

Ratio and Regression Estimation - Some Further Notes

1. Unequal Probability Sampling

Ratio and regression estimators can be applied in the situation of unequal probability sampling (see sections 7.5 and 8.2 of Thompson).

- The generalized ratio estimator of the population total τ_y is

$$\hat{\tau}_G = \frac{\hat{\tau}_y}{\hat{\tau}_x} \tau_x$$

where $\hat{\tau}_y$ and $\hat{\tau}_x$ are the Horvitz-Thompson estimators of τ_y and τ_x , respectively. As a ratio of two unbiased estimators, this estimator is not unbiased. The estimated variance is given at the top of p. 78 (equation 10).

- The generalized ratio estimator can also be used to derive an estimator of μ_y with unequal probability sampling when N is unknown by letting x be equal to 1 for all units:

$$\hat{\mu}_G = \frac{\hat{\tau}_\pi}{\widehat{N}} = \frac{\sum_{i=1}^{\nu} y_i / \pi_i}{\sum_{i=1}^{\nu} 1 / \pi_i}$$

where ν is the number of distinct units in the sample. We already looked at this estimator for PPS sampling in the notes on “Unequal Probability Sampling” (p. 25, equation 3b) using Hansen-Hurwitz estimators for the numerator and denominator, so this is a generalization to any unequal probability sampling plan. Thompson notes on p. 78 that this estimator is sometimes a useful alternative even when N is known if there is not a linear relationship between the inclusion probabilities and the y values; $\hat{\mu}_G$ may have a smaller variance than the estimator $\hat{\tau}_\pi / N$, as for the elephant example discussed on p. 78. The analogous estimator of τ_y is then

$$\hat{\tau}_G = N \hat{\mu}_G = N \frac{\hat{\tau}_\pi}{\widehat{N}} = N \frac{\sum_{i=1}^{\nu} y_i / \pi_i}{\sum_{i=1}^{\nu} 1 / \pi_i}.$$

Note that we “adjust” the estimator of τ_y by the ratio N / \widehat{N} with the idea being that \widehat{N} will tend to overestimate N when $\hat{\tau}_\pi$ overestimates τ_y and vice-versa.

- These same ideas can be extended to regression estimation yielding generalized regression estimators of μ_y and τ_y (section 8.2 of Thompson).

2. Design and Model Approaches to Sampling

The ratio and regression estimators are not design-unbiased. What does that mean? That means if we view the population of y values y_1, \dots, y_N as fixed with $\mu_y = \sum_{i=1}^N y_i / N$, then it is not necessarily true that $E(\hat{\mu}_r) = \mu_y$ or that $E(\hat{\mu}_L) = \mu_y$ (we have to say “not necessarily true” because it may be true for specific cases). This is

shown by Thompson for the ratio estimator in an example on p. 72 in section 7.2. The bias is usually small if there is a linear relationship between the x and y variables.

However, the ratio and regression estimators are model-unbiased if we assume the right model for the population. What is the model-based approach?

- In the model-based approach to sampling, we consider the y values in the population to be random variables denoted by Y_1, \dots, Y_N . That is, they are only one possible realization of a process that generated the population. Therefore, the population total $\tau_y = \sum_{i=1}^N Y_i$ and the population mean $\mu_y = \tau_y/N$ are also random variables (this is very important to remember). What it means, therefore, for an estimator of the population total to be model-unbiased, is if the expected value of the estimator is equal to the expected value of the population total (and analogously for an estimator of the population mean). Since the population total and mean are not fixed quantities but random variables under the model-based approach, we might more accurately say that our estimators are predictors of these random quantities.
- As an example, suppose we have an auxiliary variable x . One way to model y is to consider the x values x_1, \dots, x_N to be fixed, but the y values to be random variables:

$$Y_i = \beta x_i + \epsilon_i$$

where the ϵ_i 's are independent random variables with $E(\epsilon_i) = 0$. This is the linear regression model without an intercept. Under this model,

$$E(Y_i) = E(\beta x_i + \epsilon_i) = \beta x_i + E(\epsilon_i) = \beta x_i.$$

Now consider the ratio estimator of μ_y :

$$\hat{\mu}_r = r\mu_x = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i} \mu_x$$

(note that we have replaced the fixed values y_1, \dots, y_n by the random variables Y_1, \dots, Y_n in the model-based approach). Then, regardless of how our sample is chosen:

$$\begin{aligned} E(\hat{\mu}_r) &= E\left(\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i} \mu_x\right) = \frac{E(\sum_{i=1}^n Y_i)}{\sum_{i=1}^n x_i} \mu_x \quad (\text{since the } x_i\text{'s and } \mu_x \text{ are fixed}) \\ &= \frac{\sum_{i=1}^n E(Y_i)}{\sum_{i=1}^n x_i} \mu_x = \frac{\sum_{i=1}^n \beta x_i}{\sum_{i=1}^n x_i} \mu_x = \beta \mu_x. \end{aligned}$$

Note also that the expected value of the population mean is

$$E\left(\frac{1}{N} \sum_{i=1}^N Y_i\right) = \frac{1}{N} \sum_{i=1}^N E(Y_i) = \frac{1}{N} \sum_{i=1}^N \beta x_i = \beta \mu_x.$$

Since $E(\widehat{\mu}_r) = E(\tau_y)$, $\widehat{\mu}_r$ is model-unbiased for predicting the population mean under the model above.

- Note that the randomness in the model-based approach comes from the model for the Y_i 's and not from how the sample was chosen. Therefore, model-unbiasedness holds regardless of how the sample was chosen, randomly or not. However, it depends crucially on the model assumed. In the example here, this means not only that $E(Y_i) = \beta x_i$ but also that the ϵ_i are independent, regardless of how the units were chosen for the sample.
- In the design-based approach, all the randomness comes from how the sample was chosen since y_1, \dots, y_N are fixed. Therefore, design-unbiasedness of an estimator depends entirely on how the sample was chosen (e.g., SRS or whatever sampling plan is assumed).
- Extension to the regression estimator: model-unbiasedness of the regression estimator holds under the linear regression model with an intercept:

$$Y_i = \alpha + \beta x_i + \epsilon_i$$

where the ϵ_i 's are independent random variables and $E(\epsilon_i) = 0$.

- Implications for sampling design: if we truly believe in the regression model with independent errors, then we can select our sample any way we like and still preserve model-unbiasedness. In particular, we could select units in such a way to minimize the variance of our estimators. For the linear regression model without an intercept, regression theory tells us to choose the units with the largest x -values. For the linear regression model with an intercept, it tells us to choose the units with the smallest and largest x -values. Unfortunately, these designs do not allow us to assess the appropriateness of the model and could lead to large errors if the model turns out to be wrong. A compromise would be to choose units with a wide range of x -values, perhaps by stratified random sampling, stratifying by x . Most independent observers would be skeptical of a selection process without some element of randomness in it.

3. Least-squares estimators in the model-based approaches

The regression estimators of α and β in the linear regression model are the usual least-squares estimators. However, the ratio estimator of β in the regression model without an intercept is $\bar{y}/\bar{x} = \sum_{i=1}^n y_i / \sum_{i=1}^n x_i$ which is not the least squares estimator. The least squares estimator is

$$\widehat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}.$$

Why is that and why don't we use the least-squares estimator?

- First, we can show that the least-squares estimator of β in the regression model without an intercept also gives model-unbiased estimators of the population mean and total, so neither the ratio or least-squares estimator has an advantage in that respect.
- The model unbiasedness of the ratio and regression estimators (as well as the least squares estimator for the model without intercept) of the population mean and total depends only on the assumption that the ϵ_i 's are independent with mean 0. We do not need to assume that they have identical distributions. As Thompson shows on pp. 80 and 84 of the text, the ratio estimator is the best linear unbiased estimator (BLUE) of β if we assume that $\text{Var}(\epsilon_i)$ is proportional to x_i while the least squares estimator is the BLUE if we assume that the variances are all equal. Variance proportional to x is actually a quite reasonable assumption in many ratio estimation problems, or at least more reasonable than constant variance. For example, in the trees example on pp. 57-8 of the notes, it seems reasonable to assume that that the larger the number of trees in a plot, the more the actual number might vary. That is, if we had a set of plots all with photo estimates of 10 trees, we would expect the actual numbers to vary less than for a set of plots all with photo estimates of 50 trees.