TEMA 4. INFERENCIA A PARTIR DE MODELOS DE SUPERPOBLACIÓN.

1 Introducción.

La teoría clásica en muestreo de poblaciones finitas asume que la población finita que se investiga es una "población fija". Más explí citamente, se trabaja en el contexto de una población que consta de N unidades bien identificadas y una variable de estudio y con valores fijados $y_1, \ldots, y_N$, a priori desconocidos, asociados respectivamente a las unidades $1, \ldots, N$. Se asume tácitamente que las unidades poblacionales están realmente etiquetadas y que el estadístico tiene acceso al verdadero valor y_i de la i–ésima unidad mediante una encuesta.

Cuando se trabaja con poblaciones reales algunas de estas básicas consideraciones no están totalmente justificadas y hay que proceder a una revisión de los problemas de inferencia y de los métodos. Así han ido desarrollándose otras aproximaciones a la inferencia en poblaciones finitas.

Desde un punto de vista teórico la aproximación por poblaciones fijas no permite establecer estrategias óptimas. La no existencia de estimadores uniformemente de mínima varianza de la media en la clase de estimadores insesgados, en la clase de estimadores lineales insesgados y en la clase de estimadores lineales homogéneos (este último para diseños no unicluster) motivó a los estadísticos en el estudio de poblaciones finitas.

Dadas dos estrategias muestrales, generalmente no se puede establecer ningún resultado acerca de cuál de ellas es "mejor". Entonces la formulación de un modelo que estructure los valores de la variable en estudio, puede resultar positivo para concluir algunos resultados. Esta es una de las razones para proponer modelos de superpoblación en el estudio de la comparación de estrategias muestrales para poblaciones finitas. Así estos modelos se pueden ver como un instrumento para intentar obtener resultados más concretos al comparar estrategias y producir estrategias óptimas en varias

situaciones. Ya *Cochran* (1946) propuso una estructuración de los valores poblacionales para poder comparar la eficacia del muestreo sistemático frente al muestreo aleatorio simple y al muestreo estratificado.

Aunque el desarrollo de modelos de superpoblación es reciente, éstos fueron utilizados en los años cuarenta por diversos autores como *Cochran* (1939,46), *Deming* y *Stephan* (1941) y *Madow* y *Madow* (1944).

Existen no obstante diferentes opiniones sobre la justificación de la inferencia basada en modelos: algunos autores (*Neyman*, 1971) creen que deben construirse considerando sólo la aleatorización impuesta por el diseño muestral. Otros (*Barnard*, 1971; *Kalbfleich* y *Sprott*, 1969; *Royall*, 1970/71) consideran la inferencia basada en modelos no sólo deseable sino necesaria.

2 El concepto de modelo de superpoblación.

Vamos a introducir a continuación la noción de modelo de superpoblación, que es inherente a una población finita dada.

Supongamos cierto centro universitario formado por N edificios numerados de 1 a N, y se está interesado en conocer el total de dinero gastado en teléfono por la Universidad. La naturaleza de la variable a estudiar es tal que parece poco real concentrar el problema sólo en el año de referencia, y estimar el parámetro relevante. En cambio se puede considerar la naturaleza estocástica de la variable en estudio e intentar predecir el gasto total en un determinado tiempo o estimar el gasto medio sobre todo el dominio del tiempo. Podemos ver así los valores de la variable de interés como variables aleatorias $Y_1, \ldots, Y_N$ que tienen cierta distribución conjunta $\xi\left(Y_1, \ldots, Y_N\right)$ y considerar que los valores actuales (desconocidos a menos que se haga una encuesta) $y_1, \ldots, y_N$ constituyen una realización de las variables $Y_1, \ldots, Y_N$, respectivamente. La distribución $\xi(\cdot)$ va a determinar el modelo de superpoblación.

El modelo de superpoblación proporciona un resumen de nuestro conocimiento a priori de la naturaleza de la variable a estudiar en referencia a la población. Este conocimiento puede estar basado en una larga experiencia o en un juicio subjetivo personal.

3 Principales modelos de superpoblación.

Mediante un modelo de superpoblación se dan un conjunto de condiciones que definen la clase de distribuciones a la que pertenece. Estas especificaciones pueden ser muy incompletas (por ejemplo, fijando sólo los dos primeros momentos) o por el contrario más detalladas (como que tengan una distribución normal con medias y covarianzas conocidas).

Vamos a denotar por E_ξ, V_ξ, Cov_ξ a la media, varianza y covarianza de las variables aleatorias generadas por el modelo $\xi(\cdot)$, mientras que E_p, V_p y Cov_p, o (E, V y Cov si no hay confusión) a los operadores basados en el diseño muestral.

En particular, para $k, l = 1, \ldots, N$ notamos

$$\mu_k = E_\xi\left(Y_k\right),$$

$$\sigma_k^2 = V_\xi\left(Y_k\right),$$

$$\sigma_{kl} = Cov_\xi\left(Y_k, Y_l\right), k \neq l, \text{ y}$$

$$\overline{\mu} = \frac{1}{N}\sum_{k=1}^{N}\mu_k.$$

A continuación vamos a ver algunos de los modelos más usados en la práctica

3.1 Modelos Generales.

3.1.1 Modelo G_{MR} (de regresión múltiple).

La clase de medidas de probabilidad en R_N tales que Y_1, ..., Y_N son independientes y

$$\mu_k = E_\xi\left(Y_k\right) = \beta_1 + \sum_{i=2}^{q}\beta_i x_{k_i},\ k = 1, \ldots, N, \text{ y}$$
$$\sigma_k^2 = E_\xi\left(Y_k - \mu_k\right)^2 = \sigma^2 v_k,\ k = 1, \ldots, N,$$

donde $\beta_1, \ldots, \beta_q, \sigma^2$ son desconocidos, y $x_{k_2}, \ldots, x_{k_q}, v_k$ es un conjunto de números conocidos para cada k ($k = 1, \ldots, N$). ($x_1, \ldots, x_k$ son variables auxiliares conocidas)

Según este modelo se tiene que:

$$y_k = \mu_k + e_k = \beta_1 + \sum_{i=2}^{q} \beta_i x_{k_i} + e_k$$

Si $i = 1$ y $v_k = 1$ se tiene el modelo de regresión simple:

$$y_k = \mu_k + e_k = \beta_0 + \beta_1 x_k + e_k,$$

El caso especial del modelo G_{MR} donde la regresión es lineal y pasa por el origen, con una variable auxiliar sólo, se considerará separadamente.

3.1.2 Modelo G_R.

La clase de medidas de probabilidad $\mathcal{P}$ en R_N tal que $Y_1, \ldots, Y_N$ están independientemente distribuidas y

$\mu_k = E_\xi(Y_k) = \beta x_k,\ k = 1, \ldots, N$, y

$\sigma_k^2 = V_\xi(Y_k) = \sigma^2 v(x_k)\ \ k = 1, \ldots, N,$

β y σ^2 son desconocidos, $v(\cdot)$ es una función conocida y $x_1, \ldots, x_N$ números positivos conocidos. Es decir:

$$y_k = \beta x_k + e_k$$

Una suposición común en los modelos G_R es que $\sigma_k^2 = \sigma^2 v(x_k) = \sigma^2 x_k^\delta$ donde δ es conocido (el valor $\delta = 0$ corresponde a errores homocedásticos).

Cochran (1953, p 212) y *Brewer* (1936) mantienen que valores de δ entre 1 y 2 son más frecuentes que los valores fuera de este intervalo. En general se asume que el intervalo de interés es $0 \leq \delta \leq 2$.

3.1.3 Modelo G_{MRE} (de regresión estratificado).

Consideremos el modelo G_{MR} para el caso en que la población esté estratificada en L subgrupos disjuntos (estratos) de tamaño N_h, $h = 1, \ldots, L$. $(N = \sum_{h=1}^{L} N_h)$. Las unidades vienen etiquetadas por dos subíndices, h y k, denotando el estrato y la unidad dentro de este, $k = 1, \ldots, N_h$ y $h = 1, \ldots, L$. Los estratos se construyen partiendo la población de acuerdo a los valores de la variable auxiliar x. Las N_1 unidades del primer estrato corresponden a los N_1 menores valores de x, las N_2 unidades del segundo estrato corresponden a los N_2 siguientes menores valores de x y así sucesivamente.

Considerando así las medidas de probabilidad en R_N tales que Y_{hk} son independientes y

$$\mu_{hk} = \beta_{h1} + \sum_{i=2}^{q} \beta_{hi} x_{hk_i}, \text{ y}$$
$$\sigma^2_{hk} = \sigma^2_h v_{hk},$$

Es decir:

$$y_{hk} = \beta_{h1} + \sum_{i=2}^{q} \beta_{hi} x_{hk_i} + e_{hk}$$

Este modelo fué introducido por *Pfeffermann* (1984).

4 Principios de la estimación basada en modelos

En la aproximación basada en el diseño, la inferencia sobre la media $\overline{y}$ se desarrolla bajo el criterio de insesgadez bajo el diseño. La aleatorización impuesta sobre el estadístico a través del diseño muestral usado es esencial en la determinación de la estrategia óptima.

Por contra, en la aproximación predictiva las predicciones consideradas están justificadas únicamente por el modelo asumido. La insesgadez bajo el diseño no se considera necesaria. Este punto de partida es contrario a la mayoría de los clásicos de encuestas, que consideran que la aleatorización del diseño muestral es esencial en el proceso de estimación, y justifican que las inferencias basadas en el modelo no serán válidas si el modelo asumido no es real. Sin

embargo, es innegable que el modelo predictivo puede tener un gran valor potencial para la práctica de las encuestas por muestreo.

5 Definiciones: Predictores, estimadores, insesgadez.

Definición 5.1 *Un estadístico T es una función de los datos tal que para cada valor dado s de S, depende de $Y_1, \ldots, Y_N$ sólo a través de los Y_k, $k \in s$.*

Definición 5.2 *Un predictor de $\overline{Y}$ es un estadístico T usado para hacer inferencia acerca de $\overline{Y}$. El valor de T para $S = s$ e $Y_k = y_k$, $k \in s$ se denotará por t y es el estimador de $\overline{y}$.*

Definición 5.3 *Llamamos esperanza, varianza y error cuadrático medio bajo el diseño (o p–esperanza, p–varianza y p–error cuadrático medio) de un predictor T, a las funciones de las variables aleatorias*

$$E(T) = \sum_{\mathcal{S}} p(s)T,$$
$$V(T) = \sum_{\mathcal{S}} p(s)\left(T - E(T)\right)^2, \text{ y}$$
$$ECM(T) = \sum_{\mathcal{S}} p(s)\left(T - \overline{Y}\right)^2.$$

Definición 5.4 *Llamamos esperanza y varianza bajo el modelo (o ξ–esperanza y ξ–varianza) del predictor T a*

$$E_\xi(T)) = \int T d\xi,$$

$$V(T) = \int \left(T - E_\xi(T)\right)^2 d\xi.$$

Definición 5.5 *Decimos que un predictor T es p–insesgado para $\overline{Y}$ si y sólo si, para un diseño dado p, $E(t) = \overline{y}$, $\forall y = (y_1, \ldots, y_N) \in R_N$ siendo t el valor de T obtenido para $Y_k = y_k$, $k \in S$.*

Definición 5.6 *Una estrategia (p, T) se dice que es p–insesgada si T es un predictor p–insesgado para ese diseño.*

Definición 5.7 *Diremos que T es ξ–insesgado de $\overline{Y}$ si y sólo si para una distribución ξ dada, $E_\xi\left(T-\overline{Y}\right)=0$, $\forall s\in\mathcal{S}$.*

Definición 5.8 *Diremos que T es $p\xi$–insesgado de $\overline{Y}$ si y sólo si para p y ξ dados, $E_\xi E\left(T-\overline{Y}\right)=0$, $\forall s\in\mathcal{S}$.*

Observación 5.9 *Por conveniencia, a veces se dice que T es un estimador ξ–insesgado de $E_\xi\left(\overline{Y}\right)=\overline{\mu}$ si T es un predictor ξ–insesgado de $\overline{Y}$. Similarmente se dice que T es un estimador $p\xi$–insesgado de $\overline{\mu}$ si T es un predictor $p\xi$–insesgado de $\overline{Y}$.*

Observación 5.10 *Para cada diseño dado p y un modelo dado ξ, la clase de predictores $p\xi$–insesgados contiene a la clase de predictores p–insesgados y a la clase de predictores ξ–insesgados. Claramente si t es un estimador p–insesgado de $\overline{y}$, al imponer cualquier modelo de superpoblación, es un estimador $p\xi$–insesgado de $\overline{\mu}$ y un predictor $p\xi$–insesgado de $\overline{Y}$.*

6 Predicción bajo el modelo G_R

Consideramos el modelo de regresión básico G_R, asumiendo que $x_k>0$ son conocidos en toda la población. El siguiente resultado es debido a Brewer(1963) y a Royall (1970), los cuales derivan el estimador lineal y homogéneo óptimo para $\overline{Y}$.

Teorema 6.1 *Bajo el modelo G_R, el predictor lineal ξ-insesgado óptimo (predictor ξ-BLUE) viene dado para cualquier diseño d, por:*

$$T_{BR}=f_S\overline{Y}_S+(1-f_S)\widehat{\beta}\overline{x}_{\overline{s}}$$

siendo

$$\widehat{\beta}=\frac{\sum_s\frac{x_kY_k}{v(x_k)}}{\sum_s\frac{x_k^2}{v(x_k)}}$$

Nota. Si $v(x)=x$, el predictor óptimo es el clásico predictor de razón

$$T_R=\frac{\overline{x}\overline{Y}_s}{\overline{x}_s}$$

y

$$E_\xi E(T_R - \overline{Y})^2 = \frac{\overline{x}E(\frac{N\overline{x}}{\sum_s x_k} - 1)\sigma^2}{N}$$

El predictor T_R es ξ-insesgado y puede ser también insesgado bajo el diseño (p.e. si se utiliza el diseño de Lahiri-Midzuno-Sen).

7 Predicción bajo el modelo de regresión múltiple

Los resultados anteriores se pueden extender al caso de varias variables auxiliares. Así suponiendo el modelo G_{MR}, donde Y_1, ..., Y_N son independientes y

$$\mu_k = E_\xi(Y_k) = \beta_1 + \sum_{i=2}^{q} \beta_i x_{k_i},\ k = 1, \ldots, N, \text{ y}$$
$$\sigma_k^2 = E_\xi(Y_k - \mu_k)^2 = \sigma^2 v_k,\ k = 1, \ldots, N,$$

donde $\beta_1, \ldots, \beta_q, \sigma^2$ son desconocidos, y $x_{k_2}, \ldots, x_{k_q}, v_k$ es un conjunto de números conocidos para cada k ($k = 1, \ldots, N$), ($x_1, \ldots, x_k$ son variables auxiliares conocidas).

Notamos por

$E_\xi(Y_s) = X_s\beta$, $E_\xi(Y_{\overline{s}}) = X_{\overline{s}}\beta$, $V_\xi(Y_s) = \sigma^2\Sigma_s$, $V_\xi(Y_{\overline{s}}) = \sigma^2\Sigma_{\overline{s}}$, $COV_\xi(Y_s, Y_{\overline{s}}) = 0$ donde Y_s e $Y_{\overline{s}}$ son vectores columna. Además $\beta' = (\beta_1, ..., \beta_q)$ es un vector de parámetros desconocidos. Las matrices X_s y $X_{\overline{s}}$ son de orden $\nu(s) \times q$ y $(N - \nu(s)) \times q$ respectivamente. El vector fila de X_s correspondiente a la unidad k, $x_k' = (x_{k1}, .., x_{kq})$ es tal que $x_{k1} = 1$ y $x_{k2}, ..., x_{kq}$ son los valores para las variables auxiliares $x_2, ..., x_q$ en la unidad k. Las matrices Σ_s y $\Sigma_{\overline{s}}$ son $\nu(s) \times \nu(s)$ y $(N - \nu(s)) \times (N - \nu(s))$ respectivamente.

El siguiente resultado da el predictor ξ-BLU para el modelo que estamos considerando.

Teorema 7.1 *Bajo el modelo G_{MR} el predictor ξ-BLU para $\overline{Y}$ dado un diseño d, viene dado por:*

$$T_{BLU} = f_S\overline{Y}_S + (1 - f_S)m_{\overline{s}}'\widehat{\beta}_{BLU}\overline{x}_{\overline{s}}$$

donde $m'_{\bar{s}} = (m_{\bar{s}1}, ..., m_{\bar{s}q})$, $m_{\bar{s}i} = \frac{\sum_{\bar{s}} x_{ki}}{N-\nu(s)}$

$$\widehat{\beta}_{BLU} = (X'_s \Sigma_s^{-1} X_s)^{-1} (X'_s \Sigma_s^{-1} Y_s).$$

Además $E_\xi E(T_{BLU} - \overline{Y})^2$ *viene dada por la esperanza bajo el diseño de*

$$E_\xi (T_{BLU} - \overline{Y})^2 = \frac{(\sum_{\bar{s}} v_k)\sigma^2}{N^2} + (1 - f_s)(m'_{\bar{s}} X'_s \Sigma_s^{-1} X_s)^{-1} m_{\bar{s}})\sigma^2$$

8 ¿Inferencia basada en el diseño o inferencia basada en el modelo?

En los últimos años una cuestión muy debatida entre los especialistas en muestreo de poblaciones finitas es el uso o no de modelos de superpoblación. En el muestreo en poblaciones finitas la experiencia de muchos años y trabajos dice que la inferencia debe estar basada fundamentalmente en la distribución resultante de la aleatorización de la muestra más que en los modelos estadísticos. Hoy día sigue existiendo una gran controversia entre los partidarios de la inferencia basada en el diseño y los partidarios de la inferencia basada en el modelo.

La inferencia basada en el diseño tiene una serie de fortalezas frente a la basada en el modelo que hace que sea más popular entre los expertos en realización de encuestas. Este enfoque toma en cuenta la información proporcionada por el diseño muestral y proporciona buenas estimaciones para grandes muestras, sin necesidad de imponer fuertes restricciones de modelización de las variables. Por otro lado, es esencialmente asintótica y entonces su uso para muestras pequeñas puede no ser adecuado. Sin embargo en muestras grandes (como las que realizan los organismos oficiales) suele dar estimaciones con menor error que las basadas en el modelo.

En cuanto a la inferencia basada en el modelo tiene también sus ventajas:

- proporciona una aproximación unificada para la inferencia más de acorde con las principales teorías estadísticas en otras áreas,

- permite utilizar la metodología bayesiana, que no tiene cabida en la aproximación clásica,

- si el modelo está bien especificado pude dar lugar a estimaciones más precisas que las basadas en el diseño. Esto suele ocurrir en los estudios en poblaciones pequeñas, donde hay más información auxiliar de la población y recursos para su estudio con lo que se puede definir y comprobar bien el modelo de superpoblación.

La mayor debilidad de la inferencia basada en el modelo es que si el modelo está mal definido, puede dar lugar a inferencias mucho peores que las basadas en el diseño. La no robustez de los estimadores basados en el modelo es una de las razones utilizadas para preferir la inferencia basada en el diseño.

Algunos expertos en muestreo utilizan ambas filosofías, de acuerdo con el contexto. Por ejemplo, la inferencia acerca de parámetros lineales basados en grandes muestras se realiza mediante métodos basados en el diseño, mientras que los modelos son usados para problemas donde esta aproximación no funciona bien, como en la estimación con falta de respuesta o en áreas pequeñas.

En los últimos años hay intentos de reconciliar ambas teorías. Así ha aparecido un nuevo enfoque: la inferencia basada en el diseño modelo asistida.

En el enfoque modelo asistido la inferencia está basada en el diseño igual que en la aproximación clásica (el objetivo es minimizar el error cuadrático medio bajo el diseño) pero se utilizan modelos de superpoblación para motivar la elección del estimador; en particular los estimadores clásicos que incorporan información auxiliar (como los estimadores de razón y de regresión) pueden motivarse utilizando modelos lineales de superpoblación.

En los últimos años han aparecido dos metodologías importantes para el uso de información auxiliar: el método de calibración introducido por Deville y Särndal (1992) y el de verosimilitud empírica introducido por Chen y Qin (1993). Estos estimadores con buenas propiedades teóricas y prácticas son modelo asistidos pues utilizan modelos para explicar la relación entre la variable objeto de estudio y la variable auxiliar.

9 Estimadores modelo asistidos

En las encuestas por muestreo es común que la característica que queremos estudiar y, se encuentra fuertemente relacionada con una característica auxiliar x, y además, los datos sobre x son conocidos, o se pueden recoger fácilmente para todas las unidades de la población. En estas situaciones cabe plantearse cómo utilizar esta información auxiliar para mejorar las estimaciones en el sentido de construir nuevos estimadores que para el mismo tamaño muestral tengan menor error de estimación (lo que implicaría mayor precisión en las estimaciones de los parámetros), o equivalentemente que tengan el mismo error que los ya conocidos pero con un menor tamaño muestral (lo que produciría una disminución en el coste de la realización de la encuesta).

Los primeros métodos utilizados para incorporar la información auxiliar en la fase de estimación, son los llamados métodos indirectos de estimación, entre los que destacan los conocidos métodos de razón, de diferencia y de regresión. Estos estimadores no siempre garantizan que se produzca una disminución del error de muestreo respecto a los estimadores que no usan información auxiliar. La precisión que ganamos con los métodos que emplean la información auxiliar estará en función del buen uso de las hipótesis que se usen para emplear un procedimiento u otro, y el que dichas hipótesis se ajusten en mayor o menor medida al problema real.

Una alternativa a estos métodos la propuso Deville y Särndal (1992) a través de los estimadores de calibración o los estimadores de verosimilitud empirica.

9.1 Estimador de regresin generalizado

En esta seccin se construye un estimador del total poblacional que mejora sustancialmente en eficiencia al incorporar información auxiliar mediante el supuesto de que las variables de información auxiliar estn relacionadas con la variable de inters mediante un modelo lineal general que le da el nombre al estimador.

Suponiendo el modelo de regresión donde $y_1, \ldots, y_N$ son independientes y

$\mu_k = E_\xi(y_k) = \sum_{i=1}^{J} \beta_i x_{k_i},\ k = 1, \ldots, N$, y
$\sigma_k^2 = E_\xi(Y_k - \mu_k)^2 = \sigma^2 v_k,\ k = 1, \ldots, N$,
las variables ser relacionan mediante:

$$y_k = \sum_{i=1}^{J} \beta_i x_{k_i} + E_k$$

Supongamos que queremos estimar el total poblacional de la variable y, es decir, queremos estimar

$$T_y = \sum_{k \in U} y_k$$

y que $\mathbf{x} = (\mathbf{x_1}, \ldots, \mathbf{x_J})$, es un vector de variables auxiliares de forma que es perfectamente conocido para todos los elementos de la población U.

El ajuste mínimo cuadrtico de los coeficientes de regresión se obtiene a partir del problema clsico de minimización:

$$min_\beta \sum_{k \in U} \frac{(y_k - \mathbf{x}'\beta)^{\mathbf{2}}}{\sigma_k^2}$$

Derivando respecto al vector β e igualando a cero se obtiene el estimador

$$\widehat{\beta}_s = \sum_{k \in s} \frac{\mathbf{x_k y_k}}{\sigma_k^2 \pi_k} (\sum_{k \in s} \frac{\mathbf{x_{x'_k}}}{\sigma_k^2 \pi_k})^{-1} \tag{1}$$

A partir de estos valores se obtiene el estimador general de regresión:

$$\widehat{T}_{yreg} = \sum_{k \in s} d_k y_k + (T_{\mathbf{x}} - \widehat{T}_{\mathbf{XH}}) \cdot \widehat{B}_s \tag{2}$$

siendo
$\sum_s d_k \mathbf{x_k} = \widehat{\mathbf{T}}_{\mathbf{XH}}$ y $T_{\mathbf{x}} = (T_{x_1}, \ldots, T_{x_J})$ y $d_k = 1/\pi_k$ los pesos del diseo.

9.2 Estimadores de calibración

Supongamos que queremos estimar el total poblacional de la variable y, es decir, queremos estimar

$$T_y = \sum_{k \in U} y_k$$

y que $\mathbf{x} = (\mathbf{x_1}, \ldots, \mathbf{x_J})$, es un vector de variables auxiliares de forma que es perfectamente conocido para todos los elementos de la población U, entonces si consideramos el estimador de *Horvitz-Thompson* para T_y, es decir

$$\widehat{T}_{YH} = \sum_{k \in s} d_k y_k$$

lo que pretendemos es modificar los pesos $d_k = \dfrac{1}{\pi_k}$, por otros pesos ω_k, de forma que el estimador basado en dichos pesos, proporcione estimaciones perfectas para $\mathbf{x}$, es decir:

$$\sum_s \omega_k \mathbf{x_k} = \mathbf{T_x} = (\mathbf{T_{x_1}}, \ldots, \mathbf{T_{x_J}}) \tag{3}$$

y estén tan próximos como sea posible, respecto a una distancia dada, a los pesos originales d_k.

Usualmente la distancia elegida es la suma ponderada de cuadrados de las distancias, que viene dada por

$$\sum_{k \in s} \frac{(\omega_k - d_k)^2}{q_k d_k} \tag{4}$$

donde q_k son constantes positivas, entonces tenemos el siguiente problema:

- Minimizar $\displaystyle\sum_{k \in s} \frac{(\omega_k - d_k)^2}{q_k d_k}$ sujeto a la condición $\displaystyle\sum_{k \in s} \omega_k \mathbf{x_k} = \mathbf{T_x}$

Utilizando el método de los multiplicadores de Lagrange se obtienen los siguientes pesos calibrados:

$$\omega_k = d_k + d_k q_k \lambda \mathbf{x_k'} \tag{5}$$

donde

$$\lambda = T_s^{-1}(T_{\mathbf{x}} - \widehat{T}_{\mathbf{XH}})$$

supuesto que la inversa de

$$T_s = \sum_{k \in s} d_k q_k \mathbf{x_k} \mathbf{x_k}^{'}$$

existe, y siendo $\widehat{T}_{\mathbf{XH}}$, el estimador de *Horvitz-Thompson*, para el vector de variables auxiliares $\mathbf{x}$.

El estimador calibrado obtenido, así construido, viene dado por:

$$\widehat{T}_{yreg} = \sum_{k \in s} \omega_k y_k = \widehat{T}_{YH} + (T_{\mathbf{x}} - \widehat{T}_{\mathbf{XH}}) \cdot \widehat{B}_s \tag{6}$$

siendo

$$\widehat{B}_s = T_s^{-1} \cdot \sum_{k \in s} q_k \mathbf{x_k} \mathbf{y_k} \tag{7}$$

Dicho estimador, es el estimador general de regresión (véase Cassel, Särndal y Wretman 1976).

La forma de $\widehat{T}_{yreg}$ dependerá tanto del diseño muestral como de las constantes q_k elegidas. Por ejemplo, si trabajamos con una única variable auxiliar $\mathbf{x} = \mathbf{x_1}$, como q_k elegimos $q_k = \dfrac{1}{x_k}$ y trabajamos con un muestreo aleatorio simple entonces

$$\widehat{T}_{yreg} = \frac{\widehat{T}_{YH}}{\widehat{T}_{XH}} \cdot T_x$$

que es el estimador de razón.

En general, el estimador $\widehat{T}_{yreg}$, no es insesgado, pero como los pesos ω_k están próximos a d_k, es asintóticamente insesgado.
La varianza asintótica del estimador general de regresión viene dada aproximadamente, por

$$AV(\widehat{T}_{Yreg}) = \sum_{k \in U} \sum_{k \in U} \Delta_{kl} \Big(d_k E_k \Big) \Big(d_l E_l \Big) \tag{8}$$

donde $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$ y $E_k = y_k - \mathbf{x_k'B}$, siendo

$$B = \left(\sum_{k \in U} \mathbf{x_k x_k' q_k} \right)^{-1} \cdot \sum_{\mathbf{k} \in \mathbf{U}} \mathbf{q_k x_k y_k}$$

Un estimador, para esta varianza, viene dado por:

$$\widehat{V_1}(\widehat{T}_{Yreg}) = \sum_{k \in s} \sum_{l \in s} \Delta_{kl} d_{kl} (d_k e_k)(d_l e_l) \tag{9}$$

siendo $d_{kl} = \dfrac{1}{\pi_{kl}}$ y $e_k = y_k - \mathbf{x_k'} \widehat{\mathbf{B}}_\mathbf{s}$, donde $\widehat{B}_s$ viene dado por (**??**), (para un estudio detallado de estos resultados puede consultarse *Särndal*, *Swensson*, *Wretman* 1989).
Otro posible estimador para la varianza de $\widehat{T}_{yreg}$ viene dado por:

$$\widehat{V_2}(\widehat{T}_{Yreg}) = \sum_{k \in s} \sum_{k \in s} \Delta_{kl} d_{kl} (\omega_k e_k)(\omega_l e_l) \tag{10}$$

donde los pesos calibrados ω_k sustituyen a los pesos originales d_k. Este estimador suele resultar más adecuado que el anterior, cuando trabajamos con un modelo de superpoblación apropiado (véase *Deville* y *Särndal* 1992), por ejemplo:

$$E_\xi(y_k) = \beta' \mathbf{x_k} \qquad \text{y} \qquad \mathbf{V}_\xi(\mathbf{y_k}) = \sigma_\mathbf{k}^\mathbf{2}$$

A la vista de estos resultados, el caso ideal, aunque utópico, se daría cuando la relación entre las variable de interés y y las variables auxiliares es la siguiente:

$$y_k = \beta \mathbf{x_k'} \qquad \mathbf{k} \in \mathbf{U}$$

siendo β un vector de constantes. En tal caso la estimación coincide con el verdadero valor, es decir, $\widehat{T}_{yreg} = T_y$ y la varianza de $\widehat{T}_{yreg}$ es nula.