El estimador de regresión para muestreo en dos etapas

Daniel Czarnievicz

2017

Consideremos una situación en la que la población se encuentra particionada en N_I clusters, $U = \left\{U_1; \dots; U_i; \dots; U_{N_I}\right\}$, de forma tal que $\#U_i = N_i$, $U = \bigcup\limits_{i=1}^{N_I} U_i$, y $N = \sum\limits_{i=1}^{N_I} N_i$. En la primera etapa se obtiene una muestra de PSUs, s_I, de acuerdo al diseño $p_I(.)$ con probabilidades de inclusión π_{I_i} y $\pi_{I_{ij}}$. Dadas las PSU, en la segunda etapa se obtiene una muestra dentro de cada PSU, s_i, de acuerdo con el diseño $p_i(.|s_I)$, con probabilidades de inclusión $\pi_{k|i}$ y $\pi_{kl|i}$. En esta etapa se asume independencia e invarainza. La muestra queda conformada por $s = \bigcup\limits_{i \in s_I} s_i$ de tamaño $n_S = \sum\limits_{i \in s_I} n_i$. Para cada elemento en la muestra se observa la variable de interés y_k.

La información auxiliar referente a las unidades se anotará como $\mathbf{x}_k$. La información auxiliar referente a las PSU se anotará como $\mathbf{u}_i$. Quedan definidas entonces las siguientes matrices:

$$\mathbf{U}_{N_I \times J} = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1\nu} & \cdots & u_{1J} \\ u_{21} & u_{22} & \cdots & u_{2\nu} & \cdots & u_{2J} \\ \vdots & \vdots & & \vdots & & \vdots \\ u_{i1} & u_{i2} & \cdots & u_{i\nu} & \cdots & u_{iJ} \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ u_{N_I 1} & u_{N_I 2} & \cdots & u_{N_I \nu} & \cdots & u_{N_I J} \end{bmatrix} = \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_i \\ \vdots \\ \mathbf{u}_{N_I} \end{bmatrix}$$

$$\mathbf{X}_{N \times J} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1\nu} & \cdots & x_{1J} \\ x_{21} & x_{22} & \cdots & x_{2\nu} & \cdots & x_{2J} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{k1} & x_{k2} & \cdots & x_{k\nu} & \cdots & x_{kJ} \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{N\nu} & \cdots & x_{NJ} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_k \\ \vdots \\ \mathbf{x}_N \end{bmatrix}$$

Se desprenden tres posibles casos:

- **Caso A**: $\mathbf{u}_i$ está disponible para todas las PSU.
- **Caso B**: $\mathbf{x}_k$ está disponible para todos los elementos.
- **Caso C**: $\mathbf{x}_k$ está disponible solo para los elementos en las PSU seleccionadas en la primera etapa.

Caso A

Supongamos que el scatter de los N_I puntos $(t_{y_i}; \mathbf{u}_i)$ puede describirse mediante el siguiente modelo:

$$E_\xi(y_k) = \mathbf{u}_i' \boldsymbol{\beta}_I \qquad V_\xi(y_k) = \sigma^2_{I_i}$$

donde los totales por cluster, t_{y_i}, son independientes bajo el modelo especificado. $\boldsymbol{\beta}_I$ puede ser estimado por:

- en un censo: $\mathbf{B}_I = \left(\sum_{U_I} \dfrac{\mathbf{u}_i \mathbf{u}_i'}{\sigma^2_{I_i}} \right)^{-1} \left(\sum_{U_I} \dfrac{\mathbf{u}_i t_{y_i}}{\sigma^2_{I_i}} \right)$

- en una muestra: $\hat{\mathbf{B}}_I = \left(\sum_{s_I} \frac{\mathbf{u}_i\mathbf{u}_i'}{\sigma^2_{I_i}\pi_{I_i}}\right)^{-1}\left(\sum_{s_I} \frac{\mathbf{u}_i t^*_{y_i}}{\sigma^2_{I_i}\pi_{I_i}}\right)$

donde $t^*_{y_i} = \hat{t}_{y_{i\pi}} = \sum_{s_i} \frac{y_k}{\pi_{k|i}}$. Se definen los valores ajustados $t^0_{y_i} = \mathbf{u}_i'\mathbf{B}_I$, las predicciones $\hat{t}_{y_{ip}} = \mathbf{u}_i'\hat{\mathbf{B}}_I$, y ambos residuos $D_i = t_{y_i} - t^0_{y_i}$ $\forall i \in U_I$ (inobservables), $d_i = t^*_{y_i} - \hat{t}_{y_{ip}}$ $\forall i \in s_I$ (observables).

Se define el estimador de regresión como:

$$\hat{t}_{y_{rA}} = \sum_{U_I} \hat{t}_{y_{ip}} + \sum_{s_I} \frac{t^*_{y_i} - \hat{t}_{y_{ip}}}{\pi_{I_i}} = \sum_{U_I} \hat{t}_{y_{ip}} + \sum_{s_I} \frac{t^*_{y_i}}{\pi_{I_i}} - \sum_{s_I} \frac{\hat{t}_{y_{ip}}}{\pi_{I_i}} =$$

$$= \sum_{U_I} \hat{t}_{y_{ip}} + \sum_{s_I} \frac{1}{\pi_{I_i}} \sum_{s_i} \frac{y_k}{\pi_{k|i}} - \sum_{s_I} \frac{\hat{t}_{y_{ip}}}{\pi_{I_i}} =$$

$$= \sum_{U_I} \hat{t}_{y_{ip}} + \sum_{s_I} \sum_{s_i} \frac{y_k}{\pi_{k|i}\pi_{I_i}} - \sum_{s_I} \frac{\hat{t}_{y_{ip}}}{\pi_{I_i}} =$$

$$= \sum_{U_I} \hat{t}_{y_{ip}} + \sum_{s} \frac{y_k}{\pi_k} - \sum_{s_I} \frac{\hat{t}_{y_{ip}}}{\pi_{I_i}} = \sum_{U_I} \hat{t}_{y_{ip}} + \sum_{s} y_k^{\checkmark} - \sum_{s_I} \frac{\hat{t}_{y_{ip}}}{\pi_{I_i}} =$$

$$= \hat{t}_{y_\pi} + \sum_{U_I} \hat{t}_{y_{ip}} - \sum_{s_I} \frac{\hat{t}_{y_{ip}}}{\pi_{I_i}} = \hat{t}_{y_\pi} + \sum_{U_I} \mathbf{u}_i'\hat{\mathbf{B}}_I - \sum_{s_I} \frac{\mathbf{u}_i'\hat{\mathbf{B}}_I}{\pi_{I_i}} =$$

$$= \hat{t}_{y_\pi} + \left(\sum_{U_I} \mathbf{u}_i - \sum_{s_I} \frac{\mathbf{u}_i}{\pi_{I_i}}\right)' \hat{\mathbf{B}}_I$$

por lo tanto, el estimador de regresión no es más que el estimador π, más un término de ajuste, determinado por la información auxiliar sobre las PSU. El estimador también puede escribirse como un estimador-π g-ponderado, partiendo de la expresión anterior:

$$\hat{t}_{y_{rA}} = \sum_{s_I} \frac{t^*_{y_i}}{\pi_{I_i}} + \left(\sum_{U_I} \mathbf{u}_i - \sum_{s_I} \frac{\mathbf{u}_i}{\pi_{I_i}}\right)' \left(\sum_{s_I} \frac{\mathbf{u}_i\mathbf{u}_i'}{\sigma^2_{I_i}\pi_{I_i}}\right)^{-1} \left(\sum_{s_I} \frac{\mathbf{u}_i t^*_{y_i}}{\sigma^2_{I_i}\pi_{I_i}}\right) =$$

$$= \sum_{s_I} \frac{t^*_{y_i}}{\pi_{I_i}} \underbrace{\left[1 + \left(\sum_{U_I} \mathbf{u}_i - \sum_{s_I} \frac{\mathbf{u}_i}{\pi_{I_i}}\right)' \left(\sum_{s_I} \frac{\mathbf{u}_i\mathbf{u}_i'}{\sigma^2_{I_i}\pi_{I_i}}\right)^{-1} \frac{\mathbf{u}_i}{\sigma^2_{I_i}}\right]}_{g_{i s_I}} = \sum_{s_I} \frac{t^*_{y_i} g_{i s_I}}{\pi_{I_i}}$$

Para hallar el estimador de la varianza primero conviene expresar el estimador de regresión en términos del error de estimación:

$$\hat{t}_{y_{rA}} - t_y = \sum_{s_I} \frac{g_{i s_I} t^*_{y_i}}{\pi_{I_i}} - t_y \Rightarrow$$

$$\Rightarrow \hat{t}_{y_{rA}} - t_y = \sum_{s_I} \frac{g_{i s_I} t^*_{y_i}}{\pi_{I_i}} - \sum_{U_I} t_{y_i} + \sum_{s_I} \frac{g_{i s_I} t_{y_i}}{\pi_{I_i}} - \sum_{s_I} \frac{g_{i s_I} t_{y_i}}{\pi_{I_i}} \Rightarrow$$

$$\Rightarrow \hat{t}_{y_{rA}} - t_y = \sum_{s_I} \frac{g_{i s_I}}{\pi_{I_i}} (t^*_{y_i} - t_{y_i}) + \sum_{s_I} \frac{g_{i s_I} t_{y_i}}{\pi_{I_i}} - \sum_{U_I} t_{y_i} \Rightarrow$$

$$\Rightarrow \hat{t}_{y_{rA}} - t_y = \sum_{s_I} \frac{g_{i s_I}}{\pi_{I_i}} (t^*_{y_i} - t_{y_i}) + \sum_{s_I} \frac{g_{i s_I} t_{y_i}}{\pi_{I_i}} - \sum_{U_I} (\mathbf{u}_i'\mathbf{B} + D_i) \Rightarrow$$

$$\Rightarrow \hat{t}_{y_{rA}} - t_y = \sum_{s_I} \frac{g_{i s_I}}{\pi_{I_i}} (t^*_{y_i} - t_{y_i}) + \sum_{s_I} \frac{g_{i s_I} t_{y_i}}{\pi_{I_i}} - \sum_{U_I} \mathbf{u}_i'\mathbf{B} - \sum_{U_I} D_i \Rightarrow$$

$$\Rightarrow \hat{t}_{y_{rA}} - t_y = \sum_{s_I} \frac{g_{i s_I}}{\pi_{I_i}} (t^*_{y_i} - t_{y_i}) + \sum_{s_I} \frac{g_{i s_I} (\mathbf{u}_i'\mathbf{B} + D_i)}{\pi_{I_i}} - \sum_{U_I} \mathbf{u}_i'\mathbf{B} - \sum_{U_I} D_i \Rightarrow$$

$$\Rightarrow \hat{t}_{y_{r_A}} - t_y = \sum_{s_I} \frac{g_{i_{s_I}}}{\pi_{I_i}} \left(t^*_{y_i} - t_{y_i}\right) + \sum_{s_I} \frac{g_{i_{s_I}} \mathbf{u}_i'\mathbf{B}}{\pi_{I_i}} + \sum_{s_I} \frac{g_{i_{s_I}} D_i}{\pi_{I_i}} - \sum_{U_I} \mathbf{u}_i'\mathbf{B} - \sum_{U_I} D_i \Rightarrow$$

$$\Rightarrow \hat{t}_{y_{r_A}} - t_y = \sum_{s_I} \frac{g_{i_{s_I}}}{\pi_{I_i}} \left(t^*_{y_i} - t_{y_i}\right) + \sum_{U_I} \mathbf{u}_i'\mathbf{B} + \sum_{s_I} \frac{g_{i_{s_I}} D_i}{\pi_{I_i}} - \sum_{U_I} \mathbf{u}_i'\mathbf{B} - \sum_{U_I} D_i \Rightarrow$$

$$\Rightarrow \hat{t}_{y_{r_A}} - t_y = \underbrace{\sum_{s_I} \frac{g_{i_{s_I}}}{\pi_{I_i}} \left(t^*_{y_i} - t_{y_i}\right)}_{R_{A_s}} + \underbrace{\sum_{s_I} \frac{g_{i_{s_I}} D_i}{\pi_{I_i}} - \sum_{U_I} D_i}_{Q_{A_{s_I}}} \Rightarrow$$

$$\Rightarrow \hat{t}_{y_{r_A}} - t_y = Q_{A_{s_I}} + R_{A_s}$$

Luego, la varianza viene dada por:

$$V\left(\hat{t}_{y_{r_A}} - t_y\right) = V\left(\hat{t}_{y_{r_A}}\right) =$$

$$= V_I\Big[\underbrace{E_{II}\left(Q_{A_{s_I}}|s_I\right)}_{Q_{A_{s_I}}}\Big] + E_I\Big[\underbrace{V_{II}\left(Q_{A_{s_I}}|s_I\right)}_{=0}\Big] + V_I\Big[\underbrace{E_{II}\left(R_{A_s}|s_I\right)}_{=0}\Big] + E_I\Big[V_{II}\left(R_{A_s}|s_I\right)\Big] =$$

$$= \underbrace{V_I\Big(Q_{A_{s_I}}\Big)}_{V_{APSU}} + \underbrace{E_I\Big[V_{II}\left(R_{A_s}|s_I\right)\Big]}_{V_{ASSU}} = V_I\left(\sum_{s_I} \frac{g_{i_{s_I}} D_i}{\pi_{I_i}}\right) + E_I\left(\sum_{s_I} \frac{g^2_{i_{s_I}} V_i}{\pi^2_{I_i}}\right)$$

$$\text{donde } V_i = \sum\sum_{U_I} \Delta_{kl}\, y^{\checkmark}_{k|i}\, y^{\checkmark}_{l|i} \text{ con } y^{\checkmark}_{k|i} = \frac{y_k}{\pi_{k|i}}$$

Un estimador de la varianza viene dado por:

$$\star\ \hat{V}\left(\hat{t}_{y_{r_A}}\right) = \hat{V}_{APSU} + \hat{V}_{ASSU}$$

$$\star\ \hat{V}_{APSU} = \sum\sum_{s_I} \Delta^{\checkmark}_{I_{ij}} \frac{g_{i_{s_I}} d_i}{\pi_{I_i}} \frac{g_{j_{s_I}} d_j}{\pi_{I_j}} - \sum_{s_I} \frac{1}{\pi_{I_i}} \left(\frac{1}{\pi_{I_i}} - 1\right) g^2_{i_{s_I}} \hat{V}_i$$

$$\text{donde } \ d_i = t^*_{y_i} - \mathbf{u}_i'\hat{\mathbf{B}}_I$$

$$\star\ \hat{V}_{ASSU} = \sum_{s_I} \frac{g^2_{i_{S_I}} \hat{V}_i}{\pi^2_{I_i}}$$

$$\star\ \hat{V}_i = \sum\sum_{s_i} \Delta^{\checkmark}_{kl|i} \frac{y_k}{\pi_{k|i}} \frac{y_l}{\pi_{l|i}}$$

$\hat{V}_{APSU}$ puede justificarse dado que:

$$\bullet\ E_{II}(d_i^2) = E_{II}\Big[\Big(t^*_{y_i} - \mathbf{u}_i'\hat{\mathbf{B}}_I\Big)\Big]^2 \doteq E_{II}\Big[\Big(t^*_{y_i} - \mathbf{u}_i'\mathbf{B}_I\Big)\Big]^2 =$$

$$= \Big[E_{II}\big(t^*_{y_i} - \mathbf{u}_i'\mathbf{B}_I\big)^2\Big] + V_{II}\Big(t^*_{y_i} - \mathbf{u}_i'\Big) = D_i^2 + V_i \quad \forall i = j$$

$$\bullet\ E_{II}(d_i\, d_j) \doteq D_i\, D_j \quad \forall i \neq j$$

$E_{II}(\hat{V}_i) = V_i$, por lo tanto, $E_{II}(\hat{V}_{APSU}) \doteq \sum\sum_{s_I} \Delta^{\checkmark}_{I_{ij}} \frac{g_{i_{s_I}} D_i}{\pi_{I_i}} \frac{g_{j_{s_I}} D_j}{\pi_{I_j}}$ de donde obtenemos que $E(\hat{V}_{APSU}) = E_I\Big[E_{II}\big(\hat{V}_{APSU}|s_I\big)\Big] \doteq AV_{APSU}$. Por otro lado,

$$E(\hat{V}_{ASSU}) = E_I\left[E_{II}\left(\sum_{s_I} \frac{g^2_{i_{s_I}} \hat{V}_i}{\pi^2_{I_i}} \,\middle|\, s_I\right)\right] = E_I\left(\sum_{s_I} \frac{g^2_{i_{s_I}} \hat{V}_i}{\pi^2_{I_i}}\right)$$

Por lo que $\hat{V}_{ASSU}$ es insesgada para V_{ASSU}. Entonces, $\hat{V}(\hat{t}_{y_{r_A}})$ es insesgada para $V(\hat{t}_{y_{r_A}})$.

Caso B

En el caso B la información auxiliar es conocida para todas las unidades del marco. Se asume que el scatter de puntos puede ser presentado por el modelo:

$$E_\xi(y_k) = \mathbf{x}_k'\boldsymbol{\beta} \qquad V_\xi(y_k) = \sigma_k^2$$

donde $\boldsymbol{\beta}$ es estimado por $\mathbf{B} = \left(\sum_U \frac{\mathbf{x}_k\,\mathbf{x}_k'}{\sigma_k^2}\right)^{-1}\left(\sum_U \frac{\mathbf{x}_k\,y_k}{\sigma_k^2}\right)$. Los residuos vienen dados por: $E_k = y_k - y_k^0 = y_k - \mathbf{x}_k'\mathbf{B}$. El tamaño de muestra es aleatorio dado que $n = \sum_{s_I} n_i$.

$\mathbf{B}$ puede estimarse mediante $\hat{\mathbf{B}} = \left(\sum_s \frac{\mathbf{x}_k\,\mathbf{x}_k'}{\pi_k\,\sigma_k^2}\right)^{-1}\left(\sum_s \frac{\mathbf{x}_k\,y_k}{\pi_k\,\sigma_k^2}\right)$. Los valores ajustados vienen dados por $\hat{y}_k = \mathbf{x}_k'\mathbf{B}$, y los residuos por: $e_{k_s} = y_k - \hat{y}_k$. De esta forma, el estimador de regresión queda determinado como:

$$\star\ \hat{t}_{y_{r_B}} = \sum_U \hat{y}_k + \sum_s \frac{e_{k_s}}{\pi_k} = \sum_{U_I}\sum_{U_i} \hat{y}_k + \sum_{s_I} \frac{1}{\pi_{I_i}} \sum_{s_i} \frac{y_k - \hat{y}_k}{\pi_{k|i}}$$

También podemos expresar el estimador como un estimador que corrige al estimador-π:

$$\star\ \hat{t}_{y_{r_B}} = \sum_U \hat{y}_k + \sum_s \frac{e_{k_s}}{\pi_k} = \sum_U \mathbf{x}_k'\hat{\mathbf{B}} + \sum_s \frac{y_k - \hat{y}_k}{\pi_k} =$$

$$= \sum_U \mathbf{x}_k'\hat{\mathbf{B}} + \sum_s \frac{y_k}{\pi_k} - \sum_s \frac{\hat{y}_k}{\pi_k} = \sum_U \mathbf{x}_k'\hat{\mathbf{B}} + \hat{t}_{y_\pi} - \sum_s \frac{\mathbf{x}_k'\hat{\mathbf{B}}}{\pi_k} =$$

$$= \hat{t}_{y_\pi} + \left(\sum_U \mathbf{x}_k' - \sum_s \frac{\mathbf{x}_k'}{\pi_k}\right)'\hat{\mathbf{B}} = \boxed{\hat{t}_{y_\pi} + \left(\mathbf{t}_x - \hat{\mathbf{t}}_{x_\pi}\right)'\hat{\mathbf{B}}}$$

Alternativamente, podríamos expresarlo como un estimador-π g-ponderado:

$$\star\ \hat{t}_{y_{r_B}} = \hat{t}_{y_\pi} + \left(\mathbf{t}_x - \hat{\mathbf{t}}_{x_\pi}\right)'\hat{\mathbf{B}} = \sum_s \frac{y_k}{\pi_k} + \left(\mathbf{t}_x - \hat{\mathbf{t}}_{x_\pi}\right)'\hat{\mathbf{T}}^{-1}\left(\sum_s \frac{\mathbf{x}_k\,y_k}{\pi_k\,\sigma_k^2}\right) =$$

$$= \sum_s \frac{y_k}{\pi_k}\underbrace{\left[1 + \left(\mathbf{t}_x - \hat{\mathbf{t}}_{x_\pi}\right)'\hat{\mathbf{T}}^{-1}\frac{\mathbf{x}_k}{\sigma_k^2}\right]}_{g_{k_s}} = \boxed{\sum_s \frac{g_{k_s}\,y_k}{\pi_k}}$$

Para derivar su varianza, es conveniente expresar el estimador en términos de su error de estimación:

$$\hat{t}_{y_{r_B}} - t_y = \underbrace{\sum_{s_I} \frac{t_{E_i}}{\pi_{I_i}} - \sum_{U_I} t_{E_i}}_{Q_{B_{s_i}}} + \underbrace{\sum_{s_I} \frac{1}{\pi_{I_i}}\left(\sum_{s_i} \frac{g_{k_{s_B}}\,E_k}{\pi_{k|i}} - \sum_{U_i} E_k\right)}_{R_{B_s}}$$

$$\text{donde}\ \ E_k = y_k - y_k^0 = y_k - \mathbf{x}_k'\mathbf{B}\ \ \text{y}\ ,\ t_{E_i} = \sum_{U_i} E_k$$

La varianza del estimador viene dada por:

$$\star\ AV\left(\hat{t}_{y_{r_B}}\right) = AV_{BPSU} + AV_{BSSU}$$

$$\star\ AV_{BPSU} = \sum\sum_{U_I} \Delta_{I_{il}} \frac{t_{E_i}}{\pi_{I_i}} \frac{t_{E_j}}{\pi_{I_j}}$$

$$\star\ AV_{BSSU} = \sum_{U_I} \frac{V_{E_i}}{\pi_{I_i}}$$

$$\star \; V_{E_i} = \sum\sum\nolimits_{U_i} \Delta_{kl|i} \frac{E_k}{\pi_{k|i}} \frac{E_l}{\pi_{l|i}}$$

Un estimador para la varianza del estimador viene dado por:

$$\star \; \hat{V}(\hat{t}_{y_{r_B}}) = \hat{V}_{BPSU} + \hat{V}_{BSSU}$$

$$\star \; \hat{V}_{BPSU} = \sum\sum\nolimits_{s_I} \Delta^{\checkmark}_{I_{il}} \frac{\hat{t}_{E_i}}{\pi_{I_i}} \frac{\hat{t}_{E_j}}{\pi_{I_j}} - \sum\nolimits_{s_I} \frac{1}{\pi_{I_i}} \left(\frac{1}{\pi_{I_i}} - 1 \right) \hat{V}_{B_{E_i}}$$

$$\star \; \hat{V}_{BSSU} = \sum\nolimits_{s_I} \frac{\hat{V}_{B_{E_i}}}{\pi^2_{I_i}}$$

$$\star \; \hat{V}_{B_{E_i}} = \sum\sum\nolimits_{s_i} \Delta^{\checkmark}_{kl|i} \frac{g_{k_{s_B}} e_{k_s}}{\pi_{k|i}} \frac{g_{l_{s_B}} e_{l_s}}{\pi_{l|i}}$$

$$\text{donde} \;\; \hat{t}_{E_i} = \sum\nolimits_{s_i} \frac{g_{k_{s_B}} e_{k_s}}{\pi_{k|i}}$$

Caso C

En el caso C la información auxiliar es conocida solo para las unidades pertenecientes a las PSU que fueron sorteadas en la primera etapa. Se asume que el scatter de puntos puede ser presentado por el modelo:

$$E_\xi(y_k) = \mathbf{x}_k' \boldsymbol{\beta} \qquad V_\xi(y_k) = \sigma_k^2$$

donde $\boldsymbol{\beta}$ es estimado por $\mathbf{B} = \left(\sum_U \frac{\mathbf{x}_k \mathbf{x}_k'}{\sigma_k^2} \right)^{-1} \left(\sum_U \frac{\mathbf{x}_k \, y_k}{\sigma_k^2} \right)$. Los residuos vienen dados por: $E_k = y_k - y_k^0 = y_k - \mathbf{x}_k' \mathbf{B}$. El tamaño de muestra es aleatorio dado que $n = \sum_{s_I} n_i$.

$\mathbf{B}$ puede estimarse mediante $\hat{\mathbf{B}} = \left(\sum_s \frac{\mathbf{x}_k \mathbf{x}_k'}{\pi_k \sigma_k^2} \right)^{-1} \left(\sum_s \frac{\mathbf{x}_k \, y_k}{\pi_k \sigma_k^2} \right)$. Los valores ajustados vienen dados por $\hat{y}_k = \mathbf{x}_k' \mathbf{B}$, y los residuos por: $e_{k_s} = y_k - \hat{y}_k$. De esta forma, el estimador de regresión queda determinado como:

$$\star \; \hat{t}_{y_{r_C}} = \sum\nolimits_{s_I} \frac{\hat{t}_{y_{i_r}}}{\pi_{I_i}} \;\; \text{donde} \;\; \hat{t}_{y_{i_r}} = \sum\nolimits_{U_i} \hat{y}_k + \sum\nolimits_{s_i} \frac{y_k - \hat{y}_k}{\pi_{k|i}} \Rightarrow$$

$$\Rightarrow \hat{t}_{y_{r_C}} = \sum\nolimits_{s_I} \frac{t_{\hat{y}_i}}{\pi_{I_i}} + \sum\nolimits_{s_I} \frac{1}{\pi_{I_i}} \sum\nolimits_{s_i} \frac{y_k - \hat{y}_k}{\pi_{k|i}} = \sum\nolimits_{s_I} \frac{t_{\hat{y}_i}}{\pi_{I_i}} + \sum\nolimits_{s} \frac{y_k - \hat{y}_k}{\pi_k}$$

$$\text{donde} \;\; t - \hat{y}_i = \sum\nolimits_{U_i} \hat{y}_k$$

Su expresión como estimador-π g-ponderado es:

$$\star \; \hat{t}_{y_{r_C}} = \sum\nolimits_{s} \frac{g_{k_{s_C}} \, y_k}{\pi_k} \;\; \text{donde} \;\; g_{k_{s_C}} = 1 + \left[\sum\nolimits_{s_I} \frac{\mathbf{t}_{x_i} - \hat{\mathbf{t}}_{x_{i_\pi}}}{\pi_{I_i}} \right]' \hat{\mathbf{T}}^{-1} \frac{\mathbf{x}_k}{\sigma_k^2}$$

Para derivar su varianza es conveniente expresar el estimador de regresión en términos de su error de estimación:

$$\hat{t}_{y_{r_C}} - t_y = \underbrace{\sum\nolimits_{s_I} \frac{t_{y_i}}{\pi_{I_i}} - \sum\nolimits_{U_I} t_{y_i}}_{Q_{C_{s_I}}} + \underbrace{\sum\nolimits_{s_I} \frac{1}{\pi_{I_i}} \left(\sum\nolimits_{s_i} \frac{g_{k_{s_C}} E_k}{\pi_{k|i}} - \sum\nolimits_{U_i} E_k \right)}_{R_{C_s}}$$

La varianza del estimador viene dada por:

$$\star \; AV(\hat{t}_{y_{r_C}}) = AV_{CPSU} + AV_{CSSU}$$

$$\star \ AV_{CPSU} = \sum\sum_{U_I} \Delta_{I_{ij}} \frac{t_{y_i}}{\pi_{I_i}} \frac{t_{y_j}}{\pi_{I_j}}$$

$$\star \ AV_{CSSU} = \sum_{U_I} \frac{V_{E_i}}{\pi_{I_i}}$$

$$\star \ V_{E_i} = \sum\sum_{U_i} \Delta_{kl|i} \frac{E_k}{\pi_{k|i}} \frac{E_l}{\pi_{l|i}}$$

$$\text{donde} \ \ t_{y_i} = \sum_{U_i} y_k$$

Un estimador para la varianza del estimador viene dado por:

$$\star \ \hat{V}(\hat{t}_{y_{r_C}}) = \hat{V}_{CPSU} + \hat{V}_{CSSU}$$

$$\star \ \hat{V}_{CPSU} = \sum\sum_{s_I} \Delta^{\checkmark}_{I_{ij}} \frac{\hat{t}_{y_{i_\pi}}}{\pi_{I_i}} \frac{\hat{t}_{y_{j_\pi}}}{\pi_{I_j}} - \sum_{s_I} \frac{1}{\pi_{I_i}} \left(\frac{1}{\pi_{I_i}} - 1 \right) \hat{V}_i$$

$$\star \ \hat{V}_{CSSU} = \sum_{U_I} \frac{\hat{V}_{C_{E_i}}}{\pi^2_{I_i}}$$

$$\star \ \hat{V}_i = \sum\sum_{s_i} \Delta^{\checkmark}_{kl|i} \frac{y_k}{\pi_{k|i}} \frac{y_l}{\pi_{l|i}}$$

$$\star \ \hat{V}_{C_{E_i}} = \sum\sum_{s_i} \Delta^{\checkmark}_{kl|i} \frac{g_{k_{s_C}} e_{k_s}}{\pi_{k|i}} \frac{g_{l_{s_C}} e_{l_s}}{\pi_{l|i}}$$

$$\text{donde} \ \ \hat{t}_{y_{i_\pi}} = \sum_{s_i} \frac{y_k}{\pi_{k|i}}$$