Online Appendixes

Researchers who wish to conduct a sensitivity analysis for estimates made using the generalized scale-up method can therefore (1) assume values or ranges of values for K_{F_1}, K_{F_2}, c_1, c_2, c_3, δ_F, τ_F, and η_F; and (2) use Corollary D.11 to determine the resulting values of N_H. Thus, researchers can use this approach to explore the sensitivity of their estimates to all of the assumptions they had to make, individually and jointly.

E Approximate unbiasedness of compound ratio estimators

E.1 Overview

Several of the estimators we propose are nonlinear, which means that they are not design-unbiased (Sarndal et al., 1992). While ratio estimators are common in survey sampling and the bias of these estimators is commonly regarded as insignificant (Sarndal et al., 1992), several of the estimators we propose are somewhat more complex than standard ratio estimators. In fact, all of our nonlinear estimators turn out to all be special cases of a ratio of ratios (Table E.1), which is also known as a double ratio estimator (Rao and Pereira, 1968). Any double ratio can be written

$$R_d = \frac{R_1}{R_0} = \frac{\frac{\bar{y}_1}{\bar{x}_1}}{\frac{\bar{y}_0}{\bar{x}_0}} = \frac{\bar{y}_1 \bar{x}_0}{\bar{x}_1 \bar{y}_0}. \tag{E.1}$$

If we have unbiased estimators for each of the four terms, we can estimate R_d by

$$\widehat{r}_d = \frac{\widehat{\bar{y}}_1 \widehat{\bar{x}}_0}{\widehat{\bar{x}}_1 \widehat{\bar{y}}_0}. \tag{E.2}$$

In this appendix we investigate when we can expect the biases in our estimators to be small enough to be negligible; we conclude that, in practice, the bias is typically negligible when compared to sampling and non-sampling error.

E.2 The general case

We will focus on the relative bias in our estimator, $\widehat{r}_d$. The relative bias is given by

$$B_d = \frac{\mathbb{E}[\widehat{r}_d] - R_d}{R_d}. \tag{E.3}$$

B_d expresses the bias in our estimator $\widehat{r}_d$ in terms of the true value; a relative bias of 0.5, for example, means that our estimator is typically 0.5 times bigger than the true value. This is a natural quantity to consider because estimators that have small relative bias have small bias in substantive terms.

Our approach will be to follow Rao and Pereira (1968) in using a Taylor series to form an approximation to the relative bias. This is accomplished in Result E.1.

Result E.1 *(Rao and Pereira, 1968) If $\widehat{\bar{x}}_0$, $\widehat{\bar{x}}_1$, $\widehat{\bar{y}}_0$, and $\widehat{\bar{y}}_1$ are unbiased estimators, and $|(\widehat{\bar{x}}_1 - \bar{x}_1)/\bar{x}_1| < 1$ and $|(\widehat{\bar{y}}_0 - \bar{y}_0)/\bar{y}_0| < 1$, then the relative bias of the double ratio estimator, B_d, is approximated by*

$$B_d = \frac{\mathbb{E}[\widehat{r}_d] - R}{R} \approx B'_d = C_{\widehat{\bar{x}}_1,\widehat{\bar{y}}_0} - C_{\widehat{\bar{x}}_1,\widehat{\bar{y}}_1} - C_{\widehat{\bar{y}}_0,\widehat{\bar{y}}_1} - C_{\widehat{\bar{x}}_0,\widehat{\bar{x}}_1} - C_{\widehat{\bar{x}}_0,\widehat{\bar{y}}_0} + C_{\widehat{\bar{y}}_1,\widehat{\bar{x}}_0} + C^2_{\widehat{\bar{y}}_0} + C^2_{\widehat{\bar{x}}_1}, \tag{E.4}$$

where $C_{\widehat{\bar{x}},\widehat{\bar{y}}} = \frac{cov(\widehat{\bar{x}},\widehat{\bar{y}})}{\bar{x}\bar{y}}$ is the relative covariance between $\widehat{\bar{x}}$ and $\widehat{\bar{y}}$, and $C^2_{\widehat{\bar{y}}} = \frac{var(\widehat{\bar{y}})}{\widehat{\bar{y}}^2}$.

Proof: Define

$$\delta_{\widehat{\bar{x}}_0} = \frac{\widehat{\bar{x}}_0 - \bar{x}_0}{\bar{x}_0}, \tag{E.5}$$

Estimator	Reference	Form	$\widehat{\bar{x}}_0$	$\widehat{\bar{y}}_1$	$\widehat{\bar{x}}_1$	$\widehat{\bar{y}}_0$	Approx. rel. bias
$\widehat{\phi}_F$	Res. B.6	$K\widehat{\bar{x}}_0/\widehat{\bar{y}}_0$	$\sum_{i\in s_F} y_{i,\mathcal{A}_{F_1}}/\pi_i$	-	-	$\sum_{i\in s_F} y_{i,\mathcal{A}_{F_2}}/\pi_i$	$C^2_{\widehat{\bar{y}}_0} - C_{\widehat{\bar{y}}_0,\widehat{\bar{x}}_0}$
$\widehat{\bar{v}}_{H,F}$	Res. C.2	$K\widehat{\bar{x}}_0/\widehat{\bar{y}}_0$	$\sum_{i\in s_H} \tilde{v}_{i,\mathcal{A}_H\cap F}/c\pi_i$	-	-	$\sum_{i\in s_H} 1/c\pi_i$	$C^2_{\widehat{\bar{y}}_0} - C_{\widehat{\bar{y}}_0,\widehat{\bar{x}}_0}$
$\widehat{\bar{d}}_{H,F}$	Res. C.5	$K\widehat{\bar{x}}_0/\widehat{\bar{y}}_0$	$\sum_{i\in s_H} y_{i,\mathcal{A}_H\cap F}/c\pi_i$	-	-	$\sum_{i\in s_H} 1/c\pi_i$	$C^2_{\widehat{\bar{y}}_0} - C_{\widehat{\bar{y}}_0,\widehat{\bar{x}}_0}$
$\widehat{\delta}_F$	Res. C.6	$K\widehat{\bar{x}}_0/(\widehat{\bar{y}}_0\ \widehat{\bar{x}}_1)$	$\sum_{i\in s_H} y_{i,\mathcal{A}_H\cap F}/c\pi_i$	-	$\sum_{i\in s_F} y_{i,\mathcal{A}_F}/\pi_i$	$\sum_{i\in s_H} 1/c\pi_i$	$C^2_{\widehat{\bar{y}}_0} + C^2_{\widehat{\bar{x}}_1} - C_{\widehat{\bar{y}}_0,\widehat{\bar{x}}_0}$
$\widehat{\tau}_F$	Res. C.7	$K\widehat{\bar{x}}_0/(\widehat{\bar{y}}_0\ \widehat{\bar{x}}_1)$	$\sum_{i\in s_H} \tilde{v}_{i,\mathcal{A}_H\cap F}/c\pi_i$	-	$\sum_{i\in s_H} y_{i,\mathcal{A}_H\cap F}/c\pi_i$	$\sum_{i\in s_H} 1/c\pi_i$	$C^2_{\widehat{\bar{y}}_0} + C^2_{\widehat{\bar{x}}_1} - C_{\widehat{\bar{y}}_0,\widehat{\bar{x}}_0}$
$\widehat{N}_H$	Res. C.8	$K\widehat{\bar{y}}_1\widehat{\bar{x}}_0/\widehat{\bar{y}}_0$	$\sum_{i\in s_H} 1/c\pi_i$	$\sum_{i\in s_F} y_{i,H}/\pi_i$	-	$\sum_{i\in s_H} \tilde{v}_{i,\mathcal{A}_H\cap F}/c\pi_i$	$C^2_{\widehat{\bar{y}}_0} - C_{\widehat{\bar{y}}_0,\widehat{\bar{x}}_0}$
$\widehat{N}_H$	Res. C.10	$K\widehat{\bar{x}}_0/\widehat{\bar{y}}_0$	$\sum_{i\in s_F} y_{i,H}/\pi_i$	-	-	$\sum_{i\in s_F}\sum_j y_{i,A_j}/\pi_i$	$C^2_{\widehat{\bar{y}}_0} - C_{\widehat{\bar{y}}_0,\widehat{\bar{x}}_0}$

Table E.1: Description of the general form of the nonlinear estimators we propose. K is a constant, $\widehat{\bar{y}}_1$ and $\widehat{\bar{x}}_1$ are taken from s_F, while $\widehat{\bar{x}}_0$ and $\widehat{\bar{y}}_0$ are taken from s_H. Our nonlinear estimators are all special cases of the double ratio estimator, which we define and discuss below. Note that the estimator for $\widehat{N}_H$ that involves adjusting a basic scale-up estimate (Result C.10) would, in practice, take these adjustment factors from other studies; we therefore assume that these adjustment factors are independent of the quantities that go into the scale-up estimate, and treat them as constants.

with analogous definitions for $\delta_{\widehat{\bar{x}}_1}$, $\delta_{\widehat{\bar{y}}_1}$, and $\delta_{\widehat{\bar{y}}_0}$. We can express r_d as

$$\widehat{r}_d = R\frac{(1+\delta_{\widehat{\bar{y}}_1})(1+\delta_{\widehat{\bar{x}}_0})}{(1+\delta_{\widehat{\bar{y}}_0})(1+\delta_{\widehat{\bar{x}}_1})}. \tag{E.6}$$

The relative bias then becomes

$$B_d = \frac{\mathbb{E}[\widehat{r}_d] - R}{R} = \mathbb{E}\left[\frac{(1+\delta_{\widehat{\bar{y}}_1})(1+\delta_{\widehat{\bar{x}}_0})}{(1+\delta_{\widehat{\bar{y}}_0})(1+\delta_{\widehat{\bar{x}}_1})}\right] - 1. \tag{E.7}$$

The strategy is now to expand the two factors in the denominator and to then discard high-order terms. What remains will be an approximation to the true relative bias.

Recall that if $|x| < 1$ then $\frac{1}{1-x} = \sum_{i=0}^{\infty} x^i$ and, in particular, $\frac{1}{1+x} = 1 - x^2 + x^3 - \cdots$. We'll make use of this expansion for the two factors in the denominator of Equation E.7; that is, we assume that $|\delta_{\bar{y}_0}| < 1$ and $|\delta_{\bar{x}_1}| < 1$. Then we have

$$B_d = \mathbb{E}\left[(1+\delta_{\widehat{\bar{y}}_1})(1+\delta_{\widehat{\bar{x}}_0})(1-\delta_{\widehat{\bar{y}}_0}+\delta^2_{\widehat{\bar{y}}_0}-\cdots)(1-\delta_{\widehat{\bar{x}}_1}+\delta^2_{\widehat{\bar{x}}_1}-\cdots)\right] - 1 \tag{E.8}$$

If we multiply this out and retain only terms up to order 2, we obtain the following approximation:

$$B_d \approx \mathbb{E}\left[\delta_{\widehat{\bar{x}}_1}\delta_{\widehat{\bar{y}}_0} + \delta_{\widehat{\bar{x}}_0}\delta_{\widehat{\bar{y}}_1} - \delta_{\widehat{\bar{x}}_0}\delta_{\widehat{\bar{y}}_0} - \delta_{\widehat{\bar{x}}_0}\delta_{\widehat{\bar{x}}_1} - \delta_{\widehat{\bar{x}}_1}\delta_{\widehat{\bar{y}}_1} - \delta_{\widehat{\bar{y}}_0}\delta_{\widehat{\bar{y}}_1} + \delta_{\widehat{\bar{x}}_0} + \delta_{\widehat{\bar{y}}_1} - \delta_{\widehat{\bar{x}}_1} - \delta_{\widehat{\bar{y}}_0} - \delta^2_{\widehat{\bar{y}}_0} - \delta^2_{\widehat{\bar{x}}_1}\right]. \tag{E.9}$$

Since we assumed that the estimators for the individual components of r_d are unbiased, we know that

$$\mathbb{E}[\delta_{\widehat{\bar{x}}_1}] = 0, \tag{E.10}$$

We can also determine that

$$\mathbb{E}[\delta_{\widehat{\bar{x}}_1}\delta_{\widehat{\bar{y}}_1}] = \frac{\text{cov}(\widehat{\bar{x}}_1, \widehat{\bar{y}}_1)}{\bar{x}_1\bar{y}_1}, \tag{E.11}$$

and, that

$$\mathbb{E}[\delta^2_{\widehat{\bar{x}}_1}] = \frac{\text{var}(\widehat{\bar{x}}_1)}{\bar{x}_1^2}. \tag{E.12}$$

Applying these relationships to Equation E.9, we find

$$B_d \approx C_{\widehat{\bar{x}}_0,\widehat{\bar{y}}_1} + C_{\widehat{\bar{x}}_1,\widehat{\bar{y}}_0} - C_{\widehat{\bar{x}}_0,\widehat{\bar{x}}_1} - C_{\widehat{\bar{x}}_0,\widehat{\bar{y}}_0} - C_{\widehat{\bar{x}}_1,\widehat{\bar{y}}_1} - C_{\widehat{\bar{y}}_0,\widehat{\bar{y}}_1} + C^2_{\bar{x}_1} + C^2_{\bar{y}_0}, \tag{E.13}$$

which is our result. ■

Result E.1 is useful because it reveals the behavior of double ratio estimators in quite general contexts. To understand what it says a bit more intuitively, note that Result E.1 is framed in terms of the relative covariances and variances of the *estimators* $\widehat{\bar{x}}_0$, $\widehat{\bar{x}}_1$, $\widehat{\bar{y}}_0$, and $\widehat{\bar{y}}_1$. In the special case of simple random sampling with replacement, we can re-write the approximation in terms of the finite population variances and covariances and a constant, κ:

$$B'_d = \kappa\left[C_{x_1,y_0} - C_{x_1,y_1} - C_{y_0,y_1} - C_{x_0,x_1} - C_{x_0,y_0} + C_{y_1,x_0} + C^2_{y_0} + C^2_{x_1}\right], \tag{E.14}$$

where $\kappa = \left(\frac{1}{n} - \frac{1}{N}\right)$, n is our sample size, and N is the size of the population. In the case of simple random sampling, the relative bias depends upon the finite population variances of the underlying population values and the size of our sample.

For designs other than simple random sampling, there is no analogous expression as simple as Equation E.14. However, speaking roughly, if we have an idea that our

sampling plan has a typical design effect (deff) for the quantities inside the square brackets in Equation E.14, then we can see that we would simply replace the κ in Equation E.14 by $(\kappa \cdot \text{deff})$ in order to get a sense of the approximate relative bias.

Notice, also, that Result E.1 is framed largely in terms of relative covariances. When we apply Result E.1, we will often make use of the fact that the relative covariances can be expressed in terms of correlations and coefficients of variation as follows:

$$C_{\widehat{\bar{x}},\widehat{\bar{y}}} = \frac{\text{cov}(\widehat{\bar{x}},\widehat{\bar{y}})}{\bar{x}\bar{y}} = \frac{\rho_{\widehat{\bar{x}},\widehat{\bar{y}}}\sqrt{\text{var}(\widehat{\bar{x}})}\sqrt{\text{var}(\widehat{\bar{y}})}}{\bar{x}\bar{y}} \tag{E.15}$$

$$= \rho_{\widehat{\bar{x}},\widehat{\bar{y}}}\ \text{cv}(\widehat{\bar{x}})\ \text{cv}(\widehat{\bar{y}}), \tag{E.16}$$

where $\rho_{\widehat{\bar{x}},\widehat{\bar{y}}}$ is the correlation between the estimators $\widehat{\bar{x}}$ and $\widehat{\bar{y}}$, and $\text{cv}(\widehat{\bar{x}}) = \frac{\sqrt{\text{var}(\widehat{\bar{x}})}}{\bar{x}}$ is the coefficient of variation of the estimator $\widehat{\bar{x}}$. We will also make use of the fact that $C^2_{\widehat{\bar{x}}} = \text{cv}(\widehat{\bar{x}})^2$.

E.3 Applying Result E.1 to scale-up

We now apply Result E.1 to understand the biases in the nonlinear estimators we propose for realistic situations. For each particular estimator, we can simplify the expression in Result E.1. In order to do so, we first remove terms that do not appear in the estimator itself (for example, in $\widehat{\delta}_F$, there is no $\widehat{\bar{y}}_1$). Additionally, we assume that the estimates produced from a sample from the frame population and a sample from the hidden population will be independent of one another, meaning that their correlation will be 0. Table E.1 summarizes the nonlinear estimators we propose, along with the specific version of the approximate relative bias from Result E.1 that

applies.

Finally, in order to give a sense of the magnitude of the coefficients of variation and correlations found in real studies, we estimated the quantities that go into the approximate relative bias from the studies available to us. Table E.2 shows the coefficients of variation for the estimated degree (the values of $\widehat{\bar{x}}_1$ for $\widehat{\delta}_F$) in surveys from Rwanda, the United States, and Curitiba, Brazil. Further, Tables E.3 and E.4 show the relevant coefficients of variation and pairwise correlations for all remaining quantities using data from Curitiba, Brazil (currently, the only setting where we have data from a sample of the hidden population). For all values in these tables, the estimated variance of the estimators is calculated using the bootstrap methods presented in Section F.1.

Since we have both a sample from the frame population and a sample from the hidden population in Curitiba, we can compute numerical estimates of the bias of each nonlinear estimator in the context of that study. We can see that in this study bias caused by the nonlinearity of the estimator was not a big problem: in each case, the estimated approximate bias was less than one percent of the estimate (Table E.5).

To conclude, we derived an expression for the approximate relative bias in double ratio estimators in general. We then simplified the approximation for each specific nonlinear estimator that we propose. Finally, we used data from a real scale-up study in Curitiba, Brazil to estimate magnitude of the biases caused by the non-linearity of the estimators in a specific scale-up study. From these results, we conclude that theses estimators are essentially unbiased, and that sampling error and non-sampling error will dominate any bias introduced by the nonlinear form of the estimators.

$\widehat{\text{cv}}(\widehat{\bar{d}})$	source
0.05	Rwanda
0.10	Curitiba
0.02	US

Table E.2: Estimated coefficients of variation for the average degree from 3 different scale-up surveys. These play a role in the approximate relative bias for the estimate of $\widehat{\delta}_F$. Our approximation tells us that the larger these values are, the worse the relative bias will be. The estimates were computed using the rescaled bootstrap procedure.

	estimated coef. of variation
$\sum_{i \in s_H} y_{i,\mathcal{A} \cap F} / c\pi_i$	0.08
$\sum_{i \in s_H} \tilde{v}_{i,\mathcal{A} \cap F} / c\pi_i$	0.08
$\sum_{i \in s_H} 1 / c\pi_i$	0.06

Table E.3: Estimated coefficients of variation for quantities derived from a sample from the hidden population. These quantities play a role in the approximate relative bias for the estimate of all of the nonlinear estimators we propose. The estimates were computed using the respondent-driven sampling bootstrap procedure (Salganik, 2006).

F Variance estimation and confidence intervals

In addition to producing point estimates, researchers must also produce confidence intervals around their estimates. The procedure currently used by scale-up researchers begins with the variance estimator proposed in Killworth et al. (1998b):

$$\widehat{se}(\widehat{N}_H) = \sqrt{\frac{N \cdot \widehat{N}_H}{\sum_{i \in s_F} \widehat{d}_{i,U}}}, \tag{F.1}$$

and then produces a confidence interval:

$$\widehat{N}_H \pm z_{1-\alpha/2} \widehat{se}(\widehat{N}_H), \tag{F.2}$$

where $1-\alpha$ is the desired confidence level (typically 0.95), and $z_{\alpha/2}$ is the $\alpha/2$ quantile of the standard Normal distribution.