The model based approach to survey sampling

Some Notation

- $y = (y_1, \ldots, y_N)$ is the population characteristic of interest.
- $x = (x_1, \ldots, x_N)$ is a known auxiliary variable with $x_i > 0$ for all i.
- $\Delta(\cdot)$ is the sampling design.
- $\mathrm{x_{smp}}$ and $\mathrm{y_{smp}}$ are the sample observations.
- We wish to estimate $\mu(y) = \bar{Y} = \sum_{i=1}^{N} y_i/N$, the population mean.

The model

We assume that $y_1, \ldots, y_n$ are realized values of the independent random variables $Y_1, \ldots, Y_N$ where

$$E(Y_i) = \beta x_i \quad \text{and} \quad V(Y_i) = \sigma^2 v(x_i)$$

Here β and σ^2 are unknown parameters and $v(\cdot)$ is some known function. For example it could be $v(x_i) = x_i$

Let ξ denote the joint distribution of $Y_1, \ldots, Y_N$.

Discussion

For a fixed *smp* we can write

$$\mu(Y) = \frac{1}{N}\left(\sum_{i \in smp} Y_i + \sum_{j \notin smp} Y_j\right)$$

Given the observed values $\mathrm{y_{smp}}$ and $\mathrm{x_{smp}}$ if

$$\hat{\beta}_{\mathrm{smp}} = \hat{\beta}(\mathrm{y_{smp}}, \mathrm{x_{smp}})$$

is a good estimate of β then

$$\delta = \frac{1}{N}\left(\sum_{i \in smp} y_i + \hat{\beta}_{\text{smp}} \sum_{j \notin smp} x_j\right)$$

should be a good estimator or predictor of $\mu(Y)$.

We say that δ is ξ-unbiased for $\mu(Y)$ if for each *smp*

$$E_\xi(\delta(\mathrm{Y}_{\text{smp}}) - \mu(Y)) = 0$$

Note

$$\begin{aligned} E_\xi(\delta(\mathrm{Y}_{\text{smp}}) - \mu(Y)) &= \frac{1}{N} E_\xi\left(\hat{\beta}_{\text{smp}} \sum_{j \notin smp} x_j - \sum_{j \notin smp} Y_j\right) \\ &= \frac{1}{N}\left(E_\xi(\hat{\beta}_{\text{smp}}) \sum_{j \notin smp} x_j - \sum_{j \notin smp} \beta x_j\right) \\ &= \frac{1}{N}\left(E_\xi(\hat{\beta}_{\text{smp}}) - \beta\right) \sum_{j \notin smp} x_j \end{aligned}$$

Since each $x_i > 0$ the above will be equal to 0 if and only if $E_\xi(\hat{\beta}_{\text{smp}}) - \beta = 0$. That is δ will be ξ-unbiased for $\mu(Y)$ if and only if $\hat{\beta}_{\text{smp}}$ is unbiased for each sample, *smp*.

Given our model and y_{smp} and x_{smp} the best estimate of β is the weighted least squares estimate, $\hat{\beta}^*_{\text{smp}}$. This is the solution to the problem

$$\min_\beta \sum_{i \in smp} \frac{(y_i - \beta x_i)^2}{v(x_i)}$$

and it is given by

$$\hat{\beta}^*_{\text{smp}} = \left(\sum_{i \in smp} \frac{x_i y_i}{v(x_i)}\right) / \left(\sum_{i \in smp} \frac{x_i^2}{v(x_i)}\right)$$

Its model variance is given by

$$V(\hat{\beta}^*_{\text{smp}}) = \sigma^2 / \sum_{i \in smp} (x_i^2 / v(x_i))$$

In the special case when $v(x_i) = x_i$ we have

$$\hat{\beta}^*_{\text{smp}} = \frac{\sum_{i \in smp} y_i}{\sum_{i \in smp} x_i} = \frac{\bar{y}_{smp}}{\bar{x}_{smp}} \quad \text{and} \quad V(\hat{\beta}^*_{\text{smp}}) = \sigma^2 / \sum_{i \in smp} x_i$$

For a fixed *smp* the predictor becomes

$$\begin{aligned}
\delta^* &= \frac{1}{N}\left(\sum_{\text{i}\in\text{smp}} Y_i + \hat{\beta}^*_{\text{smp}} \sum_{\text{j}\notin\text{smp}} x_j \right) \\
&= \frac{1}{N}\left(\sum_{\text{i}\in\text{smp}} Y_i + \frac{\bar{Y}_{smp}}{\bar{x}_{smp}} \sum_{\text{j}\notin\text{smp}} x_j \right) \\
&= \frac{\sum_{\text{i}\in\text{smp}} Y_i}{N}\left(1 + \frac{\sum_{\text{j}\notin\text{smp}} x_j}{\sum_{\text{i}\in\text{smp}} x_j} \right) \\
&= \frac{\bar{Y}_{smp}}{\bar{x}_{smp}} \frac{\sum_{i=1}^{N} x_i}{N} \\
&= \frac{\bar{Y}_{smp}}{\bar{x}_{smp}} \bar{x}
\end{aligned}$$

which is the ratio estimator or predictor.

Next we find the model variance of this predictor for a fixed *smp*.. Let *nsmp* denote the units not in the sample and $\bar{x}_{nsmp}$ their average x value. Since

$$\begin{aligned}
\delta^* - \mu(Y) &= \frac{\bar{Y}_{smp}}{\bar{x}_{smp}} \bar{x} - \frac{\sum_{\text{i}\in\text{smp}} Y_i}{N} - \frac{\sum_{\text{j}\notin\text{smp}} Y_j}{N} \\
&= \frac{1}{N}\left(\left(\frac{N\bar{x}}{n\bar{x}_{smp}} - 1 \right) \sum_{i \in smp} Y_i - \sum_{j \notin smp} Y_j \right) \\
&= \frac{1}{N}\left(\frac{(N-n)\bar{x}_{nsmp}}{n\bar{x}_{smp}} \sum_{i \in smp} Y_i - \sum_{j \notin smp} Y_j \right)
\end{aligned}$$

we have

$$
\begin{aligned}
V_\xi(\delta^* - \mu(Y)) &= \left(\frac{(N-n)\,\bar{x}_{nsmp}}{N\,n\,\bar{x}_{smp}}\right)^2 n\,\bar{x}_{smp}\,\sigma^2 + \frac{N-n}{N^2}\bar{x}_{nsmp}\sigma^2 \\
&= \frac{N-n}{N^2}\,\sigma^2\left(\frac{N-n}{n}\frac{\bar{x}^2_{nsmp}}{\bar{x}_{smp}} + \bar{x}_{nsmp}\right) \\
&= \frac{N-n}{N^2}\,\sigma^2\,\bar{x}_{nsmp}\left(\frac{N-n}{n}\frac{\bar{x}_{nsmp}}{\bar{x}_{smp}} + 1\right) \\
&= \frac{N-n}{N^2}\,\sigma^2\,\bar{x}_{nsmp}\,\frac{\sum_{i=1}^N x_i}{n\,\bar{x}_{smp}} \\
&= \frac{1-f}{n}\frac{\bar{x}_{nsmp}}{\bar{x}_{smp}}\,\bar{x}\,\sigma^2 \\
&\widehat{=} \frac{1-f}{n}\frac{\bar{x}_{nsmp}}{\bar{x}_{smp}}\,\bar{x}\,\frac{1}{n-1}\sum_{i\in smp}\frac{(y_i - \hat{\beta}^*_{\text{smp}}\,x_i)^2}{x_i}
\end{aligned}
$$

where $f = n/N$.

Note this variance depends just on the model and not how the sample was selected. It is minimized if smp contains the n units with the largest x values.

If the design Δ is srs without replacement and n is large then δ^* is approximately design unbiased and

$$
\begin{aligned}
Var(\delta^*;\Delta) &\doteq \frac{1-f}{n}\sum_{i=1}^N \frac{(y_i - \beta^* x_i)^2}{N-1} \\
&\widehat{=} \frac{1-f}{n}\sum_{i\in smp}\frac{(y_i - \hat{\beta}^*_{\text{smp}}\,x_i)^2}{n-1}
\end{aligned}
$$

where $\beta^* = \sum_{i=1}^N y_i / \sum_{i=1}^N x_i$.