized variates for these statistics. These statistics and their standard error estimates will be forthcoming in a future release of the SUDAAN statistical software and this research was designed, in part, to test the computation of the new statistics and verify their standard error estimates. A few comments on how the standard error estimates were validated are presented in **Section 7**. In this section we present an example that was used in the testing process.

To illustrate these statistics, public use file data was used from the 2008 National Survey on Drug Use and Health (NSDUH). This survey is conducted by the Office of Applied Studies within the United States Substance Abuse and Mental Health Services Administration (SAMHSA). The primary purpose of NSDUH is to measure the

prevalence and correlates of drug use in the United States. This is a household, multistage study with data collected quarterly. The target population for this study is the civilian, noninstitutionalized population of the United States who were 12 years of age or older at the time of the survey. For additional information on this study, see (Office of Applied Studies, SAMHSA).

In this example, it was assumed that individuals who indicated they were a past year cigarette user was a survey "nonrespondent". Since cigarette use is correlated with numerous outcome measures collected in this study, assuming the cigarette users are nonrespondents resulted in a nonignorable response mechanism for most items being considered.

**Exhibit 2** presents the estimates of R-indicators assuming the response propensity was estimated using a Logistic Model and a General Exponential Model with the main effects of age group, race/Ethicity, gender, health status, metropolitan size, education and marital indicator used as predictors. The R-indicator was discussed in **Sections 2** and **3**. Also presented in this exhibit are the standard errors of the R-indicators assuming the model predicted response propensities are known without error [see equation (3.8)] and assuming the propensities are estimated from a model [equation (3.7) and (3.9)].

**Exhibit 2. R-Indicators and their Estimated Standard Error**

| Domain | Logistic Model | | | Generalized Exponential Model | | |
|---|---|---|---|---|---|---|
| | R-Indicator | Std Error Ignoring Model | Std Error Full | R-Indicator | Std Error Ignoring Model | Std Error Full |
| **Total** | .7258 | .0012 | .0064 | .7195 | .0012 | .0064 |
| **Age Group** | | | | | | |
| 12-17 | .9197 | .0006 | .0042 | .9141 | .0007 | .0041 |
| 18-25 | .7779 | .0015 | .0075 | .7692 | .0016 | .0078 |
| 26-34 | .7471 | .0027 | .0092 | .7406 | .0028 | .0092 |
| 35+ | .8010 | .0020 | .0078 | .7970 | .0021 | .0078 |
| **Race/Ethnicity** | | | | | | |
| Hispanic | .7391 | .0027 | .0065 | .7379 | .0028 | .0060 |
| NonHisp, Black | .7283 | .0026 | .0086 | .7238 | .0028 | .0086 |
| Other | .7250 | .0012 | .0066 | .7185 | .0013 | .0065 |
| **Gender** | | | | | | |
| Male | .7095 | .0016 | .0069 | .7026 | .0017 | .0069 |
| Female | .7556 | .0014 | .0067 | .7504 | .0015 | .0066 |

**Note:** Std Error Ignoring Model=Standard errors were computed assuming response propensities were known without error. Std Error Full = Standard error account for the model estimation of the response propensities.

In general the R-indicators in **Exhibit 2** are in the .7 to .9 range, with the largest R-indicators being associated with the 12-17 year-old age group. The larger R-indicator means the variability in the predicted response propensities among 12-17 year-olds is less, and therefore the nonresponse bias associated with estimates for 12-17 year-olds is likely to be less (assuming the propensity model is true.) The standard error estimates associated with the R-indicators are fairly equal between the Logistic Model and Generalized Exponential Model predictions. But note that if the term that accounts for the model prediction of the response propensity is included in the standard error estimate, the standard errors are considerably bigger than assuming the propensities are known without error. Over all, we found including the term in the variance estimate that accounts for the model-based prediction of response propensity will increase the standard error estimate by an average of 333% in this table. This large increase suggests omitting this term would likely significantly bias the standard error estimate associated with the R-indicator.

**Exhibit 3** displays the mean, standard deviation and relative standard deviation of both the response propensity and the weight adjustment. This was discussed in **Section 4**. In this exhibit, response propensities were estimated using a Logistic Model and the weight

adjustments were estimated using the Generalized Exponential Model. Similar to **Exhibit 2**, this exhibit shows that including the term that accounts for the model prediction of the response propensity (or weight adjustment) will increase the standard error estimate substantially which again demonstrates omitting this term can lead to standard error estimates that are significantly biased. On average, in this exhibit we found including the second term in the standard error estimate increases the standard error by over 320%.

**Exhibit 3. Additional Statistics and their Estimated Standard Error Using Linearization**

| Domain | Mean | | | Std Deviation | | | Relative Std Deviation | | |
|---|---|---|---|---|---|---|---|---|---|
| | Esti-mate | Std Error Ignoring Model | Std Error Full | Esti-mate | Std Error Ignoring Model | Std Error Full | Esti-mate | Std Error Ignoring Model | Std Error Full |
| **Response Propensities** $\rho_t$ | | | | | | | | | |
| **Total** | .7207 | .0012 | .0042 | .1403 | .0006 | .0032 | .1946 | .0011 | .0047 |
| **Age Group** | | | | | | | | | |
| 12-17 | .8512 | .0006 | .0033 | .0429 | .0003 | .0020 | .0504 | .0004 | .0024 |
| 18-25 | .5506 | .0014 | .0057 | .1154 | .0008 | .0039 | .2096 | .0014 | .0073 |
| 26-34 | .5966 | .0026 | .0096 | .1297 | .0014 | .0046 | .2174 | .0027 | .0097 |
| 35+ | .7640 | .0013 | .0057 | .1015 | .0011 | .0039 | .1329 | .0015 | .0055 |
| **Weight Adjustments** $\alpha_t$ | | | | | | | | | |
| **Total** | 1.4586 | .0029 | .0104 | .3831 | .0027 | .0161 | .2626 | .0016 | .0099 |
| **Age Group** | | | | | | | | | |
| 12-17 | 1.1779 | .0008 | .0046 | .0613 | .0005 | .0032 | .0521 | .0004 | .0027 |
| 18-25 | 1.9068 | .0050 | .0224 | .4512 | .0038 | .0245 | .2366 | .0018 | .0113 |
| 26-34 | 1.7757 | .0090 | .0359 | .4816 | .0087 | .0346 | .2712 | .0041 | .0151 |
| 35+ | 1.3366 | .0028 | .0114 | .2132 | .0036 | .0131 | .1595 | .0025 | .0090 |

**Note:** Response propensities were estimated using a Logistic Model. Weight adjustments were estimated using a Generalized Exponential Model. Std Error Ignoring Model=Standard errors were computed assuming response propensities were known without error. Std Error Full = Standard error account for the model estimation of the response propensities.

**Exhibit 4** displays some results from estimating the correlation, semi-partial correlation and partial correlation between response propensity and two outcome measures (past year alcohol use and any illicit drug use except marijuana.) The top half of the table shows statistics computed using the entire sample and the bottom half shows statistics computed for respondents-only. The correlations computed using the respondents-only treat the respondents as a domain $d_i \in \Omega$ of interest when applying the formulas from **Section 5**. Also, the correlations computed using the respondents-only reflect what one would likely want to do in a "typical" study – so the response propensity model was fit using the entire sample, a second model was fit to the second variable (e.g. past year alcohol use) using only the respondents and correlations were computed between propensity and the second variable among respondents-only.

The correlations between the dependent variables (column #1), between the model predictions (column #2) and between response propensity and the second variable (column #3) are negative for most of the outcome measures displayed in **Exhibit 4,** and this is expected. Past year cigarette users are considered nonrespondents in this example and we know cigarette smoking tends to be positively correlated with the outcome measures presented in these exhibits. So the negative correlation indicates as response propensity increases the probability of exhibiting one of secondary outcome measures, such as past year alcohol use, decreases. In other words, survey respondents (nosmokers) are likely to not use the substances defined by the rows in this table. The estimates in **Exhibit 4** also indicate the standard error of the correlations is relatively small for most estimates, indicating most of these correlations are statistically different from zero.

**Exhibit 4. Correlations Between Response Propensity and Selected Outcome Measures**

| Domain | #1 Correlation Between Dependent Variables | | #2 Correlation Between Model Predictions | | #3 Correlation Between Propensity and Variable | | #4 Correlation Btwn Propensity Residuals and Variable (Semi-Partial Correlation) | | #5 Correlation Between Residuals (Partial Correlation) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Corr | Std Error Full | Corr | Std Error Full | Corr | Std Error Full | Corr | Std Error Full | Corr | Std Error Full |
| **Total Sample Alcohol Use** | | | | | | | | | | |
| Total | -.2045 | .0072 | -.5719 | .0164 | -.0815 | .0088 | -.1887 | .0075 | -.1984 | .0078 |
| Age Group | | | | | | | | | | |
| 12-17 | -.4530 | .0098 | -.7751 | .0358 | -.0287 | .0126 | -.4515 | .0099 | -.4546 | .0100 |
| 18-25 | -.2800 | .0082 | -.4849 | .0331 | .0106 | .0119 | -.2895 | .0083 | -.2877 | .0082 |
| 26-34 | -.2045 | .0162 | -.3172 | .0352 | -.0382 | .0226 | -.2051 | .0165 | -.1981 | .0160 |
| 35+ | -.1047 | .0115 | -.2328 | .0284 | .1495 | .0124 | -.1431 | .0117 | -.1476 | .0120 |
| **Total Sample Any Illicit But Mrj** | | | | | | | | | | |
| Total | -.2359 | .0077 | -.5547 | .0232 | -.1656 | .0078 | -.1948 | .0077 | -.1983 | .0079 |
| Age Group | | | | | | | | | | |
| 12-17 | -.3326 | .0144 | -.7789 | .0350 | -.0398 | .0103 | -.3294 | .0144 | -.3323 | .0144 |
| 18-25 | -.2915 | .0098 | -.4972 | .0335 | -.1272 | .0101 | -.2697 | .0102 | -.2692 | .0103 |
| 26-34 | -.2343 | .0187 | -.3351 | .0373 | -.1554 | .0174 | -.2046 | .0185 | -.2008 | .0184 |
| 35+ | -.1601 | .0139 | -.2419 | .0296 | -.0729 | .0121 | -.1465 | .0141 | -.1475 | .0140 |
| **Respondents-Only Alcohol Use** | | | | | | | | | | |
| Total | n/a | | -.4530 | .0199 | -.0119 | .0125 | .0119 | .0125 | -.1415 | .0112 |
| Age Group | | | | | | | | | | |
| 12-17 | n/a | | -.7772 | .0386 | .0114 | .0123 | -.0114 | .0123 | -.1120 | .0137 |
| 18-25 | n/a | | -.4531 | .0392 | .0862 | .0161 | -.0862 | .0161 | -.1587 | .0164 |
| 26-34 | n/a | | -.2598 | .0450 | .0134 | .0284 | -.0134 | .0284 | -.0533 | .0282 |
| 35+ | n/a | | -.2089 | .0289 | .1942 | .0180 | -.1942 | .0180 | -.2279 | .0183 |
| **Respondents-Only Any Illicit But Mrj** | | | | | | | | | | |
| Total | n/a | | -.4073 | .0369 | -.0703 | .0083 | .0703 | .0083 | .0225 | .0074 |
| Age Group | | | | | | | | | | |
| 12-17 | n/a | | -.7758 | .0478 | -.0110 | .0091 | .0110 | .0091 | -.0294 | .0123 |
| 18-25 | n/a | | -.4712 | .0558 | -.0634 | .0134 | .0634 | .0134 | .0354 | .0140 |
| 26-34 | n/a | | -.2929 | .0628 | -.0667 | .0219 | .0667 | .0219 | .0531 | .0220 |
| 35+ | n/a | | -.2182 | .0347 | -.0369 | .0155 | .0369 | .0155 | .0305 | .0152 |

**Note:** Std Error Full = Standard error account for the model estimation of the predicted values. Alcohol Use and Any Illicit But Mrj refer to past year alcohol use and past year any illicit drug use except marijuana.

Notice from **Exhibit 4** that if we look at the respondents-only results, the correlation between the response propensity and a variable (column #3) and the correlation between the response propensity residuals and a variable (column #4) differ by a factor of -1 and the standard error estimates are identical. Again, this is expected. This table is examining statistics computed among respondents-only. And the residuals used in the correlation in column #4 are $(r_i - \hat{p}_i)$ which reduces to $(1 - \hat{p}_i)$ when looking at respondents-only. In column #3 the correlations are computing using $\hat{p}_i$. So the correlations in column #3 and #4 should only differ by a factor of -1.

**Exhibit 5** presents a graphical representation of the some of the results from **Exhibit 4.** This exhibit shows that the correlations between the total sample and respondent-only domains are very similar if one were to look at the correlation between the model predictions and between the model predicted response propensity and the second variable. These are columns #2 and #3 in **Exhibit 4**. However, the patterns in the partial correlations are different (column #5). **Exhibit 5** indicates, for example, that when examining past year alcohol use that the partial correlation is greatest for the 35+ years olds and the smallest for the 12-17 year-olds when computed using the total sample.

However, when looking at the respondent-only sample, the partial correlations are smallest for the 35+ year-olds and greatest for the 26-34 year-olds.

Additional empirical studies will need to be conducted, but from this simple example it appears that examining correlations by restricting the domain of interest to just the respondents is a feasible alternative when examining correlations between response propensity and second variable, however it does not appear to be useful when examining partial correlations between response propensity and a second variable if the second variable is known only known for respondents

**Exhibit 5. Comparing Correlations Computed For Total Sample and Respondents-Only, By The Four Age Groups**
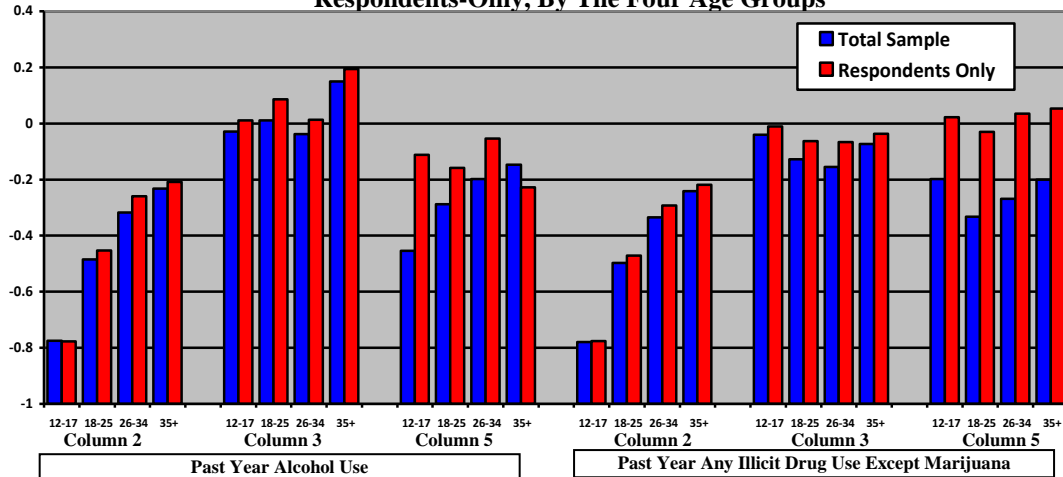


**Note:** Correlations in bar chart are taken from Columns #2, #3 and #5 in **Exhibits 4**

## 7. Some Comments about Deriving These Statistics in SAS and SUDAAN

The estimates presented in the example in **Section 6** were computed using the IML procedure in SAS 9.2. As noted in **Section 1**, all of these statistics will be available in the next release of SUDAAN which is scheduled to be released in August, 2011. The SAS code and example illustrated in **Section 6** were developed to test the new statistics that will be coming in SUDAAN.

The formulas for the linearized variance esimate of the R-indicator, the related statistics and the correlations all require the use of the information matrix associated with the model used to estimate response propensity (as well as the model used to predict a second variable if one were computing partial correlations.) Consequently, all modeling and code development was done in SAS IML to facilitate the use of matrices in the formulas. One could also use established procedures in SAS (such as SAS's REGRESS or LOGISTIC procedures) to back out needed factors for the linearized variance estimates. But this added step seems unnecessary with moderately sized data sets and when only a relatively small number of predictors are used in the models.

When programming the linearized variance estimator one would naturally seek out methods that can be used to verify the accuracy of both the mathematics and the program used to derive the linearized variates. A few things that can be done for this verification include:

• In general, variance estimates obtained using the Linearization Method will be very close to what is obtained using the Delete-1 Jackknife Method, even though these two methods use very different algorithms to derive a sampling error estimate. The Delete-1

Jackknife Method is a replication variance estimation method that involves forming replicate estimates by removing one primary sampling unit (PSU) at a time and adjusting the sample weights of the other PSUs in the same strata to account for the missing PSU. This was done for all estimates displayed in **Section 6**. For all estimates, the difference between the Taylor Series standard error estimate and the Delete-1 Jackknife standard error estimate was less than .0001.

- One can also check the linearized value of a subset of individual records using the definition of the derivative. Recall from **Section 3** that the weighted linearized variate for record $k$ associated with some statistics defined by a general function $f(x_1..x_k.., w_1..w_k..)$ was computed as $\tau_k = w_k \cdot \dfrac{\partial f}{\partial w_k}$. One can verify the computation of

the derivative for record $k$ by computing $\dfrac{\partial f}{\partial w_k} \approx \dfrac{f(x_1..x_k.., w_1..w_k + h..) - f(x_1..x_k.., w_1..w_k..)}{h}$

for some small value of $h$. This check was done with the first 100 or so records on the file for many of the statistics presented in **Section 6**.

## References

Bethlehem, J.G. (1988). "Reduction of Nonresponse Bias through Regression Estimation." Journal of Official Statistics, 4, 3, 251-260.

Biemer, Paul P. and Lyberg, Lars E. (2003) *Introduction to Survey Quality*, John Wiley & Sons, New Jersey.

Binder, D. (1995) "Taylor Linearization for Single and Two Phase Samples: A Cookbook Approach", *Proceedings of the Survey Research Methods Section*, ASA, 1995, pp. 132-137.

Bose, Jonaki (2001) "Nonresponse Bias Analysis at the National Center for Education Statistics." *Proceeding of Statistics Canada Symposium, 2001*.

Cobben, F. and Schouten, B. (2007) "An Empirical Validation of R-indicators." Discussion Paper, CBS, Voorburg. Available at http://www.cbs.nl/NR/rdonlyres/A789F70B-AF96-4635-9E3D-050EB2B3FFB0/0/200806x10pub.pdf

Office of Applied Studies, Substance Abuse and Mental Health Services Administration (SAMHSA). Additional information on the National Survey on Drug Use and Health (NSDUH) can be found here http://www.oas.samhsa.gov/nhsda.htm.

Schouten, B., Cobben, F. and Bethlehem, J. (2009) "Indicators for the Representativeness of Survey Response." *Survey Methodology, June 2009*. Vol 35, No. 1 pp 101-113.

Schouten, B. and Cobben, F. (2007), "R-indexes for the comparison of different fieldwork strategies and data collection modes." Discussion Paper 07002, Statistics Netherlands, Voorburg, The Netherlands. Available at http://www.risq-project.eu/papers/schouten-cobben-2007-a.pdf

Shah, B. V. (2002) "Calculus of Taylor Deviations" paper presented at the Joint Statistical Meetings, ASA, 2002.

Skinner, C., Shlomo, N., Schouten, B., Zhang, L. & Bethlehem, J. (2009) "Measuring Survey Quality Through Representativeness Indicators Using Sample and Population Based Information." Paper presented at the NTTS Conference, 18-20 February 2009, Brussels, Belgium. Available at http://www.risq-project.eu/papers/skinner-shlomo-schouten-zhang-bethlehem-2009-a.pdf

Research Triangle Institute (2008). SUDAAN Language Manual, Release 10.0 Research Triangle Park, NC: Research Triangle Institute.