No Respuesta

Cap. 15 - Model Assisted Survey Sampling - Erik Sarndal

Daniel Czarnievicz
Lucía Coudet

Universidad de la República

Martes 29 de Noviembre de 2017

Índice

Planteamiento del problema

Se toma una muestra S de tamaño n_s de la población finita $U = (1; \ldots; k; \ldots; N)$, bajo un diseño $p(.)$ con probabilidades de inclusión:

- $\pi_k > 0 \;\; \forall \text{k} \in U$
- $\pi_{kl} > 0 \;\; \forall k; l \in U$

Se observan los valores de la variable y_k solamente para un subconjunto de la muestra, $r \subset s$, de tamaño m_r y por lo tanto el estimador $\hat{t}_\pi$ será sesgado.

Objetivo

El objetivo es lograr estimadores que sean resistentes al sesgo y con una varianza reducida.

Un modelo de respuesta inocente

$$
\begin{array}{ll}
P(k \in r|s) = \theta_k = \theta & \forall k \in s \\
P(k; l \in r|s) = \theta_k \theta_l = \theta^2 & \forall k; l \in s
\end{array}
$$

Si hubiera respuesta completa, usaríamos el estimador de ratio:

$$\hat{t} = \frac{N}{n} \sum\nolimits_s y_k = N \frac{\sum_s y_k}{\sum_s 1} = N \bar{y}_s = N \frac{\sum_s \frac{y_k}{\pi_k}}{\sum_s \frac{1}{\pi_k}}$$

Dada la no respuesta, sumamos sobre el subconjunto de respuesta r y ajustamos los pesos:

$$\hat{t}_1 = N \frac{\sum_r \frac{y_k}{\pi_k \theta_k}}{\sum_r \frac{1}{\pi_k \theta_k}} = N \frac{\sum_r \frac{y_k}{\pi_k \theta}}{\sum_r \frac{1}{\pi_k \theta}} = N \frac{\sum_r \frac{y_k}{\pi_k}}{\sum_r \frac{1}{\pi_k}}$$

Lo anterior es equivalente a no hacer nada respecto a la no respuesta.

El sesgo del estimador

Para calcular el sesgo del estimador anterior se deben tener en cuenta los siguientes 3 casos:

Caso N°1: La RD es verdadera

- El modelo de respuesta inocente es una perfecta descripción de la verdadera distribución de respuestas (RD).
- El estimador $\hat{t}_1$ es aproximadamente insesgado, y su sesgo despreciable es debido a que es un estimador de ratio y no a la no respuesta.

Caso N°2: El modelo es falso

- El modelo anterior no es correcto y las probabilidades de respuesta son independientes pero varían individuo a individuo:
  - $P(k \in r|s) = \theta_k$
  - $P(k; l \in r|s) = \theta_k \theta_l$

El sesgo en el caso Nº 2

Sesgo

$$B(\hat{t}_1) = E(\hat{t}_1) - t \doteq N\frac{\sum_U y_k\theta_k}{\sum_U \theta_k} - t = \frac{\sum_U y_k\theta_k}{\bar{\theta}_U} - t = (N-1)\frac{S_{y\theta_U}}{\bar{\theta}_U} =$$

$$= \frac{(N-1)}{\bar{\theta}_U} R_{y\theta_U} S_{\theta_U} S_{y_U} = \frac{t}{N}(N-1) R_{y\theta_U} cv_{\theta_U} cv_{y_U}$$

Sesgo relativo

$$RB(\hat{t}_1) = \frac{B(\hat{t}_1)}{t} \doteq R_{y\theta_U} cv_{y_U} cv(\theta_U)$$

Por lo tanto, cuanto mayor sea la correlación entre la variable de interés y y la probabilidad de no respuesta θ, mayor será el sesgo relativo.

Caso N°3: Comportamiento de respuesta determinístico

- El verdadero comportamiento de respuesta es determinístico, con un estrato de respuesta U_1 y uno de no respuesta U_2 tales que:
  - los elementos k $\in U_1$ responden con probabilidad 1.
  - los elementos k $\in U_2$ responden con probabilidad 0.

Sesgo

$$B(\hat{t}_1) \doteq N_2(\bar{y}_{U_1} - \bar{y}_{U_2})$$

Por lo tanto, el sesgo crece con el tamaño del estrato de no respuesta (N_2) y la diferencia de medias entre los estratos.

Grupos de respuesta homogénea (RHG)

1. La muestra s es particionada en H_s grupos de tamaño n_h, de forma tal que:

$$s = \bigcup_{h=1}^{H_s} s_h$$

2. Se denomina r_h de tamaño m_h al subconjunto de respuesta dentro del grupo s_h, por lo tanto:

$$r = \bigcup_{h=1}^{H_s} r_h \quad \text{y} \quad m = \sum_{h=1}^{H_s} m_h$$

3. Se asume que, dado s, todos los individuos del mismo grupo presentan la misma probabilidad de no respuesta.
4. H_s varía de muestra en muestra.
5. La asignación del elemento k varía de muestra en muestra.

Probabilidades de inclusión

Probabilidades de inclusión al subconjunto de respuesta condicionales a la muestra s

- $P(k \in r|s) = \pi_{k|s} = \theta_{hs} > 0 \qquad \forall k \in s_h$
- $P(k; l \in r|s) = \pi_{k;l|s} = P(k \in r|s)P(l \in r|s) > 0 \qquad \forall k \neq l \in s$

Por lo tanto, dado s, si el modelo ajusta correctamente los datos entonces el set de respuesta se distribuye de acuerdo a un diseño STBE.

Probabilidades de inclusión condicionales a s y $\mathbf{m}$

- $P(k \in r|s; \mathbf{m}) = \pi_{k|s;\mathbf{m}} = \frac{m_h}{n_h} = f_h \quad \forall k \in s_h$
- $P(k; l \in r|s; \mathbf{m}) = \pi_{kl|s;\mathbf{m}} = \begin{cases} \frac{m_h}{n_h}\frac{(m_h-1)}{(n_h-1)} & \forall k; l \in s_h \\ \\ \frac{m_h}{n_h}\frac{m_{h'}}{n_{h'}} & k \in s_h;\ l \in s_{h'};\ h \neq h' \end{cases}$

Por lo tanto, dados s y $\mathbf{m}$, si el modelo ajusta correctamente los datos entonces el set de respuesta se distribuye de acuerdo a un diseño STSI.

Surgirán dos posibles estrategias de estimación:

1. Estimadores que solo usan los pesos.
2. Estimadores que usan los pesos y variables auxiliares.

Estimadores que usan solamente pesos

Pesos ajustados

Definimos los pesos ajustados como:

$$\frac{1}{\pi^*_k} = \frac{1}{\pi_k \, \pi_{k|s,m}} \quad \text{donde } \frac{1}{\pi_{k|s;\mathbf{m}}} \text{ es el ajuste por no respuesta.}$$

Estimador con pesos ajustados

$$\hat{t}_{c\pi^*} = \sum_r \frac{y_k}{\pi^*_k} = \sum_r \frac{\check{y}_k}{\pi_{k|s,m}} = \sum_{h=1}^{H_s} \sum_{r_h} \frac{\check{y}_k}{\frac{m_h}{n_h}} = \sum_{h=1}^{H_s} f_h^{-1} \sum_{r_h} \check{y}_k$$

$$\text{donde} \qquad f_h = \frac{m_h}{n_h}$$

Sesgo del estimador $\hat{t}_{c\pi^*}$

Para poder derivar el sesgo de $\hat{t}_{c\pi^*}$, primero estudiaremos su esperanza condicional a la muestra s:

$$E_{RD}(\hat{t}_{c\pi^*}|s) = E_m[E_{RD}(\hat{t}_{c\pi^*}|s;\mathbf{m})] = E_m\left[E_{RD}\left(\sum_{h=1}^{H_s} f_h^{-1}\sum\nolimits_{r_h}\check{y}_k \middle| s;\mathbf{m}\right)\right] =$$

$$= E_m\left(\sum_{h=1}^{H_s}\sum\nolimits_{s_h}\check{y}_k|s\right) = \sum\nolimits_{s}\check{y}_k = \hat{t}_\pi$$

Esto implica que, dada la muestra s, si el modelo ajusta correctamente, el estimador con pesos ajustados es, en promedio, igual a al que se hubiera obtenido de haber existido respuesta completa.

Observación

Para que el estimador con pesos ajustados sea calculable, debe ocurrir que la probabilidad del evento $\bar{A}_1 = \{m_h = 0 \text{ para algún } h = 1;\ldots;h;\ldots;H_s\}$ sea despreciable.

Sesgo del estimador $\hat{t}_{c\pi^*}$

Luego, dado que la esperanza condicional en s y $\mathbf{m}$ es el estimador π, la esperanza incondicional será:

$$E(\hat{t}_{c\pi^*}) = E_p \, E_{RD} \, (\hat{t}_{c\pi^*} \mid s) = E_p \, (\hat{t}_\pi) = t$$

Por lo tanto, el estimador con pesos ajustados a la no respuesta es insesgado para el total de y en U si el modelo RHG ajusta y si $P(\bar{A}_1)$ es despreciable.

La varianza del estimador $\hat{t}_{c\pi^*}$

Se considera otro evento $\bar{A}_2 = \{m_H \leq 1 \text{ para algún } h\}$

$$V(\hat{t}_{c\pi^*}) \;=\; V_p\, E_m\, E_s(\hat{t}_{c\pi^*} \,|\, s\,) \,+\, E_p\, V_m\, E_s(\hat{t}_{c\pi^*} \,|\, s\,) \,+\, E_p\, E_m\, V_s(\hat{t}_{c\pi^*} \,|\, s\,) \Rightarrow$$

$$\Rightarrow V(\hat{t}_{c\pi^*}) \;=\; \underbrace{\sum\sum\nolimits_U \Delta_{kl}\, \check{y}_k\, \check{y}_l}_{V(\hat{t}_\pi)} + \underbrace{E_p\, E_m \left(\sum_{h=1}^{H_s} \frac{n_h^2}{(1-f_h)}\, S^2_{\check{y}_{sh}} \,\middle|\, s \right)}_{\text{incremento por no respuesta}}$$

donde:

- $S^2_{\check{y}_{sh}}$ es la varianza de $\check{y}$.
- $E_p(.)$ es la esperanza respecto al diseño.
- $E_m(.|s)$ es la esperanza respecto a la distribución de $\mathbf{m}$, dada s.

Estimación de la varianza

$$\hat{V}(\hat{t}_{c\pi^*}) = \sum\sum\nolimits_r \frac{\check{\Delta}_{kl}}{\pi_{kl|s,m}}\, \check{y}_k\, \check{y}_l \,+\, \sum_{h=1}^{H_s} \frac{n_h^2}{m_h}\, (1-f_h)\, S^2_{y_{rh}}$$

Ejemplo: weighting class estimator

Supongamos el caso en que una muestra s de tamaño n es tomada de una población U mediante un diseño SI.

⋆ $\hat{t}_{c\pi^*} = \dfrac{N}{n} \displaystyle\sum_{h=1}^{H_s} n_h \, \bar{y}_{r_h} = N \, \hat{\bar{y}}_U$ conocido como *weighting class estimator*

⋆ $V(\hat{t}_{c\pi^*}) = \dfrac{N^2}{n}(1 - f)S^2_{y_U} + \dfrac{N^2}{n^2} \, E_p \left[E_m \left(\displaystyle\sum_{h=1}^{H_s} \dfrac{N_h^2}{m_h}(1 - f_h)S^2_{y_{s_h}} \middle| \, s \right) \right] = V_1 + V_2$

Nótese que el primer sumando corresponde a la varianza del estimador π bajo un diseño simple. El segundo sumando corresponde al incremento de varianza generado por la no respuesta.

Un estimador insesgado para la varianza viene dado por:

$$\star \quad \hat{V}(\hat{t}_{c\pi^*}) = \hat{V}_1 + \hat{V}_2$$

$$\star \quad \hat{V}_1 = \frac{N^2}{n}(1-f)\left[\sum_{h=1}^{H_s} \frac{n_h}{n}(1-\delta_h)S^2_{y_{r_h}} + \frac{n}{n-1}\sum_{h=1}^{H_s}\frac{n_h}{n}\left(\bar{y}_{r_h} - \hat{\bar{y}}_U\right)^2\right]$$

$$\text{donde} \quad \delta_h = \left(\frac{1 - {}^{n_h}\!/_{n}}{m_h}\right)\left(\frac{n}{n-1}\right)$$

$$\star \quad \hat{V}_2 = N^2 \sum_{h=1}^{H_s} \frac{n_h}{n}\left(\frac{1-f_h}{m_h}\right) S^2_{y_{r_h}}$$

Estimadores que usan pesos y variables auxiliares

- La intuición detrás de estos es utilizar estimadores de regresión con el objetivo de asistir la estimación mediante el uso de la información auxiliar.
- El uso de esta información auxiliar genera estimadores resistentes al sesgo y ayuda a disminuir la varianza.
- Se utilizarán 2 tipos de predicciones:
    - $\hat{y}_k = \mathbf{x}'_k \hat{\mathbf{B}}_\mathbf{r}$
    - $\hat{y}_{1k} = \mathbf{x}'_{1k} \hat{\mathbf{B}}_\mathbf{1r}$

donde:

- x_k es un vector de información auxiliar a nivel de la muestra.
- x_{1k} es un vector de información auxiliar a nivel poblacional.

El uso de estos estimadores requiere conocer:

- $\sum_{s_h} x_k \quad \forall h$
- $\sum_U x_{1k}$
- $\sum_{s_h} x_{1k} \quad \forall h$
- Los valores individuales $x_{1k} \, \forall k \in U$
- Los valores individuales $x_k \, \forall k \in r$

Coeficientes estimados $\hat{\mathbf{B}}$

$$\star \quad \hat{\mathbf{B}}_{\mathbf{r}} = \left(\sum_{h=1}^{H_s}\sum\nolimits_{r_h} \frac{\mathbf{x}_k\,\mathbf{x}_k'}{\sigma_k^2\,\pi_k^*}\right)^{-1}\left(\sum_{h=1}^{H_s}\sum\nolimits_{r_h} \frac{\mathbf{x}_k\,y_k}{\sigma_k^2\,\pi_k^*}\right) =$$

$$= \left(\sum_{h=1}^{H_s} f_h^{-1}\sum\nolimits_{r_h} \frac{\mathbf{x}_k\,\mathbf{x}_k'}{\sigma_k^2\,\pi_k}\right)^{-1}\left(\sum_{h=1}^{H_s} f_h^{-1}\sum\nolimits_{r_h} \frac{\mathbf{x}_k\,y_k}{\sigma_k^2\,\pi_k}\right)$$

$$\star \quad \hat{\mathbf{B}}_{\mathbf{1r}} = \left(\sum_{h=1}^{H_s}\sum\nolimits_{r_h} \frac{\mathbf{x}_{1k}\,\mathbf{x}_{1k}'}{\sigma_{1k}^2\,\pi_k^*}\right)^{-1}\left(\sum_{h=1}^{H_s}\sum\nolimits_{r_h} \frac{\mathbf{x}_{1k}\,y_k}{\sigma_{1k}^2\,\pi_k^*}\right) =$$

$$= \left(\sum_{h=1}^{H_s} f_h^{-1}\sum\nolimits_{r_h} \frac{\mathbf{x}_{1k}\,\mathbf{x}_{1k}'}{\sigma_{1k}^2\,\pi_k}\right)^{-1}\left(\sum_{h=1}^{H_s} f_h^{-1}\sum\nolimits_{r_h} \frac{\mathbf{x}_{1k}\,y_k}{\sigma_{1k}^2\pi_k}\right)$$

El estimador de regresión

$$\hat{t}_{cr} = \sum\nolimits_U \hat{y}_{1k} + \sum_{h=1}^{H_s} \left(\sum\nolimits_{s_h} \frac{\hat{y}_k - \hat{y}_{1k}}{\pi_k} + \sum\nolimits_{r_h} \frac{y_k - \hat{y}_k}{\pi_k^*} \right) =$$

$$= \sum\nolimits_U \hat{y}_{1k} + \sum_{h=1}^{H_s} \left(\sum\nolimits_{s_h} \frac{\hat{y}_k - \hat{y}_{1k}}{\pi_k} + f_h^{-1} \sum\nolimits_{r_h} \frac{y_k - \hat{y}_k}{\pi_k} \right)$$

2 casos particulares:

1. Información auxiliar solamente a nivel de muestra:

$$\hat{t}_{cr} = \sum_{h=1}^{H_s} \left(\sum\nolimits_{s_h} \frac{\hat{y}_k}{\pi_k} + f_h^{-1} \sum\nolimits_{r_h} \frac{y_k - \hat{y}_k}{\pi_k} \right)$$

2. Información auxiliar solamente a nivel poblacional:

$$\hat{t}_{cr} = \sum\nolimits_U \hat{y}_{1k} + \sum_{h=1}^{H_s} f_h^{-1} \sum\nolimits_{r_h} \frac{y_k - \hat{y}_{1k}}{\pi_k}$$

Varianza y estimación de la varianza

Errores y residuos π-expandidos

- $\check{E}_k = \frac{E_k}{\pi_k} = \frac{y_k - \mathbf{x}_k' \mathbf{B_s}}{\pi_k}$ con $\mathbf{B_s} = \left(\sum_s \frac{\mathbf{x}_k \mathbf{x}_k'}{\sigma_k^2 \pi_k} \right)^{-1} \left(\sum_s \frac{\mathbf{x}_k y_k}{\sigma_k^2 \pi_k} \right)$
- $\check{E}_{1k} = \frac{E_{1k}}{\pi_k} = \frac{y_k - \mathbf{x}_{1k}' \mathbf{B_1}}{\pi_k}$ con $\mathbf{B_1} = \left(\sum_U \frac{\mathbf{x}_{1k} \mathbf{x}_{1k}'}{\sigma_{1k}^2} \right)^{-1} \left(\sum_U \frac{\mathbf{x}_{1k} y_k}{\sigma_{1k}^2} \right)$
- $\check{e}_{kr} = \frac{e_{kr}}{\pi_k} = \frac{y_k - \hat{y}_k}{\pi_k}$
- $\check{e}_{1kr} = \frac{e_{1kr}}{\pi_k} = \frac{y_k - \hat{y}_{1k}}{\pi_k}$

Si el modelo ajusta correctamente los datos $\Rightarrow$ el estimador de regresión presentado es aproximadamente insesgado para el total t_y.

Varianza Aproximada

$$AV(\hat{t}_{cr}) = \sum\sum_U \Delta_{kl}\, \check{E}_{1k}\, \check{E}_{1l} + E_p\left[E_m\left(\sum_{h=1}^{H_s} \frac{n_h^2}{m_h}(1-f_h)S^2_{\check{E}_{s_h}} \middle| s\right)\right]$$

donde $S^2_{\check{E}_{s_h}}$ es la varianza de $\check{E}_k$ en el set s_h

Un estimador de la varianza

$$\hat{V}(\hat{t}_{cr}) = \sum\sum_r \frac{\check{\Delta}_{kl}}{\pi_{kl|s;\mathbf{m}}}\, \check{e}_{1k_r}\, \check{e}_{1l_r} + \sum_{h=1}^{H_s} \frac{n_h^2}{m_h}(1-f_h)\, S^2_{\check{e}_{r_h}}$$

donde $S^2_{\check{e}_{r_h}}$ es la varianza de $\check{e}_{r_h}$ sobre r_h

Cada uno de los sumandos del estimador es insesgado para su contraparte en la varianza $\Rightarrow E[\hat{V}(\hat{t}_{cr})] = V(\hat{t}_{cr})$.

Ejemplo: ratio estimator with weighting class adjustment

- Se toma una muestra s de tamaño n bajo un diseño SI.
- x_k valores positivos solamente conocidos en la muestra $s \Rightarrow$ caso especial 1.

Supongamos que el scatter de los puntos $(x_k; y_k)$ queda bien descrito por el modelo:

$$\begin{cases} E_\xi(y_k) & = & \mathbf{x}_k' \boldsymbol{\beta} \\ V_\xi(y_k) & = & \sigma^2 \mathbf{x}_k \end{cases}$$

Estimador de regresión con pesos ajustados

$$\hat{t}_{cr} = \frac{N}{n} \left(\sum\nolimits_s x_k \right) \hat{B}_r = \frac{N}{n} \left(\sum\nolimits_s x_k \right) \frac{\sum_{h=1}^{H_s} n_h \, \bar{y}_{r_h}}{\sum_{h=1}^{H_s} n_h \, \bar{x}_{r_h}}$$

Su varianza aproximada

$$AV(\hat{t}_{cr}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)S^2_{y_U} + \frac{N^2}{n^2}E_p\left[E_m\left(\sum_{h=1}^{H_s}\frac{n_h^2}{m_h}(1-f_h)S^2_{E_{s_h}}\middle| s\right)\right] = V_1 + AV_2$$

donde:

- $S^2_{E_{s_h}}$ es la varianza de los residuos $E_K = y_k - \mathbf{x}_k\,\mathbf{B_s}$ en s_h
- $\mathbf{B_s} = \frac{\sum_s y_k}{\sum_s x_k}$

Un estimador insesgado para la varianza viene dado por:

$$\star \quad \hat{V}(\hat{t}_{c\pi^*}) = \hat{V}_1 + \hat{V}_2$$

$$\star \quad \hat{V}_1 = \frac{N^2}{n}(1-f)\left[\sum_{h=1}^{H_s} \frac{n_h}{n}(1-\delta_h) S^2_{y_{r_h}} + \frac{n}{n-1}\sum_{h=1}^{H_s}\frac{n_h}{n}\left(\bar{y}_{r_h} - \hat{\bar{y}}_U\right)^2\right]$$

$$\text{donde} \quad \delta_h = \left(\frac{1 - {}^{n_h}/_{n}}{m_h}\right)\left(\frac{n}{n-1}\right)$$

$$\star \quad \hat{V}_2 = \sum_{h=1}^{H_s} \frac{n_h}{m_h}\left(1 - f_h\right) S^2_{\check{e}_{r_h}}$$

$$\text{donde} \quad e_{k_r} = y_k - \mathbf{x}_k \, \hat{\mathbf{B}}_{\mathbf{r}}$$

Ejemplo: post-estratificación (caso 1)

- Se toma una muestra s bajo un diseño SI de tamaño n, y luego se post-estratifica.
- Modelo de la media común por grupos (o estratos) para los datos:

$$\begin{cases} E_\xi(y_k) = \beta_h & \forall k \in s_h \\ V_\xi(y_k) = \sigma_h^2 & \forall k \in s_h \end{cases}$$

- El vector x_{1k} indica a qué grupo pertenece el elemento k.
- Se asume que el vector $\sum_U \mathbf{x}_{1k} = (N_1; \ldots; N_h; \ldots; N_H)$ es conocido.
- Se asume un modelo RHG para la respuesta.
- Los estratos formados son equivalentes a los grupos de respuesta homogénea (RGH).

Estimador poset-stratificado

$$\hat{t}_{cr} = \sum_{h=1}^{H_s} N_h \, \bar{y}_{rh}$$

Ejemplo: post-estratificación (caso 2)

Modelo para los datos

- Los elementos de la muestra s se clasifican en G_s grupos (estratos): $(s_1; \ldots; s_h; \ldots; s_{G_s})$ de tamaño n_g de forma tal que los valores de y_k dentro de cada grupo tengan una variación modesta alrededor de la media grupal.
- Conocemos los totales por estrato solo a nivel de muestra.
- Modelo para los datos: one-way ANOVA

$$\begin{cases} E_\xi(y_k) = \beta_g & \forall k \in g \\ V_\xi(y_k) = \sigma^2 & \forall k \in g \end{cases}$$

- La muestra es obtenida mediante un diseño SI.

Modelo para la no respuesta

- Se asume un modelo RHG para la no respuesta con H_s categorías.
- La clasificación cruzada entre estratos y grupos genera $G_s \times H_s$ categorías de clasificación, s_{gh}, de tamaño n_{gh}.
- r_{gh} es el set de respuesta del grupo s_{gh} de tamaño m_{gh}.
- La tasa de respuesta en el grupo h es $f_h = \frac{m_{\cdot h}}{n_{\cdot h}}$ donde:
  - $n_{\cdot h} = \sum_{g=1}^{G_s} n_{gh}$
  - $m_{\cdot h} = \sum_{g=1}^{G_s} m_{gh}$

Estimador del total: $\hat{t}_{cr} = \frac{N}{n} \sum_{g=1}^{G_s} n_{g.}\hat{B}_{gr}$

donde:

- $n_{g.} = \sum_{h=1}^{H_s} n_{gh}$
- $\hat{B}_{gr} = \left(\sum_{h=1}^{H_s} f_h^{-1} m_{gh} \right)^{-1} \left(\sum_{h=1}^{H_s} f_h^{-1} \sum_{r_{gh}} y_k \right)$

El mismo puede interpretarse como la suma de los totales por estrato $\hat{N}_g \times \hat{B}_{g_r}$ donde:

- $\hat{B}_{g_r}$ es el ajuste por la no respuesta estimada para la media del estrato g.
- $\hat{N}_g = {}^{N\,n_{g.}}/_{n}$ es el conteo estimado para el estrato g.

Estimador de la varianza del total: $\hat{V}(\hat{t}_{cr}) = \hat{V}_1 + \hat{V}_2$

donde $\hat{V}_1$ y $\hat{V}_2$ son los mismos que en caso anterior con residuos $e_{k_r} = y_k - \hat{B}_{g_r}$

Imputación

Supongamos que tenemos un estudio con q variables de análisis:
$\mathbf{y}_k = (y_{1k}; \ldots; y_{jk}; \ldots; y_{qk})'$ donde:

- r_j es el set de respuesta para la variable j.
- $r_u = r_1 \cup r_2 \cup \ldots \cup r_q$ es el set de los elementos que responden una o más preguntas.
- $r_c = r_1 \cap r_2 \cap \ldots \cap r_q$ es el set de los elementos que responden todas las preguntas.
- item non-response set: $r_u - r_c$ (se asume no vacío).
- unit non-response set: $s - r_u$ (se asume no vacío).

Dos opciones en cuanto a cómo utilizar la información observada y la información auxiliar:

1. **Response set approach**: la información asociada con el set de respuesta de la variable j es usada para crear estimaciones para la variable j.
2. **Clean data matrix approach**: se crea una matriz completa, la cual es utilizada para calcular estimaciones para los valores faltantes.

Response set approach

- Puede utilizarse el enfoque de ajustes ponderados visto anteriormente variable-a-variable.
- Se define un set de RHGs.
- $\check{y}_{jk}$ recibe el ajuste ${}^{n_h}/_{m_{jh}}$ si $k \in h$, donde ${}^{n_h}/_{m_{jh}}$ es la tasa de respuesta en el grupo h para el item j.
- Los RHGs pueden diferir entre items del cuestionario.
- Pueden generar estimaciones no permitidas (por ejemplo: valores negativos para variables que el investigador sabe son siempre positivas).

Clean data matrix approach

- Forma inocente: utilizar únicamente los datos observados y_k para $k \in r_c$
  - La información para las observaciones en el set $r_u - r_c$ es descartada.
  - El método solo funciona si el tamaño del set descartado es muy reducido.
  - Se utilizan métodos de imputación para crear la matriz de datos completos.
  - Los valores imputados los anotamos como: $\tilde{y}_{jk}$, los cuales son generados mediante el uso de información auxiliar.
  - Esto conlleva a una matriz completa de datos de dimensiones $n_{r_u} \times q$.
- Imputaciones para la no respuesta de unidades y la no respuesta de items:
  - Se producen estimaciones $\tilde{y}_{jk}$ para todo el set $s - r_c$.
  - El resultado es una matriz de datos de dimensiones $n_s \times q$.
- La imputación siempre produce sesgos y varianzas adicionales en las estimaciones.
- Conocemos como *imputación deductiva* a las instancias en las que un valor faltante puede ser imputado de forma perfecta $(\tilde{y}_{jk} = y_{jk})$ producto de una conclusión lógica.

Métodos de predicción imperfecta

Overall mean imputation

- Para cada item j, se asigna el mismo valor $\bar{y}_{r_j}$ a todos los valores faltantes y_{jk} en el set $r_u - r_j$.
- Puede producir estimaciones de varianza pobres.

Class mean imputation

- El set de respuesta es particionado en clases según un algoritmo de clasificación para el cual se utiliza la información auxiliar.
- Los valores faltantes son imputados con la media de la clase a la que pertenece el elemento.

Hot-Deck and Cold-Deck imputation

- **Hot-Deck**: los valores faltantes son remplazados por valores seleccionados de entre las observaciones de la encuesta.
- **Cold-Deck**: utiliza valores de otras fuentes.

Random overall imputation

- Para cada valor faltante se sortea un valor en el set de respuesta r_j.
- Este se conoce como donante.

Random imputation with classes

- Ídem que el anterior, pero los donantes son sorteados dentro de la misma clase a la que pertenece la unidad a ser imputada.

Sequential Hot-Deck

- Los donantes son seleccionados mediante "backtracking" dentro de la clase de la unidad a imputar.
- Se elige el donante "más cercano" según un criterio establecido.
- El procedimiento siempre comienza con un valor "cold-deck" para cada clase.
- Un problema de este método es que algunos donantes puede terminar siendo usados varias veces.

Distance function matching

- Para cada valor faltante y para cada item, y_{jk} es remplazado por el valor contestado por un elemento presente en la encuesta para dicho item.
- El donante es elegido mediante cercanía según alguna función de distancia, definida sobre las variables auxiliares.

Regression imputation

- Utiliza la información de los respondentes para ajustar una regresión para la variable que se desea imputar.
- Para dicha regresión se utilizan variables que se asume tienen alto poder predictivo para y_j.

Multiple imputation

- Para cada valor faltante se realizan m imputaciones.
- Se forman m data sets completos a ser analizados.
- Se utilizan pooled-variance para construir intervalos de confianza.